PLIEGO DE PRESCRIPCIONES TÉCNICAS PARA LA CONTRATACIÓN, POR PROCEDIMIENTO NEGOCIADO CON PUBLICIDAD, DEL SUMINISTRO DE AMPLIACIÓN DEL SISTEMA DE SUPERCOMPUTACIÓN HETEROGÉNEO QUE ADQUIRIÓ EL BSC-CNS EN EL AÑO 2011 MEDIANTE EL EXPEDIENTE DE CONTRATACIÓN NÚMERO CONSU02010018OP.

Requisitos Técnicos de actualización de Cluster Heterogéneo

En junio del año 2011 se adquirió un equipo de computación heterogéneo que ha permitido ofrecer servicios específicos a ciertos grupos de investigación que, con la utilización adecuada de los procesadores gráficos, pueden obtener mejor rendimiento en las ejecución de las aplicaciones con un menor coste energético que en la utilización de procesadores de propósito general.

De la misma forma, desde el BSC, en sus departamentos de investigación, se desarrollan herramientas para la utilización óptima de estos equipos, reduciendo así los costes globales.

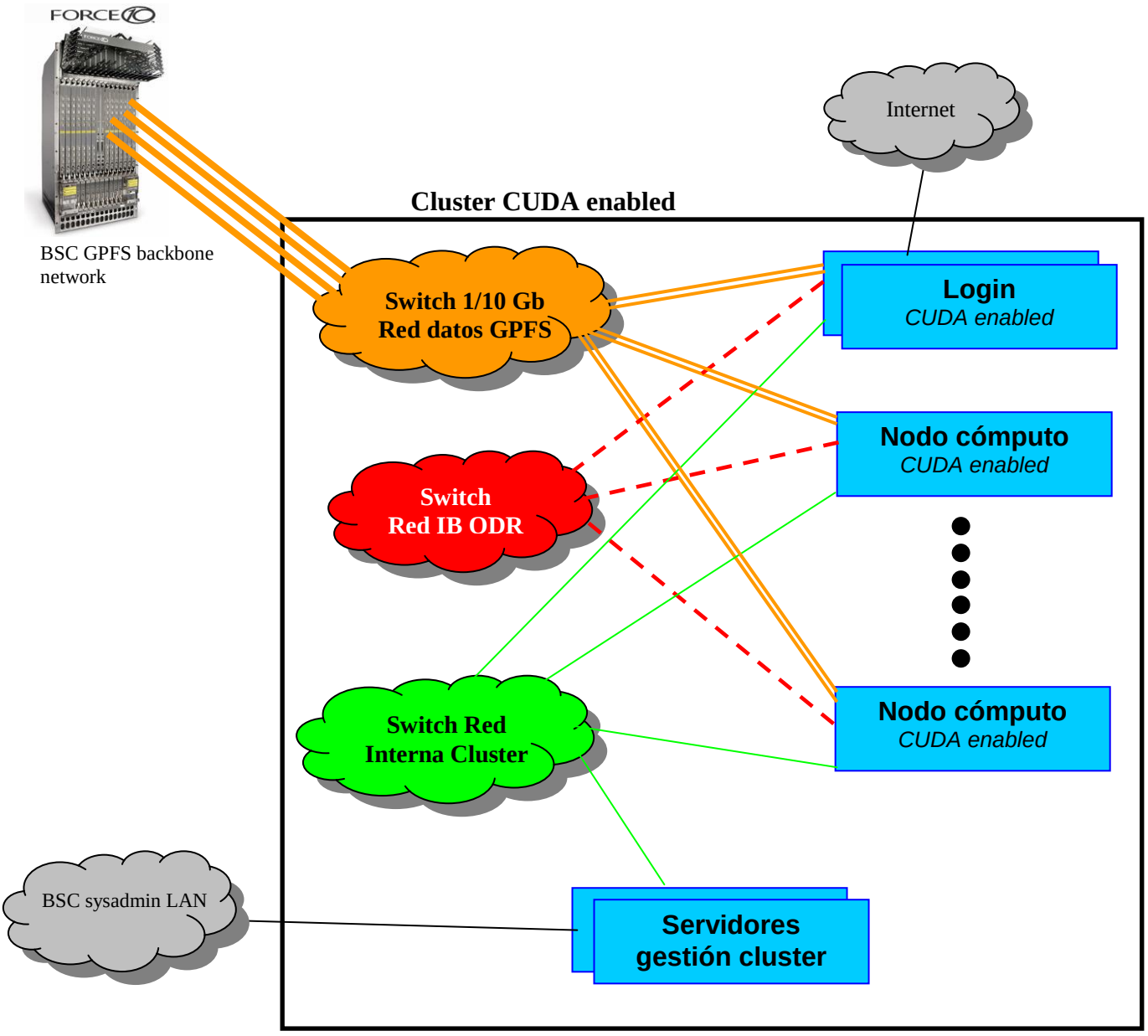
La disponibilidad de estos equipos permitirá el desarrollo de estos proyectos internos que darán soporte a proyectos de usuarios externos, en tiempo adecuado para resolver los problemas científicos que se quieren analizar. Este equipo ha permitido la participación del BSC en proyectos internacionales, que buscan dotar a Europa de computación de nivel superior al Petaflop, es decir, Exaflop.

Las nuevas tecnologías han hecho que estos procesadores gráficos hayan evolucionado, haciendo necesario poner a disposición de los investigadores equipos dotados de los nuevos K40/K80.

Se propone la ampliación del sistema heterogéneo con procesadores K40/K80 aprovechando la infraestructura existente, que incluye sistema de comunicaciones, sistema de alimentación y sistema de gestión.

Introducción

A continuación podemos ver un esquema con los componentes y redes de la configuración actual del cluster de aceleradoras gráficas que el BSC posee y desea ampliar:



A nivel general la ampliación se puede resumir en los siguientes grandes puntos:

- Reestructuración de los nodos de cómputo actuales para disponer de espacio para la instalación de los nuevos elementos de la actualización
- Inclusión de 45 nuevos nodos de cómputo con 2 GPUs K80 cada uno de ellos
- Actualización del cluster y del software de gestión a la última versión para poder disponer de las últimas ventajas de CUDA y de los entornos de programación que permiten las nuevas GPUs K80
- Desplazamiento de los racks actuales del cluster 3 posiciones a la izquierda en el CPD para dejar sitio a otras instalaciones
- Recableado y reconexionado de los racks con la nueva configuración en las nuevas posiciones en el CPD
- Conexionado de la ampliación a la actual red de gestión y red Infiniband de MPI y red de GPFS
- Reducción del cluster actual a un mínimo de 7 chasis, reinstalando el sistema operativo con imagen de la nueva parte del cluster

Por cada entrada con requisitos de las tablas siguientes se han clasificado como:

R – Representa que lo anunciado es un requerimiento que se debe cumplir en la solución presentada

D – Representa un requerimiento deseable a tener, se valorarán positivamente aquellas soluciones que lo incorporen

1. Descripción Hardware

1.1 Nodos

Ref	Descripción
R1	Mínimo de 45 nodos de cómputo con las siguientes características mínimas: - 2 procesadores x86_64 con 8 cores cada uno con 2.4 GHz de frecuencia nominal - 64 GB RAM - 1 disco SSD 240GB para sistema operativo - 2 tarjetas NVIDIA K80
R2	Cada nodo de cómputo deberá tener las siguientes interfaces de red para conectarse con el resto de componentes del cluster: - Tarjeta IB FDR (56 Gbit/s) - Mínimo de 3 interfaces Gigabit Ethernet (1 red interna y 1 red GPFS) - Interfaz Ethernet de gestión out-of-line, dicha interfaz deberá soportar el Standard IPMIv.2.0
D3	Se valorará mejoras en el número de nodos como en las características técnicas de cada nodo. Todos los nodos de cómputo ofertados deberán ser idénticos.
R4	Cada nodo de cómputo deberá de ofrecer los buses y anchos de banda óptimos para la conexión de los componentes de GPU e Infiniband.
R5	Las tarjetas IB FDR de los nodos deberán disponer de tecnologías optimizadas para tecnologías GPU en conjunto con Infiniband. Como por ejemplo, NVIDIA GPU Direct.
R6	Los nodos de cómputo ofertados para la ampliación 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, …), generación de alarmas, etc. Y deberá ser compatible con los sistemas ya en funcionamiento en el cluster actual.
R7	En el caso que con la ampliación se necesitara algún elemento extra de administración, se debería incluir dicho componente, o reconfiguración de alguno existente.
R8	Los nodos de cómputo actuales se deberán reducir a un mínimo de 7 chasis activos e integrados con la ampliación del cluster. Los chasis restantes se dejarán como material spare.

1.2. Switches y Redes

Ref	Descripción
R1	Se deberán de proveer de esquemas de conexionado de cada una de las redes que conforman el cluster una vez este instalada la ampliación
D2	En todos los switches y redes se valorará el proveer un 5% de puertos libres

Ref	Descripción
Red Interna cluster	
R3	Se deberá proveer del hardware necesario (switches, cables, etc.) para poder ampliar el cluster actual en la red interna de management del cluster. Dichos componentes deberán ser compatibles con los actuales del cluster. De la misma manera se deberá reconfigurar aquellos elementos ya existentes para soportar la ampliación.
Red IB-FDR	
R4	Se deberá proveer del hardware necesario (switches, cables, etc.) para poder establecer la red interna de alto rendimiento basada en Infiniband FDR (56 Gbit/s) entre los nuevos nodos de cómputo. Se deberá de interconectar esta nueva red Infiniband con la anterior para que tener una única visión de la red Infiniband a nivel de cluster entre los componentes actuales y la ampliación.
R5	La red IB FDR deberá ser no bloqueante entre los nodos que conforman la ampliación del cluster. La red IB FDR podrá ser bloqueante con la antigua red IB QDR del cluster.
D6	Se valorará el mínimo número de switches utilizados para formar esta red
Red de datos GPFS	
R7	Se deberá proveer del hardware necesario (switches, GBICs, cables, etc.) para poder establecer la red interna de datos GPFS con tecnología 1/10 Gigabit Ethernet de los nuevos nodos que conforman la ampliación. Dichos componentes deberán ser compatibles con los actuales del cluster.
R8	Dicha red deberá ser ampliada para conectar todos los nodos de cómputo nuevos mediante los 2 links de 1 Gbit cada uno, y la parte anterior del cluster que se mantenga en producción (7 chasis).
R9	Después de la ampliación todos los puertos de la red de GPFS deberán ser line-rate (sin sobre-suscripción) en su conexión hacia el switch Force10 que aglutina las conexiones 10 Gbit Ethernet hacia el sistema de almacenamiento del BSC.
R10	En el caso que no hubiera suficientes GBICs de Force10 para la conexión al almacenamiento del BSC después de la ampliación se deberían de proporcionar los necesarios para cumplir con el requerimiento de line-rate.

Ref	Descripción
D11	Se valorará el mínimo número de switches utilizados para formar esta red

2. Operacional

Ref	Descripción
Requerimientos operacionales	
R1	Los racks actuales del cluster de GPU se deberán desplazar 3 posiciones a la izquierda dentro del CPD. Cualquier faena derivada de este movimiento deberá venir incluida en este proyecto (recableado, reconexión eléctrica, reconexión circuito de refrigeración, etc.)
R2	Se deberá presentar en la documentación un esquema con la ocupación de los racks de los diversos equipos presentados en la solución. En ella se deberá claramente especificar el hardware que quedará en funcionamiento del cluster original y la parte que conforma la ampliación.
R3	Cada uno de los racks deberá tener un consumo máximo de 28 Kw. por rack con todos los componentes funcionando. Los nuevos componentes deberán ser compatibles con los racks que actualmente alberga el cluster de GPUs del BSC.
R4	Los nuevos nodos deberán disponer igual que los nodos actuales de un sistema de monitorización de la temperatura de tal manera que provoque la parada controlada del sistema en caso de temperatura muy alta.
R5	Los nuevos nodos de cómputo y resto de componentes de la ampliación deberán disponer de fuentes de alimentación redundadas
R6	La solución deberá incluir el montaje en racks de toda la ampliación, recableado debido al movimiento de los racks o para la conexión de los nuevos elementos de la ampliación. La empresa instaladora se deberá hacer cargo de los residuos derivados de su instalación.
R7	La ampliación propuesta deberá ser instalada en los racks existentes del cluster

3. Software

Ref	Descripción
R1	El proyecto de ampliación debe incorporar las licencias y los servicios de actualización del sistema operativo tanto de la parte actual como de la ampliación a las últimas versiones soportadas de sistema operativo como de CUDA.
R2	Se deberá aportar todo el software necesario para la gestión de todos los componentes nuevos, siendo este compatibles con los software que actualmente se usa en el cluster actual.
D3	Se valorará que la solución de clustering soporte o sea una solución 'Diskless nodes', donde el sistema operativo no reside en la máquina local de cada nodo del cluster, sino que en un disco de los servidores de gestión del cluster.
R4	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, OpenCL y CUDA. A parte del entorno de programación Open-Source proporcionado, se deberá de proporcionar el entorno de programación específico para la arquitectura proporcionada (tanto Intel como para las aceleradoras): - Intel Compilers, NVIDIA CUDA toolkit para C y Fortran (de PGI) como mínimo
R5	Se deberán de proporcionar las librerías numéricas proporcionadas por el fabricante de los procesadores como para las aceleradoras debidamente optimizadas para dicha arquitectura. (Intel MKL y NVIDIA CUDA toolkit)
R6	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.
R7	Se deberán incluir el software y las licencias de GPFS cliente para todos los nuevos nodos de cómputo
R8	Se deberá de proveer de herramientas de profiling y debugging de aplicativos secuenciales y paralelos especializados para la arquitectura de la ampliación como puede ser: Intel Vtune Amplifier

4. Mantenimiento y soporte

Esta parte describe los requerimientos y ampliaciones referidos al mantenimiento y soporte de la solución a nivel global.

Ref	Descripción
R1	Garantía y soporte de 3 años en todos los nuevos componentes (hardware y software), 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.
D2	Se valorará la extensión de la garantía / mantenimiento tanto en años de duración como cobertura y tiempo de respuesta dentro del horario de oficina.
R3	El proyecto de instalación incluirá la comprobación del buen funcionamiento, integración y óptimo rendimiento de la solución.
R4	Se exigirá un trabajo en equipo con el equipo de sistemas del BSC, para la coordinación en la instalación, configuración del cluster y solución de cualquier problema de incompatibilidad que surja o implementación de la solución.
R5	Se proporcionará (dentro de período de garantía): - Acceso a todos los software upgrade (incluyendo sistemas operativos, clientes GPFS y firmware) de todos los componentes de la solución - Punto único de soporte para el aviso de problemas e incidencias de cualquier componente que componga la solución
R6	Se exigirá soporte pro-activo, notificando y recomendado subidas de versión tanto de software como de firmware de cualquier componente de la solución. Sobretodo referente a upgrades de CUDA y drivers de NVIDIA.
R7	Se deberá entregar al final de la instalación una documentación digital en la que se describa la situación final después de la ampliación: - Descripción general de los componentes de la solución - Esquema de conexionado e IPs y valores de configuración empleados - Explicación del proceso de instalación y tareas realizadas - Explicación procedimientos para: Puesta en marcha, y disaster recovery
R8	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.
R9	Se deberán ofrecer formación durante la instalación de la solución, que cubran: - Conceptos básicos - Administración básica y procedimientos básicos de configuración - Compilación y utilización de las aceleradoras gráficas CUDA - Optimización de la solución - Solución de problemas