Pliego de prescripciones técnicas para la adquisición de un cluster de computación de alto rendimiento con destino al Servei d'Informàtica de la Universitat de València.

1. Requerimientos mínimos y detalles técnicos

A continuación, se detallarán cada uno de los componentes del cluster y se especificarán los servicios que se requieren de la empresa adjudicataria. Lo expuesto a continuación representa los requerimientos mínimos con los que debe cumplir la oferta presentada, que no será valorada en caso de no hacerlo.

1.1) Nodos de computación del cluster.

El sistema estará formado por equipos denominados nodos que serán de dos tipos: thin y fat. Todos los nodos integrarán procesadores de propósito general, no considerándose de este tipo acelerados gráficos (GPU) o la familia de procesadores Xeon Phi (KNC, KNL, ...). Se requiere que todos los procesadores soporten al menos el juego de instrucciones AVX2 y proporcionen un mínimo de 16 FLOPs por ciclo en doble precisión. Cada nodo de cómputo ofertado sólo podrá ser thin o fat y todos los nodos de un mismo tipo serán idénticos entre sí. Los procesadores ofertados en los nodos thin y fat serán del mismo fabricante y de la misma generación, ya que se pretende lograr una solución homogénea.

Se requieren un número a determinar por el licitante de nodos **thin**, en ningún caso inferior **a 18,** que tendrán las siguientes características técnicas mínimas:

- 2 procesadores de propósito general por nodo, con un mínimo de 14 y un máximo de 20 cores por procesador y una frecuencia base mínima de 2,4 Ghz (no se tendrán en cuenta frecuencias de tipo TurboBoost).
- Al menos 4 GB/core de memoria principal con tecnología DDR4 ECC. Este valor debe tomarse como un valor de referencia mínimo. La configuración de memoria en cada socket deberá ser equilibrada, de forma que se usen todos los canales de acceso a memoria soportados por el procesador. No se considerarán válidas configuraciones de memoria en las que el ancho de banda efectivo de acceso quede afectado por una incorrecta distribución de los módulos de memoria por los canales del procesador. Todos los DIMMs con que contaran las máquinas deben ser de la misma marca, frecuencia y latencia. Como mínimo los 2 sockets deberán contar con cuádruple canal de acceso a memoria con los DIMMs a una frecuencia mínima de 2400Mhz.

Para los nodos **fat**, se requieren **dos (2)** nodos con las siguientes características:

- 4 procesadores de propósito general por nodo, con un mínimo de 14 cores por procesador y una frecuencia base mínima de 2,0 Ghz (no se tendrán en cuenta frecuencias de tipo TurboBoost). Los procesadores deberán contar con 3 interconexiones punto a punto.

- Al menos 1,5 TB de memoria RAM DDR4 ECC por nodo. Si la configuración lo permite, un máximo de 2 TB de memoria RAM. La configuración de memoria en cada socket deberá ser equilibrada, de forma que se usen todos los canales de acceso a memoria soportados por el procesador. No se considerarán válidas configuraciones de memoria en las que el ancho de banda efectivo de acceso quede afectado por una incorrecta distribución de los módulos de memoria por los canales del procesador. Todos los DIMMs con que contaran las máquinas deben ser de la misma marca, frecuencia y latencia. Como mínimo todos los sockets deberán soportar cuádruple canal de memoria con los DIMMs a una frecuencia mínima de 1600 Mhz.

Se requiere que todos los nodos (thin y fat) cumplan además con las siguientes características:

- Almacenamiento local con una capacidad mínima de 2 TB, 7200 rpm, con interfaz SATAIII o superior. Cualquier disco ofertado será extraíble en caliente (hot swap) y de calidad Enterprise.
- Fuentes de alimentación redundantes con certificación 80 PLUS Platinum o superior.
- Tarjeta de conexión a una red de baja latencia para el tráfico de datos hacia el sistema de ficheros y para las aplicaciones basadas en MPI. El ancho de banda mínimo teórico de esta red será 100Gbit/s (ver apartado de Conmutadores y redes).
- Una interfaz al menos de 1 Gbit Ethernet para conectar con la red de gestión.
- Una interfaz de gestión out-of-band. En el caso de que los nodos estén empaquetados en un chasis, esta interfaz podrá ser compartida por todos los nodos del mismo. Esta interfaz debe permitir como mínimo realizar las siguientes operaciones: power on/off, recoger consola gráfica, motorización del entorno hardware, generación de alarmas y actualización de firmware.

Una vez determinados el número de nodos thin y fat, se proporcionará la siguiente información:

Rpeak thin = (número de procesadores thin * frecuencia del procesador * número de cores/procesador * FLOPs/ciclo del procesador) / 1000

Rpeak fat = (número de procesadores fat * frecuencia del procesador * número de cores/procesador * FLOPs/ciclo del procesador) / 1000

Rpeak total = Rpeak thin + Rpeak fat

Dónde:

- frecuencia del procesador (Ghz): es la frecuencia base del procesador, sin contar con frecuencias de tipo TurboBoost.
- FLOPs/ciclo: será un mínimo según pliego de 16 en doble precisión, aunque hay procesadores que superan este valor, por lo que deberá ajustarse en función del procesador ofertado. Se corresponde con el número de operaciones en coma flotante que puede ejecutar el procesador ofertado por ciclo de reloj. En el caso en que la frecuencia del procesador disminuya al usar extensiones avanzadas de coma flotante esta circunstancia se ignorará por motivos de simplicidad.
- Número de cores/procesador: se tendrán en cuenta únicamente los cores, no los threads.

Se requiere que la potencia de cálculo teórica total (Rpeak total) calculada utilizando únicamente los nodos de computación (excluyendo específicamente el nodo de gestión, el de login y los servidores del almacenamiento) sea de al menos 35 Tflops.

1.2) Nodo de login y nodo de gestión

De forma adicional a los nodos de computación, se requiere un nodo para ser usado como nodo de login. Este nodo tendrá las mismas características técnicas que un nodo thin (aunque no es necesario que esté empaquetado en el mismo chasis que el resto), pero contará con una interfaz adicional de 10 Gbit Ethernet para la conexión con la red del SIUV.

Se precisa además un nodo adicional para la gestión del cluster. Este nodo puede considerarse crítico, por lo que contará con fuentes de alimentación redundantes extraíbles en caliente y 2 discos SAS de 10krpm o superior, configurados en RAID 1, ofreciendo un tamaño mínimo de 240 GB para el sistema operativo. Este servidor ejecutará los servicios necesarios para la gestión del cluster: DNS, TFTP, SLURM, NTP, LDAP, SSH, monitorización, herramientas de clustering, etc. Dispondrá además de suficiente espacio local para almacenar la imagen del sistema operativo de todos los nodos del cluster y lo exportará por NFS. Este disco local estará configurado al menos con RAID5 o superior para asegurar la protección de los datos. Este servidor también contará con dos interfaces 10 Gbit Ethernet para la conexión con la red del SIUV y la red de gestión.

Se requiere que ambos nodos tengan las siguientes características comunes:

- Interfaz de gestión out-of-band completamente licenciada. Esta interfaz debe permitir como mínimo realizar las siguientes operaciones: power on/off, recoger consola gráfica, motorización del entorno hardware, generación de alarmas y actualización de firmware.
- Fuentes de alimentación redundantes con certificación 80 PLUS Platinum o superior.
- Cualquier disco ofertado será extraíble en caliente (hot swap) y de calidad Enterprise.

1.3) Conmutadores y redes

Se precisan varios tipos de redes para la interconexión de todos los elementos del sistema. Las denominaremos como red de baja latencia, red de gestión y red out-of-band. Será necesario además proporcionar la conexión a la red del SIUV. Se deberá proveer de esquemas de conexionado físico de cada una de la redes y describir el ancho de banda disponible a cada nivel.

- A menos que se especifique lo contrario, las redes descritas serán disjuntas a nivel físico. Todos los conmutadores dentro de una misma red serán del mismo fabricante y tendrán todos sus puertos habilitados a nivel de licencias de software, se usen o no. Los conmutadores tendrán fuentes de alimentación redundantes.
- Los conmutadores serán gestionables y por lo tanto esa interfaz será accesible desde la red de gestión.
- Se proporcionarán todos los cables necesarios para conectar los elementos ofertados y se etiquetarán de manera clara y duradera. Se incluirán también los transceptores ópticos necesarios para las conexiones que se describen.
- Todos los cables de una misma red tendrán entre ellos el mismo color, que será diferente entre todas las redes para poder identificarlas con facilidad.

- Red de baja latencia

Todos los nodos de cómputo, el login y el sistema de almacenamiento estarán conectados a esta red de interconexión. Se precisa una red de muy baja latencia y alto rendimiento, con un ancho de banda nominal de 100Gbps. La topología de esta red será fat tree, con un factor de sobresuscripción máximo de 2:1. La red debe ser de tipo offloading y no se aceptaran soluciones de conexión de tipo onloading (que requieren una mayor utilización de la CPU). Sobre esta red se encaminará el tráfico producido por las aplicaciones MPI y el tráfico del sistema de almacenamiento. El nodo de gestión no requiere acceso a esta red, pero podrá estar conectado si con ello se facilita la gestión de la misma.

- Red de gestión

Se proporcionará todo el hardware necesario para crear una red interna del cluster con tecnología de al menos 1 Gbit Ethernet. A esta red irán conectadas todos los equipos del cluster, incluidos los servidores de la red de almacenamiento. El servidor de gestión irá conectado a esta red mediante una de sus dos interfaces de 10 Gbit Ethernet mediante cobre o fibra, proporcionando el licitante el equipamiento necesario para realizar esta conexión. Sobre esta red se encaminará todo el tráfico de gestión del cluster e irá conectado con un uplink de 10 Gbit a un conmutador de la red de almacenamiento del SIUV para poder tener acceso al directorio home de los usuarios mediante NFS. Se usarán para ello los transceptores SFP+ necesarios.

- Red out-of-band

Esta red puede compartir hardware con la red de gestión, aunque en ese caso se utilizarán VLANs para dividir los dominios broadcast. A esta red irán conectadas todas las interfaces de gestión out-of-band de todos los equipos que formen la solución (incluidos el almacenamiento y login). Se incluyen aquí las interfaces IPMI ya mencionadas, los conmutadores, PDU's, etc. Sólo será accesible desde el servidor de gestión, que estará conectado mediante una interfaz de al menos 1 Gbit Ethernet o compartiendo, mediante VLAN tagging, la conexión 10 Gbit Ethernet de la red de gestión. El puerto de gestión out-of-band del servidor de gestión se conectará a la red de gestión del SIUV con un cable de la tecnología adecuada. La longitud estimada de este cable será de unos 15 metros.

- Conexión con la red del SIUV

Para dotar de conectividad externa al cluster, se solicita el suministro de un conmutador que como mínimo disponga de 16 puertos 10 Gbit Ethernet y 2 uplinks 10 Gbit Ethernet. Por motivos de compatibilidad con la electrónica de red existente en el SIUV deberá soportar los protocolos CDP (Cisco Discovery Protocol) y VTP (Vlan Trunking Protocol). A esta red irán conectados los nodos de login y gestión y los servidores de almacenamiento a 10 Gbps. Se utilizarán ambos uplink para la conexión con la red del SIUV, con los correspondientes transceptores SFP+.

1.4) Sistema de almacenamiento

Se deberá proporcionar un sistema de almacenamiento escalable y paralelo para el cluster, basado en el sistema de ficheros OpenSource Lustre Community. El sistema de almacenamiento no compartirá hardware con el sistema de gestión o los nodos de cálculo, pudiendo funcionar de forma totalmente independiente al resto de componentes.

Deberán tener acceso a este sistema de ficheros todos los nodos de cómputo, así como el nodo de login. No es necesario el acceso del nodo de gestión. El acceso a este sistema de ficheros se

producirá de dos formas. La forma principal mediante la red de baja latencia. Todos los nodos de computación de la solución accederán a este almacenamiento a través de esta red, por lo que se tendrán en cuenta los enlaces necesarios en la misma. Para otros equipos ya existentes en el SIUV, se configurará un acceso mediante Ethernet para el sistema de almacenamiento (con tecnología 10 Gbit).

Queda a discreción del licitador ofrecer una implementación adecuada, que se adapte a las necesidades de la solución propuesta. Se proporcionará un esquema de la propuesta, que deberá maximizar el ancho de banda de acceso al sistema de ficheros tanto el lectura como en escritura. Se requiere que la solución propuesta cumpla las siguientes características.

- Se espera un ancho de banda de al menos 3 GB/s en escritura y 5 GB/s en lectura en los nodos de cómputo. Se justificará adecuadamente el ancho de banda de la solución propuesta.
- La solución estará basada completamente en discos duros, quedando totalmente excluido el uso de cintas. Los discos serán extraíbles en caliente (hot swap), incluidos los discos del sistema operativo de los servidores, y se usarán discos de gama profesional o "Enterprise", preparados para funcionar 24x7 en un entorno de centro de datos y habilitados para su uso en entornos con gran cantidad de discos.
- Se requiere tolerancia a fallos en todo el sistema de almacenamiento, por lo que no debe existir un punto único de fallo. Esto implica la redundancia, que no duplicidad, de los elementos comunes de fallo como discos y fuentes de alimentación, así como de los propios servidores implicados en la gestión del sistema de ficheros. Se instalarán al menos dos servidores MDS y dos servidores OSS.
- El almacenamiento físico de metadatos (MDT) tendrá espacio suficiente para el almacenamiento de los metadatos necesarios para el sistema propuesto, configurado en RAID10 o tecnologías similares con discos SAS 10krpm o superiores. Este almacenamiento deberá ser visible por ambos servidores MDS para poder implementar una solución HA activa/pasiva.
- A nivel del almacenamiento físico de datos (OST), se espera una configuración que utilice cabinas de almacenamiento accesibles simultáneamente por ambos servidores OSS, de forma que pueda implementarse una solución HA activa/activa. La cabina empleará RAID6 (8+2) o tecnologías similares para la redundancia en el sistema de almacenamiento. La reconstrucción del RAID en caso de fallo y la sustitución de los discos fallidos por discos nuevos no debe afectar a la producción del sistema de almacenamiento. Los discos ofertados serán Near-Line SAS de 7.2 krpm o superiores. Las cabinas será ampliable al menos hasta 120 discos.
- El sistema de almacenamiento deberá proporcionar 190 TB netos (después de descontar los discos de paridad y antes de formatear el sistema de ficheros).
- Se requiere la instalación y configuración de este sistema de almacenamiento según los requisitos establecidos por el SIUV.

Una vez finalizada la instalación, se pedirá a la empresa adjudicataria que demuestre el rendimiento proporcionado usando las herramientas de benchmark típicas (como Iozone o Bonnie++).

1.5) Instalación física

Todos los equipos ofertados deberán ser instalados en los armarios (racks) que sean necesarios. Los racks serán de 42U de alto y contarán con puertas microperforadas por delante y detrás. Los racks serán instalados en el CPD del SIUV, situado en el campus de Burjassot. Si la solución propuesta no cabe completamente en un rack, se instalará en dos racks en función de los siguientes criterios:

- Sistema de almacenamiento, nodos de gestión y nodos de login en un armario.
- Resto de equipos en el otro armario.

En cualquier caso, los huecos libres que queden en el/los rack/s serán tapados mediante el uso de tapas o caratulas ciegas sin tornillos, de forma que se asegure un correcto flujo del aire a través de los equipos de computación. El flujo del aire será front-to-back y se instalarán todos los equipos de forma que se cumpla con este propósito. Se instalarán el número de PDU's adecuadas para conectar todos los equipos de la solución propuesta. La configuración de las mismas deberá proporcionar redundancia en la circunstancia de la caída del 50 % de las PDU's de un rack, estando ese 50% conectado a una acometida eléctrica distinta de la del restante 50%. Todas las PDU's serán trifásicas, con conectores IEC 60309 (conector rojo) o similar y se tendrá que proporcionar el número total para prever su instalación en el CPD.

El cableado de los racks estará ordenado y etiquetado y se dispondrá de forma que no impida el acceso a ningún componente hardware ni el flujo del aire de refrigeración. Será posible extraer cualquier pieza del cluster que sea necesario sustituir sin tener que desconectar nada más que el componente afectado.

Al finalizar la instalación física, se arrancarán todos los equipos y se comprobará que todos funcionan correctamente. Todo el firmware asociado al hardware será actualizado a la última versión disponible ofrecida por el fabricante.

1.6) Software

El cluster HPC objeto de este pliego debe funciona por completo con software OpenSource. Este requisito persigue tres objetivos básicos: minimizar el coste en licencias, compatibilidad total con otras infraestructuras HPC que tenemos y tendremos instaladas en el SIUV e independencia de la funcionalidad del cluster de posibles renovaciones futuras de licencias. No se valorarán positivamente soluciones que incluyan licencias comerciales del sistema operativo o software propietario de clustering. Así pues, se requiere una instalación software usando los siguientes paquetes de software:

- A nivel de sistema operativo, toda la solución propuesta debe funcionar con CentOS, en su última versión.
- Se requiere de la empresa adjudicataria proveer una solución software para la gestión a nivel hardware de los equipos ofertados. Este software debe permitir realizar la monitorización del hardware (recogida de datos del entorno, consumo eléctrico, estado de los equipos, generación de alertas, integración con software de monitorización tipo nagios, etc...) y la actualización del firmware de los equipos en remoto. El software será instalado en el nodo de gestión y se incluirá su funcionamiento en las sesiones de transferencia de conocimientos.
- Como ya se ha mencionado, a nivel del sistema de almacenamiento se instalará Lustre Community. Se instalará la última versión disponible y se configurará el acceso desde la red de baja latencia y la red Ethernet.
- A nivel de gestión del cluster, se ha elegido xCat como herramienta OpenSource para el despliegue de imágenes del cluster y el gestor de recursos OpenSource Slurm para la gestión de la carga de trabajo del cluster. Este software es el usado el SIUV para estas tareas, por lo que se requiere su instalación por motivos de compatibilidad con la infraestructura existente. Se requiere de la empresa adjudicataria que instale y configure el xCat para una instalación "statelite" del cluster.

- A nivel de monitorización de estado software de los nodos se usará ganglia o similar y un sistema de generación de alertas por e-mail, tipo nagios o similar.

La empresa adjudicataria se coordinará con los técnicos del SIUV para realizar la instalación, proporcionando éstos todos los datos necesarios para la misma, como la definición de las subredes a utilizar, ayuda para la configuración de los conmutadores de acceso, software a instalar, alertas a configurar, etc. Se solicita un periodo de soporte de 3 meses de la solución software instalada, contados una vez aceptado el suministro. Durante este periodo, la empresa adjudicataria proporcionará un teléfono o e-mail en que se ofrecerá soporte sobre la solución, incluidas actualizaciones de software en caso de ser necesario debido a bugs.

1.7) Garantía y mantenimiento

Se solicitan 5 años de garantía y mantenimiento de todos los componentes hardware de la solución propuesta en las siguientes condiciones:

- El soporte será ofrecido por el fabricante del hardware.
- La comunicación de la incidencia se producirá por teléfono o e-mail, preferiblemente en castellano, aunque se aceptará inglés.
- El soporte será del tipo 5 x 9 Next Business Day (NBD). Significando este nivel que, una vez identificado el problema, la pieza de reemplazo se recibirá como tarde al siguiente día laborable. La reparación será on-site, por lo que en caso de ser necesario, un técnico se desplazará a las instalaciones del SIUV para sustituir la pieza afectada.
- Los portes de envío y devolución (si procede) de los recambios y piezas defectuosas correrán a cargo del proveedor de servicios de soporte.
- Todos los equipos suministrados y los reemplazos posteriores durante el periodo de garantía deberán ser de nueva fabricación, rechazándose equipos acondicionados o de segunda mano.

La empresa que resulte ser la adjudicataria proporcionará los mecanismos de contacto necesarios para obtener el soporte hardware durante el periodo de garantía de todos los componentes de la solución, así como una relación de los números de serie de todos los equipos incluidos y su localización en el/los rack/s.

No obstante los términos generales de la garantía, el SIUV podrá considerar defectuosos aquellos productos o sus componentes (y sus reemplazos subsiguientes) que a lo largo de su vida productiva demuestren tasas de fallo juzgadas anormales. En esos casos extraordinarios, el adjudicatario aceptará realizar las actuaciones de subsanación que se detallan a continuación, sin coste adicional para el SIUV:

- Reemplazo del 100% de las unidades de aquellos productos que resulten de diseño defectuosos, entendiéndose como defectuoso que el 10% o más de las unidades suministradas de dicho producto fallan catastróficamente durante los dos primeros meses de uso, contando a partir de la fecha de recepción del equipamiento.
- Reemplazo de los equipos que resulten defectuosos de fábrica (5 o más de sus componentes de cualquier tipo fallan durante los 6 primeros meses de uso).

- Reemplazo del 100% de las unidades de componentes defectuosos (20% o más de las unidades instaladas del mismo modelo fallan durante cualquier periodo de 6 meses a lo largo del periodo de cobertura de la garantía).

En cualquier caso, los reemplazos deberán siempre realizarse con la aprobación previa del SIUV y por unidades de un modelo compatible y de prestaciones iguales o mejores que las del producto original.

2. Transferencia de conocimientos.

Una vez finalizada la instalación y configuración de los sistemas, la empresa adjudicataria impartirá varias sesiones de transferencia de conocimientos a los técnicos del SIUV encargados de la gestión futura del nuevo equipamiento científico. Estas sesiones tienen como objetivo garantizar la correcta operación de la infraestructura a lo largo de su vida útil. En estas sesiones se deberán tratar los siguientes aspectos básicos:

- Presentación de los componentes instalados en la solución. Mantenimiento del firmware de todos los componentes. Uso del software de gestión del hardware del licitante y de las interfaces de gestión out-of-band.
- Configuración y gestión del sistema de almacenamiento propuesto por el licitante. Software de gestión de las cabinas de almacenamiento. Optimización de la solución propuesta. Generación de los paquetes de instalación en los clientes.
- Configuración y gestión de la red de baja latencia. Detección y solución de problemas.
- Gestión de incidencias hardware y software a través del servicio de atención al cliente de la empresa licitante.
- Despliegue de imágenes del cluster.
- Monitorización de la solución.
- Cualquier otro tema de interés sobre la solución propuesta por el licitante.

La duración mínima de estas sesiones será de 20 horas en total. Se deberá entregar un plan de estas sesiones, adaptando el contenido de las mismas a la solución propuesta por el licitante, que incluya la duración de cada sesión. Para facilitar la transferencia de conocimiento, se entregará la documentación digital necesaria. Las sesiones tendrán lugar en las dependencias del SIUV una vez finalizada la instalación física, en castellano o inglés.

José María Ibáñez Cabanell
Catedrático de Universidad
Fecha: 15 / 05 / 2017