Universidad del País Vasco Euskal Herriko Unibertsitatea

DOC 5

PLIEGO DE PRESCRIPCIONES TÉCNICAS

I.- OBJETO DEL CONTRATO

El objeto del contrato es el suministro de nuevos nodos de cómputo, de un sistema de almacenamiento de alto rendimiento y de servidores de gestión para la ampliación del entorno de cálculo intensivo (HPC Arina) del servicio general de informática aplicada a la investigación adscrito a los SGIker de la UPV/EHU.

II.- DESCRIPCIÓN DETALLADA DEL SISTEMA

El sistema se entregará "llave en mano", esto es, con todo el hardware/software ofertado enrackado, instalado, configurado y funcionando, los nodos instalados y funcionando, al igual que los elementos de interconectividad. Se incluirán racks de 42U, de unos 200 cm de alto, 75cm de ancho y 120 cm de fondo, con guías de 19' y con robustez acorde a la carga que va a contener, PDUs, ruedas y soportes para fijarlos y con los paneles ciegos necesarios para evitar la recirculación de aire en el interior del rack. El cableado de comunicaciones saldrá por la parte superior del rack y el de potencia por debajo. El software de gestión de la solución deberá estar también instalado y configurado.

Los servidores tendrán como sistema operativo la última versión de la distribución estable de Rocky Linux.

Se incluirá un software para la monitorización del estado y rendimiento de los elementos ofertados como Icinga/Nagios o Zabbix. Se entregará configurada la monitorización del estado del equipamiento up/down, estado de ocupación de los sistemas de ficheros, funcionamiento de los discos, RAM, carga de red, carga del sistema, rendimiento IO del sistema de archivos de alto rendimiento, demonios del sistema. Se monitorizarán todos los elementos críticos para el funcionamiento del clúster.

Se incluirá configurado el software necesario para la creación de imágenes del sistema operativo de los nodos en producción y su posterior despliegue.

La solución estará integrada e interconectada con los recursos ya existentes a través de 2 conexiones SFP+ 10 Gb/s (como mínimo) a los puertos QSFP+ del switch ethernet Lenovo ThinkSystem NE1072T. Se proveerá de los cables y los conectores necesarios para esta conexión. La integración deberá de incluir la posibilidad de acceder a los otros 3 nodos de gestión actualmente instalados y a su grupo de nodos de cálculo correspondiente a través de los protocolos ssh e IPMI. Cada uno de estos 3 nodos administran el acceso y la gestión a un grupo de nodos de cálculo Intel de diferentes familias y generaciones, es decir, los nodos Xeon E5-2680 v2, Xeon E5-2680 v4 y Xeon Gold 6230 que montan sistema operativo RHEL 7.2, 7.2 y 7.7, respectivamente. Todos los nodos tienen IP fija y está instalado un DHCP en el servidor de gestión más reciente.

Se incluirá la solución configurada con la última versión del sistema de colas SLURM, que tendrá que integrar todos los recursos (i.e., los nodos de cálculo) ya existentes en el servicio. El sistema de colas SLURM se montará en el nuevo servidor de gestión, cuyo acceso estará restringido solo a los administradores del cluster. Se ajustará un sistema redundante del mismo SLURM en el servidor que hospeda ahora el sistema de colas SLURM, de manera que se active automáticamente ante una caída del nuevo servidor de colas.

Como punto de acceso se utilizarán los tres servidores actuales, ajustándose una única IP virtual con un sistema de repartición de la carga de los accesos por el software HAProxy (con algoritmo de balance Round Roby).

Se incluirá el servicio LDAP para la gestión del personal usuario, integrándolo a la situación actual del servicio y/o migrando el servicio actual al nuevo nodo de gestión.

Se incluirá la configuración de todos los elementos por direcciones IPs fijas, con relativo ajuste del servicio DHCP, para favorecer la integración con los entornos HPC ya instalados en el servicio.

Toda la configuración de la solución HPC tendrá que ser completamente compatible con los recursos actuales sin interferir con el buen funcionamiento del sistema ya existente.

Se incluirá la electrónica de red de cálculo para nodos de cómputo y almacenamiento de altas prestaciones Infiniband HDR 200 Gb/s. El switch de Infiniband Mellanox tendrá mínimo 40 puertos de 200 Gb/s, mientras que los nodos se podrán conectar a 100 Gb/s a través de cables splitters, para permitir puertos libres para futuras ampliaciones.

Se incluirá la electrónica de red privada de servicio para la administración del clúster ethernet a 10 Gb/s, que se integrará con la ya existente como descrito anteriormente. Los switches serán enterprise class.

Se incluirá una electrónica de red privada de gestión para la administración del clúster ethernet a 1 Gb/s como mínimo. Por esta red se realizará la gestión del hardware (BMC o similares, etc.).

En el caso que la solución no llenase por completo el RACK, los switches Infiniband y Ethernet deberán estar situados con una disposición adecuada de cara a posibles futuras ampliaciones de nodos de cálculo.

Los nodos y servidores contarán con una interfaz de administración out-of-band: encendido/apagado, monitorización, consola de sesión remota, etc.

El sistema será apto para situarlo sobre un suelo técnico de losetas de 600x600mm que soportan una carga máxima distribuida de 1310 kg/m2. La solución HPC será apta para un sistema de refrigeración estándar de aire y las puertas del rack tendrán que ajustarse a este requisito (no puertas de cristal o compactas).

Todos los elementos críticos para el funcionamiento (nodos, servidores, cabinas de disco y switches) deberán tener redundantes los elementos críticos como fuentes de alimentación, ventiladores, etc.

Las fuentes de alimentación redundantes de cada equipo irán conectadas a PDUs independientes en el rack. Estos pares de PDUs, asimismo, se conectarán a dos líneas eléctricas independientes del CPD. Las PDUs serán como máximo de 32A monofásico con terminador CETAC.

En la instalación y trabajo en el CPD se deberán cumplir los criterios y buenas prácticas fijadas por la UPV/EHU. Se incluyen las NORMAS PARA TRABAJOS EN EL CPD que debe cumplir todo el personal que entra en el CPD, así como la información necesaria sobre los ACCESOS PARA CARGA Y DESCARGA DE BULTOS del CPD que deberán tener en cuenta para el transporte e instalación de los equipos. Ver Anexos I y II.

Los equipos de la ampliación montarán el sistema de archivos correspondiente a la partición /home y /software actual usando el protocolo NFS y conectándose directamente al servidor NetApp DM3000H a través de 2 puertos SFP+ a mínimo 10 Gb/s. Se proveerá de los cables y los conectores necesarios para conectarse a los dos puertos SFP+.

El cableado, ya sea eléctrico o de comunicaciones, ha de ajustarse, sin cocas, lo máximo a la distancia que cubre, evitando así sobrantes excesivos.

Los servidores de gestión serán compatibles con el software de copia de seguridad Networker y se realizarán pruebas de backup y recuperación.

El hardware (i.e., todos los elementos físicos de la solución como nodos, servidores, etc.) tendrá como mínimo 5 años de garantía. Esta garantía incluirá, como mínimo, sustituciones de piezas defectuosas o estropeadas, resolución de problemas y extracción de logs, bien a través de envío de piezas de repuesto para realizar reparaciones en el caso que las piezas sean del tipo "Customer Replaceable Unit" (CRU Nivel 1) y desplazamiento y mano de obra de un técnico in situ para todos los otros casos. Por la excepcionalidad de los equipos, se requiere que durante los 3 primeros años la garantía de el/los nodo/s con GPU tenga/n una cobertura de tipo "in-situ" para todos sus componentes. Todo lo indicado no tendrá ningún coste para la UPV/EHU.

Durante el periodo de garantía de 5 años, la contratista deberá proporcionar una bolsa de 300 horas para la

resolución de problemas en la configuración de la solución HPC.

El tiempo de respuesta a incidencias será como máximo de 8 horas en días laborables entre las 09:00 y las 17:00. El servicio de resolución de incidencias deberá estar provisto de una herramienta (CALL HOME) capaz de abrir incidencias de manera sencilla con el fabricante. El fabricante de los servidores dispondrá de una plataforma o página web donde el cliente podrá descargarse los diferentes firmwares del servidor durante toda la vida del mismo.

Las licencias privativas del software (BeeGFS y otros) tendrán una validez de 5 años, en el que se incluye el soporte técnico de las mismas para la resolución de incidencias ante problemas de configuración, rendimiento, mal funcionamiento y similares.

Se facilitará toda la documentación referente a las máquinas, sistema operativo, etc., que deberá ser entregada en formato electrónico a la entrega de la solución HPC, incluyendo un fichero con todas las IPs de conexión a todos los equipos, así como usuarios y claves y un esquema detallado de las redes de conectividad.

La contratista deberá impartir un curso de formación inicial para dos técnicos de gestión del entorno HPC Arina sobre "administración de la solución propuesta" con una duración mínima de 6 horas, que se realizará dentro de los primeros diez días laborables una vez finalizada la instalación. Esta formación incluirá, como mínimo, los siguientes aspectos: configuración y mantenimiento del software propuesto; monitorización de los elementos del sistema; creación y despliegue de imágenes; sistema de ficheros de alto rendimiento; configuración, monitorización y manejo de redes (e.g., IPs, DHCP, ...).

Pasados 6 meses desde la formación inicial se realizará otro curso de formación avanzada de, como mínimo, 2 horas de duración, para resolver cuestiones surgidas durante este periodo de uso de la solución HPC.

Estos cursos se podrán realizar tanto en las instalaciones de la UPV/EHU como en remoto.

III.- DESCRIPCIÓN DETALLADA DE LOS SERVIDORES

Nodos de cálculo

Se requiere una solución de nodos con un mínimo de las siguientes características por nodo:

- 2 procesadores de última generación a 2.0 GHz de frecuencia base (sin opción Turbo o Max Boost), 48 cores y 2 sockets (2P) (Intel® Xeon® Platinum 8468 o superiores, o AMD EPYC 9454 o superiores).
- Almacenamiento local de tecnología SSD (mixed use - enterprise class) de capacidad 480 GB.
- 4 GB por core de memoria RAM, con DIMMs del tipo DDR5 distribuidas de manera óptima y balanceada ocupando todos los canales de acceso; si con esos GB por core no se pudiera tener la RAM balanceada, optimizada y con todos los canales de acceso ocupados, se instalará como mínimo la cantidad inmediatamente superior de RAM que permita dicha configuración.
- Tarjeta de servicio Ethernet a 10 Gb/s.
- Tarjeta de servicio Ethernet a 1 Gb/s para IPMI.
- Tarjeta Infiniband HDR 100 Gb/s.

Nodos de cálculo con GPUs

Se requiere una solución de o 1 nodo con 6 GPUs (ampliable a 8 GPUs) o 1 nodo con 2 GPUs (ampliable a 4 GPUs) + 1 nodo con 4 GPUs, para un total de 6 GPUs NVIDIA A100 80GB. El/Los nodo/s de cálculo tendrán las siguientes características mínimas por nodo:

- 2 procesadores iguales a los montados en los "Nodos de cálculos". Estos procesadores tienen que ser compatibles con los requisitos de potencia impuestos por las GPUs, si no pudiera ser posible por motivos de configuración del servidor se montarán procesadores justo inferiores en cores (manteniendo las otras características mínimas de los "Nodos de cálculos", incluso generación y familia).
- Las tarjetas gráficas tendrán que ser NVIDIA A100 80GB.
- Las tarjetas GPUs tendrán que estar conectadas con NVLink a pares.
- En el caso de un único nodo con 6 GPUs, el servidor tiene que estar ya predispuesto para una posible ampliación futura hasta alcanzar el valor máximo de 8 GPUs. En el caso de dos servidores, uno con 2

GPUs y uno con 4 GPUs, el primero tiene que estar ya predispuesto para una posible ampliación futura hasta alcanzar el valor máximo de 4 GPUs
- Almacenamiento local de tecnología SSD (mixed use - enterprise class) de capacidad de 480 GB.
- 1.5 TB de memoria RAM en el caso de un nodo con 6 GPUs; 1 TB de memoria RAM para cada nodo en el caso de la configuración con dos nodos (uno con 2 GPUs y otro con 4 GPUs). En ambos casos se incluirá con DIMMs del tipo DDR5 distribuidas de manera óptima y balanceada ocupando todos los canales de acceso; si con esos GB por core no se pudiera tener la RAM balanceada, optimizada y con todos los canales de acceso ocupados, se instalará como mínimo la cantidad inmediatamente superior de RAM que permita dicha configuración.
- Tarjeta de servicio Ethernet a 10 Gb/s.
- Tarjeta de servicio Ethernet a 1 Gb/s para IPMI.
- Tarjeta Infiniband HDR 100 Gb/s.

Servidor de Gestión

Se requiere un servidor de gestión de dos U, con las siguientes características mínimas:
- 2 procesadores de última generación a 2.9 GHz de frecuencia base (sin opción Turbo o Max Boost) y 16 cores (Intel® Xeon® Gold 6444Y o superiores, o AMD EPYC 9124 o superiores). Estos tienen que ser de la misma generación y familia que los exigidos en la sección anterior "Nodos de cálculo".
- 192 GB de memoria RAM, con DIMMs del tipo DDR5 distribuidas de manera óptima y balanceada ocupando todos los canales de acceso; si con esos GB no se pudiera tener la RAM balanceada, optimizada y con todos los canales de acceso ocupados, se instalará como mínimo la cantidad inmediatamente superior de RAM que permita dicha configuración.
- El sistema operativo montado en discos SSD (mixed use - enterprise class) en RAID 1, 5 o 6 con 900 GB de espacio neto mínimo.
- Por lo menos 8 conexiones a 10 Gb/s 10GBASE-T. Dos de estas bocas se agregarán para conectarse a la red privada del clúster y las otras dos para conectarse al exterior con conexiones de servicio Ethernet a 10Gb/s UTP y cableado de 10m de categoría 6A para conectar a nuestros switches.
- Tarjeta de servicio Ethernet a 1 Gb/s para IPMI.
- Tarjeta Infiniband HDR 100 Gb/s.
- El servidor deberá montar el sistema de ficheros de alto rendimiento.

Sistema de almacenamiento de alto rendimiento

Se requiere un sistema global de ficheros distribuido de alto rendimiento BeeGFS con las siguientes características mínimas:
- Por lo menos 160 TB de espacio neto.
- 5 GB/s tanto en escritura como lectura (benchrmark Iozone de la valoración).
- El sistema de ficheros se conectará y se exportará por las interfaces Infiniband y Ethernet al resto de nodos y servidores.
- El sistema de almacenamiento deberá de ser tolerante a fallos y no existirá un punto único de fallo, aplicando un protocolo adecuado de RAID6 (RAID1 para metadatos y OS). Esto implica redundancia ante la caída incluso de uno cualquiera de los servidores que proveen el almacenamiento.

Fdo.: Inmaculada Arostegui Madariaga
Vicerrectora de Investigación de la UPV/EHU