Laboratorio de Cómputo de Alto Desempeño
Comité de servicios de supercómputo

20160608_LANCAD-RFI-V01-R01

Cluster de Alto Rendimiento para el LANCAD

Solicitud de información y propuesta.

Integrado por:
Comité de servicios de supercómputo
LANCAD

Documento elaborado por CSS – LANCAD Serial 001. Revisión 01. 20160608_LANCAD-RFI	**Cluster de Alto Rendimiento para LANCAD** **Solicitud de información y propuesta.**
Laboratorio de Cómputo de Alto Desempeño	
Para información adicional sobre este documento, contactar a:	José Luis Gordillo Ruiz jlgr@super.unam.mx +52 (55) 56228529 Hector Oliver Hernandez holiver@cinvestav.mx +52 (55) 43877274 Apolinar Martinez Melchor ap0@xanum.uam.mx +52 (55) 15071043

Índice

Privacidad de la información.

La presente Solicitud de Información no constituye, en todo o en parte, compromiso alguno del Laboratorio de Cómputo de Alto Desempeño o de cualquiera de sus entidades o dependencias, de su personal académico, académico – administrativo, usuario, empleados o cualquier otro miembro de su comunidad, para la adquisición de todos o partes de los componentes tecnológicos del Cluster de Alto Rendimiento para LANCAD. Tampoco implica compromiso del Laboratorio con empresa, integrador o distribuidor alguno para la participación de estos últimos en cualquier proceso de adquisición, licitación o procedimiento administrativo en general. El LANCAD se reserva el derecho a utilizar la información proporcionada exclusivamente para los fines que impliquen el diseño, adquisición, instalación y operación de los Cluster de Alto Rendimiento para LANCAD, sin compartir la información recibida con relación a este documento con otras personas, físicas o morales.

Las instancias que proporcionen la información solicitada en este documento asumen lo anterior en el entendido de que los datos relacionados con este documento que emitan a el LANCAD serán de tipo confidencial para cualquier persona ajena a su comunidad, no significan una cotización definitiva o propuesta económica final y tampoco sustituyen los procesos definidos por la UAM en su normatividad para la adquisición de bienes y servicios. Cualquier instancia involucrada con los datos de esta Solicitud de Información y Propuesta asume que no puede liberar hacia terceros, en todo o en parte, información proporcionada por el LANCAD al respecto de este proyecto, para ningún fin o propósito. Esto último no aplica para casos en los que el emisor de información requiera integrar diversos proveedores, fabricantes o representantes, quienes también se obligarán a observar la privacidad de la información al considerarse co-partícipes del proyecto.

Propósito del documento

El propósito del presente documento es solicitar información, de las características y especificaciones técnicas, operativas, funcionales, de instalación, acondicionamiento y comerciales, a proveedores potenciales para la adquisición de equipo de supercómputo. La información que nos proporcione se revisará, analizará, y utilizará como base del estudio de mercado que diseña el LANCAD, para ejercer los recursos obtenidos de la convocatoria 2016 "Apoyos complementarios para el Establecimiento y Consolidación de Laboratorios Nacionales CONACYT", de la cual el LANCAD es beneficiada.

Alcance del documento

Obtener información de los proveedores de supercómputo con experiencia en el suministro e implementación de soluciones del área, con el fin de:

1. Informar el inicio del proceso de licitación para la adquisición del Clúster de Alto Rendimiento para LANCAD, a proveedores potenciales.
2. Establecer las condiciones operativas y técnicas para asegurar la calidad de la instalación y acondicionamiento de soluciones de supercómputo.
3. Ubicar niveles de servicios mediante los cuales se determine la satisfacción de soluciones de supercómputo.
4. Identificar la capacidad de los proveedores potenciales de equipo de supercómputo.

Entrega y métrica de la información

Los documentos de respuesta se deben enviar antes de las 17:00 horas CDT del 22 de Julio del 2016, a los correos jlgr@super.unam.mx, holiver@cinvestav.mx y ap0@xanum.uam.mx. con copia a iret@xanum.uam.mx, en formato de archivo PDF. La respuesta a la presente solicitud de información y documentos relativos deben redactarse en español, con la opción de agregar la información técnica en inglés.

El documento de información debe incluir precios, estos deben expresarse en pesos mexicanos con la vigencia estipulada para estos, y desglosados por unidad siguiendo el formato proporcionado en "Información de resumen de la propuesta".

Las propuestas técnicas serán revisadas por un grupo de especialistas, se le comunicará por correo electrónico al responsable de la propuesta los pasos que siguen para el proceso de adquisición.

Información sobre el procedimiento de adquisición.

La adquisición de equipo de supercómputo se realizará siguiendo los procedimientos de compra de la Universidad Autónoma Metropolitana (UAM), que sustentada en su carácter de institución autónoma se reserva los derechos y procedimientos de compra según se establece en el documento titulado "REGLAMENTO PARA LA ADJUDICACIÓN DE OBRAS, BIENES y SERVICIOS", disponible en la página web de la UAM.

Resumen del Clúster Xiuhcóatl.

El "cluster Híbrido de Supercómputo Xiuhcoatl" es un cluster de supercómputo de propósito general heterogéneo, en él convergen diferentes tecnologías en nodos de cómputo de dos vías con la siguiente distribución:

- 72 AMD Opteron(TM) Processor 6274 @ **2.2GHz**.
- 88 Intel(R) Xeon(R) CPU X5675 @ 3.07GHz.
- 10 Intel(R) Xeon(R) CPU X5675 @ 3.07GHz con 2 GPUs Nvidia FERMI 2070/2075.
- 12 Intel(R) Xeon(R) CPU E5-2650L v3 con un total de 28 GPU (NVIDIA K40).
- 4 Intel(R) Xeon(R) CPU E5-2650L v3 con un total de 8 Xeon Phi (7120P).
- 19 Intel(R) Xeon(R) CPU E5-2660 v3 con 3 GPU (NVIDIA K80) cada uno.

Todos los nodos de cómputo se conectan eléctricamente a PDUs marca APC modelo AP7868, y están alojados en RACKs APC modelo NETSHELTER SX de 42U de altura excepto los 19 nodos con tarjetas NVIDIA K80.

Cuenta con las siguientes redes:

- 2 Gigabit ethernet para administración y servicios ethernet del cluster.
- 1 De alta velocidad y baja latencia basada en Infiniband QDR usando el SW Grid Director 4700 Voltaire/Mellanox, el cual cuenta con 12 line-boards para hacer un total de 216 puertos.

Tiene un sistema de almacenamiento basado en LUSTRE FS presentando dos volúmenes Lustre par home y scratch de 33 TB y 17 TB respectivamente integrados de la siguiente manera:

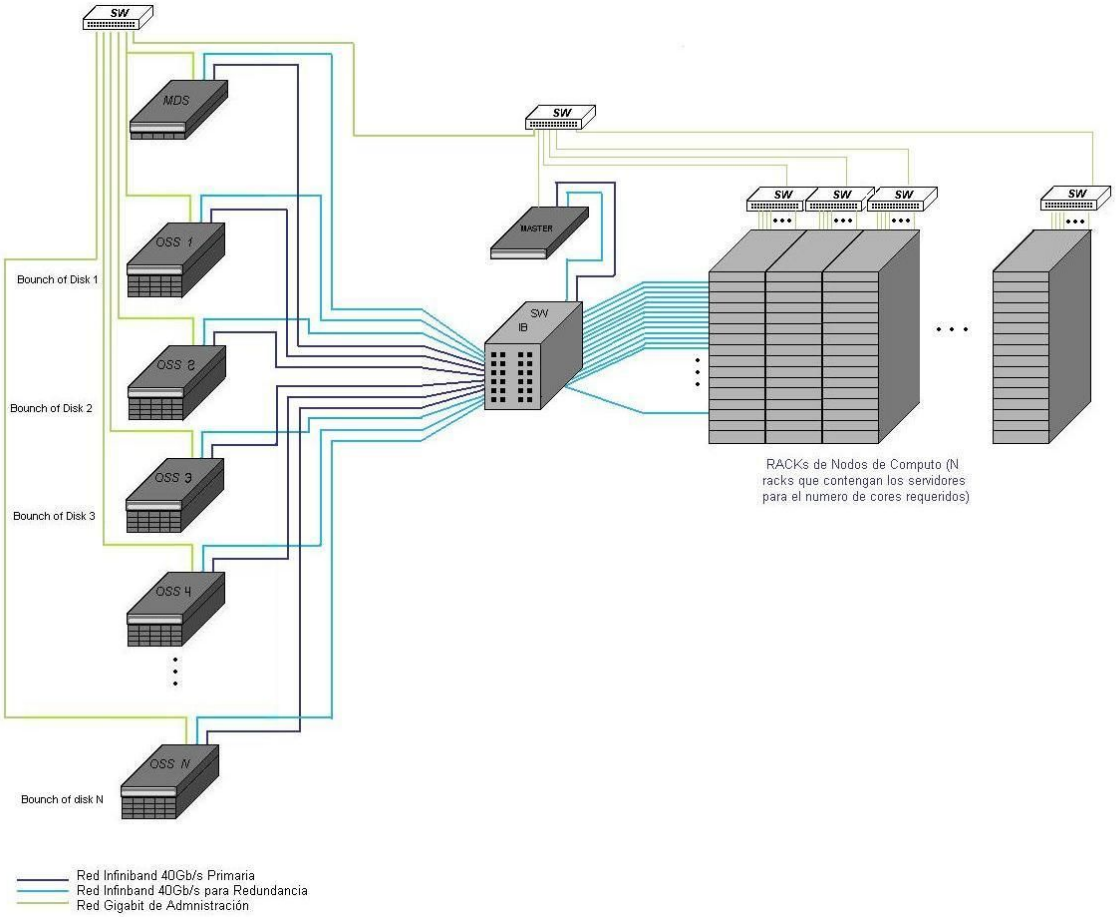
- Para HOME se cuentan con 5 OSS cada uno con 2 OST y cada OST está integrado por 1 raid 6 usando 4 discos + 1 disco de spare en cada RAID.
- Para SCRATCH se cuenta con 5 OSS cada uno con 1 OST y cada OST está integrado por 1 raid 5 usando 3 disco + 1 disco de spare en cada RAID.

- Un MDS con 2 MDTs uno de 5.8 TB y el otro de 2.5TB en RAID 6 para direccionar el Home y Scratch respectivamente.

Los elementos de SW del cluster se describen en la siguiente tabla:

Software	Especificación
Sistema Operativo	CentOS 6.X
Librerías científicas y frameworks	Open MPI, MVAPICH2, Intel MPI, Pthreads, OpenMP,, FFTW, GNU, GSL, HDF5.
Compiladores	GNU GCC, Intel compilers C y Fortran, CUDA, JDK, Python.
Parallel Filesystems and Storage	Lustre
Job Schedulers and Resource Managers	Torque y Maui
System Management	Ganglia

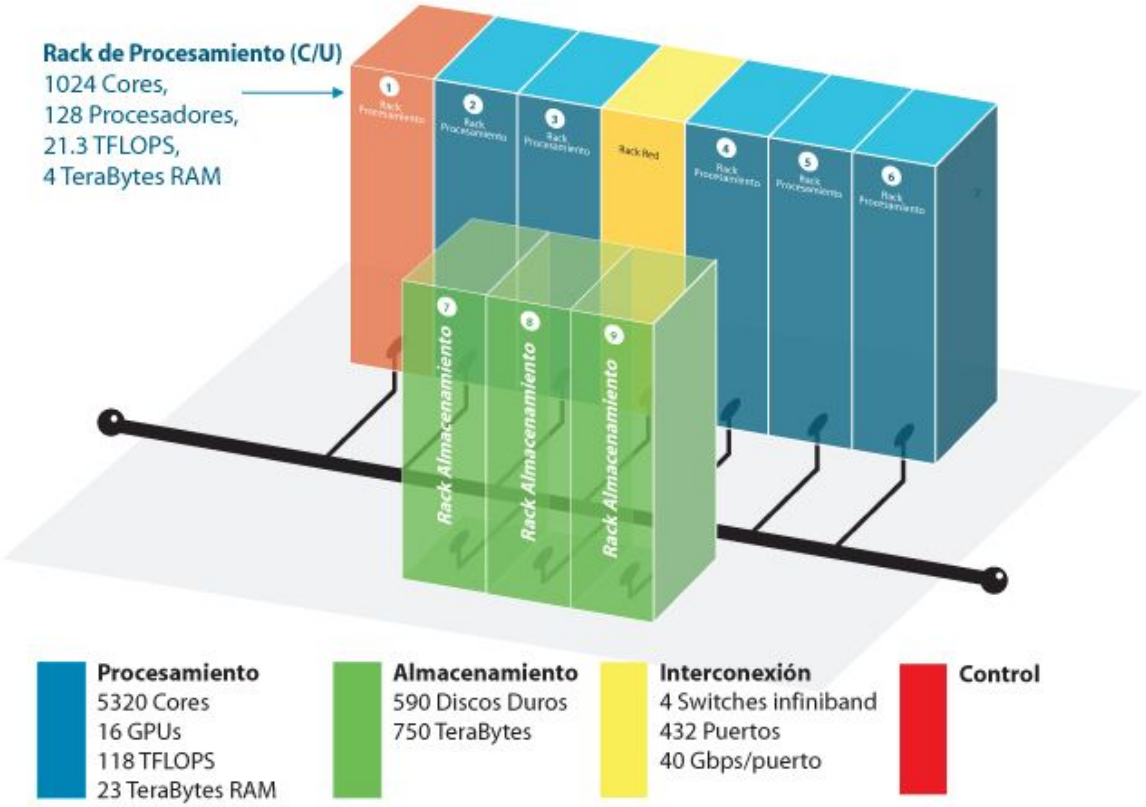
El siguiente diagrama muestra la manera en que está interconectado el cluster (la red gigabit ethernet está interconectada en estrella).

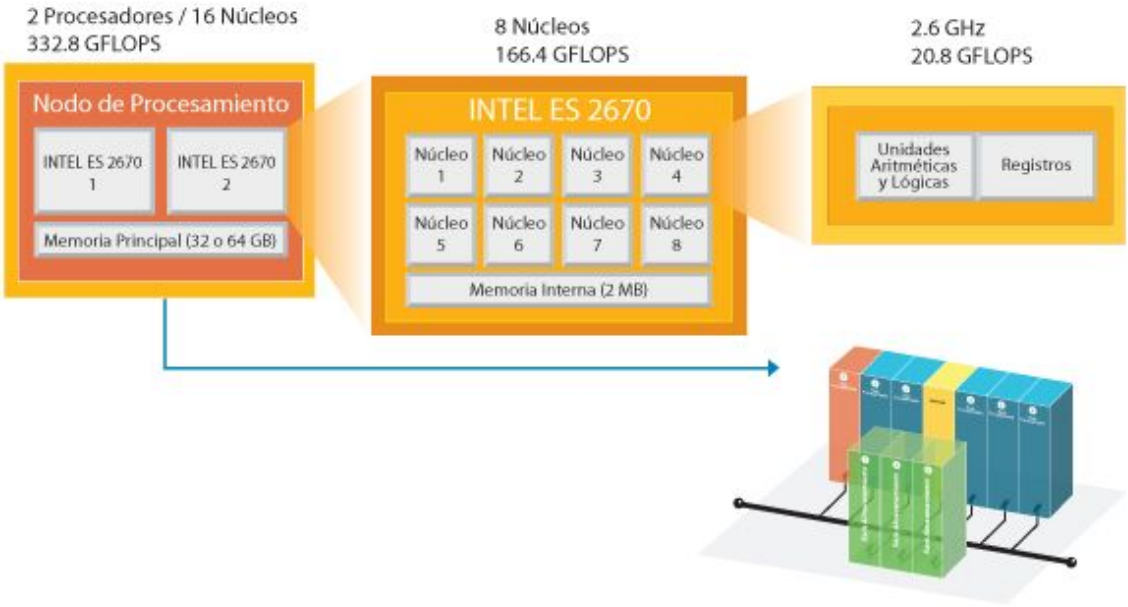


Resumen del Clúster Miztli.

Miztli es una supercomputadora basado en un sistema HP Cluster Platform 3000SL con una capacidad de procesamiento de 118 TFlops. Cuenta con 5,312 núcleos de procesamiento Intel E5-2670, 16 tarjetas NVIDIA M2090, una memoria RAM total de 15,000 Gbytes y un sistema de almacenamiento masivo de 750 Terabytes.

El equipo se compone de 344 servidores (HP Proliant SL230 y SL250), comunicados a través de una red de datos Infiniband, una red de administración ethernet, una red de consolas y varios sistemas de archivos globales.


Rack de Procesamiento (C/U)
1024 Cores,
128 Procesadores,
21.3 TFLOPS,
4 TeraBytes RAM
Rack Red
Rack Almacenamiento
Rack Almacenamiento
Rack Almacenamiento
Procesamiento
5320 Cores
16 GPUs
118 TFLOPS
23 TeraBytes RAM
Almacenamiento
590 Discos Duros
750 TeraBytes
Interconexión
4 Switches infiniband
432 Puertos
40 Gbps/puerto
Control


2 Procesadores / 16 Núcleos
332.8 GFLOPS
Nodo de Procesamiento
INTEL ES 2670 1
INTEL ES 2670 2
Memoria Principal (32 o 64 GB)
8 Núcleos
166.4 GFLOPS
INTEL ES 2670
Núcleo 1
Núcleo 2
Núcleo 3
Núcleo 4
Núcleo 5
Núcleo 6
Núcleo 7
Núcleo 8
Memoria Interna (2 MB)
2.6 GHz
20.8 GFLOPS
Unidades Aritméticas y Lógicas
Registros

Nodos de cálculo

Cada nodo cuenta con dos procesador Intel E5 2670 a 2.6 GHz con 8 cores cada uno, es decir, 16 cores por nodo, clasificados en:

- 314 nodos "regulares" con 64 GB RAM
- 10 nodos "especiales" con 256 GB RAM
- 8 nodos "GPU" con 32 GB RAM y 2 tarjetas NVIDIA M2090
- 4 nodos "de tiempo compartido" con 32 GB RAM
- 2 nodos "login" con 64 GB de RAM
- 6 nodos "de servicio" con 64 GB de RAM, 2 HDD de 1TB en RAID 1, fuentes de poder y ventiladores redundante e intercambiables en caliente.

Almacenamiento

El sistema de almacenamiento principal de Miztli es distribuido y está basado en las tecnologías SFA12K de Data Direct Networks y LUSTRE, del consorcio OpenSFS.

El dispositivo DDN SFA12K proporciona ocho dispositivos de almacenamiento a cada OSS, que se componen de 10 discos SAS/SATA y son capaces de contener hasta 7 Terabytes. El sistema de almacenamiento contiene un total de 590 discos, con fuentes de poder y ventiladores intercambiables en caliente.

Consta de dos sistemas de archivos de 220 TB y 598 TB

Interconexión

El sistema de interconexión principal en Miztli es la red de datos, la cual está compuesta por un switch core Mellanox 4700 de 324 puertos, y tres switches leaf Mellanox 4036 de 36 puertos. Cada puerto es de tecnología Infiniband QDR de 40 Gigabits por segundo, para un ancho de banda total teórico de 6.4 Terabits por segundo. A esta red se conectan todos los elementos del cluster, utilizando para ello tarjetas Mellanox ConnectX-3.

Red de administración, compuesta por switches Gigabit Ethernet 1000 Mbps, 368 puertos totales

Red de consolas, compuesta por switches Fast Ethernet 100 Mbps

Partición LANCAD 1

Compuesta por 56 nodos que contienen 2 procesadores Intel 2660-v3, cada uno con 10 cores a 2.6 GHz., es decir, cada nodo tiene 20 cores y 128 Gigabytes de RAM.

Resumen del Clúster Yoltla.

Yoltla, es una palabra proveniente del vocablo náhuatl: Yoltlamaltini, que significa "Semillero del saber o conocimiento".

El cluster Yoltla se diseñó como un conglomerado de equipos de cómputo para brindar altas prestaciones en el cómputo de alto rendimiento. Su adquisición se realizó a finales del año 2013 y, a partir de este año se ha incrementado su capacidad de procesamiento. En el 2014, con Yoltla se tenían alrededor de 788,400 días de procesamiento CPU, de los que se utilizaron 722,333, con un aprovechamiento del 92.8%. Actualmente el poder de procesamiento del Clúster Yoltla es de 163 billones de operaciones de punto flotante acumulados y medidos con la prueba de rendimiento Linpack.

El cluster Yoltla se integra por 188 nodos de cómputo basados en procesadores Intel Xeon E5-2600v2, 58 nodos de cómputo con procesadores Intel Xeon E5-2600v3 y 64 tarjetas de NVIDIA K20. La administración de los nodos se realiza con el Resource Manager SLURM en su versión 15.08 con el algoritmo de priorización multifactor2, y todos los nodos comparten un almacenamiento LUSTRE 2.5 de 60TB. El almacenamiento de Yoltla cuenta con 7 servidores de objeto, un servidor de metadato y un servidor de gestión. Esta interconectado por una red Infiniband FDR10.
La diversidad de las áreas que requieren del cómputo científico en la UAM originan que el Laboratorio de Supercomputo tenga 441 usuarios, que utilizan nuestro cluster, para análisis numérico y modelación matemática, biofisicoquímica, computación y sistemas, fisicoquímica de superficies, fisicoquímica teórica, procesamiento digital de señales e imágenes biomédicas, química analítica, química cuántica, catálisis, ecuaciones diferenciales y geometría, electroquímica, física de sistemas complejos, física teórica, física de líquidos, optimización e inteligencia artificial y química inorgánica. Por ende, en el Clúster Yoltla se ejecutan aplicaciones con diferentes propósitos y funcionamiento (ver Tabla 1).

El Clúster Yoltla despacha más de 300 tareas diarias, que varían en la solicitud y uso de recursos de hardware (memoria, procesamiento, etc.), las tareas pequeñas requieren de computadoras con una gran cantidad de núcleos de procesamiento aunque hay pocas adecuadas para el correcto uso de memorias en topología NUMA. Las tareas grandes pueden requerir arriba de 320 núcleos de procesamiento y la

granularidad de información procesada puede ser fina o media-gruesa. Para algunas aplicaciones es sumamente importante manejar latencias por debajo de 500ns.

Tabla 3. Recursos de Software disponibles en el clúster Yoltla

Software	Especificación
Sistema Operativo	CentOS 6 y 7
Librerías científicas y frameworks	Open MPI, MVAPICH2, Intel MPI, Pthreads, OpenMP, Libevent, FFTW, GNU GSL, GSL, HDF5.
Compiladores	GNU GCC, Intel compilers C y Fortran, CUDA, JDK, Python.
Aplicaciones	Nwchem, Gromacs, Onetep, Mathematica, R, Namd, Vasp, Gpaw, Qiime, Gaussian.
Herramientas	Clustershell, Pdsh, Tmux, Htop,
Parallel Filesystems and Storage	Lustre, Gluster
Job Schedulers and Resource Managers	SLURM
System Management	Ganglia, Nagios

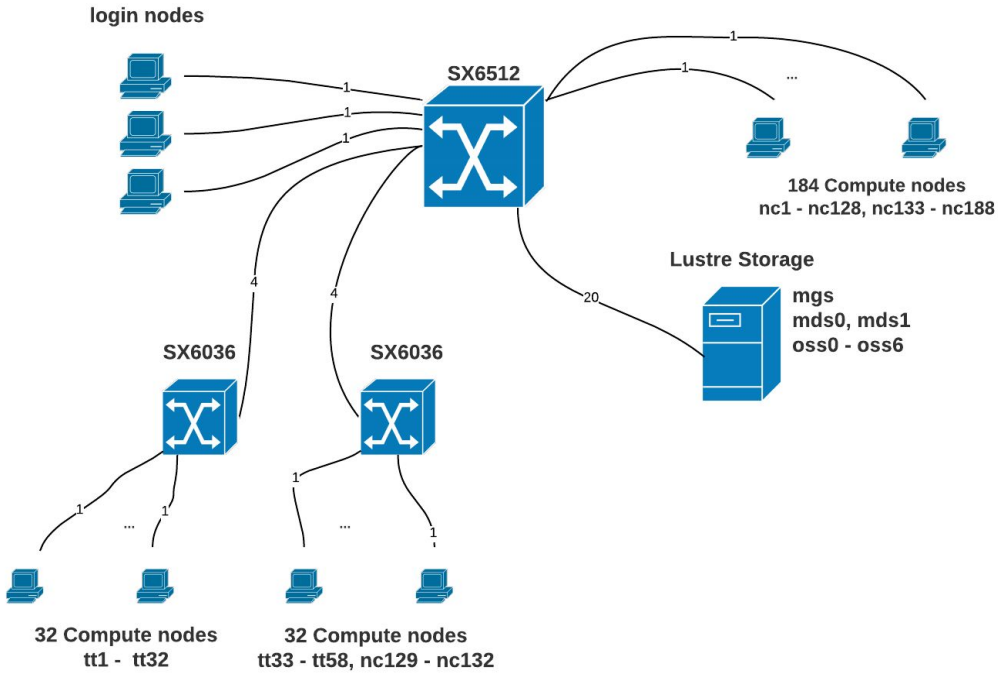


Figura 3. Topología actual cluster Yoltla

Requerimientos generales

El requerimiento del proyecto es la implementación de tres equipos de supercómputo con las siguientes características para cada una, se requiere una cotización para cada combinación de memoria solicitada:

a. 64 nodos de cómputo.
b. 2 Procesadores Intel Xeon E5-2680 v4 por nodo.
c. Con la siguiente distribución de RAM DDR4:
    i. 256 GigaBytes de RAM usando módulos de 32 GB en bus de 2400 MHz por nodo de cómputo en la distribución que optimice el máximo rendimiento de los módulos de memoria con el procesador antes mencionado.
    ii. 192 GigaBytes de RAM usando una combinación de módulos de 32GB y 16 GB en bus de 2400 MHz por nodo de cómputo en la distribución que optimice el máximo rendimiento de los módulos de memoria con el procesador antes mencionado.
    iii. 128 GigaBytes de RAM usando módulos de 16 GB en bus de 2400 MHz por nodo de cómputo en la distribución que optimice el

máximo rendimiento de los módulos de memoria con el procesador antes mencionado.

d. Al menos 2 puertos ethernet de 1 Gbps por nodo de cómputo.
e. Proveer al menos un puerto de comunicaciones Infiniband FDR de 56 Gbps o superior, por nodo.
f. 1 disco para almacenamiento local cumpliendo cualquiera de las siguientes especificaciones:
    i. SSD de al menos 200 GiB
    ii. HDD de al menos 1 TiB
g. 1 Interfaz IPMI 2.0 equivalente o superior por nodo con la capacidad de realizar sesiones Gráficas con su licenciamiento para al menos 5 años de requerirse .
h. La densidad de los nodos de cómputo deberá ser de al menos 2 nodos de cómputo por unidad de rack.
i. 1 Sistema de comunicaciones Infiniband FDR de 56Gbps o superior, que integre todos los nodos de cálculo y el nodo de servicio/acceso. Además debe permitir la capacidad de integrarse a la infraestructura actual, permitiendo un bloqueo máximo de 1:4 con el SW existente.
j. 1 Red ethernet 1 Gbps que integre todos los nodos de cálculo y el nodo de servicio.
k. 1 Red IPMI que integre todos los nodos de cálculo y el nodo de servicio.
l. Servicios de instalación y configuración e integración.
m. Pruebas de rendimiento con las aplicaciones Gromacs 5.X y NWChem 6.X en al menos 4 nodos de cómputo con características iguales a las requeridas en el presente documento.

La infraestructura de nodos de cómputo debe ser enfriada por aire.

Los *input* para las pruebas de rendimiento deben ser descargados del portal de lancad y podrán ser entregados posterior al envío de sus propuestas económicas.

La información solicitada es sobre todos los elementos necesarios para implementar el cluster. Requerimos información para nodos de cálculo A y B para fines de comparación entre las dos alternativas.

Partida	Cantidad total	Producto o componente
3	1	Red Infiniband - Switches, cableado y configuración de switches y nodos necesarios para implementar una red Infiniband FDR que incluya todos los nodos de cálculo y el nodo de servicio (partidas 1 y 2).

		- 3 (tres) años de garantía y soporte contabilizados a partir de la fecha de entrega a satisfacción de la UNAM. - Soporte técnico: 8x5 con resolución el mismo día -
4	1	Red Ethernet - Switches, cableado y configuración de switches y nodos necesarios para implementar una red ethernet 1 Gbps que incluya todos los nodos de cálculo y el nodo de servicio (partidas 1 y 2). - 3 (tres) años de garantía y soporte contabilizados a partir de la fecha de entrega a satisfacción de la UNAM. - Soporte técnico: 8x5 con resolución el mismo día -
5	1	Red IPMI - Switches, cableado y configuración de switches y nodos necesarios para implementar una red IPMI que incluya todos los nodos de cálculo y el nodo de servicio (partidas 1 y 2). - 3 (tres) años de garantía y soporte contabilizados a partir de la fecha de entrega a satisfacción de la UNAM. - Soporte técnico: 8x5 con resolución el mismo día
6	1	Rack - Rack para implementar todos los elementos de la propuesta.
7	1	Servicios - Configuración de herramienta de instalación remota de SO - Configuración de herramienta de monitoreo de HW - Configuración de redes Infiniband, Ethernet, IPMI
8	1	Integración - Integración del cluster a la infraestructura actual de cada nodo LANCAD
9	2	Licenciamiento por tres años para un host del software: Intel Parallel Studio XE Cluster Edition for Linux uno para el cluster UAM y para el cluster CINVESTAV.
10	1	Extensión en el licenciamiento para el manejo de 5 licencias académicas y flotantes por tres años del software: Intel Parallel Studio XE Cluster Edition for Linux.

ANEXO I. Formato para recepción de información.

Por favor remita su propuesta considerando el siguiente formato.

RFI	**20160608-LANCAD-RFI-V01-R01**
Fecha	
Empresa	
Domicilio fiscal	
Teléfonos	
Representante legal (Nombre y correo electrónico)	
Representante ventas (Nombre y correo electrónico)	

Partida	**Cantidad**	**Descripción**	**Precio unitario**	**Subtotal**
		Subtotal		
		Descuentos aplicables a LANCAD		
		16% IVA		
		TOTAL		

Moneda	*EXCLUSIVAMENTE EN MONEDA NACIONAL TODOS LOS VALORES*
Tiempo de entrega	Indicar entrega a partir de formalización de pedido
Vigencia	Señalar la fecha límite de vigencia de la propuesta (no indicación de días)
Otros términos y condiciones	