Diseño de una solución *High Performance Computing* para el grupo CPS

CÉSAR DAVID QUIÑONES VALIENTE

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERIAS FISICO-MECANICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
Bucaramanga
2015

Diseño de una solución *High Performance Computing* para el grupo CPS

CÉSAR DAVID QUIÑONES VALIENTE

Trabajo de Grado para optar al título de

Ingeniero Electrónico

Director

Ing. Carlos A. Angulo Julio

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERIAS FISICO-MECANICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
Bucaramanga
2015

DEDICAROTIA

Este trabajo de grado está dedicado a Dios, a mi familia por estar allí en los momentos donde el camino a este triunfo fue lo más difícil posible, gracias a mi mamá Myriam, a mi papá José y a mis hermanos Erik y María José su apoyo incondicional, también va dedicado a mi abuelo Papi, Hernando Valiente que desde el cielo ha de estar orgulloso y contento de obtener este triunfo para él, a mi gran amigo Carlos Angulo que me apoyo y fue mi guía para lograr terminar, a todos mis amigos de la universidad, Mario, Jhon, Martin Luis Carlos, Daniel, que sin su ayuda esto no sería posible, también a mis amigos de infancia Jorge y Alex, que con su apoyo también aportaron a este gran logro, a mi amigo Alfredo Orduz, que en el momento más oscuro del camino me dio la luz y poder continuar, "Viejo Lucho" este triunfo también es suyo y por último de dedico este triunfo a Johana mi novia, por estar apoyándome siempre lo que fue muy importante para terminar este gran logro, gracias por estar siempre.

AGRADECIMIENTOS

Agradezco al grupo de investigación CPS de la Universidad Industrial de Santander por su constante apoyo y confiar siempre en mí. A mi director Carlos Angulo por su presencia incondicional, sus valiosos aportes y sugerencias, también agradezco a los profesores William Salamanca y Sergio Abreo por colaborarme con la información pertinente para desarrollar este trabajo de grado.

CONTENIDO

Pág.

INTRODUCCIÓN *12*

1. HIGH PERFORMANCE COMPUTING *14*

1.1 ARQUITECTURA DE MULTIPROCESADORES SIMÉTRICOS (SMP - *SYMMETRIC MULTIPROCESSORS*) 14

1.2 ARQUITECTURA PROCESADORES VECTORIALES 15

1.3 ARQUITECTURA CLÚSTER 15

1.4 *HARDWARE/SOFTWARE* EN *HIGH PERFOMANCE COMPUTING* 16

1.5 RED HIGH PERFORMANCE COMPUTING 19

1.6 ALTA DISPONIBILIDAD DE UNA INFRAESTRUCTURA HPC 20

2. DISEÑO DEL HPC PARA EL GRUPO CPS *21*

2.1 INFRAESTRUCTURA CONVERGENTE DE MARCA RANQUEADA EN EL TOP500 23

2.2 *HEAD NODE* Y LOS NODOS CONFORMADAS CON GPGPUS 23

2.3 RED ESPECIALIZADA HPC 24

2.4 ESCALABILIDAD 25

2.5 ALTA DISPONIBILIDAD 25

2.6 ADMINISTRACIÓN REMOTA Y MONITOREO DE LOS DISPOSITIVOS 25

2.7 DISEÑO 1 25

2.8 DISEÑO 2 27

2.9 DISEÑO 3 29

2.10 DISEÑO 4 31

3. CONCLUSIONES *35*

BIBLIOGRAFÍA *36*

LISTA DE FIGURAS

Pág.

Figura 1. Multiprocesadores Simétricos 14
Figura 2. Arquitectura de procesadores vectoriales 15
Figura 3. Arquitectura HPC Cluster 16
Figura 4. Tendencias HPC mundiales en vendedores de hardware 16
Figura 5. Stack de software HP 17
Figura 6. *Stack* de software IBM 18
Figura 7. *Stack* de software Cray 18
Figura 8. *Stack* de software SGI 19
Figura 9. Desempeño Infiniband vs Gigabit Ethernet en HPC. 20
Figura 10. Tecnología Nvidia para HPC 23
Figura 11. Aprovechamiento de Nvidia con Infiniband en HPC 24
Figura 12. Anchos de Banda manejados por redes Infiniband 24
Figura 13. Diseño 1 27
Figura 14. Diseño 2 29
Figura 15. Diseño 3 31
Figura 16. Diseño 4 33

LISTA DE TABLAS

Pág.

Tabla 1. Velocidades de Infiniband 20
Tabla 2. Medidas de disponibilidad de Infraestructuras 21
Tabla 3. Requerimientos de Proyecto 22
Tabla 4. Requerimientos de Proyecto 22
Tabla 5. Comparativo de Diseños Propuestos……..34

RESUMEN

TÍTULO: Diseño de una solución High Performance Computing para el grupo CPS[1]

AUTOR: César David Quiñones Valiente[2]

PALABRAS CLAVES: HPC, High Performance Computing, GPGPU, Infraestructura Convergente, Alta disponibilidad.

DESCRIPCIÓN:

El grupo de investigación CPS se encuentra elaborando algoritmos que buscan mejorar los métodos empleados por la industria petrolera en el estudio del subsuelo con el fin de hallar nuevas reservas de hidrocarburos en nuestro país con el apoyo de COLCIENCIAS. Con el fin de disminuir el tiempo de ejecución de dichos algoritmos, surge la necesidad de implementar una infraestructura hardware de alto desempeño para poner en funcionamiento los algoritmos desarrollados por el grupo CPS con el fin de estudiar el subsuelo.

Este trabajo de grado procura realizar el diseño de una infraestructura de hardware, la cual ha de trabajar como una máquina de *High Performance Computing* (HPC) que sea capaz de manejar la gran cantidad de información que se requiere para el estudio de datos sísmicos y que permita la disminución de los tiempos necesarios para obtener resultados.

El diseño de esta infraestructura *High Performance Computing* debe cumplir con las necesidades de los datos a manejar y de la arquitectura de cómputo desarrollada por los proyectos del grupo CPS. Adicionalmente, debe permitir escalabilidad del hardware a futuro, que sea convergente y de fácil migración a infraestructuras más robustas y especializadas, para que sean utilizables por la industria del petróleo.

[1]Trabajo de Grado

[2] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones.
Director: Ing. Carlos Andrés Angulo Julio

ABSTRACT

TITLE: Designing a High Performance Computing solution for the CPS group[3]

AUTHOR: César David Quiñones Valiente[4]

KEYWORDS: HPC, High Performance Computing, GPGPU, converged infrastructure, High Availability.

DESCRIPTIÓN:

CPS research group is developing algorithms that seek to improve the methods used by the oil industry in the study of subsoil in order to find new hydrocarbon reserves in our country with the support of COLCIENCIAS. In order to reduce the execution time of these algorithms, the need to implement a high-performance hardware infrastructure to operate the algorithms developed by the CPS group to study the subsoil.

This work seeks to make the design grade of hardware infrastructure, which has to work like a machine of High Performance Computing (HPC) that can handle the large amount of information required for the study of seismic data and allows decreasing the time required to get results.

The design of this High Performance Computing infrastructure to meet the needs of data handling and computing architecture developed by CPS group projects. Additionally, you should enable future scalability of hardware that is convergent and easy migration to more robust and specialized infrastructure to be used by the oil industry.

[3]Degree Project

[4]Faculty of Physics Mechanics Engineering. Electrical, Electronics Engineering and Telecommunications School. Director: M.Sc Carlos Andrés Angulo Julio

INTRODUCCIÓN

El grupo de investigación CPS, teniendo en cuenta el desarrollo de proyectos basados en exploración petrolera por medio de algoritmos especializados, se ve en la necesidad de adquirir una infraestructura de alto desempeño con la cual se puedan poner en práctica los algoritmos desarrollados para la sísmica y así realizar las pruebas pertinentes.

Normalmente, estos algoritmos tardarían semanas en obtener resultados en una CPU, mientras que con la utilización de una infraestructura de alto desempeño, dichos resultados podrán ser obtenidos en días o incluso horas. Para esto se debe seleccionar la infraestructura especializada adecuada que permita reducir el tiempo de ejecución de los algoritmos, lo cual también generará disminución en costos, consumo de energía, entre otros.

Dicha infraestructura debe estar compuesta de dispositivos de cómputo, de almacenamiento, de comunicación y administración. Para la selección de estos dispositivos se debe tener en cuenta que el volumen de datos que se maneja en esta clase de algoritmos sísmicos es del orden de terabytes, por lo tanto, se debe tener el almacenamiento capaz de manejar esta cantidad de información con la suficiente velocidad tanto de lectura como escritura.

Así mismo, se requiere realizar simultáneamente diversos cálculos sobre cada uno de estos datos, por lo cual, esta infraestructura debe contar con GPGPUs (*General Purpose Graphics Processor Units*) como unidad de procesamiento principal. Al utilizar una GPU como dispositivo de cómputo, el tiempo de ejecución empleado en la ejecución de los algoritmos puede disminuir significativamente, ya que en estas arquitecturas es posible ejecutar varios procesos simultáneamente aprovechando la posibilidad de paralelismo que ofrecen sus múltiples núcleos.

Adicionalmente, debido a la complejidad de esta infraestructura, se deben tener las herramientas de administración que permitan que todos y cada uno de sus dispositivos sean monitoreados, de fácil acceso y administrados en su totalidad

desde una sola consola. La comunicación entre los dispositivos de dicha infraestructura será una comunicación especializada que soporte flujos de datos masivos y tolerante a fallos con el fin de garantizar que no se presente pérdida de información.

El propósito de este trabajo es diseñar una infraestructura de hardware, integrada por GPUs y dispositivos de almacenamiento, redes y administración, que posibilite reducir el tiempo de ejecución de los algoritmos desarrollados por el grupo de investigación CPS para la industria petrolera.

Este documento está organizado de la siguiente forma: En la sección 1 se muestra un estado del arte de High Performance Computing. En la sección 2 se muestran los diseños propuestos obtenidos de acuerdo a las necesidades del grupo CPS y finalmente en la sección 3 se mencionan las conclusiones obtenidas de este trabajo de grado.

1. *HIGH PERFORMANCE COMPUTING*

Hoy en día los clúster para HPC (*High Performance Computing* o Computación de Alto Desempeño) son muy populares ya que se consideran como una herramienta fundamental para el desarrollo de la investigación científica. Se basan en el alto rendimiento dado la gran demanda de procesamiento de datos, memoria y otros recursos de hardware, donde la comunicación entre ellos es muy rápida, con lo cual el tiempo de transferencia de información de una interfaz a otra se disminuye de manera considerable.

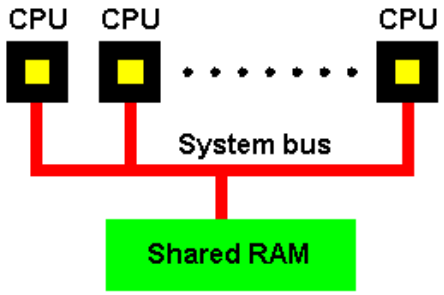
La base principal del HPC es la programación en paralelo, la cual es una metodología de programación que permite dividir un programa en subprogramas para resolver un problema determinado, permitiendo que se utilicen todos los recursos de la máquina al mismo tiempo. Por lo cual se puede decir que HPC es el uso de supercomputadores y técnicas de procesamiento en paralelo para resolver problemas de alta complejidad computacional.

Existen diversos tipos de arquitecturas HPC, entre ellas: Multiprocesadores Simétricos (SMP - *Symmetric Multiprocessors*), Procesadores Vectoriales y Clústeres.

1.1 ARQUITECTURA DE MULTIPROCESADORES SIMÉTRICOS (SMP - *SYMMETRIC MULTIPROCESSORS*)

Es un tipo de arquitectura HPC en la cual múltiples procesadores comparten la misma memoria y utilizan un bus de datos compartidos para acceder de manera simétrica a la memoria. Dependiendo de la cantidad de procesadores, generalmente dicho bus se convierte en un cuello de botella, lo cual disminuye de manera considerable el desempeño de la máquina. Suelen ser más costosos y menos escalables que la arquitectura Clúster.

Figura 1. Multiprocesadores Simétricos

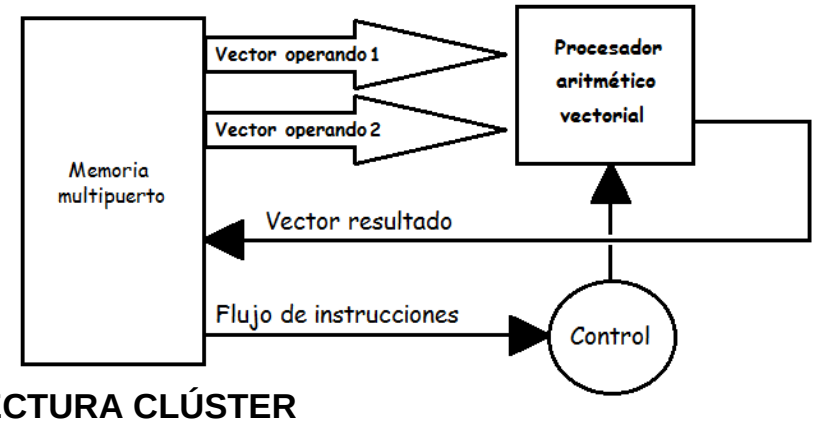


1.2 ARQUITECTURA PROCESADORES VECTORIALES

Los sistemas de procesadores vectoriales se han venido trabajando desde 1980. En la arquitectura de procesadores vectoriales, la CPU optimiza su funcionamiento utilizando arreglos vectoriales. Los procesadores vectoriales permiten que se realicen operaciones simultáneas sobre múltiples datos utilizando vectores.

Una máquina vectorial consta de una unidad escalar segmentada y una unidad vectorial. La unidad vectorial dispone de *M* registros vectoriales de *N* elementos y de unidades funcionales vectoriales (de suma/resta, multiplicación, división, carga/almacenamiento, etc.) que trabajan sobre los registros vectoriales, y un conjunto de registros escalares.

Figura 2. Arquitectura de procesadores vectoriales



1.3 ARQUITECTURA CLÚSTER

Es la predominante en las soluciones HPC recientes. La arquitectura clúster consiste en un conjunto de procesadores (los cuales son identificados como nodos) que poseen su propia CPU, memoria, dispositivos I/O, sistema operativo y comunicación con otros nodos de ser necesario.

Existen tres tipos de clúster: *Fail-over cluster* o clúster de alta disponibilidad, clúster de balanceo de carga, *High Performance cluster* o clúster de alto desempeño.

- Clúster de alta disponibilidad:

  Es aquel clúster compuesto de dos nodos o tres nodos, donde uno de los nodos está activo mientras los otros están pasivos realizando un monitoreo constante del activo. En caso de falla, el nodo pasivo se convierte en el activo y toma el control de los servicios permitiendo que el sistema de misión crítica siga funcionando.

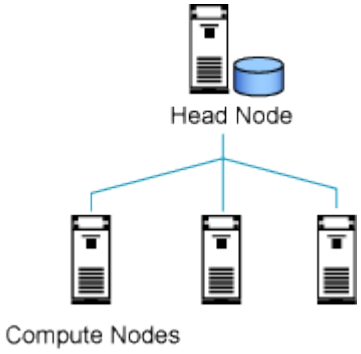
- Clúster de balanceo de carga

  Este tipo de clúster tiene varios nodos entre los cuales se reparten la carga de trabajo de acuerdo a las necesidades de los usuarios. Son usualmente utilizados para servidores web que sirven de *host*.

- Clúster de alto desempeño

  Son utilizados para ejecutar programas que funcionan en paralelo y requieren alta computación para obtener resultados de especial interés para la comunidad científica. Estos programas, que tomarían una gran cantidad de tiempo en un solo PC, se realizarían de manera más rápida y confiable en este tipo de clúster.

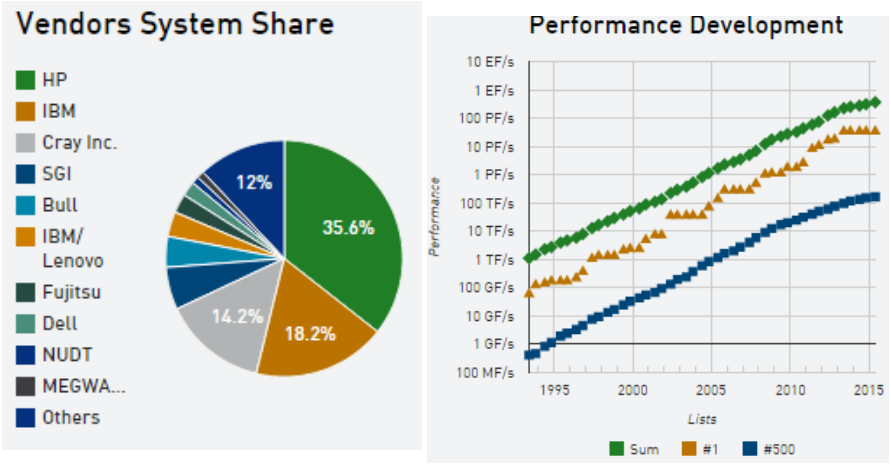
Figura 3. Arquitectura HPC Cluster



1.4 *HARDWARE/SOFTWARE* EN *HIGH PERFOMANCE COMPUTING*

En la Figura 4 son mostrados los vendedores de hardware con mejor desempeño para HPC de acuerdo a TOP500[5]. Se puede observar que las 4 empresas que venden más sistemas HPC son: HP, IBM, Cray Inc. y SGI.

Figura 4. Tendencias HPC mundiales en vendedores de hardware



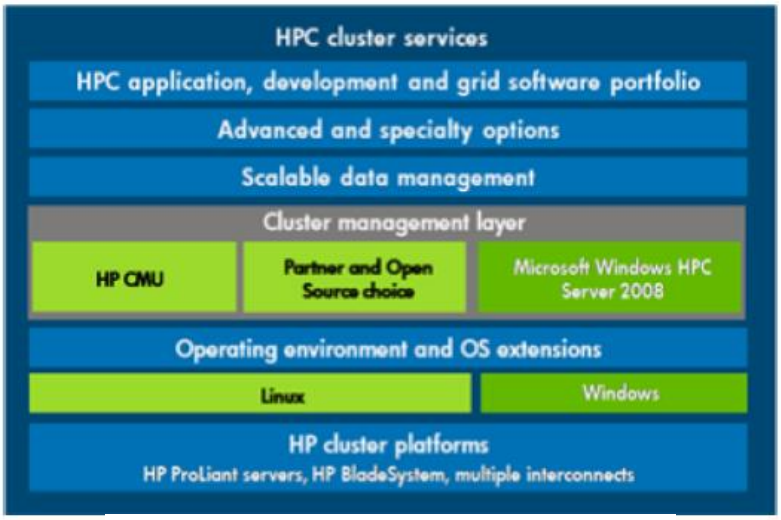
[5] TOP500: Ranking mundial de los 500 supercomputadores con mejor rendimiento en el mundo, http://www.top500.org/

A continuación se describirá el paquete software o *stack* de software[6] que utiliza cada empresa líder en HPC:

- HP (Hewlett Packard)

En la Figura 5 vemos el *stack* de software que ofrece esta empresa para sus clústeres. Ofrecen sistema operativo Linux y Windows, y tienen software preparado por ellos.

Figura 5. Stack de software HP



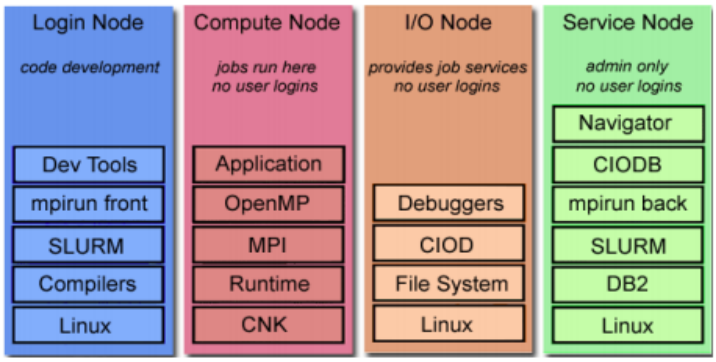
Librerías	Intel math kernel library	
MPI	IntelMPI	OpenMPI
Compiladores	Intel compiler collection	
Sheduler	SLURM	
SO	Linux	

- IBM

En la Figura 6 vemos los distintos nodos que IBM usa en sus clústeres y su *stack* de software, si se ve como un supercomputador donde todos los nodos son iguales. IBM usa en los nodos de computación un *kernel* [7] llamado *Compute Node Kernel* (CNK), que delega la tarea de entrada/salida a otros nodos que corren Linux.

[6] *Stack* de software: Conjunto de software que son utilizados por una infraestructura de hardware para conseguir un fin común.
[7] *Kernel:* Es el núcleo del sistema operativo, permite que trabajen el hardware y el software de manera conjunta realizando una administración de ambos recursos.

Figura 6. *Stack* de software IBM



Librerías	MASS	FFTW	Netlib
MPI	MPI		
Compiladores	GCC	IBM Compilers	
Sheduler	SLURM		
SO	Linux		

- Cray Inc.

En la Figura 7 se destaca la cantidad de software propio que usan y la cantidad de software que ofrecen.

Figura 7. *Stack* de software Cray

Performance Monitoring	HPCC	Perfctr	IOR	PAPI/IPM	netperf	
Development Tools	Cray Compiler Environment (CCE)	Intel® Cluster Studio	PGI (PGI CDK)	GNU		
Application Libraries	Cray LibSci	MVAPICH2	OpenMPI	Intel® MPI (Cluster Studio)		
Resource Management/ Job Scheduling	Grid Engine Integrated	SLURM Integrated	MOAB	Altair PBS Professional	IBM Platform LSF	Torque/Maui
File System	NFS	Local FS (ext3, ext4, XFS)	PanFS	Lustre		
Provisioning	Cray® Advanced Cluster Engine (ACE™) Management Software					
Cluster Monitoring	Cray ACE (iSCB and OpenIPMI)					
Remote Power Management	Cray ACE					
Remote Console Management	Cray ACE					
Operating System	Linux (Red Hat, CentOS, SUSE)					

- SGI

El *stack* de software usado es el mostrado en la Figura 8. Cuenta con librerías optimizadas para las Xeon Phi que usan. También han desarrollado OpenMC, que es una capa de abstracción del tipo OpenMP, CUDA u OpenCL para poder facilitar el uso de los procesadores y las Xeon Phi del nodo a la vez.

Figura 8. *Stack* de software SGI

Librerías	Intel math kernel library
MPI	MPICH
Compiladores	Intel compiler collection
Sheduler	SLURM
SO	Linux

1.5 RED HIGH PERFORMANCE COMPUTING

Debido a la gran densidad de información que se intercambia entre los nodos, se hace necesaria una red muy robusta que permita que la velocidad de procesamiento no se vea mermada por la velocidad de transmisión de los datos obtenidos en el procesamiento.

Las infraestructuras modernas de *High Performance Computing* son extremadamente dependientes del ancho de banda de interconexión. También es primordial tanto la baja latencia en dichas comunicaciones entre nodos como la administración desde el nodo principal.

En las infraestructuras HPC, el tráfico de entrada/salida es de gran volumen dada la cantidad de información y tamaño de archivos involucrados. En soluciones en que la cantidad de archivos no es común, se debe tener en cuenta que existe un tráfico de mensajes para sincronización, por división de tareas, intercambio de datos entre procesadores, por lo cual el ancho de banda necesario es realmente grande y con una concurrencia muy alta. Es por esto que se requiere una red especializada, en la cual la latencia sea menor en comparación de las redes no especializadas. Para el entorno HPC se recomienda manejar al menos latencia de un rango entre 30 y 100 µs **[Symetric Computing]**

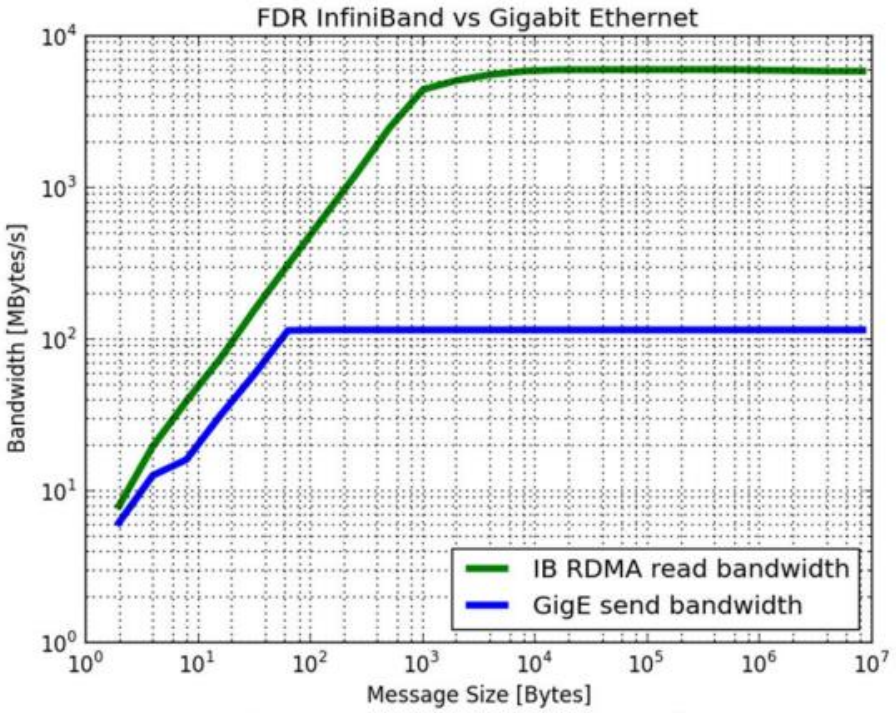
Para un clúster HPC se pueden utilizar dos tipos de redes: Ethernet e Infiniband. Las redes Ethernet, siendo estas redes las más económicas del mercado que se pueden implementar en la actualidad, requieren de comprobaciones de error continuas que aumentan su latencia. Además, sus velocidades de trasmisión no son las suficientes para un clúster de alto desempeño y de gran intercambio de datos, según la aplicación.

Estas redes manejan velocidades de 10Mbps, 100Mbps, 1Gbps y 10Gbps. La velocidad de 10Gbps es obtenida con módulos de cobre especializados o por conexiones de Fibra Óptica utilizando módulos SFP+. Este tipo de red es utilizada para administración de dispositivos, por lo cual en un clúster HPC su mayor utilización sería en acceso administrativo y de monitoreo de los dispositivos que conformen la infraestructura.

Las redes InfiniBand, son las sugeridas para entornos con infraestructura clúster HPC, manejando los estándares necesarios para el intercambio de información

denso, a baja latencia y con mucho mayor ancho de banda en comparación a las redes Ethernet como lo muestra la Figura 9.

Figura 9. Desempeño Infiniband vs Gigabit Ethernet en HPC.



Tomada de [KALCHER]

Las redes Infiniband utilizan la agregación de canales permite anchos de bandas muy altos. Estas agregaciones se pueden observar en la Tabla 1, logrando enlaces a velocidades reales de hasta 56,25Gbps. Es por esto que la red para una infraestructura HPC debe ser Infiniband con el fin de aprovechar las velocidades de procesamiento de los nodos.

Tabla 1. Velocidades de Infiniband

InfiniBand link type	Signaling Rate	Data Rate
4 x DDR (Double Data Rate)	20 Gbit/s	16 Gbit/s
4 x QDR (Quad Data Rate)	40 Gbit/s	32 Gbit/s
4 x FDR (Fourteen Data Rate)	56.25 Gbit/s	54.55 Gbit/s

1.6 ALTA DISPONIBILIDAD DE UNA INFRAESTRUCTURA HPC

Ante la importancia y lo crítico de cada uno de los procesamientos, la disminución de tiempos de meses y años a días, hacen que durante el tiempo de ejecución de los algoritmos en la infraestructura no puede existir un punto de fallo en el hardware,

ya que se perderían muchas horas de trabajo, energía dedicada al HPC y tiempo de estudio, que al día de hoy corresponden a pérdida de conocimiento y a dinero.

Por lo cual, una infraestructura convergente para trabajar como *High Performance Computing* debe ser altamente disponible. Dicha disponibilidad equivale a tener redundancia en fuentes de alimentación, en respaldo energético, en respaldo de la información evitando fallas de discos y conectividad por falla en equipos de comunicación entre nodos.

Para garantizar la continuidad del funcionamiento de la infraestructura se realiza un monitoreo continuo por medio de *heartbeat* a todos y cada uno de los dispositivos y verificando que los sistemas redundantes trabajen de manera normalizada. La alta disponibilidad o *high availability* es medida de acuerdo a la disponibilidad que deben tener los servicios o procesos que funcionan sobre la infraestructura como se muestra en la siguiente tabla:

Tabla 2. Medidas de disponibilidad de Infraestructuras

Disponibilidad	Tiempo de caída
90%	16.8 horas
99%	1.68 horas
99.9%	10.1 minutos
99.99%	1.01 minutes
99.999%	6.05 segundos
99.9999%	604.8 milisegundos

De acuerdo a la cantidad de nueves se mejora la disponibilidad, por lo tanto se requiere mayor y mejor infraestructura con el propósito de garantizar la disponibilidad, lo cual implica costos altos.

2. DISEÑO DEL HPC PARA EL GRUPO CPS

El diseño de la Infraestructura Convergente *High Performance Computing* del Grupo CPS ha de cumplir con los requisitos obtenidos del estudio de la tecnología actual y con los requerimientos de los proyectos desarrollados en el grupo de investigación.

De las investigaciones realizadas por el grupo de investigación CPS, tienen particular interés las relacionadas con los siguientes temas:

1. Inversión sísmica para la detección de modelos de velocidades en zonas geológicas complejas.
2. Optimización del tiempo de ejecución del algoritmo RTM 3D sobre una plataforma GPU.

Estas investigaciones presentan los requerimientos mostrados en las tablas 3 y 4 respectivamente.

Tabla 3. Requerimientos del Proyecto 1: Inversión Sísmica.

HEAD NODE	
Memoria RAM	128GB
Procesador	2,4GHz por núcleo
NODOS	
Memoria RAM	256GB
Procesador	2,6GHz por núcleo
RAM de GPGPU	14,5GB
Almacenamiento	8.66 TB

Tabla 4. Requerimientos del Proyecto 2: Optimización algoritmo RTM.

HEAD NODE	
Memoria RAM	128GB
Procesador	2,4GHz por núcleo
NODOS	
Memoria RAM	32GB
Procesador	2,6GHz por núcleo
RAM de GPGPU	14,5GB
Almacenamiento	8.66 TB

Por lo tanto, el diseño de la infraestructura HPC para el grupo de investigación CPS debe cumplir con los siguientes aspectos:

- Infraestructura Convergente de marca ranqueada en el top500
- *Head Node* y los Nodos conformadas con GPGPUs
- Red especializada HPC
- Escalabilidad
- Alta disponibilidad

- Administración Remota y Monitoreo de los dispositivos.

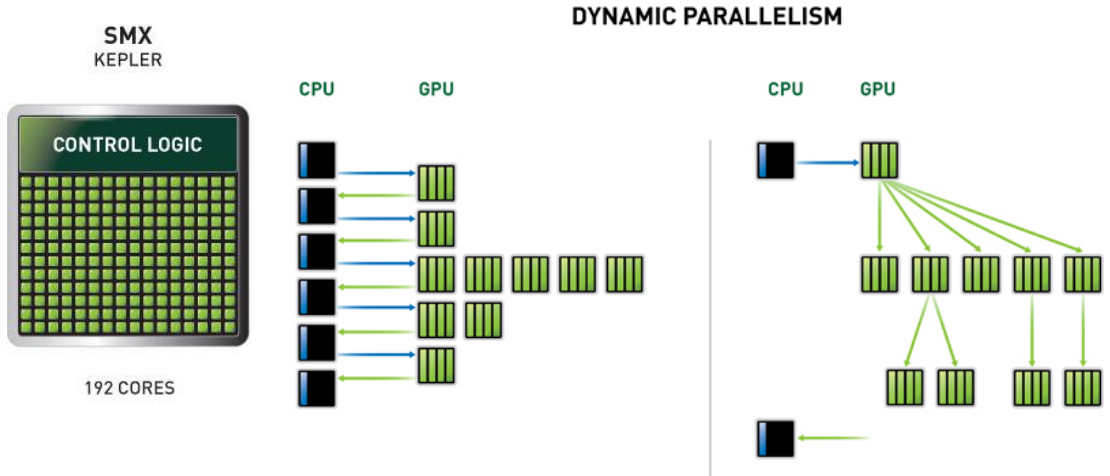
2.1 INFRAESTRUCTURA CONVERGENTE DE MARCA RANQUEADA EN EL TOP500

Los diseños son basados en la infraestructura que le permitirá migrar a una HPC más robusto como lo es GUANE[8]. Dicha infraestructura posee la estructura empleada por Hewlett Packard (HP), una de las mejores ranqueadas en el top500. Esta es la única marca donde todos y cada uno de los elementos que lo conforman trabajan de manera convergente, es decir, mantienen el mismo estándar y la comunicación, administración y monitoreo entre los servidores (computo), almacenamiento (discos) y redes (comunicación).

2.2 *HEAD NODE* Y LOS NODOS CONFORMADAS CON GPGPUS

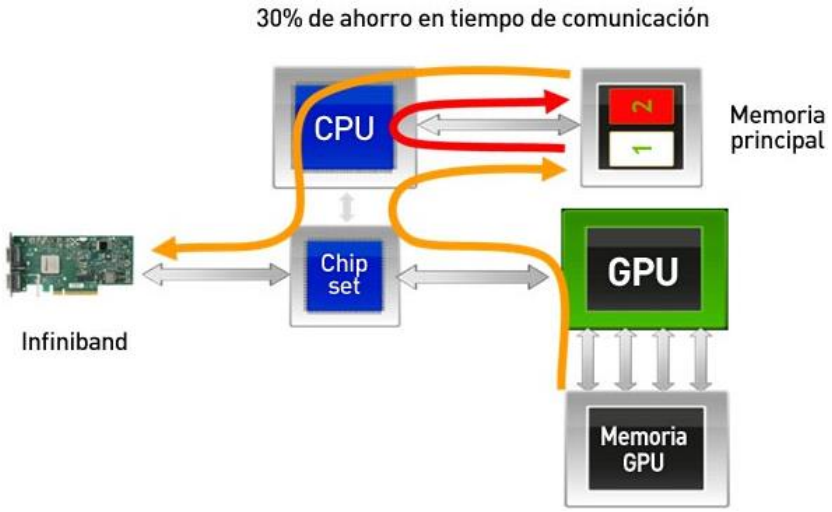
Se escogió el modelo HPC Clúster, el cual está conformado por un nodo principal y varios nodos de cómputo en los que se realizará el procesamiento. Este tipo de clúster es el que permite sacar mejor desempeño al hardware implementado. Este modelo de clúster está basado las arquitecturas desarrolladas por el grupo CPS y también permite su migración sencilla a GUANE ya que este último también trabaja con este modelo de HPC. Se seleccionaron las GPU Nvidia Tesla K40 las cuales manejan una memoria RAM disponible de 12GB, las cuales están entre las de mejor desempeño del mercado, con su arquitectura de hardware especialmente diseñada para el trabajo en paralelo como lo muestra la Figura 10. También poseen acelerador de arquitecturas de software como lo es CUDA, además por medio del uso de redes especializadas Infiniband es capaz de mejorar los tiempos de comunicación entre nodos (Figura 11)**[NVIDIA]**

Figura 10. Tecnología Nvidia para HPC



[8] Infraestructura Convergente propiedad de la UIS, especializada en High Performance Computing

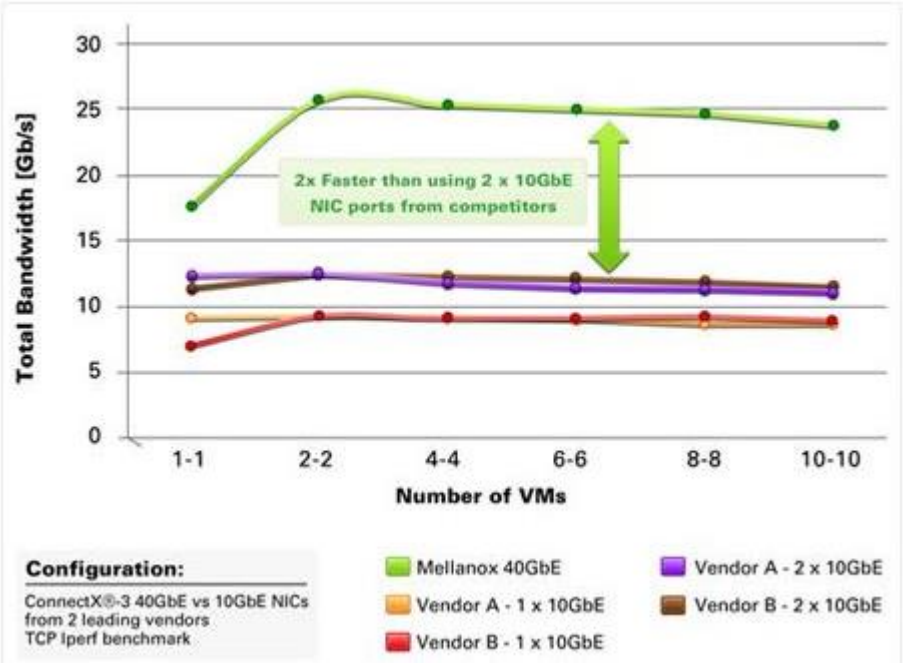
Figura 11. Aprovechamiento de Nvidia con Infiniband en HPC



2.3 RED ESPECIALIZADA HPC

Se seleccionó como red especializada una red Infiniband la cual es la más idónea para *High Performance Computing*, agregándole tarjetas FlexibleLOM con tarjetas Infiniband tanto a los nodos como al nodo principal y un *switch* Infiniband administrable. Con este diseño se pueden manejar velocidades de alrededor de 40Gb/s.

Figura 12. Anchos de Banda manejados por redes Infiniband.



2.4 ESCALABILIDAD

Toda la infraestructura diseñada es totalmente escalable, es decir, permite crecimiento a futuro.

- El chasis el SL6500 donde se ubican los nodos de cómputo permite que se puedan agregar más nodos al HPC, lo cual significa crecimiento en poder de cómputo.
- El chasis del nodo principal permite crecimiento en almacenamiento agregando más discos, más memoria y también permite colocar procesadores muchos más potentes, con tarjetas PCI express para adicionar GPGPUs si se requiere mayor poder de cómputo

2.5 ALTA DISPONIBILIDAD

La infraestructura propuesta es altamente disponible ya que posee redundancia en potencia, almacenamiento y redes. Cada servidor posee fuentes de alimentación redundantes:

- Nodo principal: posee dos fuente redundantes
- Chasis nodos de cómputo: poseen 4 fuentes, las cuales permiten que fallen hasta dos de las fuentes que lo componen, también posee redundancia en el sistema de enfriamiento.
- Se diseña también con dos UPS con el fin de tener circuitos independientes en caso de que falle una de ellas o uno de los circuitos al cual está conectado, el otro siga funcionando.

2.6 ADMINISTRACIÓN REMOTA Y MONITOREO DE LOS DISPOSITIVOS

Debido a la convergencia de los equipos HP, se permite la administración, monitoreo y mantenimiento de ser necesario utilizando una sola herramienta, llamada *One View*. También por medio de una tarjeta especializada y totalmente independiente a los servidores permite, su monitoreo y administración sin necesidad que el sistema operativo esté en funcionamiento, esta herramienta se llama *Insight Lights Out*, ILO.

De acuerdo a estos requerimientos se generan 4 diseños:

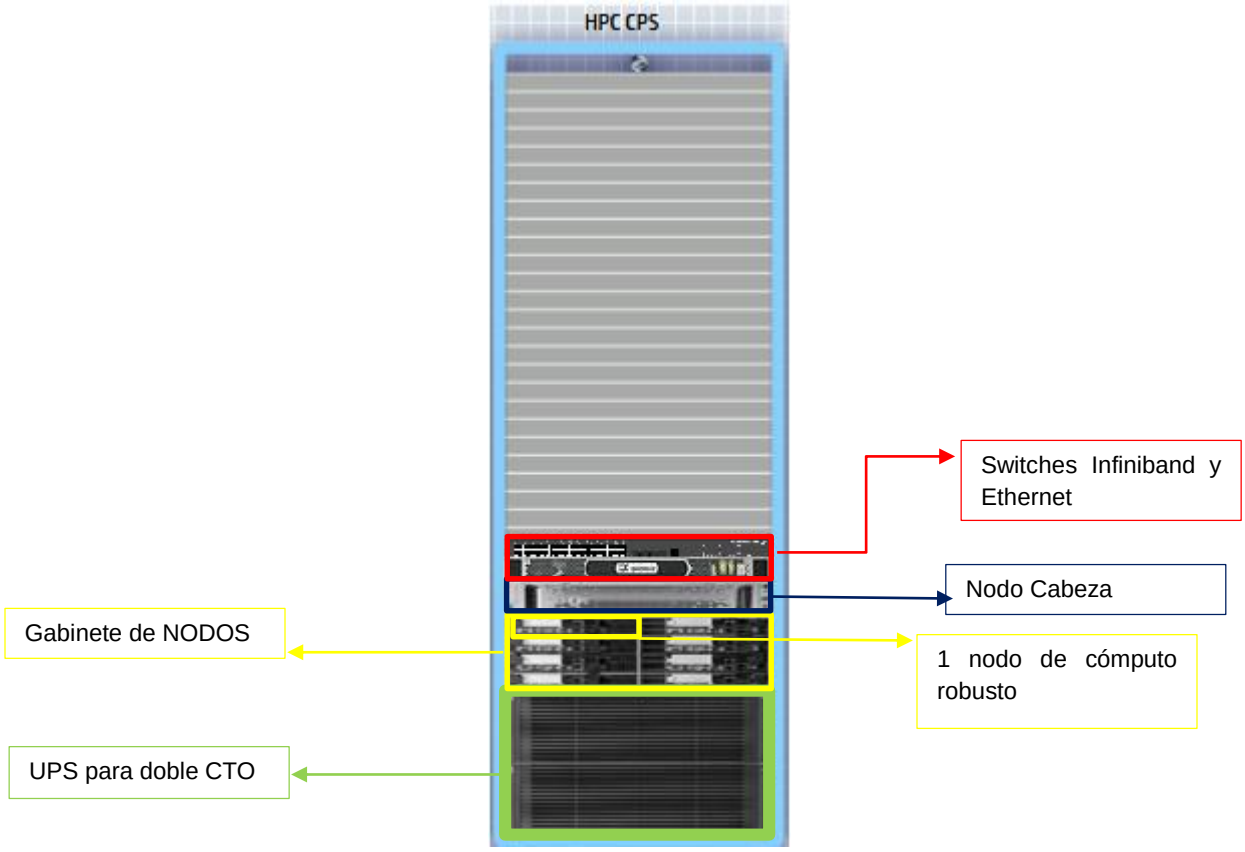
2.7 DISEÑO 1

Este diseño corresponde a un cluster conformado con un solo nodo muy robusto y con la memoria una gran cantidad de memoria en el mismo nodo.

HEADNODE	CANTIDAD
DL380p G8	
HP DL380 Gen9 Intel® Xeon® E5-2620v3 (2.4GHz/6-core/15MB/85W) Processor Kit	2
Memoria DDR4-2133 16GB MODULE	8
Flexible LOM HP InfiniBand FDR 2-port 545QSFP Adapter	1
Doble tarjeta de red PCI-express 1Gbps	1
Fuente redundante 800W gold	2
HP ILO REMOTE 1Gbps	1
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise 3yr Warranty Hard Drive	2
HP 1.2TB 6G SAS 10K rpm SFF (2.5-inch) SC Dual Port Enterprise 3yr Warranty Hard Drive	12
Smart Array Controller RAID 1/1+0 5/5+0 6/6+0	
NODOS	
SL6500	
Fuente de redundantes HP 1200W/1500W Common Slot Platinum Plus Hot Plug Power Supply Kit	4
4 Fans redundantes	4
COMPUTE NODE	
SL250s G8	
Intel® Xeon® E5-2650v2 (2.6GHz/8-core/20MB/95W)	2
HP 16GB (1x16GB) Dual Rank x4 PC3-14900R (DDR3-1866) Registered CAS-13	16
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise	2
GPU Nvidia K40	2
Flexible-LOM Infiniband QLOGIC de 40Gbps dual port	1
Doble tarjeta de red PCI-express 1Gbps	1
HP ILO REMOTE 1Gbps	1
Smart Array RAID 1/1+0	1
NETWORKING	
Switch 10/100/1000 HP 5800-24G Switch	1
Switch Qlogic Infiniband	1
Soportes HP	
HP Proactive Care	1
POWER	
R7000 Uninterruptible Power System (UPS)	2
HP PDUs	2
HP PDU de PDUs	2

RACK HP 42u	1

Figura 13. Diseño 1



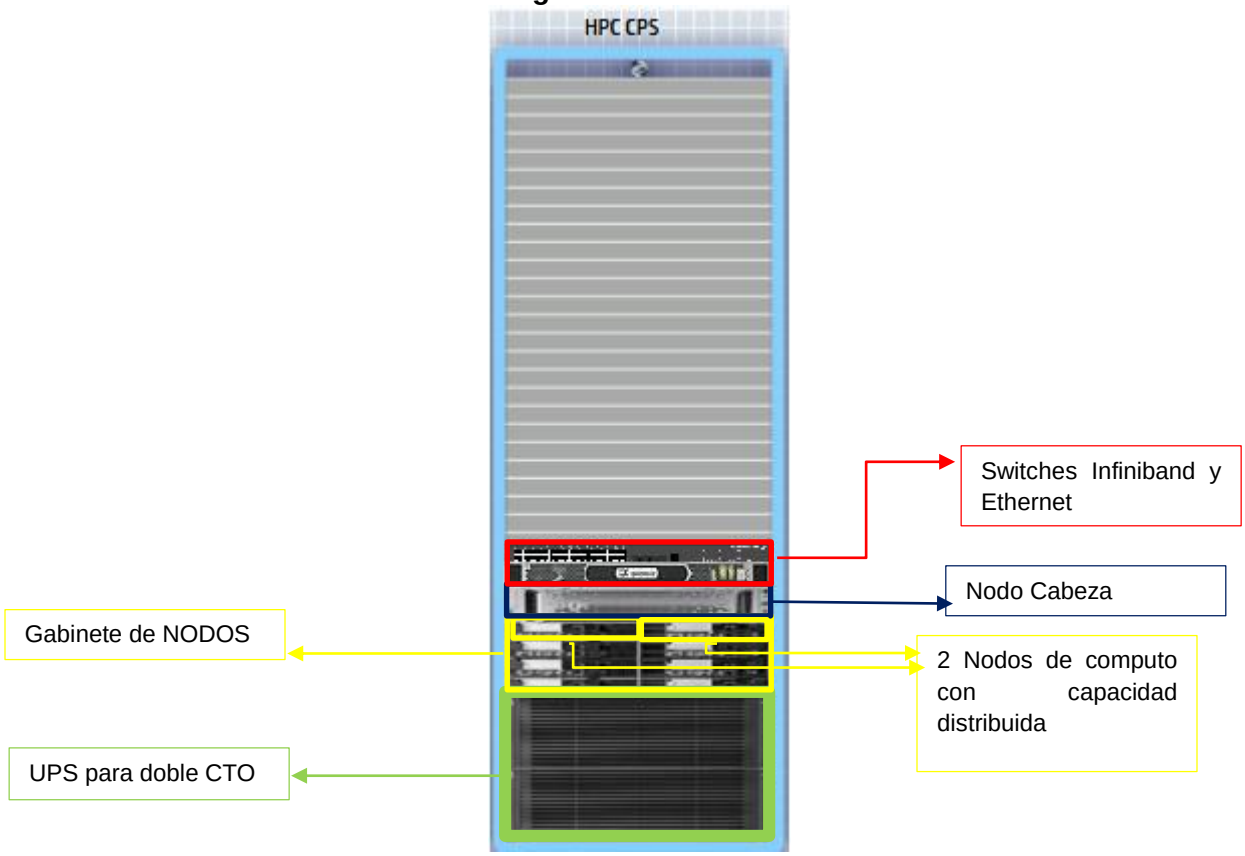
2.8 DISEÑO 2

Este clúster está conformado por nodo principal y dos nodos de cómputo con la memoria distribuida entre ellos y con 4 GPGPUs también distribuidos entre ellos.

HEADNODE	CANTIDAD
DL380p G8	
HP DL380 Gen8 Intel® Xeon® E5-2620v3 (2.4GHz/6-core/15MB/85W) Processor Kit	2
Memoria DDR4-2133 16GB MODULE	8
Flexible LOM HP InfiniBand FDR 2-port 545QSFP Adapter	1
Doble tarjeta de red PCI-express 1Gbps	1
Fuente redundante 800W gold	2
HP ILO REMOTE 1Gbps	1
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise 3yr Warranty Hard Drive	2

HP 1.2TB 6G SAS 10K rpm SFF (2.5-inch) SC Dual Port Enterprise 3yr Warranty Hard Drive	12
Smart Array Controller RAID 1/1+0 5/5+0 6/6+0	1
NODOS	
SL6500	
Fuente de redundantes HP 1200W/1500W Common Slot Platinum Plus Hot Plug Power Supply Kit	4
4 Fans redundantes	4
COMPUTE NODE 1	
SL250s G8	
Intel® Xeon® E5-2650v2 (2.6GHz/8-core/20MB/95W)	2
HP 16GB (1x16GB) Dual Rank x4 PC3-14900R (DDR3-1866) Registered CAS-13	8
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise	2
GPU Nvidia K40	2
Flexible-LOM Infiniband QLOGIC de 40Gbps dual port	1
Doble tarjeta de red PCI-express 1Gbps	1
HP ILO REMOTE 1Gbps	1
Smart Array RAID 1/1+0	1
COMPUTE NODE 2	
SL250s G8	
Intel® Xeon® E5-2650v2 (2.6GHz/8-core/20MB/95W)	2
HP 16GB (1x16GB) Dual Rank x4 PC3-14900R (DDR3-1866) Registered CAS-13	8
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise	2
GPU Nvidia K40	2
Flexible-LOM Infiniband QLOGIC de 40Gbps dual port	1
Doble tarjeta de red PCI-express 1Gbps	1
HP ILO REMOTE 1Gbps	1
Smart Array RAID 1/1+0	1
NETWORKING	
Switch Qlogic Infiniband	1
Switch 10/100/1000 HP 5800-24G Switch	1
Soportes HP	
HP Proactive Care	1
POWER	
R7000 Uninterruptible Power System (UPS)	1
HP PDUs	2
HP PDU de PDUs	2
RACK HP 42u	1

Figura 14. Diseño 2



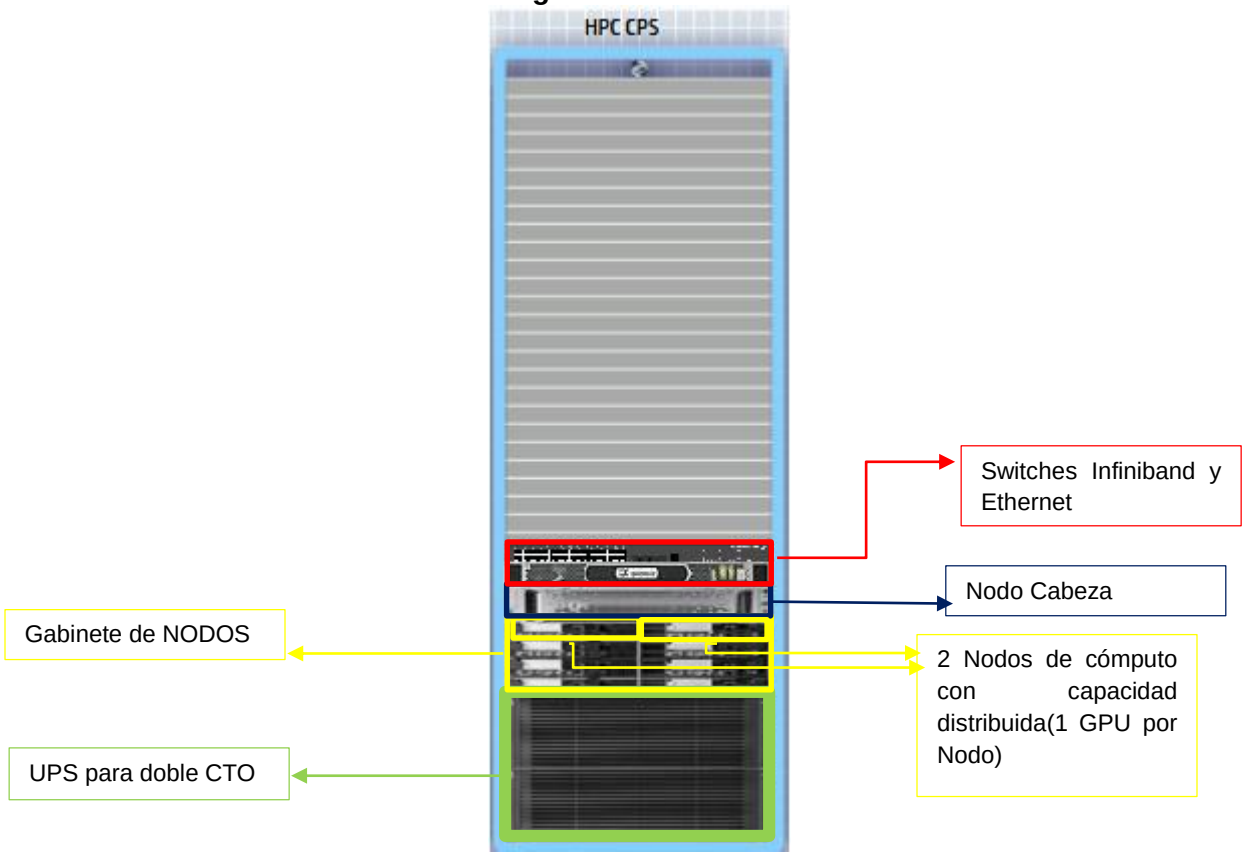
2.9 DISEÑO 3

Este diseño está conformado por nodo principal y dos nodos de cómputo con la memoria distribuida entre ellos y con dos GPGPU también distribuidos entre ellos.

HEADNODE	**CANTIDAD**
DL380p G8	
HP DL380 Gen8 Intel® Xeon® E5-2620v3 (2.4GHz/6-core/15MB/85W) Processor Kit	2
Memoria DDR4-2133 16GB MODULE	8
Flexible LOM HP InfiniBand FDR 2-port 545QSFP Adapter	1
Doble tarjeta de red PCI-express 1Gbps	1
Fuente redundante 800W gold	2
HP ILO REMOTE 1Gbps	1
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise 3yr Warranty Hard Drive	2
HP 1.2TB 6G SAS 10K rpm SFF (2.5-inch) SC Dual Port Enterprise 3yr Warranty Hard Drive	12
Smart Array Controller RAID 1/1+0 5/5+0 6/6+0	1
NODOS	

SL6500	
Fuente de redundantes HP 1200W/1500W Common Slot Platinum Plus Hot Plug Power Supply Kit	4
4 Fans redundantes	4
COMPUTE NODE 1	
SL250s G8	
Intel® Xeon® E5-2650v2 (2.6GHz/8-core/20MB/95W)	2
HP 16GB (1x16GB) Dual Rank x4 PC3-14900R (DDR3-1866) Registered CAS-13	8
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise	2
GPU Nvidia K40	1
Flexible-LOM Infiniband QLOGIC de 40Gbps dual port	1
Doble tarjeta de red PCI-express 1Gbps	1
HP ILO REMOTE 1Gbps	1
Smart Array RAID 1/1+0	1
COMPUTE NODE 2	
SL250s G8	
Intel® Xeon® E5-2650v2 (2.6GHz/8-core/20MB/95W)	2
HP 16GB (1x16GB) Dual Rank x4 PC3-14900R (DDR3-1866) Registered CAS-13	8
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise	2
GPU Nvidia K40	1
Flexible-LOM Infiniband QLOGIC de 40Gbps dual port	1
Doble tarjeta de red PCI-express 1Gbps	1
HP ILO REMOTE 1Gbps	1
Smart Array RAID 1/1+0	1
NETWORKING	
Switch Qlogic Infiniband	1
Switch 10/100/1000 HP 5800-24G Switch	1
Soportes HP	
HP Proactive Care	1
POWER	
R7000 Uninterruptible Power System (UPS)	1
HP PDUs	2
HP PDU de PDUs	2
RACK HP 42u	1

Figura 15. Diseño 3



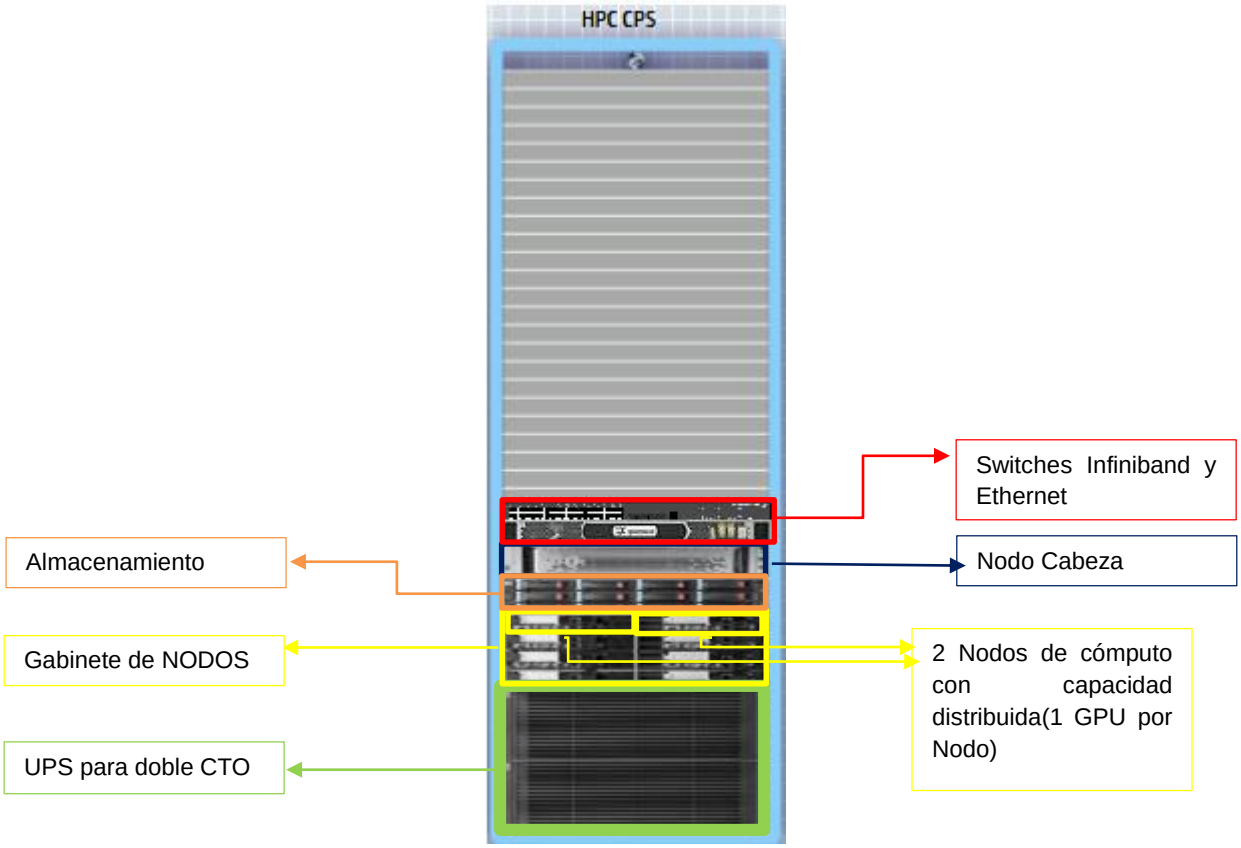
2.10 DISEÑO 4

En este diseño el clúster está conformado por nodo principal, dos nodos de cómputo con la memoria distribuida entre ellos, dos GPGPU y un almacenamiento especializado.

HEADNODE	CANTIDAD
DL380p G8	
HP DL380 Gen8 Intel® Xeon® E5-2620v3 (2.4GHz/6-core/15MB/85W) Processor Kit	2
Memoria DDR4-2133 16GB MODULE	8
Flexible LOM HP InfiniBand FDR 2-port 545QSFP Adapter	1
Doble tarjeta de red PCI-express 1Gbps	1
Fuente redundante 800W gold	2
HP ILO REMOTE 1Gbps	1
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise 3yr Warranty Hard Drive	2
HP 1.2TB 6G SAS 10K rpm SFF (2.5-inch) SC Dual Port Enterprise 3yr Warranty Hard Drive	12
Smart Array Controller RAID 1/1+0 5/5+0 6/6+0	1

NODOS	
SL6500	
Fuente de redundantes HP 1200W/1500W Common Slot Platinum Plus Hot Plug Power Supply Kit	4
4 Fans redundantes	4
COMPUTE NODE 1	
SL250s G8	
Intel® Xeon® E5-2650v2 (2.6GHz/8-core/20MB/95W)	2
HP 16GB (1x16GB) Dual Rank x4 PC3-14900R (DDR3-1866) Registered CAS-13	8
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise	2
GPU Nvidia K40	1
Flexible-LOM Infiniband QLOGIC de 40Gbps dual port	1
Doble tarjeta de red PCI-express 1Gbps	1
HP ILO REMOTE 1Gbps	1
Smart Array RAID 1/1+0	1
COMPUTE NODE 2	
SL250s G8	
Intel® Xeon® E5-2650v2 (2.6GHz/8-core/20MB/95W)	2
HP 16GB (1x16GB) Dual Rank x4 PC3-14900R (DDR3-1866) Registered CAS-13	8
HP 300GB 6G SAS 10K rpm SFF (2.5-inch) SC Enterprise	2
GPU Nvidia K40	1
Flexible-LOM Infiniband QLOGIC de 40Gbps dual port	1
Doble tarjeta de red PCI-express 1Gbps	1
HP ILO REMOTE 1Gbps	1
Smart Array RAID 1/1+0	1
NETWORKING	
Switch Qlogic Infiniband	1
Switch 10/100/1000 HP 5800-24G Switch	1
Almacenamiento	
HP MSA 2040 Storage(doble controladora)	1
HP P2000 3TB 6G SAS 7.2K LFF (3.5- inch) Dual Port MDL	5
Soportes HP	
HP Proactive Care	1
POWER	
R7000 Uninterruptible Power System (UPS)	1
HP PDUs	2
HP PDU de PDUs	2
RACK HP 42u	1

Figura 16. Diseño 4



2.11 COMPARATIVO DE DISEÑOS PROPUESTOS

Tabla 5. Comparativo de Diseños Propuestos

	DISEÑO 1	DISEÑO 2	DISEÑO 3	DISEÑO 4
CANTIDAD DE GPUS Y NODOS	1 Nodo y 2GPUs	2 Nodos y 4GPUs	2 Nodos 2GPUs	2 Nodos 2GPUs
ALMACENAMIENTO	En Headnode	En Headnode	En Headnode	En SAN
PERFORMANCE				
ESCALABILIDAD EN PROCESAMIENTO				
ESCALABILIDAD EN ALMACENAMIENTO				
VELOCIDADES DE COMUNICACIÓN				
DISPONIBILIDAD				

3. CONCLUSIONES

Se obtuvo la información necesaria y actualizada para la realización de un diseño ajustado a las necesidades del grupo de investigación CPS, con las características que permitirán la utilización correcta del hardware y su compatibilidad con las arquitecturas de cómputo para las simulaciones sísmicas.

Se seleccionaron los dispositivos idóneos para conformar una infraestructura convergente de High Performance Computing que permitan realizar las pruebas y simulaciones, que permitan la migración a infraestructuras más robustas como GUANE.

Se han diseñado 4 alternativas para la adquisición de la infraestructura convergente High Performance Computing que necesita el Grupo de investigación CPS donde se cumplen con los requerimientos de los proyectos que van a trabajar sobre dicha infraestructura y también los requerimientos necesarios para el correcto funcionamiento de un HPC, como lo es escalabilidad, alta disponibilidad, convergencia, administración remota y monitoreo continuo de toda la infraestructura.

BIBLIOGRAFÍA

ARAYA-POLO, M.; CABEZAS, J.; HANZICH, M.; PERICAS, M.; RUBIO, F.; GELADO, I.; SHAFIQ, M.; MORANCHO, E.; NAVARRO, N.; AYGUADE, E.; CELA, J. M. and VALERO, M. “Assessing Accelerator-Based HPC Reverse Time Migration,” IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 1, pp. 147–162, Jan. 2011.

KALCHER, Sebastian. DON’T FORGET THE “FABRIC”, pp. 3–4, 2013.

NVIDIA, NVIDIA’s Next Generation CUDATM, Compute Architecture: Kepler TM GK110/210, pp 15-20, 2014

ORTEGA CARRASCO, Cristobal. Diseño y evaluación de un clúster HPC: Software de sistema, pp. 22–23, 2014

SCHROEDER, Bianca; GIBSON, Garth A. A large-scale study of failures in high-performance-computing systems

SYMMETRIC COMPUTING. Distributed Symmetric Multi-Processing (DSMP), pp. 4-8 2013