Tarántula: Una grid de clusters de cómputo para el desarrollo de aplicaciones paralelas y en grid en México

R. Rivera-López[1*], A. Rodríguez-León[1], M. A. Cruz-Chávez[2], I. Y. Hernández-Baez[3]
[1]*Laboratorio de Cómputo Intensivo, Instituto Tecnológico de Veracruz, M. A. de Quevedo 2779, Col. Formando Hogar, 91860,Veracruz, Ver., México*
[2]*Centro de Investigación en Ingeniería y Ciencias Aplicadas, Universidad Autónoma del Estado de Morelos, Av. Universidad 1001 Col. Chamilpa, 62209, Cuernavaca, Morelos, México*
[3]*Departamento de Informática, Universidad Politécnica del Estado de Morelos, Boulevard Cuauhnáhuac 566, Col. Lomas del Texcal, 62550, Jiutepec, Morelos, México.*
**rrivera@itver.edu.mx*
***Área de participación**: Sistemas Computacionales*

Resumen

En este artículo se describen los elementos que se utilizaron para la implementación de un grupo de clusters de computadoras que se integraron para formar una grid de cómputo de alto desempeño entre el Instituto Tecnológico de Veracruz (ITVer), la Universidad Autónoma del Estado de Morelos (UAEM) y la Universidad Politécnica del Estado de Morelos (UPEMor). Estos clusters sirven como infraestructura de hardware y software para el desarrollo de proyectos de investigación en cómputo paralelo y distribuido en estas instituciones. En este documento se describe el esquema de comunicaciones que se utiliza para distribuir tareas entre los procesadores de la grid, y se subraya la importancia del uso de clusters para resolver problemas computacionalmente complejos y se describen los proyectos de investigación que se han desarrollado utilizando esta infraestructura en las tres instituciones.

***Palabras clave**: Cómputo Paralelo, Computación Grid*

Abstract

In this paper the elements for the construction of a groups of computer clusters in the Instituto Tecnológico de Veracruz (ITVer), the Universidad Autónoma del Estado de Morelos (UAEM) and the Universidad Politécnica del Estado de Morelos (UPEMor). These clusters are used as software and hardware infrastructure for the development of research projects in parallel and distributed computing for these institutions. This paper describes the communication scheme used to distribute tasks among the processors in the grid, and underline the importance of using clusters to solve computationally complex problems and describes the research projects that have been developed using this infrastructure.

Introducción

Aún cuando el incremento del poder de cómputo de una computadora personal crece de manera vertiginosa por los avances tecnológicos, existen en la actualidad problemas cuyo tiempo de procesamiento hace prácticamente imposible su resolución en tiempos razonables por computadoras con procesadores de gran velocidad, ya que su tiempo de ejecución crece exponencialmente en función al tamaño de los datos de entrada. La teoría de la complejidad computacional estudia la eficiencia de los algoritmos en la resolución de problemas y ha propuesto que los problemas se clasifican en problemas P, NP y NP-Completos. Los problemas NP-Completos son aquellos problemas que en la actualidad se conocen solo algoritmos ineficientes para resolverlos, por lo que la comunidad científica ha abordado varios esquemas de desarrollo para tratar de reducir su tiempo de resolución, como es el estudio teórico de los algoritmos, la hibridación de los algoritmos y la distribución de las tareas de un algoritmo en varios procesadores.

El cómputo paralelo es la rama de la computación que se enfoca en la ejecución de más de una instrucción de un algoritmo de manera simultánea usando mas de un procesador en una computadora [Bertsekas y Tsitsiklis,

1997]. El cómputo paralelo ha permitido que se puedan implementar soluciones eficientes a problemas que son computacionalmente complejos cuya solución utilizando un algoritmo secuencial consume un tiempo de cómputo excesivo, para una gran variedad de áreas de conocimiento que abarcan desde simulaciones para científicos aplicaciones de ingeniería hasta aplicaciones comerciales con procesamiento de transacciones y minería de datos [Grama y col. 2003]. Inicialmente el cómputo paralelo se implementó en supercomputadoras, pero en la actualidad el uso de grupos de computadoras personales (o clusters) ha renovado el uso intensivo del cómputo paralelo. Se dice que un cluster es un tipo de sistema de procesamiento distribuido o paralelo que consiste de una colección de computadoras personales interconectadas trabajando unidas como un solo recurso integrado [Morrison, 2003]. Para aprovechar los recursos de cómputo distribuidos geográficamente, en la actualidad las organizaciones han empezado a diseñar redes de clusters, la cuales se denomina una grid, que implica agrupaciones de computadoras unidas por redes de área extensa (WAN).

En México, desde hace pocos años algunas instituciones de investigación y de educación superior han invertido en la construcción de clusters de computadoras para ser aplicadas en el desarrollo de proyectos de investigación en una gran cantidad de áreas de conocimiento (por ejemplo, en [Acosta y col., 2004] y [Valdez y Campos, 2008]). En el ITVer, a partir del año 2003, se inició con el uso de un cluster de computadoras denominado Nopal, en diferentes proyectos de investigación del Departamento de Sistemas y Computación. En la actualidad, con el apoyo de la Dirección General de Educación Superior Tecnológica (DGEST), del Consejo Nacional de Ciencia y Tecnología (CONACyT), del Programa de Mejoramiento del Profesorado (PROMEP) y de la Corporación Universitaria para el Desarrollo del Internet, A.C. (CUDI), se ha construido varios clusters de alto rendimiento para apoyar el desarrollo de proyectos de investigación. Aunado a la iniciativa de crear un laboratorio nacional de grids por las principales instituciones de educación superior en México, el ITVer, la UAEM y la UPEMor se abocaron a compartir los recursos de sus clusters de alto desempeño implementando una grid de cómputo comunicada utilizando internet 2, la cual se denomina Tarántula.

Este documento presenta una descripción de los clusters que constituyen la grid Tarántula, así como un perfil de hardware y su rendimiento, y una descripción de los proyectos de investigación que se han desarrollado y que utilizan estos clusters.

Cómputo Paralelo

En el pasado para realizar proyectos que requerían de un poder de cómputo mayor al que podían ofrecer las computadoras personales se recurría al uso de supercomputadoras. De acuerdo con [Suppi, 2010], una supercomputadora puede definirse como una computadora con capacidades de cálculo muy superiores a las de los equipos de cómputo comunes, pueden procesar enormes cantidades de información en poco tiempo pudiendo ejecutar millones de instrucciones por segundo, están destinadas a una tarea específica y poseen una capacidad de almacenamiento muy grande, además cuentan con sistemas especiales de enfriamiento, sin embargo, éstas tienen un altísimo costo, por lo que su uso generalmente se limita a organismos militares, gubernamentales y grandes centros de investigación. Su aplicación se encuentra en las áreas científicas, como es la simulación de procesos naturales como los cambios climáticos, modelaje molecular, simulaciones físicas, criptoanálisis, entre otros.

Una alternativa al uso de supercomputadoras fue el uso de clústers de computadoras, debido a que su implementación resulta más barata haciéndoles accesibles a un mayor número de usuarios. De acuerdo a [Suppi, 2010], se denomina clúster a la agrupación de computadoras que trabajan con un fin común. Estas computadoras agrupan hardware, redes de comunicación y software para trabajar conjuntamente como si fueran un único sistema. Existen muchas razones atrayentes para realizar estas agrupaciones, pero la principal es poder efectuar el procesamiento de la información de forma más eficiente y rápida como si fuera un único sistema. Generalmente, un clúster trabaja sobre una red de área local (LAN) y permite una comunicación eficiente si las máquinas se encuentran dentro de un espacio físico próximo. Una concepción mayor del concepto de clúster es la llamada grid, donde el objetivo es el mismo, pero implica agrupaciones de computadoras unidas por redes de área extensa (WAN). Algunos autores consideran el grid como un clúster de clústers en un sentido 'global'.

Tarántula: Una grid experimental para resolver problemas complejos

En el año 2008, con la colaboración de los cuerpos académicos de Optimización y Software de la UAEM (UAEMOR-CA-87) y de Cómputo Intensivo Aplicado a la Ingeniería (ITVER-CA-1), se participó en un proyecto para el uso de Internet 2 patrocinado por la Corporación Universitaria para el Desarrollo de Internet (CUDI) donde se proponía el Estudio de Modelos Teóricos de tipo NP-completos en el Laboratorio Nacional de GRIDS de Súper Cómputo, Utilizando Algoritmos Evolutivos de Optimización con Técnicas de Procesamiento Distribuido, pero debido a las dificultades técnicas para la integración de dicho laboratorio, se determinó implementar una miniGrid entre las dos instituciones académicas, dando origen a la primera versión de la miniGrid Tarántula (figura 1). En el 2009 se integró a los trabajos de la grid la UPEMor, por lo que la grid Tarántula se encuentra configurada como se presenta en la figura 2.

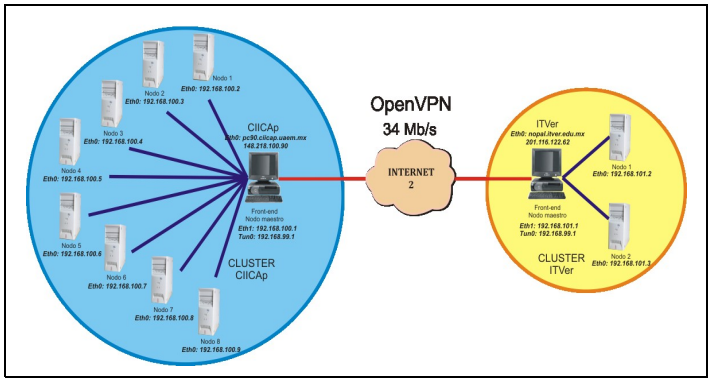

Figura 1.- Primera versión de la grid Tarántula.

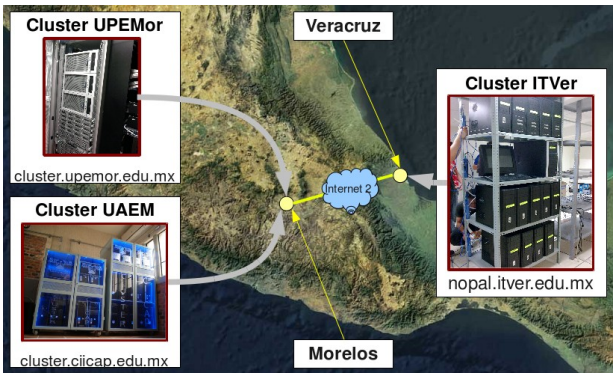

Figura 2.- Versión actual de la grid Tarántula.

El escenario para la integración de los clusters en una grid experimental es que se encuentran alejados geográficamente, además de ser administrados localmente y son por si mismo entidades independientes. Este panorama de integración presentó múltiples dificultades, la más difícil de ellas, fue como resolver la comunicación entre los diferentes nodos que no pertenecen al mismo dominio, esto es, que no pertenecen al mismo Cluster y poder responder a la pregunta ¿como alcanzar los nodos de otro Cluster que se encuentra alejado geográficamente. La idea fue poder ver en forma transparente los Clusters unidos como uno sólo y poder ejecutar programas MPI en esta grid. Los clusters que se unen en la grid Tarántula son el cluster Nopal del ITVer, el cluster CIICAp de la UAEM, y el cluster UPEMor de la misma institución

El cluster Nopal en el IT Veracruz

La versión actual del cluster Nopal tiene 15 computadoras, diez con procesadores Intel Pentium 4 dual-core y cinco con procesadores Intel Prentium 4 cuad-core, constituyendo un cluster de cuarenta nodos. El front-end tiene un velocidad de procesamiento de 3.2 Ghz y las otras computadoras tienen una velocidad de 2.33 Ghz. La capacidad de almacenamiento del cluster es de 1.2 TB y una memoria total de 57 GB. El sistema operativo es Rocks 5.2 que es una distribución de Linux para clusters de alto rendimiento basado en CentOS 5 (figura 3).

El cluster CIICAp en la UAEM

El cluster CIICAp (figura 4) cuenta con un front-end con procesador Pentium 4 a 2.8 Ghz, y 18 nodos esclavos procesador Intel Celeron Dual Core a 2.0 Ghz. La capacidad de almacenamiento del cluster es de 2.9 TB y una memoria total de 36 GB. También usa el sistema operativo Rocks 5.2.

El cluster UPEMor

Este cluster (figura 5) cuenta con un front-end Sunfire x2270w con procesador Intel Core 2 Duo a 2.6 Ghz y 4 nodos esclavos con las mismas características. La capacidad de almacenamiento del cluster es de 2.5 TB y una memoria total de 20 GB. También usa el sistema operativo Rocks 5.2.

Configuración de los Clusters

Los componentes básicos para construir un Cluster de alto rendimiento con soporte para programación paralela con OpenMPI y MPICH esta basado en la configuración de la red privada, de los sistema de archivos de red usando Network File System (NFS), de los servicios de información de red usando Network Information Service

(NIS), de las llaves públicas para habilitar la ejecución de trabajos remotos, de MPI para usar de OpenMPI y MPICH y de Ganglia para el monitoreo en tiempo real del Cluster.

Figura 3.- El cluster Nopal del ITVer.

Figura 4.- El cluster CIICAp de la UAEM.

Figura 5.- El cluster UPEMor.

La integración de todos estos componentes es lo que da vida a un Cluster de alto rendimiento como se muestra en la figura 6.

Configuración de la grid Tarántula

El software intermedio para la administración de recursos de la red (middleware) está configurado como una red privada virtual (VPN) que, a través de la definición de un túnel de comunicaciones, comparte los recursos de los clusters configurados en diferentes segmentos de red [Mache, 2006]. La Grid está conformada por tres clusters de alto rendimiento, alejados geográficamente y que se unen a través de una Red Privada Virtual red-a-red utilizando OpenVPN. En principio, los clusters contienen una diferente subred, el cluster CIICAp se configuró como una subred 192.168.100.0, el cluster Nopal como una subred 192.168.101.0 y el cluster UPEMor se configuró como una subred 192.168.102.0 (figura 7).

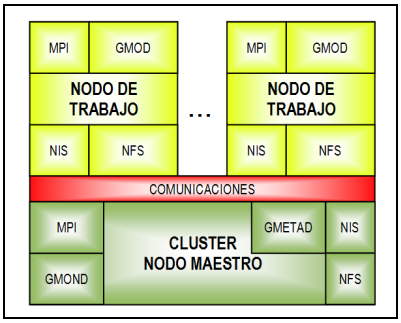

Figura 6.- Componentes de un cluster.

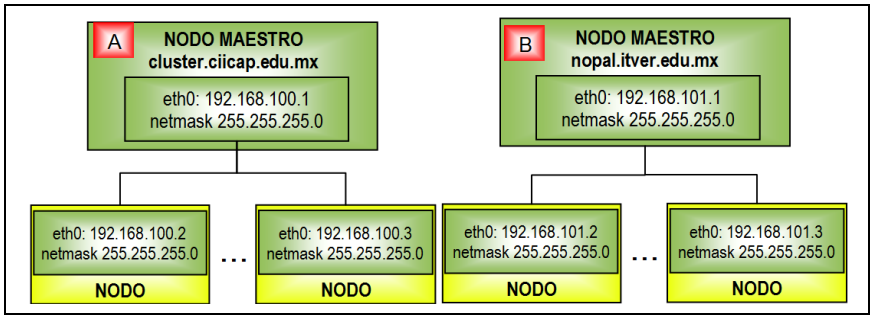

Figura 7.- Clusters en diferentes subredes

Configuración de OpenVPN: OpenVPN es una solución de conectividad basada en software que ofrece conectividad punto a punto a través de Internet con validación de usuarios y computadoras conectadas remotamente. OpenVPN es una implementación vía software para construir una red privada virtual sin necesidad de routers vía Hardware, soporta diferentes medios de autenticación como certificados, usuarios y contraseñas.

La configuración de OpenVPN se puede resumir en tres tipos: máquina a máquina, road warrior y red a red. Al configurar OpenVPN varias redes separadas y alejadas geográficamente pueden unirse y alcanzar mutuamente sus recursos, la comunicación entre ambas redes viajará encriptada una vez que salga de los servidores de OpenVPN y hasta que llegue al otro extremo. La idea está basada en poder usar un único canal de comunicación llamado túnel para enviar todas las comunicaciones de un extremo a otro y viceversa, permitiendo ver las máquinas conectadas al otro extremo como si estuvieran unidas físicamente a través de un cable.

Para poder configurar una conexión red a red es necesario definir quien será servidor y quienes los clientes, en la grid Tarántula se usa el cluster CIICAp como servidor y los otros como clientes, donde las únicas computadoras conectadas son los front-end de cada cluster. Para poder alcanzar los nodos que están atrás de los extremos de la VPN es necesario implementar encaminamiento de paquetes.

Configuración servidor: La instalación del middleware OpenVPN se puede hacer vía compilación de los archivos fuente o vía instalación de paquete RPM. En esta grid esta configuración se basó en la descarga, desempaquetado, compilado, construcción de binarios y edición del archivo de configuración. Posteriormente es necesario configurar la ejecución del middleware de forma que sea automático durante el arranque de sistema. En este punto ya esta listo para esperar las conexiones de parte del cliente (figura 8).

Configuración cliente: La instalación del middleware del lado del cliente es igual al procedimiento usado en el servidor, con la excepción de que el archivo de configuración es diferente, ya que este indica a donde debe conectarse mientras que el lado servidor indica en donde estará esperando recibir conexiones (figura 9).

```
[root@cluster ~]# cat /etc/openvpn/server.conf
port 1723
proto tcp
dev tun
ca ca.crt
cert server.crt
key server.key
dh dh1024.pem
server 192.168.99.0 255.255.255.0
ifconfig-pool-persist ipp.txt
client-config-dir ccd
route 192.168.101.0 255.255.255.0
client-to-client
push "route 192.168.101.0 255.255.255.0"
push "route 192.168.100.0 255.255.255.0"
keepalive 10 120
comp-lzo
user nobody
group nobody
persist-key
persist-tun
status openvpn-status.log
verb 4
```

Figura 8.- Archivo de configuración para el Servidor OpenVPN.

```
root@nopal:~> cat /etc/openvpn/client1.conf
client
dev tun
proto tcp
remote 148.218.100.90 8080
resolv-retry infinite
nobind
#Las dos siguientes opciones no van en windows
user nobody
group nobody

persist-key
persist-tun
ca ca.crt
cert client1.crt
key client1.key
comp-lzo
verb 4
```

Figura 9.- Archivo de configuración para el Cliente OpenVPN.

Esquema de comunicación para la grid Tarántula

Con la intención de resolver problemas computacionalemente completos utilizando todos los recursos de la grid Tarántula, se define un esquema de comunicación entre los nodos de la grid [Rodríguez y col, 2010]. Este esquema de comunicación tiene dos etapas: la primera fase define el procedimiento para la combinación de los datos procesados entre todos los nodos en un cluster, utilizando los métodos de MPI y la segunda fase define la transmisión de un grupo de datos procesados entre los clusters de la grid utilizando el protocolo FTP para reducir el tráfico en de la red.

Fase 1. El envío de los segmentos entre los nodos en un cluster
La figura 10 muestra que cada nodo (1, 2, ..., n) en un solo cluster (Nopal, CIICAp o UPEMor), después de aplicar el proceso sobre los datos, envía los resultados a los otros nodos del cluster, utilizando los métodos MPI_Send() y MPI_Recv(). En la figura 10, s_1 representa los datos procesados en el nodo 1, s_2 es el segmento de datos procesados por el nodo 2, y así sucesivamente. El método MPI_Barrier() se utiliza para asegurar que todos los nodos en un cluster han compartido sus resultados antes de aplicar, utilizando el protocolo FTP, la transmisión del conjunto de segmentos entre los clusters de la grid Tarántula.

Fase 2. El envío de los segmentos entre los cluster de la grid
Para combinar los segmentos de datos entre los clusters de la grid Tarántula, es necesario compartir los segmentos entre todos los nodos de cada grupo. Si los segmentos se transmiten de forma independiente para cada nodo de la grid, aumentaría el tráfico de red y esto podría reducir el rendimiento de los programas. Para reducir este tráfico se utiliza un nodo de cada clustercomo responsable de enviar, a través de FTP, el grupo de

los resultados entre los clusters (figura 11). Los nodos seleccionados en cada cluster se utilizan para crear un archivo con los segmentos de resultados de este cluster. Este archivo es enviado a los otros clusters de la grid protocolo FTP. Utilizando un mensaje de MPI a través de la VPN, el nodo seleccionado en el cluster se comunica los otros para indicar que la transmisión del archivo se ha completado.

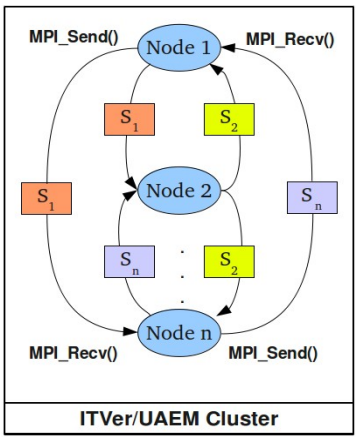

Figura 10.- Combinación de datos procesados en un cluster

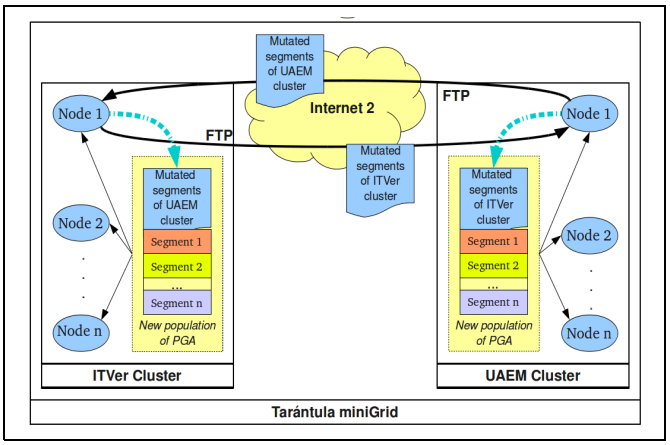

Figura 11.- Envío de segmentos de resultados entre los cluster de la grid Tarántula.

Este esquema permite reducir el tráfico entre los cluster de la grid, ya que al pasar la información por internet, se está sujeto a las condiciones cambiantes del ancho de banda y de la calidad del servicio. Una de las condiciones que se deben estudiar es el incremento de las transacciones vía FTP al incluir mas clusters a la grid.

Proyectos de Investigación

La grid Tarántula se diseñó para que sea utilizando por cualquier investigador de las instituciones involucradas. La tabla 1 muestra una relación de los proyectos de investigación que se han desarrollado utilizando la infraestructura de la grid Tarántula. En estos proyectos se han integrado muchos alumnos de licenciatura y posgrado de las tres instituciones, así como se han producido mas de quince artículos y cinco tesis de licenciatura y posgrado.

Tabla 1.- Proyectos de Investigación que utilizan la grid Tarántula.

Título del Proyecto	**Responsables**
Estudio de Modelos Teóricos de tipo NP-completos en el Laboratorio Nacional de GRIDS de Súper Cómputo, utilizando Algoritmos Evolutivos de Optimización con Técnicas de Procesamiento Distribuido	Dr. Abelardo Rodríguez León (ITVer) Dr. Marco A. Cruz Chávez (UAEM) Financiado por: CUDI-Conacyt
Sistema MultiAgente para Distribución por demanda de vídeo de alta definición sobre Internet 2	Dr. Abelardo Rodríguez León (ITVer) Financiado por: DGEST
Estudio Experimental del Efecto del Paso de Mensajes en ambiente Grid para el Desarrollo de Sistemas que Tratan con Problemas NP-Completos.	Dr. Abelardo Rodríguez León (ITVer) Dr. Marco A. Cruz Chávez (UAEM) M.C. Irma Y. Hernández Báez (UPEMor) Financiado por: CUDI-Conacyt
Estudio de Modelos para la Asignación de Recursos en Talleres de Manufactura en Ambiente Grid con Estancias de Tamaño Grande	Dr. Marco A. Cruz Chávez (UAEM) Financiado por: UAEM
Construcción y Fortalecimiento de una MiniGrid en el Estado de Morelos para Proyectos de Investigación en e-Ciencia	Dr. Marco A. Cruz Chávez (UAEM)

Tabla 1.- Proyectos de Investigación que utilizan la grid Tarántula.

Título del Proyecto	Responsables
	Financiado por: Conacyt
Optimización por un Algoritmo Genético en Ambiente Grid, para la Asignación de Recursos en una Cadena de Suministros	Dr. Marco A. Cruz Chávez (UAEM) Financiado por: UAEM

Conclusiones y Trabajo Futuro

La siguiente etapa de crecimiento de la grid Tarántula es la creación de la Grid Morelense de Alto Rendimiento [Grid Morelense, 2010]. La Grid Morelense de Cómputo de Alto Rendimiento busca incrementar el número de Clusters conectados a la grid Tarántula con la integración de otras instituciones como la Universidad Recnológica Emiliano Zapata y el Instituto Tecnológico de Zacatepec (figura 12).

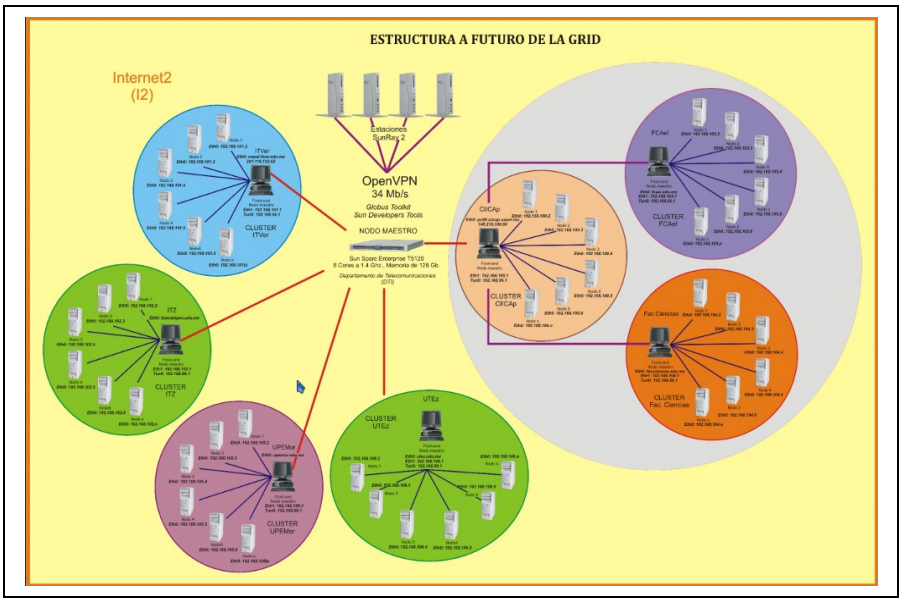

Figura 12.- Proyecto de la Grid Morelense de Alto Desempeño.

Agradecimientos

La construcción de la grid Tarántula ha sido apoyada con el financiamiento de CONACyT, Promep, CUDI, la UAEM, la UPEMor y la Dirección General de Educación Superior Tecnológica de la Secretaría de Educación Pública de México.

Referencias

1. Bertsekas D.P., y Tsitsiklis J.N.: Parallel and Distributed Computation: Numerical Methods, Athena Scientific (1997).
2. Grama, A., Gupta, A., Karypis, G., y Kumar, V.: Introduction to Parallel Computing, 2nd. Edición, Addison Wesley (2003).
3. Morrison, R. S.: Cluster Computing. Architectures, Operating Systems, Parallel Processing & Programming Languages (2003).
4. Acosta, R., García-Ruiz, M., Banda. C.A., Barajas, O.A., Ramírez, J.M., Reyes, P.D., y Bustos, C.R.: Implementación de un Cluster de alto rendimiento como herramienta para resolver problemas de cómputo científico, en Memorias de la 3a. Conf. Iberoamericana en Sistemas, Cibernética e Informática, CICSI 2004, Julio de 2004.

5. Valdez, A.G., y Campos, G.: Creación de un Cluster de Linux utilizando Knoppix. Sociedad de la Información, No. 23, Junio de 2008.
6. Suppi Boldrito, R., Clustering, Universitat Oberta de Catalunya (2010).
7. Página principal del Laboratorio Grid de Alto Rendimiento (http://www.gridmorelos.uaem.mx:8080/).
8. Mache, J., Tyman, D., Pinter, A., Allick, C.: Performance Implications of Using VPN Technology for Cluster Integration and Grid Computing, Int. Conf. on Networking and Services, pp. 75-80 (2006).
9. Rodriguez-León, A., Cruz-Chávez, M. A., Rivera-López, R., Ávila-Melgar, E. Y., Juárez-Pérez, F., Cruz-Rosales, M. H.: A Communication Scheme for an Experimental Grid in the Resolution of VRPTW using an Evolutionary Algorithm, Proc. of Electronics, Robotics and Automotive Mechanics Conf., CERMA2010, IEEE-Computer Society, pp. 108 – 113 (2010).

Autorización y renuncia

Los autores del presente artículo autorizan al Instituto Tecnológico de Orizaba (ITO) para publicar el escrito en el libro electrónico del coloquio de investigación multidisciplinaria, en su edición 2011. El ITO o los editores no son responsables ni por el contenido ni por las implicaciones de lo que está expresado en el escrito.