Implementación y Operación de un Cluster HPC utilizando Laboratorios de Computadoras en Horarios de Inactividad

Ernesto Sale[1], Sebastián Rodríguez[2]

[1,2]Grupo de Investigación en Tecnologías Informáticas Avanzadas

Facultad Regional Tucumán - Universidad Tecnológica Nacional

Rivadavia 1050 - (4000) S.M. de Tucumán

[1]ernesto.sale@gitia.org, [2]sebastian.rodriguez@gitia.org

Resumen: *Este trabajo presenta un reporte preliminar de una plataforma para facilitar a alumnos, docentes e investigadores el acceso a clusters HPC a costos accesibles utilizando recursos existentes en las instituciones donde desempeñan sus actividades.*

Los laboratorios de informática de las universidades constituyen un reservorio de potencia de procesamiento. Estos pueden ser utilizados para conformar un cluster de cálculo de alto desempeño en los horarios que no se encuentran ocupados.

De esta manera se puede al mismo tiempo reducir los costos de poseer un cluster HPC, ampliar el abanico de servicios y posibilidades que la institución brinda a sus miembros y aumentar la utilización de los laboratorios mejorando su amortización y por ende justificando mejor la inversión.

Palabras Claves: *Cluster HPC, cluster de alto rendimiento, utilización recursos existentes entorno educativo, cluster by night, cluster en laboratorio.*

Abstract: *This work presents a preliminary report on a platform to implement an affordable HPC cluster to provide access to students, teacher and investigators, by using resources already available in the institutions where they perform their activities.*

The universities' computer laboratories are a processing power reservoir. They can be used to build a high performance computing cluster during their inactive periods of time.

This way it is possible at the same time to reduce HPC cluster's cost, to widen the institution's variety of services and possibilities for its members, and to increase the lab usage, improving its amortization and thus better justifying the investment.

Keywords: *Cluster HPC, high performance cluster, use existing resources educational environment, cluster by night, lab cluster.*

INTRODUCCIÓN

A medida que la investigación avanza en diversos campos encarando tareas más complejas, trae consigo mayores requisitos de procesamiento de datos. Por otro lado el avance en las herramientas informáticas y la recolección de datos hacen posible enfocar los problemas desde nuevos ángulos, siempre y cuando se posea la capacidad de procesamiento adecuada. Esta es la realidad de aplicaciones tales como predicción de clima, predicción de consumo eléctrico de poblaciones, simulaciones de plegamiento de proteínas, diseño de automóviles, predicción financiera, simulación de terremotos, exploración petrolera y data mining, entre muchas otras, las cuales manejan cantidades de datos y/o algoritmos que requieren recursos de computación a gran escala (Engineering and Physical Sciences

Research Council, 2004 y 2005; Bose et al., 2010).

Esto presenta una oportunidad para que centros de investigación e instituciones educativas universitarias se desarrollen y mantengan actualizados en la provisión de altas capacidades de cálculo y a la vez la amenaza de quedar relegados como meros espectadores en el avance de los conocimientos en diversas áreas como las mencionadas previamente.

Originalmente esta capacidad era provista por equipos y sistemas que resultaban costosos, a medida del cliente, fabricados por empresas como Cray, Convex, Tera Computer, SGI, IBM y HP.

En 1994, desde el Goddard Space Flight Center de la NASA, hicieron su aparición los clusters del proyecto Beowulf (Gropp et al., 2003; Ridge et al., 1997) que integran PCs, Linux y una red de comunicación altamente difundida (Ethernet) para formar una computadora paralela y que bajaron dramáticamente la barrera para la posesión de una supercomputadora (Adams y Vos, 2002).

No obstante hay numerosas instituciones que siguen impedidas de superar esta barrera por no contar con los recursos necesarios (dinero, equipos, PCs, personal capacitado) para construir y operar un High Performance Computing cluster (en adelante llamado cluster HPC) dedicado.

Sin embargo, muchas mantienen vastos laboratorios de computación donde están congregados todos los ingredientes físicos necesarios para la operación de un cluster HPC tipo Beowulf que podrían ser utilizados a tales fines en los horarios que no están siendo ocupados para sus funciones tradicionales, por ejemplo, durante la noche o los días feriados.

Si bien la capacidad de un sistema como el sugerido sería inferior a la de un sistema con recursos dedicados, es justo lo indicado para numerosas instituciones que, por razones de volumen de trabajo y/o de presupuesto, no pueden justificar el costo de tener un cluster HPC dedicado y como consecuencia sus alumnos, docentes e investigadores no pueden desarrollar tareas de I+D+i que requieran alta capacidad de procesamiento. Por otro lado, quienes realmente necesiten hardware dedicado, pueden utilizar este sistema como una herramienta de medición que les permita justificar objetivamente dicha necesidad. Adicionalmente serviría para capacitar los recursos humanos necesarios para administrar y mantener el cluster.

Para lograr la implementación de un sistema como el mencionado es preciso resolver algunos problemas. El acceso físico a las máquinas es uno de ellos, dado que típicamente los laboratorios e incluso los edificios que los contienen se encuentran cerrados por las noches y los días Domingo y feriados todo el día. Otro problema es la configuración de los equipos que debe ser ajustada con cuidado para no afectar el normal funcionamiento del laboratorio para su función principal. A ésto se puede agregar restricciones derivadas de políticas de los administradores de los laboratorios (por ejemplo no modificar el contenido del disco rígido) y de los administradores de la red (por ejemplo impidiendo acceso desde Internet), entre otros.

Este reporte presenta un enfoque para transformar un laboratorio de computadoras en un cluster HPC durante sus horarios de inactividad y los resultados parciales de su implementación en uno de los laboratorios de la Facultad Regional Tucumán de la Universidad Tecnológica Nacional.

El objetivo de este trabajo es proponer una solución que:

1. Permita el empleo de equipamiento e infraestructura existentes.

2. Sea escalable.

3. Sea reutilizable (en otras instituciones educativas, por ejemplo).

4. No requiera conocimientos avanzados de sistemas operativos distribuidos para su configuración, administración y operación.

5. Posea una interfaz para el usuario sencilla de utilizar.

6. Sea útil para la investigación y la docencia.

7. Utilice software gratuito y de código abierto posible de mejorar.

8. Posea robustez y presente un desempeño estable y predecible.

9. Sea lo más independiente posible de la administración del laboratorio y lo menos invasiva posible de su configuración preestablecida.

La estructura de este documento presenta una breve reseña sobre clusters HPC y los trabajos relacionados, una descripción estructural y funcional de la solución propuesta, la plataforma de evaluación utilizada para prueba del concepto, los resultados obtenidos y por último las conclusiones y líneas de trabajo futuro.

TRABAJOS RELACIONADOS

En este trabajo sólo serán evaluadas las soluciones gratuitas y de código abierto buscando respectivamente el menor costo posible y brindar un aspecto más de utilidad a la docencia e investigación habilitando la posibilidad que estos sistemas sean evaluados y mejorados.

RESEÑA DE CLUSTERS HPC

En esta sección se intenta explicar qué es HPC y cómo se implementa, sus características sobresalientes, las necesidades que atiende, las oportunidades que genera, su propósito y las soluciones alternativas a su uso.

Para una mejor comprensión se puede proveer en este punto, como una definición aproximada y a priori del concepto de cluster, como un conjunto de máquinas que trabajan concertadamente entre sí (Slater, 1997).

COMPUTACIÓN DE ALTO DESEMPEÑO - HIGH PERFORMANCE COMPUTING

Desde sus comienzos teóricos la informática (tratamiento automático de datos por medio de máquinas) fue vista como una forma de superar las limitaciones humanas en lo que a velocidad, complejidad y capacidad de cálculo se refiere. Dicho de otro modo, al programar máquinas para que realicen los cálculos por nosotros podemos obtener resultados confiables procesando cantidades de datos, complejidades de algoritmos y en tiempos humanamente imposibles, accediendo así a información que está más allá de nuestro alcance natural. Y esta necesidad de llegar más lejos siempre estuvo empujando los límites para aumentar la potencia de esta herramienta. Ya sea para probar nuevos enfoques, para incrementar el detalle de alguna simulación o incluso con fines publicitarios (por ejemplo Deep Blue de IBM vs. Kasparov), hay una constante avidez para expandir las capacidades de las computadoras a nuestra disposición.

Las aplicaciones para este tipo de maquinaria son numerosas, al punto que no resultaría práctico citarlas en extensión, pero para tener una idea de las dimensiones y variedad se mencionan como ejemplos las siguientes:

Finanzas: aplicaciones para medir riesgos (incluyendo catástrofes como terremotos, huracanes, enfermedades, etc.), estimar tasaciones, simular balances, generadores de números aleatorios, puentes brownianos, soluciones de ecuaciones diferenciales parciales, exploración y extracción de información de big data.

Defensa e Inteligencia: visualización y análisis geoespacial, “recuperación” de contraseñas, análisis de video para reconocimiento y seguimiento de objetos, búsqueda de objetos en videos e imágenes en vivo y diferido, modelado de propagación de señales electromagnéticas para ambientes urbanos y terrenos complejos.

Industriales: CAD, dinámica de fluídos, simulaciones y análisis para mecánicas estructurales, paquetes multifísicos, simuladores para diseño de circuitos digitales de radio frecuencia, microondas

y alta velocidad, simulaciones de campos electromagnéticos en circuitos, softwares para acelerar el proceso de litografía, modelado electromecánico. Aplicaciones para la industria del entretenimiento y televisión, tales como modelado 3D, animación y rendering, corrección de colores, restauración y remastering de imágenes, efectos visuales, edición y trascodificación de video, sistemas de gráficos en el aire y climatológicos. Aplicaciones para la industria petrolera, tales como modelaje, procesamiento e interpretación sísmica y modelaje de reservorios.

Investigación: aplicaciones bioinformáticas para mapeo de secuencias, alineación de nucleótidos, cuantificar la contribución de la variación genética en la variación genotípica, aplicaciones para ciencias de los materiales.

Hay que tener en cuenta que esta carrera por la potencia y/o capacidad de cálculo no es tal en otros casos, como sucede en general en el mercado de las computadoras de escritorio, donde es la industria quien intenta impulsar los avances en sus clientes/usuarios, cuyas aplicaciones en general (excepto los juegos) y a diferencia de las mencionadas previamente, ocupan un porcentaje relativamente pequeño de las capacidades de los equipos en los cuales se desempeñan.

El objetivo es, en síntesis, resolver problemas "grandes" queriendo decir que, por la cantidad de datos a procesar y/o los algoritmos involucrados, excederían la capacidad de una computadora de escritorio (en RAM, por ejemplo) o bien que la magnitud del tiempo necesario para finalizar el cálculo resultaría poco práctica.

La computación de alto desempeño o high performance computing (HPC, por su nombre en inglés), parece no estar claramente definida. Tampoco su relación con la supercomputación (supercomputing). Buscando en Internet se encuentran cosas tales como:

High-performance computing: (n.) A branch of computer science that concentrates on developing supercomputers and software to run on supercomputers. A main area of this discipline is developing parallel processing algorithms and software: programs that can be divided into little pieces so that each piece can be executed simultaneously by separate processors (http://www.webopedia.com/TERM/H/High_Performance_Computing.html).

High-performance computing (HPC) is the use of parallel processing for running advanced application programs efficiently, reliably and quickly. The term applies especially to systems that function above a teraflop or 1012 floating-point operations per second. The term HPC is occasionally used as a synonym for supercomputing, although technically a supercomputer is a system that performs at or near the currently highest operational rate for computers. Some supercomputers work at more than a petaflop or 1015 floating-point operations per second (http://searchenterpriselinux.techtarget.com/definition/high-performance-computing/).

High Performance Computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business (http://insidehpc.com/hpc-basic-training/what-is-hpc/).

The term high performance computing (HPC) refers to any computational activity requiring more than a single computer to execute a task (http://www.hpcwales.co.uk/what-is-hpc/).

Wikipedia (http://en.wikipedia.org/wiki/HPC) directamente redirige a la página dedicada a supercomputación.

Se puede observar que algunas fuentes situan la supercomputación y HPC como alternativas, otras como la Wikipedia sostienen que un cluster

HPC es parte de la supercomputación y otras (como webopedia) dicen lo inverso. Para entender la relación y el porqué de la diversidad a la hora de definirla ayuda revisar algo de la historia de estas disciplinas.

La supercomputación fue el enfoque original para encarar estos problemas "grandes" mencionados y consistía en la utilización de una computadora de capacidades ampliamente superiores a un equipo convencional, diseñada específicamente a medida del cliente y del problema, como por ejemplo ILLIAC IV de la Universidad de Illinois y Burroughs Corp., NORC (Naval Ordenance Research Calculator) de IBM, CM-1 de Thinking Machines y Cray-3 de Cray Inc. El objetivo es alcanzar el límite de velocidad y desempeño de las tecnologías disponibles al momento de su diseño. Hay que tener en cuenta que la arquitectura especial de una supercomputadora requiere otros componentes especiales también, como sistemas de refrigeración, alimentación, software y servicio técnico, y que el mercado para este tipo de dispositivos es relativamente pequeño, todo lo cual impacta en su costo de adquisición y operación. Adicionalmente estas condiciones también aceleran la obsolescencia de una supercomputadora debido a que el hardware a medida puede limitar la aplicación de nuevas técnicas en ella y el software a medida dificulta su empleo para resolver otros problemas, con lo cual una vez que se deja de usarla para su propósito inicial, incluso cuando el equipo siga siendo una máquina de cálculo excelente y competitiva para muchas aplicaciones, lo más rentable puede ser simplemente dejar de usarla. Esto es así al punto que, al momento de la redacción, el fabricante de supercomputadoras Cray Inc. ofrece a sus clientes un programa para hacerse cargo de los equipos para su reciclaje (http://www.cray.com/support/product-recycling).

Con el correr del tiempo fue surgiendo la alternativa de clustering HPC, tal vez oficializada en 1994 con el cluster del proyecto Beowulf de la NASA. Sin buscar el rigor de una definición, podríamos decir de la computación de alto desempeño o HPC que es una práctica que busca desarrollar capacidad de procesamiento agregando las capacidades de varias máquinas (llamadas nodos) disponibles comercialmente (off-the-shelf) por medio de la técnica de clustering de procesamiento distribuido (ver la sección de Clusters HPC, más adelante) de forma tal que su desempeño sea muy superior al de una estación de trabajo típica. Surge de la oportunidad brindada por los avances en la capacidad de procesamiento y E/S (entrada/salida) de las estaciones de trabajo, por los avances en las prestaciones de las redes de comunicaciones y por la disminución en los precios en estos rubros, debido a la escala de su producción y de la necesidad de contar con alto poder de cálculo pero bajando los costos asociados a una supercomputadora mencionados previamente. El hecho de utilizar arquitecturas ampliamente disponibles y conocidas, infraestructura y tecnologías de interconexión bien establecidas y software de propósito general (usualmente basado en Linux, pero también hay versiones de Microsoft Windows, por ejemplo) hace esto último posible. Las principales desventajas del enfoque son alta latencia y bajo ancho de banda en la comunicación entre los nodos relativamente a lo que se encuentra en las supercomputadoras coetáneas.

La dificultad para establecer la relación entre HPC y supercomputación posiblemente deviene de que, con el tiempo, las ventajas de la técnica de clustering la hicieron suficientemente atractiva para los fabricantes de supercomputadoras como para ser adoptada por ellos. En la actualidad las supercomputadoras también son clusters de procesamiento distribuido, es decir, buscan agregar potencia de cálculo de equipos individuales a través de una red de comunicaciones. De esta forma la supercomputación obtiene una modularidad y escalabilidad impensable anteriormente que amplía sus usuarios potenciales. En reciprocidad algunas tecnologías propietarias y de alto desempeño utilizadas en supercomputación, por ejemplo en interconexión

(Myrinet, QsNet, InfiniBand), se tornaron más disponibles y fueron adoptadas en diversos grados en los clusters HPC. Como ejemplos, muy conocidos, de las supercomputadoras más grandes actuales podemos citar Sequoia, MareNostrum y RoadRunner de IBM, XT5 y CS-Storm de Cray Inc. y Tianhe-2 de la Universidad Nacional de Tecnología de Defensa de China.

Haciendo un somero análisis comparativo de estas dos alternativas podemos decir que la supercomputación puede ser la solución más eficiente desde el punto de vista energético y temporal, pero es muy costosa. En cambio HPC busca ser más accesible desde el punto de vista económico, aunque su característica genérica introducida por los componentes que utiliza puede hacerlo menos eficiente en términos de consumo de energía y throughput.

CLUSTERS HPC

Los clusters son grupos de computadoras (llamadas nodos) interconectadas y configuradas con un propósito específico que típicamente es alto desempeño (también conocidos como HPC, para aplicaciones de cálculo intensivo) o bien alta disponibilidad (o HA, para gran tolerancia a fallas). Los clusters HPC pretenden acelerar un cálculo complejo realizando una partición de éste de modo que cada nodo pueda ejecutar una parte. Así un cluster de 10 nodos podría idealmente resolver un sistema de ecuaciones en un décimo del tiempo que tardaría un único nodo.

En la realidad hay varias restricciones que impiden alcanzar ese ideal. Una de ellas es que no todos los cálculos pueden ser resueltos en paralelo. Cada problema a resolver, cada cálculo a realizar, cada programa a ejecutar tiene distinto grado de paralelismo. Esto hace que cada caso tenga un distinto factor de aceleración al ser procesado en un cluster. Otra es la red de interconexión y sus características como ancho de banda y latencia. También están el tipo de procesador utilizado en los nodos, su relación con cantidad y ancho de banda de la RAM, el sistema de E/S y el overhead causado por el sistema operativo subyacente, entre otras.

Por todo esto, para intentar comparar este tipo de sistemas se utilizan 2 métricas. Ambas son expresadas en operaciones de punto flotante (típicamente de precisión simple) por segundo (FLOPS). La primera es el desempeño máximo teórico (Rpeak) que se calcula como la suma de las capacidades máximas teóricas de cada procesador. La segunda es un valor arrojado por un benchmark que se convirtió en un estándar de facto para este tipo de medición llamado Linpack[1].

De hecho hay un sitio web llamado Top 500 (http://www.top500.org) que periódicamente actualiza una lista de los 500 sistemas de computadoras más poderosos del planeta que utiliza el resultado de Linpack como único parámetro de comparación[2].

Los clusters HPC de tipo Beowulf (Swendson, 2005; Brown, 2003), nombrados como el héroe de un poema que cuenta con la fortaleza de muchos guerreros y su misión es destruir a cierto monstruo[3], se caracterizan por estar integrados por componentes de uso masivo (commodity components). Los componentes incluyen las computadoras, que son típicamente PCs pero pueden ser dispositivos de propósito más específico como celulares o consolas de video juegos (Taha et al., 2010); el hardware de red, típicamente ethernet; el software de base, que tiene como principal participante a sistemas operativos GNU/Linux (Ferreira et al., 2001; Adams y Vos, 2002; Ridge et al., 1997) y middleware que se encarga de tareas administrativas. Es en la utilización de este tipo de hardware y

[1] http://www.netlib.org/linpack/ [Online; Mayo de 2015].
[2] The Linpack Benchmark — TOP500 Supercomputing Sites. http://www.top500.org/project/linpack/ [Online; Mayo de 2015].
[3] WordSpy: Beowulf cluster. http://wordspy.com/words/Beowulf-cluster.asp [Online; Mayo de 2015].

software precisamente donde se inclina la balanza de la relación precio/performance hacia los clusters Beowulf (Ridge et al., 1997) en comparación con las soluciones alternativas.

SOLUCIONES PARA CLUSTERING EN GENERAL

Para software de clustering se evaluarán únicamente aquellos sistemas que sean gratuitos, de código abierto y basados en Linux por su amplio soporte de hardware, por la vasta disponibilidad de documentación referida al tema de clustering HPC, por permitir modificación, depuración y subproyectos y por tener el menor impacto en el presupuesto de la institución. Se consideraron las implementaciones SSI debido principalmente al interés en contar con balance de carga, migración y checkpointing (la capacidad de tomar una imagen instantánea de un proceso para luego reanudar su ejecución desde ese punto) de procesos. Entre ellas se destacaron OpenMosix[4], OpenSSI[5] y Kerrighed[6].

Renaud Lottiaux y otros comparan en un trabajo (Lottiaux et al., 2004) los 3 sistemas operativos mencionados. En la actualidad, sin embargo, el proyecto OpenMosix está cerrado desde Marzo de 2008 y su continuación, LinuxPMI, tenía problemas de estabilidad y ya no se encuentra en la web[7]. El proyecto OpenSSI en su versión de desarrollo más reciente al momento de la redacción (1.9.6, lanzada en Febrero de 2010) utiliza un kernel demasiado viejo, 2.6.12, presentando problemas de drivers para el hardware. Su versión estable (1.2.2) fue lanzada en Marzo de 2005. Para este trabajo Kerrighed parece ser el mejor candidato porque su última versión al momento de esta redacción (3.0.0 lanzada en Junio de 2010) está basada en el kernel de Linux 2.6.30 y soporta arquitecturas de 64 bits y es, por esto, el que más probabilidades tiene de soportar el hardware de las máquinas del laboratorio local empleado para las pruebas.

ENFOQUES PARA REUTILIZAR RECURSOS

Debido a que este trabajo propone la confección de un cluster HPC utilizando recursos existentes y habiendo evaluado el software específico de clustering, cabe evaluar el estado del arte en reutilización de recursos de computación. Entre las técnicas existentes destacaremos algunas a continuación.

El modelo de computación voluntaria, utilizado por ejemplo en el proyecto SETI@home[8], basado en BOINC[9], voluntarios anónimos instalan un cliente en su computadora o dispositivo móvil que se conecta mediante Internet a un servidor que maneja la paralelización de las tareas donando así poder de procesamiento. La ventaja principal es que se cuenta con un potencial muy grande de usuarios, por ende de capacidad de cálculo, a un costo relativamente bajo. Las principales desventajas son la dependencia de los usuarios que deben ser convencidos para instalar el cliente, hay que confiar en su "buen comportamiento", esto es, que no interfieran con el proceso devolviendo resultados incorrectos (voluntaria o involuntariamente), el impacto en la performance o el consumo de energía de sus estaciones de trabajo asociado, el hecho de que se donan sólo los ciclos de CPU inactivos, que su equipo debe estar encendido y la dependencia de internet. Todo estos factores hacen además difícil de predecir el rendimiento del sistema. Como desventaja adicional hay que mencionar que no es apropiado para trabajos que presentan una baja relación procesamiento/comunicación (Vlădoiu y Constantinescu, 2009).

[4] http://www.openmosix.org/. [Online; Mayo de 2015].
[5] http://openssi.org/ [Online; Mayo de 2015].
[6] http://www.kerrighed.org/ [Online; Mayo de 2015].
[7] http://linuxpmi.org/trac/wiki/Status/ [Online; Agosto de 2014].
[8] http://setiathome.berkeley.edu [Online; Mayo de 2015].
[9] Berkeley Open Infrastructure for Network Computing (BOINC). http://boinc.berkeley.edu [Online; Mayo de 2015].

El modelo de Desktop Grid o Network of Workstations (Adams y Vos, 2002) es similar a computación voluntaria, sólo que se utiliza en intranets, por ejemplo en una empresa o institución, haciendo que los usuarios no sean anónimos y posibilitando más control (Vlădoiu y Constantinescu, 2009) y mejor predicción de disponibilidad de recursos (Salinas et al., 2011). Su funcionamiento en intranets complejas es muy dependiente de la configuración y administración de la red.

En el cluster de estaciones de trabajo virtualizadas un software de virtualización divide los recursos de hardware a bajo nivel asignando una cierta cantidad de CPUs, memoria, interfaces de red y otros a una máquina virtual que será nodo de un cluster, por ejemplo la utilizada en Parker Aerospace (Engineering, 2011). Es más barato agregar hardware en las PCs existentes y evita compartir recursos con el usuario de la PC, pero requiere BIOS (Basic Input/Output System) y hardware especial (chipset que soporte Intel(R) VT-d, por ejemplo) y posiblemente software propietario.

Las imágenes LiveCD de clusters HPC proveen en general un cluster basado en MPI (Message Passing Interface). En algunas como PelicanHPC[10], el "nodo maestro" arranca desde el CD y el resto por red (PXE - Preboot eXecution Environment), mientras que en otras como Cluster by Night[11], todos los nodos deben arrancar desde el CD.

Los sistemas de despliegue, provisionamiento y administración de clusters se encargan de todos los pasos mencionados, aunque la automatización requiere trabajo adicional. Si bien tienen opción para usar sistema de archivos en red, no consideran otros puntos de conflicto con administradores de laboratorio, como los servidores DHCP por ejemplo. Como instancias de esta categoría podemos citar OSCAR[12], Warewulf[13], oneSIS[14] y Perceus[15]. Este último requiere adquirir licencia para activar ciertas características. Serán evaluados en el futuro para considerar una posible integración con el presente trabajo.

ARQUITECTURA GENERAL DEL SISTEMA

A los fines de permitir la reproducción de esta experiencia facilitaremos en esta sección los elementos de hardware, software, topologías, configuraciones, criterios y demás elementos que se utilizaron en la elaboración de este trabajo y conforman el producto desarrollado.

REQUISITOS DE HARDWARE Y SOFTWARE

Es preciso contar con una computadora que tendrá a su cargo la ejecución de la máquina virtual que constituye el servidor del sistema de cluster a cargo de los servicios de DHCP, TFTP (Trivial File Transfer Protocol) y NFS (Network File System) en la red del laboratorio. Aunque puede ejecutarse con menos recursos de procesador, RAM y disco, los requisitos mínimos recomendados para dicho host son un procesador AMD Sempron 3200+ o equivalente, 512 MB RAM, 30 GB HDD e interfaz de red Ethernet 100 Mbps. Como software de virtualización se recomienda VMware Server (gratuito) que ejecuta en sistemas operativos Linux, sin embargo se observó indistinto desempeño usando VMware Player (igualmente gratuito).

Las computadoras del laboratorio, que serán los nodos del cluster, deben poseer hardware compatible con el kernel de Kerrighed 3.0 (Linux 2.6.30) o del sistema operativo de cluster que se desee ejecutar (ver sección Cualidades destacadas) y

[10] http://pareto.uab.es/mcreel/PelicanHPC/ [Online; Mayo de 2015].
[11] http://zacharski.org/cluster-by-night [Online; Mayo de 2015].
[12] http://oscar.openclustergroup.org/ [Online; Mayo de 2015].
[13] http://warewulf.lbl.gov/trac [Online; Mayo de 2015].
[14] http://onesis.org [Online; Mayo de 2015].
[15] http://perceus.org/ [Online; Mayo de 2015].

deben ser capaces de iniciar desde la red (a través de PXE) y "despertar" desde la red por medio de Wake-On-LAN, preferentemente desde el estado S5.

Todas ellas deberán contar con al menos una interfaz de red Ethernet de por lo menos 100 Mbps de capacidad.

Relativo a la conectividad de red es necesario que exista una infraestructura de conexión Switched Fast Ethernet como mínimo entre el servidor y todos los nodos.

CONFIGURACIÓN DEL LABORATORIO

Todos los nodos deben tener su BIOS protegido con contraseña para dificultar modificaciones, Wake-On-LAN activado y la placa de red como principal dispositivo de arranque. Al interrumpirse el suministro de energía eléctrica la máquina debe recordar el estado o encenderse (ver sección Operación).

En el ámbito del laboratorio no debe haber otro servidor ofreciendo arranque por red (PXE). De otro modo será necesario ajustar la configuración de los servidores, por ejemplo, apagando y encendiendo los servicios en los horarios necesarios.

Cada uno de los ajustes mencionados en esta sección debe ser coordinado con los administradores del laboratorio y es precisamente su baja cantidad lo que determina el alto grado de independencia de la administración que esta solución presenta.

SERVICIOS NECESARIOS EN LA RED LOCAL

Para iniciar el cluster los nodos deben usar como dispositivo de arranque la placa de red. La imagen de arranque es provista por un servidor TFTP. Dicha imagen de arranque será un archivo que indica arrancar desde el disco rígido local en horarios de uso "normal" (llamemos así al horario en que el laboratorio es tradicionalmente utilizado para su propósito original) y en los horarios de inactividad será la imagen de arranque del cluster.

Un servidor DHCP debe ser configurado para indicar la ubicación de la imagen de arranque a los nodos. Si es el único servidor DHCP sirviendo pedidos de arranque vía PXE no es necesario que sea autoritativo y puede coexistir con otros servidores DHCP en la misma red sin interferir. Esto es debido a que el servidor DHCP del cluster sólo es necesario para el arranque por medio de PXE dado que los nodos pueden recibir IPs estáticos a través de la línea de comando del kernel pasada por PXELINUX al momento del arranque.

Dado que las máquinas deben tener configurada la red como primera opción de dispositivo de arranque, es conveniente que estos dos servicios (TFTP y DHCP) estén permanentemente activos para evitar la espera del timeout del intento de arranque a través de PXE por parte del nodo.

El sistema de archivos del cluster también está en la red, en un servidor NFS, el cual, por razones de seguridad, conviene que no esté activo hasta el momento de operar el cluster. En el diseño de esta solución se decidió emplear arranque desde la red y sistema de archivos de red para no utilizar el disco rígido de las máquinas (nodos) y de este modo lograr que el cluster represente la menor invasión posible en el ambiente del laboratorio y la mayor independencia posible de sus administradores. Esta técnica además evita fallas por problemas con los discos y permite apagarlos por software para ahorrar energía. Como efectos adversos hay consideraciones de seguridad (NFS no es un servicio seguro por sí mismo), la red ya no es utilizada exclusivamente para conexión entre nodos sino también para las operaciones del sistema de archivos y por lo tanto sus desempeños (tanto el de la comunicación entre nodos como el del sistema de archivos) pueden verse afectados, en una medida dependiente del grado de utilización de cada uno de estos recursos por parte de la aplicación; por otro lado

la red incrementa la latencia y puede resultar en cuello de botella para las operaciones del sistema de archivos; finalmente el servidor NFS también puede ser cuello de botella. No hubo posibilidad de evaluar hasta el momento la magnitud de este efecto comparando el sistema propuesto con uno que utilice el disco rígido local de cada nodo pero las pruebas realizadas hasta el momento muestran un rendimiento aceptable, considerando las circunstancias y limitaciones de este enfoque. En caso de ser necesario un mayor rendimiento de sistema de archivos, por ejemplo, debería realizarse un acuerdo con el/los administradores del laboratorio para ocupar una porción del disco rígido local de cada computadora o utilizar redes separadas para el sistema de archivos y la comunicación entre nodos.

Todos estos servicios están configurados y en ejecución en la máquina virtual (mencionada en la sección Requisitos de hardware y software), que puede ser vista como un "nodo maestro" (head node) del cluster, y será llamada así, aunque estrictamente no lo sea.

La estructura descrita hasta acá puede ser visualizada gráficamente en el esquema de la Figura 1.

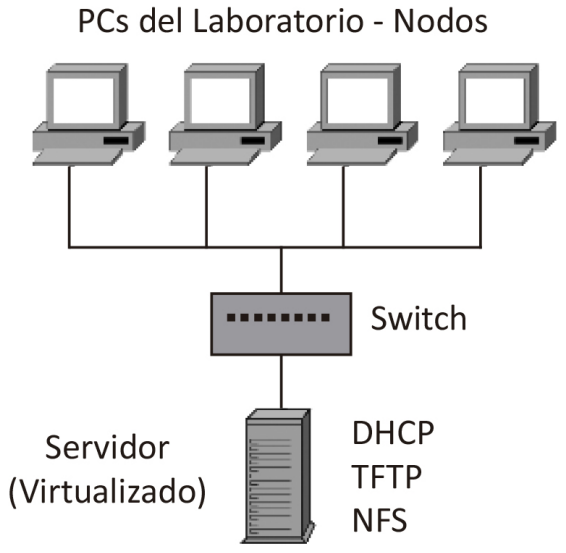

Figura 1 - Estructura del cluster en un laboratorio de computadoras.

OPERACIÓN DEL CLUSTER

Al finalizar el horario de uso normal del laboratorio las máquinas deben ser apagadas. Luego de cambiar la imagen de arranque del servidor TFTP a la del cluster deben ser encendidas nuevamente.

De esta forma el sistema operativo del cluster cargará en todas ellas.

Todo este proceso puede estar automatizado por scripts para evitar la necesidad de la intervención de un ser humano físicamente presente en el laboratorio.

El cluster debe ser configurado para apagarse momentos antes de finalizado el horario designado (idealmente realizando checkpoint de los procesos en ejecución) y la imagen de arranque del servidor TFTP debe ser cambiada a la de operación normal al mismo tiempo.

Los cortes de suministro de energía eléctrica pueden presentar un desafío particular: los nodos pueden no ser capaces de procesar la orden de arranque por red (Wake-On-LAN) en el estado en que se encuentran al momento de regresar la alimentación (S5). En este caso habrá que considerar configurar el BIOS para que recuerde el estado de la máquina y así encienda el equipo nuevamente cuando corresponda.

ACCESO AL CLUSTER

Para acceso de administración y mantenimiento puede ser necesario configurar rutas en la red para lograr el acceso desde fuera del laboratorio (ya sea dentro o fuera de la red local, dependiendo de las necesidades y posibilidades) a la máquina que ejecuta los servicios necesarios y al cluster.

Para el acceso de los usuarios del cluster se propone, considerando seguridad y usabilidad, la implementación de un sitio web con mecanismos de autenticación y autorización apropiados que permita "subir" los trabajos a ejecutar en el cluster y "bajar" los resultados una vez disponibles.

Asimismo es recomendable que este sitio web se encuentre alojado en una máquina distinta al nodo maestro y, de ser posible, ni siquiera conectada directamente con la red del cluster. De este modo este servidor bien podría estar en una DMZ o contratado en algún proveedor de hosting y sería el intermediario entre los usuarios y el cluster evitando toda interacción directa entre ellos (ver Figura 2).

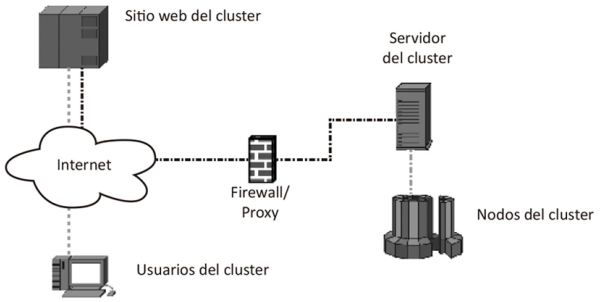

Figura 2 - Acceso al cluster desde el exterior para usuarios a través de Internet.

PRESENTACIÓN DE TRABAJOS

Los trabajos a ejecutar (programas y datos) deberán ser subidos en una forma normalizada y claramente documentada, por ejemplo empacada y comprimida en un único archivo de formato tar.bz2 o tar.gz, y con un script o archivo ejecutable principal con un nombre definido específicamente.

De igual modo debe establecerse el mecanismo para devolver los resultados. Una alternativa es establecer que el sistema comprima y suba al servidor web todo el contenido de un directorio de nombre convenido.

Una imagen de máquina virtual del sistema será provista para que los usuarios puedan realizar pruebas de ejecución y depuración antes de enviar el trabajo al cluster.

EJECUCIÓN DE TRABAJOS Y ENTREGA DE RESULTADOS

Los trabajos serán planificados con un orden según criterios a determinar por los administradores y llamados a ejecución por el host que provee de usuario correspondiente. Un email será enviado al propietario del trabajo para informar del evento.

Como último paso, por seguridad, puede disponerse el reinicio del cluster desde una imagen de respaldo de su sistema de archivos para evitar que la ejecución de un trabajo de usuario afecte la de los siguientes.

CUALIDADES DESTACADAS

A continuación se enuncian las ventajas que ofrece el enfoque elegido en este trabajo.

FLEXIBILIDAD

El método propuesto es flexible en el sentido en que permite la ejecución de diversos softwares de base y de clustering. En este caso fue evaluado conveniente utilizar Kerrighed 3.0 pero puede también servir para versiones anteriores de Kerrighed y podría servir para cualquier otro sistema de clustering existente o futuro siempre que sea compatible con el hardware de las máquinas.

Incluso las restricciones de sistema de archivos y arranque en red podrían ser superadas con el apoyo de los administradores de los laboratorios. Esto podría habilitar para ejecutar sistemas que deben estar instalados localmente y no soportan arranque por red usando (paradójicamente) una imagen de arranque por red que indique la partición del disco a iniciar.

Desde otro aspecto de la flexibilidad se destaca que los usuarios del cluster pueden disponer scripts para instalar el software que sea necesario para ejecutar sus programas, subir el software en formato binario al servidor web o incluso puede evaluarse la implementación de un mecanismo para que usuarios avanzados puedan subir el sistema de archivos completo a ser usado por el cluster.

ESCALABILIDAD

El sistema es adaptable a las capacidades de los laboratorios, ya sea que se agreguen o eliminen nodos o bien que se cambie su configuración interna (siempre y cuando el hardware sea soportado por el kernel).

PORTABILIDAD

Tener el nodo maestro o servidor del cluster en una máquina virtual ``empaqueta'' la solución facilita su distribución y trae consigo la ventaja de poder ser ejecutado en una gran variedad de entornos.

SENCILLEZ

Bajo ciertas condiciones poner en marcha un sistema como el propuesto puede ser tan simple como encender una máquina virtual. Otras tareas administrativas tales como respaldo y restauración también se benefician de la virtualización.

Para el usuario, utilizar el cluster es ``subir'' su trabajo a un sitio web y esperar el aviso de resultados via email junto con instrucciones para descargarlos. Se dispone además un mecanismo muy cómodo y conveniente para pruebas y depuración (ver sección Presentación de trabajos) donde nuevamente la virtualización hace su aporte.

SEGURIDAD

Si bien se hicieron consideraciones de seguridad en el diseño de la solución, por ejemplo en las secciones 3.5 y 3.7, un análisis en detalle queda como trabajo futuro para diseñar e implementar mecanismos que eviten usos inadecuados tales como que el cluster provea acceso desde el exterior por medio de túneles o que sea utilizado para enviar spam, entre otros.

PLATAFORMA DE EVALUACIÓN

A modo de ejemplo y de guía se presenta a continuación el listado de las características más relevantes del cluster HPC formado con las máquinas de uno de los laboratorios de la Facultad Regional Tucumán (Universidad Tecnológica Nacional).

Sin embargo debe quedar claro que el objetivo del trabajo no es detallar instrucciones sobre la configuración ni operación de un cluster HPC sino de proponer un método razonable y conveniente para convertir un laboratorio de computadoras en un cluster HPC durante las horas libres de uso.

El cluster posee una arquitectura híbrida (cluster de SMPs[16]) de nodos homogéneos y pudo ser iniciado con un máximo de 20 nodos debido a que el laboratorio donde las pruebas para este trabajo fueron desarrolladas posee esa cantidad de computadoras.

Cada nodo está equipado con un procesador Intel Core 2 Duo E7400 con 2 núcleos de 2,8 GHz

(arq. 64 bits) y 2 GB de RAM en configuración dual-channel. El chipset del motherboard es Intel G31. La interconexión es en topología estrella a través de una red Switched Fast Ethernet provista por un switch 3Com de 24 bocas de la línea Baseline Switch 2024 (3C16471B). La placa madre posee una interfaz de red PCI Express Gigabit Ethernet RTL8111/8168B y BIOS que permite arranque desde la red (via PXE) y Wake-On-LAN.

El servidor se encuentra en una máquina virtual ejecutada por VMware Server 2.0.1 con recursos de 256 MB de RAM, 10 GB de disco rígido y una interfaz Fast Ethernet. La máquina física es una estación de trabajo Sun Microsystems Ultra 20 que cuenta con un procesador Dual-Core AMD Opteron(tm) Processor 1210 de 1,8 GHz con 2 GB de RAM, 600 GB de disco rígido y dos interfaces de red Gigabit Ethernet (aunque sólo una está conectada).

[16]Symmetric MultiProcessing.

RESULTADOS

Se desarrolló una imagen de máquina virtual que permite iniciar el cluster presentado como ejemplo en un laboratorio de computadoras que cumpla con los requisitos especificados sin interferir con las funciones normales del laboratorio en horarios tradicionales.

El arranque y apagado del cluster pudo ser automatizado exitosamente usando los enfoques descritos en la sección Arquitectura general del sistema propuesto.

El sistema operativo elegido, Kerrighed, fue capaz de operar con el hardware del laboratorio y el sistema de archivos en red presentando un desempeño aceptable, indistinguible desde el punto de vista de experiencia de usuario de un Debian 6.0 (también conocido como Squeeze) sin terminal gráfica instalado en disco local. Su estabilidad se vio comprometida al intentar utilizar memoria remota y con la desconexión de uno o más nodos del cluster. La migración de procesos funciona correctamente y el process checkpointing no pudo ser evaluado en profundidad hasta el momento.

Para hacer pruebas de rendimiento del cluster sobre MPI (OpenMPI versión 1.5) se utilizó LINPACK, el benchmark con el cual se miden los desempeños de las supercomputadoras para clasificar en top500.org. Los parámetros de esta prueba todavía no han sido ajustados para obtener máximo rendimiento. El mejor valor obtenido hasta el momento es de 21,29 Gigaflops2[17] en un cluster de 8 PCs (16 núcleos en total) que tomó 2003,7 segundos en ejecutar y que en relación con el resultado obtenido para una prueba similar en sólo 1 nodo de 2 núcleos (5,09 Gflops) nos da un rendimiento del 52,3 %.

El grado de independencia de la administración del laboratorio fue el esperado. El funcionamiento del cluster no fue afectado por rupturas de discos rígidos ni instalaciones de diversos sistemas operativos en ellos.

Un esquema de pruebas más intensivo no ha sido posible por no contar hasta el momento con la autorización pertinente para reconfigurar el BIOS de las máquinas del laboratorio según lo pautado en la sección Configuración del laboratorio.

CONCLUSIONES

Las pruebas preliminares realizadas hasta el momento demuestran que es posible la reutilización de recursos de un laboratorio de computadoras para la implementación de un cluster HPC. El rendimiento cercano al 50% obtenido en el LINPACK no debe interpretarse desde un punto de vista cuantitativo debido a que es susceptible de ser mejorado y, por otro lado, es esencialmente el resultado de un benchmark y el desempeño del cluster será distinto según la aplicación a ejecutar en cada caso. Desde el punto de vista cualitativo, en cambio, nos dice que el arreglo de computadoras está en condiciones de resolver problemas usando MPI, lo que se traduce en utilidad para la investigación y la docencia, para hacer cálculos de tal complejidad que excedan las capacidades de los equipos singulares disponibles para hacer pruebas preliminares, ensayar soluciones novedosas, justificar y dimensionar solicitudes de mayor poder de procesamiento (clusters dedicados), capacitar personal de administración y mantenimiento, desarrollar software y soluciones para HPC, entre otros usos posibles.

Es importante notar que la infraestructura de hardware empleada en este desarrollo es básica y elemental a cualquier laboratorio de computadoras de la actualidad y que sus componentes de software o bien son “libres” o existen alternativas libres (en el caso de VMware, por ejemplo). Esto significa que la inversión ya se encuentra realizada donde exista un

[17] Flops: floating point operations per second - operaciones de punto flotante por segundo.

laboratorio por lo cual podríamos decir que el rendimiento costo-beneficio es óptimo, incluso cuando las prestaciones del sistema no lo sean. Hay que tener en cuenta que el eje del presente trabajo no es la optimización de rendimientos sino la reutilización de recursos existentes.

La continuación de este trabajo de aquí en adelante se enfocará en evaluar el aporte a los objetivos que pueden realizar los sistemas mencionados en la sección Enfoques para reutilizar recursos y buscar otros sistemas de administración para clusters, planificadores de procesos y gestores de colas que sean factibles de ser integrados en el esquema de funcionamiento propuesto. Habrá que considerar el desarrollo propio de aplicaciones de ser necesario.

Por otro lado se buscará un modo conveniente de integrar diversos sistemas como el propuesto (ejecutando en distintos laboratorios) junto con esquemas de Desktop Grid y computación voluntaria (también mencionados en la sección Enfoques para reutilizar recursos) para configurar con ellos un grid de alcance inicialmente local a la facultad.

Finalmente habrá que estudiar en mayor profundidad los aspectos de seguridad del cluster teniendo en cuenta que puede ser tanto objetivo como herramienta de ataques dado que es, a fin de cuentas, un sistema destinado a ejecutar código ajeno.

Por tratarse el presente de un trabajo en progreso, aún restan desafíos y problemas por resolver; no obstante los avances parciales aquí expuestos significan un paso en la dirección correcta hacia el objetivo principal, que es reducir el impacto de la escasez de recursos en las instituciones educativas por medio de su mejor utilización para ayudar a investigadores, docentes y alumnos a permanecer en la carrera del conocimiento.

AGRADECIMIENTOS

Departamento Sistemas - UTN - FRT: P. Nazar, A. J. Nasrallah; TICs: J. Arias, G. Correa; Cátedra de Sistemas Operativos - UTN - FRT: L. R. de la Zerda, E. Loandos; Otros: F. Villacis Postigo, A. Will.

REFERENCIAS

Engineering and Physical Sciences Research Council, "High End Computing Terascale Resources (HEC-ToR) Scientific Case", Engineering and Physical Sciences Research Council, (2004).

Bose, Crosswell, Hamilton and Mesa, "Piloting sustainable hpc for research at columbia", NSF Workshop on High Performance Computing Center Sustainability, (2010).

Gropp, Lusk and Sterling, "Beowulf Cluster Computing with Linux". The MIT Press, second edition, (2003).

Ridge, Becker, Merkey and Sterling, "Beowulf: Harnessing the power of parallelism in a pile-of-pcs". In Proceedings, IEEE Aerospace, pages 79–91, (1997).

Adams and Vos, "Small-college supercomputing: building a beowulf cluster at a comprehensive college". In Proceedings of the 33rd SIGCSE technical symposium on Computer science education, SIGCSE '02, pages 411–415, New York, NY, USA, (2002). ACM.

Slater, "Give Me Your Clustered Masses" . Revista CIO (www.cio.com); 10 (12), 92, (1997).

Swendson, "The Beowulf Howto". The Linux Documentation Project. http://tldp.org/HOWTO/html_single/Beowulf-HOWTO, (2005) [Online; Mayo de 2015].

Brown. "What's a Beowulf?". Engineering a Beowulf-style Compute Cluster. Physics Department. Duke University. http://www.phy.duke.edu/~rgb/brahma//beowulf_book/node9.html, (2003). [Online; Mayo de 2015].

Taha, Yalamanchili, Bhuiyan, Jalasutram, Chen

and Linderman, "Neuromorphic algorithms on clusters of playstation 3s". WCCI 2010 IEEE World Congress on Computational Intelligence, pages 3040–3049, (2010).

Ferreira, Kettmann, Thomasch, Silcocks, Chen, Daunois, Ihamo, Harada, Hill, Bernocchi and Ford. "Linux HPC Cluster Installation". IBM Redbooks, first edition (ISBN: 9780738422787), (2001).

Lottiaux, Boissinot, Gallard, Vallee and Morin, OpenMosix, "OpenSSI and Kerrighed: A Comparative Study. Research Report". RR-5399, INRIA, (2004).

Vlădoiu, Constantinescu, "Availability of computational resources for desktop grid computing". BULETINUL UniversităŃii Petrol - Gaze din Ploieşti, volume LXI, pages 71–76, (2009).

Salinas, Garino and Zunino, "Sistema de predicción y evaluación de disponibilidad operativa de recursos en desktop grids". 40JAIIO - HPC 2011, pages 105–116, (2011).

Engineering, editor, "Speed Product Development via Virtual Workstation Clustering". Hewlett Packard Development Company, (2011).

Engineering and Physical Sciences Research Council, "International Review of Research Using HPC in the UK", Engineering and Physical Sciences Research Council, (ISBN 1-904425-54-2), (2005).

Fecha de Presentación: 15/05/15
Fecha de Aprobación: 08/06/15