Plataformas Actuales para Computación de Alto Rendimiento
Current Platforms for High Performance Computing

Gilberto Díaz
Centro de Cálculo Científico de la Universidad de Los Andes (CeCalCULA), Venezuela
gilberto@ula.ve
y
Departamento de Computación, Escuela de Sistemas, Facultad de Ingeniería
Universidad de Los Andes, Mérida 510,, Venezuela

Resumen

En 1976 Seymour Cray y los ingenieros de CDC sacaron crearon la Cray 1 con 100 MFLOPs. Desde ese momento se acuñó el término Supercomputación. En la década de los años 90 los microprocesadores comienzan a exhibir los adelantos tecnológicos incorporados en los supercomputadores. También aparecen avances en redes, sistemas de almacenamiento y software. Así, la distancia entre supercomputadores y sistemas tradicionales comienza a desaparecer y el término Computación de Alto Rendimiento comienza a ser utilizado. Asi mismo, Los Clusters de PCs se hicieron populares en los 90 con el proyecto Beowulf. *La idea consiste en armar un máquina, con gran poder de cómputo, interconectando PCs y usando software libre como Linux, MPI o PVM, a un costo considerablemente menor que el de supercomputadores comparables. Hoy en día esta tecnología lidera la computación de alto rendimiento.*

In 1976 Seymour Cray and the engineers from CDC build the Cray 1 with 100 MFLOPs. Since that moment the term "Supercomputing" appears in the computer's world. In the 90s many of the technological advance in supercomputers were incorporated into the microprocessors. Computer Networks, storage systems and software had significant advances as well. Thus, the performance gap disappears and, the term "High Performance Computing (HPC)" started to be used. In the same decade, Linux Clusters became popular because the Beowulf project. The idea is build an unexpensive virtual parallel machine connecting a set of PCs and using free software: Linux, MPI, PVM, Nowadays, this technology leads the HPC world.

1. Introducción

En 1976 Seymour Cray y los ingenieros de Control Data Corporation (CDC) diseñaron la Cray 1 con una capacidad de 100 MFLOPs. Desde ese momento se acuñó el término Supercomputación. De acuerdo a A.J Forty [1] el término Supercomputación está dirigido a aquellas máquinas que alcanzan un rendimiento significativamente superior al esperado por las tecnologías de sus días. La supercomputación se popularizó en los años 80 con el crecimiento de Cray y CDC. En la década de los 90 se incorporaron muchos avances tecnológicos a los microprocesadores y la distancia entre los supercomputadores y los sistemas tradicionales comenzó a desaparecer. Entre estos avances encontramos: memoria *bit-paralela*, aritmética *bit-paralela*, memoria *cache*, búsqueda adelantada de datos en instrucciones (*lookahead*) y encauzamiento de instrucciones y datos (*pipelining*) También en los 90 los PCs comienzan a exhibir la capacidad de las estaciones de trabajo, sus precios se hacen muy asequibles y los costos de los equipos de redes disminuyen significativamente. Así mismo, las prestaciones en las tecnologías de redes y sus bajos costos propician su expansión. Por otro lado, el surgimiento de *LINUX*, un sistema operativo completamente abierto originalmente desarrollado por el Finlandés *Linus Torvalds* y luego mediante la colaboración de un sin número de voluntarios alrededor del mundo, compatible con *UNIX* y capaz de correr sobre PCs, permite finalmente satisfacer las demandas de computación a una fracción del costo asociado a los supercomputadores. En 1994 *Donald Becker, Thomas Stirling* y su equipo en la *NASA*, utilizaron una técnica para agrupar un conjunto de PCs, logrando obtener una eficiencia comparable a los supercomputadores. A este sistema lo denominaron *Beowulf* y se convirtió en el modelo de los *clusters* de PCs para la computación de alto rendimiento.

2. Arquitecturas

Las computación desde sus inicios se convirtió en una alternativa indispensable para los métodos tradicionales de la investigación científica[2]: análisis teórico y experimentos de laboratorio. El enfoque computacional frecuentemente ofrece soluciones a problemas que por su complejidad son prácticamente imposibles de manejar desde el punto de vista teórico. Así mismo, el computador puede proporcionar información que no está disponible en el laboratorio y también puede proveer nuevas formas de análisis a los problemas estudiados. El grupo de profesionales que mantiene el sitio *top500*[1] ofrece información sobre las quinientas (500) máquinas más rápidas del planeta utilizadas en investigaciónes de diferente índole. Ellos utilizan *linkpack*[2] un benchmark que consiste en un sistema denso de ecuaciones lineales. Las estadísticas de Junio del 2007 del *top500* muestran las diferentes arquitecturas más utilizadas actualmente para la Computación de Alto Rendimiento (gráfico de la figura 1)

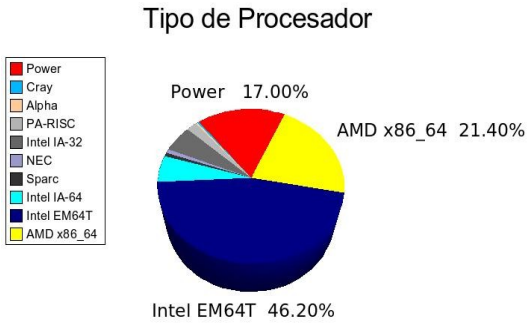

figura 1: Procesadores del top500.org

Las tecnologías Intel EM64T y AMD x86_64 proporcionan registros de 64 bits, uso de apuntadores y tipo de datos enteros de 64 bits, hasta 1 TB de espacio de memoria en la plataforma y espacio de direccionamiento virtual de 64 bits.. En Enero del 2007 Intel introdujo el transistor de 45nm lo cual permite tener hasta 400 millones de transistores en los procesadores Dual-Core y 800 millones de transistores en los QuadCore. Los actuales procesadores Intel de 3GHz tienen capacidad de hasta 12 GFLOPs y en los QuadCore podemos tener hasta 48 GFLOPs. Capacidades similares se pueden obtener con procesadores AMD. Otras arquitecturas como POWER y Cray ofrecen plataformas ideales para la computación de alto rendimiento las cuales lideraron en la década de los 90. Sin embargo, como se observa en el gráfico anterior no son ampliamente utilizadas.

[1] *http://www.top500.org*
[2] *http://www.netlib.org*

Otro factor interesante que interviene en la computación de alto rendimiento es el sistema operativo utilizado por estas aquitecturas. El gráfico de la muestra los distintos sabores de sistemas operativos de las quinientas máquinas más rápidas del planeta para Junio del 2007.

3. ¿Por qué Clusters de PCs?

En los 90 los PCs comienzan a exhibir la capacidad de las estaciones de trabajo y sus precios se hacen muy asequibles. Así mismo, en el área de redes de computadores, el rendimiento de los equipos de comunicación y sus bajos costos permiten implantar redes en muchas instituciones, particularmente en universidades y otros institutos académicos. Por otro lado, el surgimiento de *LINUX*, un sistema operativo gratuito originalmente desarrollado por el Finlandés *Linus Torvalds* y luego mediante la colaboración de un sin número de voluntarios alrededor del mundo, compatible con *UNIX* y capaz de correr sobre PCs, permite finalmente satisfacer las demandas de computación a una fracción del costo asociado a los supercomputadores. Todo esto, ofreció una serie de condiciones favorables para que en 1994 *Donald Becker, Thomas Stirling* y su equipo en la *NASA*, utilizaran una técnica para agrupar un conjunto de PCs, logrando obtener una eficiencia comparable a los supercomputadores. A este sistema lo denominaron *Beowulf* y se convirtió en el modelo de los *clusters* de PCs para la computación de alto rendimiento.

Actualmente estas condiciones han influenciado significativamente la forma de hacer computación de alto rendimiento y los clusters de PCs dominan el campo según las últimas estadísticas presentadas en el *top500.* El gráfico de la figura 2: Arquitecturas más utilizadas según el top500 ilustra esta afirmación.

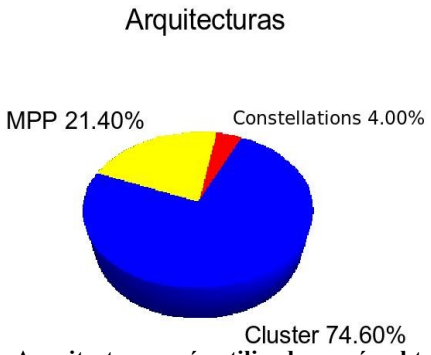

figura 2: Arquitecturas más utilizadas según el top500

Además de presentar un rendimiento comparable al de máquinas más sofisticadas a una fracción de su costo, los clusters de PCs poseen otras ventajas que los

hacen muy apropiados para la Computación de Alto Rendimiento:

- Ensamblaje: no se requiere tener un doctorado en computación y años de experiencia para ser capaz de construir un cluster. Hoy en día estudiantes de bachillerato son capaces de ensamblar PCs. Las partes se pueden comprar por separado: tarjeta madre, procesador, tarjeta de video, tarjeta de sonido, disco duro, lectrora/escritora de CDs, monitor, teclado, fuente de poder, etc., de acuerdo a los gustos y necesidades, y al acoplarlas tienen un PC hecho a la medida. Con ciertos conocimientos adicionales de redes pueden armar un cluster.
- Mantenimiento: dado que los elementos que forman un cluster se encuentran fácilmente en el mercado (son componentes de producción masiva y por lo tanto de bajo costo), al fallar alguno de ellos se puede reemplazar sin mayores inconvenientes.
- Alta disponibilidad: Si algún componente falla, solamente el nodo asociado permanece fuera de servicio. El resto de las PCs pueden atender solicitudes provenientes de los clientes. Existen mecanismos para evitar el envío de solicitudes a los nodos que tienen fallas y así evitar que un cliente piense que el servicio ya no esta disponible.
- Tolerancia a fallas: Los clusters están formados por PCs individuales interconectados por una red y en la gran mayoría de los casos no es necesario poner fuera de servicio todo el cluster para reemplazar un componente, sino solo el nodo (el PC o máquina) al que esta asociado el componente. Por lo general los servidores especializados constan de CPUs interconectados por redes especiales dentro de una misma caja. Reemplazar un componente implica apagar totalmente la máquina. Además hay que esperar que llegue el experto de la compañía con la pieza muy particular y costosa para reparar el supercomputador.
- Hospedaje: debido al alto costo de los servidores especializados, estos deben ser albergardos en centros especiales. Dentro de estos centros, ellos se encuentran en salas muy particulares con sistemas de aire acondicionado, filtrado de aire, ductos de enfriamiento, cableados y sistemas de protección eléctrica especiales, etc. Además de que deben contar con administradores, consultores, personal de mantenimiento, etc., con una preparación importante en el manejo de estas máquinas. Los clusters requieren de alojamientos mucho más modestos con el requerimiento principal de poseer un sistema eléctrico adecuado. En muchos casos ni siquiera hace falta poseer aire acondicionado.
- Modernización y expansión: por la esencia misma de lo que son los servidores especializados, cuando un centro recibe una de esta máquinas dentro de poco aparecen nuevos modelos. Actualizar los supercomputadores se traduce en comprar los nuevos modelos. Por lo particular que son los supercomputadores, expandir sus capacidades de memoria, almacenamiento en disco, numero de CPUs, etc., se traduce en inversiones sustanciales. Como los clusters están compuestos por elementos disponibles de múltiples fabricantes y debido a la compatibilidad que estos tratan de mantener con las diferentes generaciones de una misma familia de componentes, se hace sencillo modernizarlos. Actualizar el cluster con CPUs más potentes puede ser tan sencillo como sacar un CPU de la tarjeta madre e instalar otro, o quizás, reemplazar la tarjeta madre y el CPU conservando el resto de los componentes: memoria, tarjetas de video, etc. Expandir la capacidad de memoria y de almacenamiento en disco no requiere de inversiones sustanciales dado el bajo costo de estos componente. Añadir CPUs implica agregar PCs de fácil adquisición.
- Escalamiento: Si la demanda de un servicio crece, no es necesario desechar la plataforma existente para incorporar máquinas más poderosas y asi aumentar las prestaciones del servicio. Agregar más nodos al cluster es muy fácil, estos pueden tener mejores prestaciones que los nodos actuales o simplemente tener las mismas características. En cualquier caso, se incrementa la capacidad de atención de solicitudes del servicio.

Otro factor importante que influye de forma significativa es el sistema operativo. Una vez más, según el *top500* el sistema operativo más utilizado actualmente en las quinientas máquinas más rápidas del mundo el GNU/Linux. (gráfico de la figura 3: Sistemas Operativos más utilizados según top500)

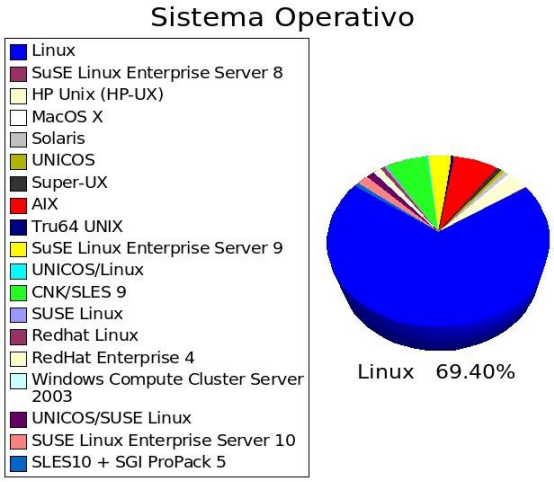

figura 3: Sistemas Operativos más utilizados según top500

Por todo esto, podemos decir que la mejor opción para la computación paralela y el cálculo intensivo son los clusters Linux.

4. Características de los Clusters Linux

En esta sección se describe brevemente los detalles de hardware y software que constituyen los elementos que integran un cluster Linux.

4.1. Hardware del Cluster

Un cluster Linux es una red de nodos, donde cada uno de ellos es un computador personal común. Por esto, los nodos constituyen el elemento principal del cluster, los cuales son responsables de todas las actividades asociadas con la ejecución de los programas de aplicación y de dar soporte al software especializado presente en los clusters. Según la función que cumplen los nodos pueden ser ubicados dentro de las siguientes categorías.

- Ejecución de instrucciones.
- Almacenamiento rápido de información temporal.
- Alta capacidad de almacenamiento de información persistente.
- Comunicación con ambientes externos incluyendo otros nodos

A la hora de diseñar un cluster se deben considerar los siguientes factores para seleccionar el hardware apropiado para los nodos de acuerso a los tipos de programas y aplicaciones que se ejecutarán:

El procesador: constituye toda la lógica requerida para la ejecución del conjunto de instrucciones, gestión de la memoria, operaciones enteras y punto flotante, y el manejo dela memoria cache.

Los nodos generalmente contienen procesadores Intel x86 o AMD. La utilización de otro tipo de procesador es permitido, sin embargo, no se consideran de uso común, ya que se elimina una de las principales características de cluster Linux (uso de componentes comunes), la cual permite reemplazar de forma fácil y con bajos costos cualquier componente del sistema.

Las máquinas con más de un procesador (Simetric MultiProcessor o SMP) son utilizadas comúnmente en clusters debido a la gran capacidad de prestaciones que proporcionan. Sin embargo, la velocidad de los buses de las tarjetas madres no tienen la capacidad necesaria para dar apoyo a arquitecturas SMP, lo que representa un cuello de botella entre los diferentes medios de almacenamiento y el procesador

Las instituciones a las que pertenecen deben estar escritas en letra tipo itálica e indicar el país. Recuerde que los nombres de las instituciones deben estar centrados respecto a cada uno de los autores. Es muy importante incluir la dirección de correo electrónico de cada uno de ellos. Después de esta información, es necesario dejar dos líneas en blanco, igualmente en tamaño de letra de 12 puntos.

La Memoria: de un computador personal es el sistema de almacenamiento más cercano al procesador. Las características deseables de la memoria son: rapidez, bajo costo y gran capacidad. Desafortunadamente, los componentes disponibles hasta ahora, solo poseen una combinación de cualquiera dos de estas características. Los sistemas de memoria modernos utilizan una jerarquía de componentes implementados con diferentes tecnologías que juntos, y en condiciones favorables, logran obtener las tres características. A pesar de todo esto, la capacidad de almacenamiento de memoria se ha incrementado considerablemente, cuadriplicándose cada tres años aproximadamente, mientras que su costo ha sufrido un constante decremento.

Las memorias constituidas por semiconductores dieron un cambio significativo a la predominancia de los medios de almacenamiento magnéticos de los años 70. Actualmente hay dos tipos de memoria de semiconductores: memoria estática de acceso aleatorio (SRAM[3]), la cual se caracteriza por ser muy rápida pero de capacidad moderada, y la memoria dinámica de acceso aleatorio (DRAM) cuya capacidad de almacenamiento es considerable pero opera de forma más lenta.

La memoria estática es implementada con celdas de bits fabricadas con circuitos flip-flop de transistores múltiples. Estos circuitos activos pueden cambiar su estado y ser accesados rápidamente, sin embargo, su consumo de energia es significativo. Este tipo de memoria es empleada en aquellas partes del sistema donde se requieren medios de almacenamiento rápidos tales como memorias cache L1 y L2.

La memoria dinámica de celdas de bits es fabricada con capacitores y transistores de puentes simples. Estos capacitores almacenan una carga en forma pasiva y las operaciones de acceso a cada celda consume esta carga, además, el aislamiento de los capacitores no es perfecta y la carga se pierde con el tiempo auque no sea accedida. Por esto, debe ser restablecida con cierta frecuencia la carga de los condensadores, lo cual implica tiempos de accesos mayores. Dentro de esta categoría podemos encontrar algunas variaciones como memoria dinámica tipo Extended Data Ouput (EDO DRAM) que proporciona un esquema de buffer interno modificado que mantiene los datos en la salida más tiempo que las DRAM convencionales. El otro tipo de memoria dinámica es la memoria dinámica síncrona (SDRAM, Syncronous DRAM) que implementa un

[3] Más detalles de este y otros términos se pueden encontrar en *http://www.webopedia.com*

modo de cauce por etapas que permite iniciar un segundo ciclo de acceso antes de ser completado el ciclo anterior.

Actualmente se cuenta con una variación de las SDRAM, *Double Data Rate (DDR SDRAM)*, las cuales doblan el acho de banda de la memoria y transfieren el doble de datos por ciclo de reloj.

Toda la información presente en los sistemas de memoria se pierde una vez que el computador se apaga o se reinicia. Es por eso que son necesarios los sistemas de almacenamiento secundarios tales como CD roms, floppies, discos duros, etc. De éstos el único medio realmente necesario es el disco duro ya que el resto de los dispositivos por lo general no se usan en ambientes de calculo intensivo.

Los discos duros mantienen copia del sistema operativo, programas y datos, así, se cuenta con un medio de almacenamiento para mantener grandes cantidades de información. Existen dos interfaces principales utilizadas para manejar discos duros: *IDE* y *SCSI*. Originalmente, las interfaces *IDE*, de menor rendimiento, tenían predominancia en el mercado de PCs debido al bajo costo en relación a los discos *SCSI*. Sin embargo, el actual abaratamiento de los costos de las interfaces *SCSI* han permitido incorporarlas en las nuevas tarjetas principales, aunque todavía los precios no son comparables con los discos *IDE*. Los discos duros *IDE* son mucho más fáciles de configurar que los discos duros *SCSI*. Por otro lado, es posible instalar hasta 7 discos duros *SCSI* en un PC y solamente 2 discos *IDE*. Así mismo, es común encontrar primero en el mercado discos duros SCSI de mayor capacidad que *IDE*. Actualmente existe una nueva tecnología en esta área *Serial ATA, (SATA)* que ofrece una tasa de transferencia de información comparable a los discos *SCSI* y a un costo comparable a los discos *IDE*. Por lo general, se recomienda utilizar discos *SCSI* en situaciones de altas tasas de operaciones de lectura y escritura, por ejemplo, directorios hogares, y utilizar discos *IDE* en situaciones de bajo número de accesos como por ejemplo espacios dedicados al sistema. Actualmente podemos encontrar otra tecnología en discos duros *Serial Attached SCSI (SAS)* que proporciona las mejores prestaciones en almacenamiento y durabilidad pero con costos superiores o otras alternativas. *SAS* es compatible con dispositivos *SATA* y los complementa agregando puertos duales, full duplex y direccionamiento de dispotivos[4].

La red de interconexión: convierte a un conjunto de computadores personales en un sólo sistema. Además, proporciona el acceso remoto al cluster y a sus servicios. Originalmente, fue posible crear clusters Linux debido a la disponibilidad de tecnología de red debajo costo y ancho de banda moderado. Ethernet fue el protocolo por excelencia utilizado en los inicios de los cluster, pero en la actualidad existe una gran variedad de tecnologías que pueden ser utilizadas para construir clusters Linux. Sin embargo, la relación costo-rendimiento Giga-Ethernet proporciona la mejor opción para implementar la red de un cluster. Otra razón para seleccionar esta topología de red es la facilidad para proporcionar escalabilidad a la hora de agregar nuevos nodos al cluster. Podemos encontrar tecnologías apropiadas como Myrinet e InfinitBand si se desea reducir la latencia en la transferencia de los mensajes entre los nodos.

4.1.1. Arquitectura del Cluster: Como hemos dicho, un cluster Linux difiere poco de una red de PCs. En la figura 4: Arquitectura de un Cluster Linux se muestra el esquema tradicional de la configuración de hardware de un cluster. La principal diferencia entre una red de PCs y un cluster es que en éste último se tiene una red privada para el intercambio de mensajes y los usuarios no ingresan a los nodos de cálculo. De esta manera se tienen todos los recursos dedicados al cálculo.

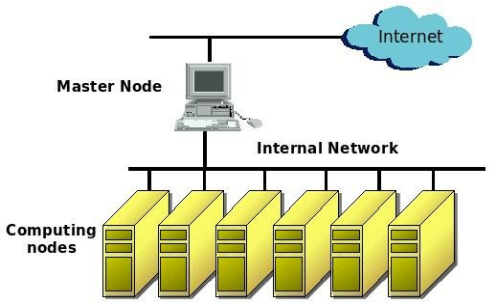

figura 4: Arquitectura de un Cluster Linux

4.2. Software del Cluster

GNU/Linux es el sistema operativo que brinda el soporte a todo el software utilizado en los clusters. Bibliotecas paralelas de código abierto (MPICH y LAM) implantan el estándar MPI y permiten la programación paralela de las aplicaciones y proveen el ambiente de ejecución paralela. Existe también otro paradigma de pase de mensajes implantado a través de PVM. Debemos contar con sistemas que gestionen colas de trabajos y que manejen eficientemente el balanceo de carga entre los nodos para aprovechar de una mejor forma los recursos. Distintas aplicaciones del mundo de software libre han sido adaptadas para facilitar las tareas de instalación, configuración, administración y supervisión de los clusters. Entre ellas existen distribuciones completas que satisfacen todas estas necesidades: NPACI Rocks, Oscar, Syld, etc. La figura 5: Arquitectura de Software muestra la arquitectura de software de un cluster Linux.

[4] *http://www.scsita.org*

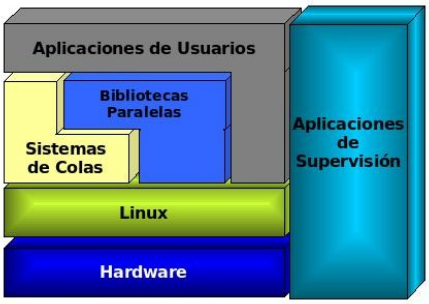

figura 5: Arquitectura de Software

5. Conclusiones

La redundancia natural que proveen los nodos de un cluster lo convierten en una plataforma confiable para la computación de alto rendimiento. Así mismo, su alta escalabilidad, posibilidad de incorporación de nuevos recursos sin eliminar los existentes, inversión inicial y software de bajo costo convierten a los clusters Linux en la opción con el mejor factor precio-rendimiento del mercado. También podemos concluir que el desarrollo de aplicaciones de software libre y avances en tecnologías de procesadores, memoria, discos y redes para PCs han propiciado la difusión de esta tecnología en todos los campos donde se requiere el uso de cómputo intensivo.

10. Referencias

[1] A.J. Forty (Chair), "Future Facilities for Advanced Research Computing" *The report of a Joint Working Party*, June 1985, Advisory Board for the Research Councils, Computer Board for Universities and Research Councils, University Grants Committee) SERC, 1985.

[2] R.G. Evans and S. Wilson (eds), *Supercomputational Science*, Plenum Press, New York, 1990.

[3] Spector, D. Building LINUX Clusters, O'Reilly, Sebastopol, California, 2002.

[4] Wilkinson, B., y Allen, M. Parallel Programming: Techniques and Applications Using Networked Workstations and Parallel Computers, Prentice Hall, Upper Saddle River, New Jersey, 1999.

[5] Gordon, B., y Gray J. High Performance Computing: Crays, Clusters, and Centers. What Next? Technical Report: MSR-TR-2001-76. Microsoft Corporation. 2001. http://reserach.microsoft.com/pubs

[5] Foster, I. Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering, Addison-Wesley, New York, 1995.

[6] Geist, A., Suderam, V., y otros. PVM: Parallel Virtual Machine. A User's Guide and Tutorial for Networked Parallel Computing, MIT Press, Massachusetts, 1994.

[7] Cormer, D. Computer Networks and Internet. 2da. Edicion. Prentice Hall, Upper Saddle River, New Jersey, 1999.

[8] Hoeger, H. Introducción a la Computación Paralela. Reporte Interno. Centro Nacional de Cálculo Cientifico Universidad de Los Andes. Mérida, Venezuela, 1997. http://www.cecalc.ula.ve/documentacion/manuales_tutoriales.html

[9] Snir, M., Dongarra, J., y otors. MPI: The Complete Reference. MIT Press, Massachusetts, 1996.

[10] Linux Rules Supercomputers, Daniel Lyons, 03.15.05, http://www.forbes.com/home/enterprisetech/2005/03/15/cz_dl_0315linux.html

[11] Moshe Bar, Stefano Cozzini, Mauricio Davini, Alberto Marmodoro OpenMosix vx Beowulf: A case of study INFM Democritos (Italy), OpenMosix Project, Department of Physics University of Pisa (Italy)

10.1 Sitios en Internet

1) *CeCalCULA*: http://www.cecalc.ula.ve/
2) Laboratorio Argon http://www.mcs.anl.gov/dbpp/
3) Netlib http://www.netlib.org/pvm3/book/pvm-book.html/
4) Información sobre *MPI: Message Passing Interfae*: http://www.mcs.an.gov/mpi/index.html
5) Sitio con los 500 supercomputadores más poderosos: http://www.top500.com
6) National HPCC Software Exchange: http://nhse.npac.syr.edu/
7) Enciclopedia electrónica de términos computacionales: http://www.webopedia.com
8) Beowulf: http://www.beowulf.org