Comunicaciones para Cómputo Paralelo Intercluster

Walter Aróztegui, Fernando L. Romero, Fernando G. Tinetti[1]

Instituto de Investigación en Informática LIDI (III-LIDI)
Facultad de Informática – UNLP

Centro de Técnicas Analógico-Digitales (CeTAD)
Facultad de Ingeniería - UNLP
waroz@graffiti.net, fromero@lidi.info.unlp.edu.ar, fernando@info.unlp.edu.ar

CONTEXTO

Esta línea de Investigación forma parte de dos de los Subproyectos dentro del Proyecto "Sistemas Distribuidos y Paralelos" acreditado por la UNLP y de proyectos específicos apoyados por CyTED, CIC, Agencia e IBM.

RESUMEN

Esta línea de investigación se orienta a resolver el problema general de las comunicaciones entre clusters usados para cómputo paralelo. La idea e partida es todo lo conocido de cómputo paralelo en clusters. En este sentido, utilizar más de un cluster para cómputo paralelo se puede considerar como una extensión natural del procesamiento paralelo en plataformas de cómputo distribuidas. En esta línea de investigación se busca, por lo tanto, resolver el problema de comunicar procesos en diferentes computadoras que eventualmente pueden pertenecer a dos o más clusters y, además, caracterizar el rendimiento de las mismas.

Aunque las comunicaciones entre clusters pueden ser consideradas *triviales* con el uso de Internet, la situación se complica cuando se deben tener en cuenta las características de seguridad involucradas. En este sentido, se deben resolver los problemas de seguridad de manera sustentable en cuanto a que se puedan aplicar estas soluciones a ámbitos de administración disjuntos y colaborativos. Como mínimo, se deben establecer las políticas a manejar en cada cluster local involucrado y quizás algunas relacionadas con los mecanismos de ruteo.

Desde el punto de vista del rendimiento, la situación es bastante más complicada. Normalmente, las comunicaciones entre diferentes clusters son compartidas con tráfico estándar de Internet de las instituciones involucradas. En este contexto, es necesario por lo menos caracterizar los intervalos de tiempo de mayor congestión y/o el rendimiento esperable a lo largo del tiempo de uso de la interconexión entre los clusters.

Keywords: *Comunicación de Procesos, Carcterización de Rendimiento, Rendimiento de Comunicaciones, Sistemas Paralelos y Distribuidos, Paralelismo en Clusters e Intercluster.*

1. INTRODUCCION

Desde las primeras propuestas de uso de clusters para cómputo paralelo o al menos distribuido, ha quedado clara la importancia de las comunicaciones [5] [3]. Esta situación no es nueva, ya que es conocida en el ámbito de procesamiento paralelo clásico [1] [2]. Una de las primeras propuestas de uso libre para resolver las comunicaciones fue PVM (Parallel Virtual Machine) [5] y, de hecho, marcó muchas de las características de lo que luego se estandarizó como MPI (Message Passing Interface) [6]. Las

[1] Investigador Asistente CICPBA

primeras implementaciones de MPI de uso libre fueron LAM/MPI [7] [13] y MPICH [9] [10] que, por supuesto, implementan el estándar MPI y resuelven satisfactoriamente el problema de comunicar procesos que se ejecutan en diferentes computadoras de un cluster.

Una vez establecidas las bibliotecas precedentes, la tendencia ha sido y es caracterizar el rendimiento de las comunicaciones [4]. Más recientemente, con las múltiples propuestas de hardware de comunicaciones, la idea ha sido caracterizar su rendimiento para su comparación [12]. En el contexto de uso de más de un cluster para cómputo distribuido, todavía se está, de alguna manera, en la etapa de propuestas, como la de grid [8].

Sin llegar al ámbito más genérico de grid computing, la idea inicial de esta línea de investigación es la de utilizar dos o más clusters para resolver un problema en paralelo [15]. En este sentido, se tienen dos líneas de acción que corresponden a dos ámbitos en principio diferentes pero necesarios para cómputo intercluster: seguridad (con la consecuente necesaria sustentabilidad de las soluciones técnicas propuestas/ adoptadas) y rendimiento caracterizable y, en el mejor de los casos, optimizado.

2. LINEAS DE INVESTIGACION Y DESARROLLO

Tal como se enuncia en la sección anterior, en principio de deben resolver dos problemas en principio no relacionados: seguridad y rendimiento. Entre estos dos problemas se deber resolver, por supuesto, la comunicación entre procesos que se ejecutan en clusters diferentes.

Inicialmente, la idea es investigar cómo llegar a tener en dos clusters un entorno de lógico de ejecución como el de un único cluster. En este sentido, se cuenta con lo básico de una red local en un cluster y a partir de allí se construye lógicamente una plataforma de cómputo paralelo. La Fig. 1 muestra este punto de partida, donde todo lo que se tiene es una red local con computadoras estándares de escritorio.

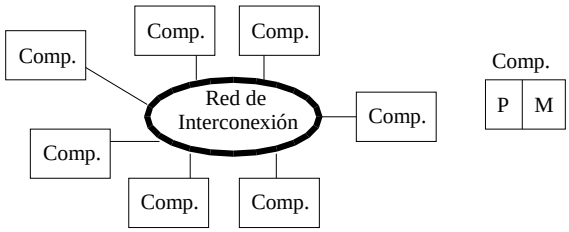

Figura 1: Red Local.

A partir de esta red local, en cada computadora se usan las bibliotecas de desarrollo y ejecución de programas de cómputo paralelo. La gran mayoría de las veces estas bibliotecas son, en realidad, para pasaje de mensajes entre procesos. La Fig. 2 muestra esquemáticamente cómo, desde el punto de vista lógico, se *transforma* una red local en una computadora paralela de pasaje de mensajes.

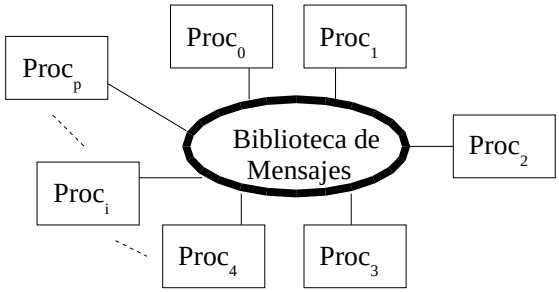

Figura 2: Cómputo Paralelo con Pasaje de Mensajes en una Red Local.

Con esta idea inicial, el objetivo es ahora transformar dos clusters (o redes locales) en una única plataforma de cómputo paralelo. Aunque en principio puede resultar sencillo de implementar, se deben tomar varias decisiones de diseño de la solución adoptada, por ejemplo:

- Utilización de los protocolos de red estándares y de relativo bajo nivel (TCP/IP, por ejemplo).
- Utilización de middleware ya desarrollado adaptándolo a las características

de cómputo intercluster (VPN: Virtual Private Network, por ejemplo).
- Utilización de herramientas que están en desarrollo (IMPI: Interoperable MPI, por ejemplo [11]).

Parte de esta tarea consiste en analizar ventajas y desventajas, tratando de cuantificar las similitudes y diferencias. Por otro lado, también se debe tener en cuenta la evolución de las herramientas que se seleccionen y/o propongan.

Desde la perspectiva de procesamiento paralelo, la idea de utilizar varias computadoras (estén en uno o varios clusters) consiste en el arranque de procesos remotos más la posibilidad de comunicaciones con esos procesos remotos. Cuando hay varios clusters involucrados, las estrategias actuales de seguridad hacen que ninguna de estas tareas es sencilla, tal como lo muestra esquemáticamente la Fig. 3 para dos clusters.

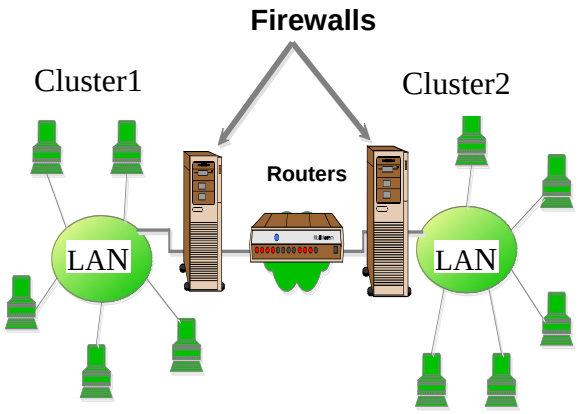

Figura 3: Interconexión de Dos Clusters.

Por otro lado, la caracterización de rendimiento también intenta seguir los lineamientos clásicos del contexto de cómputo paralelo. En este contexto, la modelización mínima del rendimiento de las comunicaciones se realiza estimando experimentalmente dos parámetros: latencia (o tiempo de arranque de las comunicaciones) y ancho de banda. Con estos parámetros, se modeliza el tiempo necesario para la comunicación de un mensaje de *n* datos como:

$$t(n) = \alpha + \beta\, n$$

donde α es el tiempo de startup y β es la inversa del ancho de banda. Lamentablemente, en el contexto actual de comunicaciones intercluster, el tráfico de estas comunicaciones tendrá que competir con tráfico estándar de Internet, dado que:

- Es muy difícil o casi imposible que se tenga una red de interconexión de uso exclusivo entre los dos clusters.
- Es muy poco probable contar con calidad de servicio en la o las redes involucradas para la interconexión de los dos clusters.

Esto significa que la modelización anterior varía a lo largo del tiempo. Más específicamente, tanto el tiempo de latencia como el ancho de banda dependen del tráfico con el que se tiene que *competir* para llegar desde un cluster al otro.

Además de las características dinámicas del rendimiento de las comunicaciones entre los clusters, se deben tener en cuenta otros factores que impactan sobre el rendimiento. El más importante desde la perspectiva de cómputo paralelo se da por, justamente, la diferencia entre la visión lógica del sistema de cómputo (básicamente la de la Fig. 2) y la realidad física del mismo (básicamente la de la Fig. 3). Mientras que en una aplicación paralela todos los procesos se pueden comunicar entre sí *idealmente* con las mismas características de rendimiento, cuando hay más de un cluster involucrado evidentemente no se puede implementar de esta manera. Más específicamente, se debe tener en cuenta que las comunicaciones *locales* a un cluster son significativamente mejores (en rendimiento) que las comunicaciones entre máquinas de diferentes clusters.

3. RESULTADOS OBTENIDOS/ESPERADOS

Inicialmente, se estudiaron diferentes formas de arrancar procesos de manera segura en diferentes computadoras. A partir del análisis de varias posibilidades, se llegó relativamente rápido a que la mejor forma (y la más estándar) es la provista por ssh

[14]. Posteriormente, se estudiaron las bibliotecas estándares de pasaje de mensajes y se llegó a que, aunque el problema de disparo remoto puede ser fácilmente resuelto por/con ssh, las comunicaciones entre procesos no son tan sencillas de resolver desde la perspectiva de los firewalls [15].

Por otro lado, se estudió el problema de analizar el rendimiento de las comunicaciones intercluster. En este sentido, además de llevar a cabo experimentación específica para dos clusters también se delinearon las estrategias/metodologías para que esta experimentación se pueda llevar a cabo en general [16]. La idea en este sentido es contar, al menos, con información acerca de la evolución en el tiempo de los índices de latencia y ancho de banda a partir de los cuales se pueda derivar el rendimiento de las comunicaciones en un intervalo de tiempo determinado. Aún más, esta experimentación también puede ser utilizada para una estimación preliminar de la confiabilidad de la conexión entre los clusters.

En resumen, la idea es contar con un conjunto de herramientas para:

- Desarrollar y ejecutar aplicaciones paralelas en varios clusters. En este sentido, ya se ha llegado a que las bibliotecas de pasaje de mensajes estándares no son suficientes en sí mismas, aún con la utilización de ssh. Los problemas a resolver son básicamente de seguridad/interacción con los firewalls.
- Analizar el rendimiento de la red de interconexión y/o estimar el rendimiento para un período de tiempo dado. En este sentido, ya se cuenta con una metodología mínima pero efectiva que también se puede utilizar para obtener información sobre la confiabilidad de la interconexión. De todas maneras, se debe extender esta metodología y/o desarrollar una herramienta de software lo más *automática* posible.
- Optimizar la comunicación entre los procesos que se ejecutan en diferentes clusters. Si bien podría considerarse dentro de la herramienta de desarrollo y ejecución de aplicaciones mencionada antes, se estima *a priori* que dentro de esta herramienta las comunicaciones entre los clusters deben ser optimizadas por razones de rendimiento. Debe recordarse el rendimiento de la interconexión entre clusters suele ser al menos un orden de magnitud menor que el rendimiento intracluster.

4. FORMACION DE RECURSOS HUMANOS

En esta línea de I/D existe cooperación a nivel nacional e internacional. Inicialmente se tiene una posible tesis de maestría y está abierta la posibilidad para varias Tesinas de Grado de Licenciatura.

5. BIBLIOGRAFIA

[1] Akl S., The Design and Analysis of Parallel Algorithms, Prentice-Hall, Inc., 1989.

[2] Akl S., Parallel Computation: Models and Methods, Prentice-Hall, Upple Saddle River, 1997.

[3] Anderson T., D. Culler, D. Patterson, and the NOW Team, “A Case for Networks of Workstations: NOW”, IEEE Micro, Feb. 1995.

[4] S. Araki, A. Bilas, C. Dubnicki, J. Edler, K. Konishi, and J. Philbin, “User-space communication: A quantitative study”, In *SC98: High Performance Networking and Computing*, November 1998.

[5] Dongarra J., A. Geist, R. Manchek, V. Sunderam, Integrated pvm framework supports heterogeneous network computing, Computers in Physics, (7) 2, pp. 166-175, April 1993.

[6] MPI Forum, “MPI: a message-passing

interface standard", International Journal of Supercomputer Applications, 8 (3/4), pp. 165-416, 1994.

[7] Burns G., R. Daoud, J. Vaigl, "LAM: An Open Cluster Environment for MPI", Proceedings of Supercomputing Symposium, pp. 379-386, 1994. Available at http://www.lammpi.org/download/files/lam-papers.tar.gz

[8] Foster I., The Grid: Blueprint for a New Computing Infrastructure, 2nd Edition, Morgan Kaufmann, 2004. ISBN: 1-55860-933-4.

[9] Gropp W., E. Lusk, "Sowing MPICH: A Case Study in the Dissemination of a Portable Environment for Parallel Scientific Computing", The International Journal of Supercomputer Applications and High Performance Computing, Vol. 11, No. 2, pp. 103-114, Summer 1997,

[10] Gropp W., E. Lusk, N. Doss, A. Skjellum, "A high-performance, portable implementation of the MPI message passing interface standard", Parallel Computing, Vol. 22, No. 6, pp. 789-828, Sep. 1996.

[11] IMPI Steering Committee, IMPI - Interoperable Message-Passing Interface DRAFT March 22, (NIST) National Institute of Standards and Technology, 1999. Disponible en: http://impi.nist.gov/.

[12] Liu J., B. Chandrasekaran, W. Yu, J. Wu, D. Buntinas, S. Kini, P. Wyckoff, and D. K. Panda, "Micro- Benchmark Performance Comparison of High-Speed Cluster Interconnects", IEEE Micro, January/ February, 2004.

[13] Squyres J. M., A. Lumsdaine, "A Component Architecture for LAM/MPI", Proceedings, 10th European PVM/MPI Users' Group Meeting, pp. 379-387, 2003, Venice, Italy, Springer-Verlag Lecture Notes in Computer Science 2840, September/ October 2003.

[14] Tinetti F. G., Aróztegui W., "Instalación y Configuración de ssh para Cómputo Intercluster", Reporte Técnico PLA-002-2005, III-LIDI, Facultad de Informática, UNLP, CeTAD, Facultad de Ingeniería, UNLP, Argentina, Junio 2005. Disponible en https://lidi.info.unlp.edu.ar/~fernando/publis/intercl1.pdf

[15] Tinetti F. G., Aróztegui W., "Bibliotecas de Pasaje de Mensajes y Cómputo Intercluster", Reporte Técnico PLA-003-2005, III-LIDI, Facultad de Informática, UNLP, CeTAD, Facultad de Ingeniería, UNLP, Argentina, Septiembre 2005. Disponible en https://lidi.info.unlp.edu.ar/~fernando/publis/portsrep.pdf

[16] Tinetti F. G., Aróztegui W., "Perfil Preliminar de las Comunicaciones Intercluster", Reporte Técnico PLA-001-2006, III-LIDI, Facultad de Informática, UNLP, CeTAD, Facultad de Ingeniería, UNLP, Marzo de 2006. Disponible en https://lidi.info.unlp.edu.ar/~fernando/publis/pprep.pdf