Capacidad de Comunicaciones Disponible para Cómputo Paralelo en Redes Locales Instaladas

Fernando G. Tinetti
fernando@ada.info.unlp.edu.ar

Antonio A. Quijano
quijano@ing.unlp.edu.ar

Grupo de Investigación en Procesamiento Paralelo[1]
Instituto Tecnológico de Buenos Aires
Av. Eduardo Madero 399
(1106) Buenos Aires, Argentina

Resumen

Uno de los grandes desafíos del cómputo paralelo en las redes locales instaladas y que se pueden aprovechar como máquinas paralelas lo constituyen las comunicaciones. Si bien existen numerosos reportes y se ha aceptado que la mayoría de las máquinas de las redes locales instaladas no tienen carga de procesamiento de manera continua, no se tiene mucha información con respecto a la capacidad disponible de comunicaciones en estas redes que pueden ser afectadas por varios factores como el propio hardware de cada computadora, el cableado (*wiring rules*) de la red y todo el tráfico de información relacionado con Internet que se genera tanto desde las aplicaciones interactivas (como un navegador de red) como desde aplicaciones que se asumen en *background* (como los servidores de correo electrónico o de páginas web). En este artículo se presentan las consideraciones necesarias para establecer una metodología de estudio de la capacidad de comunicaciones disponible para las aplicaciones paralelas que se ejecutan sobre redes locales instaladas. También se presentan algunos resultados preliminares con respecto a la capacidad de comunicaciones entre computadoras cuando se considera que la red está básicamente libre de tráfico interactivo y por lo tanto se puede considerar que es totalmente disponible para las aplicaciones paralelas a resolver.

Palabras Clave: Cómputo Paralelo, Aprovechamiento de Redes Locales para Cómputo Paralelo, Pasaje de Mensajes, Rendimiento de una Red de Interconexión, Carga de Comunicaciones.

1.- Introducción

Tanto en el contexto del aprovechamiento optimizado de las redes locales (o LAN: Local Area Network) para distribución de cómputo secuencial como para cómputo paralelo, es aceptado desde hace varios años que las computadoras están inactivas durante un gran porcentaje del tiempo en que se pueden utilizar [8] [1]. La disponibilidad de este tipo de redes que se pueden aprovechar para optimizar la tarea de distribución de cómputo secuencial (proveyendo un cluster de *alta productividad* [2] por medio de ambientes como CONDOR [9]), y también para cómputo paralelo ha sido y es muy favorecida por factores tales como:

- El muy bajo costo de las computadoras de escritorio, además de su crecimiento constante en cuanto a capacidad de almacenamiento y rendimiento de cómputo. Esto ha llevado a considerar el cómputo paralelo en clusters como el *mejor* en cuanto a la relación costo/rendimiento.
- El muy bajo costo de interconexión (tanto en hardware como en software) de las redes locales, que en su gran mayoría utilizan la norma Ethernet [7].

1 GIPP, Departamento de Coordinación de Investigación y Desarrollo, ITBA.

- El crecimiento en cuanto a instalación de redes locales ha sido constante en los últimos años y esto reduce aún más todos los costos involucrados, desde el costo mismo de cada máquina como el costo del hardware de interconexión (placas, cables, hubs y/o switches, personal técnico capacitado, etc.).
- Las bibliotecas de pasaje de mensajes para desarrollo y ejecución de programas paralelos o distribuidos como lo son PVM [3] y las implementaciones de uso libre de la norma MPI [14] [13]. La propia distribución del hardware de cómputo (computadoras de escritorio estándares) a utilizar en paralelo implica casi de manera directa la utilización del modelo de pasaje de mensajes para las programación de aplicaciones paralelas y en este sentido las bibliotecas mencionadas no hacen más que proveer una implementación de uso libre de este modelo de programación para las redes de computadoras.

Existen numerosos reportes donde se muestra la evolución de la utilización de recursos de cómputo (básicamente computadoras) de una LAN en relación, por ejemplo, con los días de un mes completo de cómputo [2]. Sin embargo, no es tan clara la situación con los recursos de comunicaciones, donde no solamente se tiene que tener en cuenta la capacidad física de la red de interconexión de computadoras sino que además esto recursos de interconexión pueden ser utilizados por las computadoras que no intervienen en el procesamiento de una aplicación paralela.

Uno de los mayores inconvenientes de rendimiento del procesamiento paralelo en los clusters de computadoras es el rendimiento de las comunicaciones [10] y esto implica de manera directa la necesidad de conocer de la manera más precisa posible la capacidad de comunicaciones de la red de interconexión. El análisis de la capacidad de la red de interconexión es muy complejo en el contexto de las redes de computadoras instaladas que se utilizan para cómputo paralelo, dado que existen varios factores difíciles de acotar y/o estimar *a priori*, como lo son:

- El tráfico de las aplicaciones que se ejecutan en las computadoras de la red local que no son utilizadas para las aplicaciones paralelas. Se reconoce en este contexto que las aplicaciones paralelas utilizan las computadoras libres de carga de cómputo y que en la misma red de interconexión pueden existir otras computadoras que se utilizan para otro/s propósito/s. Se podrían considerar diferentes fuentes de generación de tráfico, tales como:
    - Servicios propios de la red local, tales como: acceso a datos y/o sistemas de archivos compartidos e impresoras compartidas.
    - Acceso a Internet, tales como los servicios interactivos asociados a los *navegadores* o *browsers* y los servicios no interactivos de los servidores de correo electrónico.
- El hardware de interconexión, y más específicamente el cableado de la red tiene algunas variaciones importantes dependiendo, por ejemplo, del uso de switches o hubs.
- La sobrecarga de las propias herramientas de desarrollo y ejecución de software paralelo de las cuales no se conoce mucho *a priori* en lo que se refiere a su implementación.

Por lo tanto, aunque existen algunas posibilidades de modelización del rendimiento del hardware y de hecho las propias placas de interconexión de red que se instalan en cada computadora son de un rendimiento específico (Ethernet de 100 Mb/s, por ejemplo), no es posible contar *a priori* con un modelo de rendimiento de las comunicaciones entre procesos de una aplicación paralela. Es por esto que se debe recurrir a la experimentación para intentar la aproximación del rendimiento *real* de las comunicaciones entre las computadoras instaladas y que se pueden aprovechar para cómputo paralelo.

En la siguiente sección se enuncian algunas de las características más importantes a tener en cuenta tanto en la evaluación de rendimiento de las comunicaciones en general como en la monitorización de la carga de comunicaciones de una red local en particular en la que se ejecutarán aplicaciones paralelas junto con otras aplicaciones de usuario. En la tercera sección se dan las ideas principales para el desarrollo y la implementación de metodología de monitorización y estimación de la carga de comunicaciones de una red local instalada. En la cuarta sección se muestran algunos resultados obtenidos al evaluar el rendimiento de las comunicaciones en una red en particular, en el

contexto de considerar que no existen aplicaciones interactivas utilizando la red de interconexión o con grandes requerimientos de comunicaciones. En la última sección se dan las conclusiones y las principales líneas de acción para el trabajo futuro en esta área. Finalmente, se dan las referencias bibliográficas a las que se hace referencia a lo largo de este artículo.

2.- Evaluación de Rendimiento y de Carga de Comunicaciones en una LAN

Desde el punto de vista de las aplicaciones paralelas que se ejecutan sobre las redes locales instaladas, se podrían distinguir dos aspectos del rendimiento de comunicaciones de la red de interconexión:

- El máximo rendimiento posible de obtener al transferir datos desde un proceso de usuario a otro. Es importante distinguir este máximo real de la máxima capacidad de comunicaciones del que proporciona el hardware utilizado (Ethernet 10 Mb/s, 100 Mb/s, 1 Gb/s). En la evaluación a nivel de procesos de usuario se tiene en cuenta, por ejemplo, la sobrecarga (*overhead*) que impone la biblioteca de pasaje de mensajes, mas la de la pila de protocolos de los sistemas operativos (TCP, UDP, IP) más la sobrecarga de la interfase con el hardware de comunicaciones (*drivers* de las placas de red). Es importante en este punto que la red de comunicaciones se utilice solamente con este propósito, es decir que esté libre de otro tráfico de datos que no sean los de los propios experimentos de evaluación.
- La capacidad de comunicaciones a lo largo de un día normal de procesamiento en la red local, donde pueden coexistir las aplicaciones paralelas que se ejecutan en las máquinas libres o disponible con las demás aplicaciones de usuario en las demás computadoras que están siendo utilizadas por sus usuarios. De alguna manera, esta es la verdadera carga de la red que se debe monitorear a lo largo de cada día normal de trabajo, dado que cuando la red está libre de transferencias de usuarios las aplicaciones paralelas no generan ningún inconveniente.

Usualmente al máximo rendimiento de la red local se le ha denominado directamente *rendimiento*, y la también usualmente la *carga de comunicaciones* ha hecho referencia a la utilización real de la capacidad de rendimiento o también al nivel de congestionamiento de la red de interconexión. Si, por ejemplo, se tiene una red Ethernet de 10 Mb/s, desde el punto de vista del hardware se tiene el rendimiento de 10 Mb/s y la carga que imponen las aplicaciones puede variar entre 0 (no se utiliza la red de interconexión) hasta el propio máximo del hardware de 10 Mb/s. En todos los casos se debe recordar que, como se aclaró antes, el máximo rendimiento de las comunicaciones entre procesos de usuario generalmente es menor que el del hardware y la diferencia se halla utilizando modelos y métodos experimentales.

Desde el punto de vista de las aplicaciones paralelas, la carga de la red de interconexión normalmente se evalúa con respecto al máximo posible para tener una idea de rendimiento y de relación de cómputo/comunicaciones (o granularidad) real o disponible a lo largo del día. Desde el punto de vista del aprovechamiento de las instalaciones para cómputo paralelo y de los usuarios de las computadoras instaladas en las redes locales, es muy importante que la carga que introducen las aplicaciones paralelas en la red de interconexión no sea tan grande como para que los usuarios que *coexisten* se vean afectados en cuanto al retardo de las comunicaciones de sus aplicaciones (con los sistemas de archivos compartidos, por ejemplo). Como efecto colateral, la monitorización de la carga de comunicaciones de la red de interconexión también puede ser aprovechada para la evaluación de actualizaciones de hardware (si la red está continuamente congestionada, por ejemplo) y la verificación de funcionamiento de las placas de red y del cableado (para reemplazar hardware, por ejemplo).

La evaluación del máximo rendimiento de la red de interconexión no presenta dificultades en términos generales, asumiendo que se modeliza el rendimiento por la vía de la experimentación. El

modelo de rendimiento más comúnmente aceptado es el que utiliza los índices de tiempo de latencia (o *startup*) y ancho de banda asintótico. Con estos dos índices se estima el tiempo de transmisión para los mensajes de longitud *n* (unidades en *bytes*, *floats*, o algún otro tipo de dato básico) con la Ec. (1)

$$t(n) = \alpha + n/\beta \quad (1)$$

donde α es el tiempo de latencia o *startup* de las comunicaciones y β es el ancho de banda asintótico de la red de interconexión [6] [5] [11] [4] [12]. Sin embargo, la estimación de α y β no es una tarea sencilla en general cuando se asume que las aplicaciones paralelas se ejecutan con otras aplicaciones de usuario en la red local que pueden generar tráfico sobre la red interconexión.

Dado que el objetivo es analizar el rendimiento de la red de interconexión de una red local que se aprovecha para tareas de cómputo paralelo, se distingue el análisis de los índices α y β en dos situaciones distintas de la red de interconexión:

1. Disponibilidad completa, es decir que no hay aplicaciones de usuarios de la red local que tengan la necesidad de utilizar la red de interconexión. Se podría asociar directamente a los períodos *clásicos* de inactividad *completa* de las computadoras. También se podría mencionar la utilización de las computadoras de manera aislada de la red, pero esta forma de uso de las computadoras de una LAN es muy poco frecuente y por lo tanto se puede descartar *a priori*.
2. Existencia tráfico en la red local perteneciente a las tareas de usuarios de las máquinas de la red local que no deberían notar la *sobrecarga* de la red de interconexión impuesta por los mensajes de las aplicacione/s paralela/s.

En todos los casos, los índices α y β proporcionan la información mínima indispensable necesaria para las aplicaciones paralelas con respecto al rendimiento de la red de interconexión, Sin embargo, los los índices α y β que se obtienen en el contexto de la disponibilidad completa de la red de interconexión (sin otro tráfico de usuarios de las computadoras de la LAN) proporcionan información muy importante con respecto a:

- Como se aclaró antes, la sobrecarga de todo lo relacionado con la comunicación entre procesos de usuario, desde las capas de software del sistema operativo (drivers, pila de protocolos, etc.) hasta las bibliotecas de pasaje de mensajes entre procesos de una aplicación paralela (PVM, implementación de MPI, o la que se utilice).
- La carga de comunicaciones se debe relacionar de manera directa con la capacidad máxima de comunicaciones entre procesos. Si, por ejemplo, no se conoce el máximo ancho de banda entre los procesos de usuario, se puede llegar a conclusiones erróneas respecto de la carga de la red de procesadores en la cual se puede incluir la sobrecarga de las diferentes capas de software mencionadas anteriormente sin haber ningún tipo de tráfico en la red. Suponiendo que se tiene, por ejemplo, una red de Ethernet de 10 Mb/s y que la sobrecarga de las capas de software de comunicación entre procesos hace que el máximo ancho de banda sea de 8 Mb/s esto no significa que el 20% de diferencia sea debido a la carga de la red de interconexión.

Por otro lado, los índices α y β que se obtienen en el contexto de la existencia de tráfico de usuarios de las computadoras de la LAN tienen otras características y proporcionan otro tipo de información:

- Como mínimo pueden variar en el tiempo (a lo largo del día), ya que no necesariamente la carga de una red local es la misma en todo instante de tiempo y esta variación hace más complicada la modelización de rendimiento. Por otro lado, esto implica de manera directa que el tipo de aplicaciones paralelas que se pueden ejecutar pueden depender de los valores de α y β en un instante o período dado.
- Como se aclaró antes, se debe tener en cuenta para que las aplicaciones paralelas que aprovechan

las computadoras libres de la red local *también* utilicen solamente la capacidad *disponible* de la red de interconexión. De otra manera, el tráfico en la red que introducen las aplicaciones paralelas pueden producir una degradación significativa del rendimiento de las aplicaciones de los usuarios de la red local. Los requerimientos que se realicen desde un navegador, por ejemplo, pueden demorar varias veces más de lo que es estándar y los usuarios pueden notar la degradación en el tiempo de respuesta.

3.- Metodología de Evaluación de Rendimiento y de Carga de Comunicaciones en una LAN

A partir de lo explicado en la sección anterior, tanto el problema de la evaluación de rendimiento como de la carga de una red de comunicaciones se enfoca hacia la obtención de los valores de los índices α y β. En el caso de la evaluación de rendimiento (o rendimiento *máximo*), el objetivo es encontrar los valores de α y β con la red libre (sin otro tráfico de datos), y en el caso de la identificación de la carga de comunicaciones se debería encontrar la forma en que varían los valores de los índices α y β en el tiempo, dependiendo del tráfico que los usuarios de las computadoras de la red local generan desde sus aplicaciones. Es interesante notar que el tiempo de startup de comunicaciones es afectado no solamente por las computadoras que intervienen en una transferencia de datos sino también por el tráfico en la red, y esto se debe al mismo protocolo Ethernet que es denominado CSMA/CD (Carrier Sense Multiple Access / Collision Detect) [7] y su forma de resolver los conflictos de acceso al único medio de comunicaciones (bus lógico). Dado que el contexto de búsqueda de los valores de α y β es diferente, se define una metodología para encontrar experimentalmente los valores correspondientes a la red libre, sin más tráfico que el de los experimentos mismos y otra metodología para la búsqueda de la variación de α y β dependiendo de la carga de tráfico sobre la red de comunicaciones de las aplicaciones de usuario.

3.1 Rendimiento de la Red de Comunicaciones (Red Libre)

El método *clásico* de búsqueda de los valores de startup y ancho de banda de las comunicaciones se puede aplicar de manera directa en este contexto. Es conocido como *método ping-pong* por la utilización de dos procesos entre los que se envían-reciben un conjunto de datos. Básicamente, el método se puede describir con el pseudocódigo simplificado que cada uno de los procesos ejecuta, que se muestra en la Fig. 1

```
Proceso ping
    iniciar timer
    send a pong
    recv de pong
    t = tiempo de send-recv
    tiempo de com. = t / 2
End ping
```

```
Proceso pong
    recv de ping
    send a ping
End pong
```

Figura 1: Pseudocódigo de los Procesos ping-pong.

La estimación del valor del tiempo de startup de las comunicaciones, α, se realiza con mensajes de longitud mínima o de longitud 0 (si la biblioteca de pasaje de mensajes lo permite).La estimación del valor de ancho de banda asintótico de la red de comunicaciones, β, se obtiene a partir del tiempo

de comunicación de mensajes de la mayor longitud posible, donde se asume que el tiempo de startup carece de importancia en cuanto a magnitud relativa con respecto al total de tiempo de transferencia de datos. En todos los casos se toman recaudos para evitar factores que distorsionan las medidas, tales como el acceso a los datos físicos (memoria virtual-cache) y la falta de sincronización de los procesos. Normalmente se promedian los valores obtenidos de una cantidad *estadísticamente estable* de experimentos.

3.2 Carga o Tráfico de la Red de Comunicaciones (Red *Parcialmente* Utilizada)

Si bien el método ping-pong se puede aplicar de manera medianamente directa, se deben evaluar las consecuencias que el tráfico de los propios experimentos puede causar. Si se asume que en la red local se tienen otros usuarios con sus correspondientes aplicaciones, en principio ni las aplicaciones paralelas ni los propios experimentos de evaluación de carga de la red de comunicaciones deberían afectar a estos usuarios-aplicaciones. Esto significa que se debe controlar y mantener al mínimo posible el tráfico sobre la red de comunicaciones que generan los experimentos diseñados para la obtención de los valores de α y β. En este sentido, conviene identificar con claridad los tamaños de los mensajes utilizados y la cantidad de mensajes que los experimentos generan. La carga de tráfico que se impone sobre la red de comunicaciones depende de la cantidad de mensajes por unidad de tiempo (o frecuencia de los experimentos) y de la cantidad y longitud de los mensajes que se transmiten.

La frecuencia de los experimentos se relaciona de manera directa con la precisión con la cual se necesita conocer la variación de la capacidad de la red de interconexión disponible para las aplicaciones paralelas. Se podría considerar que una hora es período suficientemente preciso para asumir que la carga sobre la red de comunicaciones es relativamente constante, y por lo tanto los experimentos se llevarán a cabo una vez por hora. Expresado de otra manera, se asume que la carga de la red de interconexión es constante por períodos de una hora de duración o lo que es lo mismo: la capacidad de la red de interconexión disponible para las aplicaciones paralelas es la misma por períodos de una hora de duración. Es importante recordar este hecho dado que, por un lado no se pueden asumir períodos mucho más largos de estabilidad en cuanto a carga de comunicaciones sobre la red por el dinamismo propio de los usuarios, y por otro lado las aplicaciones paralelas que excedan el tiempo de ejecución de una hora deberían tener en cuenta que la red de comunicaciones no será la *misma* en cuanto a rendimiento aunque sea la misma en cuanto a capacidad de interconexión de computadoras.

Como se explica antes, los mensajes para la identificación del tiempo de startup deben ser de longitud mínima o cero, con lo que en realidad no implican grandes problemas de tráfico a menos que se generen miles o millones de estos mensajes. El problema más grande respecto al tráfico generado sobre la red de comunicaciones está asociado a los mensajes utilizados para la identificación del ancho de banda asintótico (en este caso *disponible*) de la red. Si bien se acepta que se deberían utilizar mensajes con una cantidad de datos relativamente grande, no hay una propuesta uniforme en cuanto a longitud de los mensajes. De los experimentos realizados en el contexto de la red libre, se puede obtener de manera más o menos sencilla la cantidad de datos mínima con la cual se obtiene una buena estimación del ancho de banda de la red. De hecho, en el contexto de la red libre se pueden utilizar mensajes arbitrariamente grandes y con una cantidad de mensajes mucho mayor de la necesaria para la estimación de los índices de rendimiento, y estos pueden ser aprovechados en el contexto de la red de comunicaciones compartida con otros usuarios-aplicaciones.

A partir de la identificación de la mínima longitud de mensajes que se puede utilizar para la estimación del ancho de banda de las comunicaciones, *Lm*, se puede conocer el tráfico que generan los experimentos sobre la red de interconexión, y está dado por la Ec. (2) para cantidad de bytes por

hora asumiendo que *Lm* está dada en bytes.

$$Tráf_{B/h} = m\# \times Lm\ /h \tag{2}$$

donde *m#* es la cantidad de mensajes que se generan por los experimentos pong-pong. A partir de ahora se puede hacer una relación entre el ancho de banda que se utiliza para los experimentos con respecto al ancho de banda físico de la red de interconexión. Dado que normalmente el ancho de banda físico está dado en 10^6 bits por segundo (Ethernet 10 Mb/ ó 100 Mb/s, por ejemplo), conviene reescribir la Ec. (2) en términos de Mb/s, tal como lo muestra la Ec. (3)

$$Tráf_{Mb/s} = ((m\# \times Lm \times 8)/3600/10^6)\ /s \tag{3}$$

Por lo tanto si se decide, por ejemplo, que se puede utilizar un 10% del ancho de banda físico de la red de interconexión para los experimentos, se determina de manera unívoca la cantidad de mensajes que los experimentos podrían generar por hora, de acuerdo con la Ec. (4)

$$m\# = (AF \times\ 0.1 \times\ 3600 \times\ 10^6)\ /\ Lm\ /\ 8 \tag{4}$$

donde *AF* es el ancho de banda físico de la red de interconexión. Además, los experimentos ping-pong se pueden distribuir a lo largo de cada hora para evitar una "congestión temporaria" de la red de interconexión que puede perturbar a los demás usuarios-aplicaciones (recordar que cada ping-pong implica dos mensajes).

4. Resultados de Evaluación de una Red de Interconexión

En esta sección se muestran los resultados obtenidos con una red de interconexión libre y con los cuales se obtuvo también el valor *Lm* mencionado previamente. El detalle de las máquinas utilizadas para la ejecución de los experimentos se muestra en la Tabla 1, y la biblioteca de pasaje de mensajes utilizada fue PVM.

	Descripción Hard	***Sistema Operativo***	***Reloj***	***Memoria***	***Red***
1)	PC-PIII	Linux 2.2.12-20	550 MHz	64 MB	Ethernet 10/100 Mb/s
2)	SGI Indigo 2	IRIX 6.2	175 MHz	512 MB	Ethernet 10 Mb/s
3)	PC-Pentium	Linux 2.2.12-20	233 MHz	64 MB	Ethernet 10 Mb/s
4)	Sun SPARC- Station 2	Solaris 2.5.1	40 MHz	48 MB	Ethernet 10 Mb/s
5)	Sun SPARC- Station 4	Solaris 2.5.1	110 MHz	32 MB	Ethernet 10 Mb/s

Tabla 1: Máquinas Utilizadas en los Experimentos.

El cableado de la red de interconexión incluye múltiples segmentos de red Ethernet de 10 Mb/s interconectados con hubs y switches, y estas cinco máquinas son una porción relativamente pequeña de la cantidad total de máquinas interconectadas. Todos los experimentos se llevaron a cabo desde la primera máquina, que es la que tiene la mejor placa de interfase de red y se utilizaron mensajes de longitudes 10^2, 10^3, 10^4, 10^5, 10^6 y 10^7 bytes (no hay problemas de congestión de la red dado que la red se utiliza básicamente libre en estos experimentos). Los resultados de la experimentación en

cuanto a tiempo de comunicaciones (en segundos) desde la primera máquinas hacia cada una de las demás se muestran en la Tabla 2.

n	2)	3)	4)	5)
10^2	0.01	0.01	0.11	0.20
10^3	0.01	0.01	0.16	0.21
10^4	0.04	0.05	0.22	0.26
10^5	0.23	0.47	0.47	0.67
10^6	2.63	6.24	3.22	10.72
10^7	51.40	64.25	78.70	56.37

Tabla 2: Tiempos de Comunicación con PVM.

Se puede notar en la Tabla 2 que el tiempo de comunicaciones para 100 bytes constituye una buena aproximación del tiempo de startup de comunicaciones dado que es casi idéntico al tiempo de comunicaciones para 1000 bytes para todas las máquinas. Con respecto a este tiempo de startup se tienen dos consideraciones muy importantes:

1. Es muy alto con respecto a 0.5 milisegundos que se toma como referencia para las redes Ethernet de 10 Mb/s [11]. Queda claro que sin estos experimentos, el tiempo de latencia sería una incógnita dado que PVM establece una alta sobrecarga en cuanto a latencia.
2. Es dependiente de las máquinas que intervienen en las transferencias de datos, trasladando de alguna manera la heterogeneidad en cuanto a capacidad de cálculo a las comunicaciones (al menos en cuanto a tiempo de startup).

Para la estimación de ancho de banda asintótico conviene expresar los valores de la Tabla 2 en función de MB/s, tal como lo muestra la Tabla 3.

n	2)	3)	4)	5)
10^2	0.031	0.015	0.002	0.001
10^3	0.186	0.128	0.012	0.009
10^4	0.542	0.422	0.088	0.074
10^5	0.812	0.404	0.402	0.284
10^6	0.726	0.306	0.593	0.178
10^7	0.371	0.297	0.242	0.338

Tabla 2: MB/s de Comunicación con PVM.

En este caso, se tienen algunos datos interesantes, tales como:

- La mejor tasa de transferencia de datos suele darse para 10^6 bytes, no para 10^7 que es la máxima longitud de las utilizadas. Se deben al menos una consideración al respecto, dado es muy difícil que en una aplicación paralela se tengan mensajes de 10^7 bytes, con lo cual esta longitud no parece muy representativa en cuanto a rendimiento de las comunicaciones que las aplicaciones.
- Se podría tomar 10^6 como la longitud mínima, *Lm*, con la cual se obtiene una estimación del ancho de banda apropiada dado que las diferencias con los mensajes de 10^5 no es muy significativa (al menos no tan significativa como entre las demás longitudes de mensajes).
- Se confirma que la heterogeneidad de las máquinas se traduce casi directamente a rendimiento heterogéneo de las comunicaciones aunque todas las máquinas están comunicándose a 10 Mb/s y

esto no debería suceder.

Si se tiene en cuenta ahora que : $Lm = 10^6$, AF=10 Mb/s (Ethernet 10 Mb/s) y se decide utilizar al 10% del ancho de banda físico de la red de interconexión para la experimentación con la red *parcialmente* utilizada, al utilizar la Ec. (4) para conocer la cantidad de mensajes que se deberían generar se tiene

$$m\# = (10 \times 0.1 \times 3600 \times 10^6) / 10^6 / 8 = 3600/8 = 450$$

lo cual implica 225 ping-pong por hora. Como se aclaró antes, dado que 450 mensajes pueden significar una "carga temporaria" sobre la red muy grande si se llevan a cabo sin intervalos intermedios, conviene distribuirlos uniformemente a lo largo de una hora, con lo que se tendrían los intervalos de tiempo entre los cuales se ejecutan los experimentos individuales.

5. Conclusiones y Trabajo Futuro

Se presenta en este artículo una forma clara y definida de evaluación del rendimiento de la red de comunicaciones para las aplicaciones paralelas tanto para cuando la red de comunicaciones está totalmente disponible como para cuando la red de comunicaciones está siendo compartida por otros usuarios-aplicaciones de la red local. En este sentido, se trata de establecer no solamente la sobrecarga impuesta por la biblioteca de pasaje de mensajes elegida para la ejecución de los programas paralelos sino también una forma de estimación de la capacidad disponible de la red de interconexión en el tiempo dependiendo de otros factores, externos a las propias aplicaciones paralelas que se utilizan para aprovechar la red local.

Se presenta en este artículo la evaluación de una red local en particular, donde se muestra que al menos al utilizar PVM, los índices de rendimiento de las comunicaciones no solamente son mucho mayores que los esperables en términos del hardware o incluso de protocolos de bajo nivel como TCP/IP sino que además la heterogeneidad del hardware de cómputo se traduce casi de manera directa a diferentes índices de rendimiento dependiendo de las computadoras entre las que se transfieren los datos.

De acuerdo a los resultados obtenidos en la experimentación con la red libre se establece la cantidad de mensajes que se podrían generar dependiendo del porcentaje del ancho de banda físico que se puede utilizar para la evaluación de la carga (o *capacidad libre*) de la red de interconexión. También se presenta una forma sencilla de evitar "congestiones temporales" de la red por la ejecución de los experimentos con la coexistencia de otros usuarios-aplicaciones en la red local.

Este artículo presenta datos preliminares en el sentido que se deberían realizar una serie extensiva de experimentos para llegar a la estimación de la carga de la red local durante, por ejemplo, todos los días de un mes (al menos todos los días laborales). El tiempo y las condiciones cambiantes de usuarios-aplicaciones suelen presentar situaciones muy complejas de resolver, pero hasta el momento no se ha presentado otra metodología o forma de estimación de la capacidad de una red local para ser aprovechada por aplicaciones paralelas/distribuidas. Por lo tanto, la primera extensión a este artículo debería ser la presentación e interpretación de los resultados de experimentación a mediano-largo plazo con la metodología presentada en este artículo.

Una de las tareas que también deben resolverse a partir del presente artículo consiste en el establecimiento de alguna forma de caracterización y/o aprovechamiento de la información que se pretende obtener para ser utilizada por las aplicaciones paralelas. En principio, por ejemplo, si se llega a que se tienen disponibles solamente 1 Mb/s (en una red de 10 Mb/s, por ejemplo) durante una determinada cantidad de horas del día, deber ser claro que las aplicaciones paralelas que pueden ser ejecutadas con rendimiento aceptable son justamente aquellas que necesitan 1 Mb/s o menos en cuanto a la transferencia de datos entre sus procesos.

Agradecimientos

Al Director del Departamento de Coordinación de Investigación y Desarrollo, Ing. Víctor Padula Pintos por su constante apoyo hacia nuestra labor.

Bibliografía

[1] Anderson T., D. Culler, D. Patterson, and the NOW Team, "A Case for Networks of Workstations: NOW", IEEE Micro, Feb. 1995.
[2] Basney J., M. Livny, "Deploying a High Throughput Computing Cluster", in R. Buyya Ed., High Performance Cluster Computing: Architectures and Systems, Vol. 1, Prentice-Hall, Upper Saddle River, NJ, USA, pp. 116-134, 1999.
[3] 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.
[4] Foster I., Designing and Building Parallel Programs, Addison-Wesley, Inc., 1995. *Versión html* disponible en http://www-unix.mcs.anl.gov/dbpp
[5] Hockney R., M. Berry (eds.), "Public International Benchmarks for Parallel Computers", Scientific Programming 3(2), pp. 101-146, 1994.
[6] Hockney R., C. Jesshope, Parallel Computers 2, Adam Hilger, Bristol and Philadelphia, IOP Publishing Ltd., 1988.
[7] Institute of Electrical and Electronics Engineers, Local Area Network - CSMA/CD Access Method and Physical Layer Specifications ANSI/IEEE 802.3 - IEEE Computer Society, 1985.
[8] Litzkov M., "Remote UNIX - Turning Idel Workstations into Cycle Servers", Proc. Of the 1987 Usenix Summer Conference, Phoenix, Arizona, 1987.
[9] Litzkov M., M. Livny, "Experience with the CONDOR Distributed Batch System", IEEE Workshop on Experimental Distributed Systems, Huntsville, Alabama, Oct. 1990.
[10] Nagendra B., Rzymianowicz, "High Speed Networks", in R. Buyya Ed., High Performance Cluster Computing: Architectures and Systems, Vol. 1, Prentice-Hall, Upper Saddle River, NJ, USA, pp. 204-245, 1999.
[11] Pacheco P., Parallel Programming with MPI, Morgan Kaufmann, San Francisco, California, 1997.
[12] Wilkinson B., Allen M., Parallel Programming: Techniques and Applications Using Networking Workstations, Prentice-Hall, Inc., 1999.
[13] LAM/MPI (Local Area Computing / Message Passing Interface) Home Page
http://www.mpi.nd.edu/lam
[14] MPICH Home Page http://www-unix.mcs.anl.gov/mpi/mpich/