Sincronización de Relojes Orientada a Evaluación de Rendimiento en Clusters

Fernando L. Romero, Fernando G. Tinetti[1]
Instituto de Investigación en Informática LIDI (III-LIDI)
Facultad de Informática – UNLP
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, CICPBA, Agencia Nacional de Promoción Científica y Tecnológica e IBM.

RESUMEN

En el presente trabajo se exponen avances en las líneas de investigación sobre la sincronización de relojes en ambientes distribuidos, orientada a implementar herramientas que permitan realizar pruebas de rendimiento a través de la instrumentación de código, inicialmente en ambientes de clusters, para ser luego extendida a otros ambientes distribuidos. Se han desarrollado herramientas específicas para los casos en que las existentes no satisfacen los requerimientos.

Keywords: *Sincronización de Procesos, Relojes Distribuidos, Rendimiento e Instrumentación, Sistemas Paralelos y Distribuidos, Paralelismo en Clusters e Intercluster, Sincronización Interna y Externa .*

1. INTRODUCCION

En la actualidad, tanto los sistemas de cómputo como los de comunicación con los que interactúan, disponen de hardware orientado a la medición del tiempo. Dichos dispositivos de hardware en algunos casos se utilizan para disponer de una referencia horaria en alguna escala de tiempo. Para estos casos, se hace necesaria la sincronización de esta referencia. Los requerimientos de exactitud con que se realizan esta sincronización y las mediciones han ido creciendo desde el segundo, milisegundo [6][7], hasta llegar a exigir microsegundos [10][9] [8].

Dicha exactitud es necesaria en el caso de la presente línea de investigación a fin de lograr cotas de errores aceptables que hagan posible estimaciones de rendimiento en ambientes de red local de aplicaciones paralelas. Se debe tener en cuenta que en una red local Ethernet los tiempos de comunicación son del orden de decenas de microsegundos, por lo que determinaciones de tiempo transcurrido entre la partida y el arribo de un mensaje requieren sincronizaciones con diferencias de microsegundos en las máquinas que intercambian mensajes, para poder contar con una medida con un error aceptable.

Por otro lado, en los últimos tiempos ha aparecido la necesidad de determinar la posición de equipos móviles con respecto a otros nodos fijos con los que tienen comunicación, sin necesidad de utilizar GPS (geoposicionamiento satelital) y para ello también se requiere una resolución del orden de microsegundos en la sincronización entre dichos nodos [10].

---

[1] Investigador Asistente, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

En todas las circunstancias es deseable que el registro de tiempos implique la menor carga posible de procesamiento, como también conocer la magnitud del error que se comete en la medición. En el caso de mediciones de rendimiento, sería deseable que la tarea de sincronización se lleve a cabo fuera del tiempo en que se ejecuta el programa que se está monitorizando. Por otro lado, la cantidad de máquinas a sincronizar puede llevar a que el tiempo que se insume en la sincronización sea excesivo. Debido a esto, se ensaya la posibilidad de sincronización mediante mensajes de *broadcast*, para dar solución a los problemas de escalabilidad.

Para el presente proyecto, la sincronización se lleva a cabo sin incluir hardware adicional al del sistema. Las comunicaciones involucradas en la sincronización deberán utilizar la red de interconexión entre computadoras que ya existe, y las mediciones de tiempo deben utilizar lo que provee cada sistema de cómputo para ese fin.

El objetivo final es contar con una herramienta de instrumentación para programas paralelos que:

- Pueda ser usada inicialmente en un cluster de PC's, con la posibilidad de ser extendido a clusters en general y luego en plataformas distribuidas aún más generales.
- Sea de alta resolución, es decir que se pueda utilizar para medir tiempos cortos, del orden de microsegundos.
- No altere el funcionamiento de la aplicación bajo prueba, o que la alteración sea mínima y conocida por la aplicación.
- Utilice en forma predecible la red de interconexión. Más específicamente, se puedan determinar, desde la aplicación, los intervalos de tiempo en los cuales se utilizará la red. De esta forma, se puede *desacoplar* el uso de la red de interconexión, ya que habrá intervalos de tiempo usados para la sincronización e intervalos de tiempo utilizados para la ejecución de programas paralelos.

2. LINEAS DE INVESTIGACION Y DESARROLLO

La posibilidad de realizar evaluación de rendimiento sin introducir hardware especial exige como requisito previo que cada computadora cuente con un oscilador físico de frecuencia relativamente constante. A partir de este oscilador físico se derivan los relojes lógicos que son los que se sincronizan [4]. De las limitaciones de estos relojes dependerá lo que se pueda lograr, por lo que se realizan experimentos a fin de caracterizar estos relojes en cuanto a exactitud, estabilidad, confiabilidad, resolución y demás atributos.

Se trata de determinar una referencia fija en el tiempo a partir de la cual se contabiliza el tiempo en cada computadora. Dicha referencia se debe comunicar a través de la red de interconexión. Debido a la varianza en los tiempos de comunicaciones, es difícil determinar este valor más allá de cierta exactitud. Para la medición de este valor se puede recurrir a la estadística de los valores de ida-vuelta (*round trip time*) de un mensaje y asumir que solo serán validos los valores de referencia trasmitidos en un tiempo de transmisión igual a la moda de los tiempos de transmisión. Ello lleva a un necesario intercambio de mensajes entre una máquina *servidora* que proporciona la referencia y los *clientes* que la reciben, intercambio que complica la escalabilidad del modelo. Dicha complicación tiene dos razones:

1) El modelo cliente/servidor implica centralizar tareas en el servidor, lo que al aumentar la cantidad de clientes, el servidor se transforma en un cuello de botella.
2) Las confirmaciones en los intercambios de mensajes con referencias de tiempo y la consiguiente bidireccionalidad de los mismos produce una saturación de mensajes en la red.

Por ello se realizan estudios para lograr la comunicación de las referencias a través de mensajes *broadcast*.

Otro aspecto a tener en cuenta son las diferentes frecuencias de los relojes de las computadoras que se sincronizan. Suponiendo que dichas frecuencias difieren proporcionalmente en una constante, para el cálculo de dicha constante se realizan mediciones sobre la evolución de los ciclos en cada reloj de cada computadora para un periodo determinado por el servidor que fija la referencia inicial. Dicho periodo debe ser los más grande posible para poder apreciar con la necesaria precisión la constante a calcular. A partir del momento de determinar esta constante, por tratarse de relojes de cuarzo cuya frecuencia de oscilación puede ser alterada por cambios ambientales tales como la temperatura, se debe establecer qué estabilidad a largo plazo tiene la frecuencia. Para hallar esta estabilidad en forma absoluta debiera disponerse de un reloj perfecto o lo más exacto posible. Pero como lo que interesa en estimaciones de rendimiento es la estabilidad en la relación entre los relojes, se realizaron experimentos de estabilidad a corto, mediano y largo plazo sobre dicha estabilidad.

La finalidad de las mediciones de rendimiento es la de lograr optimizar código de aplicaciones paralelas. Estas aplicaciones deben intercambiar datos, lo cual significa que en el caso de la recepción, se bloquea el proceso hasta que el encargado de trasmitir el mensaje no lo envía. Para saber en qué lugar del código se encuentra una máquina respecto de otra, se debe contar con una referencia única de tiempo en todas las máquinas, o sea relojes sincronizados.

3. RESULTADOS OBTENIDOS/ESPERADOS

A partir del estudio de los algoritmos básicos e implementaciones existentes [1] [2] [6], se realizaron un conjunto de experimentos a fin de poder compararlas. Como conclusión de dichos experimentos se desarrolló una biblioteca *timings* para mediciones de tiempo a nivel local y una biblioteca *synchro*, que funciona en forma distribuida y permite sincronizar en hora y frecuencia los relojes de *timings*. Estas bibliotecas además proveen una caracterización del error, dada por sus diversos componentes:

- Error en la estimación del tiempo de comunicación entre máquinas de las referencias.
- Error debido a la precisión del reloj local de la biblioteca desarrollada (*timings*).
- Error debido a latencia propia del sistema operativo.

Como extensiones futuras, siempre es deseable la sincronización externa de los relojes [3]. Este paso está muy ligado también a la posibilidad de utilizar más de un cluster de computadoras para cómputo paralelo y en este contexto la sincronización de los relojes va más allá del análisis de rendimiento con el objetivo de optimizarlo.

Por otro lado, se realizaron experimentos tendientes a estimar los problemas de escalabilidad del algoritmo implementado en la biblioteca *synchro*. Si bien se sabe que por el modelo cliente/servidor que utiliza la biblioteca de sincronización desarrollada (*synchro*) se presentará un problema de escalabilidad, se trata de determinar cuantitativamente el grado de incidencia del aumento de la cantidad de clientes a sincronizar. En estos experimentos se mide:

- El tiempo para sincronizar de 2 a 16 máquinas entre sí, mostrando un mínimo de 50 milisegundos y un máximo de 176.
- Se midió el tiempo de inicio del sistema, lo cual implica la creación de interfaces de comunicaciones y mediciones a nivel local y de la red para la puesta en marcha de la biblioteca de referencias de tiempo a nivel local (*timings*). Estos tiempos van de un mínimo de 1290

milisegundos para 2 máquinas hasta 6558 milisegundos para 16 máquinas.

También se realizaron pruebas tendientes a verificar la estabilidad de frecuencia de los relojes de cuarzo en el tiempo similares a las reportadas en [5]. El experimento requeriría de una referencia de tiempo perfecta o al menos con la menor deriva de frecuencia posible, tal como un reloj atómico o la hora UTC. Debido a que no se cuenta con estos elementos, se toma como referencia el reloj de cuarzo del *timer* de una PC, y con el se miden las variaciones del reloj TSC (*Time Stamp Counter*). Si bien no se puede medir el error en términos absolutos, se puede determinar cuánto varía la diferencia de deriva. Dado que en la línea de investigación que se sigue no se requiere de referencias externas, sino solo que las referencias estén ajustadas a la escala de tiempos que determine el servidor, es válido medir esta variación de frecuencia y pensar que el error será proporcional a dicha diferencia. Los resultados obtenidos son similares a los de [5], de menos de 10 ppm (partes por millón) en un período de varias horas.

Para la definición de futuras extensiones, se tienen planificados algunos experimentos para estimar:

- La mejora proporcionada por la comunicación de la referencia de tiempo utilizando comunicación *broadcast*.
- Más exhaustivamente la estabilidad de los relojes de las máquinas y su deriva, a fin de determinar la relación del intervalo entre sincronizaciones con el error de sincronización esperado.
- La utilidad de un servidor de hora con sistema operativo de tiempo real para disminuir la latencia impuesta por el SO al menos del lado del servidor.

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] G. Coulouris, Dollimore J., Kindberg T., Sistemas Distribuidos. Conceptos y Diseño, 3ª edición. Pearson Educación, 2001, ISBN: 8478290494.

[2] F. Cristian. "Probabilistic Clock Synchronization", *Distributed Computing*, 3: 146–158, 1989.

[3] Fetzer C., Christian F., "Integrating External and Internal Clock Synchronization", June 1996.

[4] K. H. Kim, Im C., Athreya P, "Realization of a Distributed OS Component for Internal Clock Synchronization in a LAN Environment", Proc. ISORC 2002, IEEE 5th Int'l Symp on Objetoriented Realtime distributed Computing, Washington, D.C., April 2002, pp. 263270.

[5] X. Luo, “TSC-I2: A Lightweight Implementation for Precision-Augmented Timekeeping”, reporte técnico de University of Illinois at Chicago, http://tsc-xluo.sourceforge. net/TSC-I2_Tech_Report.pdf

[6] D. L. Mills, "Measured performance of the Network Time Protocol in the Internet System". ACM Computer Communication Review 20, Jan. 1990. pp. 6575.

[7] D. L.Mills, "A Brief History of NTP Time: Confessions of an Internet Timekeeper". ACM Computer Communications Review 33, 2 (April 2003), pp 922.

[8] D. L. Mills, Kamp P.H., "The Nanokernel", Proc. Precision Time and Time Interval (PTTI) Applications and Planning Meeting (Reston VA, November 2000).

[9] F. L. Romero, W. Aróztegui, F. G. Tinetti, "Sincronización de Relojes en Ambientes Distribuidos", XII Congreso Argentino de Ciencias de la Computación (XII CACIC) Octubre 2006.

[10] A. Vallat, D. Schneuwly, ”Clock Synchronization in Telecommunications via PTP (IEEE 1588)” , Frequency Control Symposium, 2007 Joint with the 21st European Frequency and Time Forum. IEEE International. May 29 2007-June 1 2007.pp. 334-341. ISSN: 1075-6787. ISBN: 978-1-4244-0647-0