

Procesamiento Paralelo y Distribuido en tratamiento masivo de datos

**Armando De Giusti¹, Marcelo Naiouf², Cecilia Sanz³,
Fernando Tinetti⁴, Laura De Giusti⁵, Rodolfo Bertone⁶**

*LIDI - Laboratorio de Investigación y Desarrollo en Informática⁷.
Facultad de Informática. Universidad Nacional de La Plata.*

Palabras Clave

Procesamiento Paralelo y Distribuido. Programas Paralelos. Análisis de performance. Balance de carga. Escalabilidad. Métricas del paralelismo. Aplicaciones de tratamiento masivo de datos.

Resumen

Investigar el procesamiento paralelo y distribuido aplicado a tratamiento masivo de datos.

Los temas fundamentales propuestos en el proyecto se refieren a la especificación, transformación, optimización y verificación de algoritmos concurrentes ejecutables en sistemas paralelos/distribuidos, la optimización de clases de soluciones en función de modelos de arquitectura multiprocesador y las métricas de complejidad y eficiencia relacionadas con el procesamiento paralelo.

Se trabaja experimentalmente son cuatro modelos de arquitectura multiprocesador, disponibles/accesibles en el LIDI.

Interesa especialmente la aplicación de estas investigaciones al tratamiento masivo de datos de imágenes (genéricas, médicas y de cultivos), de datos numéricos (cómputo científico) y de bases de datos distribuidas.

Introducción

El procesamiento ha evolucionado hacia el paralelismo prácticamente desde el inicio mismo de las computadoras digitales. En la actualidad resulta innegable la importancia y el creciente interés en el procesamiento distribuido y paralelo dentro del espectro de la Ciencia de la Computación.

Existen diversas razones para esta realidad, entre las que se pueden citar:

- El crecimiento de la potencia de cómputo, dado en la evolución de la tecnología de los componentes y en las arquitecturas de procesamiento (supercomputadoras, hipercubos de procesadores homogéneos, grandes redes de procesadores no-homogéneos, procesadores de imágenes, de audio, etc.)
- La transformación y creación de algoritmos que explotan al máximo la concurrencia implícita en el problema a resolver, de modo de distribuir el procesamiento minimizando el tiempo total de respuesta. Naturalmente esta transformación también debe adaptarse a la arquitectura física de soporte.
- La capacidad del cómputo distribuido/paralelo de reducir el tiempo de procesamiento en problemas de cálculo intensivo o de grandes volúmenes de datos.
- La necesidad de tratar con sistemas de tiempo real distribuidos, donde los requerimientos en relación al tiempo de respuesta son críticos.

¹ Investigador Principal CONICET. Profesor Titular Ded. Exclusiva. degusti@lidi.info.unlp.edu.ar

² Profesor Titular. mnaiouf@lidi.info.unlp.edu.ar

³ Profesor Adjunto. csanz@lidi.info.unlp.edu.ar

⁴ Profesor Adjunto. fernando@lidi.info.unlp.edu.ar

⁵ Jefe de Trabajos Prácticos. ldgiusti@lidi.info.unlp.edu.ar

⁶ Profesor Adjunto. pbertone@lidi.info.unlp.edu.ar

⁷ LIDI - Facultad de Informática. UNLP - Calle 50 y 115 1er Piso, (1900) La Plata, Argentina.
TE/Fax +(54) (221) 422-7707. <http://lidi.info.unlp.edu.ar>

- El límite físico alcanzado por las computadoras secuenciales, que en algunos casos torna inaceptable el tiempo para resolver determinados problemas, y hace que la solución paralela sea la única factible.
- La existencia de sistemas en los que no es tan importante la velocidad de cómputo sino la necesidad de resolver problemas en más de una ubicación física a la vez, capacidad que puede asociarse rápidamente a una configuración paralela
- Las posibilidades que el paradigma paralelo ofrece en términos de investigación de técnicas para el análisis, diseño y evaluación de algoritmos.

Es fundamental referirse a un algoritmo paralelo mencionando el modelo de computación paralela para el que se lo diseñó. En el cómputo secuencial, el modelo RAM ha sido prácticamente aceptado como estándar. En el procesamiento distribuido/paralelo no existe un modelo teórico unificador. Esto se debe a que los diferentes modelos paralelos (PRAM para memoria compartida, BSP para memoria distribuida, LogP para redes de procesadores asincrónicos, etc), enfatizan determinados aspectos a costa de otros. Una razón más sutil es que no es probable una máquina paralela universal; para cada aplicación existe una máquina óptima. Por otro lado, existen distintas alternativas de implementación sobre sistemas de cómputo paralelo con hardware de procesamiento homogéneo o heterogéneo y con acoplamiento fuerte o débil de sus componentes.

Por esto, lo correcto es hacer referencia a *sistemas paralelos* como la combinación de algoritmo y arquitectura. La performance obtenida en un sistema paralelo está dada por una compleja relación en la que intervienen una gran cantidad de factores (el tamaño del problema, la arquitectura de soporte, la distribución de procesos en procesadores, la existencia o no de un algoritmo de balanceo de carga, etc).

Por otra parte, existe un gran número de métricas para evaluar sistemas paralelos (tiempo efectivo, speed-up, eficiencia, producto procesador-tiempo, escalabilidad, isoeficiencia).

En este marco, interesa evaluar la performance de distintas clases de aplicaciones sobre diferentes arquitecturas reales para obtener resultados concretos. El énfasis está en adecuar los resultados teóricos ideales en las métricas a la realidad de la implementación de determinada clase de problema sobre distintas plataformas. Este interés se basa en que muchos sistemas paralelos no alcanzan su capacidad teórica, y las causas de esta degradación son muchas y no siempre fáciles de determinar.

El análisis sobre plataformas disponibles permite estudiar el impacto que tienen algunos de estos factores sobre las implementaciones, y adecuar las métricas clásicas a las mismas. En particular puede estudiarse la influencia de las estrategias de distribución de procesos y datos, y la carga (estática o dinámica) asignada a cada procesador sobre el speedup, la eficiencia y la escalabilidad. Más allá de las características teóricas/algorítmicas de las aplicaciones, el rendimiento real y por lo tanto el tiempo necesario para resolver los problemas siempre o en la mayoría de los casos dependerá de la arquitectura de procesamiento elegida para la implementación.

Muchas disciplinas entre las que se encuentran el reconocimiento de patrones en tiempo real, el tratamiento y transmisión de video en tiempo real y la visión por computadora, requieren un importante esfuerzo en la investigación de algoritmos paralelos aplicables en áreas tales como robótica, industria manufacturera, ingeniería forestal y medicina. Interesa investigar la optimización de algoritmos paralelos para el reconocimiento de patrones de similitud, enfocando particularmente aplicaciones orientadas a grandes bases de datos de imágenes. En particular interesa el problema de *análisis de similitud de imágenes* basado en el estudio de la transformada Haar-Wavelet multidimensional que permite obtener una “firma digital” (codificación) de la imagen en estudio (que sea invariante a las traslaciones, rotaciones y escalado) y buscar otras imágenes que tengan una codificación similar dentro de una Base de Datos. La complejidad del procesamiento con esta técnica requiere dividir la imagen en regiones

y a su vez procesar la transformada wavelet dentro de cada región mediante una ventana deslizante de tamaño variable, lo que obliga a estudiar la implementación de algoritmos paralelos que permitan tener tiempos de respuesta razonables, más aún si el análisis de similitud se requiere que sea realizado en tiempo real. El objetivo es desarrollar una arquitectura de software paralela que permita hacer análisis de similitud de imágenes (por ejemplo fotográficas o de video) en tiempo real.

Otras áreas de interés incluyen el análisis de imágenes hiperspectrales de cultivos, los cálculos sobre matrices, y los problemas de Reconocimiento de patrones complejos.

Temas de Investigación y desarrollo

- Optimización del empleo de recursos de procesamiento.
 - ✓ Balance de carga, tanto en procesamiento como en comunicaciones.
 - ✓ Distribución óptima de procesos y procesadores.
 - ✓ Migración de datos y/o procesos en forma dinámica.
 - ✓ Ajuste dinámico de código paralelo.
- Paralelización de aplicaciones.
 - ✓ Transformación de algoritmos.
 - ✓ Adaptación de algoritmos a modelos de arquitectura.
 - ✓ Especificación y verificación de sistemas paralelos de tiempo real
- Métricas de paralelismo.
 - ✓ Análisis de complejidad de los algoritmos secuenciales y paralelos.
 - ✓ Análisis de rendimiento de las soluciones paralelas.
 - ✓ Caracterización de paradigmas de procesamiento paralelo y modelos de performance asociados.
- Arquitecturas de procesamiento paralelo.
 - ✓ Rendimiento en función de la capacidad potencial de procesamiento de las arquitecturas de cómputo paralelo (tal como supercomputadoras paralelas y redes o clusters de estaciones de trabajo)
 - ✓ Reconfiguración dinámica del hardware y/o del software para adaptarse a la aplicación y/o proponer tolerancia a fallas
 - ✓ Diseño de hardware paralelo orientado a clases de aplicaciones.
- Aplicaciones específicas
 - ✓ Análisis de similitud de imágenes
 - ✓ Análisis de imágenes hiperspectrales a áreas agrícolas
 - ✓ Aplicaciones médicas de Redes neuronales paralelas
 - ✓ Computo científico sobre arquitecturas distribuidas
 - ✓ Análisis de rendimiento y replicación en Bases de Datos distribuidas

Equipamiento de experimentación

Los modelos de arquitectura experimental propuestos (y disponibles para el proyecto) son cuatro: arquitecturas multiprocesador homogénea fuertemente acopladas tipo cubo de transputer, arquitecturas multiprocesador distribuida (homogénea y heterogénea) débilmente acoplada tipo NOW, multiprocesadores dedicados a tratamiento de imágenes basados en DSPs y arquitecturas multiprocesador de memoria compartida distribuida tipo SGI Origin 2000 (Clementina).

Bibliografía Básica

- ACM Siggraph (Conferencias realizadas anualmente desde 1995)
- Akl S, "Parallel Computation. Models and Methods", Prentice-Hall, Inc., 1997.
- Baxes G.. "Digital Image Processing". Wiley, 1994
- Bell, David; Grimson, Jane, "Distributed Database Systems", Addison Wesley, 1992
- Brinch Hansen, P., "Studies in computational science: Parallel Programming Paradigms", Prentice-Hall, Inc., 1995.
- Date, C.J., "Introducción a los sistemas de Bases de Datos". Addison Wesley 1994.
- Glassner A., "Principles of Digital Image Synthesis", Morgan Kaufmann, 1995.
- Gonzalez and Woods, "Digital Image Processing", Addison-Wesley, 1992
- Gupta A., Kumar V., "Performance properties of large scale parallel systems", Journal of Parallel and Distributed Computing, November 1993.
- Hwang K., "Advanced Computer Architecture: Parallelism, Scalability, Programability", McGraw, 1993.
- IEEE Transactions on Image Processing (colección de revistas 1990-2002)
- IEEE Transactions on Parallel and Distributed Processing (colección de revistas 1990-2002)
- Jain Anil, "Fundamentals of Digital Image Processing", Prentice Hall, 1989
- Jensen. "Introductory Digital Image Processing. A remote sensing perspective" 2da edición. Prentice Hall. 1996
- Kohonen T., "Self-Organizing Maps", Springer-Verlag, 1997
- Kosko B., "Neural Networks and Fuzzy Systems", Prentice Hall, 1992
- Kumar V., Grama A., Gupta A., Karypis G., "Introduction to Parallel Computing. Design and Analysis of Algorithms", Benjamin/Cummings, 1994.
- Larson, J., "Database Directions. From relational to distributed, multimedia, and OO database Systems". Prentice Hall. 1995
- Lawson H., "Parallel processing in industrial real time applications", Prentice Hall 1992.
- Leighton F. T., "Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes", Morgan Kaufmann Publishers, 1992.
- Maravall Gomez-Allende, "Reconocimiento de Formas y Visión Artificial", Addison-Wesley Iberoamericana, 1994.
- Miller R., Stout Q. F., "Algorithmic Techniques for Networks of Processors", CRC Handbook of Algorithms and Theory of Computation, M. J. Atallah, ed, 1998.
- Morse F., "Practical Parallel Computing", AP Professional, 1994.
- Peddle. "An Empirical comparison of evidential reasoning, linear discriminant analysis, and maximum likelihood algorithms for alpine land cover classification". Canadian Journal of Remote Sensing. Vol.19. Nro.1- 1993
- Peddle. "Knowledge Formulation for Supervised Evidential Classification". Photogrammetric Engineering & Remote Sensing. Vol. 61- Nro. 4 – 1995
- Silbershatz, F., "Fundamentos de las Bases de Datos".. Mc Graw Hill. 1998.
- Sima D, Fountain T, Kacsuk P, "Advanced Computer Architectures. A Design Space Approach", Addison Wesley Longman Limited, 1997.
- Thomas M Lillesand, Ralph W Kiefer. "Remote Sensing and Image Interpretation" 3rd Edition. John Wiley. 1994
- Tinetti F., De Giusti A., "Procesamiento Paralelo. Conceptos de Arquitectura y Algoritmos", Editorial Exacta, 1998.
- Zomaya A., "Parallel Computing. Paradigms and Applications", Int. Thomson Computer Press, 1996.