Revista Digital Universitaria
10 de marzo 2010 • Volumen 11 Número 3 • ISSN: 1067-6079

ARTICULO

EL SUPERCÓMPUTO EN 2010

René Luna, Luis Villa, Marco Ramírez
Centro de Investigación en Computación del Instituto Politécnico Nacional

10 de marzo 2010 • Volumen 11 Número 3 • ISSN: 1067-6079

Resumen

Las aplicaciones que dominan el mercado del software demandan mayor capacidad de cómputo de forma permanente en prácticamente todos los campos del conocimiento: multimedia, procesamiento digital de señales, 3D, aplicaciones dedicadas, reconocimiento de patrones, astrofísica, simulación. Esta relación que existe entre el desempeño de los procesadores y los requerimientos de las aplicaciones, ha sido un patrón que se ha mantenido desde que, en 1954, hizo su entrada estelar la primera computadora que se fabricó de forma masiva y que tenía capacidad para realizar operaciones de punto flotante. Estamos hablando de la IBM 704, misma que podía ejecutar 40,000 instrucciones por segundo. Cuarenta y seis años después, en el 2010, prácticamente cualquier computadora portátil (o laptop) puede realizar millones de operaciones por segundo. Esta mejora en el desempeño se ha debido en gran parte al desarrollo de la tecnología del silicio, que ha permitido la miniaturización de los transistores, incrementando así el número de transistores que pueden integrarse en el diseño de los nuevos procesadores (un mayor número de transistores significa una mayor funcionalidad y capacidad en la misma área).

Palabras clave: procesadores, desempeño, transistores, funcionalidad.

Introducción

Desde hace más de 20 años, los arquitectos de computadoras han utilizado el incremento del número de transistores y de la velocidad que éstos pueden alcanzar, para duplicar el rendimiento cada 18 meses, como fue vaticinado en 1965 por el co-fundador de Intel, Gordon Moore. Hoy en día, la tecnología ha seguido poniendo a disposición de los arquitectos, un número mayor de transistores, sin embargo, el paradigma de cómputo que se había venido utilizando ya no pudo transformar este recurso tecnológico en el rendimiento con la misma tendencia seguida desde 1965. Este fenómeno es también conocido como “Moore’s Gap”, ya que hay un vacío en cuanto al rendimiento esperado como resultado del avance tecnológico. Dos son las razones principales que provocaron este fenómeno: los modelos de programación siguen siendo fundamentalmente secuenciales, lo que hace cada vez más difícil encontrar paralelismo en las aplicaciones que se ejecutan en los nuevos procesadores. Con el objetivo de extraer el máximo paralelismo a nivel de instrucción, los diseñadores implementaron, dentro del procesador, sistemas de control muy complejos (segmentación, renombrado, ejecución fuera de orden, predicción de saltos, pre-búsquedas de instrucciones) sin embargo, el rendimiento obtenido estaba muy por debajo de lo esperado. Esta complejidad en cuanto a la arquitectura de los procesadores se enfrentó a un nuevo problema: la disipación de calor. Los niveles de calor alcanzados por los procesadores son tan altos que resulta inviable (desde el punto de vista económico y funcional) seguir manteniendo el mismo paradigma computacional que había logrado llegar al siglo XXI. De esta forma, la industria, prácticamente en su totalidad, decidió que su futuro estaba en el cómputo paralelo.

La industria vio que su única opción viable era reemplazar su modelo de uni-procesador complejo e ineficiente, por un modelo de multiprocesador sencillo y eficiente. Esta estrategia dio entrada a lo que hoy conocemos como procesadores multinúcleo. Por un lado, se resolvía el problema de la complejidad del uniprocesador y por el otro, se utilizaría el mayor número de transistores para incrementar el número de procesadores o núcleos cada 18 meses, mientras se siguiera sosteniendo lo vaticinado por Gordon Moore.

En una arquitectura multinúcleo, cada procesador contiene múltiples procesadores (o núcleos), donde cada uno de éstos cuenta con su unidad de proceso independiente. El estilo de computación para este tipo de procesadores es MIMD (Multiple Instruction Multiple Data), es decir, que se tiene la capacidad de ejecutar más de una instrucción de manera concurrente y que, cada una de estas instrucciones, será ejecutada utilizando múltiples secuencias de datos, llevándonos a esquemas de paralelismo de procesos o hilos.

Esta estrategia también impactó a los sistemas de Supercómputo (o supercomputadoras) considerados como las computadoras con mayor capacidad de cómputo y diseñadas para atender necesidades computacionales complejas, que requieren tiempos de cómputo muy grandes. A diferencia de las Supercomputadoras anteriores, este tipo de sistemas se integra con miles de procesadores económicos conectados en paralelo, de ahí que se les haya dado el nombre de supercomputadoras masivamente paralelas. Cerca del 20% de los sistemas incluidos en el TOP500 son integrados o construidos en arquitecturas tipo cluster ensambladas con componentes de bajo costo y comerciales (commodity components). De este 20% , prácticamente todos, utilizan procesadores de 32 bits de Intel y AMD, con plataformas Linux como Sistema Operativo.

Sin embargo, que el modelo de programación por excelencia siga siendo secuencial ahora nos enfrenta a un viejo y muy conocido problema: sí las habilidades de la mayoría de los programadores se basan en modelos secuenciales de programación, ¿cómo paralelizaremos eficientemente las aplicaciones para que utilicen estos sistemas masivamente paralelos con múltiples núcleos?

Programación paralela

La meta más agresiva del paralelismo de hoy en día es hacer que los programas sean eficientes, portables y escalables (que se adapten al incremento del número de núcleos que son integrados al sistema), pero sobre todo, que sea fácil la programación de los mismos como lo es actualmente la tarea de escribir programas para computadores secuenciales. Esto debe hacerse procurando que el esfuerzo de migrar aplicaciones a modelos paralelos sea mínimo. Para lograr esta meta, la investigación en desarrollo de software es fundamental.

La investigación en desarrollo de software en México continúa sin recibir la atención adecuada. La evolución del hardware tiene que ser acompañada por una investigación robusta en desarrollo de software. Sólo de esta forma darán frutos los efímeros esfuerzos realizados por instituciones como el IPN, UNAM, UAM e IMP, que han invertido varios de sus recursos en la adquisición de infraestructuras computacionales importantes en el contexto Mexicano (pero no lo suficiente como para estar en los primeros 100 del TOP500). De no darse este impulso, no se asegura la capacidad de migrar a modelos paralelos, aplicaciones claves que nos impactan día a día como son: la industria petrolera, el medio ambiente, la simulación y la predicción de tráfico, los análisis biológicos y la simulación molecular.

Si la programación de aplicaciones paralelas nos es productiva, el progreso se retardará y, por lo tanto, se reducirá el número de programas que puedan explotar los recursos computacionales de las nuevas arquitecturas multinúcleos. Es importante hacer notar que, por el momento, el éxito en esta área ya no está en el hardware, recurso casi inalcanzable hace 20 años, sino en nuestra capacidad de investigar y formar recursos humanos de alto nivel que puedan explotar para nuestro beneficio esta capacidad de cómputo.

Los cursos de programación que se imparten en las universidades mexicanas tienen que evolucionar para cubrir las nuevas expectativas, no sólo en el campo del supercómputo moderno, sino también en los nuevos nichos de mercado que se abrieron con los procesadores gráficos, utilizados fundamentalmente en aplicaciones multimedia (GPU, Cell). Los estudiantes necesitan aprender los fundamentos y las diferentes técnicas de la programación paralela, con la finalidad de que puedan explotar la tecnología computacional actual y la del futuro.

El impacto de la ley de Moore en los procesadores multinúcleos

Si cada núcleo que se incluye en un procesador multinúcleo utiliza el mismo número de transistores, entre mas núcleos se integren en un mismo procesador, el número de transistores se multiplicará en la misma

10 de marzo 2010 • Volumen 11 Número 3 • ISSN: 1067-6079

proporción. Utilizando el corolario de la Ley de Moore, podríamos predecir que el número de núcleos se duplicará cada 18 meses y dado que procesadores con dos y cuatro núcleos (*dual y quad cores) están actualmente disponibles en el mercado, querría decir que dentro de doce años se estarían ofreciendo procesadores con 1K (1024) núcleos por procesador.*

Debido al problema de consumo de potencia y por ende, a la disipación de calor mencionada en el inicio de este artículo, el número de transistores que representan 1K núcleos acentuaría este problema. Suponiendo que cada núcleo consuma una potencia de 5 watts (muy por debajo de la realidad actual), 1K núcleos representaría un consumo global de 5K watts. Para darnos una idea de la magnitud del problema, basta recordar que en nuestras casas se recomienda reemplazar los focos que consumen mucha energía (de 50 o 60 watts) por los nuevos focos ahorradores de 10 watts! De esta forma, el problema del consumo de energía en arquitecturas multinúcleos es también un área en donde la investigación juega un papel muy importante.

La evolución del hardware también tiene que ser acompañada por una investigación robusta en cuanto a la reducción de consumo de energía. La lección aprendida en la arquitectura de los procesadores "mono núcleo" es que la inversión en transistores (recursos) que se haga en un procesador debe estar sustentada con una mejora en el rendimiento con la misma proporción del porcentaje de transistores que se requieran para alcanzar ese rendimiento. Esto nos conduce a un replanteamiento de la arquitectura de los procesadores, ya que el modelo de las arquitecturas de los procesadores de hoy en día no garantiza la relación transistores-rendimiento requerida.

El SyAPAR

El panorama que vemos a corto y mediano plazo en el SyAPAR (Laboratorio de Simulación y Algoritmos Paralelos) del Centro de Investigación en Computación del IPN, evidencia la urgencia de un replanteamiento del perfil que estamos formando en nuestros alumnos. Para ello estamos siguiendo la estrategia del Par Lab de la Universidad de Berkeley, en donde el trabajo de investigación se realiza sobre aplicaciones reales, dejando de lado el esquema tradicional de trabajar en modelos que, quizás, nunca se lleguen a probar en las aplicaciones que realmente requieren de una mejora en su rendimiento. Para ello seleccionaremos aplicaciones que sean un reto como la simulación biomolecular, la modelación física, la modelación de proteínas, la modelación del flujo vehicular y de la calidad del aire, la modelación meteorológica, la modelación geológica, el cambio climático.

El replanteamiento del SyAPAR no sólo se da en sus expectativas, sino que representa todo un proceso de actualización de los planes de estudio de los cursos de programación que se imparten dentro de los programas de Maestría y Doctorado, en donde también se incluye la incorporación de especialistas bajo el esquema de contratación de excelencia del IPN y la vinculación con grupos o instituciones que estén realizando investigación y desarrollo tecnológico de primer nivel.

Conclusiones

A través de la información y de las reflexiones que exponemos en este artículo, intentamos proporcionar el panorama global de una de las muchas descripciones que se puedan hacer del Supercómputo en el 2010. Un Supercómputo que, tecnológicamente, cambió muy rápido y que, en muchos casos, tomó por sorpresa a sus usuarios y a los que forman a los usuarios. Subrayamos la necesidad de formar a los futuros programadores con las bases y las técnicas que les permitan explotar la capacidad computacional que ya está disponible a través de las arquitecturas o procesadores multinúcleos. También planteamos la necesidad de redirigir la investigación que se realiza en la línea del supercómputo de tal forma que los programas

sean eficientes, portables y escalables (que se adapten al incremento del número de núcleos que sean integrados al sistema). Ahora bien, esta investigación debe, por sobre todo, hacer más fácil la programación de los mismos, Y finalmente en cuanto a la arquitectura de los procesadores multinúcleos, hablamos de la necesidad de buscar nuevos mecanismos que permitan reducir su consumo de energía, ya que sería inviable mantener la tendencia que este rubro ha mantenido en los procesadores que existen hoy en día.

Referencias

K. Asanovic, R. Bodik, J. Demmel, T. Keaveny, K. Keutzer, J. D. Kubiatowicz, N. Morgan, D. A. Patterson, K. Sen, J. Wawrzynek, D. Wessel, and K. A. Yelick. "A View of the Parallel Computing Landscape". Communications of the ACM, Vol. 52, No. 10, (2009) 56-67.

Wolffe, G. and Trefftz, C. 2009. "Teaching parallel computing: new possibilities". J. Comput. Small Coll. 25, 1 (2009), 21-28.

H. W. Meuer. "The TOP500 Project. Looking Back over 15 Years of Supercomputing Experience". PIK - Praxis der Informationsverarbeitung und Kommunikation. Volume 31, (2008) 122–132.

Penry David A. "Multicore Diversity : A Software Developer's Nightmare". Operating systems review, vol. 43, no2, (2009) 100-10.

Anant Agarwal, Markus Levy. "The kill rule for multicore". Annual ACM IEEE Design Automation Conference archive Proceedings of the 44th annual Design Automation Conference. San Diego, California. 2007.

R. Ramaswami and K. N. Sivarajan. The Future of Supercomputing: An Interim Report. Committee on the Future of Supercomputing, Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, National Research Council, National Academies Press.

Fan, Z., Qiu, F., Kaufman, A., and Yoakum-Stover, S. 2004. "GPU Cluster for High Performance Computing". In Proceedings of the 2004 ACM/IEEE Conference on Supercomputing, Conference on High Performance Networking and Computing. IEEE Computer Society, Washington, DC. (2004).

Valero, M. "A european perspective on supercomputing". In Proceedings of the 23rd international Conference on Supercomputing.Yorktown Heights, NY, USA, 2009.

10 de marzo 2010 • Volumen 11 Número 3 • ISSN: 1067-6079