Paralelización de Aplicaciones Econométricas que requieren Estimación de los Modelos de Elección Discreta

Mariano Trigila, Ricardo Di Pasquale

Ingeniería en Informática / Facultad de Ciencias Fisicomatemáticas e Ingeniería / Pontificia Universidad Católica Argentina
Alicia Moreau de Justo 1600 (Edificio San Jose) 2° Piso, Sector Investigadores, Box 39, Buenos Aires, Argentina
marianotrigila@gmail.com
ricardo.dipasquale@gmail.com

Resumen

En las últimas dos décadas, el uso eficiente del hardware para aplicaciones científicas fue creciendo en dificultad. Además, muchas de estas aplicaciones requieren mejorar el rendimiento del procesamiento y tratar datos masivos. Estos son sistemas complejos de implementar y en especial para aquellos que no son especialista en computación. Es necesario desarrollar e implementar abstracciones de programación de alto nivel que permitan modelos de programación simples de usar. Las aplicaciones econométricas que requieren modelos de elección discreta son aplicaciones de este tipo. En este proyecto, se aplicarán las metodologías para crear abstracciones de programación de alto nivel para extender el *framework* R, para estas aplicaciones. Los objetivos del proyecto son cubrir las necesidades de: 1) procesamiento y soporte de resolución de problemas computacionales en econometría de otros grupos de investigación; y 2) transferencia del conocimiento de las tecnologías existentes, del diseño y desarrollo de aplicaciones en el área del procesamiento paralelo para la formación de recursos humanos; y para la actualización de graduados en informática y de otras áreas en este tipo de procesamiento que requieren las nuevas aplicaciones de gran escala. Los primeros resultados de este proyecto que comenzó, en junio de 2015, son dos cursos sobre computación paralela.

Palabras clave: econometría, computación paralela, clústeres de bajo costo.

Contexto

A partir de la finalización del proyecto de instalación del clúster de computadoras para procesamiento paralelo en el laboratorio LAPPA de la Facultad de Fisicomatemática e Ingeniería de la Pontificia Universidad Católica Argentina (UCA), el presente proyecto da comienzo a la investigación con énfasis en la integración de computación de alto rendimiento (HPC: High Performance Computing), la investigación y desarrollo en el área de algoritmos aplicados para brindar la formación y asistencia de procesamiento paralelo que requiere actualmente la comunidad de investigación y enseñanza de la UCA y de otras instituciones. Este proyecto se centra en paralelización de aplicaciones econométricas. El principal resultado de

este proyecto será la formación de recursos humanos para la investigación, la enseñanza y la asistencia a otros grupos de investigación de la Facultad de Fisicomatemáticas e Ingeniería de la UCA y de la comunidad.

Introducción

El primer desafío, y el más controvertido para el uso de la computación paralela, consiste en que la programación secuencial sigue siendo la habilidad más habitual de los programadores actuales. El segundo desafío son los sistemas operativos y compiladores. Estos fueron creciendo y evolucionaron resistentes al cambio que supone incluir el paralelismo. El tercer desafío final es como medir el rendimiento en lenguajes y frameworks de programación paralela sin que dependa estrictamente del programador o de sugerencias de otros que realizaron trabajos similares. [1].

Frente a esta situación de interrogantes, Asanovic et al. [1] de la Universidad de Berkeley plantean dos preguntas: ¿Cómo se puede desarrollar tecnología para beneficiar a los programadores actuales de la revolución paralela? y ¿Cómo se puede innovar rápidamente sin compiladores y sistemas operativos apropiados? El problema se agrava aún más si el investigador no es del área de computación.

Por lo tanto, varios laboratorios de investigación en computación paralela decidieron comenzar con la paralelización de aplicaciones para clientes [1] [2], [3], [4], [5], [6]. El enfoque es empezar con las aplicaciones, luego el software y finalmente el hardware [1]. El objetivo es encontrar una abstracción de más alto nivel para razonar acerca de los requerimientos de una aplicación paralela. La idea es definir los requerimientos de una aplicación de tal forma que no sea excesivamente dependiente del tipo o clase de aplicación a la cual pertenece. En la Universidad de Berkeley [7], proponen que se definan métodos para capturar los requerimientos comunes de clases de aplicaciones razonablemente separadas de implementaciones individuales.

Idealmente, el objetivo es obtener buen rendimiento en el procesamiento paralelo de las aplicaciones del futuro tanto en arquitecturas con múltiples computadoras como en arquitecturas multinúcleos.

Para seleccionar las aplicaciones se sugieren cinco criterios [1]:

1. que sea desafiante en términos de impacto social y de uso en la actividad profesional;
2. que sea factible a corto plazo y que tenga potencial a largo plazo;
3. que requiera speedup significativo o moderado sobre una plataforma más eficiente y adecuada al trabajo que debe realizarse;
4. que cubra las posibles plataformas y productos que son habituales para ese tipo de problema;
5. que genere nuevas tecnologías para otras aplicaciones;
6. que involucre a expertos del área de la aplicación en el diseño, uso y evaluación de la tecnología desarrollada.

Una de las clases de aplicaciones propuesta corresponde a las aplicaciones "vergonzosamente paralelas" (*embarrassingly parallel*) [7]. Las tareas de análisis y simulación estadística en áreas tales como bioingeniería y econometría son de este tipo [8]. En particular, la Universidad de Berkeley centra el estudio sobre una de las aplicaciones representativas que es *Monte Carlo* [9]. Esta clase de aplicaciones paralelas, en particular en el

área de econometría, son requeridas por otros grupos investigación en la UCA, para su aplicación en educación y para asistir a la comunidad en general.

Estas aplicaciones tienen como característica que los cálculos dependen de resultados estadísticos de múltiples repeticiones de ensayos y las comunicaciones entre subtareas paralelas no son dominantes. Estas aplicaciones vergonzosamente paralelas son consideradas fáciles de paralelizar sobre múltiples computadoras o múltiples núcleos de una computadora, sin embargo no son problemas paralelos clásicos. Por esta razón necesitan herramientas especiales y en general, estudios de rendimiento (*speedup*) y escalamiento específicos para encontrar una paralelización eficiente que reduzca el tiempo necesario para realizar los análisis [10]. Por consiguiente, el escaso uso de computación paralela se puede explicar por las habilidades informáticas que se requieren para desarrollar este tipo de aplicaciones paralelas.

Actualmente, existen versiones paralelas de software comúnmente usado para econometría, por ejemplo, la versión *Stata/MP* de *Stata*, *Parallel Computing Toolbox* de *Matlab*, o *MP Connect Setup* de *SAS*. Cada uno de estos productos paralelos publicita que los programas estadísticos pueden ser implementados en paralelo con muy pocas modificaciones de sus métodos existentes y que reducen el tiempo de procesamiento drásticamente en comparación con la versión secuencial estándar. Las versiones paralelas de estos productos y sus licencias de uso son costosas y no suelen estar disponibles en universidades y centros de investigación.

En la literatura reciente hay múltiples artículos que demuestran el creciente interés en la computación paralela para el área de estadística y en especial usando R [11], [12]. R es un proyecto de software libre que se distribuye bajo la licencia GNU GPL y está disponible para todos los sistemas operativos de PC. R proporciona una amplia gama de herramientas estadísticas, permitiendo que el desarrollador las extienda definiendo sus propias funciones en R e implemente bibliotecas de algoritmos más complejos de carga dinámica usando C, C++ y Fortran. R puede integrase con distintas bases de datos y tiene una capacidad gráfica importante [13]. Sin embargo, R por sí mismo no permite ejecución en paralelo, sino que requiere de alguno de los paquetes existentes que permitan distribuir cálculos sobre múltiples computadoras o de su integración con plataformas Big Data como Apache Spark.

En el reporte técnico de Schmidberger et al. [14], se reseñan dieciséis diferentes paquetes que proveen soporte paralelo para R y se compara su estado de desarrollo, la tecnología paralela empleada, usabilidad, aceptación y rendimiento. El problema que subsiste es que requieren que el usuario/desarrollador realice el armado, configuración y administración del clúster para resolver un problema en particular. Además, la generalización es compleja ya que no todos los paquetes corren en cualquier configuración de hardware y sistema operativo, requieren programación en *MPI* o *PVM* para correr sobre un clúster y otros son específicos para *grid*.

Líneas de Investigación, Desarrollo e Innovación

El objetivo de este proyecto es la formación de recursos humanos en el área de procesamiento paralelo, para asistencia a otros grupos de investigación y su aplicación a la enseñanza de grado y

posgrado Se toma como meta la investigación en paralelización de bibliotecas del lenguaje R para aplicaciones econométricas de modelos de elección discreta con el enfoque que invita adoptar la Universidad de Berkeley [7], [1] y otros centros de computación paralela [6], [5].

Además, el tema es apropiado para la formación de recursos humanos de un proyecto conformado por investigadores que se están formando en investigación y becarios porque cumple con las líneas de investigación general, brinda ejemplos y permite estudiar exhaustivamente características, limitaciones y soluciones específicas que requieren generalización.

Las líneas de investigación son:

1. El estudio de las características de las aplicaciones econométricas y sus desafíos en la generalización. Se estudiarán los modelos de regresión logística multinomial en el contexto de modelos de elección discreta usando el framework R.
2. El estudio de las características de la computación paralela, su impacto sobre el *speedup*, sus dificultades para su generalización y los limites que imponen las plataformas de hardware y software sobre las soluciones a problemas econométricos del tipo "vergonzozamente paralelo". En este proyecto, el estudio se centra sobre clústeres del tipo *Beowulf* y NOW (*Network of Workstations*).
3. El estudio de las metodologías de diseño para crear soluciones y abstracciones exitosas en sistemas distribuidos paralelos.
4. El desarrollo de nuevas bibliotecas y extensiones de métodos para el proyecto R. El aporte se centra en bibliotecas para aplicaciones que requieren modelos de regresión logística multinomial que usan los principios de maximización, tales como estimación de máxima verosimilitud.

Resultados y Objetivos.

El proyecto se inició en junio de 2015. Se publicaron dos artículos en congresos internacionales como resultado de los estudios realizados sobre computación paralela para clústeres de bajo costo y clústeres NOW. Estos artículos corresponden a los resultados de instalación del clúster *Beowulf* LAPPA de la UCA y su adecuación para este proyecto [15]. El segundo artículo corresponde a un curso de postgrado para graduados en carreras de computación. Es un curso de diseño y programación de aplicaciones en diferentes ambientes paralelos (MPI, Apache Hadoop y Spark) [16].

Además, se han dicado cursos de extensión a la comunidad sobre MPI y OPENMP, cumpliendo con uno de los objetivos principales del proyecto que es la transferencia al área educativa para la formación de profesionales en el área.

En la línea de investigación correspondiente a las aplicaciones econométricas, se empezará a trabajar con asistencia de un grupo de investigación que usa econometría para su proyecto y que será el usuario inicial de las extensiones al proyecto R. Simultáneamente, se continúa investigando en el área de computación paralela.

Una vez caracterizados el ambiente de desarrollo y las aplicaciones econométricas se comenzará con el estudio de la metodologia de diseño y de abstracción, para la implementación de las bibliotecas correspondientes.

Formación de Recursos Humanos

El equipo actual consta de dos investigadores. Uno de ellos está realizando su doctorado en otra institución.

En la Facultad de Ciencias Físicomatemáticas e Ingeniería, durante 2015, se desarrollaron Trabajos Finales (Tesinas) de alumnos de grado de Ingeniería en Informática en temas relacionados con las temáticas del proyecto. Como resultado se ha completado una tesina y hay 2 tesinas actualmente en curso.

Referencias

[1] Asanovic, K., Bodik, R., Demmel, J., Keaveny, T., Keutzer, K., Kubiatowicz, J., Morgan, N., Patterson, D., Sen, K., Wawrzynek, J., Wessel, D. and Yelick, K. 2009. "A View of the Parallel Computing Landscape". *Commun. ACM*. 52, 10 (Oct. 2009), 56–67.

[2] Oxford University, Department of Statistics, http://www.stats.ox.ac.uk/research (1/15/2016).

[3] Edinburg University School of Mathematics, http://www.maths.ed.ac.uk/research (1/15/2016)

[4] UCLA University, Idre (Institute for Digital Research and Education) http://www.hoffman2.idre.ucla.edu/ (1/15/2016).

[5] Pacific Northwest National Laboratory, U.S. Department of Energy, High-Performance Computing: *http://hpc.pnnl.gov/*. (1/15/2016).

[6]: A Strategy for Research and Innovation through High Performance Computing, Editors Mark Sawyer, Business Development and Project Manager, EPCC Mark Parsons, Executive Director, EPCC; Associate Dean for e-Research, University of Edinburg, PlanetHPC is supported under Objective "Computing Systems" of Challenge 3 "Components and Systems" of the ICT Programme of the European Commission, http://cordis.europa.eu/fp7/ict/computing/documents/planethpc-strategy.pdf (1/15/2016)

[7] Asanovic, K., Bodik, R., Catanzaro, B. C., Gebis, J. J., Husbands, P., Keutzer, K., Patterson, D. A., Plishker, W.L., Shalf, J., Williams, S. W. and Yelick, K.A. 2006. *The Landscape of Parallel Computing Research: A View from Berkeley*. Technical Report #UCB/EECS-2006-183. EECS Department, University of California, Berkeley.

[8] Grama, A., Karypis, G., Kumar, V., and Gupta A. 2003 *Introduction to Parallel Computing*. Pearson Education, second edition.

[9] Aspuru-Guzik A, Salomon-Ferrer, R., Austin, B., Perusquia-Flores, R., Griffin, M.A., Oliva, R.A., Skinner, D., Domin, D. and Lester, Jr., W.A. 2005. "Zori 1.0: A Parallel Quantum Monte Carlo Electronic Package," *Journal of Computational Chemistry*, 26, 8, (Jun. 2005), 856–862.

[10] Bischl, B., Lang, M., Mersmann, O., Rahnenführer, J., Weihs, C., 2012, *Computing on high performance clusters with R: Packages BatchJobs and BatchExperiments*, Technical Report, TU Techniche Universität Dortmund.

[11] Eugster, M. J. A., Knaus, J., Porzelius, C., Schmidberger, M. & Vicedo, E. 2011, Hands-on tutorial for parallel computing with R, *Computational Statistics* 26, 219–239.

[12] Hayfield, T. & Racine, J. S. 2012, *Parallel nonparametric kernel smoothing methods for mixed data types*, http://cran.r-project.org/web/packages/npRmpi/npRmpi.pdf .

[13] Project R. https://www.r-project.org/ (2/29/2016)

[14] Schmidberger M, Morgan M, Eddelbuettel D, Yu H, Tierney L, Mansmann U. 2009. "State of the Art in Parallel Computing with R." *Journal of Statistical Software*, 31, 1, 1–27. http://www.jstatsoft.org/v31/i01/ (1/15/2016).

[15] Trigila, M., 2015, "Cluster Lappa: Implementation of a Low Cost Cluster For Parallel Processing". *WCSEIT'2015 II World Congress on Systems Engineering and Information Technologies.* Vigo, España, 13. https://issuu.com/council.copec/docs/ba_wcseit2015_full (2/29/2016).

[16] Trigila, M., Di Pasquale, R., 2016, "Teaching Parallel Computing with Low-Cost Cluster". *INTERTECH 2016: XIV International Conference on Engineering and Technology Education*, Salvador, Bahía, Brasil,