Optimización de Algoritmos y Aplicaciones Paralelas en Sistemas Heterogéneos Mediante el Uso Combinado de Modelos Formales de Cómputo y Comunicaciones

Idioma Español

Investigadores:

- Juan Antonio Rico Gallego (Investigador Principal). Universidad de Extremadura.
- Juan Carlos Díaz Martín. Universidad de Extremadura.
- Juan Luis García Zapata. Universidad de Extremadura.
- Javier Corral García. Fundación COMPUTAEX.
- Jesús Calle Cancho. Fundación COMPUTAEX.
- Carmen Calvo Jurado. Universidad de Extremadura.
- Jesús Álvarez Llorente. Universidad de Extremadura.

Descripción:
Los sistemas hardware utilizados en computación paralela de altas prestaciones son heterogéneos. Centros de cálculo y grandes instalaciones han incorporado máquinas que combinan procesadores multi-core y aceleradores (GPUs, etc.) conectados mediantes diferentes canales de comunicación, fundamentalmente memoria compartida y redes de altas prestaciones como Infinband. El objetivo de esta propuesta es doble: incrementar el rendimiento de supercomputación actual, crecientemente heterogénea, y al mismo tiempo reducir su consumo energético.

Las aplicaciones que ejecutan en estas plataformas lo hacen usualmente en una secuencia de fases de computación y comunicación. El código de las mismas está compuesto por una serie de patrones o kernels, como por ejemplo una multiplicación de matrices o una trasformada de Fourier. Estos kernels son ejecutados de forma paralela por procesos que ejecutan sobre hardware heterogéneo, y por tanto, con diferentes capacidades de cómputo. En el actual estado del arte, el tiempo de ejecución del kernel se optimiza mediante una distribución no uniforme de la carga de cómputo, asignando a cada proceso el trabajo que pueda realizar de forma que la carga global quede equilibrada, a fin de que todos los procesos lleguen a la vez a la fase de comunicación de resultados, evitando que los más rápidos esperen a los más lentos.

En los sistemas heterogéneos que encontramos en las instalaciones actuales de supercomputación, el criterio de equilibrado basado en la carga es necesario pero es insuficiente, y esta es la cuestión que acomete la propuesta. El problema es ahora bien conocido y comienza a ser objeto de atención por parte de la comunidad investigadora. La distribución no uniforme de la carga de trabajo no sólo produce diferencias en el volumen de datos que cada proceso transmite, sino que también determina la utilización de los diferentes canales de comunicación del sistema.

En consecuencia, encontrar una asignación de carga óptima, exige no sólo conocer la capacidad de carga de los procesadores implicados, sino también estimar el coste de las comunicaciones derivado de cada asignación. Si actualmente la búsqueda de la mejor configuración se hace a través de costosos tests que hacen un uso extensivo de los recursos de cómputo del sistema, en este proyecto se propone la utilización de un modelo analítico de predicción de coste de comunicaciones desarrollado en la Universidad de Extremadura (UEX) y el University College de Dublin (UCD) de Irlanda. El modelo se denomina τ-Lop. Evaluado en clusters multi-core con redes de altas prestaciones, τ-Lop ha sido publicado en dos revistas de reconocido prestigio (primer cuartil y primer decil), y ha sido objeto de una tesis doctoral con mención europea en la UEX. La propuesta busca financiación para extenderlo a plataformas heterogéneas. Consideramos que la aproximación basada en τ-Lop posibilitará automatizar el proceso de asignación óptima de carga en arquitecturas heterogéneas, evitando el consumo intensivo de recursos que requieren los tests, y obteniendo una mejora significativa del rendimiento durante la ejecución de la aplicación. Ello redundará en el ahorro apreciable de costes tanto computacionales como energéticos en las instalaciones de supercomputación.

El objetivo último del proyecto, no obstante, es eminentemente práctico. Consiste en el desarrollo de una herramienta de usuario que proponga una distribución de carga teniendo en cuenta los modelos computacional y de comunicaciones de la aplicación. La integración de ambos modelos es la parte central de la propuesta. Supondría el primer modelo de estas características que no solo puede ser usado en aras del rendimiento y el consumo energético, sino en simuladores y algoritmos de optimización.

La propuesta parte del trabajo realizado por el investigador principal en el desarrollo de su tesis doctoral. Cuenta con un equipo multidisciplinar que incluye expertos de diferentes campos, y trabaja habitualmente con investigadores nacionales e internacionales para afrontar un proyecto con dos hitos principales, el desarrollo de un modelo formal mixto de asignación de carga, y su posterior aplicación en forma de una herramienta software usable por el usuario final de un centro de supercomputación.

Fuentes de financiación:

- Proyecto cofinanciado por la Junta de Extremadura y el Fondo Europeo de Desarrollo Regional (FEDER) de Extremadura al 80 %, dentro del Objetivo Temático 01 “Refuerzo de la investigación, el desarrollo tecnológico y la innovación”, a través de la convocatoria de ayudas destinadas a la realización de proyectos de investigación, orientados hacia las áreas estratégicas de la economía regional contempladas en el V Plan Regional de I+D+i (2014-2017), en los centros públicos de I+D+i de la Comunidad Autónoma de Extremadura, al amparo del Decreto 68/2016 de 6 de junio.

URL del envío: http://www.cenits.es/proyectos/optimizacion-algoritmos-aplicaciones-paralelas-sistemas-heterogeneos-mediante-uso