Resumen

La actual era multinúcleo y la futura era *manycore/manythread* generan grandes retos en el área de la computación incluyendo, entre otros, la programación paralela productiva o la gestión eficiente de la energía. El último objetivo es alcanzar las mayores prestaciones limitando el consumo energético y garantizando una ejecución confiable. El incremento del número de contextos hardware de los sistemas hace que el planificador se convierta en un componente importante para lograr este objetivo debido a que existen múltiples formas diferentes de planificar las aplicaciones, cada una con distintas prestaciones debido a las interferencias que se producen entre las aplicaciones. Seleccionar la planificación óptima puede proporcionar importantes mejoras de prestaciones.

Esta tesis se ocupa de las interferencias entre aplicaciones, cubriendo los problemas que causan en las prestaciones y equidad de los sistemas actuales. El estudio empieza con procesadores multinúcleo monohilo (Intel Xeon X3320), sigue con multinúcleos con soporte para la ejecución simultanea (SMT) de dos hilos (Intel Xeon E5645), y llega al procesador que actualmente soporta un mayor número de hilos por núcleo (IBM POWER8). La disertación analiza los principales puntos de contención en cada plataforma y propone algoritmos de planificación que mitigan las interferencias que se generan en cada uno de ellos para mejorar la productividad y equidad de los sistemas.

En primer lugar, analizamos la contención a lo largo de la jerarquía de memoria. Los estudios realizados revelan la alta degradación de prestaciones provocada por la contención en memoria principal y en cualquier cache compartida. Para mitigar esta contención, proponemos diversos algoritmos de planificación cuya idea principal es distribuir los accesos a memoria a lo largo del tiempo de ejecución de la carga y las peticiones a las caches entre las diferentes caches compartidas en cada nivel.

Las altas interferencias que sufren las aplicaciones que se ejecutan simultáneamente en un núcleo SMT, sin embargo, no solo afectan a las prestaciones, sino que también pueden comprometer la equidad del sistema. En esta tesis, también abordamos la equidad en los actuales multinúcleos SMT. Para mejorarla, diseñamos algoritmos de planificación que

estiman el progreso de las aplicaciones en tiempo de ejecución, lo que permite priorizar los procesos con menor progreso acumulado para reducir la inequidad.

Finalmente, la tesis se centra en la contención entre aplicaciones en el sistema IBM POWER8 con un planificador simbiótico que aborda la contención en todo el núcleo SMT. El planificador simbiótico utiliza un modelo de interferencia basado en pilas de CPI que predice las prestaciones para la ejecución de cualquier combinación de aplicaciones en un núcleo SMT. El número de posibles planificaciones, no obstante, crece muy rápido y hace inviable explorar todas las posibles combinaciones. Por ello, el problema de planificación se modela como un problema de teoría de grafos, lo que permite obtener la planificación óptima en un tiempo razonable.

En resumen, esta tesis aborda la contención en los recursos compartidos en la jerarquía de memoria y el núcleo SMT de los procesadores multinúcleo. Identificamos los principales puntos de contención de tres sistemas con diferentes arquitecturas y proponemos algoritmos de planificación para mitigar esta contención. La evaluación en sistemas reales muestra las mejoras proporcionados por los algoritmos propuestos. Así, el planificador simbiótico mejora la productividad, en promedio, un 6.7 % con respecto a Linux. En cuanto a la equidad, el planificador que considera el progreso consigue reducir la inequidad de Linux a una tercera parte. Además, dado que los algoritmos propuestos son completamente software, podrían incorporarse como políticas de planificación en Linux y usarse en servidores a pequeña escala para obtener los beneficios descritos.