Resumen

Los procesadores multinúcleo actuales cuentan con recursos compartidos entre los diferentes núcleos. Dos de estos recursos compartidos, la cache de último nivel y el ancho de banda de memoria principal, pueden convertirse en cuellos de botella para el rendimiento. Además, con el crecimiento del número de núcleos que implementan los diseños más recientes, la red dentro del chip también se convierte en un cuello de botella que puede afectar negativamente al rendimiento, ya que las redes tradicionales pueden encontrar limitaciones a su escalabilidad en el futuro cercano.

Prácticamente la totalidad de los diseños actuales implementan jerarquías de memoria que se comunican mediante rápidas redes de interconexión. Esta organización es eficaz dado que permite reducir el número de accesos que se realizan a memoria principal y la latencia media de acceso a memoria. Las caches, la red de interconexión y la memoria principal, conjuntamente con otras técnicas conocidas como la prebúsqueda, permiten reducir las enormes latencias de acceso a memoria principal, limitando así el impacto negativo ocasionado por la diferencia de rendimiento existente entre los núcleos de cómputo y la memoria. Sin embargo, compartir los recursos mencionados es fuente de diferentes problemas y retos, siendo uno de los principales el manejo de la interferencia entre aplicaciones. Hacer un uso eficiente de la jerarquía de memoria y las caches, así como contar con una red de interconexión apropiada, es necesario para sostener el crecimiento del rendimiento en los disenõs tanto actuales como futuros. Esta tesis analiza y estudia los principales problemas e inconvenientes observados en estos dos recursos: la cache de último nivel y la red dentro del chip.

En primer lugar, se estudia la escalabilidad de las tradicionales redes dentro del chip con topología de malla, asćómo ésta puede verse comprometida en próximos disenõs que cuenten con mayor número de núcleos. Los resultados de este estudio muestran que, a mayor número de núcleos, el impacto negativo de la distancia entre núcleos en la latencia puede afectar seriamente al rendimiento del procesador. Como solución a este problema, en esta tesis proponemos una de red de interconexión óptica modelada en un entorno de

simulación detallado, que supone una solución viable a los problemas de escalabilidad observados en los diseños tradicionales.

A continuación, esta tesis dedica un esfuerzo importante a identificar y proponer soluciones a los principales problemas de diseño de las jerarquías de memoria actuales como son, por ejemplo, el sobredimensionado del espacio de cache privado, la existencia de réplicas de datos y rigidez e incapacidad de adaptación de las estructuras de cache. Aunque bien conocidos, estos problemas y sus efectos adversos en el rendimiento pueden ser evitados en procesadores de alto rendimiento gracias a la enorme capacidad de la cache de último nivel que este tipo de procesadores típicamente implementan. Sin embargo, en procesadores de bajo consumo, no existe la posibilidad de contar con tales capacidades y hacer un uso eficiente del espacio disponible es crítico para mantener el rendimiento. Como solución a estos problemas en procesadores de bajo consumo, proponemos una novedosa organización de jerarquía de dos niveles cache que utiliza una red de interconexión óptica. Los resultados obtenidos muestran que, comparado con diseños convencionales, el consumo de energía estática en la arquitectura propuesta es un 60 % menor, pese a que los resultados de rendimiento presentan valores similares.

Por último, hemos extendido la arquitectura propuesta para dar soporte tanto a aplicaciones paralelas como secuenciales. Esta extensión no es trivial, ya que la propuesta adapta dinámicamente el espacio de cache requerido por las aplicaciones en ejecución, y es bien conocido que aplicaciones paralelas y secuenciales presentan coomportamientos muy diferentes. Además, el manejo de la coherencia de memoria se vuelve más complejo, ya que los módulos de cache pueden ser asignados y desasignados a cualquier núcleo en cualquier momento de la ejecución. Los resultados obtenidos con la esta nueva arquitectura muestran un ahorro de hasta el 78 % de energía estática en la ejecución de aplicaciones paralelas.

En resumen, esta tesis estudia y propone soluciones a los problemas derivados de la compartición de caches y recursos de comunicación dentro del chip en los procesadores multinúcleo actuales. Los dos organizaciones de cache propuestas reducen el consumo de energía tanto estática como dinámica en comparación a otras arquitecturas convencionales, a la vez que consiguen un rendimiento similar.