Resumen

En los últimos años, la creciente necesidad de la capacidad de cómputo ha supuesto un reto que ha llevado a la industria a buscar arquitecturas alternativas a los procesadores superescalares con ejecución fuera de orden convencionales, con el objetivo de incrementar la potencia de cómputo con una mayor eficiencia energética.

Las GPU, que hasta hace apenas una década se dedicaban exclusivamente a la aceleración de los gráficos en los computadores, han sido una de las arquitecturas alternativas más utilizadas durante varios años para alcanzar el mencionado objetivo. Una de las características particulares de las GPU es su gran ancho de banda para acceder a memoria principal, lo que les permite ejecutar un gran número de hilos de forma muy eficiente. Esta característica, así como su elevada potencia computacional ejecutando operaciones de coma flotante, ha originado la aparición del paradigma de computación denominado *GPGPU computing*, paradigma en el que las GPU realizan cómputo de propósito general. Las citadas características convierten a las GPU en dispositivos especialmente apropiados para la ejecución de aplicaciones masivamente paralelas que tradicionalmente se habían ejecutado en procesadores convencionales de altas prestaciones.

El trabajo desarrollado en esta tesis persigue ayudar a mejorar las prestaciones de las GPU en la ejecución de aplicaciones GPGPU. Con este fin, como primer paso, se realiza un estudio de caracterización donde se identifican las características más importantes de estas aplicaciones desde el punto de vista de la jerarquía de memoria y su impacto en las prestaciones. Para ello, se utiliza un simulador detallado ciclo a ciclo donde se modela la arquitectura de una GPU reciente. El estudio revela que es necesario modelar de forma más detallada algunos componentes críticos de la jerarquía de memoria de las GPU para obtener resultados precisos.

Los resultados obtenidos muestran que las prestaciones alcanzadas pueden variar hasta en un factor de 3× dependiendo de cómo se modelen estos componentes críticos.

Por este motivo, como segundo paso antes de elaborar la propuesta de mejora, el trabajo se centra en determinar qué componentes de la jerarquía de memoria de la GPU necesitan modelarse con mayor detalle para mejorar la precisión de los resultados del simulador, y en mejorar los modelos existentes de estos componentes. Además, se realiza un estudio de validación que compara los resultados obtenidos con los modelos mejorados contra los de una GPU comercial real. Las mejoras implementadas reducen la desviación de los resultados del simulador sobre los resultados reales alrededor de un 96%.

Finalmente, una vez mejorada la precisión del simulador, en esta tesis se presenta una propuesta innovadora, denominada FRC (siglas en inglés de *Fetch and Replacement Cache*), que mejora en gran medida la potencia computacional de la GPU, gracias a que aumenta el paralelismo en el acceso a memoria principal. La propuesta incrementa el número de accesos en paralelo a memoria principal mediante la aceleración de la gestión de las acciones de búsqueda y reemplazo relacionadas con los accesos que fallan en la cache. La propuesta FRC se basa en una pequeña estructura cache auxiliar que descongestiona el subsistema de memoria eficientemente, aumentando las prestaciones de la GPU hasta un 118% de media respecto al sistema base. Además, también reduce en 57% el consumo energético de la jerarquía de memoria.