Resumen

La jerarquía de caches y la red en el chip (NoC) son dos componentes clave de los chip multiprocesadores (CMPs). La mayoría del trafico en la NoC se debe a mensajes que las caches envían según lo que establece el protocolo de coherencia. La cantidad de tráfico, el porcentaje de mensajes cortos y largos y el patrón de tráfico en general varían dependiendo de la geometría de las caches y del protocolo de coherencia. La arquitectura de la NoC y la jerarquía de caches están de hecho firmemente acopladas, y estos dos componentes deben ser diseñados y evaluados conjuntamente para estudiar cómo el variar uno afecta a las prestaciones del otro. Además, cada componente debe ajustarse a los requisitos y a las oportunidades del otro, y al revés. Normalmente diferentes clases de mensajes se envían por diferentes redes virtuales o por NoCs con diferente ancho de banda, separando mensajes largos y cortos. Sin embargo, otra clasificación de los mensajes se puede hacer dependiendo del tipo de información que proveen: algunos mensajes, como las peticiones de datos, necesitan campos para almacenar información (dirección del bloque, tipo de petición, etc.); otros, como los mensajes de reconocimiento (ACK), no proporcionan ninguna información excepto por el ID del nodo destino. Esta segunda clase de mensaje no necesita de mucho ancho de banda: la laténcia es mucho mas importante, dado que el nodo destino está bloqueado esperando su recepción. En este trabajo de tesis se desarrolla una red dedicada para trasmitir la segunda clase de mensajes; la red es muy sencilla y rápida, y permite la entrega de los ACKs con una laténcia de pocos ciclos de reloj. Reduciendo la laténcia y el tráfico en la NoC debido a los ACKs, es posible:

- acelerar la fase de invalidación en fase de escritura en un sistema que usa un protocolo de coherencia basado en directorios
- mejorar las prestaciones de un protocolo de coherencia basado en broadcast, hasta llegar a prestaciones comparables con las de un protocolo de directorios pero sin el coste de área debido a la necesidad de almacenar el directorio
- implementar un mapeado dinámico de bloques a las caches de último nivel de forma eficiente, con el objetivo de acercar al máximo los bloques a los cores que los utilizan

El objetivo final es obtener un co-diseño de NoC y jerarquía de caches que minimice los problemas de escalabilidad de los protocolos de coherencia. En esta tesis se exploran diferentes alternativas para una entrega rapida de los ACKs y las oportunidades

que ofrece al protocolo de coherencia. Combinando los mecanismos presentados en un sistema final, se obtiene una arquitectura adaptable dinamicamente a los requisitos de múltiples aplicaciones en un entorno virtualizado.