Abstract

The cache hierarchy and the Network-on-Chip (NoC) are two key components of chip multiprocessors (CMPs). Most of NoC traffic is due to messages exchanged by the caches according to the coherence protocol. The amount of traffic, the percentage of short and long messages and the traffic pattern in general depend on the cache geometry and the coherence protocol. NoC architecture and the cache hierarchy are indeed tightly coupled, and these two components should be designed and evaluated together to study how varying one's design affects the other one's performance. Furthermore, each component should adjust to match the requirements and exploit the performance of the other one, and vice versa. Usually, messages belonging to different classes are sent through different virtual networks or through NoCs with different bandwidth, thus separating short and long messages. However, other classification of the messages can be done, depending on the type of information they provide: some messages, like data requests, need fields to store information (block address, type of request, etc.); other messages, like acknowledgement messages (ACKs), do not need to specify any information except for the destination node. This second class of messages do no require high NoC bandwidth: latency is far more important, since the destination node is typically blocked waiting for their reception. In this thesis we propose a dedicated network which is able to transmit this second class of messages; the dedicated network is lightweight and fast, and is able to deliver ACKs in a few clock cycles. By reducing ACKs latency and the NoC traffic, it is possible to:

- speed-up the invalidation phase during write requests in a system which employs a directory-based coherence protocol
- improve the performance of a broadcast-based coherence protocol, reaching performance which is comparable to that of a directory-based protocol but without the additional area overhead due to the directory
- implement an efficient and dynamic mapping of cache blocks to the last-level cache banks, aiming to map blocks as close as possible to the cores which use them

The final goal is to obtain a co-design of the NoC and the cache hierarchy which minimizes the scalability problems due to coherence protocols. In this thesis we explore the different design alternatives for fast network delivery and coherence protocol opportunities. The best mechanisms, combined on a final system, allow for a truly dynamic and customizable architecture in an environment with multiple applications demanding partitioning of resources.