Literature Report Assignment

The Locality-Aware Adaptive Cache Coherence Protocol

Cache is a small high-speed memory that stores data from some frequently used addresses of main memory. Maintaining the consistency of shared resource data stored in multiple local caches, called cache coherence, is a common drawback in a multiprocessing system. In this paper, the authors propose a scalable and efficient Locality-Aware Adaptive Coherence Protocol to address the problem of how to enable memory access whose latency and energy are low while preventing data replication and costly data movement. In this protocol, advantages of private catching and remote access are combined to retain the convenience of shared memory. Its implementation is built on top of a baseline system, an organization which include compute pipeline, private L1-I and L1-D caches and logically shared physically distributed L2 cache with integrated directory managed by Reactive-NUCA’s data placement and migration mechanisms and ACKwise limited directory coherence protocol. When a memory that misses the private cache is requested by a core, this coherence protocol works in one of two ways: 1) it brings the entire cache line using a traditional directory protocol, or 2) it just accesses the requested word at the shared cache location. In order to make this decision, it is based on the spatio-temporal locality of a particular data block. Finally, it is concluded that the protocol comes up with three potential advantages, which are better private cache utilization, reduced on-chip communication, and reduced memory access latency and energy with the overall energy consumption in a 64-core multicore by 25% and an improvement of the complete time by 15%, respectively. Furthermore, it provides an efficient locality tracking hardware and a simple implementation with no additional network for deadlock avoidance.

Reference:
URL - http://dl.acm.org/citation.cfm?id=2485967