Summary Part 1
Managing Wire Delay in Large Chip-Multiprocessor Caches

Mrinal Nath (ID: 3307043)
April 18, 2005

What are the problems solved by this paper?
Techniques like block migration and transmission line caches have been proposed to manage the wire delay due to long wires in large uniprocessor L2 caches. Hardware-directed prefetching is a technique to tolerate the long latencies associated with accessing L2 caches and/or memory. However, some of these techniques do not directly apply to a CMP design. For e.g., block migration is not as effective for CMPs as it is for unprocessors and TLCs have limited bandwidth, which can be a bottleneck in CMPs. This paper examines the three techniques – prefetching, block migration and TLCs – for managing L2 cache latency in a large CMP. It proposes a hybrid combination of these three techniques to get the best performance.

What are the approaches attempted by this paper?
First the authors consider the impact of hardware-directed prefetching to tolerate the variable latency of accessing a NUCA cache. They evaluate L2 prefetching and L1&L2 prefetching. Then they show that block migration does not perform well in a CMP environment (without prefetching). They also show that TLC improve performance across all benchmarks, but are limited by their bandwidth constrains. Finally, they propose a hybrid cache design in which all three techniques are combined.

What are the main conclusions of this paper?
Prefetching is a effective technique to hide and tolerate the variable latency of accessing NUCA caches. Block migration does not work well in a CMP due to the sharing of blocks between processors and also due to the need of a good search mechanism to reduce the miss latency. Transmission lines consistently improve performance by reducing the latency of data communication, but their bandwidth is limited. A hybrid design which tries to exploit the best of each technique seems to be the appropriate solution.