Adaptive Mechanisms and Policies for Managing Cache Hierarchies in Chip Multiprocessors
A critique on Aniket Bharadswadkar’s paper presentation in class

This paper studied adaptive mechanisms in private cache based designs to reduce off-chip traffic by discussing simple architectural extensions and adaptive policies for managing the L2 and L3 cache hierarchy in a CMP system. They propose the use of a small history table to provide hints to the L2 caches as to which lines are resident in the L3 cache. They propose the use of a L3 level of cache for holding the evicted lines from the L2 cache, avoiding write back of clean lines to memory, maximizing L2- L2 transfers thus using peer L2 lines for writing back. They also examine the performance benefits of allowing write backs from L2 caches to be placed in neighboring, on-chip L2 caches instead of forcing them to be absorbed by the L3 cache.

Strengths:

1) The new L3 cache is bound to reduce the miss rate and would certainly improve performance.
2) They used an L2 snarf table to identify locally evicted blocks that might be reused soon. Upon local eviction, such blocks are kept on-chip via write-backs to peer caches. The host cache will replace either invalidated or shared clean blocks to make room for them, potentially reducing expensive off-chip misses when they are reused later.
3) To reduce memory traffic and decrease memory latency unnecessary writes like writing back clean lines is reduced.
4) It suggests ways to improve the on chip memory usage by making use of peer L2 caches to write to while writing back.
5) The Write Back History Table (WBHT) method produces accurate predictions between 60% and 75% of the time.
   This can have a positive impact on performance and is measured to be as good as 13%.
6) The paper has also studied the challenges and proposed modifications needed for the coherence protocols to implement L2 snarfing.

Weaknesses:

1) This scheme is limited by only supporting multithreaded workloads.
2) The paper talks about the potential negative effects of the use of WBHTs and selective L2 write back scheme. But nothing has been proposed to overcome these.
   Due to the slight increase in the time a write back may reside in the write back queue while the WBHT is consulted. This may result in the queue becoming full. If the write back queue becomes full, misses to the L2 cache will be blocked and will have to wait for an open slot.
3) It wasn’t apparent if the performance of the techniques would be the same for increased number of cores. The paper did not address simulations for greater number of cores and did not take into account power issues.
4) The paper did not give an insight on the architectural impact of the introduction of the new L3 cache in their designs and also its impact on area and power.
5) Alternate L3 organizations and policies, including having separate buses for chip-private L3 caches and memory could have been discussed.
6) The paper proposes to introduce a history table for each L2 cache. This would greatly restrict the number of cores we can have on a chip or alternatively is bound to increase the chip size.
7) L2 snarfing might decrease the off-chip accesses but definitely increases the on chip accesses and this costs power and heating. These issues have not been dealt with in the paper.