Paper : Reducing DRAM latencies with an integrated memory hierarchy design
Presenter : Pravin Dalale

This paper describes techniques to reduce level-two miss latencies for memory-intensive applications that are not bandwidth bound. Among many things, it proposes the implementation of *scheduled region prefetching* in the level-two cache on a Direct Rambus memory system. Given below is a short critique of the paper that lists down it's major strengths and weaknesses:

Strengths

1. The authors do a very comprehensive analysis to identify areas such as block size, channel width, address mapping, prefetching and cache size, where improvements could be made to reduce DRAM latencies.
2. The authors propose an efficient address mapping scheme that leverages the spatial locality in applications resulting in reduced row-buffer misses and bank conflicts. They continue to prove its effectiveness by showing an average performance improvement of 16%.
3. The authors give a good detail on the prefetching methodology they suggest, covering the various aspects of it including the insertion/replacement, scheduling and queuing policies.

Weaknesses

1. The paper has little relevance for multi-processor architectures where a shared L2 cache architecture could lead to increased overheads in the form of cache coherence and thereby increased levels of contention, making prefetching less effective.
2. Most of the proposed schemes make little or no difference to the performance of bandwidth bound applications. In the paper the author fails to show performance improvement to more than half of the 26 SPEC2000 benchmarks used for evaluation
3. Increased prefetching could result in developing hot spots on chip. The author does not refer to anything related to this problem. Also, the author pins hopes of lowering memory latencies on larger on-chip caches. This has slowed down today due to various factors including the power wall.

References

1. Doug Berger et al. Reducing DRAM Latencies with an Integrated Memory Hierarchy Design

***