Abstract:
Last-level caches mitigate the high latency of main memory. A good cache replacement policy enables high performance for memory intensive programs. To be useful to industry, a cache replacement policy must deliver high performance without high complexity or cost.

I propose a novel last-level cache replacement algorithm with approximately the same complexity and storage requirements as tree-based PseudoLRU, but with performance matching state of the art techniques such as dynamic re-reference interval prediction (DRRIP) and protecting distance policy (PDP). The algorithm is based on PseudoLRU, but dynamically adapts its insertion and promotion policy. It has slightly less than one bit of overhead per cache block, compared with two or more bits per cache block for competing policies.

In this talk, I give the motivation behind the algorithm in the context of LRU with improved placement and promotion, then develop this motivation into a PseudoLRU-based algorithm, and finally give a version using set dueling to allow adaptivity to changing program behavior. I show that, with a 16-way set-associative 4MB last-level cache, an adaptive PseudoLRU insertion and promotion algorithm yields a geometric mean speedup of 5.6% over true LRU over all the SPEC CPU 2006 benchmarks using far less overhead than LRU or other algorithms. On a memory-intensive subset of SPEC, the technique improves geometric mean speedup by 15.6%. I show that the performance is comparable to state-of-the-art replacement policies that consume more than twice the area.

Bio: Daniel A. Jimenez is an Associate Professor in the Department of Computer Science and Engineering at Texas A&M University. Previously he was a full Professor and Chair of the Department of Computer Science at UTSA and Associate Professor in the Department of Computer Science at Rutgers University. His research focuses on microarchitecture and low-level compiler optimizations. He introduced and developed the perceptron branch predictor which has inspired the design of two implemented microarchitectures: the AMD "Bobcat" core and the Oracle SPARC T4. Daniel earned his B.S. (1992) and M.S. (1994) in Computer Science at UTSA and his Ph.D. (2002) in Computer Sciences at UT Austin. From 2002 through 2009, Daniel was an Assistant and later Associate Professor in the Department of Computer Science at Rutgers. In 2005 and again in 2010 Daniel took sabbatical leaves at the Technical University of Catalonia (UPC) in Barcelona, Catalonia, Spain. He returned to Texas to take a position at UTSA and later Texas A&M University. He was General Chair of the 2011 IEEE HPCA conference. He is an NSF CAREER award recipient, IEEE Senior Member, and ACM Distinguished Scientist.