Rollback-Free Value Prediction with Approximate Loads

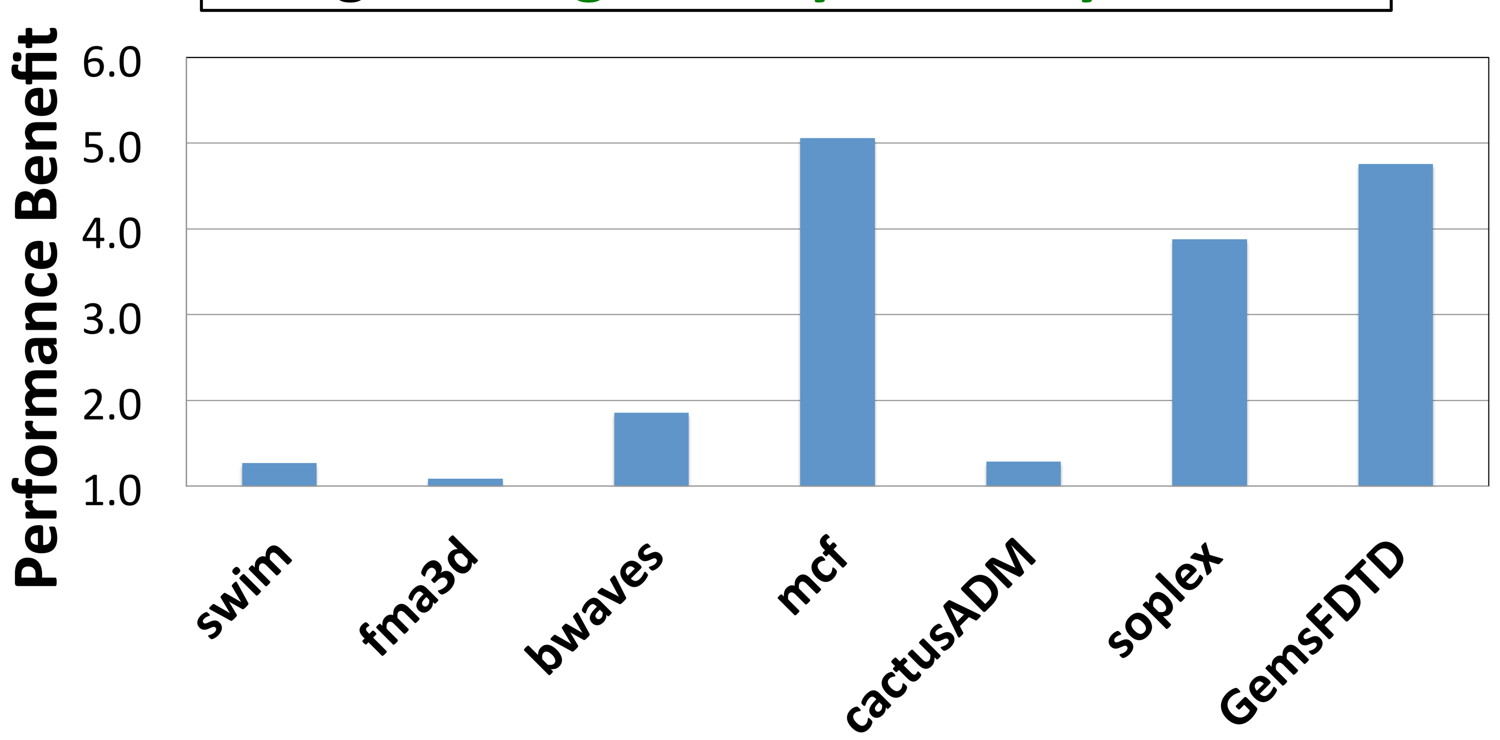
Bradley Thwaites, Gennady Pekhimenko, Amir Yazdanbakhsh, Jongse Park, Girish Mururu
Hadi Esmaeilzadeh, Onur Mutlu, Todd C. Mowry

Carnegie Mellon University

Motivation

Perfect Prediction

Mitigate **long latency memory** accesses



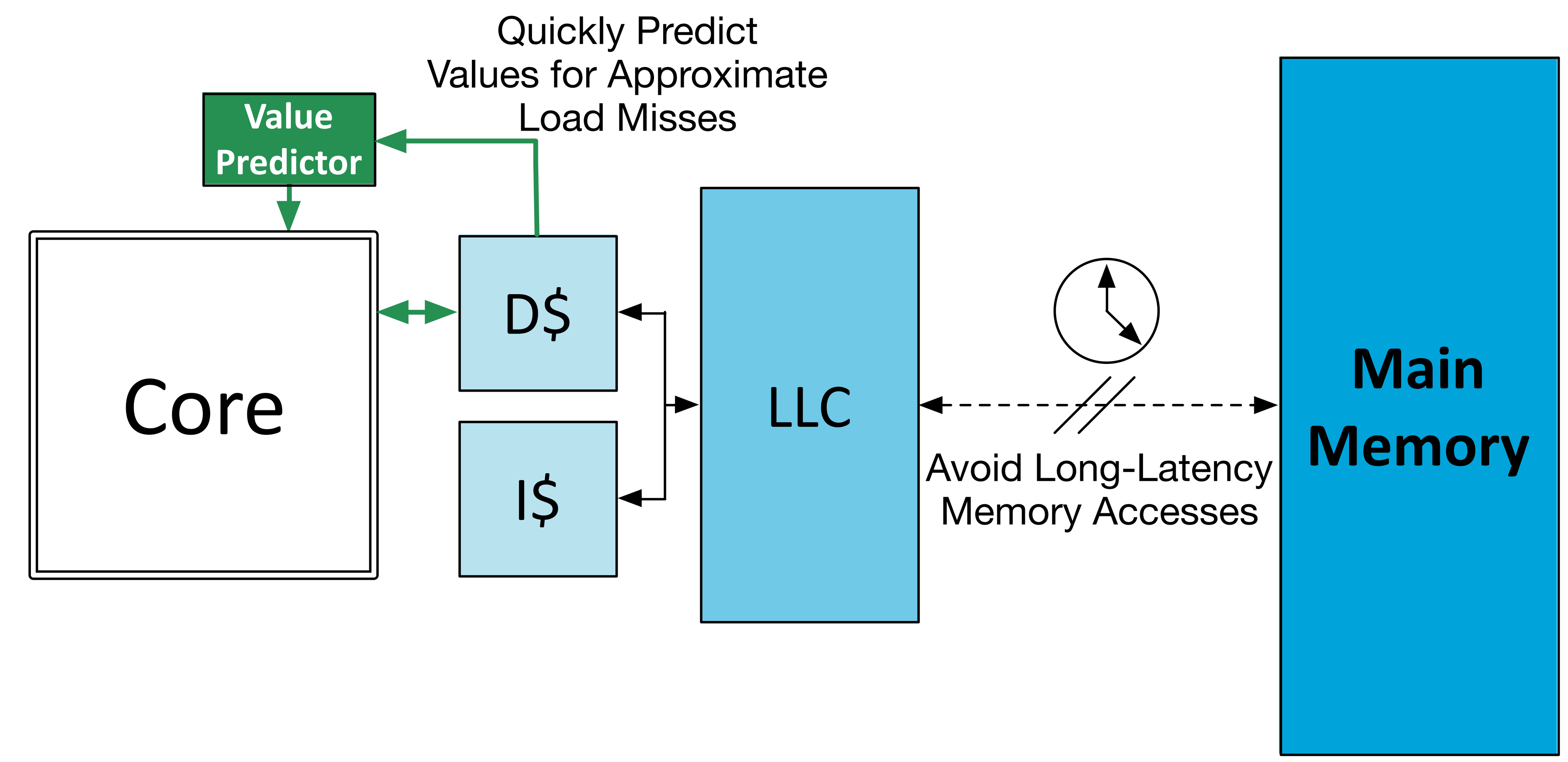
RFVP Overview

Microarchitecturally-triggered approximation

Predict the value of an approximate **load** when it **misses** in the cache

Do not **check** for mispredictions

Do not **rollback** from mispredictions



Design Principles

Maximize opportunities for performance and energy **benefits**

Minimize the adverse effects of approximation on **quality degradation**

Design Challenges

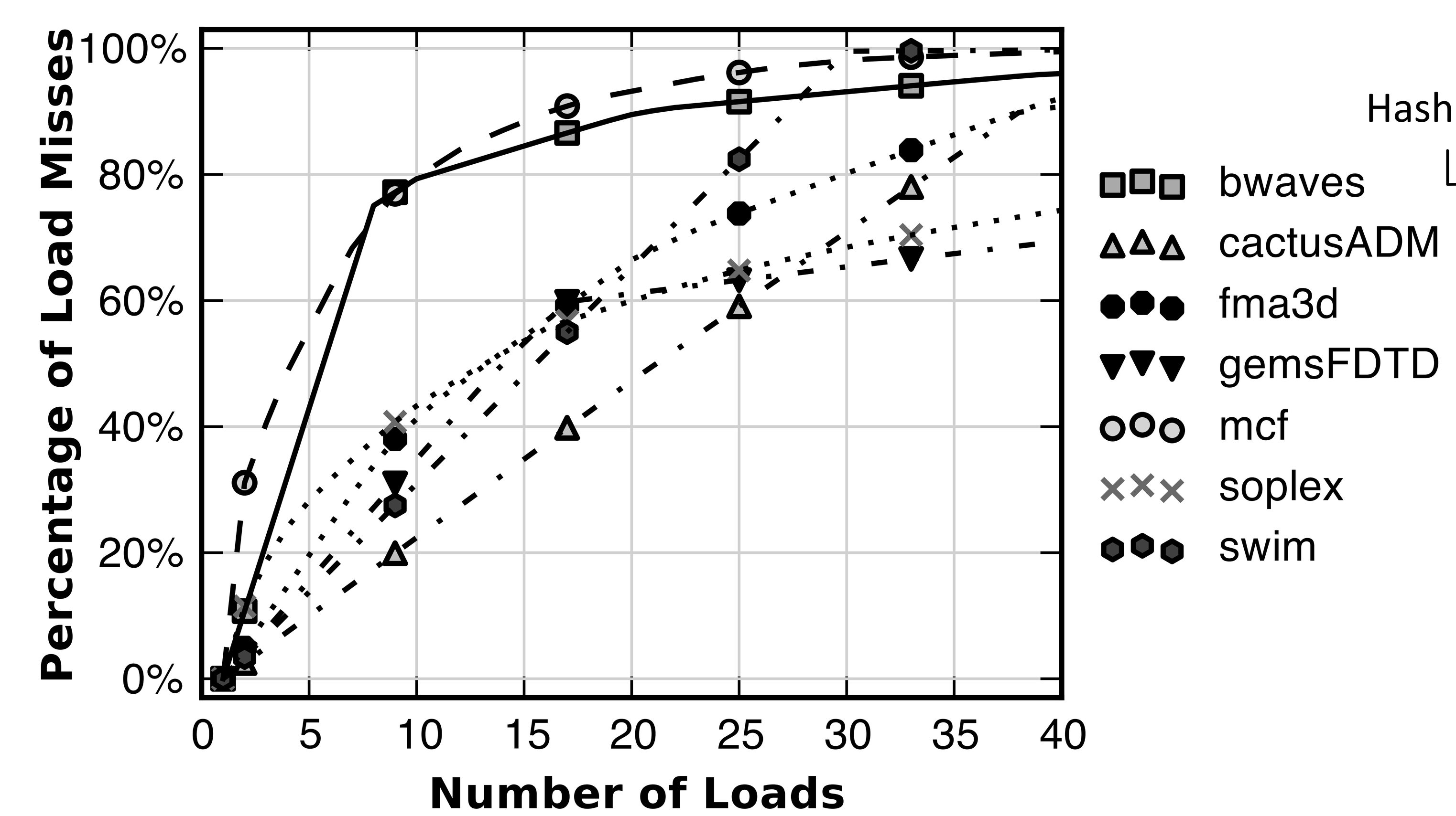
Target Performance-Critical Safe Loads

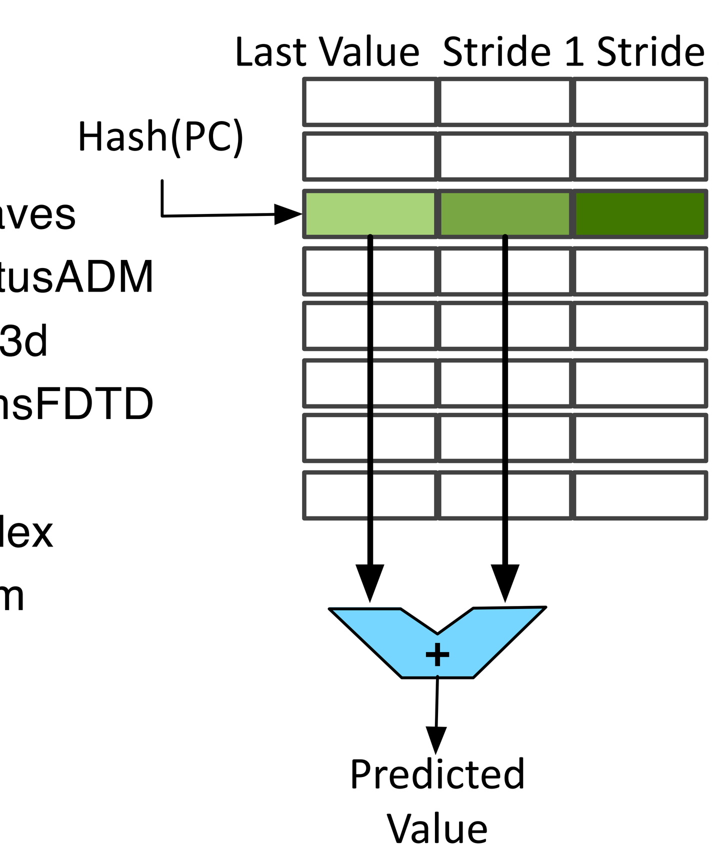
- Profile-directed compilation
- Usually, < 32 loads cause 80% of cache misses

Utilize Fast-Learning Predictors

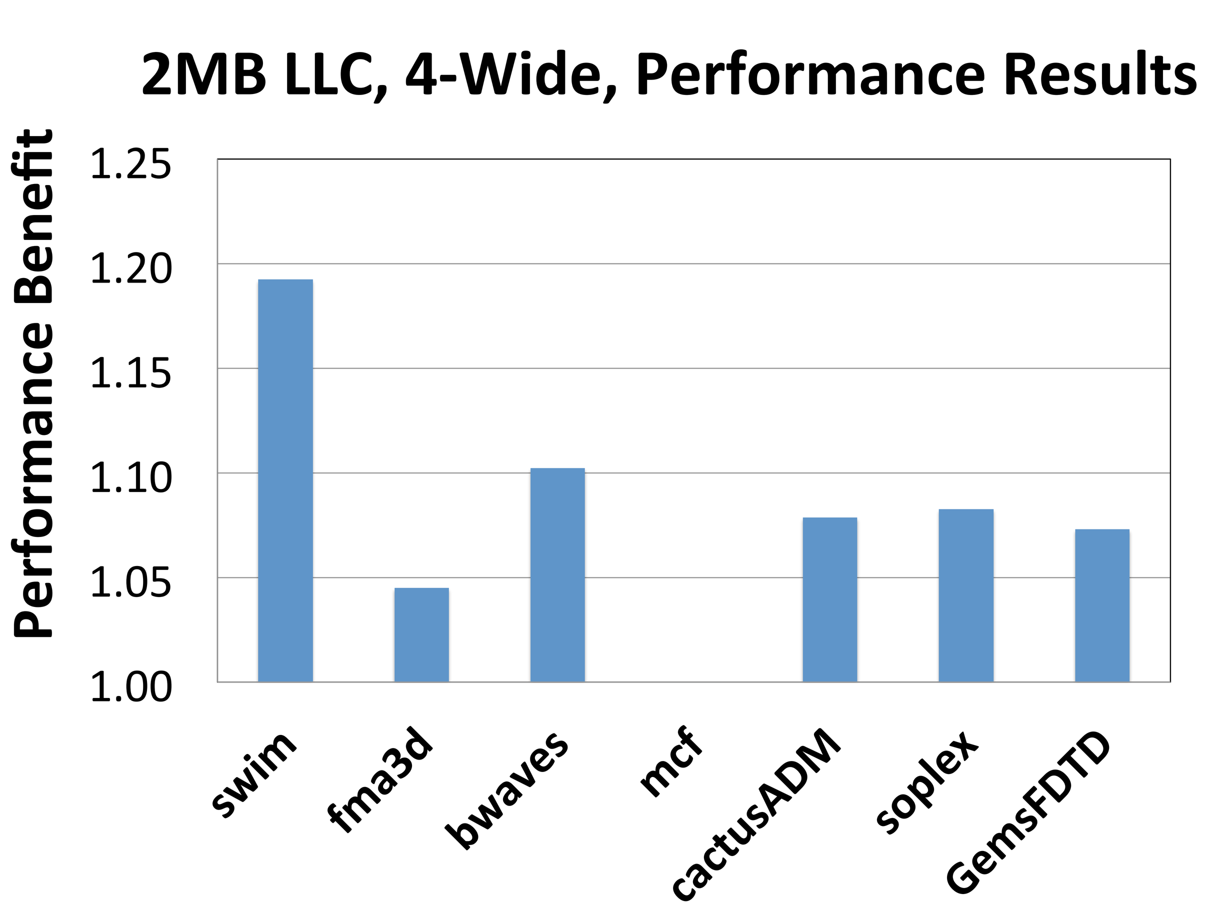
- Two-delta stride predictor
- Prediction: table lookup plus an addition

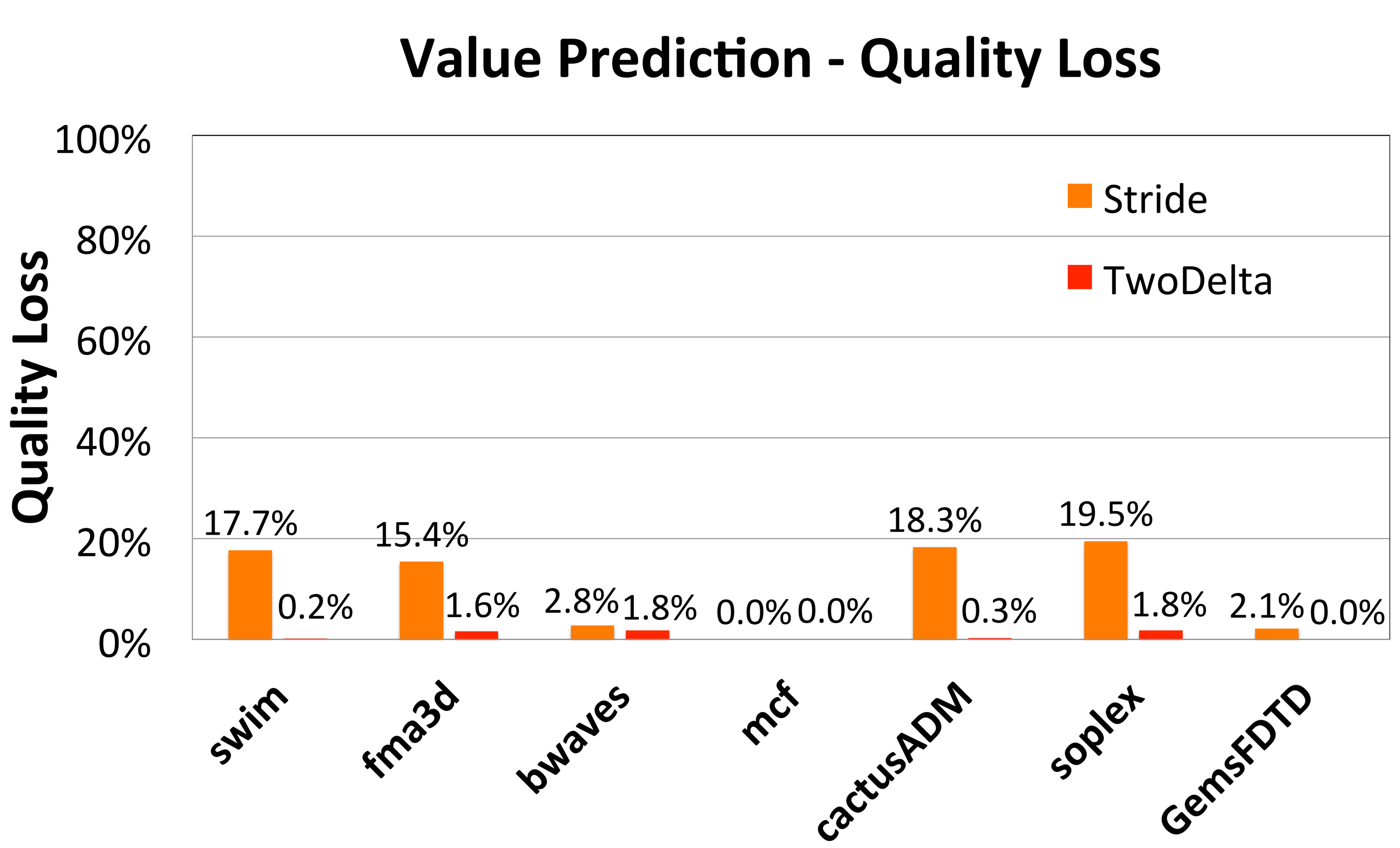
Integrate RFVP with existing architecture





Key Experimental Results





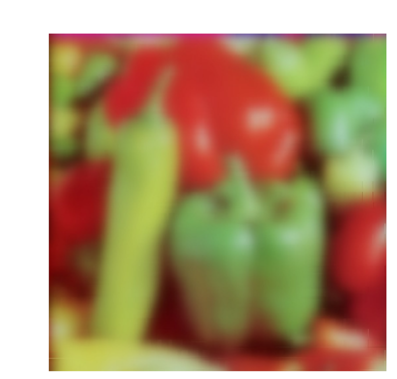
Ongoing Work

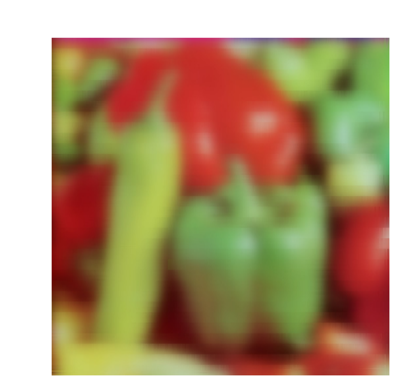
Mitigate both **Memory Wall** and **Bandwidth Wall**

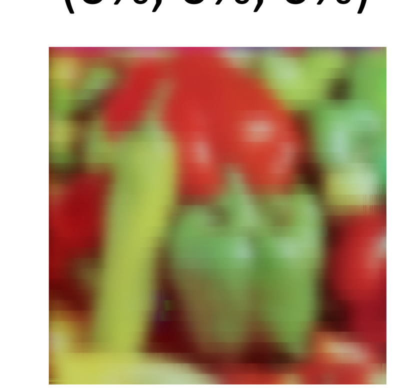
Extend rollback-free value prediction to GPUs

Drop a fraction of the missed requests

Preliminary results: Up to 2x improvement in energy and performance with only 10% quality degradation


(0%, 0%, 0%)


(1.06x, 1.03x, 2%)


(1.13x, 1.06x, 6%)


(1.27x, 1.10x, 11%)

(Performance Gain, Energy Gain, Quality Loss)