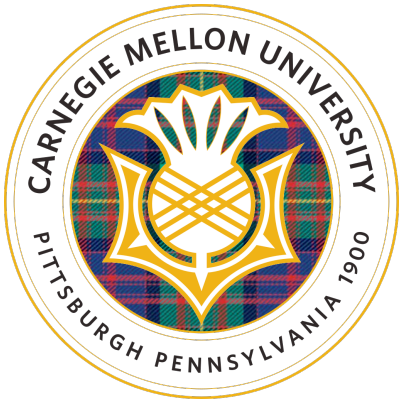
Rollback-Free Value Prediction with Approximate Loads

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

Georgia Institute of Technology
Carnegie Mellon University



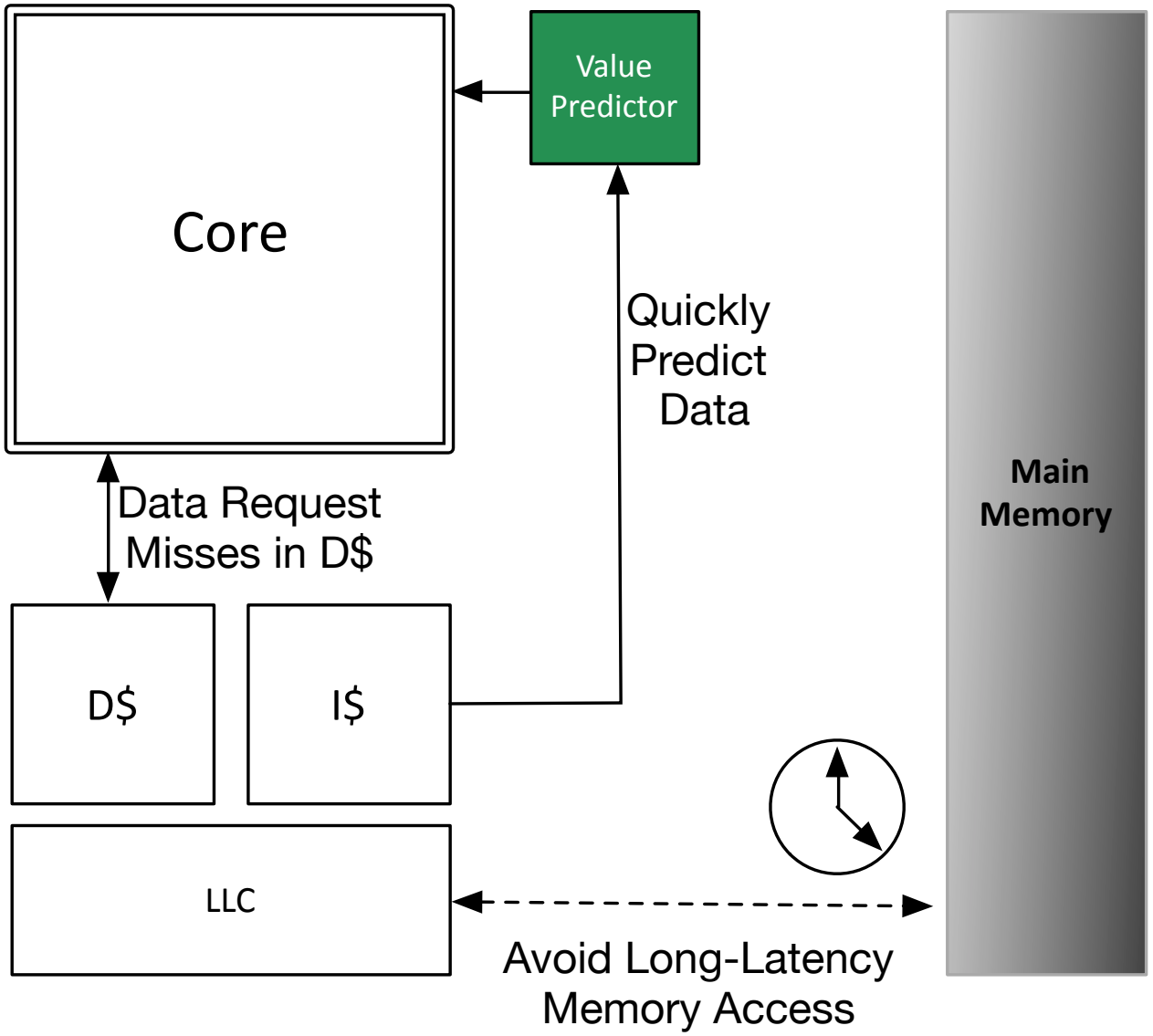
Mitigating Memory Wall with Approximation

Rollback-Free Value Prediction

- 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

Mitigate **long latency memory** accesses

Rollback Free Value Prediction



Design Principles

Maximize opportunities for performance and energy **benefits**

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

Design Challenges and Solutions

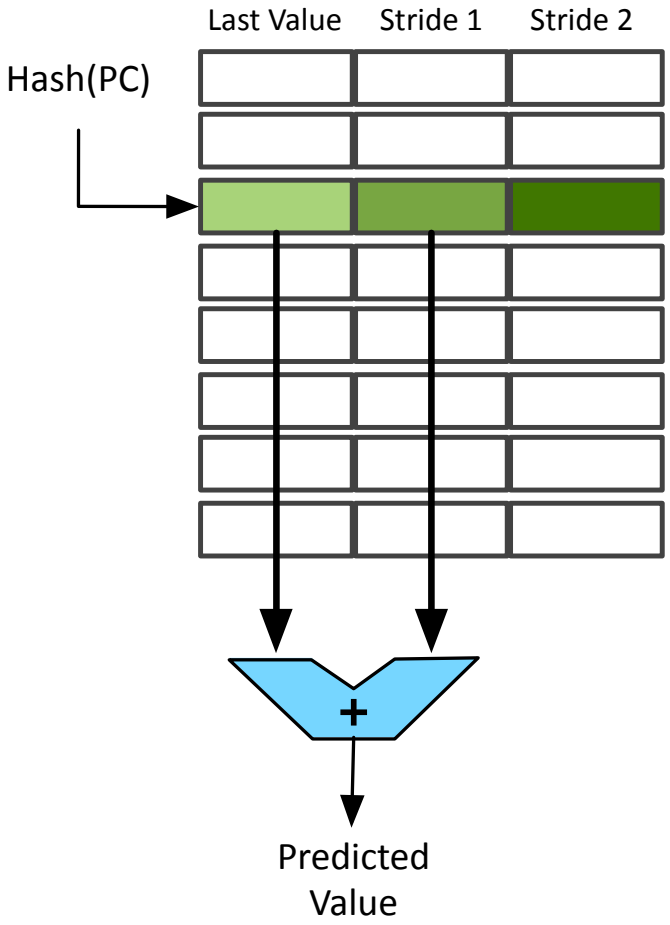
Target Performance-Critical Safe Loads

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

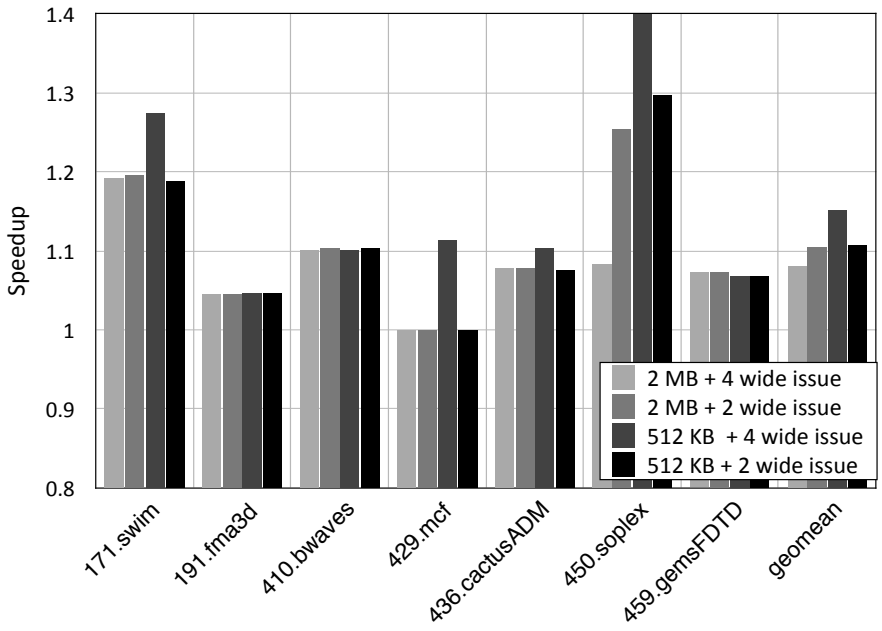
Utilize Fast-Learning Predictors

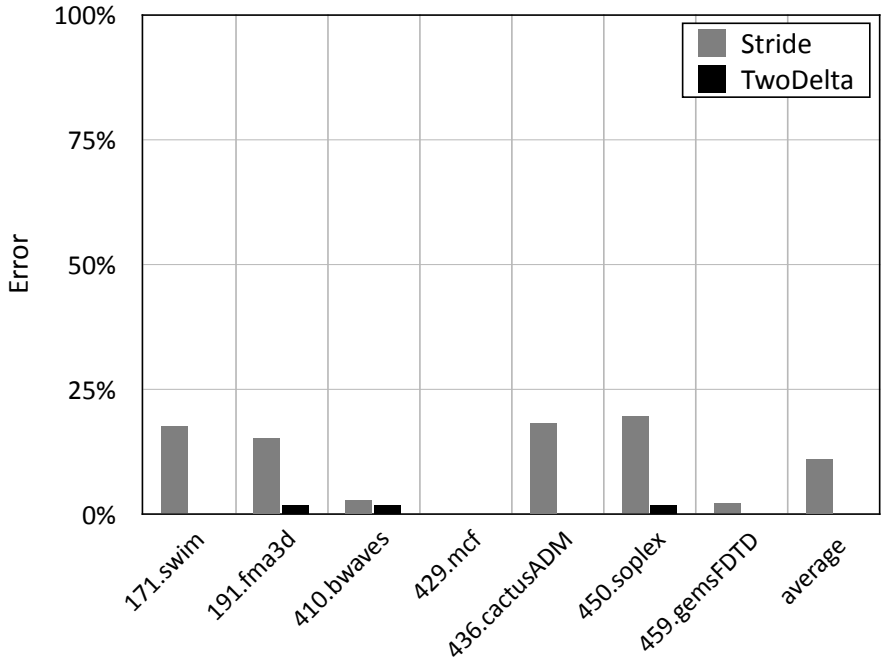
- Two-delta stride predictor
- Prediction: table lookup plus an addition

Integrate RFVP with existing architecture



Experimental Results with a Modern OoO Processor





Performance Improvement:

8% → **19%**

Average Maximum

Quality Loss:

0.8% → **1.8%**

Average Maximum

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