Concorde: Fast and Accurate CPU Performance Modeling with Compositional Analytical-ML Fusion

Arash Nasr-Esfahany, Mohammad Alizadeh, Victor Lee
Hanna Alam, Brett W. Coon, David Culler, Vidushi Dadu
Martin Dixon, Henry M. Levy, Santosh Pandey
Parthasarathy Ranganathan, Amir Yazdanbakhsh

Motivation

1. CPU Simulation

Microarchitecture simulation is a key tool in design and exploration, but we lack fast and accurate performance models.

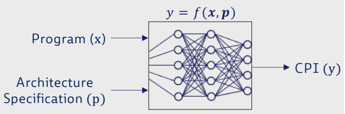
Cycle-level simulators are slow.
Analytical Models are fast, but not accurate.

Motivation

2. Prior Work

Ignores problem structure

✗ High sample complexity
✗ Bulky neural networks
✗ Slow training and inference
✗ $\mathcal{O}(\#\text{instrs})$


$y = f(\boldsymbol{x}, \boldsymbol{p})$
Program (x)
Architecture Specification (p)
CPI (y)

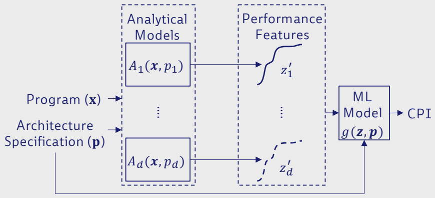
SimNet: Accurate and High-performance Architecture Simulation using Deep Learning, ACM SIGMETRICS/IFIP PERFORMANCE '22

TAO: Re-Thinking DL-based Microarchitecture Simulation, ACM SIGMETRICS/IFIP PERFORMANCE '24

Design

3. Compositional Analytical-ML Fusion

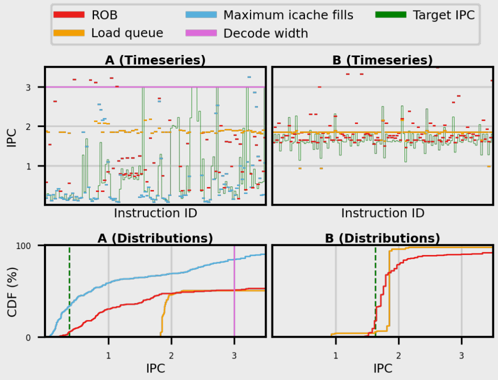
Multiple lightweight models work together to progressively achieve high fidelity with low computational complexity.


Analytical Models
$A_1(\boldsymbol{x}, p_1)$
⋮
$A_d(\boldsymbol{x}, p_d)$
Performance Features
z'_1
⋮
z'_d
Program ($\mathbf{x}$)
Architecture Specification ($\mathbf{p}$)
ML Model
$g(\boldsymbol{z}, \boldsymbol{p})$
CPI

Design

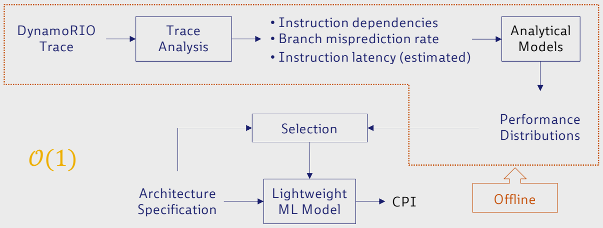
4. Performance Features

Per-resource analytical modeling produces a rich performance characterization of a program.



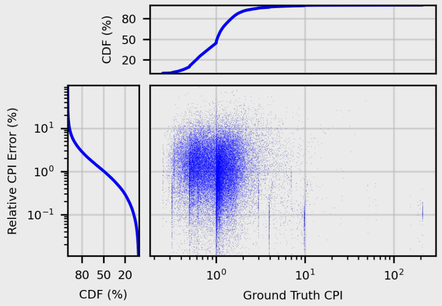
Design

5. Concorde


DynamoRIO Trace
Trace Analysis
• Instruction dependencies
• Branch misprediction rate
• Instruction latency (estimated)
Analytical Models
Performance Distributions
Selection
$\mathcal{O}(1)$
Architecture Specification
Lightweight ML Model
CPI
Offline

Evaluation

6. Concorde is fast and accurate



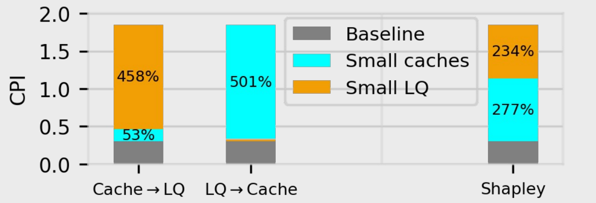
2%
CPI prediction error

$10^5\times$
speedup over cycle-level simulation

Application

7. Fine-Grained Performance Attribution

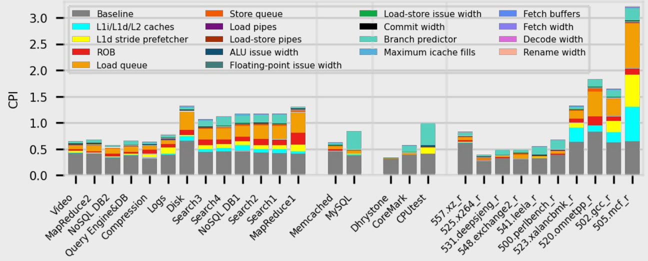
Shapley Value: A fair, order-independent attribution



Case Study

8. Large-Scale Sensitivity Analysis

143M 100k-instruction segment CPI evaluations, in just an hour!


CPI
Baseline
L1i/L1d/L2 caches
L1d stride prefetcher
ROB
Load queue
Store queue
Load pipes
Load-store pipes
ALU issue width
Floating-point issue width
Load-store issue width
Commit width
Branch predictor
Maximum icache fills
Fetch buffers
Fetch width
Decode width
Rename width
Video
MapReduce2
NoSQL DB2
Query Engine&DB
Compression
Logs
Disk
Search3
Search4
NoSQL DB1
Search2
Search1
MapReduce1
Memcached
MySQL
Dhrystone
CoreMark
CPUtest
557.xz_r
525.x264_r
531.deepsjeng_r
548.exchange2_r
541.leela_r
500.perlbench_r
523.xalancbmk_r
520.omnetpp_r
502.gcc_r
505.mcf_r