Dynamic Speculation Control of Modern Processors

Sankara Prasad Ramesh, Dean Tullsen, Daniel A. Jiménez, Gilles Pokam
Motivation

- Highly Speculative and Wide Machines prevalent

**Aggressive Front End**
- Fetch/Branch Directed Prefetching (FDIP) (Decoupled Frontend)
- Improved Branch Predictors

**Larger ROB Size**
- Intel Golden Cove – 512 entries
- AMD Zen4 – 320 entries

**Larger BTB**
- ARM Neoverse – 8K entries
- Intel Golden Cove -12K entries

**Increasing Width**
- ARM Neoverse – 16 inst/cycle → 2 taken/cycle
- 6 - wide machines most common

**Datacenter Workloads**
- Large code footprints stressing I-Cache
- Higher impact of Data dependent branches
Speculation Control

- Degree of Speculation (DoS) = \frac{\text{insts fetched}}{\text{inst committed}}
- Mean \sim 2.2X \Rightarrow \text{High overallocation during fetch}
- With FDIP enabled DoS increases by 23% on average

Simulation Setup

- Execution driven simulation on gem5
- The CPU Model has FDIP support included
- Statistics dumped every 20,000 committed instructions for fine-grained analysis
- Benchmark suites – spec17, Dacapo, Renaissance, tailbench
Dynamic Throttling

▪ How to Throttle
  ▪ Limit number of Low Confident Outstanding Branches
  ▪ Confidence Table – JRS style confidence estimator
  ▪ Important to not limit high confident branches

▪ When to Throttle
  ▪ Detect phase behavior of programs where speculation is useful
  ▪ Train a predictor whose features are the different Program Counters from different features
Dynamic Throttling

▪ How to Throttle
  ▪ Limit number of Low Confident Outstanding Branches
  ▪ Confidence Table – JRS style confidence estimator
  ▪ Important to not limit high confident branches

▪ When to Throttle
  ▪ Detect phase behavior of programs where speculation is useful
  ▪ Train a predictor whose features are the different Program Counters from different features