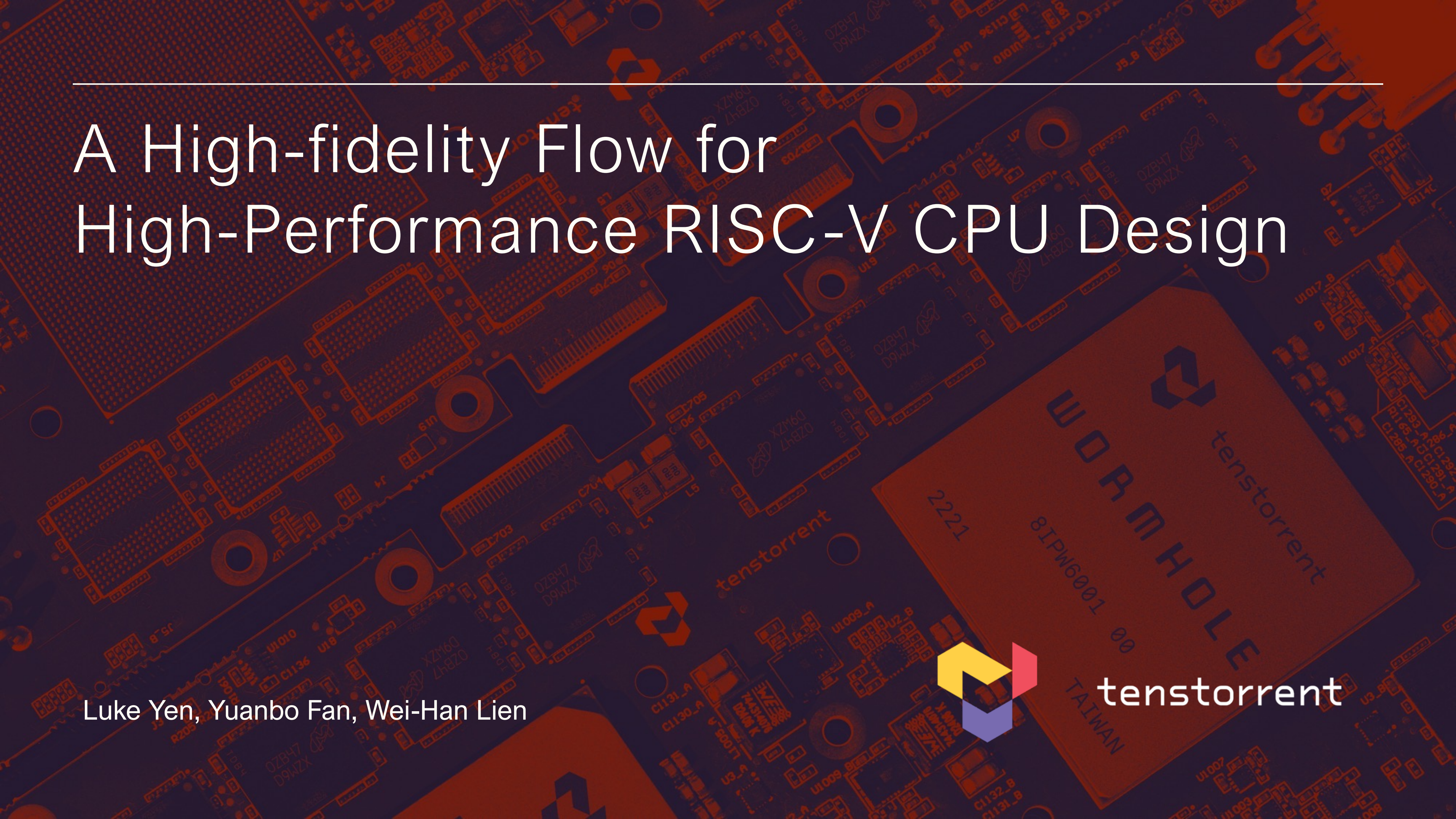


A High-fidelity Flow for High-Performance RISC-V CPU Design

Luke Yen, Yuanbo Fan, Wei-Han Lien

tenstorrent

Challenges in high-performance RISC-V CPU Design

Tedious Process

- Years of efforts
  - Design, implementation & tape-out
- Many turn-arounds
  - uArch tuning based on PPA goals
- IP progress tracking
  - Internal verification & regression
  - External visibility & validation

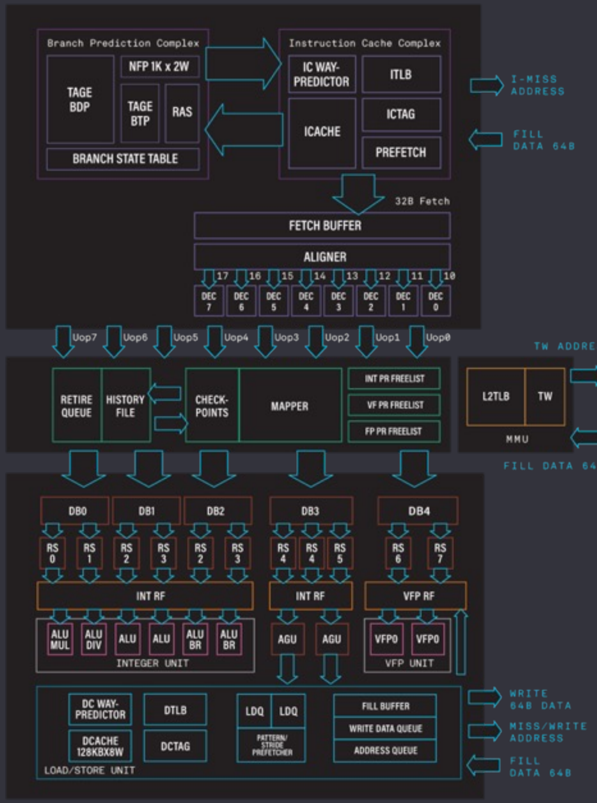
Performance Projections

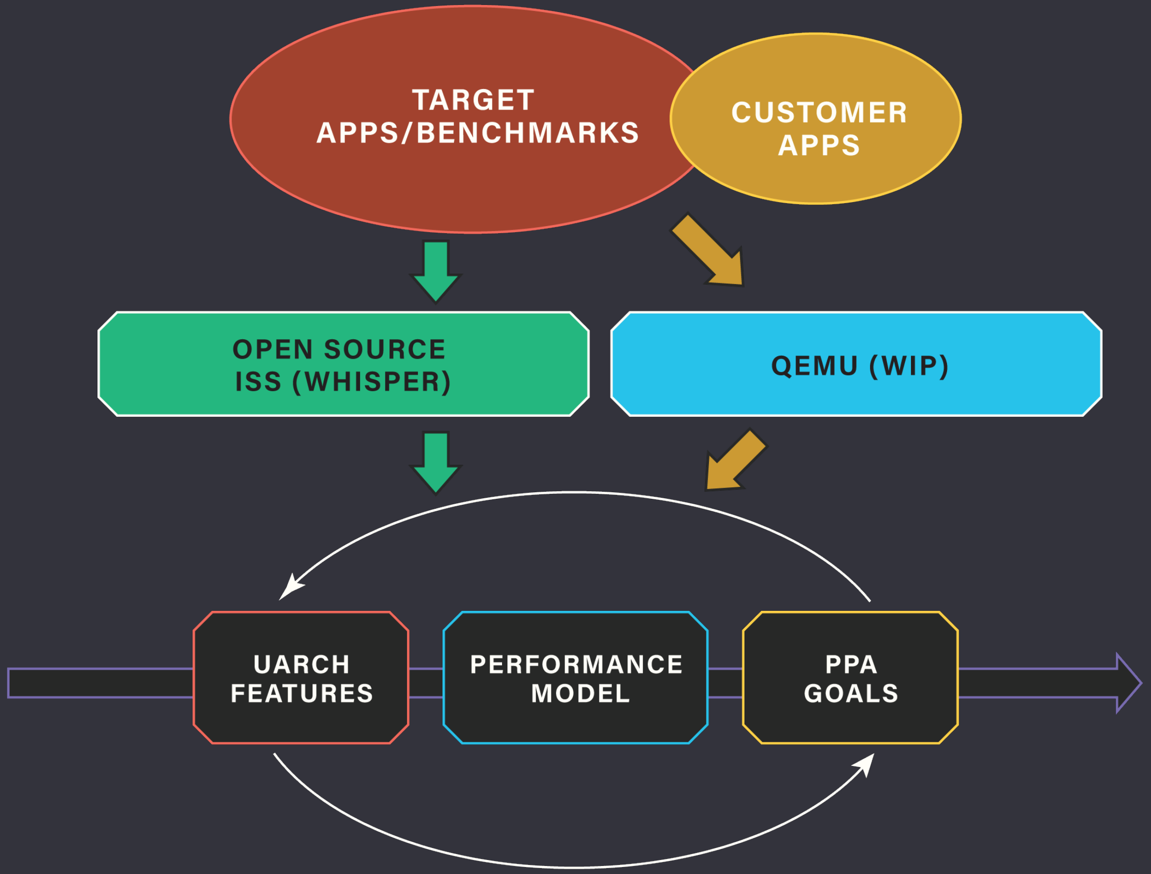
- Starting from pre-silicon stage
- Target applications/benchmarks

Collaboration

- Customized features
  - Branch predictor & prefetcher
- Sub-system
  - Cache & memory
- Vector unit design

Methodology





TARGET APPS/BENCHMARKS

CUSTOMER APPS

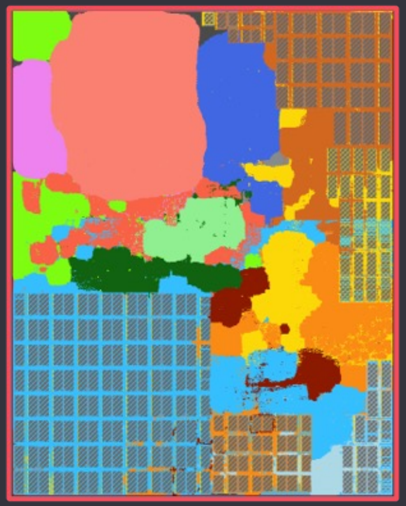
OPEN SOURCE ISS (WHISPER)

QEMU (WIP)

UARCH FEATURES

PERFORMANCE MODEL

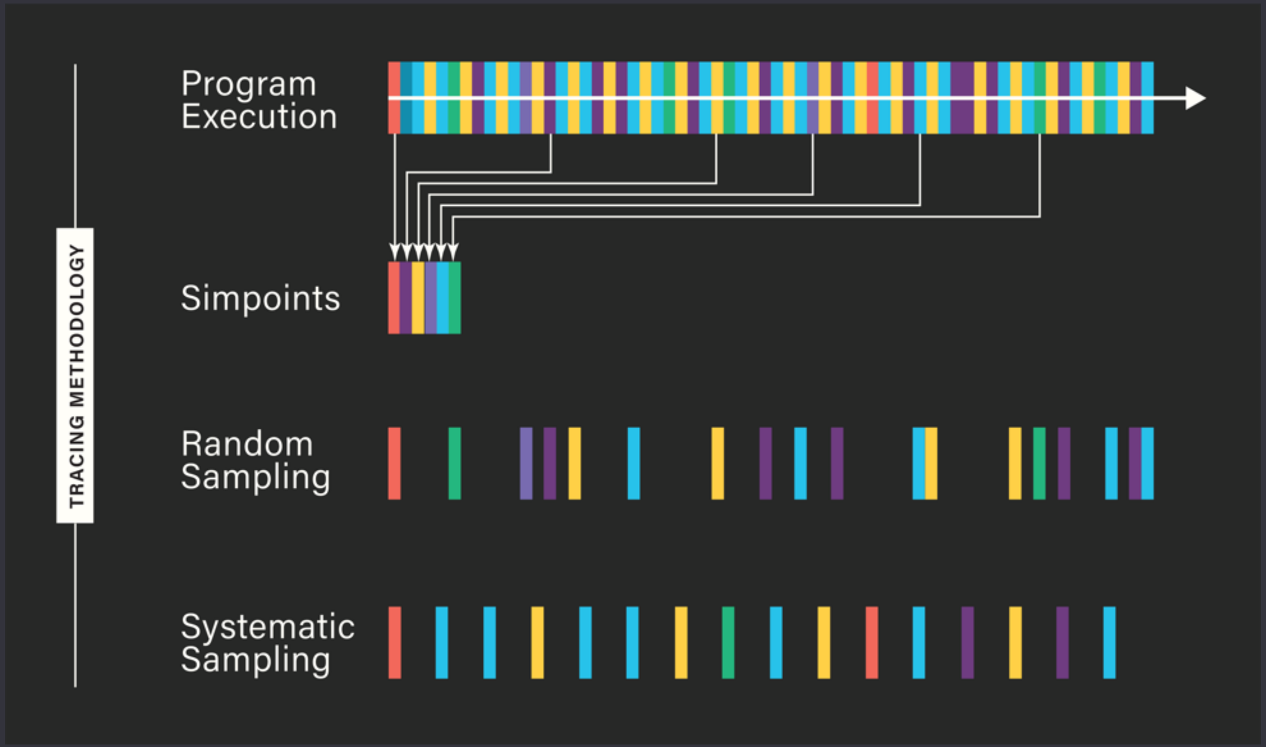
PPA GOALS



Instruction Set Simulator (i.e. Whisper)

- Available: [Github Link]

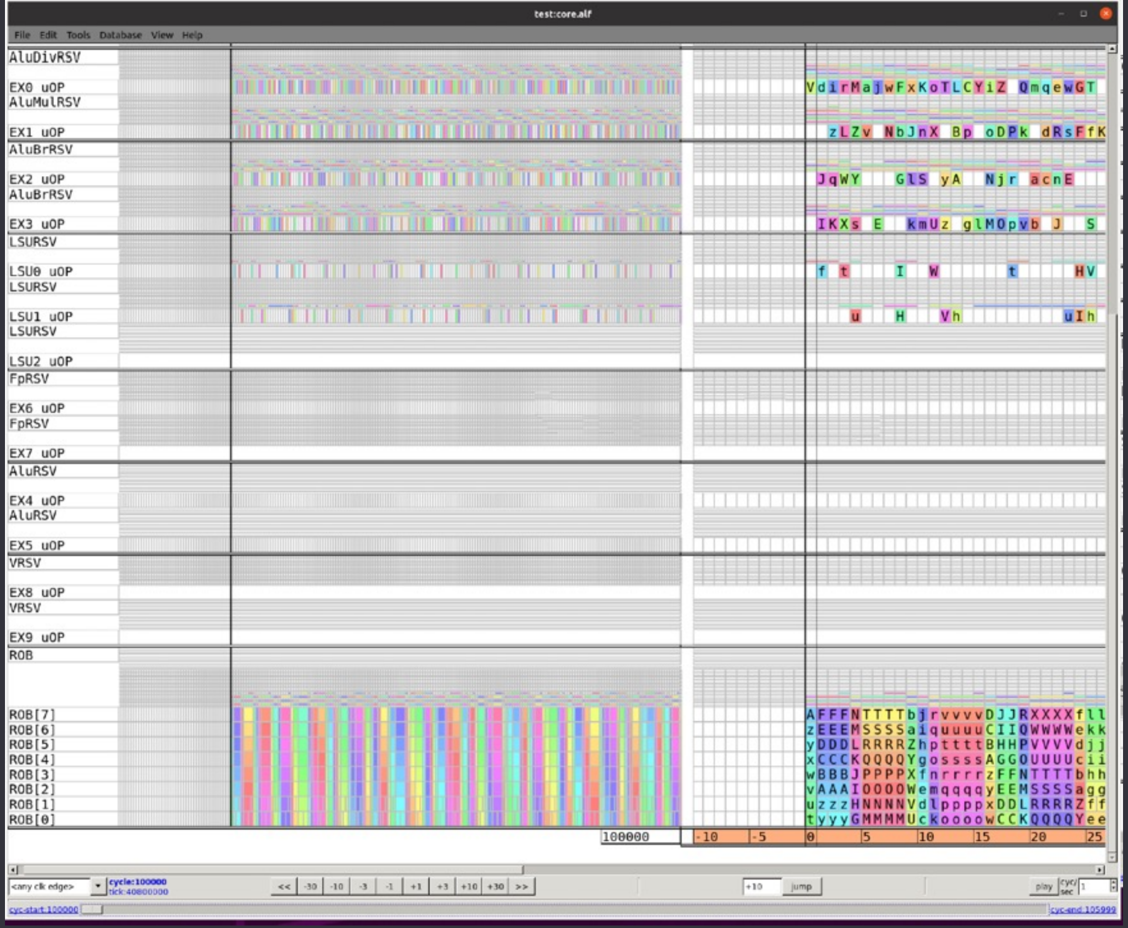
- Simulation
  - ➢ Speed: ~100M instructions per second
  - ➢ Static trace generation
  - ➢ Co-simulation (with performance model)
- Tracing Methodology
  - ✓ Random/systematic sampling
  - ✓ Representative phases (i.e. simpoints)
  - ✓ Warmup traces (for branch and I/D cache)
- Limitation
  - ❑ Static linking
  - ❑ Single-thread simulation



tenstorrent Confidential

Cycle-Accurate Performance Model

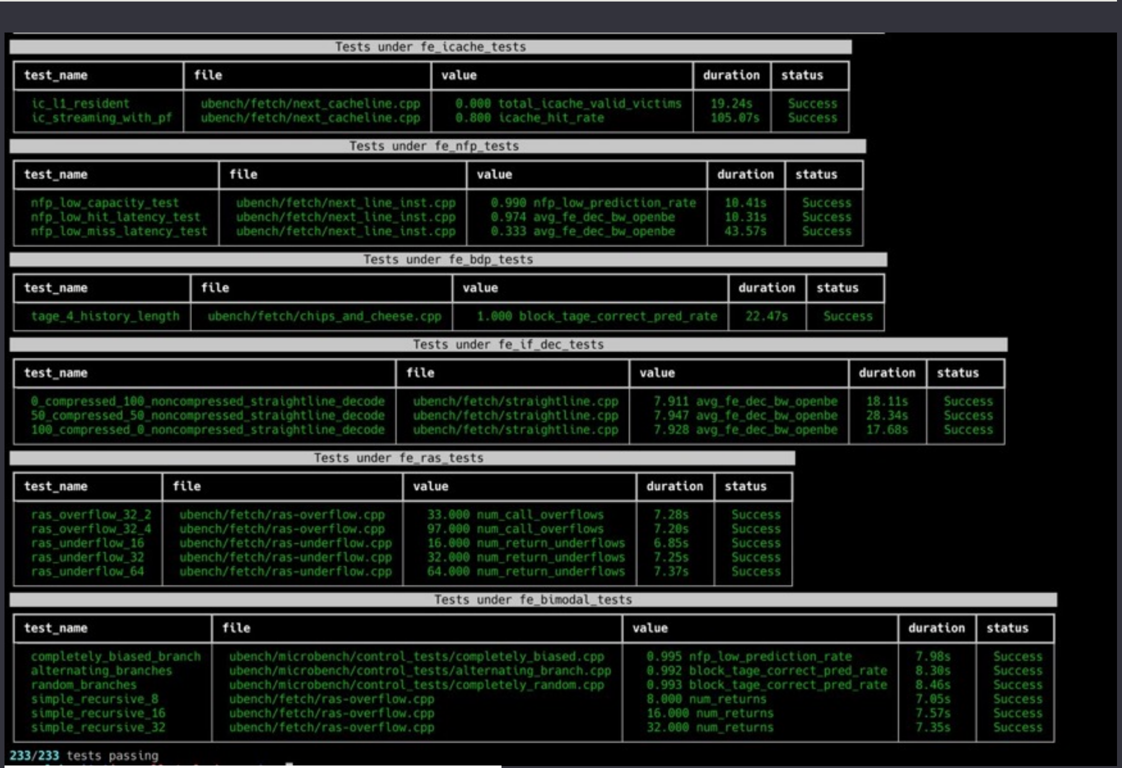
- Cycle-accurate, event-driven Model
    - Highly flexible and configurable
    - Trace-driven simulation
- Performance Metrics
    - Such as IPC, total execution cycles, cache misses, TLB misses, branch mispredicts
- Visualization & Debug
    - Pipetrace: a visual representation of the pipeline execution over time



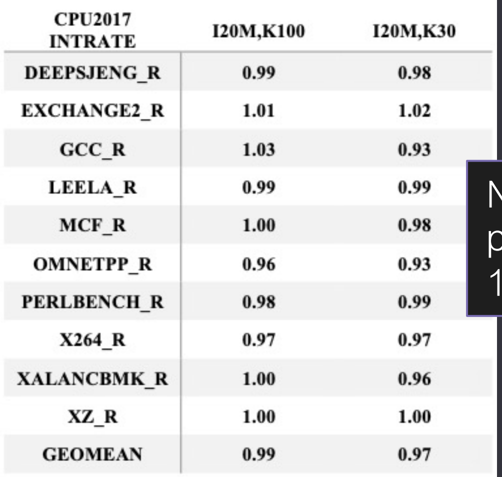
PPA-orientated Tuning

Tradeoff: Projection Accuracy vs. Simulation Speed

- The tracing flow consists of several steps including:
  - ✓ Whisper simulation
  - ✓ Benchmark profiling
  - ✓ Snapshot generation
  - ✓ Trace generation
- Customized microbenchmark
  - C/C++ tests
  - Major uarch components & timing paths
  - Critical perf. metrics (e.g. data cache hit latency, issue bandwidth)
- Calibration & Validation
  - ❖ Tradeoffs: Low projection error vs 10x simulation speedup
  - ❖ End-to-end co-simulation



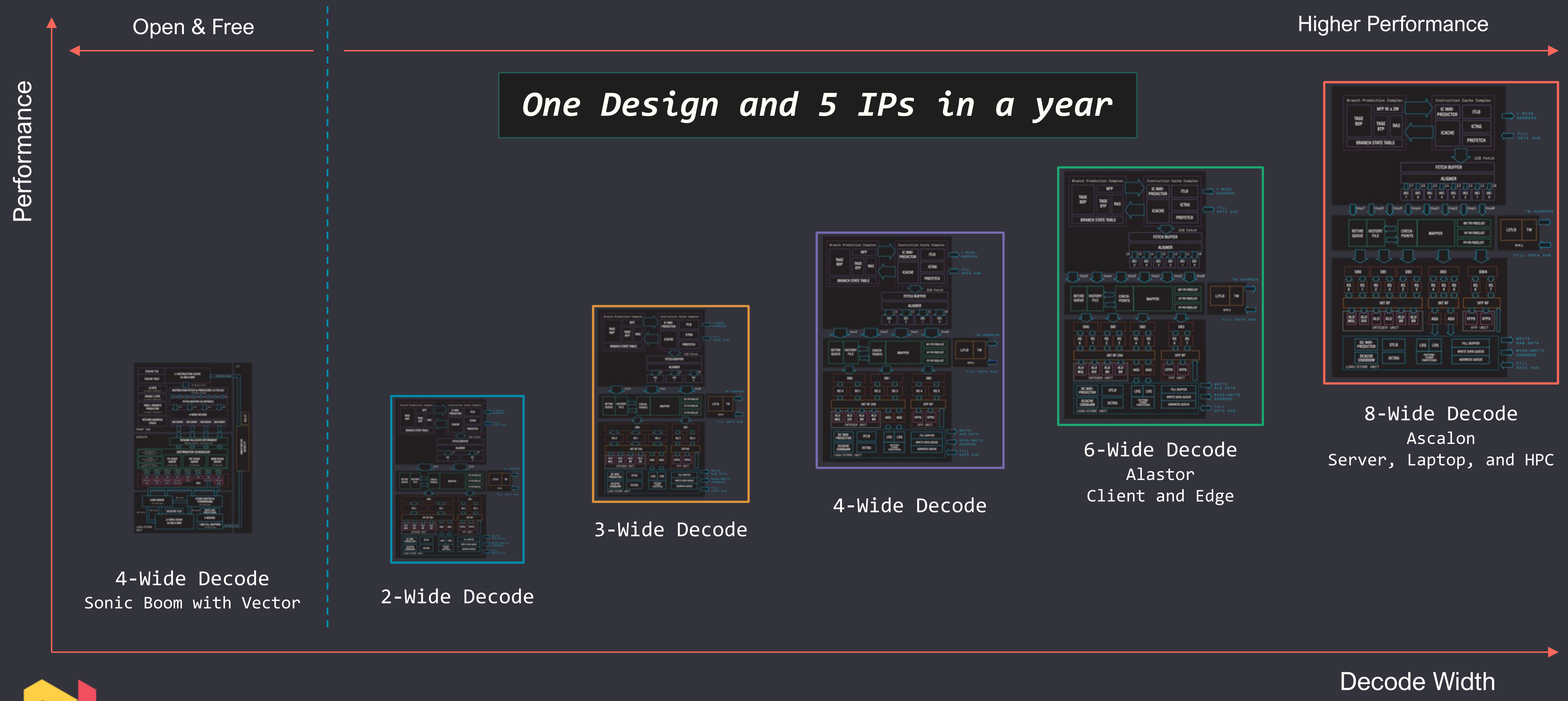
```
                              Tests under fe_icache_tests
| test_name            | file                          | value                            | duration | status  |
|  ic_l1_resident       |  ubench/fetch/next_cacheline.cpp |  0.000 total_icache_valid_victims |  19.24s   |  Success |
|  ic_streaming_with_pf |  ubench/fetch/next_cacheline.cpp |  0.800 icache_hit_rate            |  105.07s  |  Success |
                              Tests under fe_nfp_tests
| test_name                | file                          | value                        | duration | status  |
|  nfp_low_capacity_test    |  ubench/fetch/next_line_inst.cpp |  0.990 nfp_low_prediction_rate |  10.41s   |  Success |
|  nfp_low_hit_latency_test |  ubench/fetch/next_line_inst.cpp |  0.974 avg_fe_dec_bw_openbe    |  10.31s   |  Success |
|  nfp_low_miss_latency_test|  ubench/fetch/next_line_inst.cpp |  0.333 avg_fe_dec_bw_openbe    |  43.57s   |  Success |
                              Tests under fe_bdp_tests
| test_name              | file                             | value                              | duration | status  |
|  tage_4_history_length  |  ubench/fetch/chips_and_cheese.cpp |  1.000 block_tage_correct_pred_rate |  22.47s   |  Success |
                              Tests under fe_if_dec_tests
| test_name                                          | file                          | value                      | duration | status  |
|  0_compressed_100_noncompressed_straightline_decode |  ubench/fetch/straightline.cpp |  7.911 avg_fe_dec_bw_openbe |  18.11s   |  Success |
|  50_compressed_50_noncompressed_straightline_decode |  ubench/fetch/straightline.cpp |  7.947 avg_fe_dec_bw_openbe |  28.34s   |  Success |
|  100_compressed_0_noncompressed_straightline_decode |  ubench/fetch/straightline.cpp |  7.928 avg_fe_dec_bw_openbe |  17.68s   |  Success |
                              Tests under fe_ras_tests
| test_name          | file                            | value                          | duration | status  |
|  ras_overflow_32_2  |  ubench/fetch/ras-overflow.cpp   |  33.000 num_call_overflows      |  7.28s    |  Success |
|  ras_overflow_32_4  |  ubench/fetch/ras-overflow.cpp   |  97.000 num_call_overflows      |  7.20s    |  Success |
|  ras_underflow_16   |  ubench/fetch/ras-underflow.cpp  |  16.000 num_return_underflows   |  6.85s    |  Success |
|  ras_underflow_32   |  ubench/fetch/ras-underflow.cpp  |  32.000 num_return_underflows   |  7.25s    |  Success |
|  ras_underflow_64   |  ubench/fetch/ras-underflow.cpp  |  64.000 num_return_underflows   |  7.37s    |  Success |
                              Tests under fe_bimodal_tests
| test_name                | file                                                 | value                              | duration | status  |
|  completely_biased_branch |  ubench/microbench/control_tests/completely_biased.cpp   |  0.995 nfp_low_prediction_rate      |  7.98s    |  Success |
|  alternating_branches     |  ubench/microbench/control_tests/alternating_branch.cpp  |  0.992 block_tage_correct_pred_rate |  8.30s    |  Success |
|  random_branches          |  ubench/microbench/control_tests/completely_random.cpp   |  0.993 block_tage_correct_pred_rate |  8.46s    |  Success |
|  simple_recursive_8       |  ubench/fetch/ras-overflow.cpp                          |  8.000 num_returns                  |  7.05s    |  Success |
|  simple_recursive_16      |  ubench/fetch/ras-overflow.cpp                          |  16.000 num_returns                 |  7.57s    |  Success |
|  simple_recursive_32      |  ubench/fetch/ras-overflow.cpp                          |  32.000 num_returns                 |  7.35s    |  Success |
233/233 tests passing
```



CPU2017 INTRATE	I20M,K100	I20M,K30
DEEPSJENG_R	0.99	0.98
EXCHANGE2_R	1.01	1.02
GCC_R	1.03	0.93
LEELA_R	0.99	0.99
MCF_R	1.00	0.98
OMNETPP_R	0.96	0.93
PERLBENCH_R	0.98	0.99
X264_R	0.97	0.97
XALANCBMK_R	1.00	0.96
XZ_R	1.00	1.00
GEOMEAN	0.99	0.97

Note: normalized by performance projected by 100M simpoints

Tenstorrent RISC-V O-o-O Processor Family



Open & Free

Higher Performance

Performance

One Design and 5 IPs in a year

4-Wide Decode
Sonic Boom with Vector

2-Wide Decode

3-Wide Decode

4-Wide Decode

6-Wide Decode
Alastor
Client and Edge

8-Wide Decode
Ascalon
Server, Laptop, and HPC

Decode Width

Future Collaboration

- Instruction Set Simulator (Whisper)
- Target applications/benchmarks & sharable traces
- Compiler development & optimization