Design space-aware statistical simulation with machine learning

Thomas Flynn

10/7/2021

Brookhaven National Laboratory
U.S. DEPARTMENT OF ENERGY

Introduction

- Design space exploration: How to evaluate the performance of an application on many architecture configurations?
  - Use simpler surrogate applications
  - Use statistical simulation?
- Some of program intervals may not be sensitive to the architecture changes, and re-running them on each architecture may be wasteful
- Identify important program intervals for performance modeling
- Learn a mapping from interval performance to application performance

Brookhaven National Laboratory

Statistical simulation

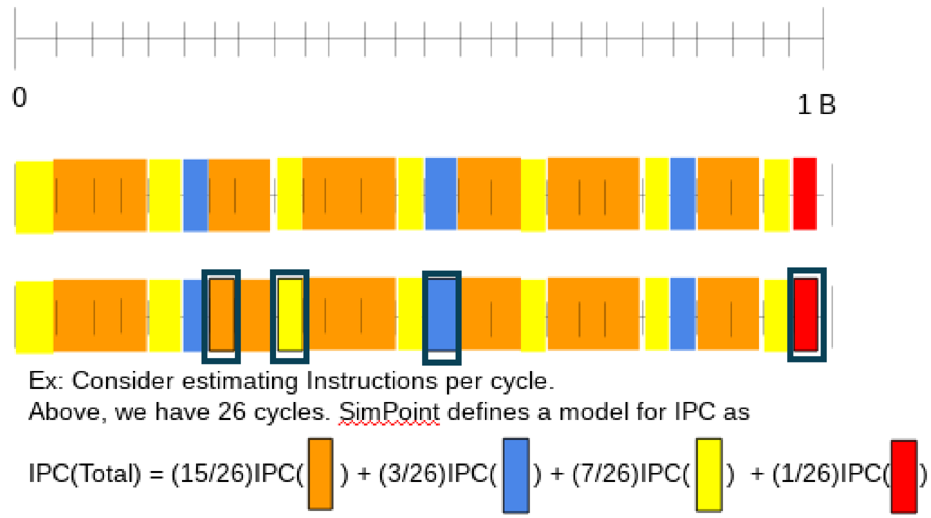
Statistical simulation speeds up simulation by only performing detailed simulation on small portions of a program and extrapolating the results for a whole-program estimate (1-2)

SimPoint divides a program into intervals and groups intervals together based on their behavior similarity

- Measured with basic block features (that is, if two intervals involve execution of many of the same basic blocks then they are similar)

Detailed simulation is performed on a representative interval of each cluster, and fast forwarding is used between these intervals.

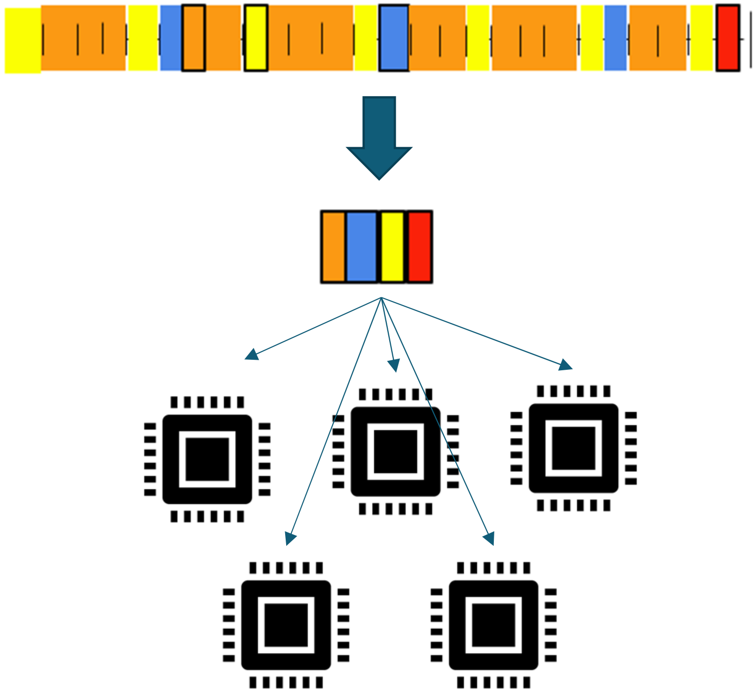
Combine measurements on representative intervals to get whole program estimate



1. T. M. Conte, M. A. Hirsch, and K. N. Menezes, “Reducing state loss for effective trace sampling of superscalar processors,” in Proceedings International Conference on Computer Design. VLSI in Computers and Processors, pp. 468–477. (1996)
2. Hamerly, G., Perelman, E., Lau, J., & Calder, B. SimPoint 3.0: Faster and More Flexible Program Phase Analysis. *J. Instr. Level Parallelism, 7*. (2005).

Brookhaven National Laboratory

Design space exploration



- A typical application of statistical simulation would be to estimate program performance over many configurations
- For each configuration, run the small number of intervals, and use a weighted sum to estimate performance
- However, this may cause some difficulties:
  - Since the intervals in a cluster are only similar, and not equal, it's not clear that a simple weighted sum is the most accurate model
  - Some intervals may not be sensitive to the design space variables

Brookhaven National Laboratory

Design-space aware statistical simulation

- Run SimPoint to obtain a set of phases
- Correlate phase performance profiles with design parameters
- Toss out uncorrelated phases
- Among "interesting" phases, build a model of phase → application performance

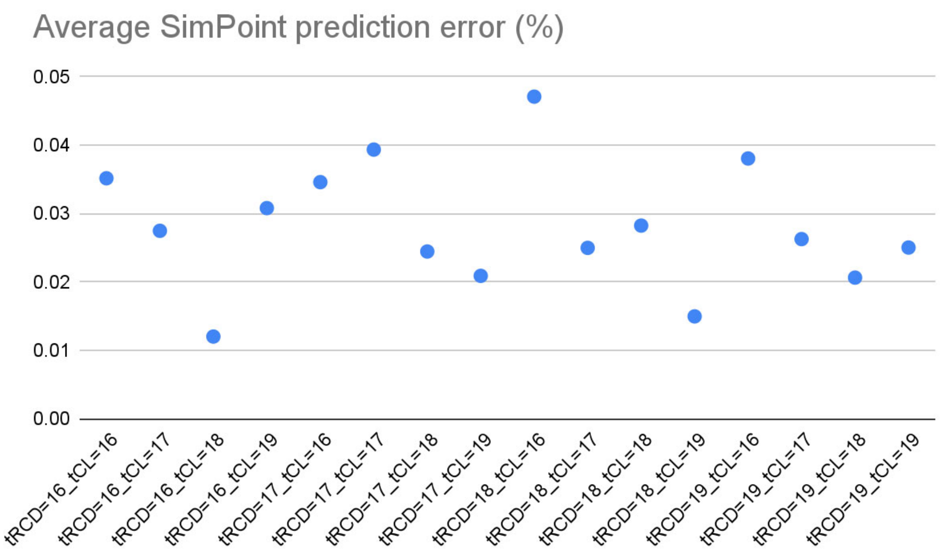
- Run SimPoint to obtain a set of **n** phases and representative intervals $\mathbf{p_1, \ldots, p_n}$.
- Choose **m** design settings $\mathbf{v_1, \ldots, v_m}$ (a small portion of overall space)
- Obtain performance **r** of the **n** intervals on each of **m** configurations.
  - Evaluate $\mathbf{r(i, v_k)}$ for **i=1…n** and **k=1…m**
- For each interval, find correlation **c(i)** between $\mathbf{[\ r(i, v_1), r(i, v_2), \ldots, r(i, v_m)\ ]}$ and $\mathbf{[\ v_1, v_2, \ldots, v_m\ ]}$
- Keep the top **k** intervals in terms of correlation, discarding the rest.
  - Represented by k indices $\mathbf{g_1, g_2, \ldots, g_k}$
- Run the application **q** at the **m** configurations, obtaining performance measurements $\mathbf{r(q, v_1), r(q,v_2), \ldots, r(q,v_m)}$
- Learn a mapping $\mathbf{f : R^k \rightarrow R}$ from the interval performances to application performances (e.g. using ML)
  - **for j=1…m** , $\mathbf{f(\ r(g_1, v_j), r(g_2,v_j), \ldots , r(g_k, v_j)\ ) = r(q,v_j)}$
- Prediction: Given a new configuration **v**, run the top **k** intervals to get performance result, then plug into ML model.
  - Evaluate $\mathbf{r(g_1,v), r(g_2,v), \ldots r(g_k,v)}$
  - Estimate runtime as $\mathbf{r(q,v) \sim f(r(g_1,v), r(g_2,v), \ldots r(g_k,v))}$

Brookhaven National Laboratory

Application to the SPEC benchmarks

Evaluation of our methodology to predict CPI in SPEC programs

- Evaluated ground truth of benchmark runtimes.
- Apply SimPoint to get a baseline prediction estimate
- Next steps are to apply DSA-SimPoint, and compare results.
- Later, consider DL-based clustering approaches, based on SimNet



Brookhaven National Laboratory