Correlation Analysis of Performance Counters in High Performance Computing

Avaneesh Ramesh[1], Joseph Sheils[2], Rachel Wang[3], Tanzima Islam[4]
[1]Westwood High School, [2]Marshall University, [3]University of California, Berkeley, [4]Texas State University

Abstract

The collection of data from hardware performance counters in High Performance Computing (HPC) systems can yield extensively large datasets, which can hinder analysis processes. To solve this issue, we propose a framework for detecting and visualizing common information among performance counters. The challenge in this instance is making the algorithm compatible to an extensive range of datasets and ensuring that the visual output is easily understandable for users. To confront this, we convert numerical output from the information sieve algorithm into a node-link networkx graph which allows the user to understand how their applications are interacting with their underlying systems.

Background

With the rise of edge devices and the large amount of data being collected, there is a need for efficient computation, pattern detection, and decision making. High-Performance Computing (HPC) can quickly process a large amount data from these edge devices. Hardware performance counters on HPC systems can help finish computation faster. However, there can be thousands of such counters, which can make the process computationally intractable. Hence, we need to automatically detect and reduce all correlated counters.

Problem

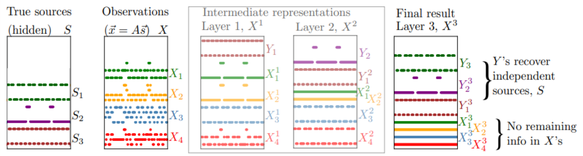
- Too many counters in HPC, thus data collection is inefficient, this process needs to be streamlined.
- Data is numerous and redundant, finding correlation allow for unique sets to represent all of the data.

Approach

- For each application:
  1. Collect extensive hardware performance counters
  2. Correlate performance counters so that users can monitor a reduced set of performance events => reducing data collection time.
- Information sieve – An algorithm to correlate performance counters by extracting common information among them.
- Graph network visualization: Visualize correlations and communities as a graph.

Framework

Aggregation of common information using the Information Sieve algorithm:



- The figure above visually describes how the Information Sieve algorithm aggregates common information among sources (X) which are described by latent factors (Y).
- The learning process is unsupervised, since no label information is used.

The multivariate mutual information first introduced as "total correlation":

$$TC(X) \equiv D_{KL}\left(p(x)||\prod_{i=1}^{n} p(x_i)\right) = \sum_{i=1}^{n} H(X_i) - H(X)$$

The reduction in multivariate information in X after conditioning on Y:

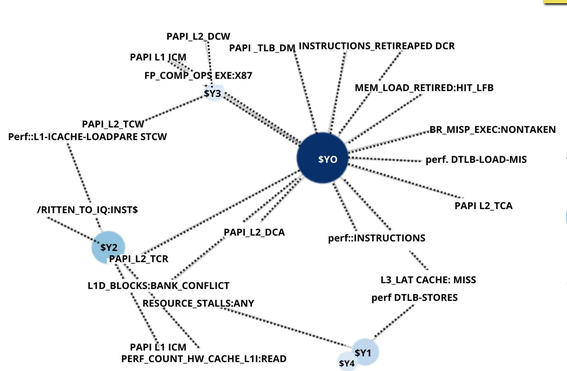
$$TC(X;Y) \equiv TC(X) - TC(X|Y) = \sum_{i=1}^{n} I(X_i;Y) - I(X;Y).$$

Our correlation analysis script outputs latent factor groups (e.g., Y_0), and the performance counters. It also describes the total correlation for each latent factor and each counter's contribution in that group.

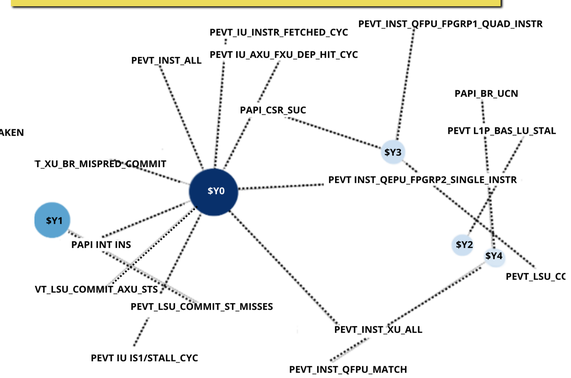
Y_0: 18.207
GRBM_COUNT: 0.938
GRBM_GUI_ACTIVE: 0.938
GPUBusy: 0.039
MemUnitBusy: 0.789

This output file serves as an input to our visualization script, seen in "Preliminary Result"

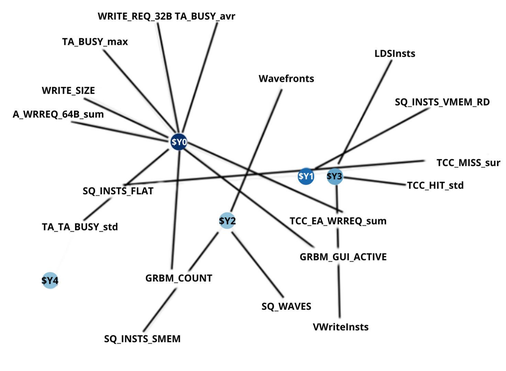
Preliminary Result



OpenMC on Cab



OpenMC on Vulcan



RNN on AMD POD

Experimental Setups:

- Applications: OpenMC: nuclear reactor simulation code; Recurrent Neural Network (RNN): Machine Learning model used to identify patterns from sequence data.
- Systems: Vulcan: IBM BG/Q supercomputer at LLNL; Cab: Intel Sandy Bridge supercomputer at LLNL; AMD POD - AMD COVID-19 HPC cluster at TxState with MI-50 GPUs.

Observation:

- A large number of performance counters are correlated, hence selecting a few can sufficiently to convey the same information.
- Graph network visualization is effective in uncovering correlated counters. The outcome of our research can inform users to select a small subset of counters (e.g., 5 instead of 50) to measure and predict the execution time of an application.
- **Our work can be applied to a large variety of data analytics problems since the process is agnostic of the names and meanings of the features (X).**

Future Plan

- Make the visuals interactive
- Analyze more applications

References

- Ver Steeg, G., & Galstyan, A. (2016, June). The information sieve. In *International Conference on Machine Learning* (pp. 164-172). PMLR.