Perspector: Benchmarking Benchmark Suites

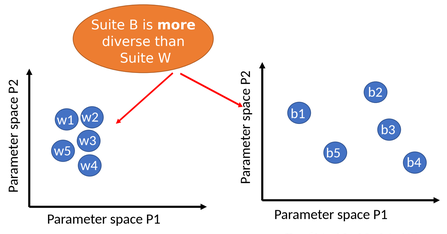
Sandeep Kumar, Abhisek Panda, and Smruti R. Sarangi
Indian Institute of Technology Delhi

❶ Contribution:

We present Perspector, a novel way to benchmark diverse benchmark suites using rigorously mathematically defined scores that capture their characteristics. The scores obtained thus are aligned with what others have qualitatively observed. These scores can be used to compare benchmark suites and select the best one for a given purpose, and they can also be used to select a subset of the workloads from a suite.

❷ Aim: Use as few benchmarks as possible to evaluate a custom system.

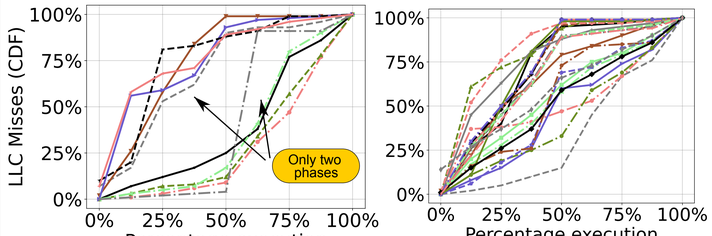
1. Workloads should be diverse



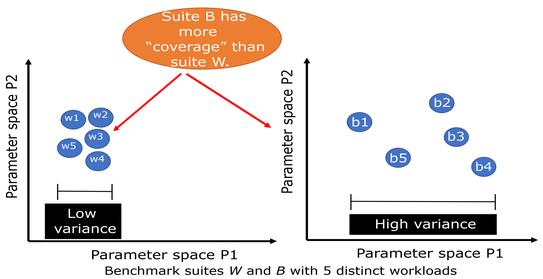
$W = \{w1, w2, w3, w4, w5\}$ $B = \{b1, b2, b3, b4, b5\}$

Benchmark suites W and B with 5 distinct workloads

2. Workloads should have distinct phases

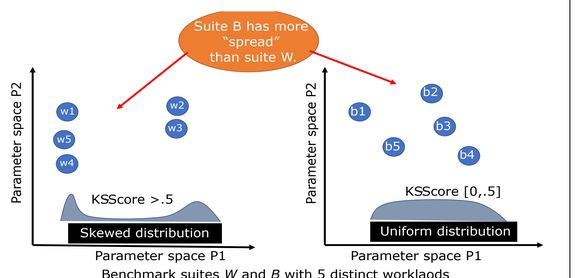


3. Workloads have a wide coverage



Benchmark suites W and B with 5 distinct workloads

4. Workloads should be uniformly spread out



Benchmark suites W and B with 5 distinct worklaods

❸ Scoring

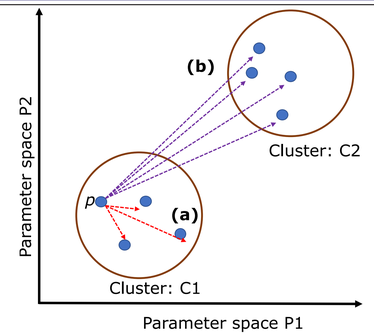
Cluster Score

Silhouette score (SS) is a measure of the quality of clusters.

$$SS = \frac{(\text{inter\_cluster\_dist} - \text{intra\_cluster\_dist})}{\max(\text{inter\_cluster\_dist}, \text{intra\_cluster\_dist})}$$

intra_cluster_dist is the average distance within clusters (a).
inter_cluster_dist is the average distance between clusters (b).
[-1 to 1] with -1: bad cluster and 1: good cluster

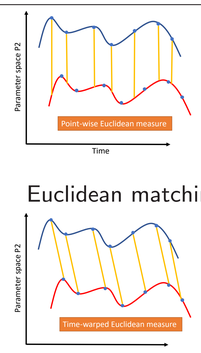
Low score is desired



Trend Score

- Dynamic Time Warping (DTW) compares two time-series sequences that may be shifted or dilated w.r.t each other.
- Optimal alignment by warping one of the sequences in the time domain

High score is desired



Coverage & Spread Score

1. **Variance $\propto$ coverage score**
   High score is desired

2. Just variance can be misleading. **Spread** should be close to uniform.

- KS (Kolmogorov-Smirnov) [0, 1] score is used to measure this.
- Measures the distance from the uniform distribution.

Low score is desired

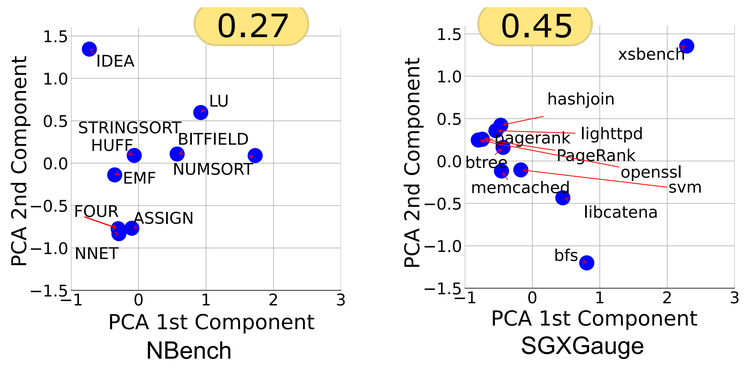
❹ Perspector Evaluation

Suites		
PARSEC	SPEC'17	Ligra
LMbench	Nbench	SGXGauge

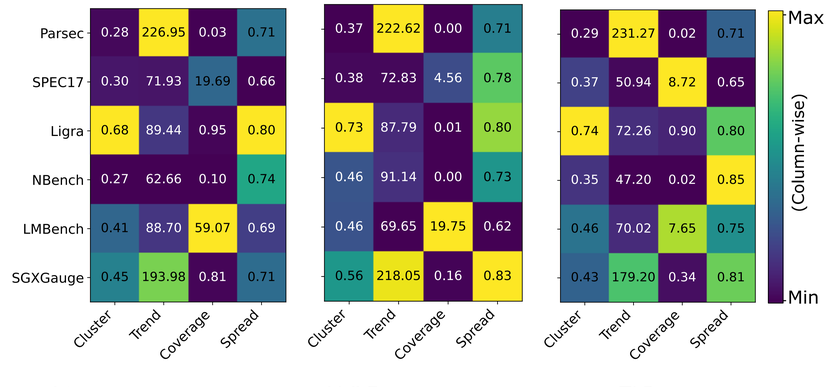
Hardware Settings
Xeon E-2186G CPU, 3.80 GHz, CPUs:6 Cores, 2 HT
L1: 384 KB, L2: 1536 KB, L3: 12 MB
DRAM: 32 GB, Disk: 1 TB (HDD)

Hardware Counters	
cpu-cycles	branch-instructions
dtlb_{load|store}_misses.walk_pending	
stalls_mem_any	page-faults
dTLB-loads dTLB-stores	dTLB-misses
LLC-loads LLC-stores	LLC-misses

Testbed: Benchmark suites, hardware settings, and hardware counters.



Nbench vs SGXGauge for cluster score: Nbench outperforms SGXGauge in terms of cluster score. This indicates that Nbench is more widespread and SGXGauge tends to cluster together.



All events LLC Events TLB events

Results: Perspector scores for different workloads with (a) all the hardware counters, (b) only LLC events, and (c) only TLB events. The scores are consistent with the characteristics of the workloads.