



Comparison of the SPEC CPU Benchmarks with 499 Other Workloads using Hardware Counters

dr. Lodewijk Bonebakker

Computer Architecture and Performance Group
Sun Microsystems Laboratories

Outline of this talk

- Context
 - representativeness of benchmarks
 - research question
 - some background theory
- Methodology
 - data collection
 - data reduction
 - analysis
- Results
 - workload set composition
 - representativeness

Context

- Relevance of SPEC CPU
 - > marketplace relevance compels chip vendors to demonstrate good results,
 - > has become the de-facto standard benchmark suite for processor architecture evaluation,
 - > heavily used in simulators evaluating future processor designs, thus greatly influencing these future processor architectures.
- But...
 - > composition based on solicited input, final selection committee decision,
 - > focused on compute intensive applications,
 - > little quantitative evaluation of representativeness.

Research question

- Paper addresses two questions:
 - > how can we efficiently and quantitatively evaluate representativeness of a benchmark set relative to a set of real workloads
 - > using such a methodology, how does SPEC CPU do?
- Relevance:
 - > an efficient quantitative method can be used to improve benchmark set composition by comparing benchmarks to *real* workloads,
 - > redundancy in the benchmark set can be avoided.

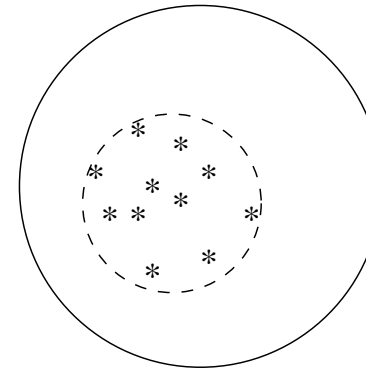
Some theoretical background

- Benchmark sets drive simulators
 - > should be a good functional representative for a given universe of real workloads
 - > should yield the same distribution of system resource utilization as real workloads

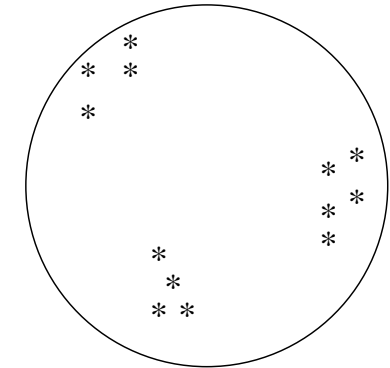
- Central characteristics
 - > size, completeness, density, granularity and redundancy. (dujmovic 2001)

- Representative set:
 - > full coverage, captures diversity, while non-redundant and complete.

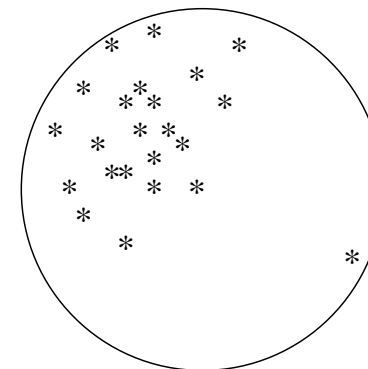
Insufficient size



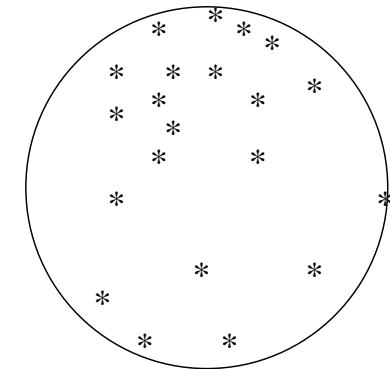
Excessive redundancy



Workload outlier



Non-uniform distribution



Methodology

- Quantitatively expressing workload representativeness
 - > express workloads in collectable metrics,
 - > span a workload space using these metrics,
 - > if needed, reduce dimensionality,
 - > define representativeness as close Euclidean proximity within the spanned workload space.
- This is an established method:
 - > reducing benchmark set similarity in simulation: Eeckhout (2005a,b)
 - > applied to SPEC CPU 2000 in simulation only,
 - > we extend this to all workloads by using hardware counter sampling as input metrics.

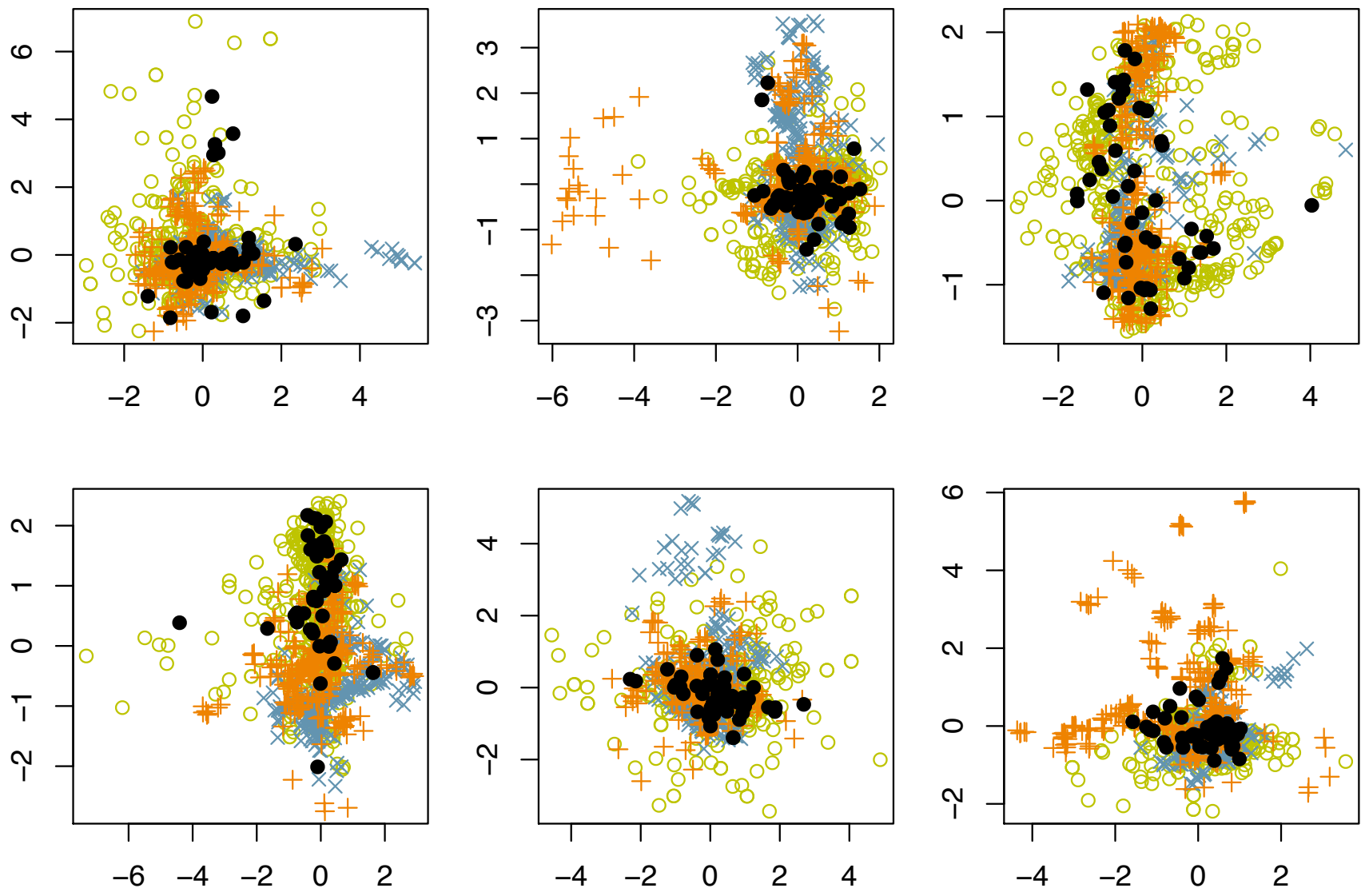
Recipe

- Collect hardware counters for large number of workloads
- Reduce workloads into a single number representation per hardware counter metric
- Span workload space with all hardware counters.
- Reduce dimensionality using *ICA*.
- Determine representativeness from distribution along principal axes.
- (Choose optimal benchmark set)

Practical application

- 1089 workloads collected
 - > 260 SPEC CPU2000 (base, peak, 1, 2, 4, 8, 12 cpu)
 - > 330 SPEC CPU2006 (idem)
 - > 447 commercial workloads
 - > 52 other benchmarks
 - > each collection took 1800 seconds
- 900 MHz UltraSPARC III+ processors, 8MB L2
- 1-64 processor systems
- workloads normalized to a single processor representation before comparison (bonebakker, 2007)

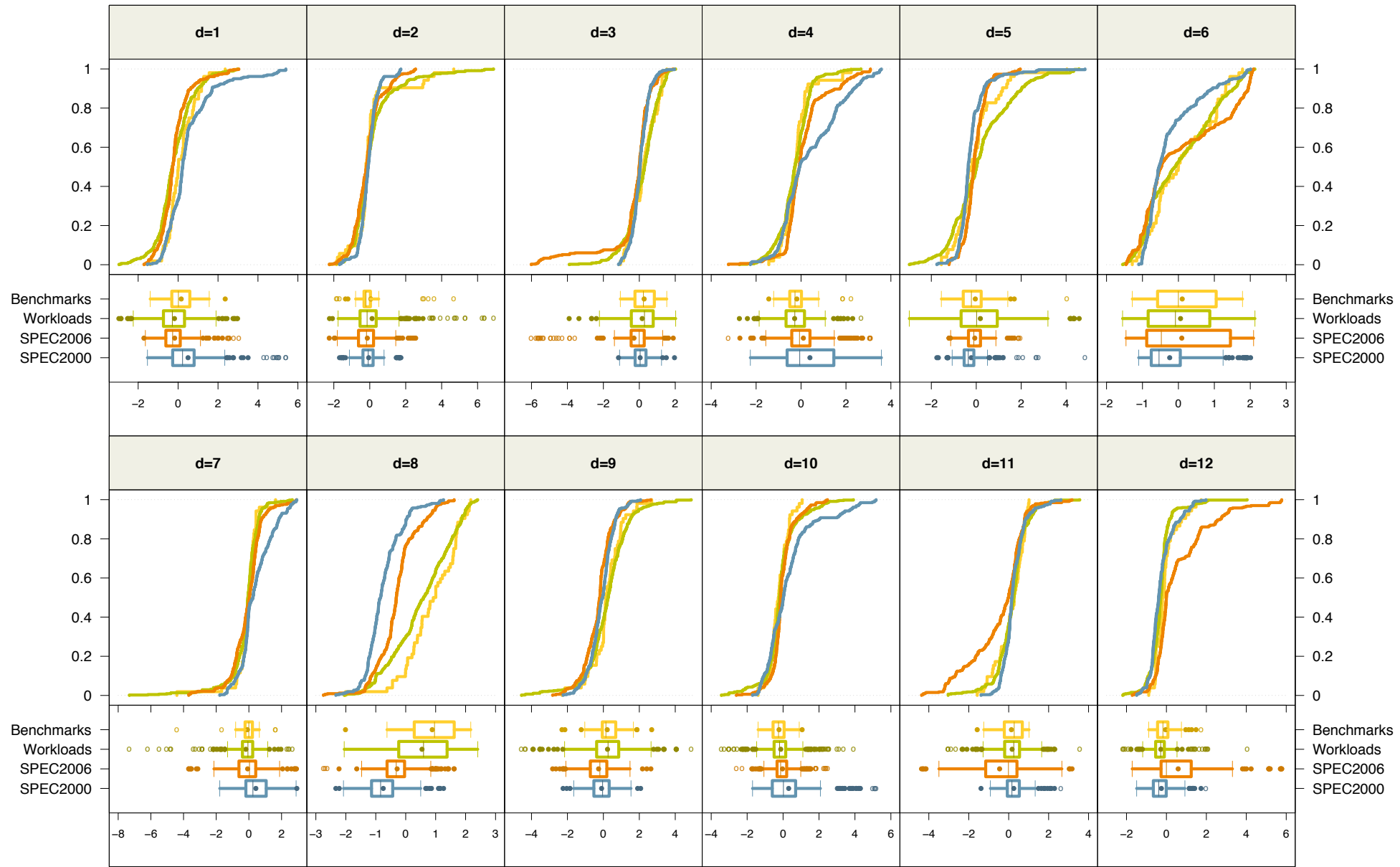
Results - workload space



○ workloads • benchmarks × SPECcpu 2000 + SPEC CPU 2006

Results - distribution

A comparison of SPEC CPU with workloads and benchmarks



Practical application - conclusions

- SPEC CPU does not provide full representativeness along all 12 independent axes
 - > commercial workloads have coherence traffic and more I/O
- SPEC CPU 2006 and 2000 also differ significantly
 - > six independent axes are different
- The set of collected additional benchmarks is more representative of the collected workload set.
 - > bootstrap indication of method applicability

Conclusions

- Processor hardware counters allow for quantitative comparison of workloads.
- Processor hardware counters are quite efficient and enable analysis of real workloads in their natural habitat.
- Workload similarity can be used to quantitatively select an optimal set of representative benchmarks.
 - > reduce redundancy
 - > achieve optimal coverage of target domain
 - > quantitative evaluation of different architectures



lodewijk.bonebakker@sun.com