# Statistical Performance Comparisons of Computers

**Tianshi Chen**[1], Yunji Chen[1], Qi Guo[1], Olivier Temam[2], Yue Wu[1], Weiwu Hu[1]

[1]State Key Laboratory of Computer Architecture,
Institute of Computing Technology (ICT), Chinese Academy of Sciences,
Beijing, China

[2]National Institute for Research in Computer Science and Control (INRIA),
Saclay, France

HPCA-18, New Orleans, Louisiana

Feb. 28th, 2012

# Outline

# Performance comparisons of computers: the tradition

- We need...
  - A number of benchmarks (e.g., SPEC CPU2006, SPLASH-2)
  - Basic performance metrics (e.g., IPC, delay)
  - Single-number performance measure (e.g., geometric mean; "War of means" [Mashey, 2004]
- The danger
  - Performance variability of computers
  - Example
    - 10 subsequent runs of SPLASH-2 on a commodity computer
    - Geometric mean performance speedups, over an initial baseline run are 0.94, 0.98, 1.03, 0.99, 1.02, 1.03, 0.99, 1.10, 0.98, 1.01
  - **Deterministic trend** *vs.* **Stochastic fluctuation**
  - **We need to estimate the confidence/reliability of each comparison result!**

# An example

- Quantitative performance comparison: estimating the performance speedup of computer "PowerEdge T710" over "Xserve" (using SPEC CPU2006 data collected from SPEC.org)
- Speedup obtained by comparing their geometric mean SPEC ratios: 3.50
- Confidence of the above speedup, obtained by our proposal: 0.31 (*If we don't estimate the confidence, we would not know that the comparison result is rather dangerous*)
- Speedup obtained by our proposal: 2.23 (with the confidence 0.95)

# Performance comparisons of computers: the tradition

- Traditional solutions: basic *parametric* statistical techniques
  - Confidence Interval
  - $t$-test [Student (W. S. Gosset), 1908]
- Preconditions
  - Performance measurements should be *normally-distributed*
  - Otherwise, number of performance measurements must be *large enough* [Le Cam, 1986]
    - **Lindeberg-Lévy Central Limit Theorem**: let $\{x_1, x_2, \ldots, x_n\}$ be a size-$n$ sample consisting of $n$ measurements of the same non-normal distribution with mean $\mu$ and finite variance $\sigma^2$, and $S_n = (\sum_{i=1}^{n} x_i)/n$ be the mean of the measurements (i.e., sample mean). When $n \to \infty$ ,

$$\sqrt{n}\,(S_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{1}$$

    - **Our practices**: **20–30** benchmarks (e.g., SPEC CPU2006), each is run for **3 (or fewer)** times

# Another example

- Consider ratios of two commodity computers $A$ (upper) and $B$ (lower) on SPECint2006 (collected from SPEC.org)

| perlbench | bzip2 | gcc | mcf | gobmk | hmmer | sjeng | libquantum | h264ref | omnetpp | astar | xalancbmk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26.4 | 20.4 | 22.2 | 45.4 | 25.4 | 34.5 | 27.0 | **633.6** | 46.0 | 22.8 | 21.7 | 30.1 |
| 13.9 | 9.47 | 10.0 | 14.3 | 10.3 | 8.31 | 9.28 | 13.91 | 19.8 | 9.08 | 8.43 | 14.7 |

- Intuitive observation: $A$ beats $B$ on all 12 benchmarks
- Paired $t$-test: at the confidence level $\geq 0.95$, $A$ does not significantly outperform $B$!
- Reason: $t$-statistic is constructed by the sample mean and the variance .

# Another example

Why?

- $t$-statistic is constructed by the sample mean and the variance. The shape of a non-normal and skewed distribution will be stretched if we consider it to be normal.
- The performance score of $A$ is incorrectly considered to obey the normal distribution $\mathcal{N}(79.63, 174.67^2)$ $(79.63 \pm 174.67)$. In other words, the performance score of $A$ takes a large probability to be negative !
- But in fact, the performance scores of $A$ are in the interval $(20, 634)$.

# Another example

- Consider ratios of two commodity computers $A$ (upper) and $B$ (lower) on SPECint2006 (collected from SPEC.org)

| perlbench | bzip2 | gcc | mcf | gobmk | hmmer | sjeng | libquantum | h264ref | omnetpp | astar | xalancbmk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26.4 | 20.4 | 22.2 | 45.4 | 25.4 | 34.5 | 27.0 | **633.6** | 46.0 | 22.8 | 21.7 | 30.1 |
| 13.9 | 9.47 | 10.0 | 14.3 | 10.3 | 8.31 | 9.28 | 13.91 | 19.8 | 9.08 | 8.43 | 14.7 |

- Paired $t$-test: at the confidence level $\geq 0.95$, $A$ does not significantly outperform $B$!
- In practice, parametric techniques are quite vulnerable to performance outliers which apparently break the normality
- Performance outliers are common (e.g., specialized architecture performing very well on specific applications)!

# Outline

# Settings

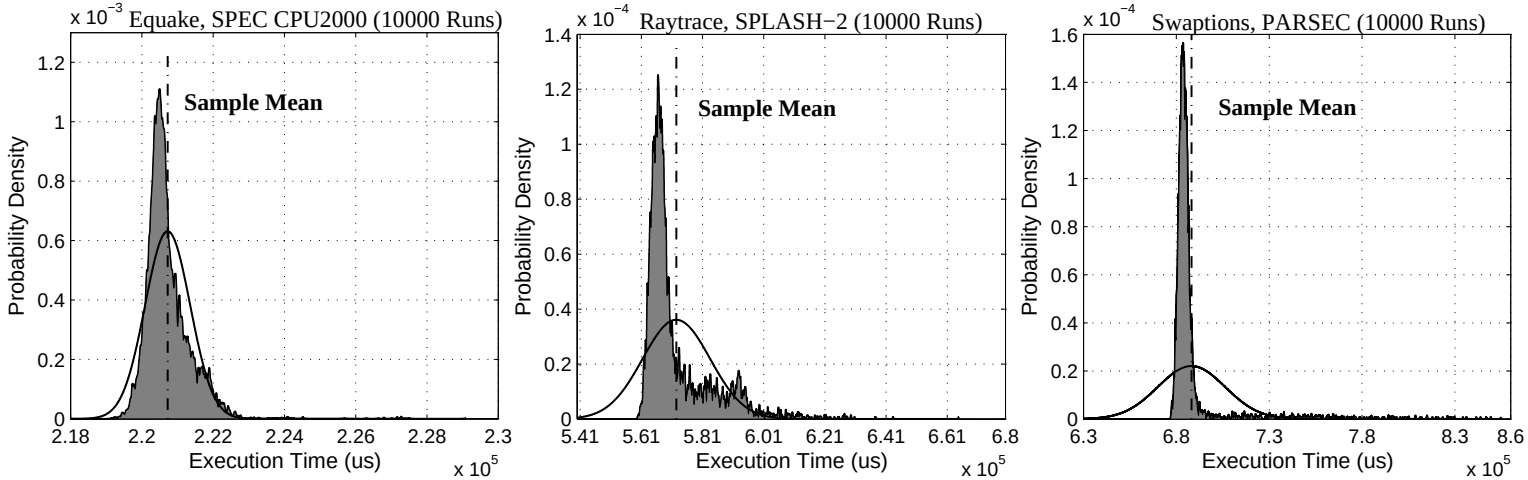
- Commodity computers
  - Intel i7 920 (4-core 8-thread), 6 GB DDR2 RAM, Linux OS
  - Intel Xeon dualcore, 2GB RAM, Linux OS
- Benchmarks
  - SPEC CPU2000 & CPU2006
  - SPLASH-2, PARSEC
  - KDataSets (MiBench) [Guthaus et al., 2001; Chen et al., 2010]
- Online repository of SPEC.org

# We need to study...

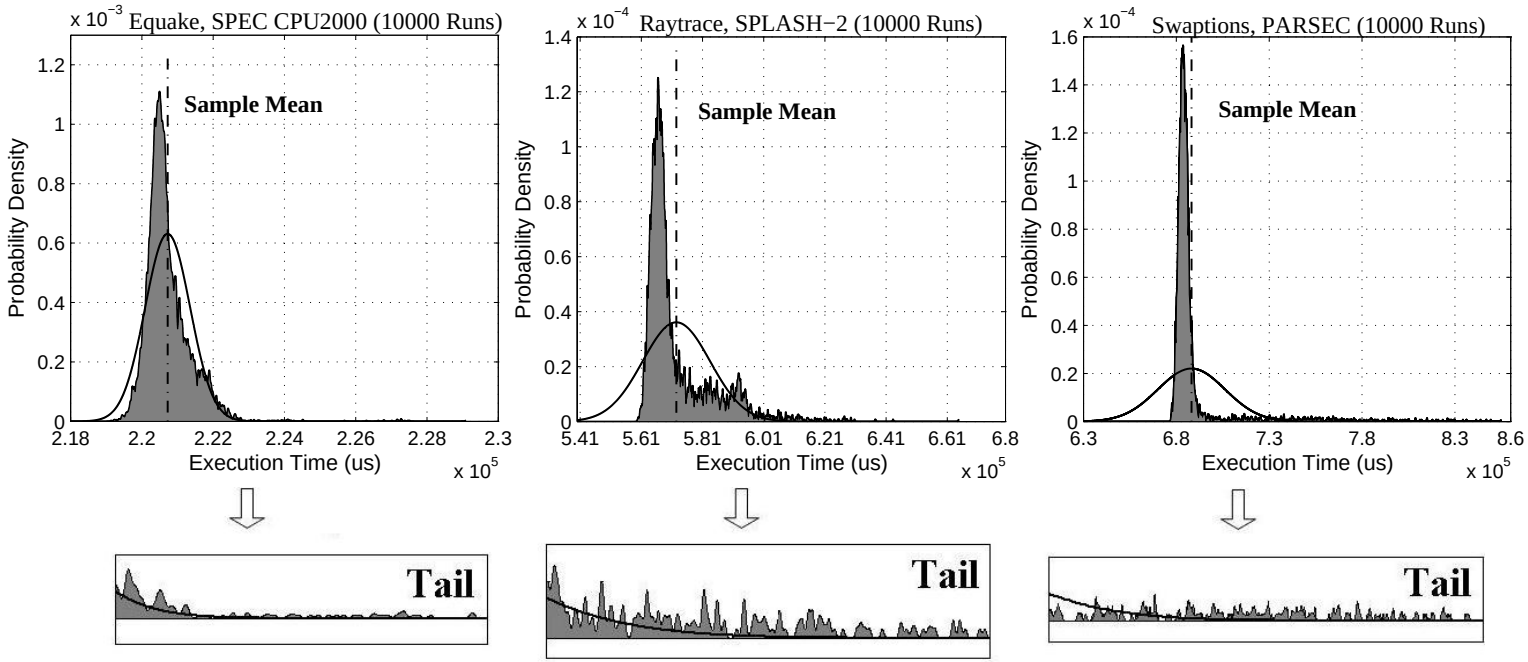
- Do performance measurements distribute normally?
- If not, whether the common number of performance measurements is *large enough* (for making the Central Limit Theorem applicable)?
- If we get two "No" above, how to carry out performance comparisons?

# Do performance measurements distribute normally?

- Naive Normality Fitting (NNF) assumes that the execution time distributes normally, and estimates the normal distribution
- Kernel Parzen Window (KPW) [Parzen, 1962] directly estimates the real distribution of execution time
- If KPW-curve≠NNF-curve, then the execution time does not obey a normal law



# Do performance measurements distribute normally?




- Long tails, especially for multi-threaded benchmarks. The distributions of the execution times on *Raytrace* and *Swaptions* seem to be *power-law*.
- It's hard for a program (especially multi-threaded ones) to execute faster than a threshold, but easy to be slowed down by, for example, data races, thread scheduling, synchronization order, and contentions of shared resources.

# Do performance measurements distribute normally?

Do cross-benchmark performance measurements distribute normally?

- CPU2006 data of 20 computers collected from from SPEC.org
- Statistical normality test
- At the confidence level of 0.95, the answer is *significantly* "No" to all 20 computers over SPEC CPU2006, 19 out of 20 over SPECint2006, 18 out of 20 over SPECfp2006

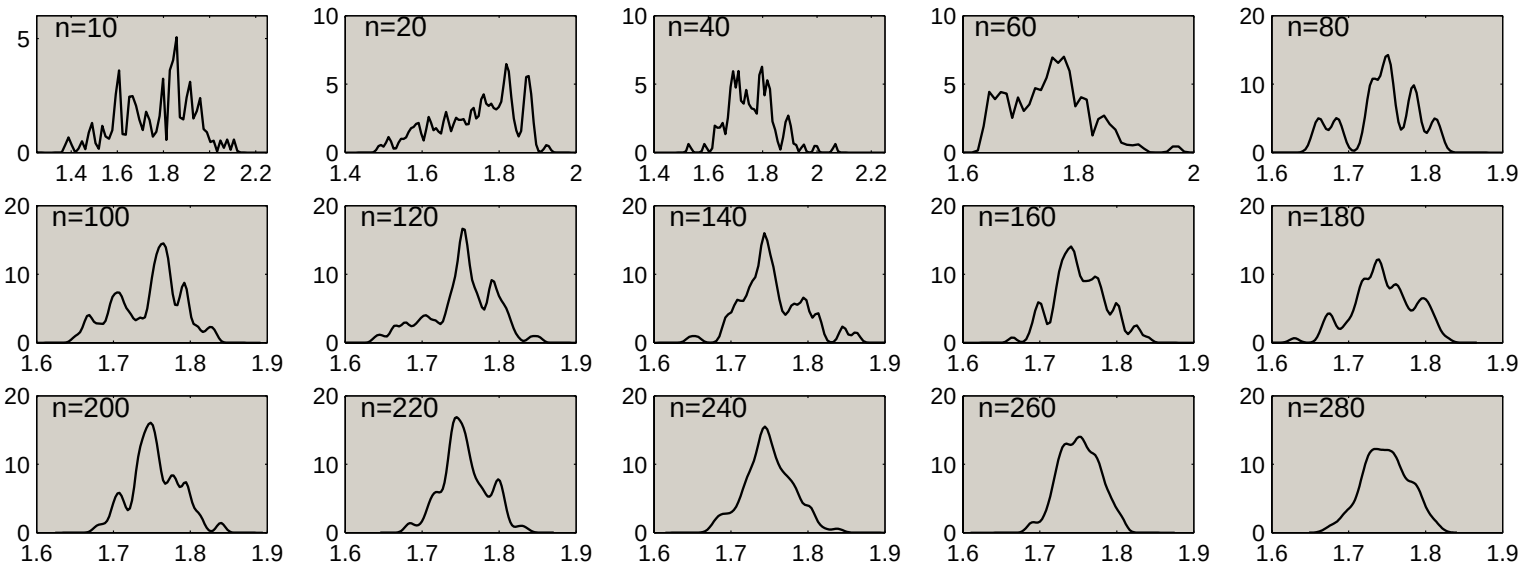
# Whether the Central Limit Theorem (CLT) is applicable?

Briefly, the CLT states that the mean of a sample (with a number of measurements) distributes normally when the sample size (number of measurements in the sample) is sufficiently-large .

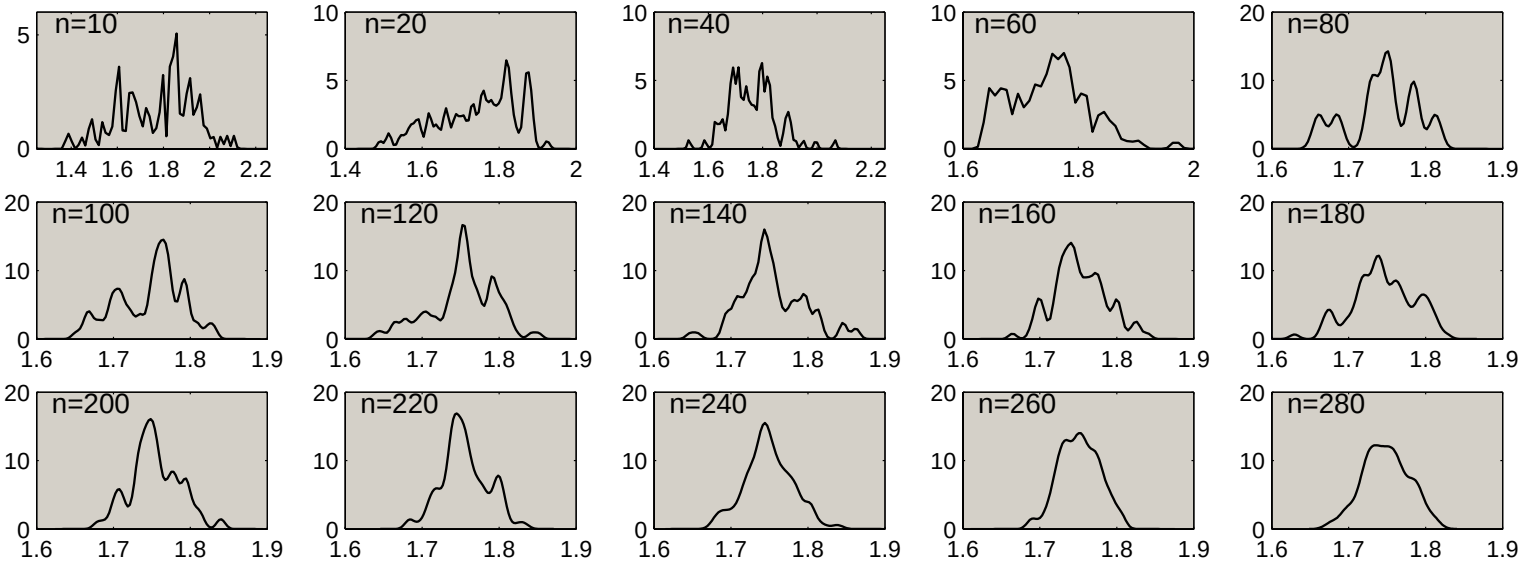
- How large is "sufficiently-large"?
- Empirical study on performance data w.r.t. KDataSets
  - 32,000 different combinations of benchmarks and data sets (thus 32,000 IPC scores) are available
  - Randomly collect 150 samples from the 32,000 scores, each consists of $n$ randomly selected scores
  - 150 observations of the sample mean have been enough for exhibiting the normality (if the normality holds)
  - The sample size $n$ is set to 10, 20, 40, 60, . . . , 240, 260, 280 in 15 different trials, respectively

# Whether the Central Limit Theorem (CLT) is applicable?

By the Kernel Parzen Window (KPW) technique, we draw the distribution curves (probability density function) of the mean performance w.r.t. the 15 trials, respectively.

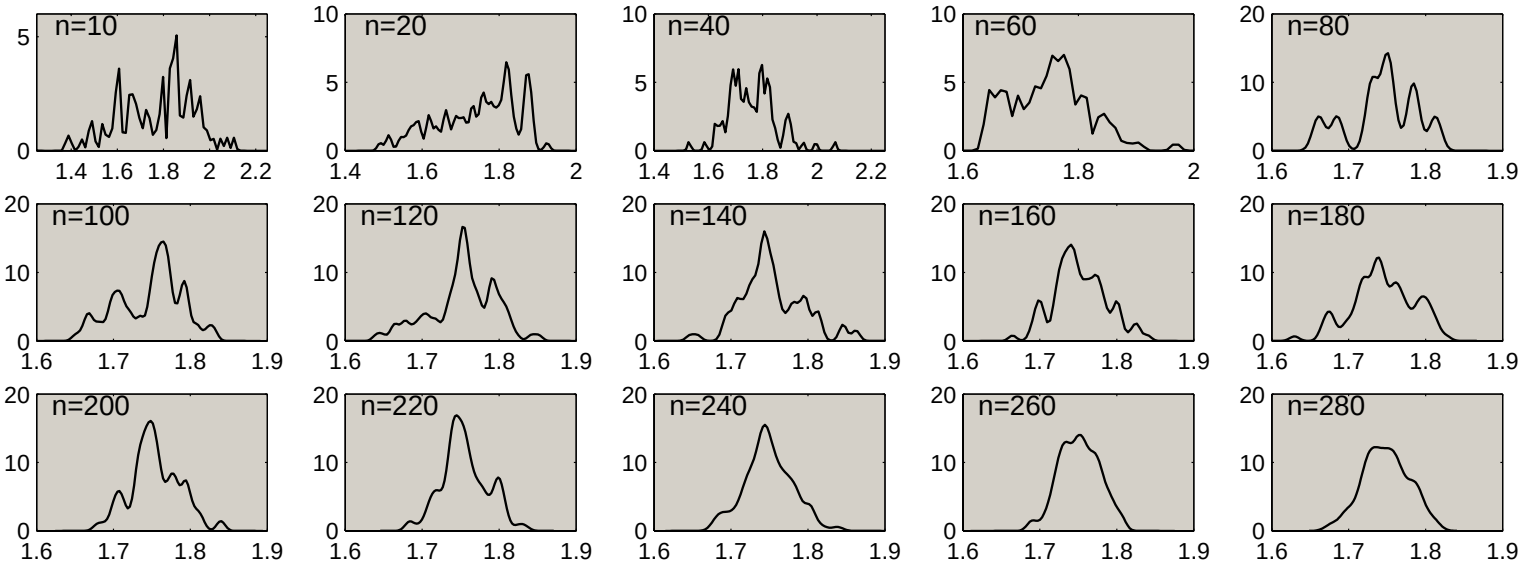


# Whether the Central Limit Theorem (CLT) is applicable?




- $n < 160$, significantly non-normal
- $n \geq 240$, promising approximation of normality

# Whether the Central Limit Theorem (CLT) is applicable?




- At least for KDataSets, a sample may have to contain $\geq 160$ in order to make the mean performance normally distribute
- Current practices: we usually have only $< 30$ performance measurements (e.g., SPEC CPU2006, SPLASH-2, PARSEC)

# How to conduct performance comparison?

- Non-normal distribution of performance measurements
- Number of measurements is not sufficiently large
- Aforementioned comparison task

| perlbench | bzip2 | gcc | mcf | gobmk | hmmer | sjeng | libquantum | h264ref | omnetpp | astar | xalancbmk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26.4 | 20.4 | 22.2 | 45.4 | 25.4 | 34.5 | 27.0 | **633.6** | 46.0 | 22.8 | 21.7 | 30.1 |
| 13.9 | 9.47 | 10.0 | 14.3 | 10.3 | 8.31 | 9.28 | 13.91 | 19.8 | 9.08 | 8.43 | 14.7 |

# Outline

# Non-parametric statistical tests

Non-parametric techniques are

- "*Distribution-free methods, which do not rely on assumptions that the data are drawn from a given probability distribution*" [Wikipedia]
  - Do not assume normally distributed performance measurements
  - Do not need lots of performance measurements for applying the CLT

Two famous non-parametric tests:

- Wilcoxon Rank-Sum Test (uni-benchmark comparisons) & Wilcoxon Signed-Rank Test (cross-benchmark comparisons) [Wilcoxon, 1945]
- Using rankings of data instead of sample mean
- Larger performance gaps count more

# Wilcoxon Signed-Rank Test for cross-benchmark comparison

- NULL hypothesis: "the performance of $A$ is equivalent to that of $B$"
- Alternative hypothesis ( the conclusion we want to make ):
  - One-tail: "$A$ outperforms $B$" or "$B$ outperforms $A$"
  - Two-tail: "the performance of $A$ is not equivalent to that of $B$"
- In addition to the concrete conclusion, the test can offer us the corresponding confidence

# Wilcoxon Signed-Rank Test for cross-benchmark comparison

- On the $i$-th $(i = 1, \ldots, n)$ benchmark, calculate $d_i = a_i - b_i$, where $a_i$ and $b_i$ are scores of computers $A$ and $B$ on the $i$-th benchmark, respectively
- Rank $d_1, d_2, \ldots, d_n$ according to an ascending order of their absolute values
- Calculate the signed-rank sums of $A$ and $B$ by

$$R_A = \sum_{i:d_i>0} \text{Rank}(d_i) + \frac{1}{2} \sum_{i:d_i=0} \text{Rank}(d_i),$$

$$R_B = \sum_{i:d_i<0} \text{Rank}(d_i) + \frac{1}{2} \sum_{i:d_i=0} \text{Rank}(d_i),$$

- Estimate the confidence based on $R_A$ and $R_B$

# Non-parametric Hierarchical Performance Testing (HPT)

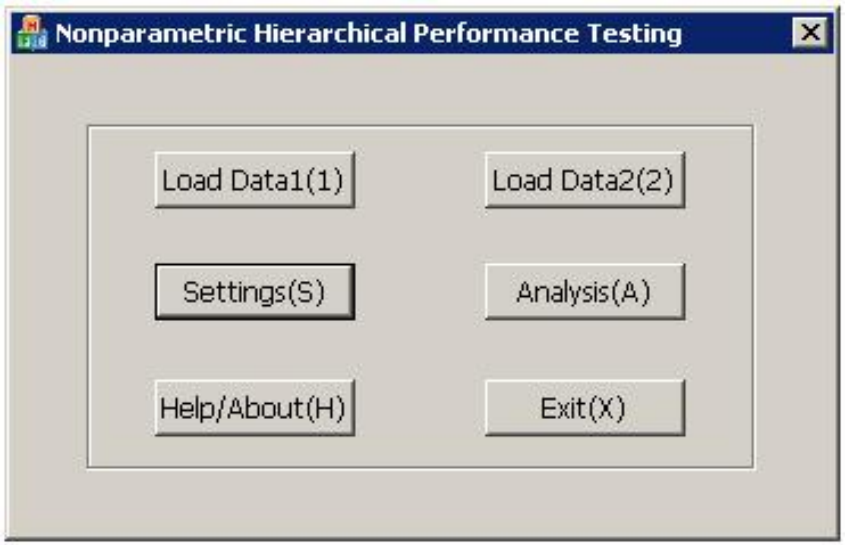
- Conduct Wilcoxon Rank-Sum Test to compare the performance of two computers on each benchmark
- Comparison result on the $i$-th benchmark can be taken into account in cross-benchmark comparison only if the difference between the performance of two computers on that benchmark is significant (otherwise, $d_i$ will be set to 0 w.r.t. the $i$-th benchmark)
- Conduct Wilcoxon Signed-Rank Test to compare the performance of two computers in cross-benchmark comparison
- Estimate the confidence that one computer outperforms another

# Quantitative comparison using HPT

How to estimate the performance speedup of computer $A$ over computer $B$)?

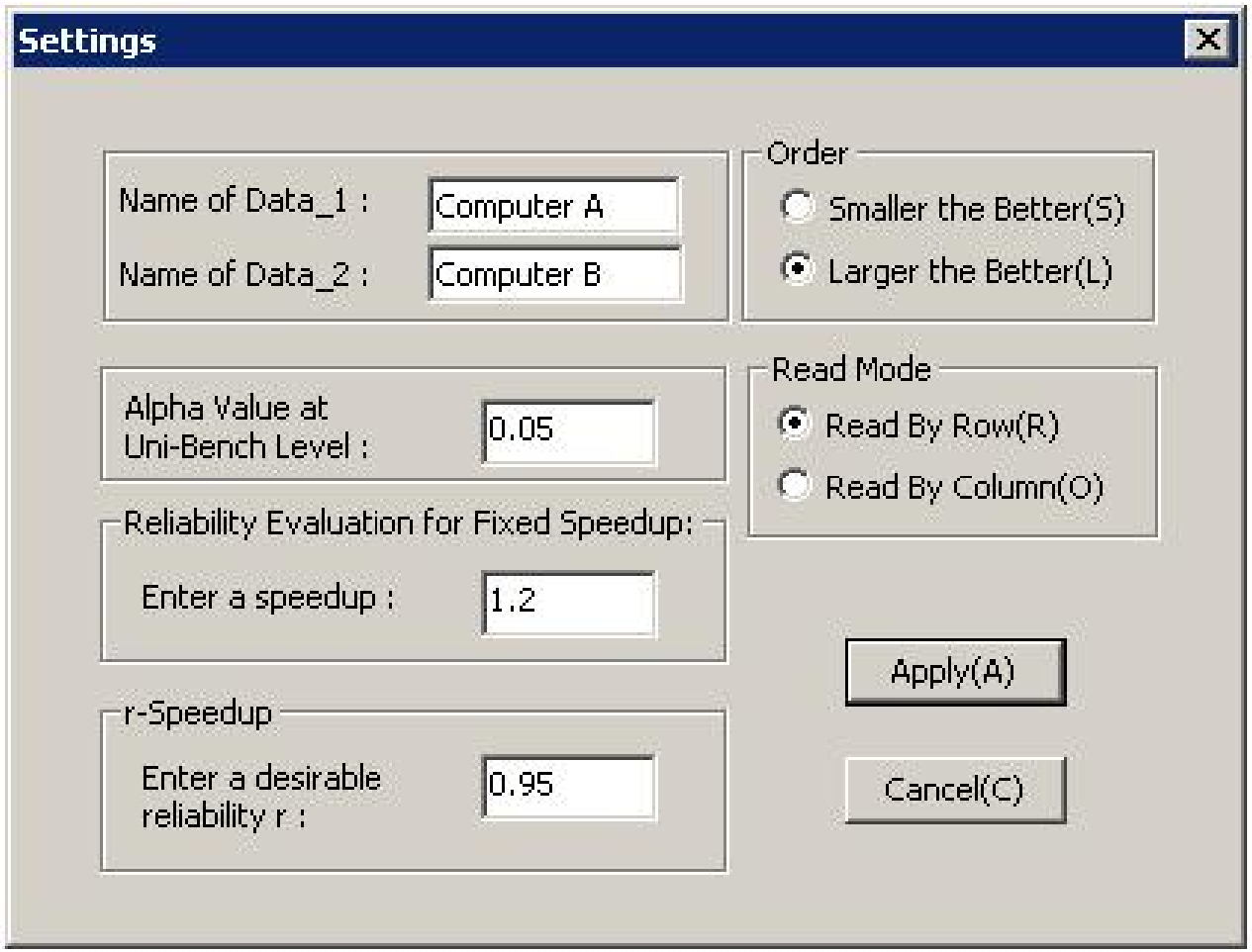
- $r$-Speedup: fix the confidence level to be $r \in [0, 1]$, then estimate the maximal performance speedup that can pass the HPT
  - Shrink all performance scores of computer $A$ by $\gamma$ ($\gamma \geq 1$); Assume a virtual computer $A_\gamma$ taking those reduced scores
  - Check whether $A_\gamma$ significantly outperforms $B$ at the confidence level $r$
  - If yes, increase $s$ by a fixed small step-size $\kappa$ and repeat the above procedure
  - Otherwise, return $r$-Speedup$=s-\kappa$

# Open-source software: http://novel.ict.ac.cn/tchen/hpt/

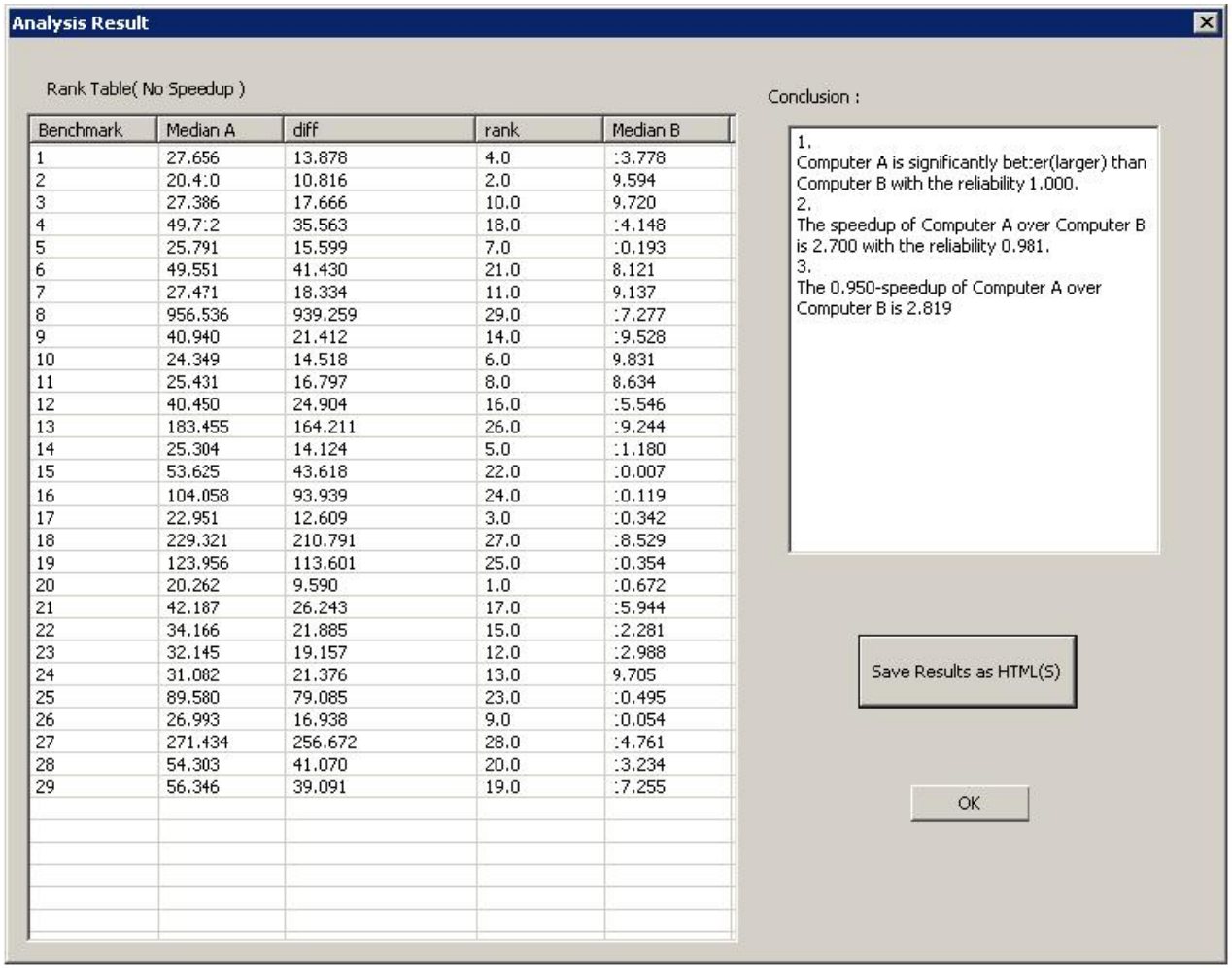



- Input: Performance scores (.txt, .csv)
- Settings: Speedup-under-test & confidence
- Output: Report (.html)
- Qualitative & quantitative comparisons

# Open-source software: http://novel.ict.ac.cn/tchen/hpt/



# Open-source software: http://novel.ict.ac.cn/tchen/hpt/




| Benchmark | Median A | diff | rank | Median B |
|---|---|---|---|---|
| 1 | 27.656 | 13.878 | 4.0 | :3.778 |
| 2 | 20.4:0 | 10.816 | 2.0 | 9.594 |
| 3 | 27.386 | 17.666 | 10.0 | 9.720 |
| 4 | 49.7:2 | 35.563 | 18.0 | :4.148 |
| 5 | 25.791 | 15.599 | 7.0 | :0.193 |
| 6 | 49.551 | 41.430 | 21.0 | 8.121 |
| 7 | 27.471 | 18.334 | 11.0 | 9.137 |
| 8 | 956.536 | 939.259 | 29.0 | :7.277 |
| 9 | 40.940 | 21.412 | 14.0 | :9.528 |
| 10 | 24.349 | 14.518 | 6.0 | 9.831 |
| 11 | 25.431 | 16.797 | 8.0 | 8.634 |
| 12 | 40.450 | 24.904 | 16.0 | :5.546 |
| 13 | 183.455 | 164.211 | 26.0 | :9.244 |
| 14 | 25.304 | 14.124 | 5.0 | :1.180 |
| 15 | 53.625 | 43.618 | 22.0 | :0.007 |
| 16 | 104.058 | 93.939 | 24.0 | :0.119 |
| 17 | 22.951 | 12.609 | 3.0 | :0.342 |
| 18 | 229.321 | 210.791 | 27.0 | :8.529 |
| 19 | 123.956 | 113.601 | 25.0 | :0.354 |
| 20 | 20.262 | 9.590 | 1.0 | :0.672 |
| 21 | 42.187 | 26.243 | 17.0 | :5.944 |
| 22 | 34.166 | 21.885 | 15.0 | :2.281 |
| 23 | 32.145 | 19.157 | 12.0 | :2.988 |
| 24 | 31.082 | 21.376 | 13.0 | 9.705 |
| 25 | 89.580 | 79.085 | 23.0 | :0.495 |
| 26 | 26.993 | 16.938 | 9.0 | :0.054 |
| 27 | 271.434 | 256.672 | 28.0 | :4.761 |
| 28 | 54.303 | 41.070 | 20.0 | :3.234 |
| 29 | 56.346 | 39.091 | 19.0 | :7.255 |

# Recall the performance comparison of computers $A$ (upper) and $B$ (lower) on SPECint2006

| perlbench | bzip2 | gcc | mcf | gobmk | hmmer | sjeng | libquantum | h264ref | omnetpp | astar | xalancbmk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26.4 | 20.4 | 22.2 | 45.4 | 25.4 | 34.5 | 27.0 | **633.6** | 46.0 | 22.8 | 21.7 | 30.1 |
| 13.9 | 9.47 | 10.0 | 14.3 | 10.3 | 8.31 | 9.28 | 13.91 | 19.8 | 9.08 | 8.43 | 14.7 |

Paired $t$-test: we cannot conclude that "$A$ significantly outperforms $B$" at the confidence level $\geq 0.95$!

# Performance comparison of computers $A$ and $B$ on SPECint2006

Nonparametric Hierarchical Performance Testing

A VS. B

A wins | B wins | Tie

Rank Table(Without Speedup)

| Benchmark | A | Diff | Rank | B |
|---|---|---|---|---|
| 1 | 26.400 | 12.500 | 3.0 | 13.900 |
| 2 | 20.400 | 10.930 | 1.0 | 9.470 |
| 3 | 22.200 | 12.200 | 2.0 | 10.000 |
| 4 | 45.400 | 31.100 | 11.0 | 14.300 |
| 5 | 25.400 | 15.100 | 6.0 | 10.300 |
| 6 | 34.500 | 26.190 | 9.0 | 8.310 |
| 7 | 27.000 | 17.720 | 8.0 | 9.280 |
| 8 | 633.600 | 619.690 | 12.0 | 13.910 |
| 9 | 46.000 | 26.200 | 10.0 | 19.800 |
| 10 | 22.800 | 13.720 | 5.0 | 9.080 |
| 11 | 21.700 | 13.270 | 4.0 | 8.430 |
| 12 | 30.100 | 15.400 | 7.0 | 14.700 |

HPT: $A$ significantly outperforms $B$ at the confidence level 1.

# What is the performance speedup of $A$ over $B$ on SPECint2006?

A VS. B

A wins | B wins | Tie

Rank Table(Speedup 2.239)

| Benchmark | A | Diff | Rank | B |
|---|---|---|---|---|
| 1 | 11.791 | -2.109 | 8.0 | 13.900 |
| 2 | 9.111 | -0.359 | 2.0 | 9.470 |
| 3 | 9.915 | -0.085 | 1.0 | 10.000 |
| 4 | 20.277 | 5.977 | 10.0 | 14.300 |
| 5 | 11.344 | 1.044 | 4.0 | 10.300 |
| 6 | 15.409 | 7.099 | 11.0 | 8.310 |
| 7 | 12.059 | 2.779 | 9.0 | 9.280 |
| 8 | 282.983 | 269.073 | 12.0 | 13.910 |
| 9 | 20.545 | 0.745 | 3.0 | 19.800 |
| 10 | 10.183 | 1.103 | 5.0 | 9.080 |
| 11 | 9.692 | 1.262 | 7.0 | 8.430 |
| 12 | 13.444 | -1.256 | 6.0 | 14.700 |

HPT: The performance 0.95-speedup of $A$ over $B$ is 2.239 ($A$ is 2.239 times faster than $B$, with the confidence 0.95).


J. R. Mashey, "War of the benchmark means: time for a truce", **ACM SIGARCH Computer Architecture News** 32(4), 2004.


Student (W. S. Gosset), "The probable error of a mean", **Biometrika** 6(1), 1908.


L. Le Cam, "The Central Limit Theorem Around 1935", **Statistical Science** 1(1), 1986.


M. Guthaus, J. Ringenberg, D. Ernst, T. Austin, T. Mudge, and R. Brown, "Mibench: A free, commercially representative embedded benchmark suite", in **Proceedings of the IEEE 4th Annual International Workshop on Workload Characterization (WWC)**, 2001.


Y. Chen, Y. Huang, L. Eeckhout, G. Fursin, L. Peng, O. Temam, and C. Wu, "Evaluating iterative optimization across 1000 datasets", in **Proceedings of the 2010 ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'10)**, 2010.


E. Parzen, "On estimation of a probability density function and mode", **Annals of Mathematical Statistics** 33, 1962.


F. Wilcoxon, "Individual comparisons by ranking methods", **Biometrics** 1(6), 1945.

Software available at:

http://novel.ict.ac.cn/tchen/hpt/

Q & A