

Precise Regression Benchmarking with Random Effects: Improving Mono Benchmark Results

Tomas Kalibera and Petr Tuma

DISTRIBUTED SYSTEMS RESEARCH GROUP

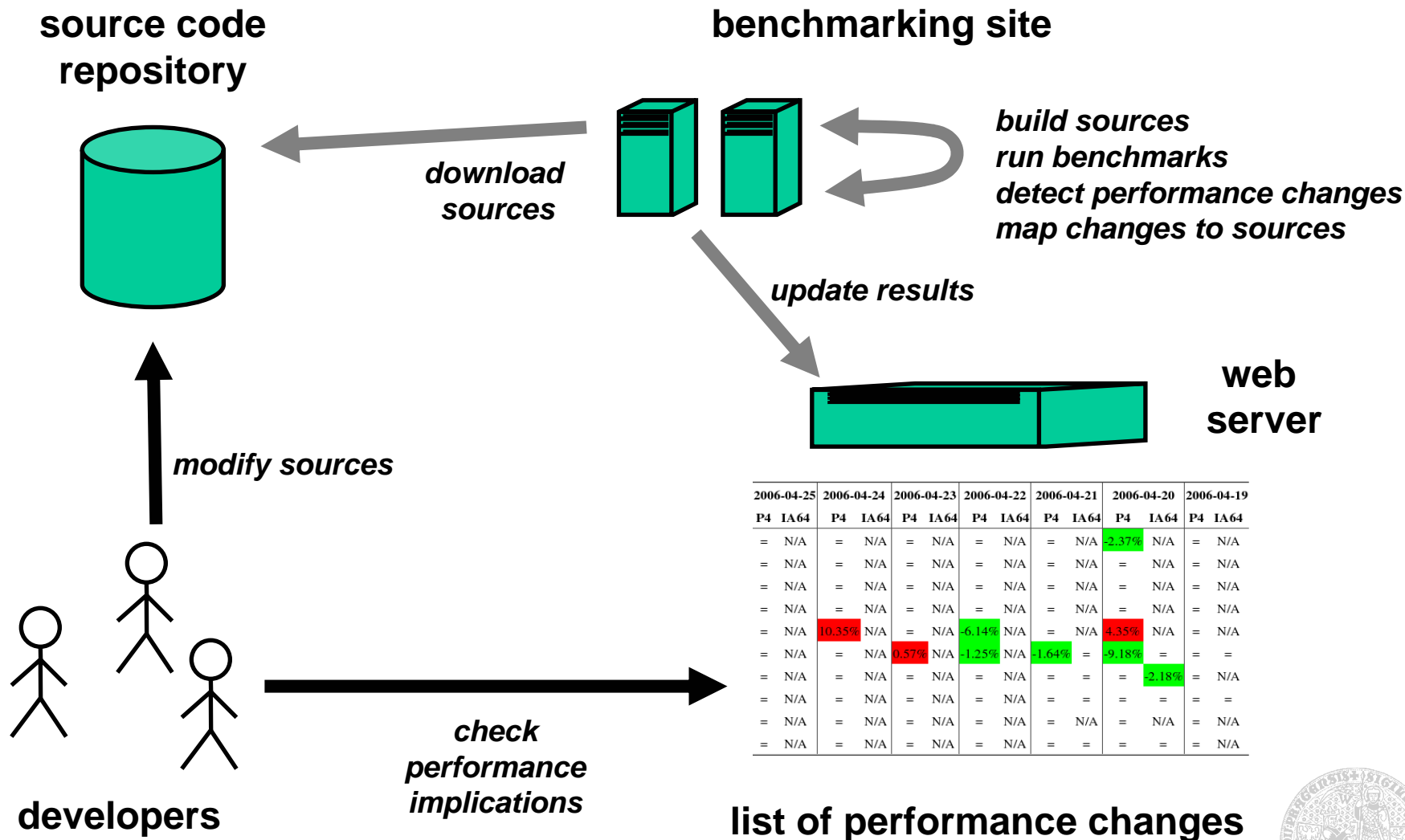
<http://nenya.ms.mff.cuni.cz>

CHARLES UNIVERSITY PRAGUE

Faculty of Mathematics and Physics



Regression Benchmarking

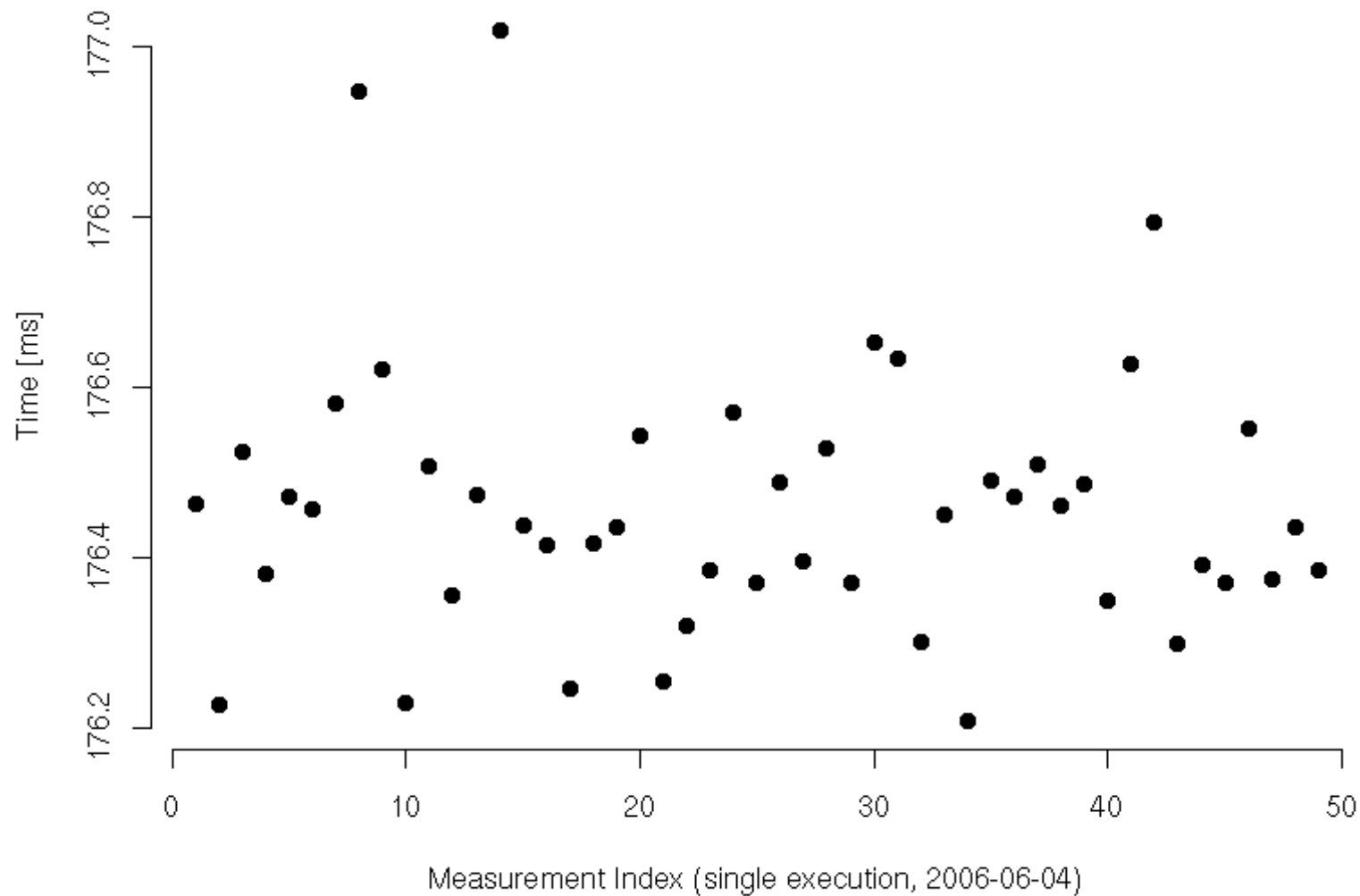


Regression Benchmarking in Mono Project

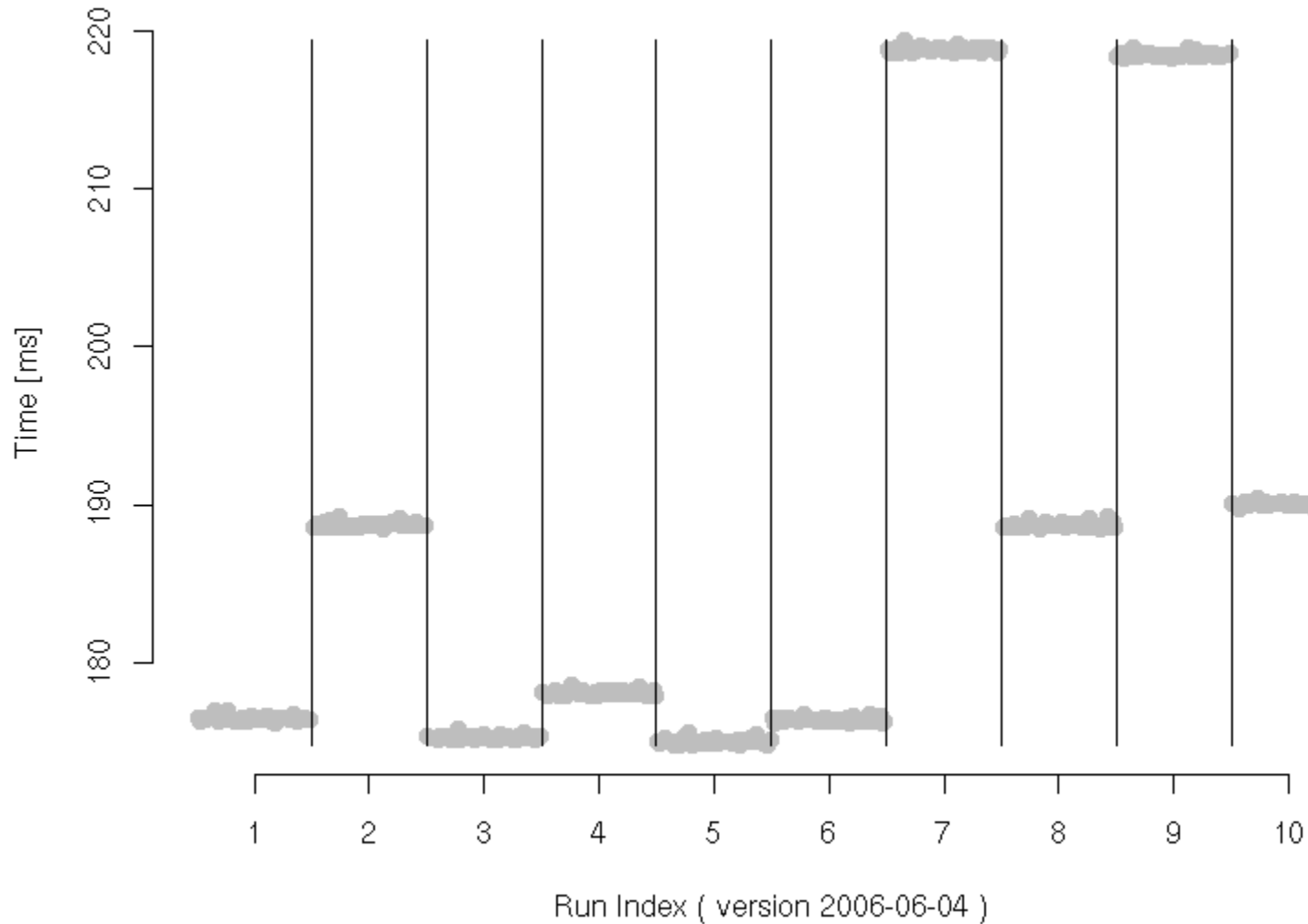
- Mono
 - Open-source .NET implementation
 - 3M lines of code, active development
 - <http://www.mono-project.com>
- Mono Regression Benchmarking Project
 - Automated benchmarking of daily versions
 - Automated detection of performance changes
 - Results since August, 2004
 - <http://nenya.ms.mff.cuni.cz/projects/mono>



Random Effects in Measured Operations



Random Effects in Benchmark Executions



Random Effects in Benchmark Executions

- Process startup
 - Allocating memory for code
 - Allocating global data structures
- Process run-time
 - Dynamic memory allocation / de-allocation
- Consequences
 - Random location in physical memory
 - Different numbers of memory cache misses



different performance



Random Effects in Compilation/Building

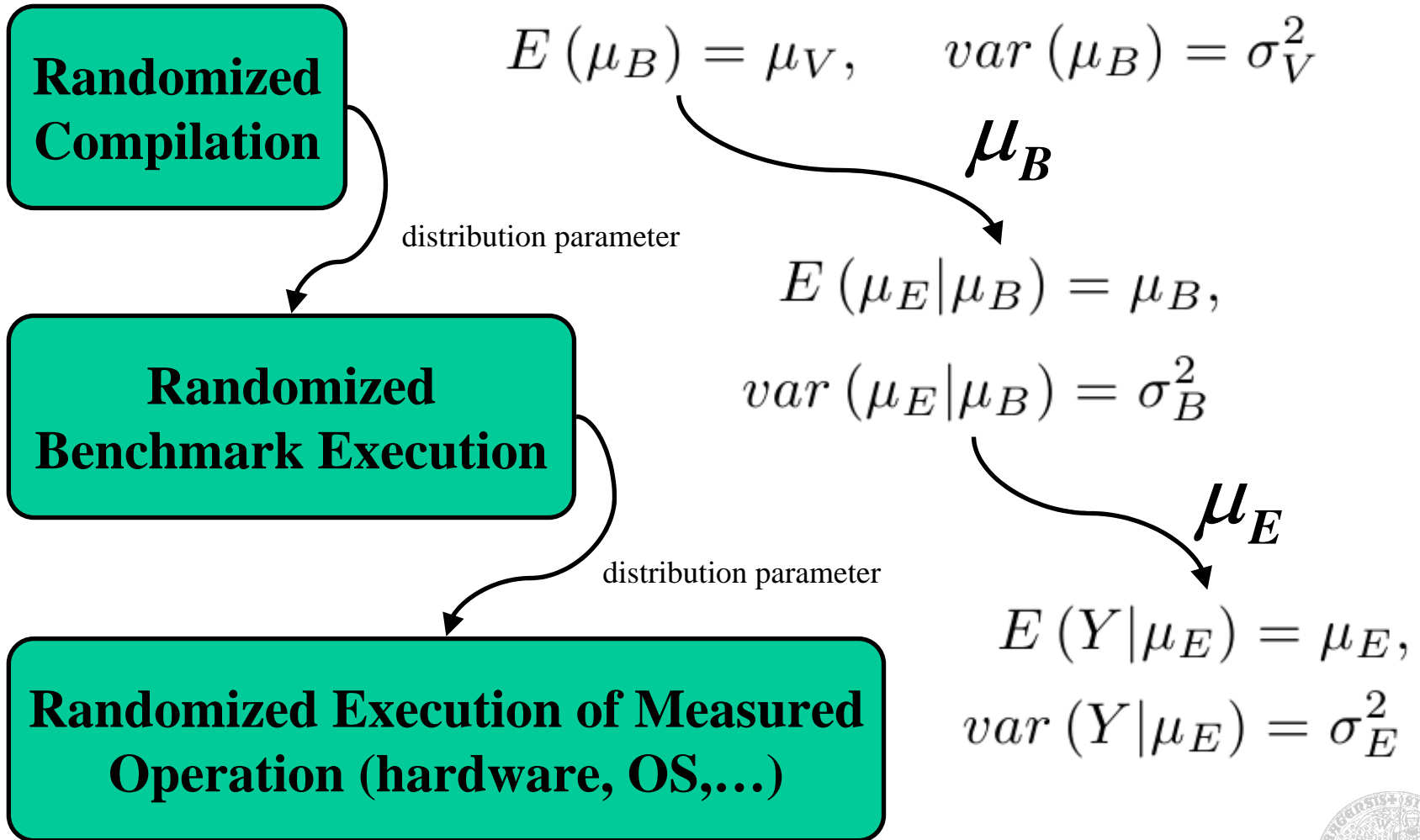
- Randomized Binary Layout
 - Random symbol naming (GCC) + symbol names based linking
 - More complex causes exist (Mono)
- Consequence
 - Different data structures / code parts in memory pages
 - Different numbers of memory cache misses, set at compilation/build time



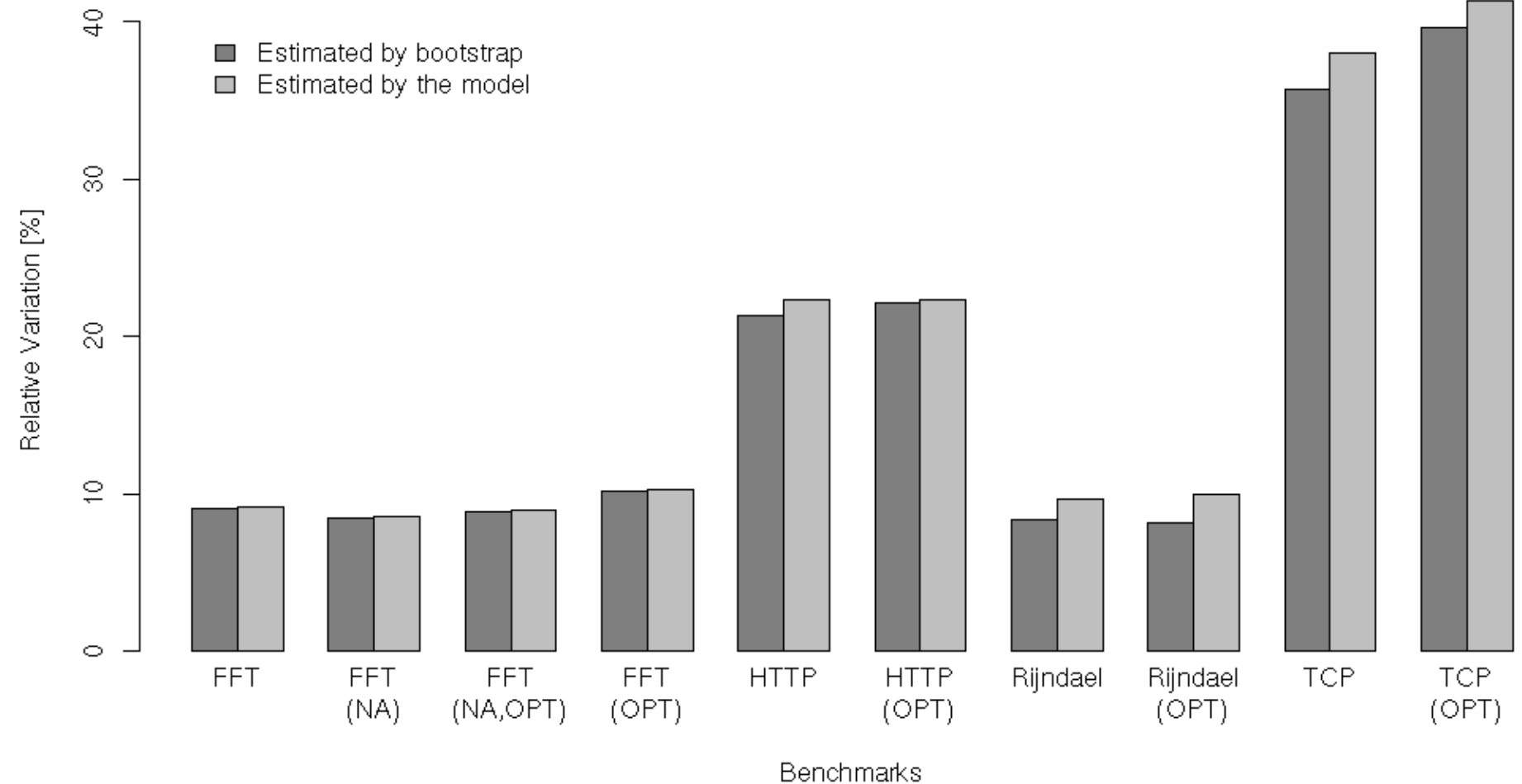
different performance



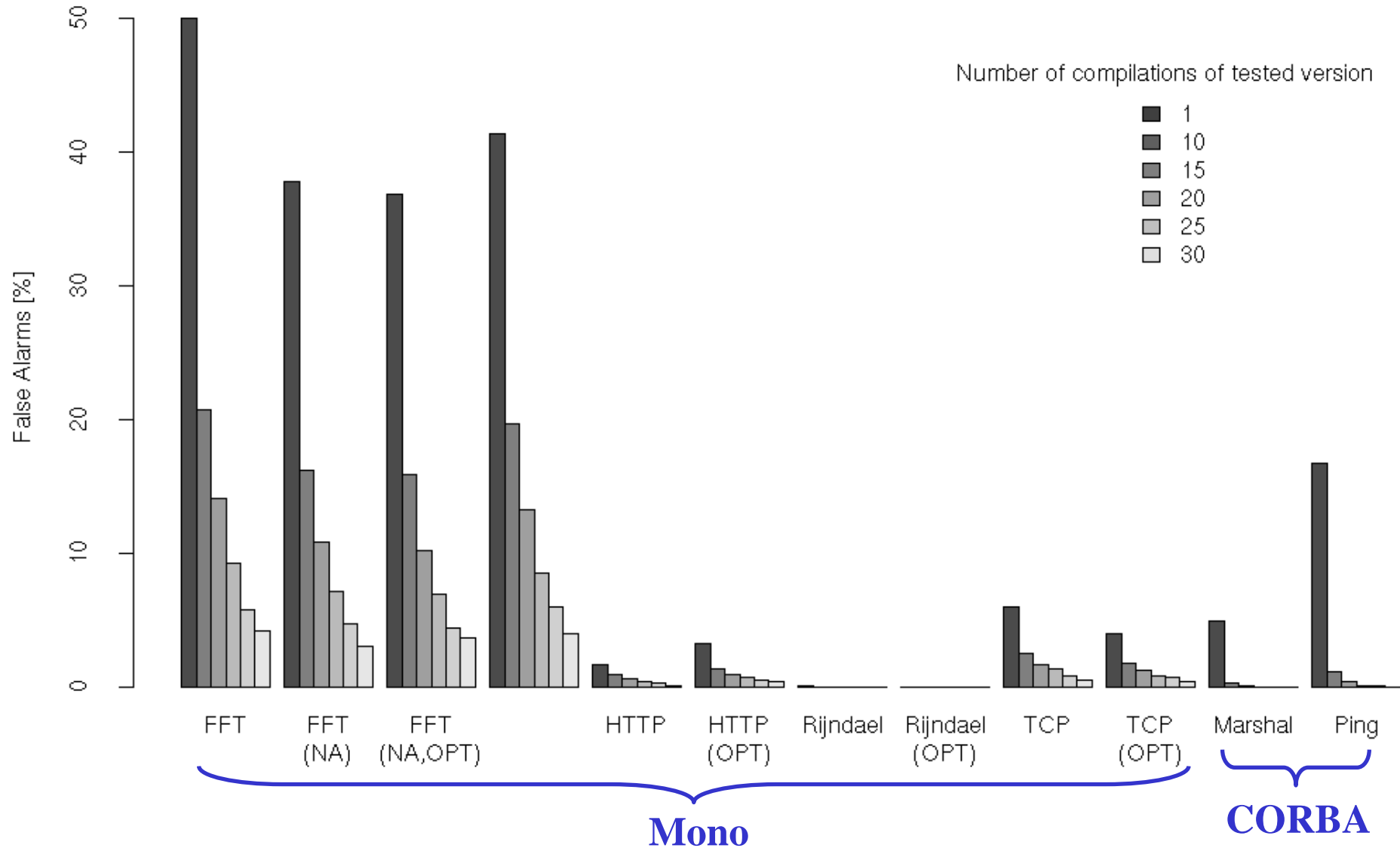
Statistical Model of Random Effects



Statistical Model: Experimental Evaluation



Random Effects in Regression Benchmarking

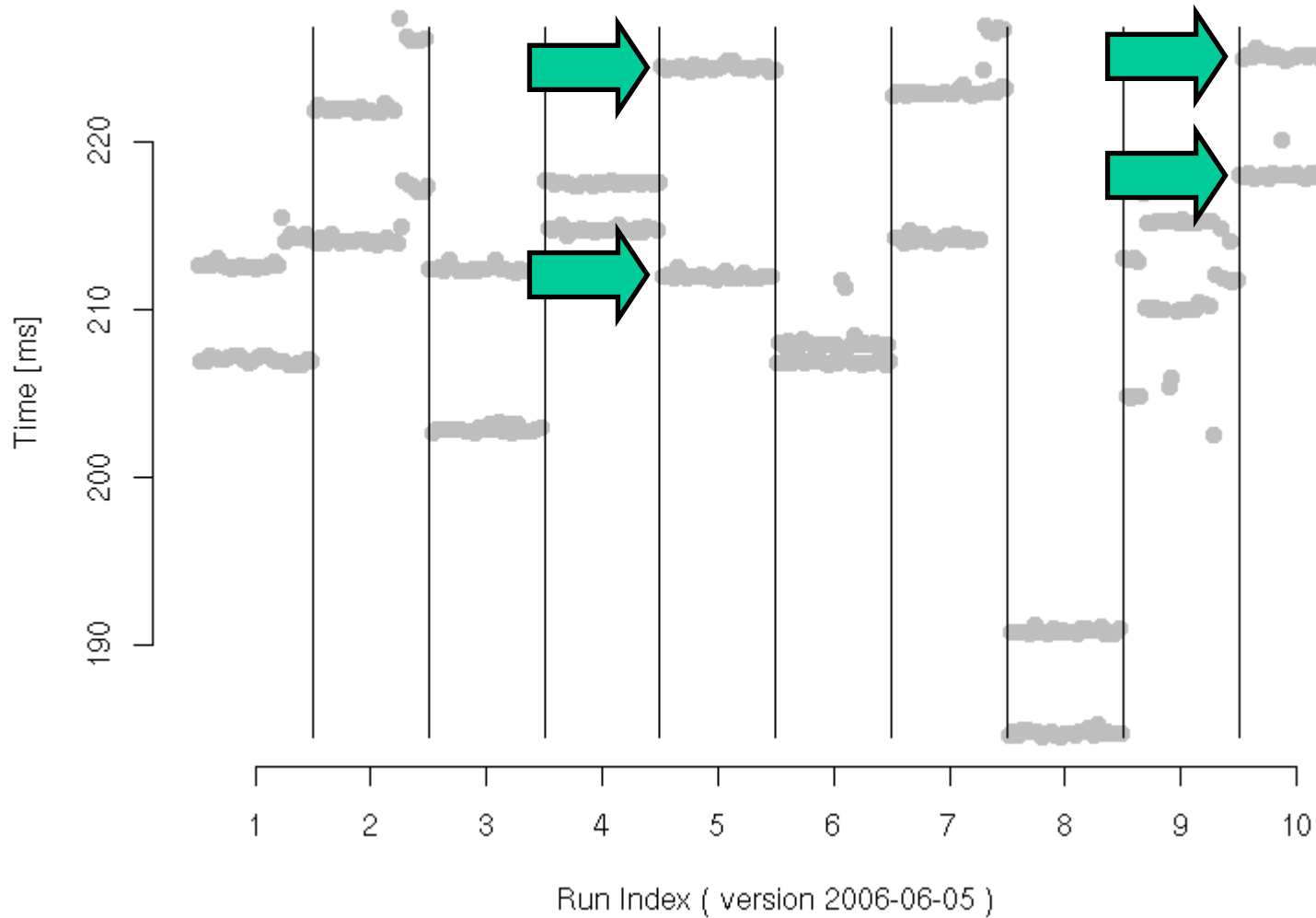


Conclusion

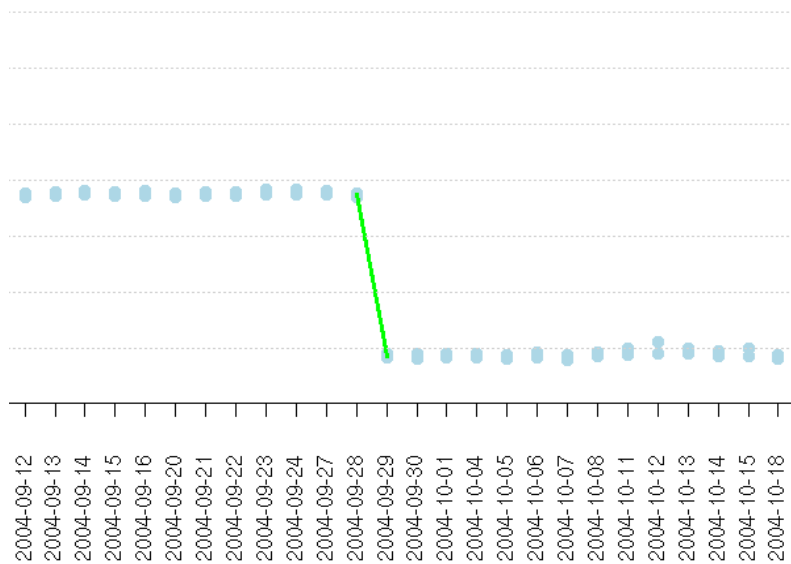
- Random effects impact performance
 - At compilation/build time
 - At process start-up, run-time
- Precise (regression) benchmarking can be achieved by statistical methods
 - Repeating compilations, executions, measurements
 - Using our statistical method to evaluate results and estimate their precision



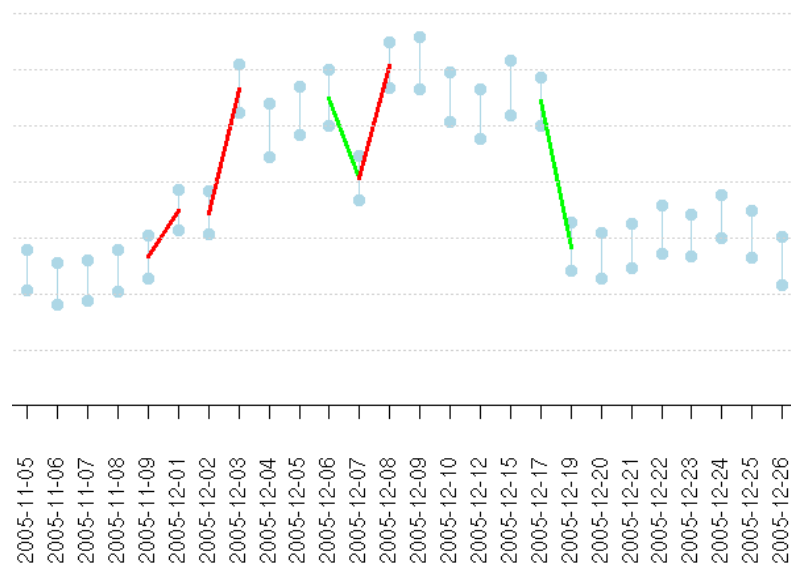
Future Work: Random Effects in Operations



Future Work: Model & Change-Point Detection



Likely change, no more measurements needed



Strange changes, more measurements might help



Publications

- Kalibera, T., Bulej, L., Tůma, P.: *Automated Detection of Performance Regressions: The Mono Experience*, MASCOTS 2005, IEEE
- Kalibera, T., Bulej, L., Tůma, P.: *Quality Assurance in Performance: Evaluating Mono Benchmark Results*, SOQUA 2005, LNCS
- Kalibera, T., Bulej, L., Tůma, P.: *Benchmark Precision and Random Initial State*, SPECTS 2005, SCS
- Bulej, L., Kalibera, T., Tůma, P.: *Repeated Results Analysis for Middleware Regression Benchmarking*, *Performance Evaluation: An International Journal*, Elsevier, 2005

