Multi-Level Repetition Benchmarking

Marek Cuchý

marek.cuchy@agents.fel.cvut.cz

B(E)4M36ESW

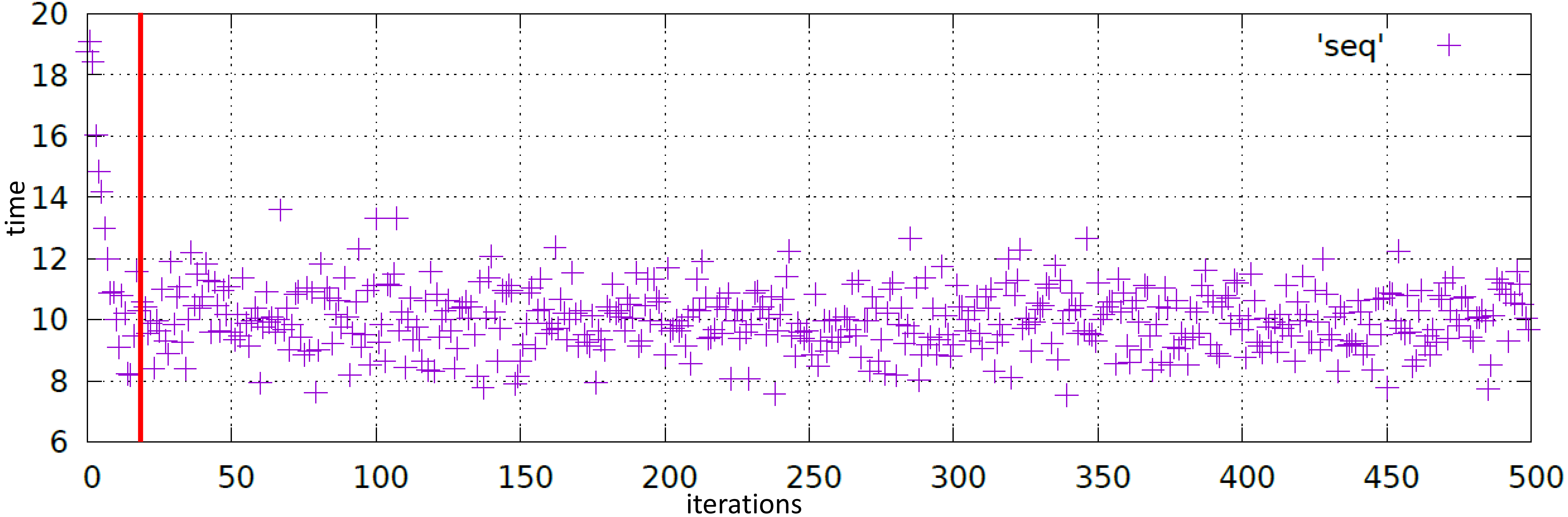
March 9, 2020

Multi-Level Repetition

- Variance in measurements may occur at higher levels
  - ➔we need to repeat measurements at least on the level of the variance
- The highest level is the most important one
- Levels:
  1. Iteration – the smallest possible measurement (e.g. loop body)
  2. Execution – running of the program
  3. Compilation (stable in Java)

Warm-up

- Measurements useful only after reaching a steady/independent state
- E.g. by manual inspection of sequence plot of several executions



Number of Repetitions

- Simplified to two levels $(n = 2)$ - Java
  - 1 – iteration
  - 2 – execution
- What is the optimal count on the lower levels to increase precision?
  - Lower levels – less time needed to do more repetitions
- Run dimensioning experiment
  - e.g. 30 execution and 40 iterations excl. warm-up
- Calculate unbiased variance estimators for each level T_1^2, T_2^2

$$T_1^2 = S_1^2$$

Biased variance estimators

$$T_2^2 = S_2^2 - \frac{S_1^2}{\boldsymbol{r}_1}$$

Number of iterations used for the dimensioning experiment

Biased Variance Estimators

$$\mathbf{S_i^2} = \frac{1}{\prod_{k=i+1}^{n} \mathbf{r_k}} \frac{1}{\mathbf{r_i} - 1} \sum_{j_n=1}^{r_n} \cdots \sum_{j_i=1}^{r_i} \left(\bar{\mathbf{Y}}_{j_n \ldots j_i \bullet \cdots \bullet} - \bar{\mathbf{Y}}_{j_n \ldots j_{i+1} \bullet \cdots \bullet} \right)^2$$

- For two levels ($n = 2$):
  - S_1^2 - mean of execution variances
  - $S_2^2 = S_n^2$ - variance of execution means
- See [2] Chapter 6.1 for details and example

Number of Repetitions

$$\forall i, 1 \leq i < n, \quad r_i = \left\lceil \sqrt{\frac{c_{i+1}}{c_i} \frac{T_i^2}{T_{i+1}^2}} \right\rceil$$

$$T_1^2 = S_1^2$$
$$T_2^2 = S_2^2 - \frac{S_1^2}{\boldsymbol{r}_1}$$

- c_1 - single measurement (iteration) duration after warm-up
- c_2 - execution cost (time to reach independent state – warm-up)
- r_i - optimal repetition count on i-th level
- What if $T_i^2 \leq 0$?
  - i-th level induces very little variance ➔ can be skipped
- What about r_n???
  - More highest level repetitions always increase precision

Number of Repetitions – Example

Matrix Multiplication

Notebook

i5-7200U (2.5GHz), 8GB RAM

```
measureMultiply
S1 = 857.2065800776355
S2 = 138.16871702761134
T1 = 857.2065800776355
T2 = 116.73855252567046
r1 = 8

measureMultiply1D
S1 = 2187.8572504410527
S2 = 672.0203176226894
T1 = 2187.8572504410527
T2 = 617.3238863616631
r1 = 4

measureMultiplyTrans
S1
S2
T1
T2
r1 = 7
```

Desktop

i5-8500 (3GHz), 32GB RAM

```
measureMultiply
S1 = 456.22120240936977
S2 = 73.46727479932864
T1 = 456.22120240936977
T2 = 62.061744739094394
r1 = 8

measureMultiply1D
S1 = 254.87119015897457
S2 = 5.73656707647479
T1 = 254.87119015897457
T2 = -0.6352126774995739
r1 = -1

measureMultiplyTrans
r1 = -1
```

For each environment (HW, OS, …) and benchmark (implementation) different setup is required

Execution Time + Effect Size Confidence Interval

$$\overline{Y} \quad \pm \quad t_{1-\frac{\alpha}{2},\nu} \sqrt{\frac{1}{r_n(r_n-1)} \sum_{j_n=1}^{r_n} \left(\overline{Y}_{j_n \underbrace{\bullet \cdots \bullet}_{n-1}} - \overline{Y} \right)^2}$$

$$\overline{Y} \quad \pm \quad t_{1-\frac{\alpha}{2},\nu} \sqrt{\frac{S_n^2}{r_n}}$$

- n - number of levels
- r_n - number of repetition on the highest level
- $\overline{Y}$ - mean across all measurements (excl. warm-up)
- $(1-\alpha)$ – confidence interval (e.g. 95% confidence → $\alpha = 0.05$)
- $t_{1-\frac{\alpha}{2},\nu}$ - $\left(1-\frac{\alpha}{2}\right)$-quantile of the t-distribution with $\nu = r_n - 1$ degrees of freedom, can be found in a table

Speed-Up Ratios with Confidence Interval

$$\frac{\overline{Y} \cdot \overline{Y'} \mp \sqrt{\left(\overline{Y} \cdot \overline{Y'}\right)^2 - \left(\overline{Y}^2 - h^2\right)\left(\overline{Y'}^2 - h'^2\right)}}{\overline{Y}^2 - h^2}$$

$$h = \sqrt{t^2_{\frac{\alpha}{2},\nu} \frac{S_n^2}{r_n}} \qquad h' = \sqrt{t^2_{\frac{\alpha}{2},\nu} \frac{S'^2_n}{r_n}}$$

- $\overline{Y}, \overline{Y'}$ - means of the compared implementations
- h, h' - half-widths of the confidence intervals for the single implementations (you can reuse the values from prev. slide)
- S_n^2, S'^2_n - biased variance estimator of the n-th level

References

[1] Kalibera, T. and Jones, R. E. (2013) Rigorous Benchmarking in Reasonable Time. In: ACM SIGPLAN International Symposium on Memory Management (ISMM 2013), 20–12 June, 2013, Seattle, Washington, USA. http://kar.kent.ac.uk/33611/

[2] Kalibera, T. and Jones, R. E. (2012) Quantifying performance changes with effect size confidence intervals. Technical Report 4–12, University of Kent