

Detection of Performance Changes in MooBench Results Using Nyrkiö on GitHub Actions

Shinhyung Yang¹ David Georg Reichelt² Henrik Ingo³ Wilhelm Hasselbring¹

¹Kiel University

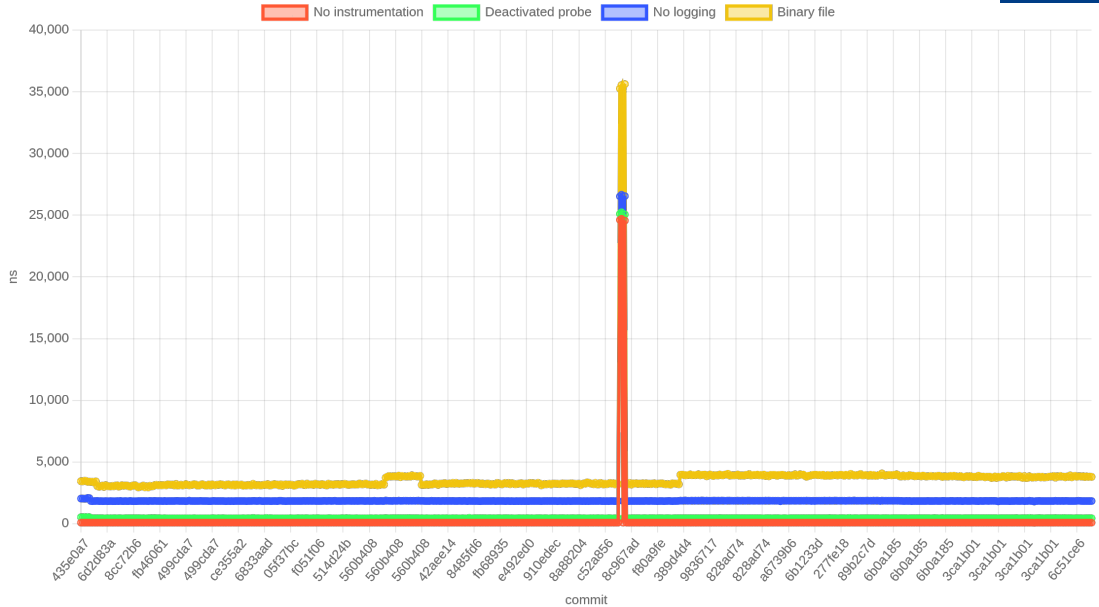
²Lancaster University Leipzig / URZ Leipzig

³Nyrkiö Oy

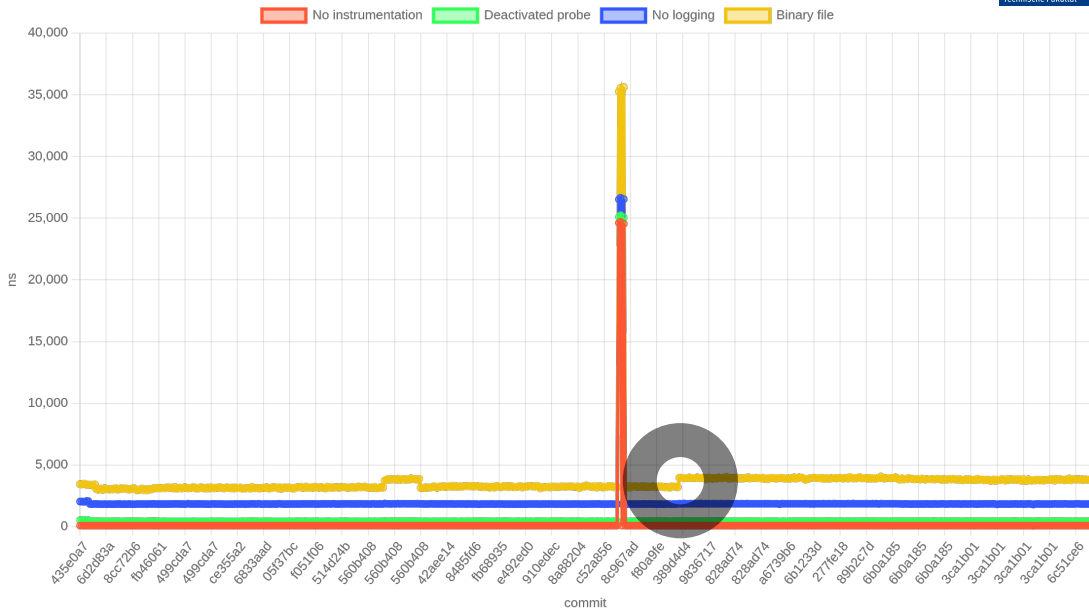
Nov 4, 2025
SSP 2025, Kiel

- ▶ Problem Statement
- ▶ Motivation
- ▶ Background
- ▶ Experiment & Evaluation
- ▶ Future Works & Conclusion

Problem Statement



Problem Statement



How We Came to This

- ▶ MooBench is deployed on GitHub CI/CD [RJvH24]
- ▶ Continuous benchmarking
 - ▶ MooBench on Jenkins CI/CD¹: run once per week
 - ▶ MooBench on GitHub CI/CD²: run once per day and by git-push
 - ▶ Using *github-action-benchmark*³ for continuous visualization
- ▶ We manually checked the performance decrease by inspecting the diagram

- Need for automated regression benchmarking → Nyrkiö
 - Investigate performance regression → Experiments and evaluation

¹<https://kieker-monitoring.net/performance-benchmarks/>

²<https://kieker-monitoring.github.io/moobench/dev/bench/>

³<https://github.com/benchmark-action/github-action-benchmark>

Nyrkiö Change Detection Service

- ▶ Change point detection using E-Divisive algorithm [MJ14]
 - ▶ Detects change points in the given numerical sequence
- ▶ Easy-to-integrate to a GitHub Project's CI/CD as a GH action⁴

uses: nyrkio/change-detection@HEAD

with:

name: 'Kieker-java'

tool: 'customSmallerIsBetter'

output-file-path: output.json

fail-on-alert: true

github-token: \${ secrets.GITHUB_TOKEN }

nyrkio-token: \${ secrets.NYRKIO_JWT_TOKEN }

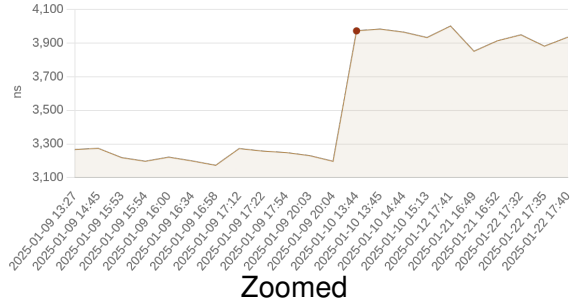
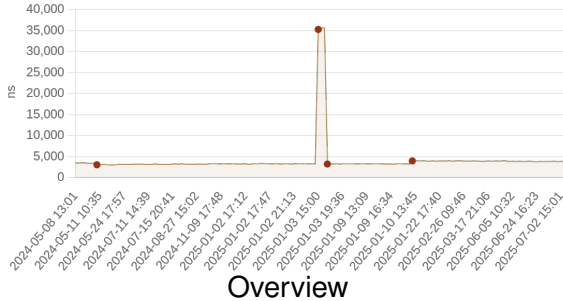
auto-push: true

- ▶ Dashboard service on an own homepage⁵

⁴<https://github.com/nyrkio/change-detection>

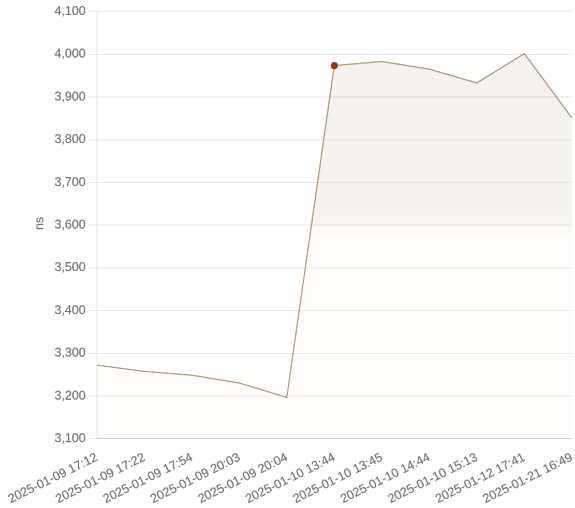
⁵<https://nyrkio.com/frontpage>

Validate Our Findings with Nyrkiö (1/2)



- ▶ The first three changes were temporary
- ▶ The fourth change is permanent

Validate Our Findings with Nyrkiö (2/2)



59d51e3b0aa10eab56d8c047cdc0314d30512741 @
2025-01-10 13:44

Also add r for benchmark Kieker-java

<https://github.com/shinhyungyang/moobench/> (main) /
commit/59d51e3b0aa10eab56d8c047cdc0314d30512741

shinhyungyang/moobench/main/Kieker-java

Binary file: 3973.32 ns

index: 111

time: 1736513043

forward_change_percent: 23

magnitude: 0.225171

mean_before: 3241.252500

stddev_before: 34.921836

mean_after: 3971.090000

stddev_after: 22.687677

pvalue: 0.000000

Prev

Next

Nyrkiö for Regression Benchmarking

- ▶ Before Nyrkiö:
 - ▶ All previous data manually uploaded using `curl` command
- ▶ After Nyrkiö:
 - ▶ New data automatically uploaded via Nyrkiö
 - ▶ Nyrkiö notifies regressions as a GH issue⁶ :-)

<> Code Issues Pull requests Actions Projects Wiki Security Insights

Nyrkiö detected change points, on 2025-06-07 05:59 #1 Closed

New changes since 2025-05-24T05:59:02

Commit time	Commit	Test name	Metric
	abc1b07	Pinpoint-java	[Pinpoint Basic](https://nyrkio.com/result/Pinpoint-java?commit=abc1b079752f0b184cf8fcbf519fe517a4a0ba7a&timestar Basic)

 Nyrkiö only files one issue per commit. Please check nyrkio.com to see if there are more changes found than what is reported here.

Create sub-issue 😊

⁶<https://github.com/shinhyungyang/moobench/issues/1>

Investigating Performance Changes (1/2)

- ▶ What made the performance change?

Ubuntu-latest workflows will use Ubuntu-24.04 image (Sep 17, 2024)^a

- ▶ “This change will be rolled out over a period of several weeks beginning December 5th and will complete on January 17th, 2025.”
- ▶ “**R**: Removed from the Ubuntu 24.04 image due to maintenance reasons.”

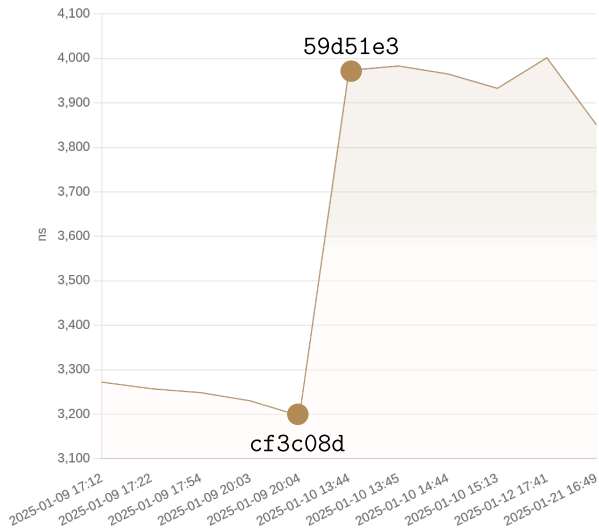
^a<https://github.com/actions/runner-images/issues/10636>

- ▶ MooBench uses **R** software and it was affected by the roll-out of Ubuntu 24.04.
- ▶ Most of changes to MooBench around the performance change was to manually install R.

- ▶ The next step was to find difference between Ubuntu 22.04 and Ubuntu 24.04
 - ▶ Set 1: Ubuntu 22.04 on GitHub-hosted VM
 - ▶ Set 2: Ubuntu 24.04 on GitHub-hosted VM
 - ▶ Set 3: Ubuntu 22.04 on self-hosted VM
 - ▶ Set 4: Ubuntu 24.04 on self-hosted VM

Verify: “Our commits did not incur regression”

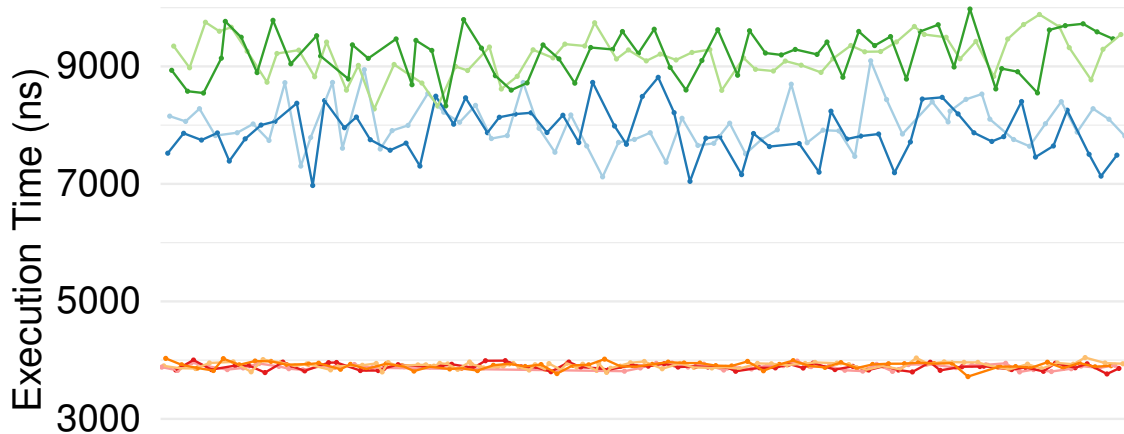
- ▶ Also, make sure our software change did not make the change



Verify: “Our commits did not incur regression”

- ▶ Also, make sure our software change did not make the change
- ▶ Commit `cf3c08d`
 - ▶ Set 1: Ubuntu 22.04 on GitHub
 - ▶ Set 2: Ubuntu 24.04 on GitHub
 - ▶ Set 3: Ubuntu 22.04 on self-hosted VM
 - ▶ Set 4: Ubuntu 24.04 on self-hosted VM
- ▶ Commit `59d51e3`
 - ▶ Set 5: Ubuntu 22.04 on GitHub
 - ▶ Set 6: Ubuntu 24.04 on GitHub
 - ▶ Set 7: Ubuntu 22.04 on self-hosted VM
 - ▶ Set 8: Ubuntu 24.04 on self-hosted VM

Verify: “Our commits did not incur regression”



Verify: “Our commits did not incur regression”

► Evaluating Significance of Difference

- Shapiro-Wilk normality test
Check that the two compared data series are normally distributed
- $w \approx 1.00$ confirms normality
- Paired t-test
Compare two input data series and check whether their difference is significant
- $p < 0.05$ confirms the difference is significant

Verify: “Our commits did not incur regression”

► Comparing MooBench results:

► on 2 different git-commits:

cf3c08d and 59d51e3

► on 4 different execution environments:

22.04/self, 24.04/self, 22.04/GitHub, and 24.04/GitHub

● 22.04/self

● 24.04/self

● 22.04/GitHub

● 24.04/GitHub

● 22.04/self

● 24.04/self

● 22.04/GitHub

● 24.04/GitHub

$w = 0.962\ 460$

$w = 0.983\ 340$

$w = 0.978\ 520$

$w = 0.984\ 860$

$p = 0.072\ 288$

$p = 0.647\ 910$

$p = 0.648\ 089$

$p = 0.488\ 497$

1. All w values are close to 1.00

Therefore, the two compared data series are normally distributed.

2. All p values are > 0.05

Therefore, the two compared data series are **not significantly different**.

Validate our assumption: Ubuntu version change incurs this regression

- ▶ For the comparison of two Ubuntu versions on each different VM, we used an aggregated data series of cf3c08d and 59d51e3. I.e.,

22.04 vs. 24.04: (w, p)

- ▶ GitHub-hosted: (● ●) vs. (● ●): $(0.988\ 280, 0.001,966,284)$
- ▶ Self-hosted: (● ●) vs. (● ●): $(0.988\ 590, 1.368,409 \times 10^{-50})$

1. All w values are close to 1.00
Therefore, the two compared data series are normally distributed.
2. All p values are < 0.05
Therefore, the two compared data series are **significantly different**.

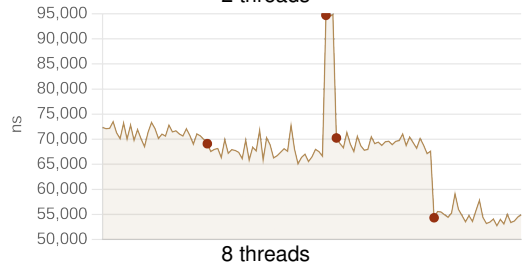
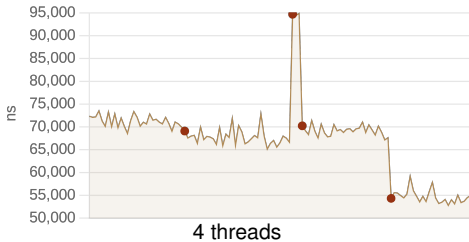
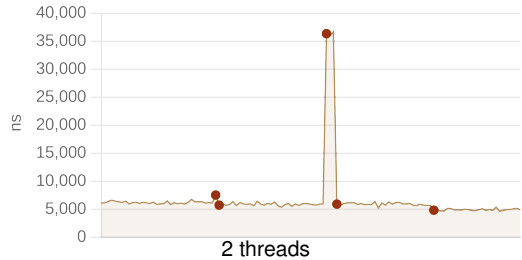
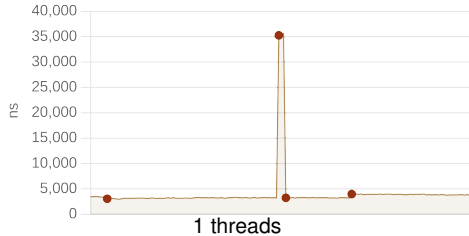
What Are Next Steps?

- ▶ Find out software changes between
 - ▶ Ubuntu 22.04 and 24.04 versions on
 - ▶ GitHub-hosted VMs and
 - ▶ self-hosted VMs
 - ▶ Linux Kernel versions⁷

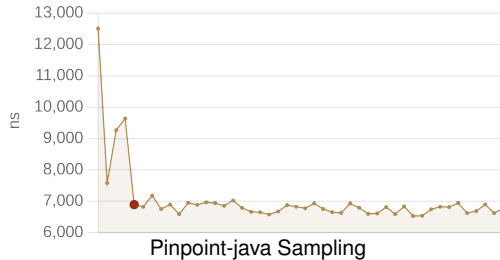
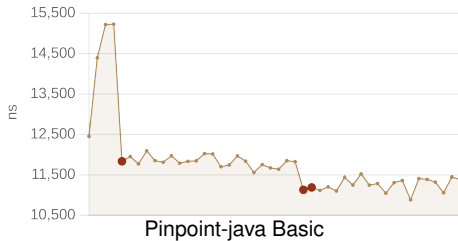
⁷<https://shinhyungyang.github.io/moobench/dev/bench/>

Thank you! Any questions?

Regression and Scalability: Kieker AspectJ



Other Tracing Technologies: Pinpoint



References I

Appendix ▷ ▷

- [MJ14] David S. Matteson and Nicholas A. James, *A nonparametric approach for multiple change point analysis of multivariate data*, Journal of the American Statistical Association **109** (2014), no. 505.

- [RJvH24] David Georg Reichelt, Reiner Jung, and André van Hoorn, *Overhead measurement noise in different runtime environments*, SSP '24, 2024, PID: 20.500.12116/45533.