Taming Timeout Flakiness: An Empirical Study of SAP HANA

Alexander Berndt, Sebastian Baltes, Thomas Bach

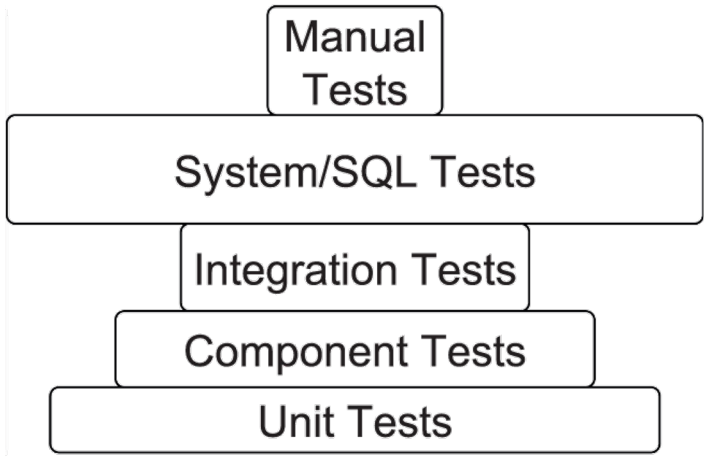
2024, April 16th

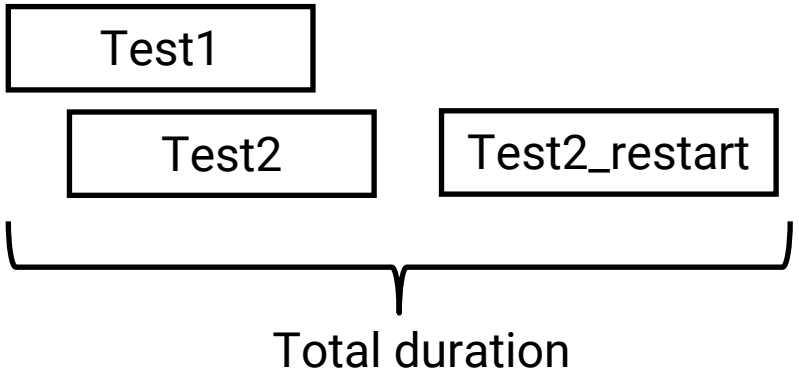
Informal Flakiness Definition

A test can be considered flaky when it exhibits both **passing** and **failing** results for the same code.

Testing SAP HANA

- large-scale database management system
- flaky failures affect **99% of CI-runs** at SAP HANA pre-submit testing
- standard strategy: **Restart flaky tests**
- **but:** additional computational resources, delay for developers





Motivation & Goals

1. understand test flakiness at SAP HANA
   - focus on system tests in the pre-submit stage
2. analyze major contributing factor
   - how much flakiness is caused by this factor?
3. provide actionable insights
   - how can we improve the current situation in practice?

Data Mass-testing

- Problem: tests are executed only **once for one code revision**
- use **idle resources** on HANA's testing infrastructure over the weekend
- **repeatedly execute** test suite on the same code
- increase timeout values for Adjusted Timeout Value dataset

Data Set	# Tests	# Test Executions	# Code Revisions
MT	744	558,423	17
ATV	701	363,169	7

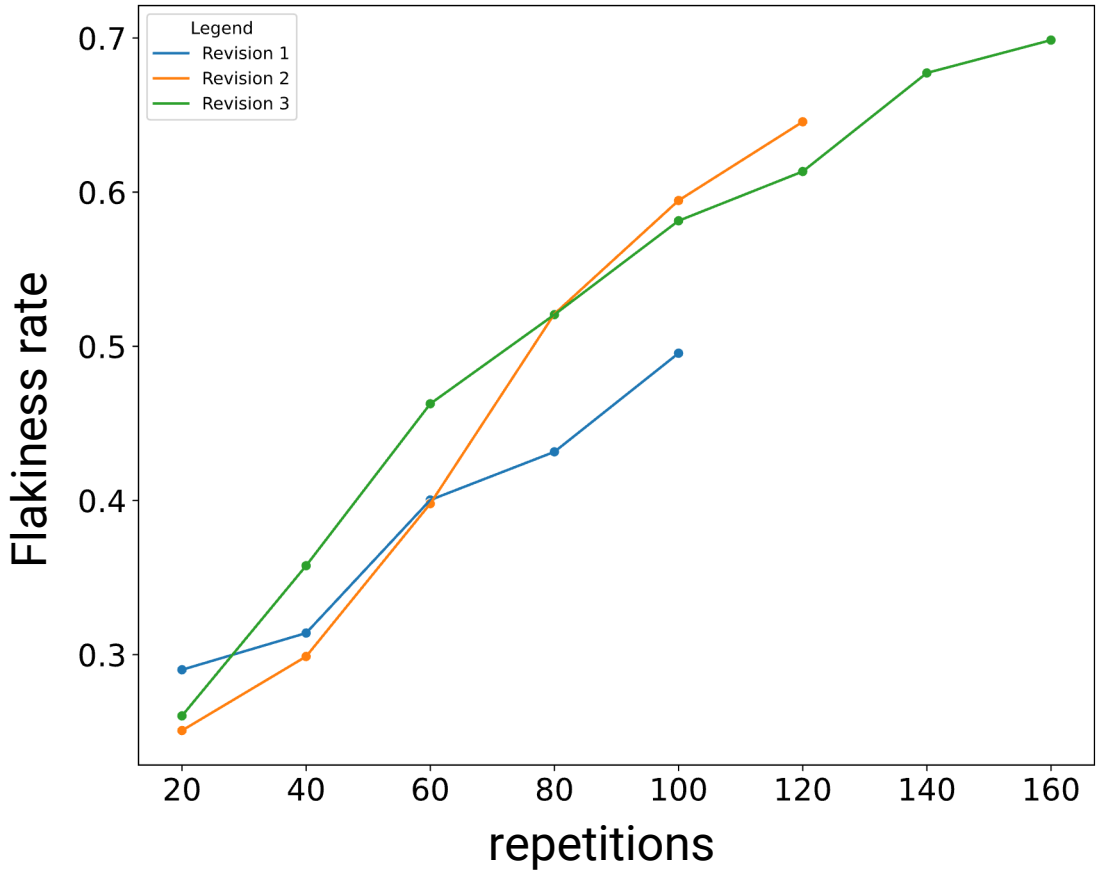
Masstesting (MT) and Adjusted Timeout Values (ATV) dataset

Evaluation

RQ1: What level of test flakiness can we observe in SAP HANA's system tests and what can we identify as a major contributing factor to flakiness?

- flakiness rate
  - $\frac{\#Flaky\ Tests}{\#Executed\ Tests}$
- **49% to 70% flakiness rate** in masstesting

Answer: The overall level of flakiness depends on the **number of test repetitions**.



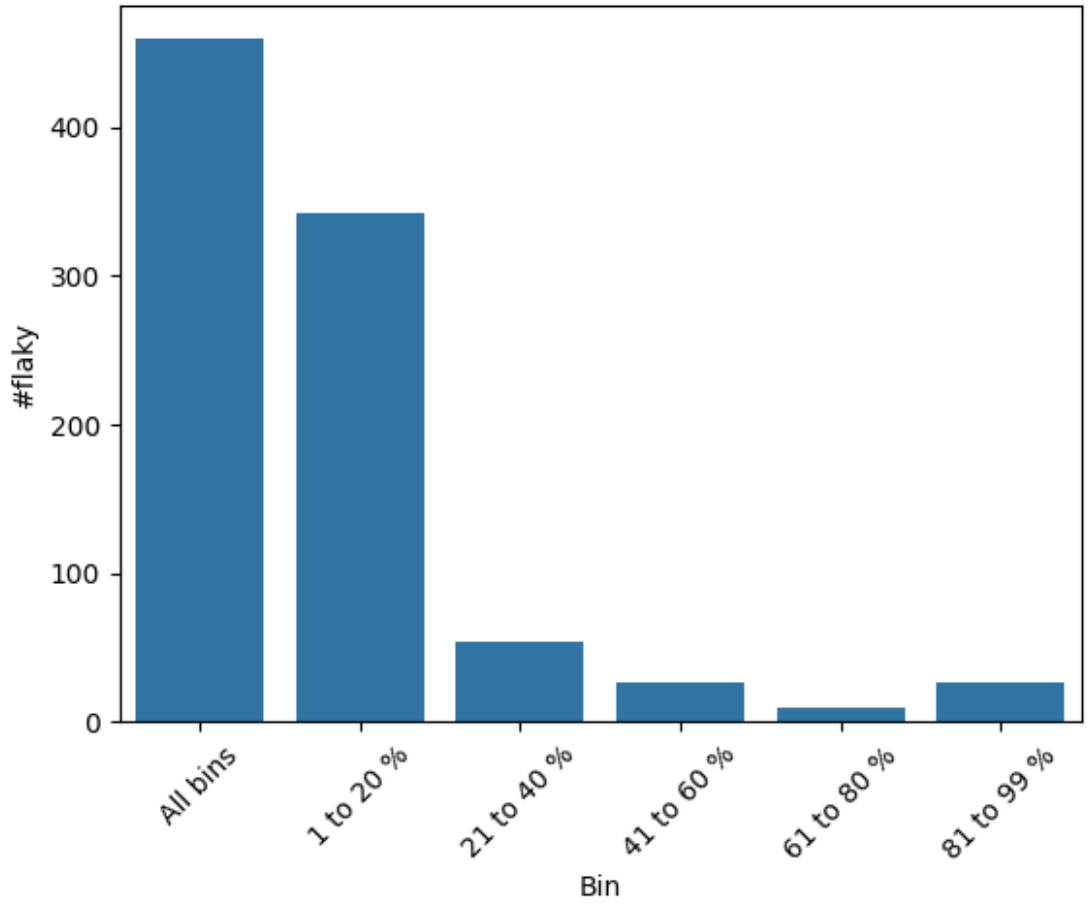
Evaluation

RQ1: What level of test flakiness can we observe in SAP HANA's system tests and what can we identify as a major contributing factor to flakiness?

- most flaky tests **fail rarely**
  - **90%** of flaky tests fail only in 1-20% of executions
- **70%** of the flaky failures caused by **timeouts**

Answer: Timeouts are the major contributing factor to test flakiness at SAP HANA.

Flaky Failure Frequency Bins for Revision 3



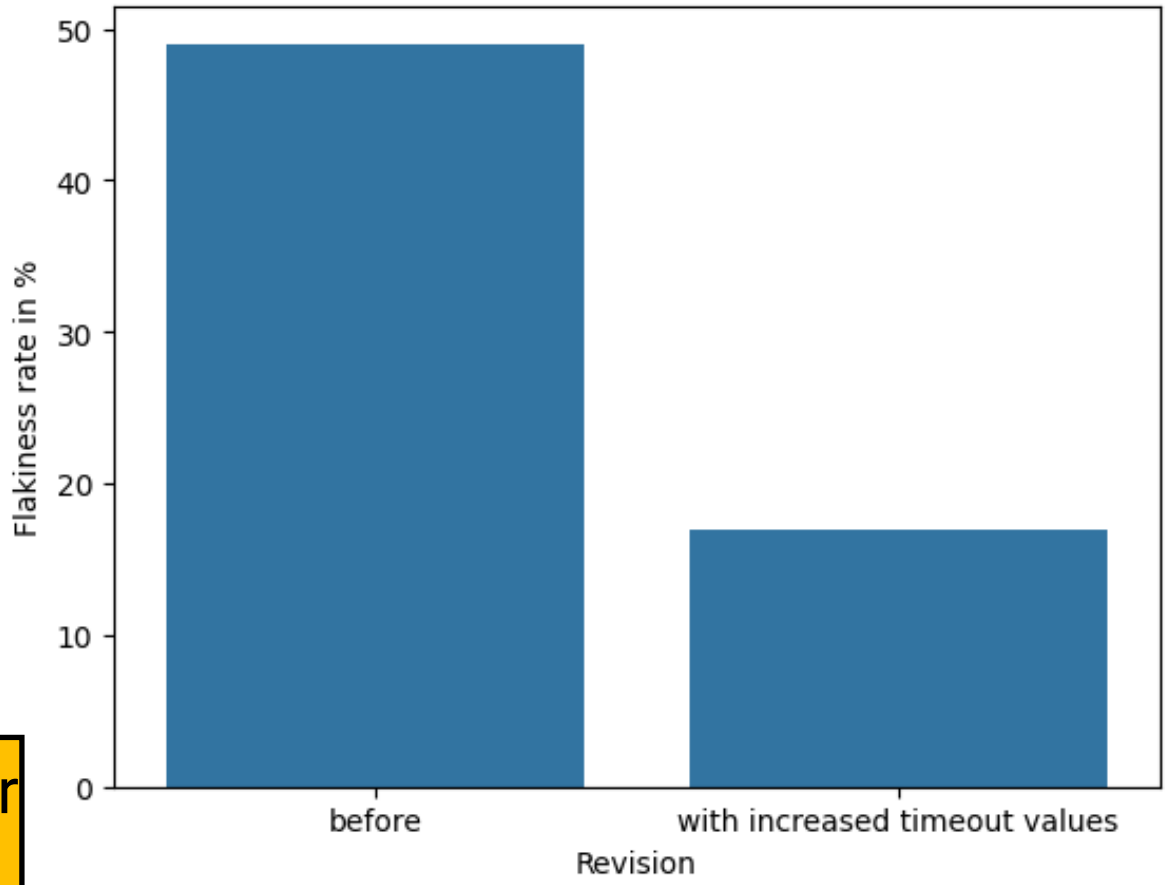
Evaluation

RQ2: What impact does increasing timeout values have on test flakiness in context of SAP HANA?

- increasing timeout values by factor 10 **reduces flakiness notably**
- E.g. for 100 repetitions, flakiness rate drops **from 49% to 17%**
- **but**: 10% flaky failures remain timeouts

Answer: Increasing timeout values by factor 10 **reduces test flakiness** by 65%.

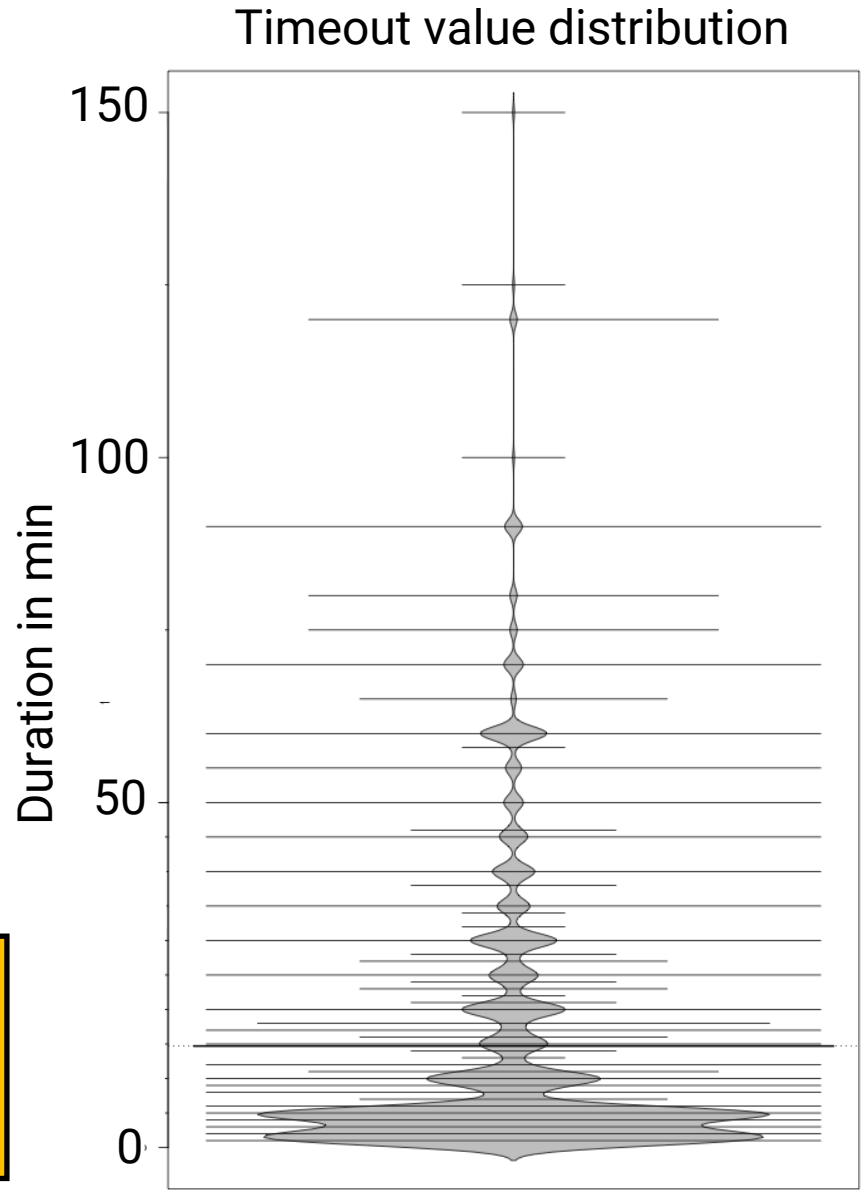
Flaky rate after 100 repetitions before (left) and after (right) timeout increase



Evolution of Max Duration Values

RQ3: How do developers commonly adjust timeout values in the context of SAP HANA?

- study version history from 2016 to 2023
- identify commits that adjust timeout values



Answer: Most common values are **33 % to 100 % for increases,** and **50% for reductions**.

Evaluation

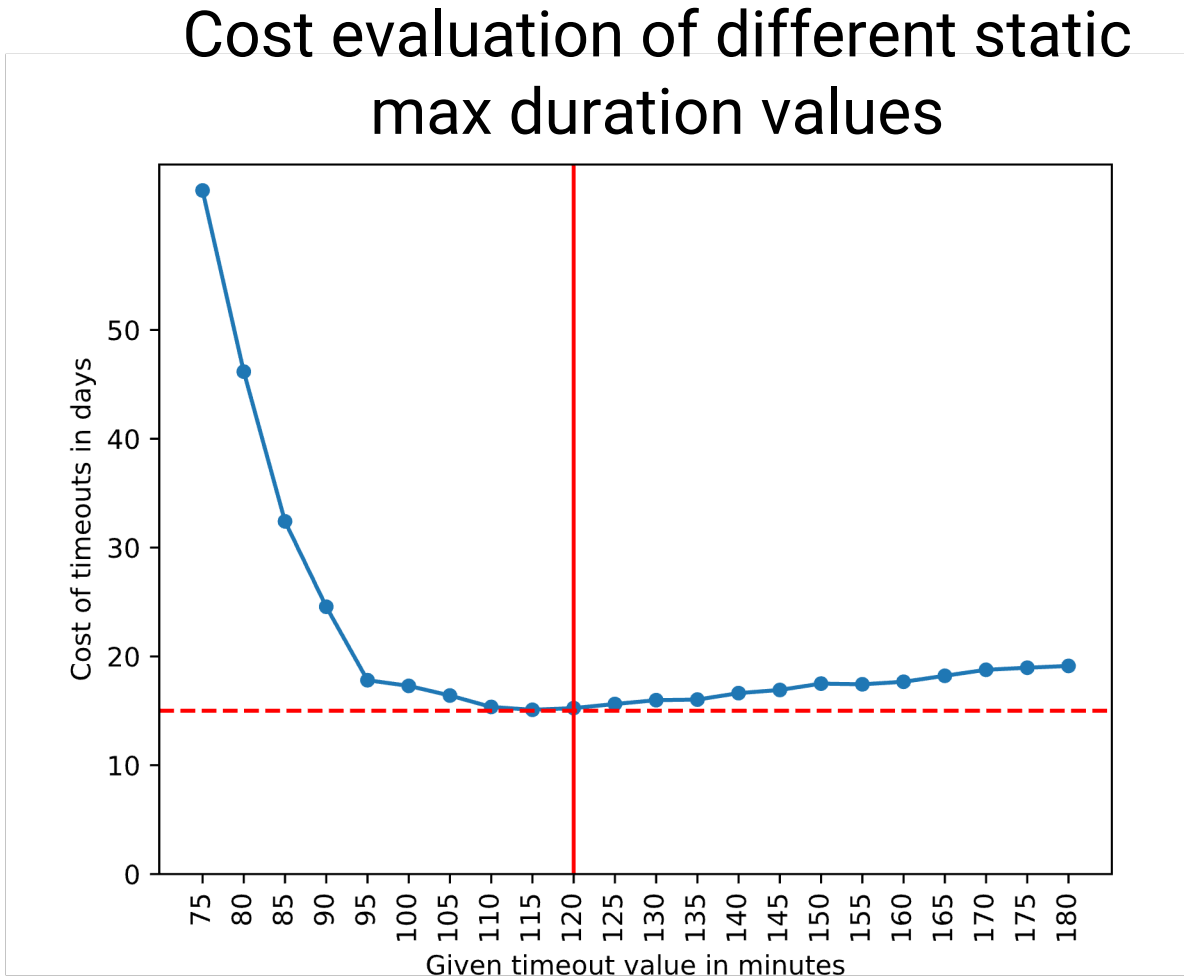
RQ4: *To what degree can we optimize the timeout values with respect to their average test execution cost?*

- **recap**: Increasing timeout values reduces timeout flakiness
- **but**: also allows for longer test execution times, e.g., hanging tests
- **trade-off** between ***average execution time*** and the **probability of a flaky timeout**.

Evaluation

RQ4: To what degree can we optimize the timeout values with respect to their average test execution cost?

- identify cost-optimal static timeout value
- evaluate on Adjusted Timeout Value dataset
- cost-optimal value of **2 hours** reduces timeout flakiness **by 99.5%**

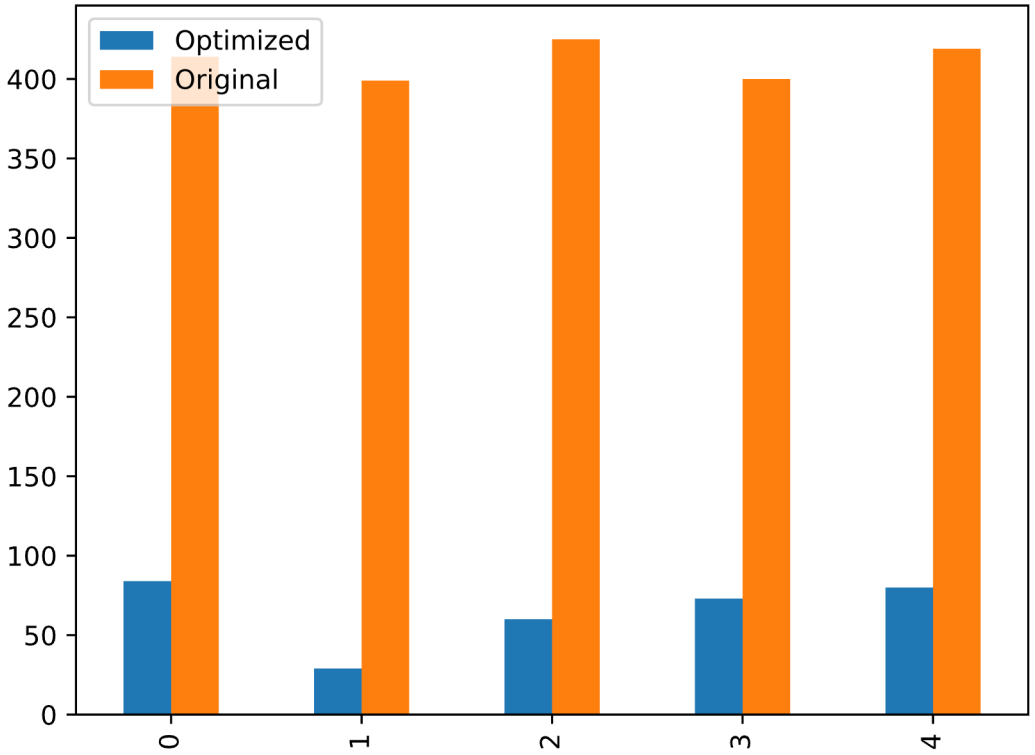


Evaluation

RQ4: *To what degree can we optimize timeout values with respect to average test cost?*

- model trade-off as an **optimization problem**
- calculate **dedicated timeout value** for every test

Comparison of resulting number of timeouts

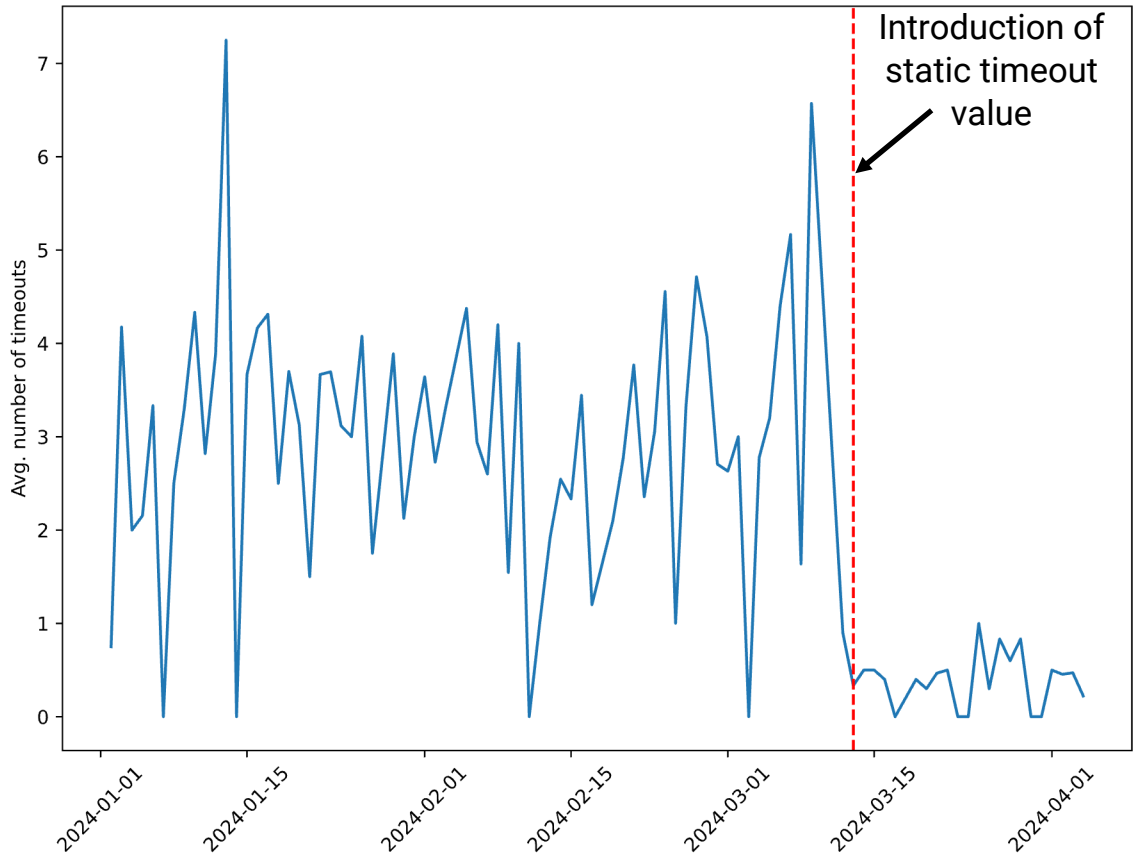


Answer: Our optimization approach **reduces timeout flakinesss by 80%** while **reducing the median timeout value** from 15 to 11 min.

Static timeout values in practice

- started with roll out **two weeks** ago
- **collect information** on effects of global timeout value
- notably **less timeouts** since introduction
- roll out to **main development branches** currently being discussed

Number of timeouts / test run / day



Conclusion

- flakiness definition has **little practical use (RQ1)**
- flakiness rate **converges to 1** when test repetitions go towards infinity **(RQ1)**
- timeout values can cause **additional cost**
- cost-optimal timeout values **can increase testing efficiency (RQ2, RQ4)**
  - **but:** complex implementation for optimization
  - launched project to implement **static global timeout value of 2 hours**
- **Contact:** Alexander Berndt alexander.berndt@students.uni-mannheim.de

We are hiring!

- Sebastian Baltes currently has open positions for PhD students at the University of Bayreuth, Germany.
- **Contact:** Sebastian Baltes (sebastian.baltes@uni-bayreuth.de)