Duet Benchmarking

Improving Measurement Accuracy in the Cloud

Lubomír Bulej
Vojtěch Horký
Petr Tůma

François Farquet
Aleksandar Prokopec

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

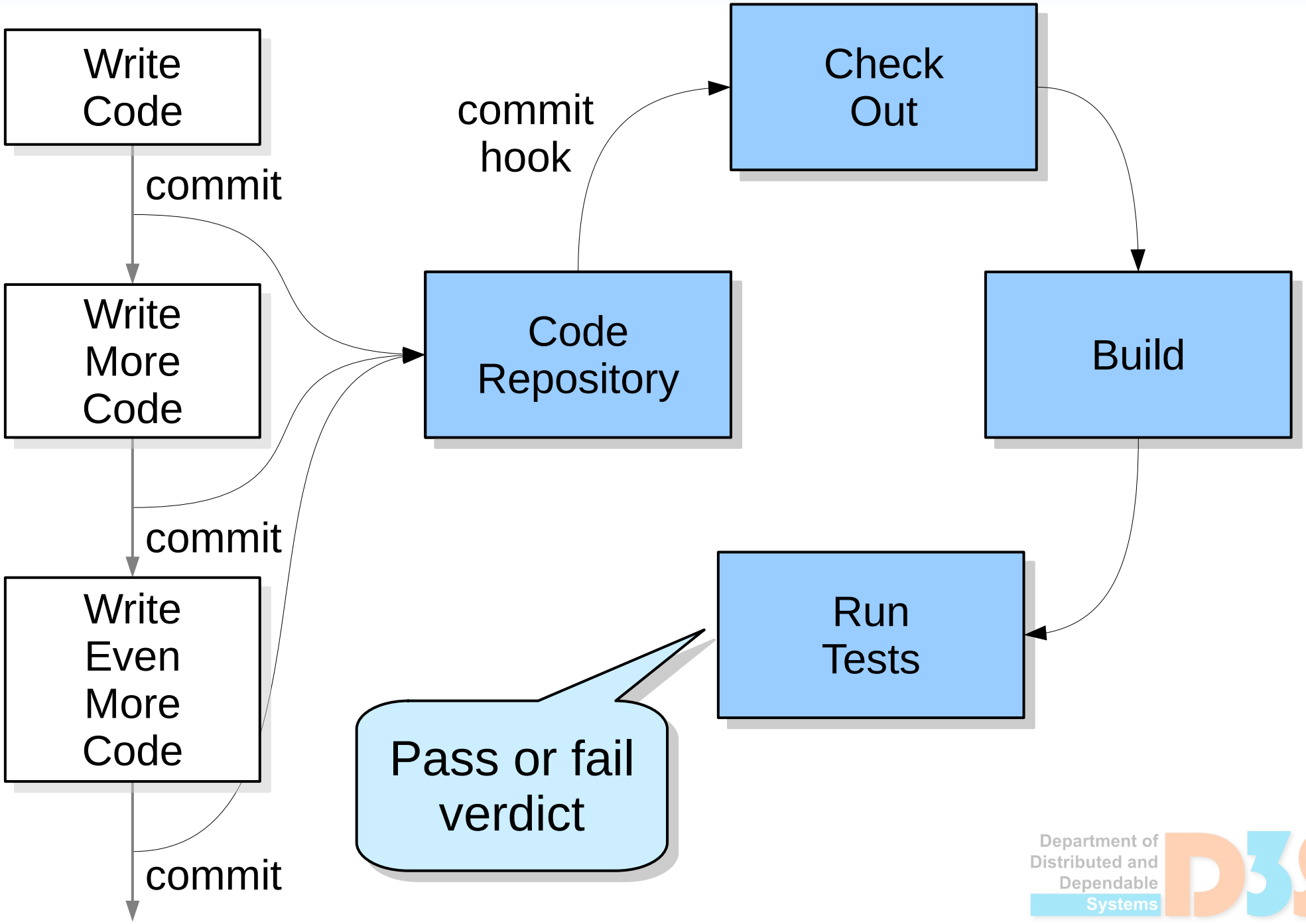
ORACLE®
Labs

Software Regression Testing …

… of Performance

Department of
Distributed and
Dependable
Systems

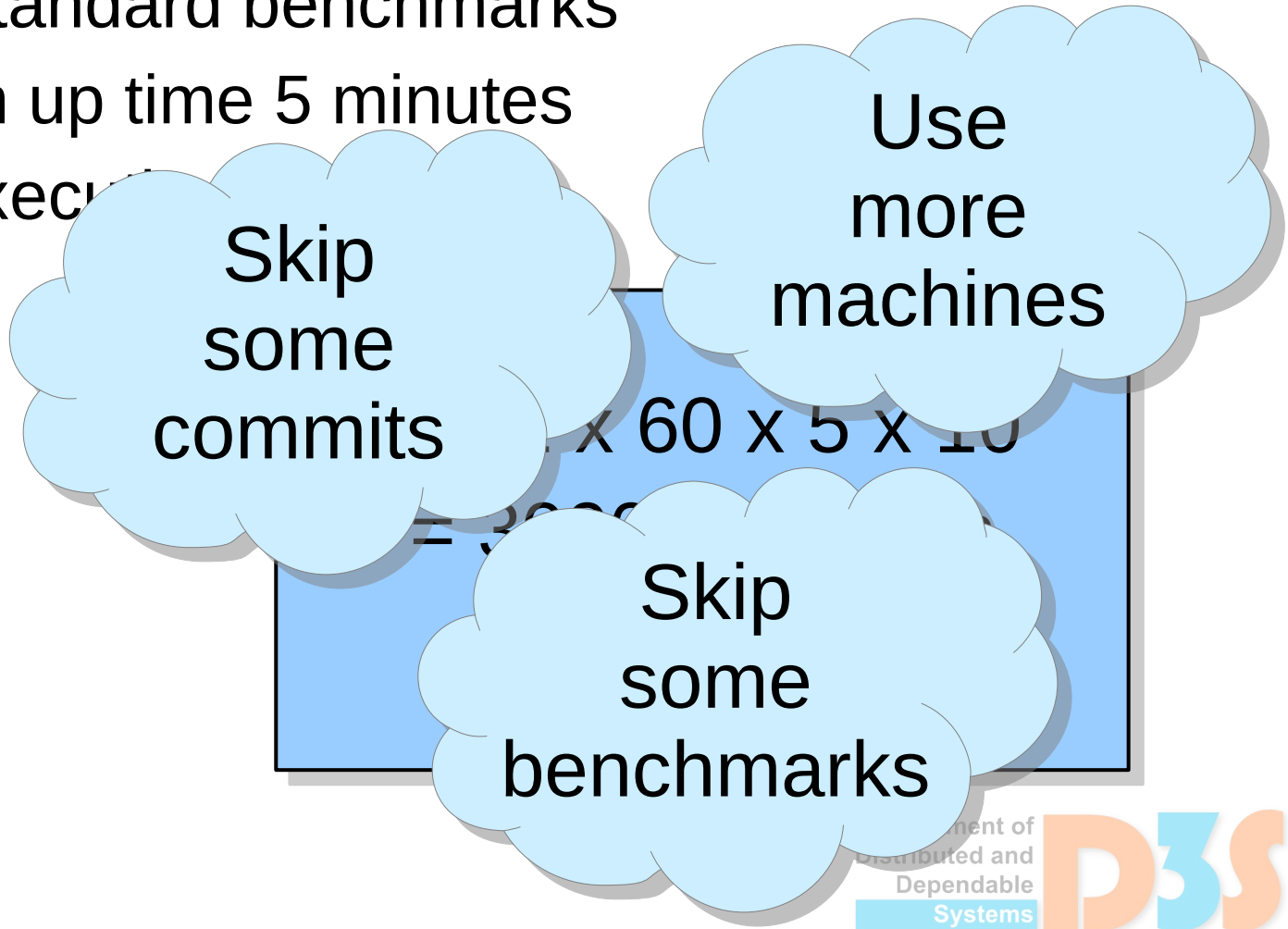
Common Testing Pipeline



Project Context

Graal Java JIT+AOT Compiler

- Currently ~5 merge commits per day
- Bare minimum testing JDK 8 + JDK 11
- Running ~60 standard benchmarks
- Minimum warm up time 5 minutes
- Minimum 10 execu



Where to Go for More Machines ?

… to the Cloud !

Cloud Resource Sharing

Amazon Elastic Cloud Instance Types

- **t3.nano** 2 vCPU @ 5% power, 512MB RAM
- **t3.medium** 2 vCPU @ 20% power, 4GB RAM
- **m5.large** 2 vCPU 8GB RAM
- ...

Or you can forgo the virtualizatio

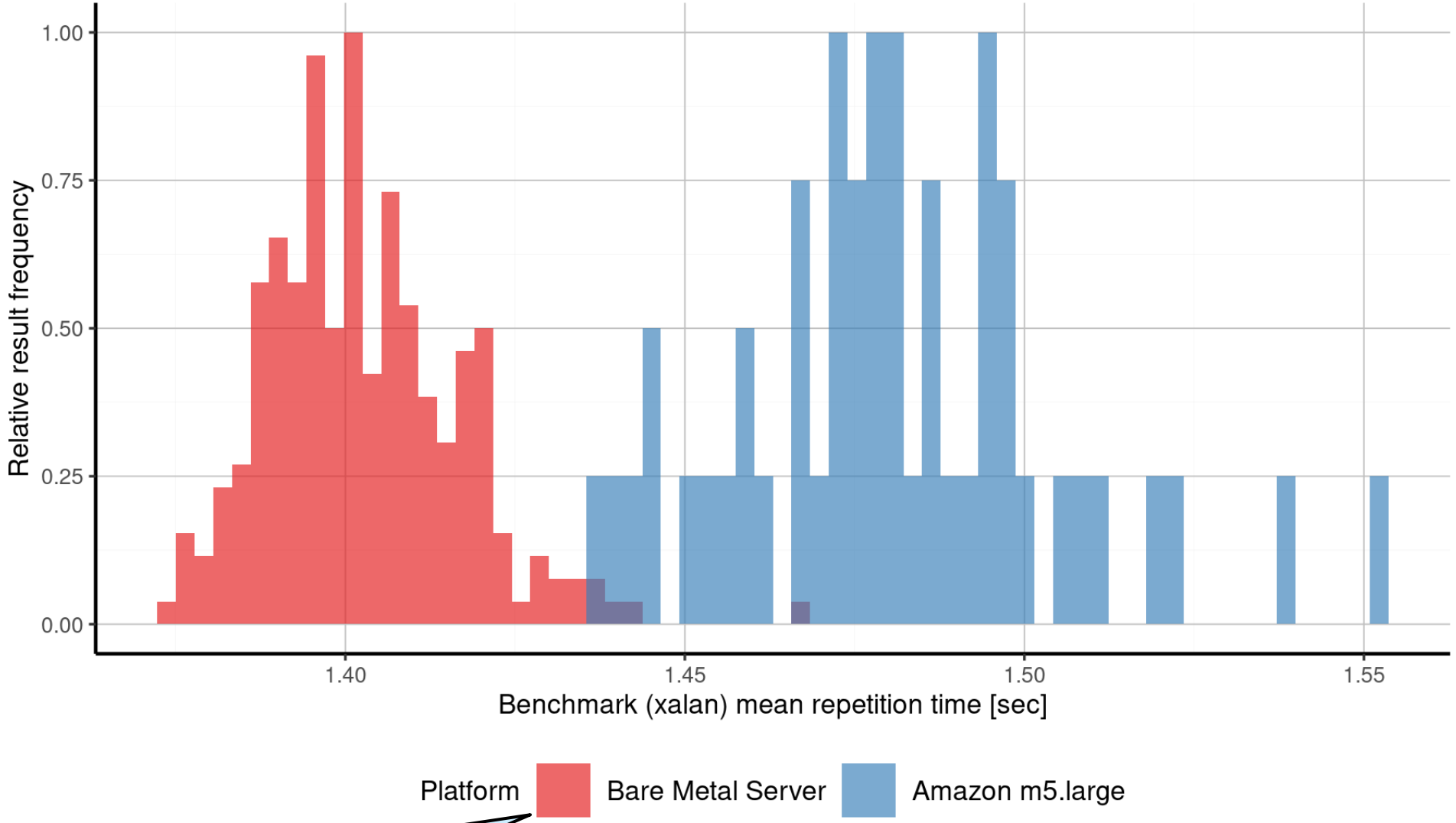
- **m5.metal** 96 threads 48 cores 3
- Likely the same Intel Xeon Platinu 5M

This might perhaps somewhat disrupt measurements

Envelope estimate

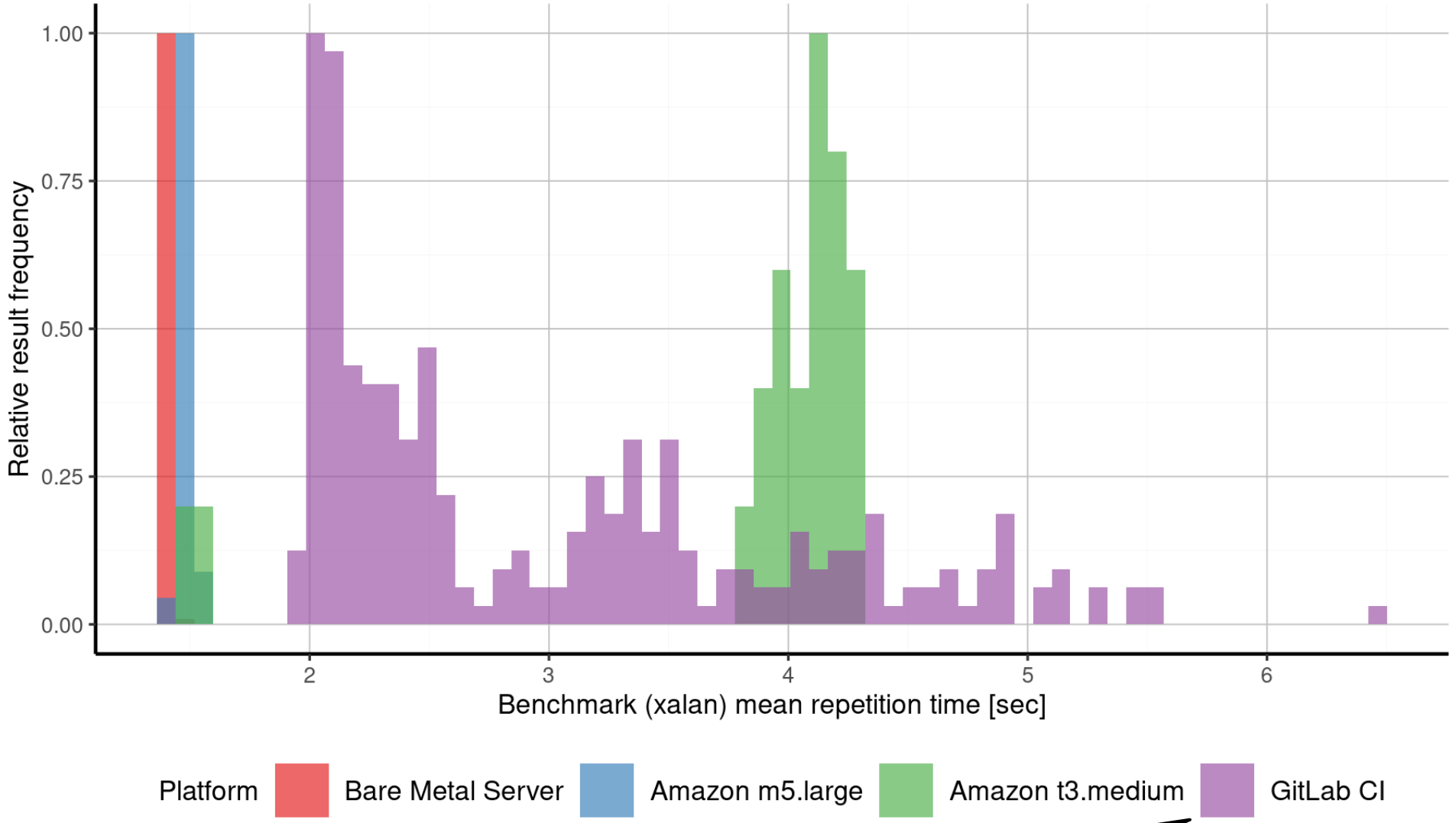
- CPU 48 cores / 5% = **960 instances**
- RAM 384 GB / 512 MB = **768 instances**

… Effect of Resource Sharing



99% CI for the mean is **~61%** bigger

… Effect of Resource Sharing



99% CI for the mean is **~1800%** bigger

Resource Management ...

… Should Be Fair !

Is Resource Management Fair ?

Hyperthreading

- Intel says it *“maximizes use of execution units”*

Bursty processor scheduling

- Amazon says *“one CPU credit is equal to 100% utilization for one minute”* (in any combination) and *“credits are accrued and spent at millisecond resolution”*

Memory caches ?

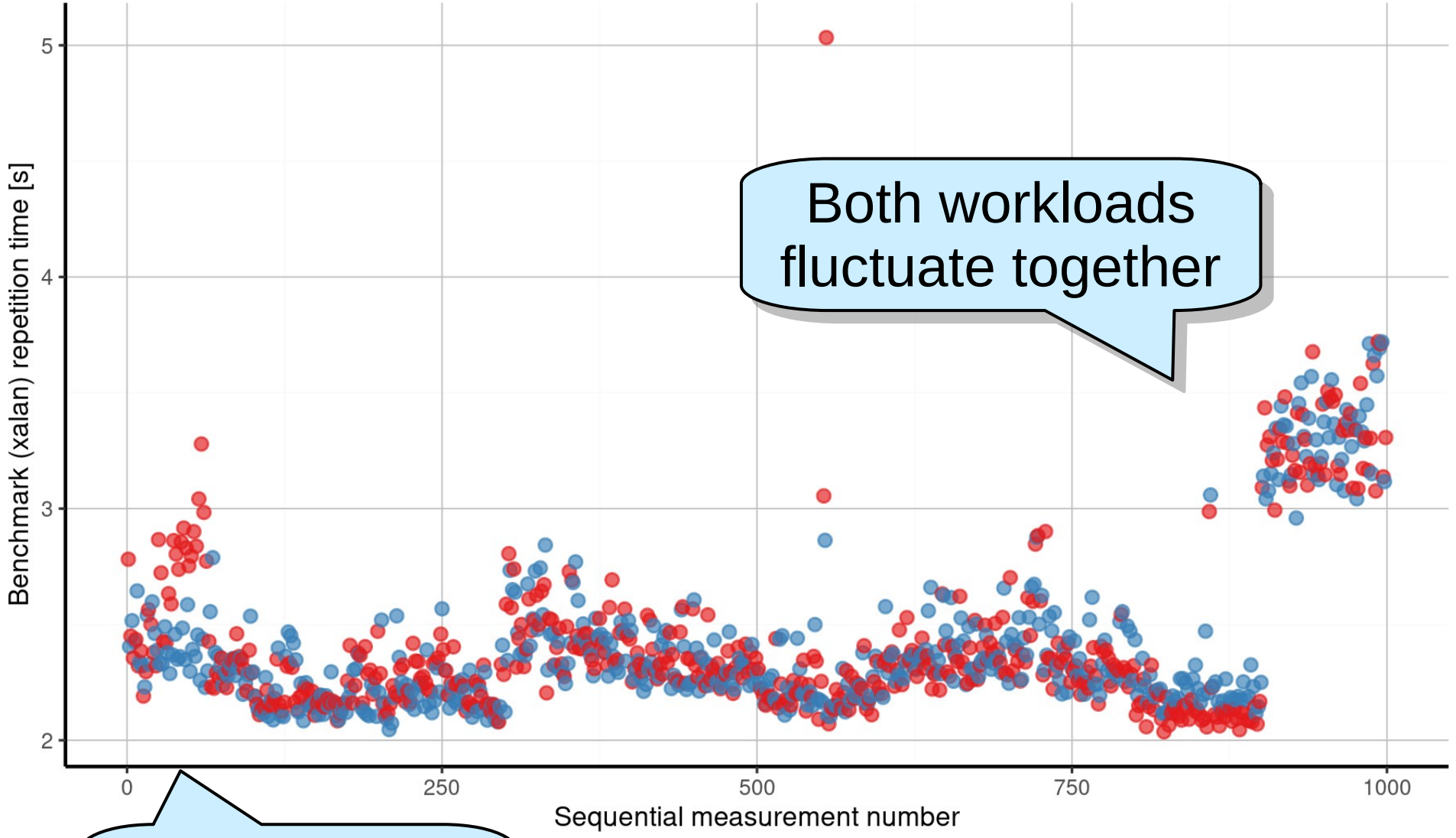
Memory bandwidth ?

Thermal budget ?

Would it be fine if some instances were **systematically** disadvantaged ?

Department of Distributed and Dependable Systems

Two Measurements In Parallel



How To Use This ?

Look at ratios instead of absolute values

- Assumes effects are multiplicative
- Ratios are what people want to know

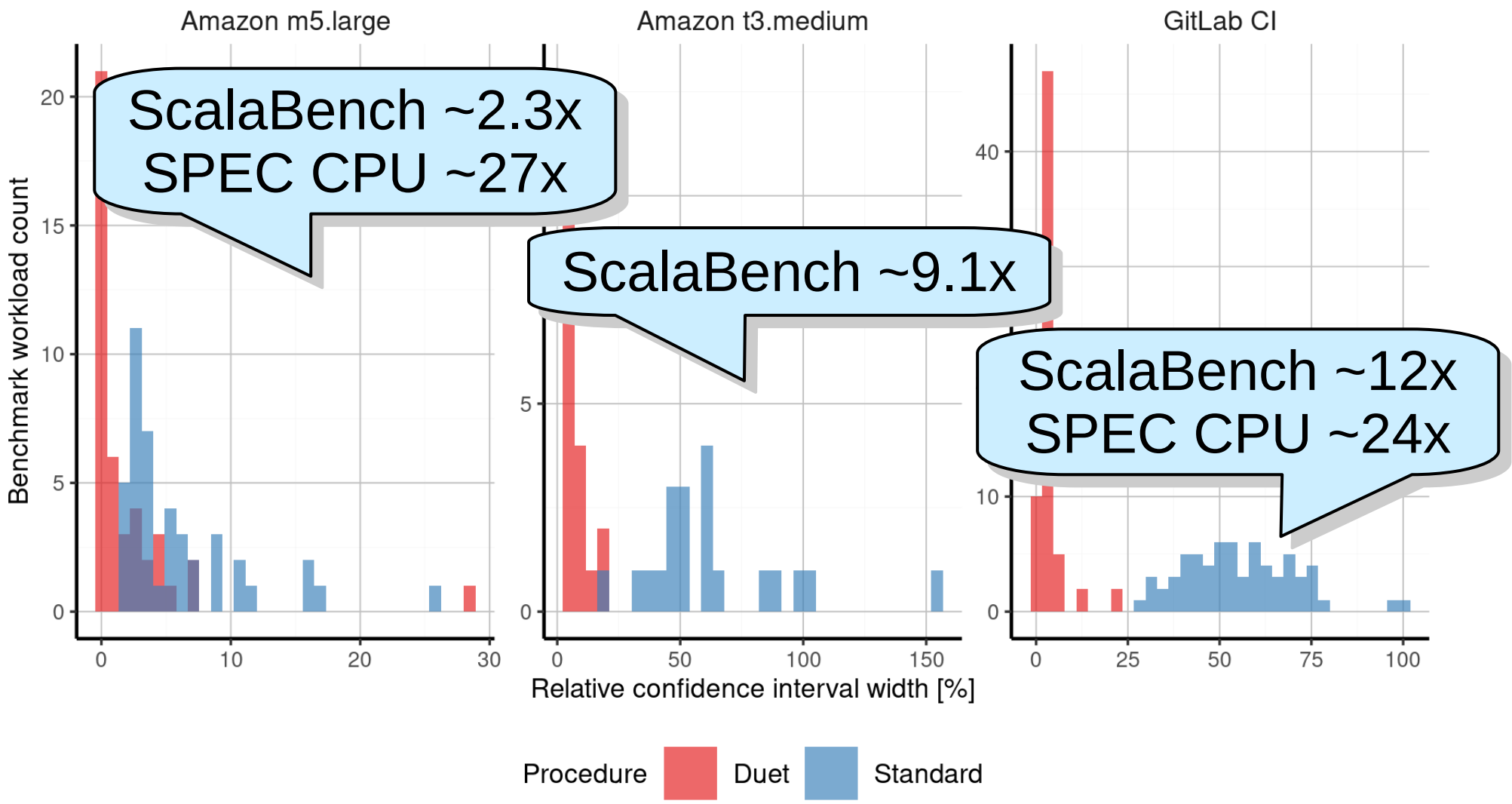
“We want to reliably detect 5% slowdowns …”

Confidence intervals using bootstrap

Compare with sequential measurements

- Confidence interval width relative to mean
- Not quite apples-to-apples but gives some intuition

How Much More Accurate ?



… More Done

Does duet benchmarking work
because of **synchronized interference** ?

Does duet benchmarking address
interference due to **resource sharing** ?

Does duet benchmarking measure
performance differences **accurately** ?

…

Thank You !

Complete paper at https://arxiv.org/abs/2001.05811

For more information visit http://d3s.mff.cuni.cz

Department of
Distributed and
Dependable
Systems