File System Trace Replay Methods Through the Lens of Metrology

Thiago Emmanuel Pereira, Francisco Vilar Brasileiro, Lívia Sampaio
temmanuel@copin.ufcg.edu.br, fubica@computacao.ufcg.edu.br, livia@computacao.ufcg.edu.br

Federal University of Campina Grande,
Brasil

Very popular but ...

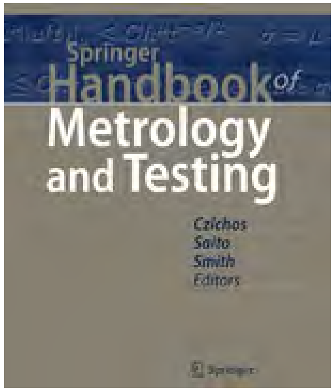
- Many ad hoc trace replay tools – no description about their design and implementation
    - Impossible to reproduce results.

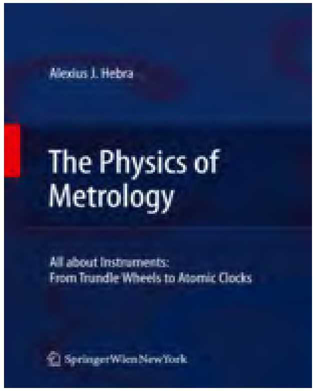
"How to do this accurately is still an open question, and the best we can do right now is take results with a degree of skepticism" - Traeger, A., Zadok, E., Joukov, N., & Wright, C. P. (2008). A nine year study of file system and storage benchmarking. *ACM Transactions on Storage (TOS)*

Before creating new methods, how good are current trace based methods?

Our take

A metrology case study

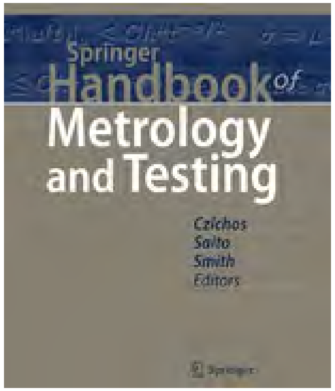


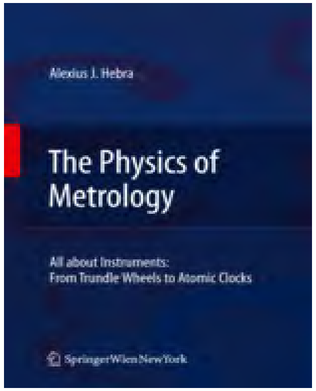


Our take

A metrology case study

Single-laboratory





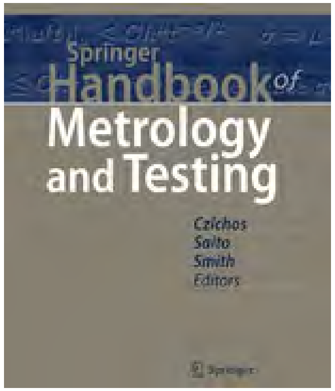
Our take

A metrology case study

Single-laboratory

Inter-laboratories

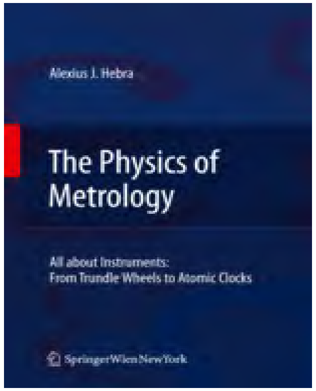
Different operators
Different instruments
Different environment



Springer Handbook of Metrology and Testing

Czichos
Saito
Smith
Editors

Springer



Alexius J. Hebra

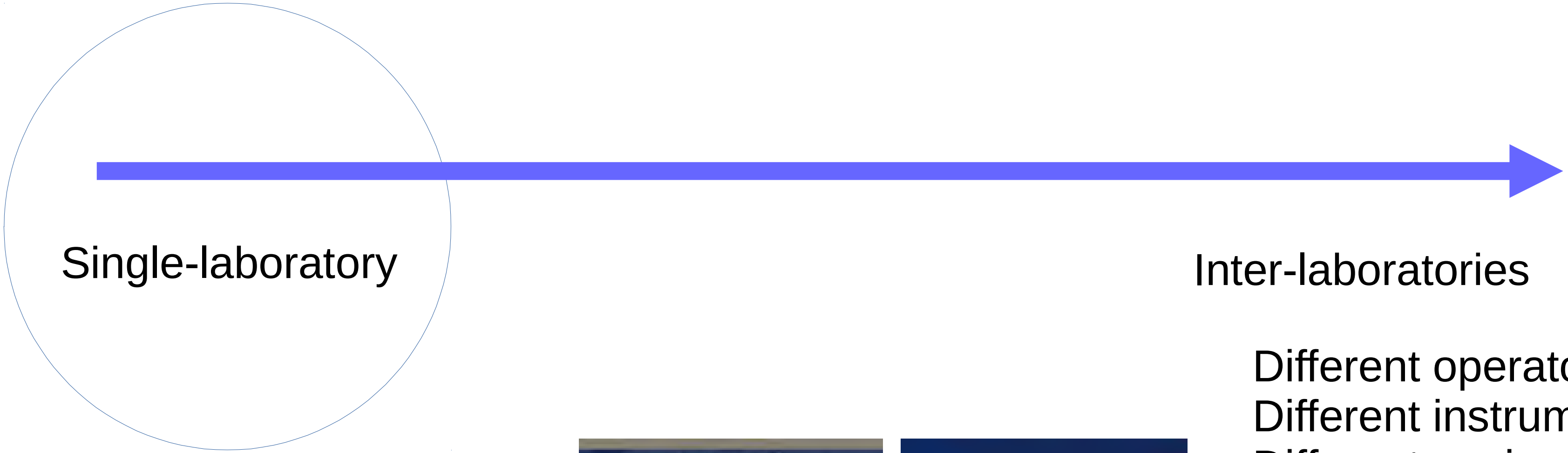
The Physics of Metrology

All about Instruments:
From Trundle Wheels to Atomic Clocks

SpringerWienNewYork

Our take

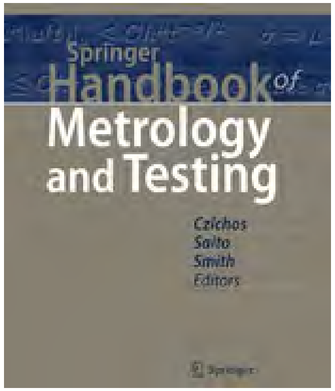
A metrology case study



Single-laboratory

Inter-laboratories

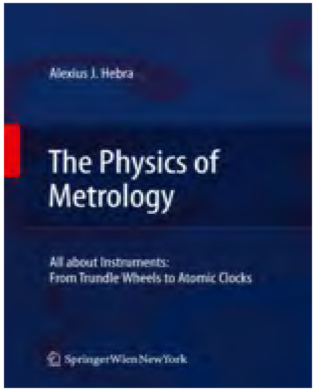
Different operators
Different instruments
Different environment



Springer Handbook of Metrology and Testing

Czichos
Saito
Smith
Editors

Springer



Alexius J. Hebra

The Physics of Metrology

All about Instruments:
From Trundle Wheels to Atomic Clocks

SpringerWienNewYork

Single-lab testing

1. Define the measurand
   - The quantity intended to be measured
2. Specify the measurement procedure
3. Identify the uncertainty sources
4. Conduct the measurement characterization
   - In terms of bias, precision, sensitivity, resolution, etc.
5. Perform the calibration (or mitigation of measurement errors)
6. Calculate the measurement uncertainty
   - An interval **[y – u, y + u]** within the true value of measurand **y** are expected to be.

Measurand

File system response time

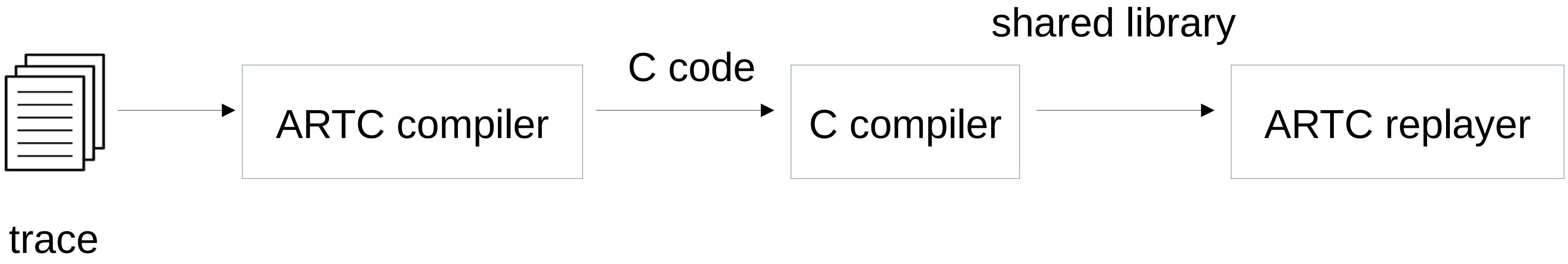
Measurement procedure

Instruments

- ARTC replayer (compilation-based)
- TBBT replayer (event-based)

1. *Weiss, Zev, et al*. **"Root: Replaying multithreaded traces with resource-oriented ordering."** SOSP. ACM, 2013.
2. Zhu, Ningning, Jiawu Chen, and Tzi-Cker Chiueh. **"TBBT: scalable and accurate trace replay for file server evaluation."** FAST,2005.

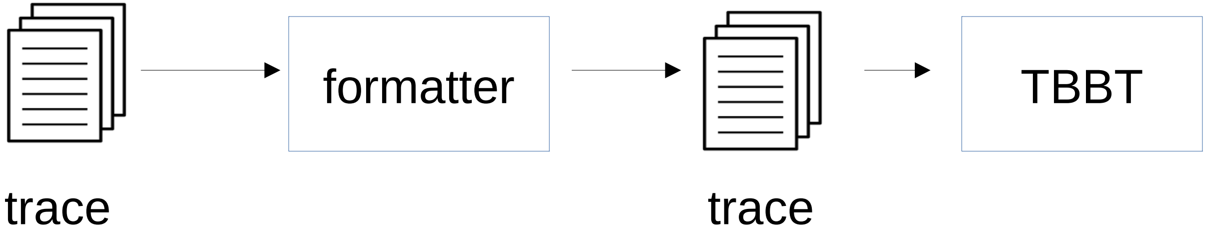
ARTC Replayer



1. *Weiss, Zev, et al*. **"Root: Replaying multithreaded traces with resource-oriented ordering."** SOSP. ACM, 2013.

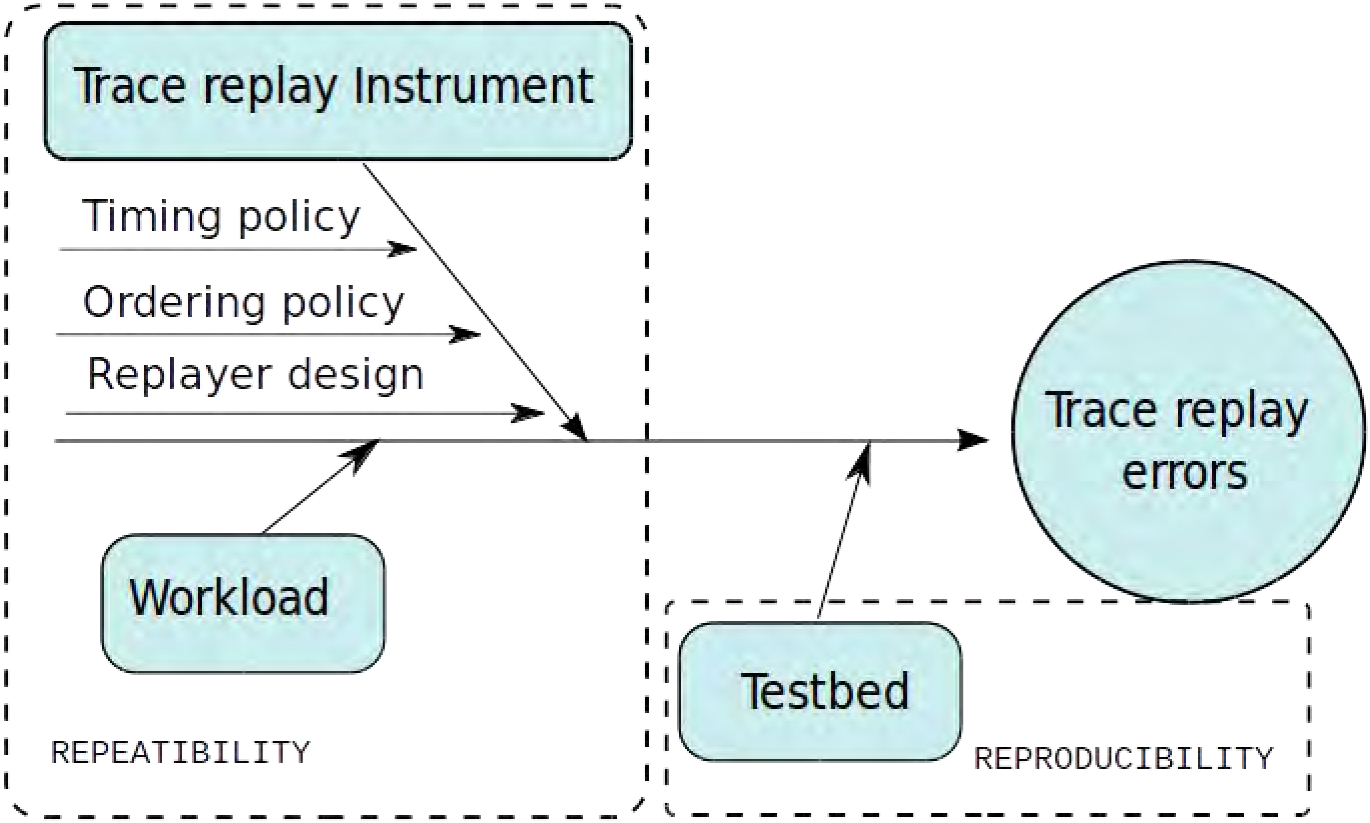
TBBT Replayer

Based on TBBT design, running as a real time process to be less **sensitive**

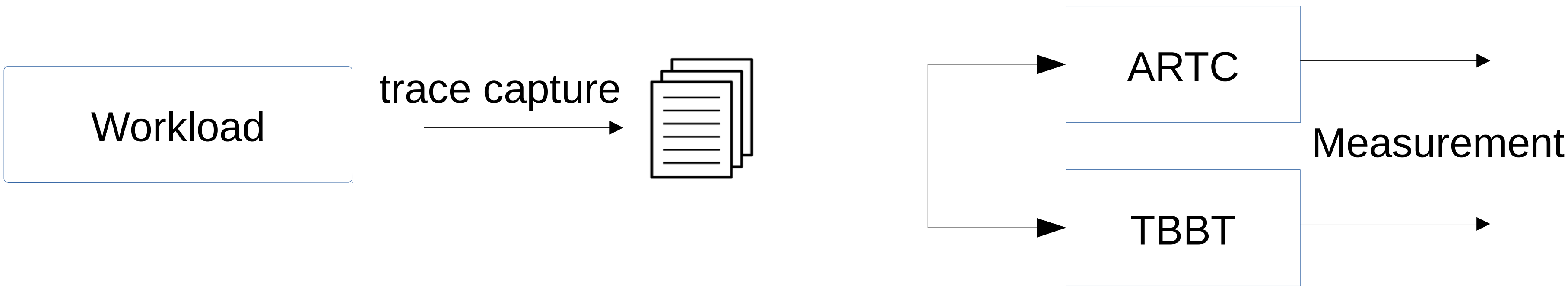


1. Zhu, Ningning, Jiawu Chen, and Tzi-Cker Chiueh. **"TBBT: scalable and accurate trace replay for file server evaluation."** FAST,2005.
2. Tarihi, Mojtaba, Hossein Asadi, and Hamid Sarbazi-Azad. **"DiskAccel: Accelerating Disk-Based Experiments by Representative Sampling."** SIGMETRICS , 2015.

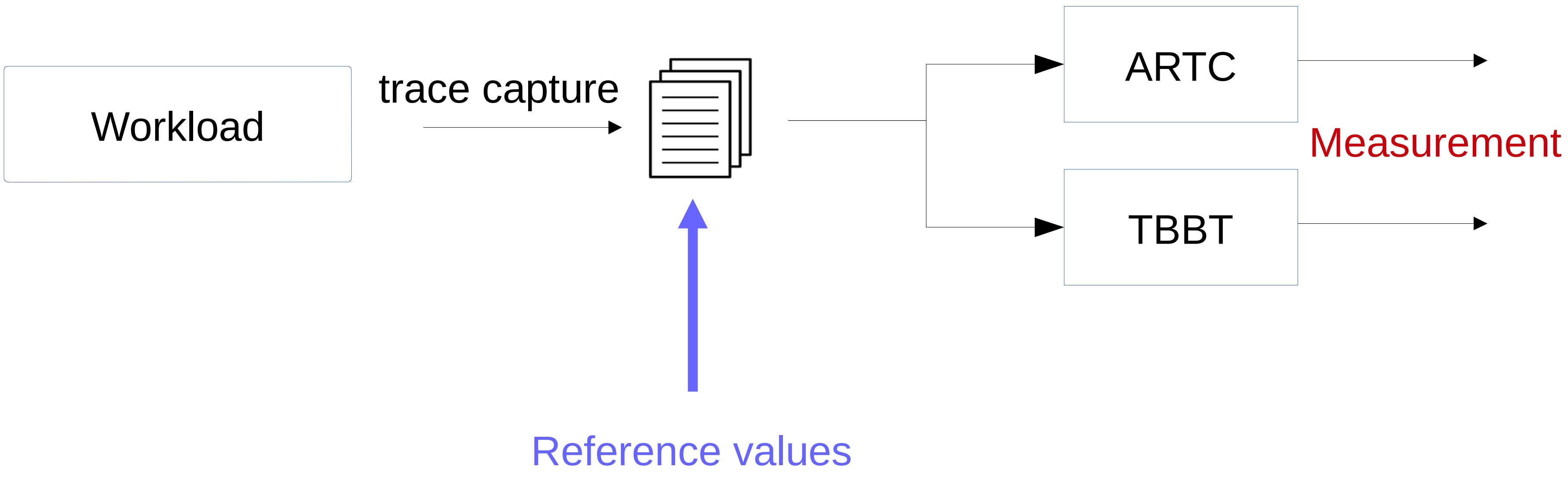
Uncertainty sources



Characterization

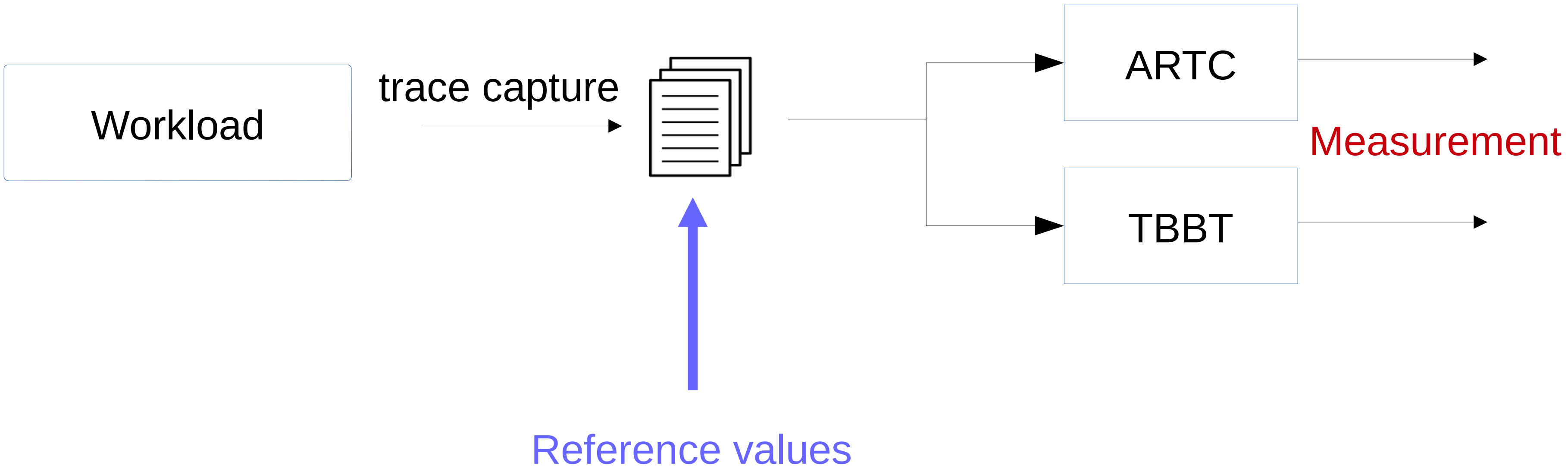


Characterization



Characterization

- Microbenchmark (5k ops, 4k chunks, [1-4] threads)
  - Random read (RR), Random write (RW)
  - Sequential read (SR), Sequential write (SW)
- Filebench fileserver workload



Workload

trace capture

ARTC

TBBT

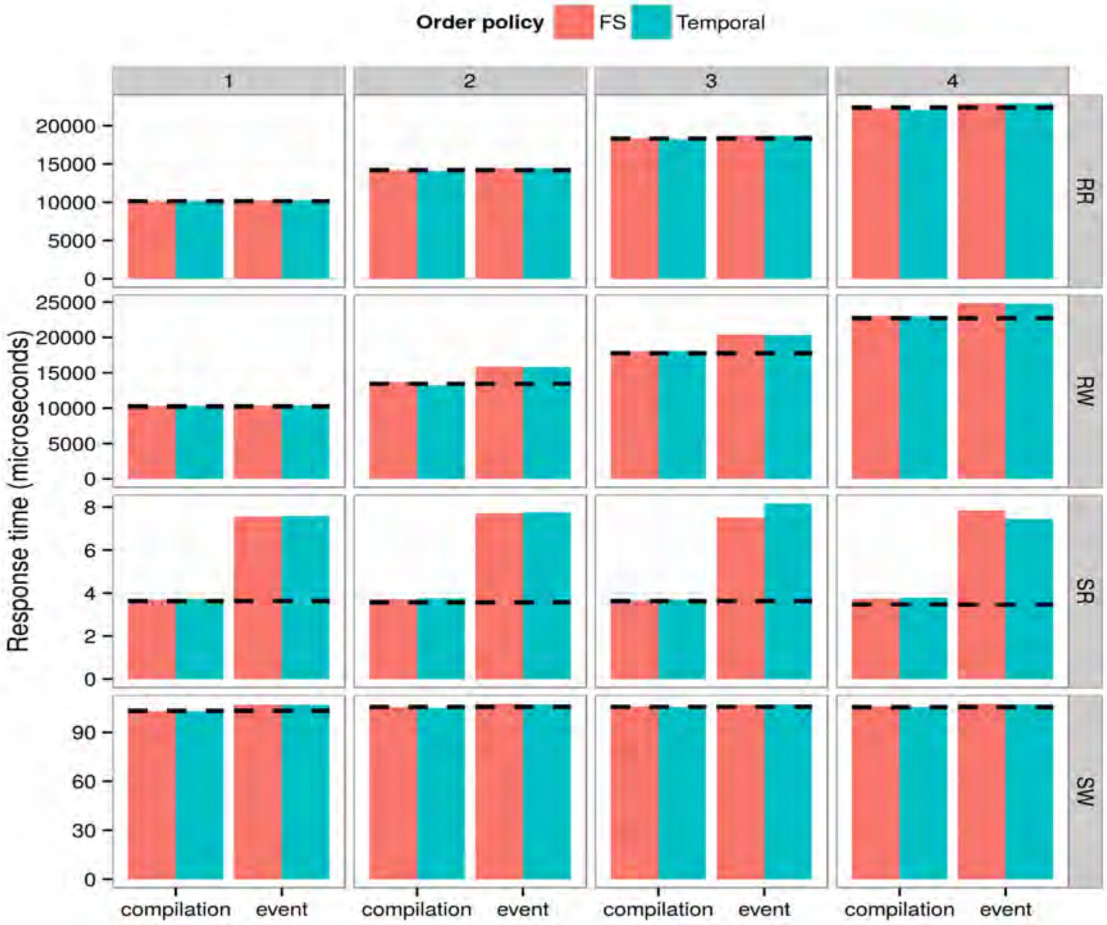
Measurement

Reference values

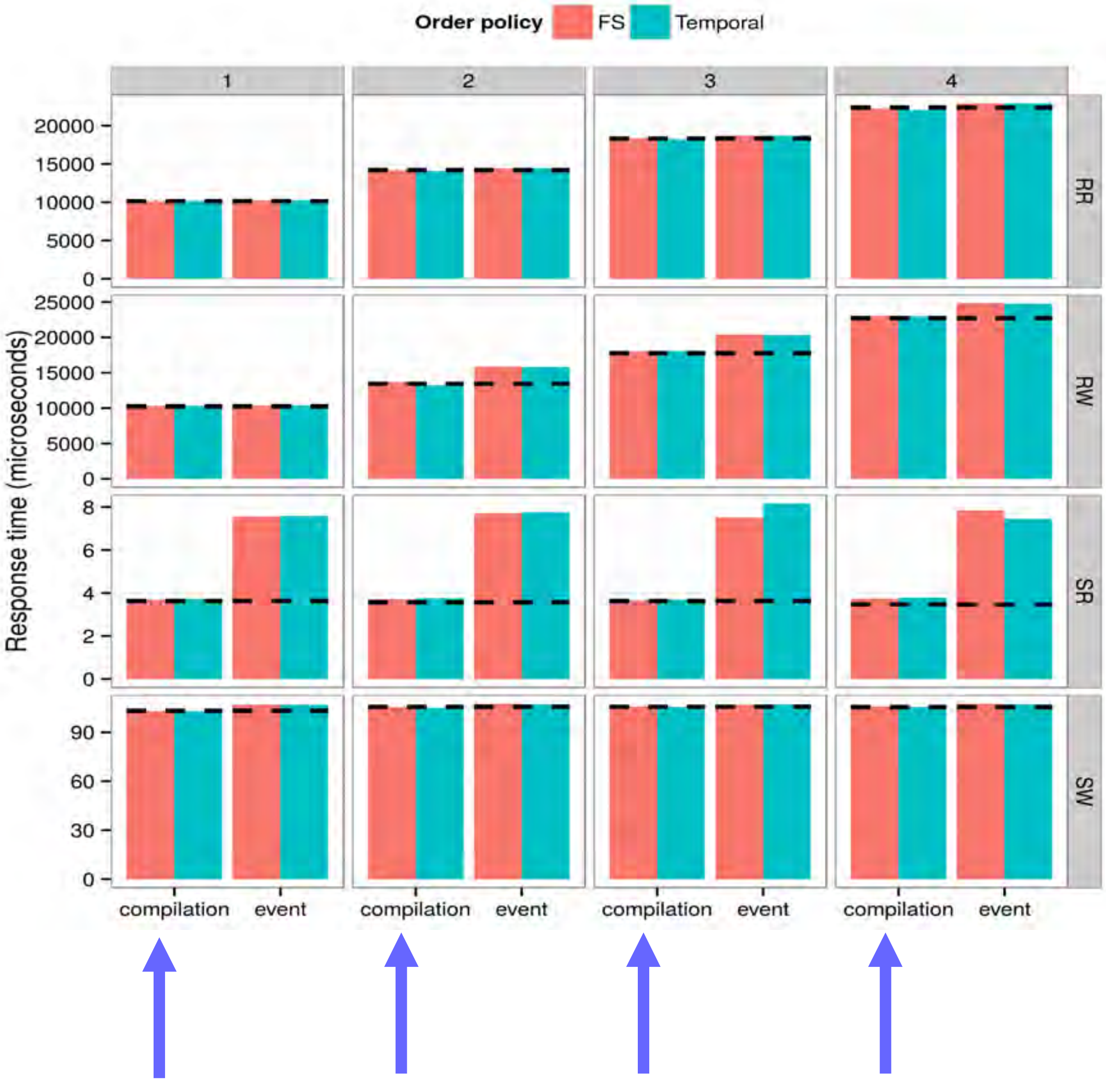
Characterization

Microbenchmark

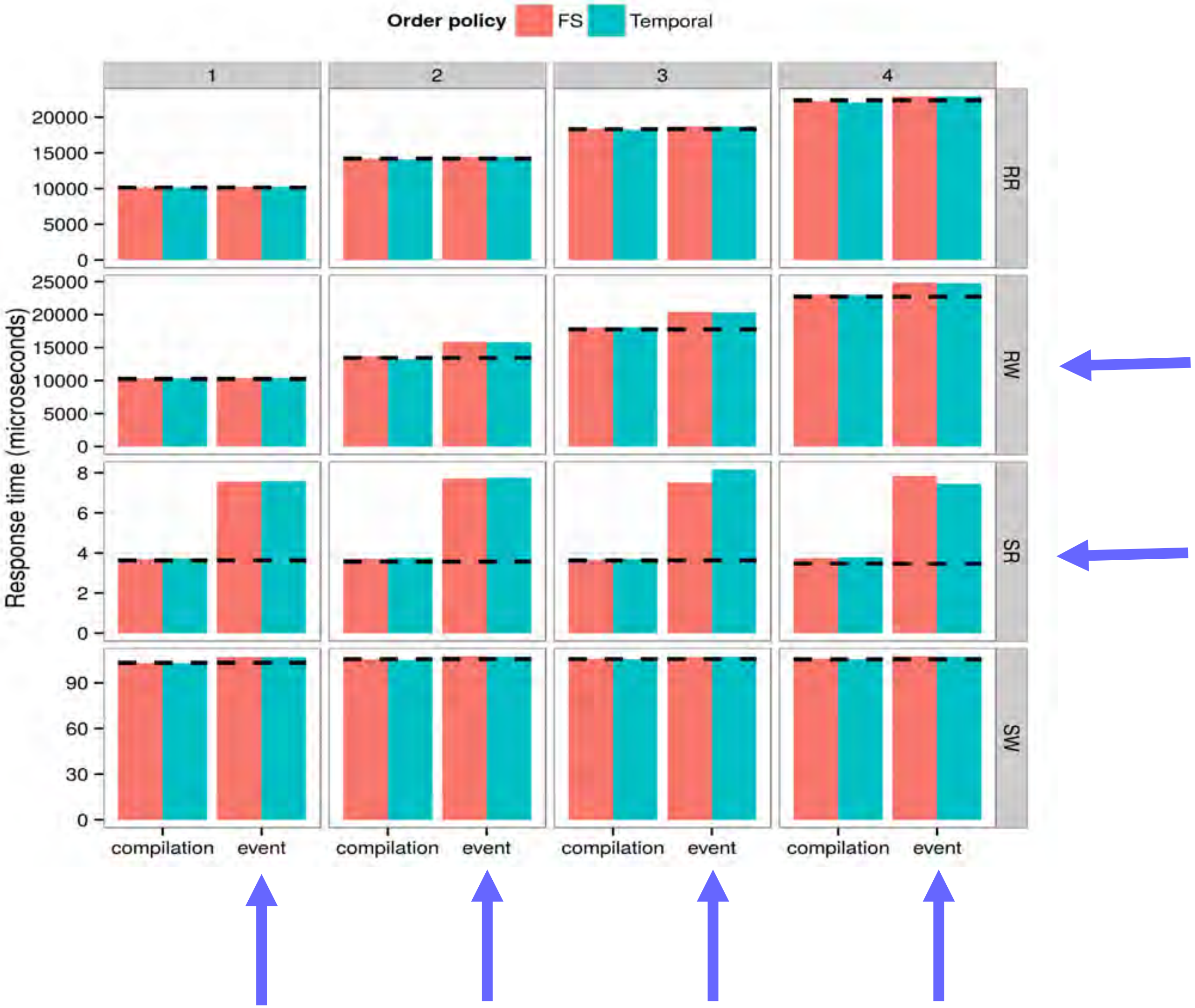
Bias characterization



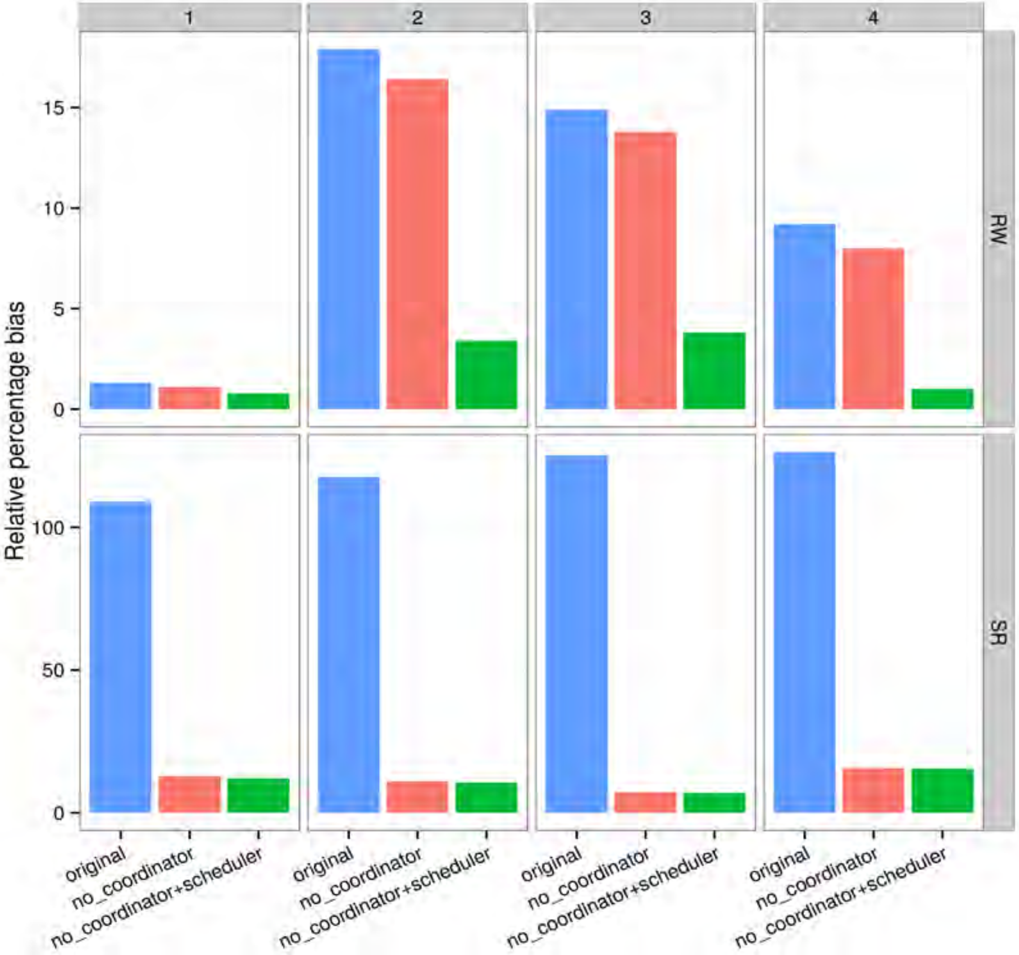
Bias characterization



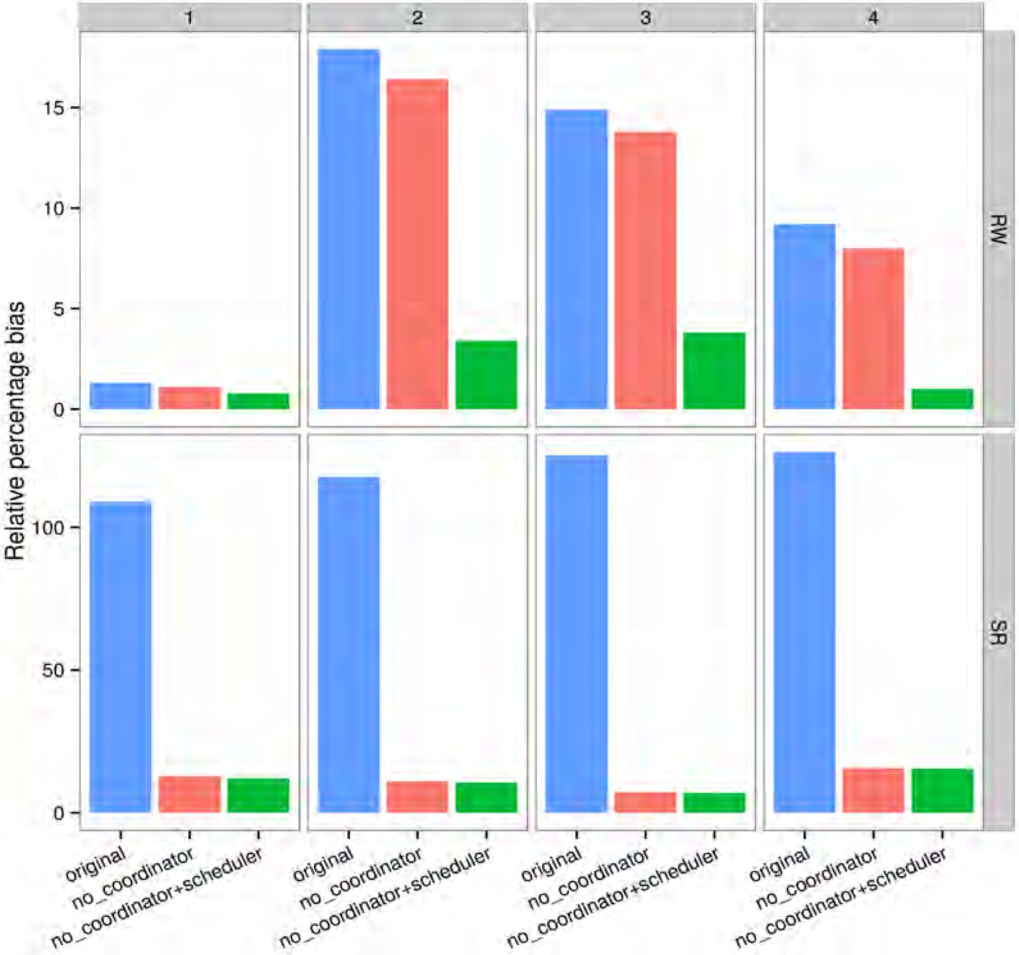
Bias characterization



TBBT improvements



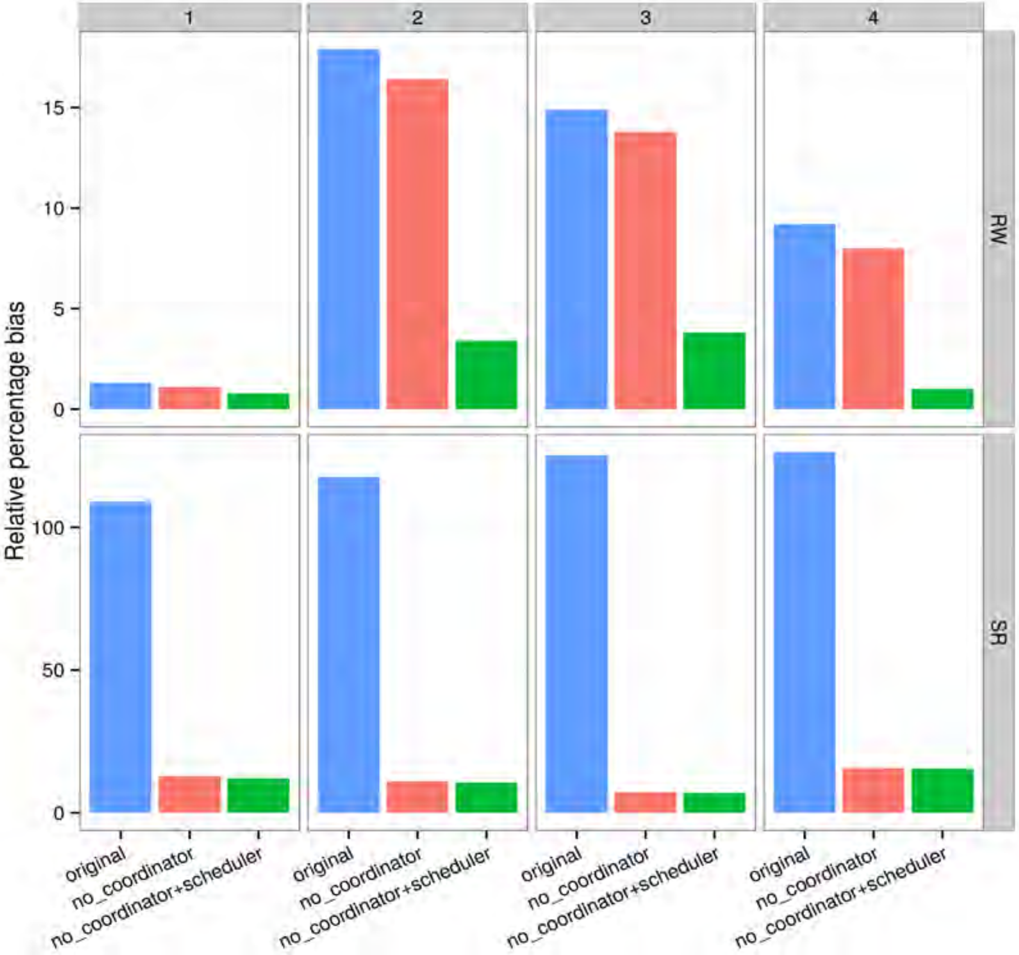
TBBT improvements



TBBT coordinator overhead

TBBT coordinator overhead

TBBT improvements



TBBT coordinator overhead

Real time scheduler

TBBT coordinator overhead

Uncertainty

Workload	TBBT	ARTC
Random read	22579.0 ± 2.4% (22891.6 ± 4.8%)	22243.5 ± 1.8%
Random write	22946.1 ± 3.2% (24807.6 ± 18%)	23076.0 ± 4.1%
Sequential read	4.0 ± 32.9% (7.8 ± 253%)	3.7 ± 18.6%
Sequential write	105.6 ± 1.3% (107.7 ± 4.2%)	105.8 ± 0.6%

Uncertainty

Before TBBT improvements ARTC is a clear winner

Workload	TBBT	ARTC
Random read	22579.0 ± 2.4% (22891.6 ± 4.8%)	22243.5 ± 1.8%
Random write	22946.1 ± 3.2% (24807.6 ± 18%)	23076.0 ± 4.1%
Sequential read	4.0 ± 32.9% (7.8 ± 253%)	3.7 ± 18.6%
Sequential write	105.6 ± 1.3% (107.7 ± 4.2%)	105.8 ± 0.6%

Uncertainty

TBBT improvements are affective

Workload	TBBT	ARTC
Random read	22579.0 ± 2.4% (22891.6 ± 4.8%)	22243.5 ± 1.8%
Random write	22946.1 ± 3.2% (24807.6 ± 18%)	23076.0 ± 4.1%
Sequential read	4.0 ± 32.9% (7.8 ± 253%)	3.7 ± 18.6%
Sequential write	105.6 ± 1.3% (107.7 ± 4.2%)	105.8 ± 0.6%

Uncertainty

How to choose between replayers?

Workload	TBBT	ARTC
Random read	22579.0 ± 2.4% (22891.6 ± 4.8%)	22243.5 ± 1.8%
Random write	22946.1 ± 3.2% (24807.6 ± 18%)	23076.0 ± 4.1%
Sequential read	4.0 ± 32.9% (7.8 ± 253%)	3.7 ± 18.6%
Sequential write	105.6 ± 1.3% (107.7 ± 4.2%)	105.8 ± 0.6%

Characterization

Filebench fileserver workload

Characterization

Filebench fileserver workload

- 4 threads
- creat, delete, append, read, write, stat
- variable file sizes
- Wholefile read and write

Uncertainty

	TBBT	ARTC	Reference
Read	20.73 ± 118.27%	27.21 ± 92.72%	50.72
Write	50.45 ± 79.81%	69.79 ± 33.79%	83.95

Uncertainty

	TBBT	ARTC	Reference
Read	20.73 ± 118.27%	27.21 ± 92.72%	50.72
Write	50.45 ± 79.81%	69.79 ± 33.79%	83.95

Replayed response time appears better than reference

TBBT and ARTC memory footprints are smaller than filebench footprint, thus more cache hits

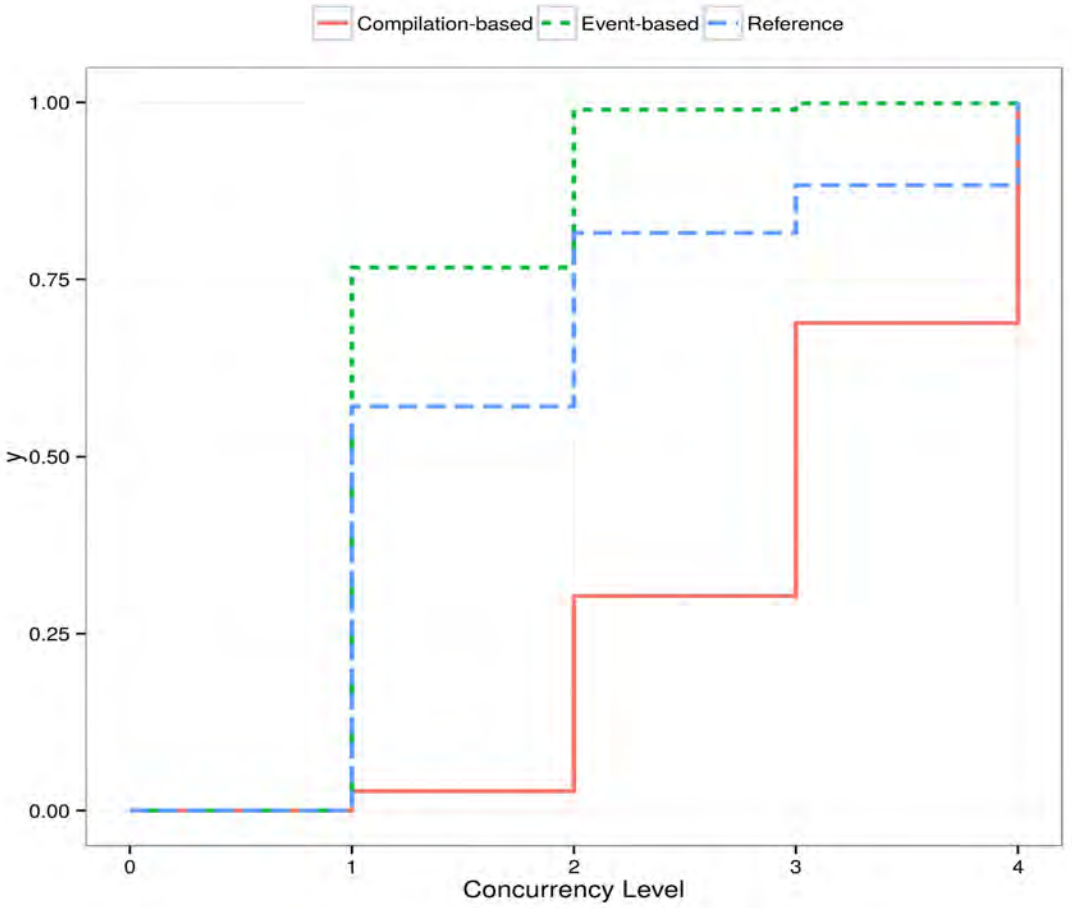
Uncertainty

	TBBT	ARTC	Reference
Read	20.73 ± 118.27%	27.21 ± 92.72%	50.72
Write	50.45 ± 79.81%	69.79 ± 33.79%	83.95

TBBT response time appears better than ARTC response time

Uncertainty

Replayers are not able to match captured workload concurrency



Conclusions

Metrology can help:

- Choosing the best instrument for the job (based on the measurement uncertainty)
    - The TBBT replayer, in some cases, is equivalent to the ARTC replayer.
- Improving tools and best practices
    - Event-based replayer needs improvement
    - Changes in OS scheduler policy may affect sensitive metrics.
- Spotting uncertainty sources
    - Differences in experimental environment, such as the amount of available memory, are likely to hurt reproducibility.