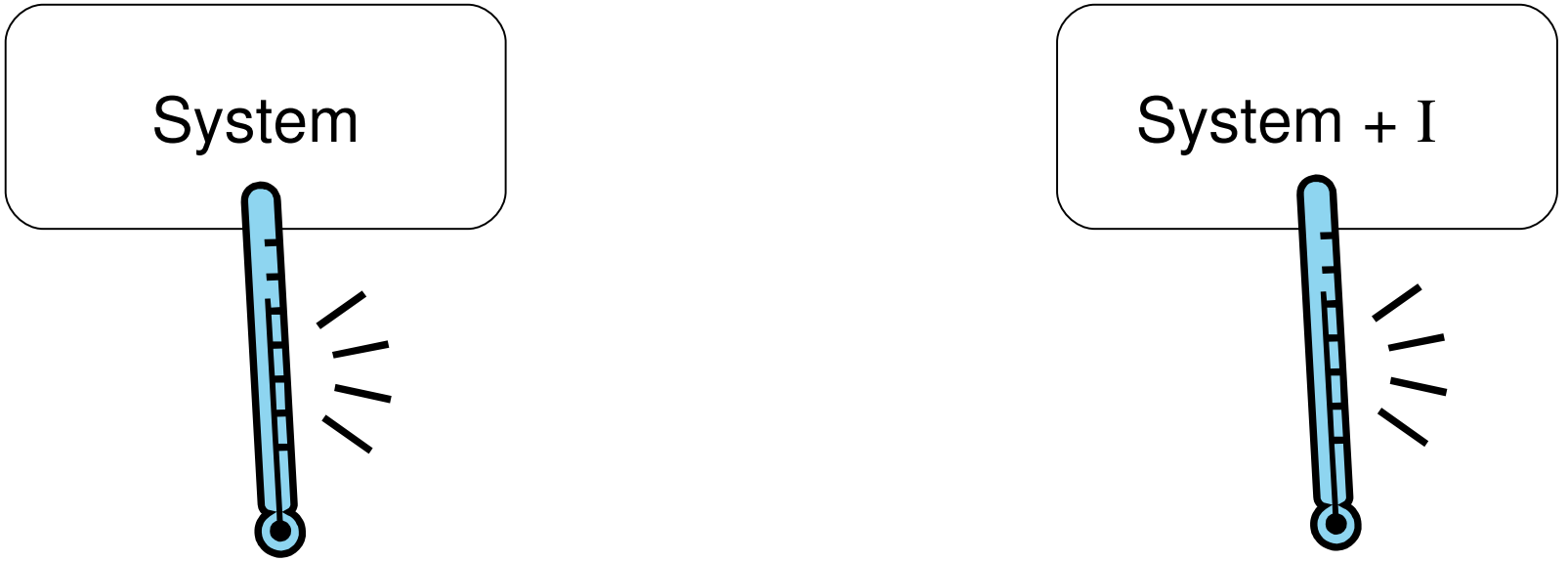
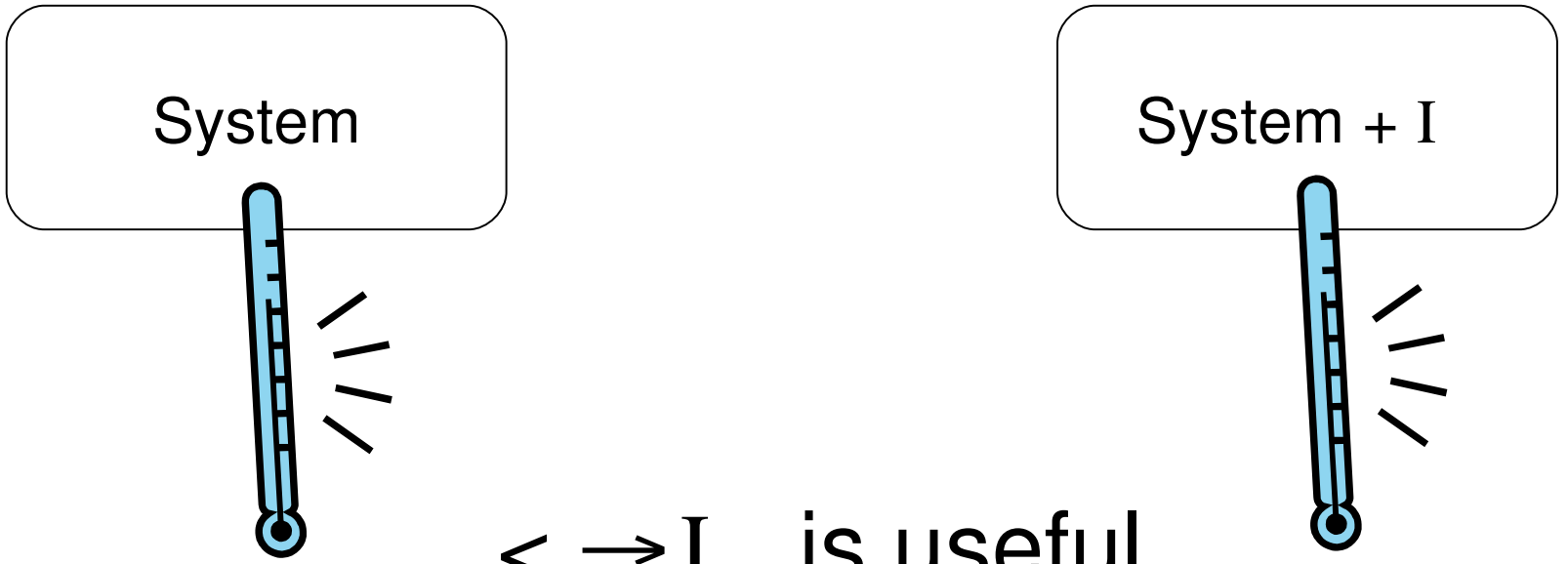
Producing wrong data without doing anything obviously wrong!

Todd Mytkowicz	**:**	**U. Colorado**
Amer Diwan	**:**	**U. Colorado**
Matthias Hauswirth	**:**	**U. Lugano**
Peter F. Sweeney	**:**	**IBM Research**

Evaluating innovations in computer systems



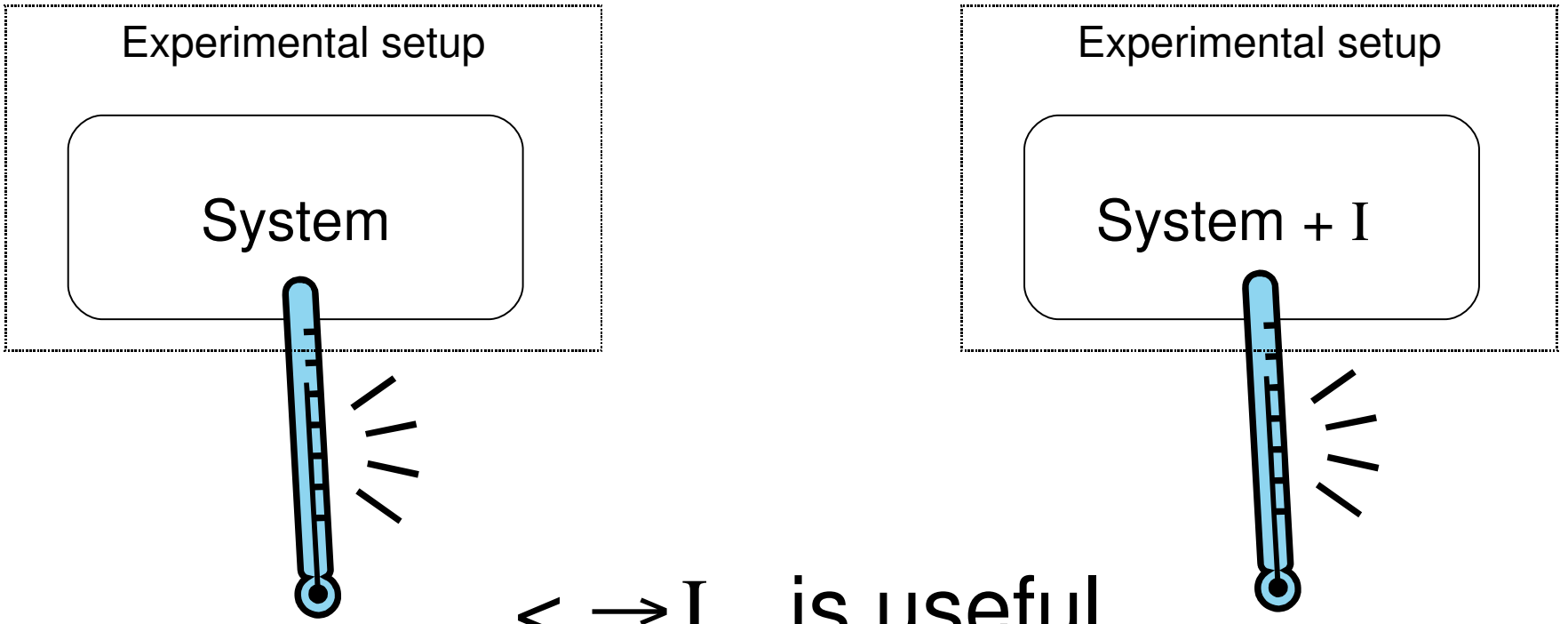
Evaluating innovations in computer systems



$< \rightarrow I$ is useful

$> \rightarrow I$ is not useful

Evaluating innovations in computer systems



$< \rightarrow I$ is useful

$> \rightarrow I$ is not useful

But what if the data are biased? Wrong conclusions!

Methodology

- SPEC CPU 2006 C programs
- Intel Core 2 (2.4GHz)
  - Linux 2.6.25
  - gcc 4.2
  - Papi 3.5.1 / perfmon 2.8
- Best Practices
  - Unloaded machine
  - Multiple runs
  - Confidence intervals

Example of bias in 400.perlbench

System = gcc -O2
System + I = gcc -O3

Example of bias in 400.perlbench

System = gcc -O2
System + I = gcc -O3

Amer:
speedup = 1.18 ± 0.0002

Conclusion: O3 is good

Example of bias in 400.perlbench

System = gcc -O2
System + I = gcc -O3

Amer:
speedup = 1.18 ± 0.0002

Conclusion: O3 is good

Todd:
speedup = 0.84 ± 0.0002

Conclusion: O3 is bad

Example of bias in 400.perlbench

System = gcc -O2
System + I = gcc -O3

Amer:
speedup = 1.18 ± 0.0002

Conclusion: O3 is good

Todd:
speedup = 0.84 ± 0.0002

Conclusion: O3 is bad

Why does this happen?

Differences in our experimental setup

Differences in our experimental setup

Amer:

HOME=/home/amerdiwan/

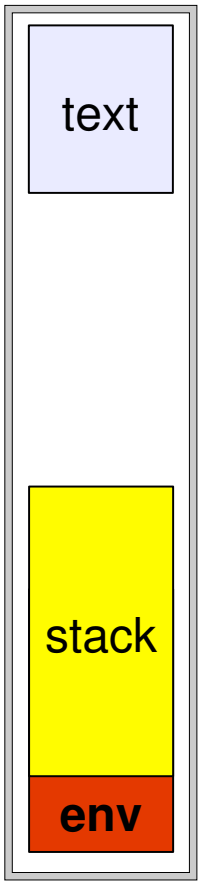
Todd:

HOME=/home/toddmytkowicz

Differences in our experimental setup

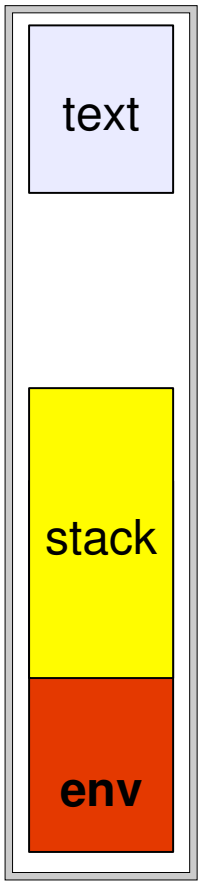
Amer:

HOME=/home/amerdiwan/



Todd:

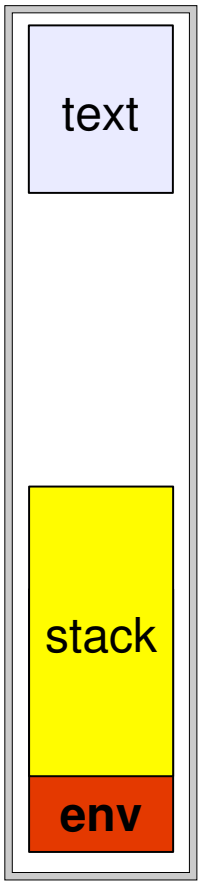
HOME=/home/toddmytkowicz



Differences in our experimental setup

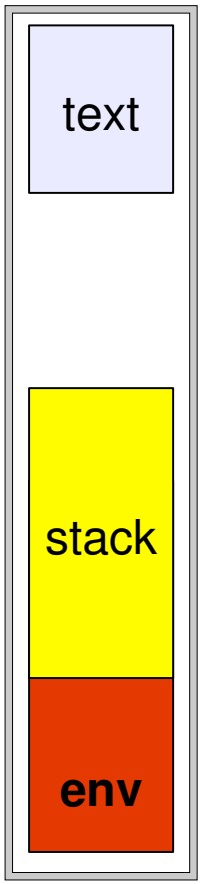
Amer:

HOME=/home/amerdiwan/



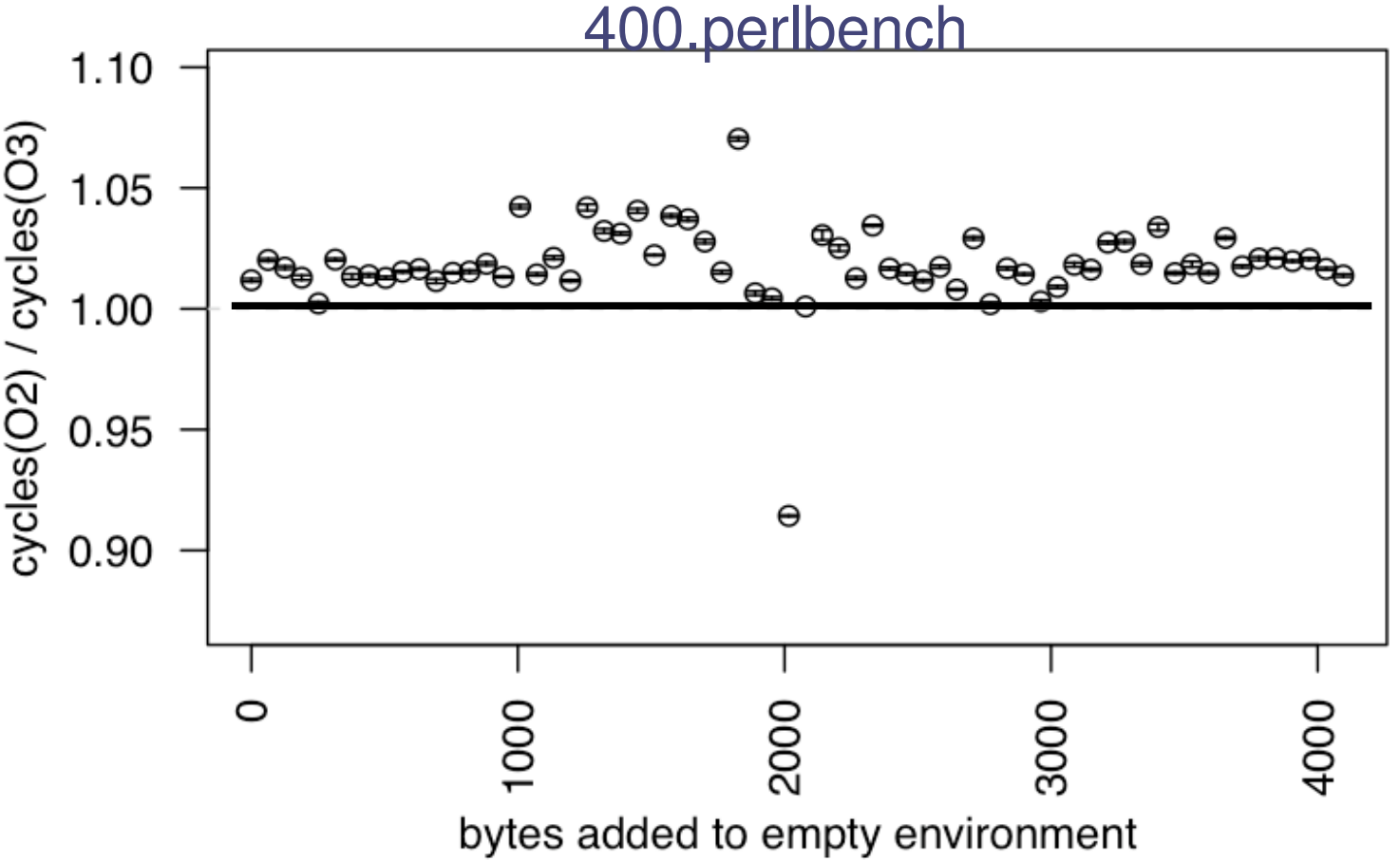
Todd:

HOME=/home/toddmytkowicz

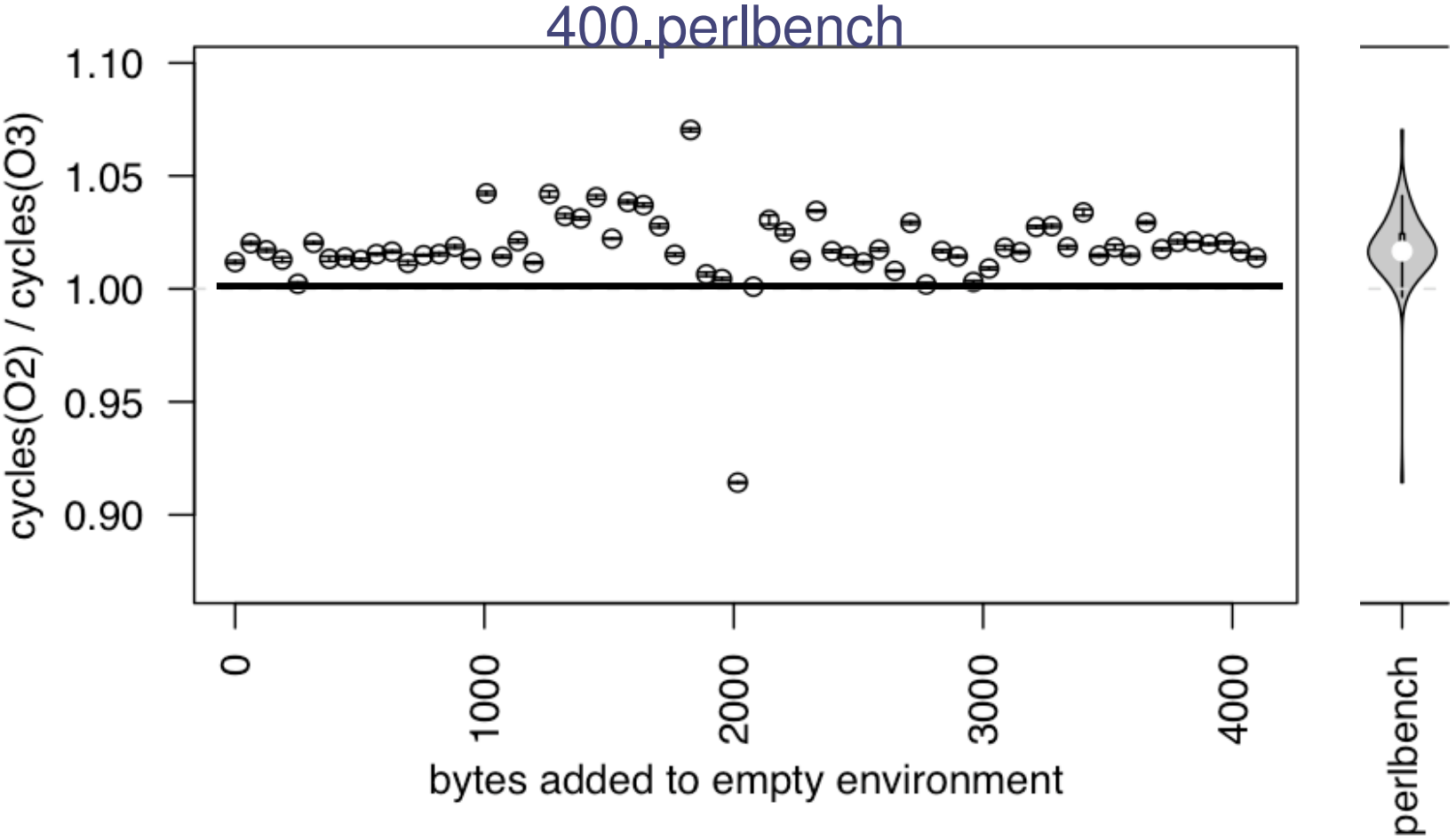


Could this be the source of bias?

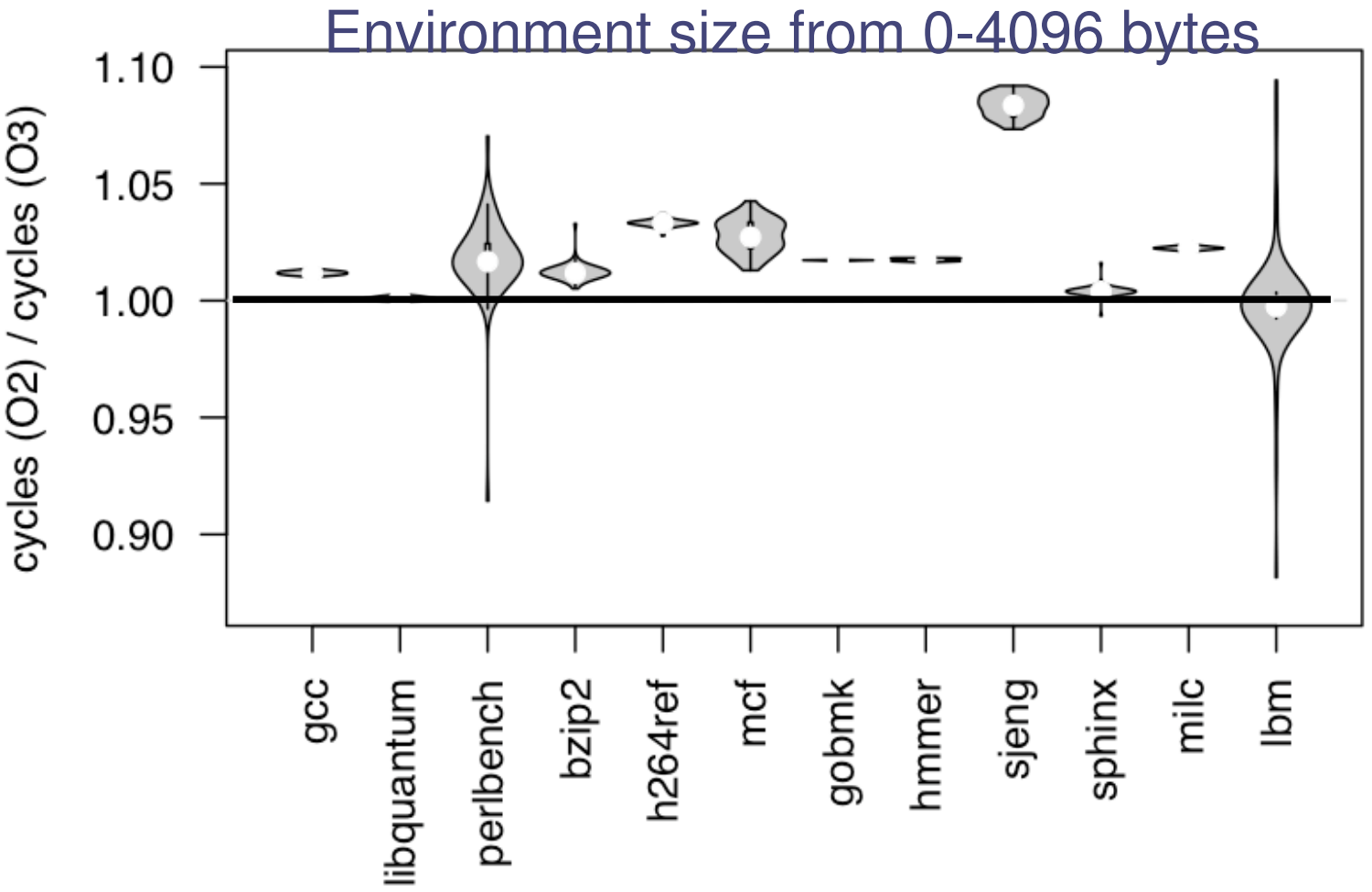
Bias from size of UNIX environment



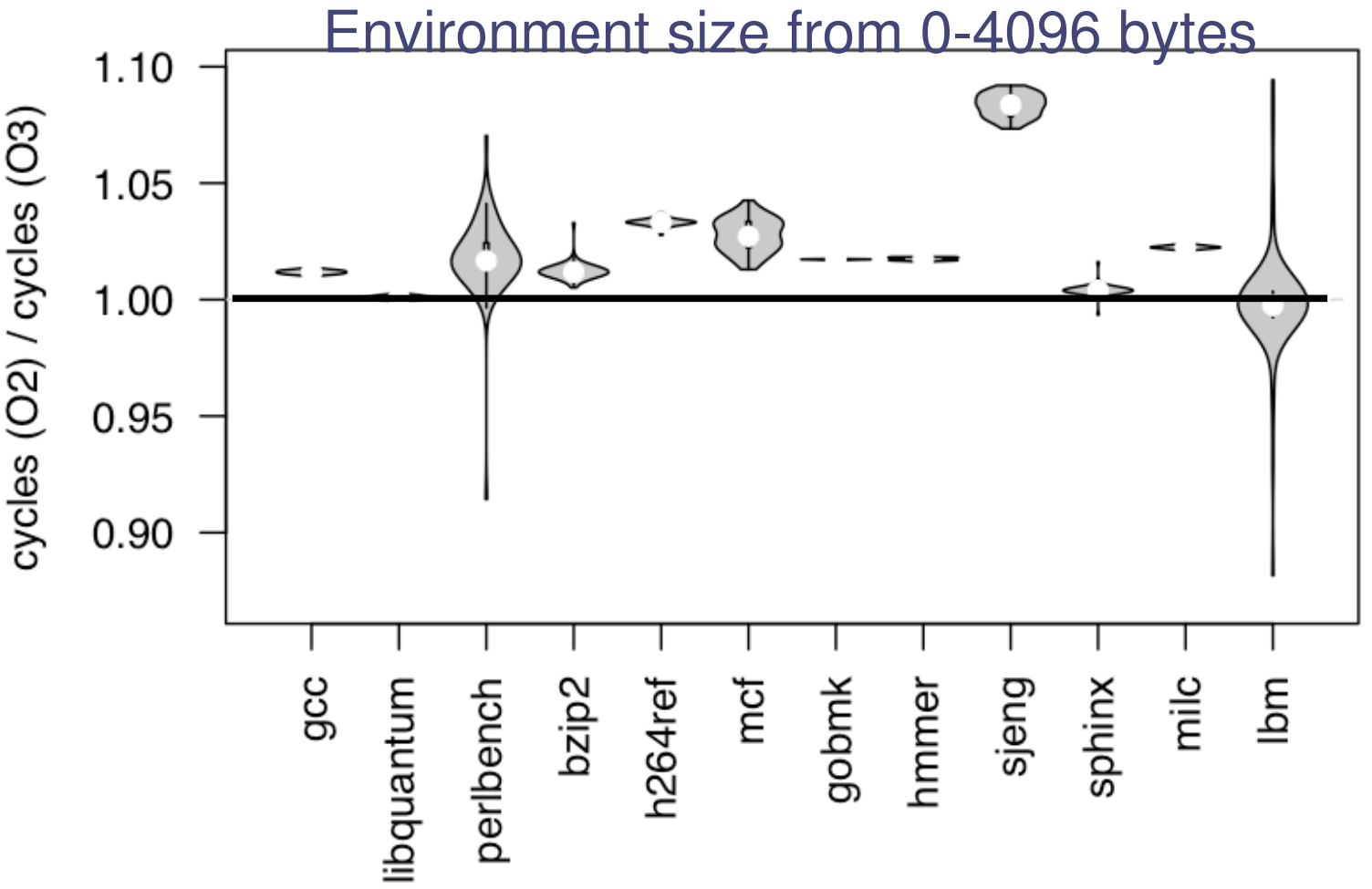
Bias from size of UNIX environment



Bias from size of UNIX environment



Bias from size of UNIX environment



The setting of irrelevant environment variables can lead to biased conclusions

By using an empty UNIX environment,
Amer and I now agree.

But for perlbench, we still differ...

Other differences in our experimental setup

Amer:

$> ld A.o B.o

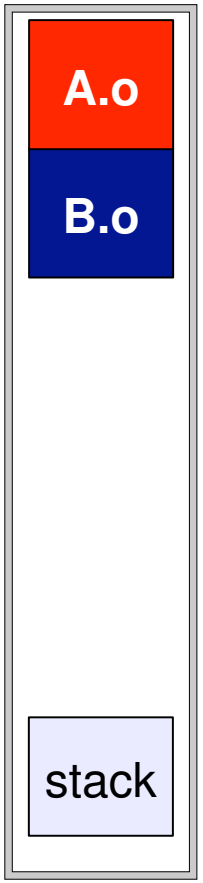
Todd:

$> ld B.o A.o

Other differences in our experimental setup

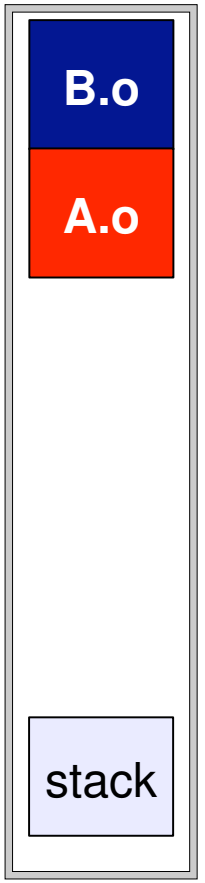
Amer:

$> ld A.o B.o



Todd:

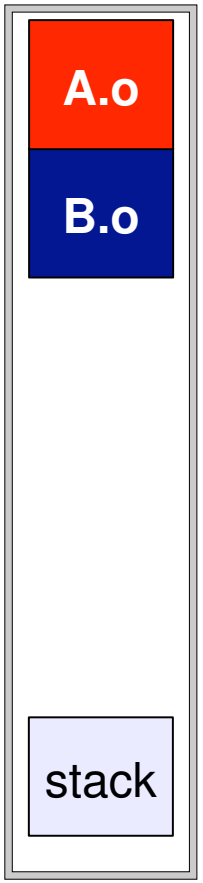
$> ld B.o A.o



Other differences in our experimental setup

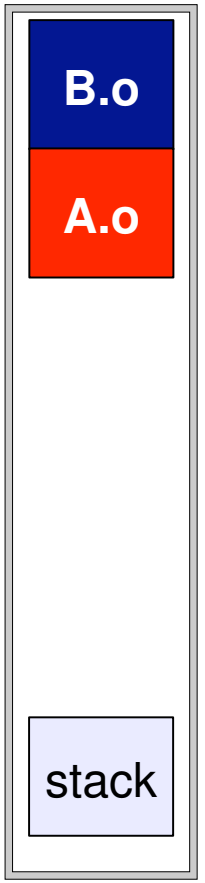
Amer:

$> ld A.o B.o



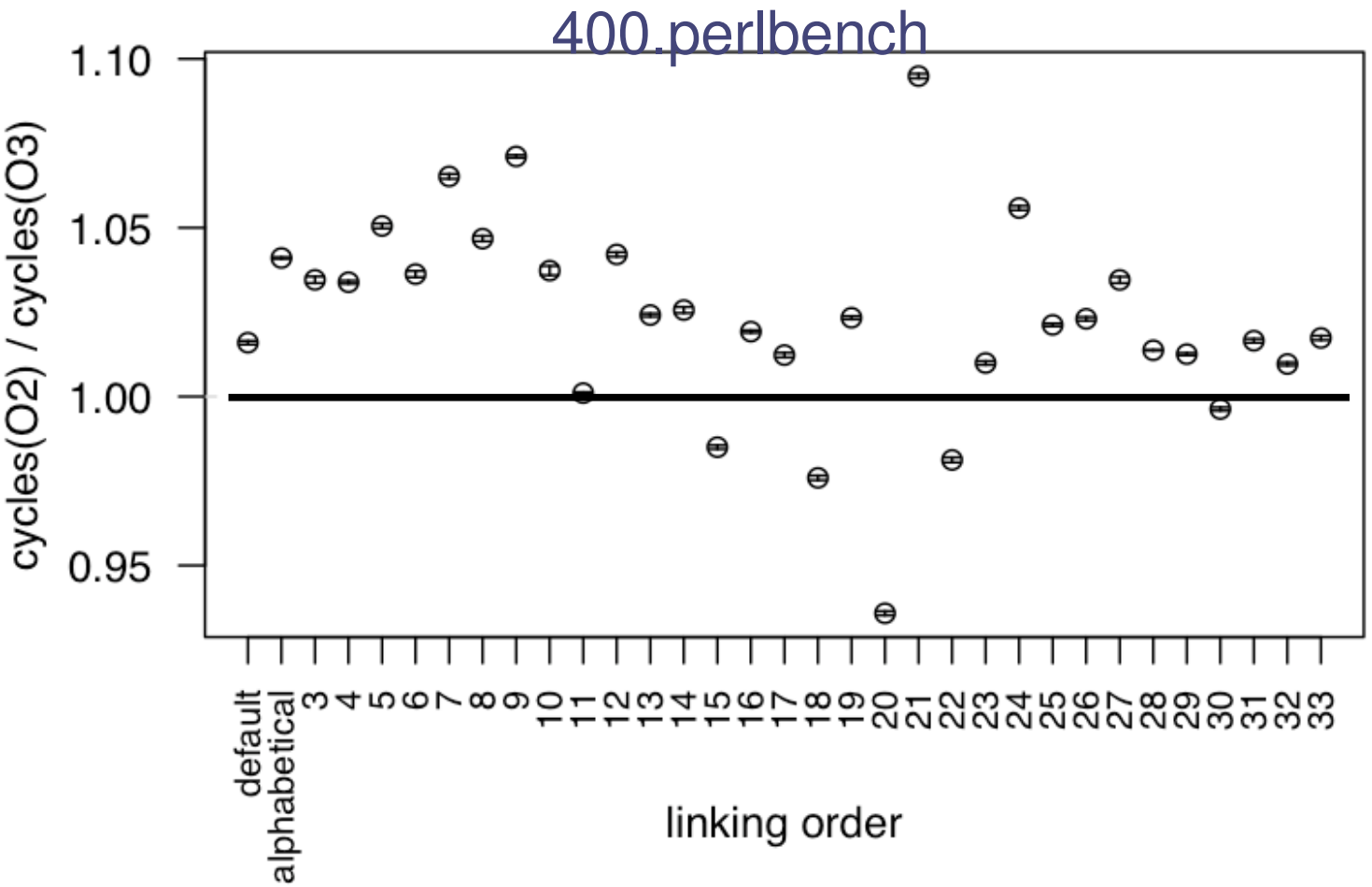
Todd:

$> ld B.o A.o

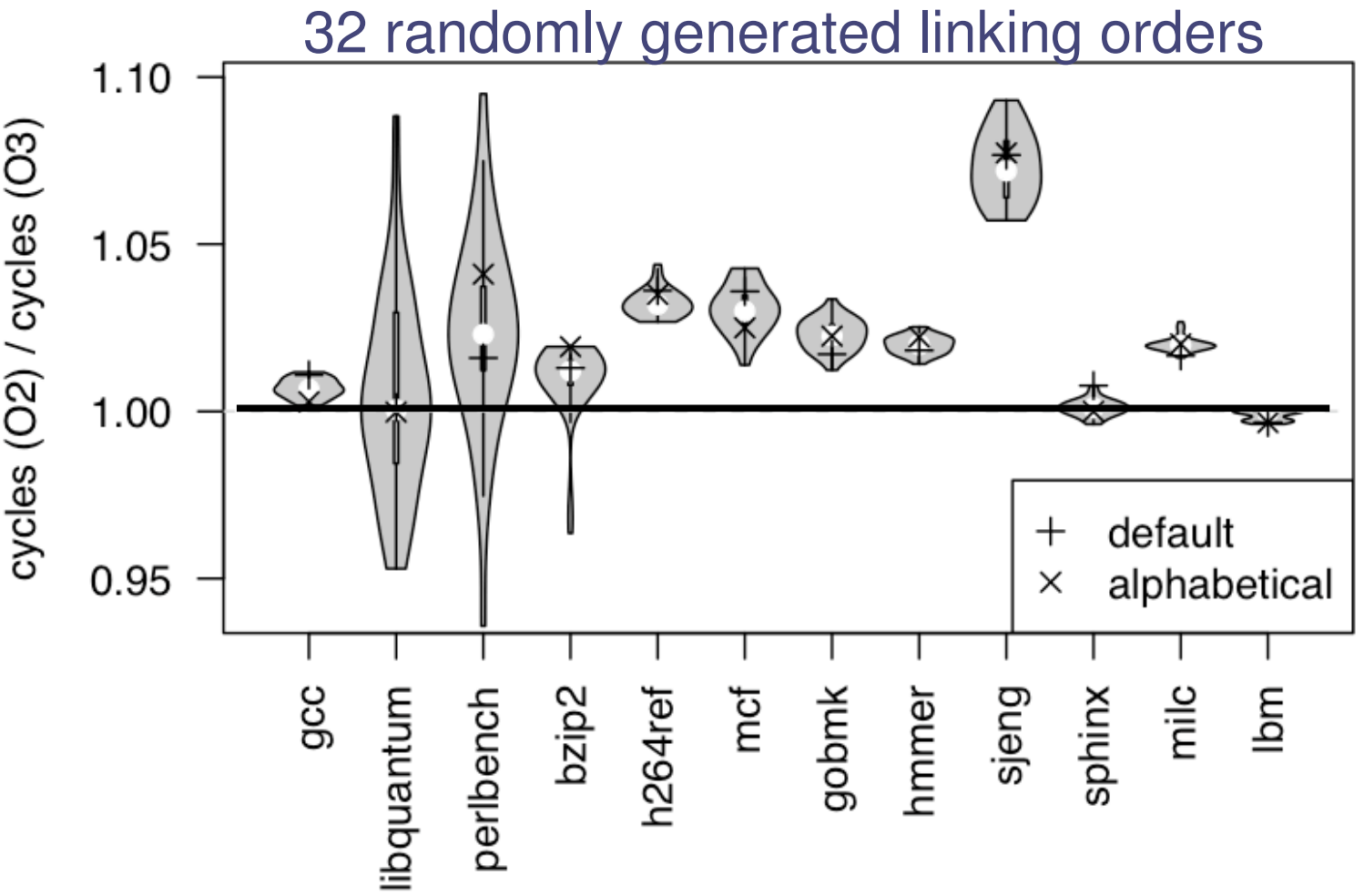


Could this be the source of bias?

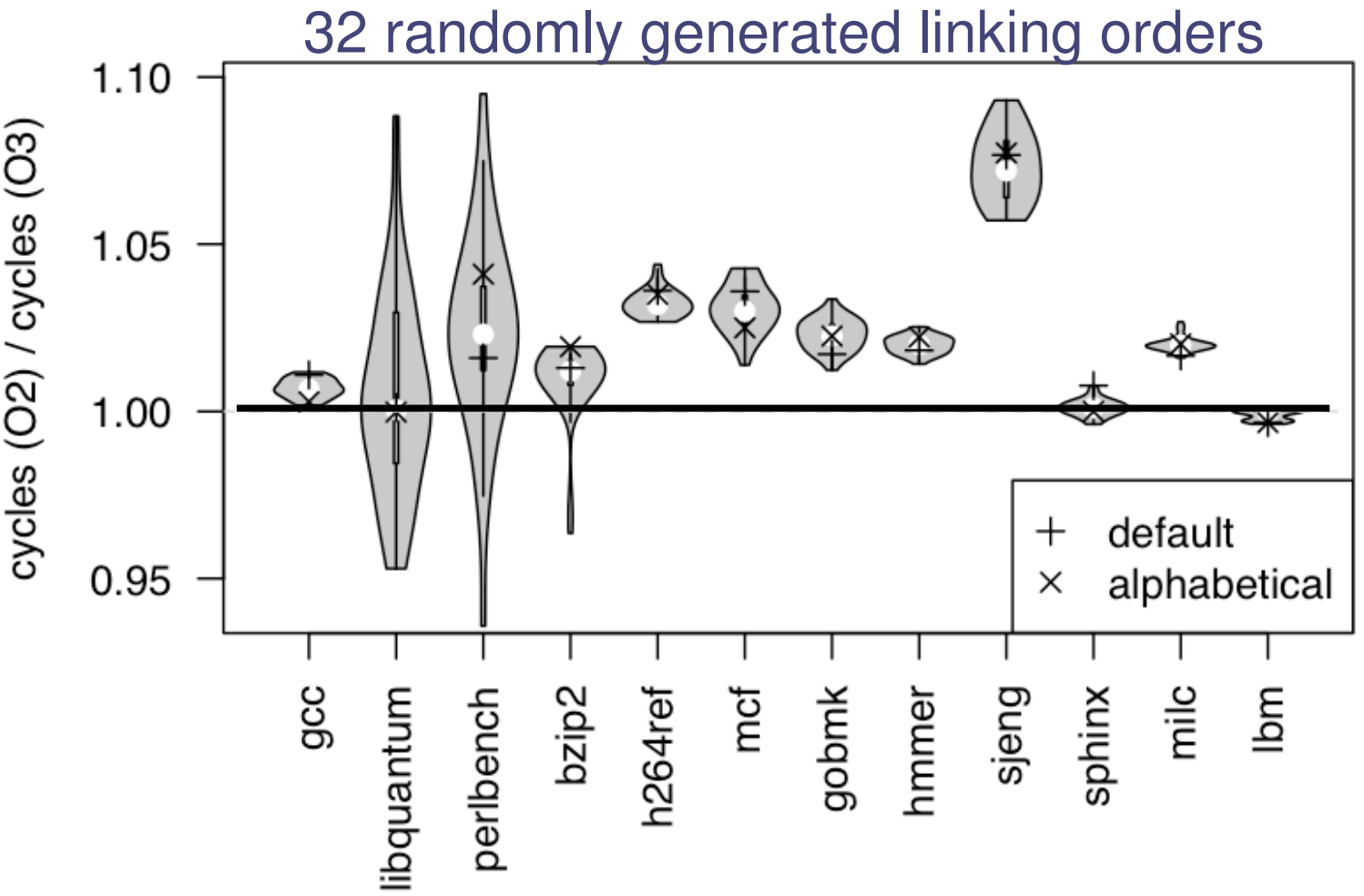
Bias from linking order



Bias from linking order



Bias from linking order

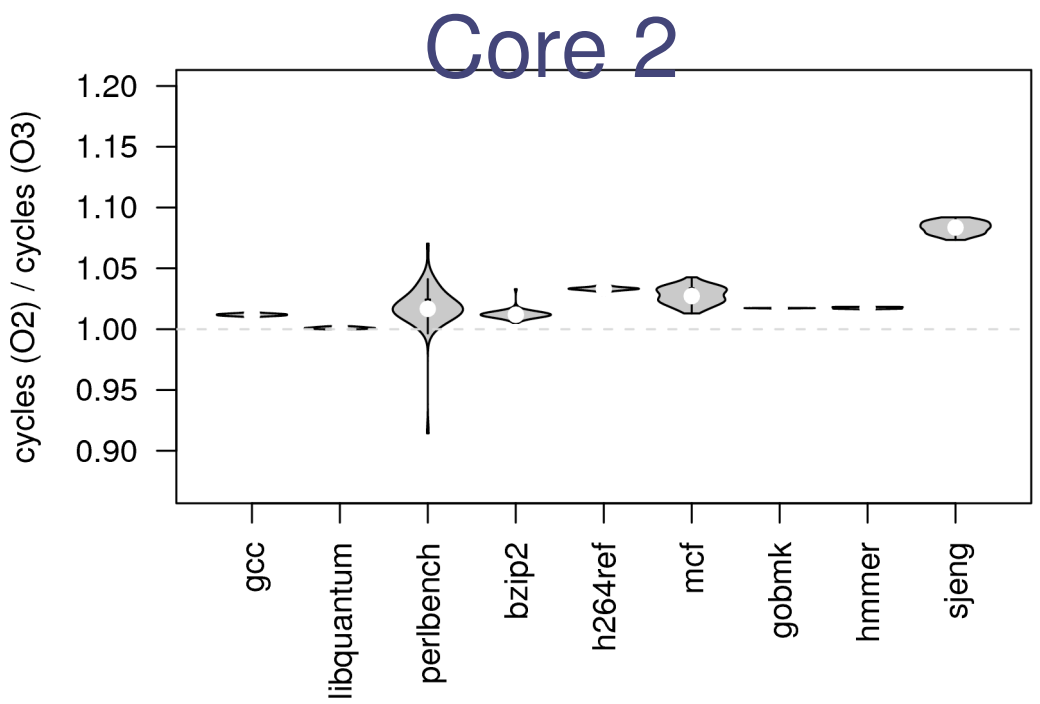


Order of .o files can lead to contradictory conclusions

Are we just showing you corner cases?

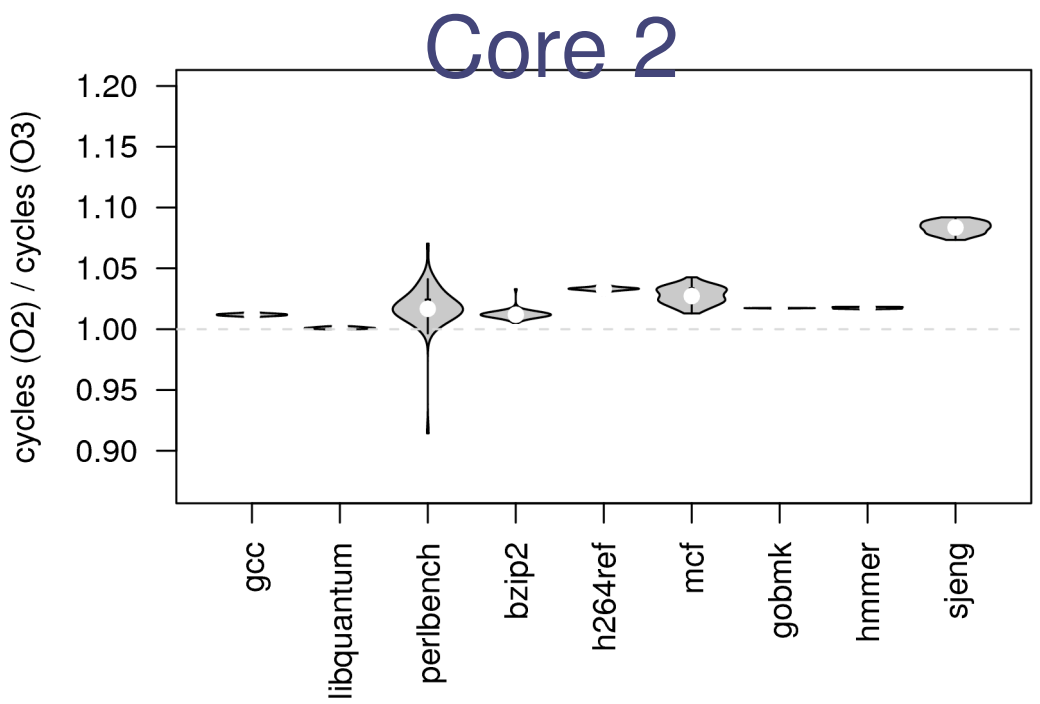
Are we just showing you corner cases?

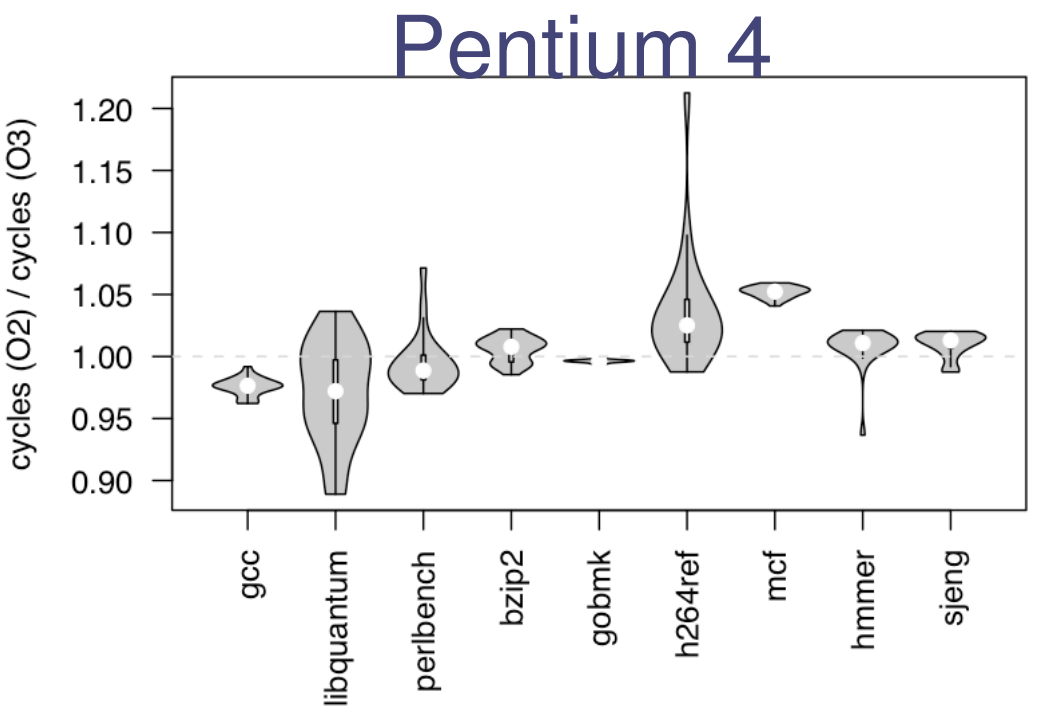
No: Bias occurs on multiple microprocessors



Are we just showing you corner cases?

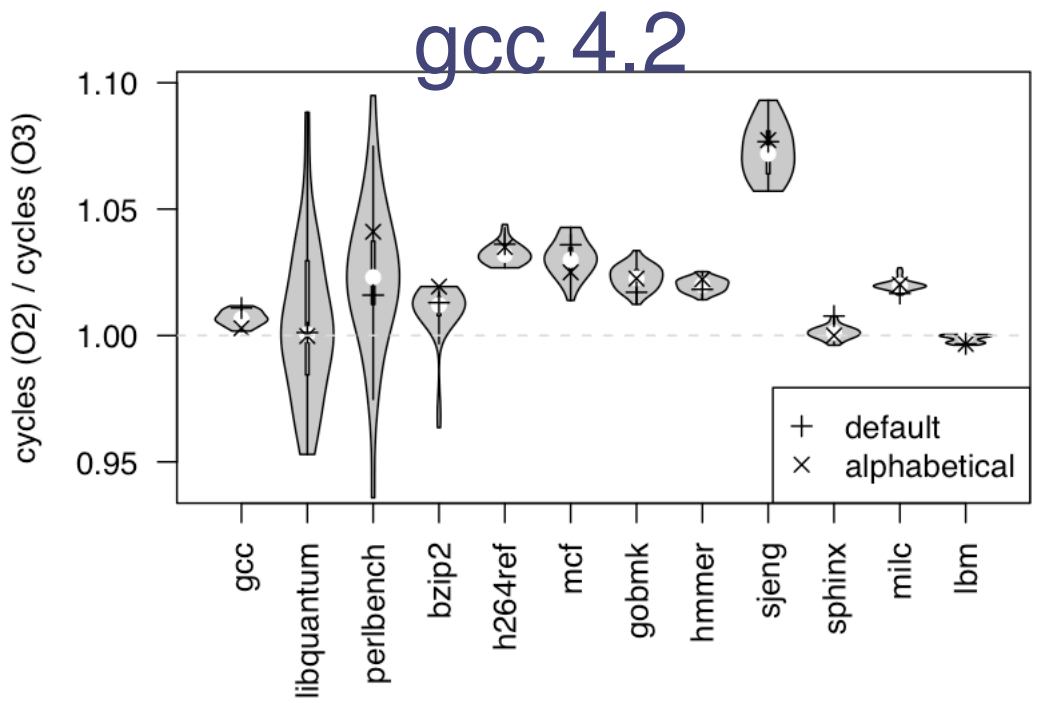
No: Bias occurs on multiple microprocessors





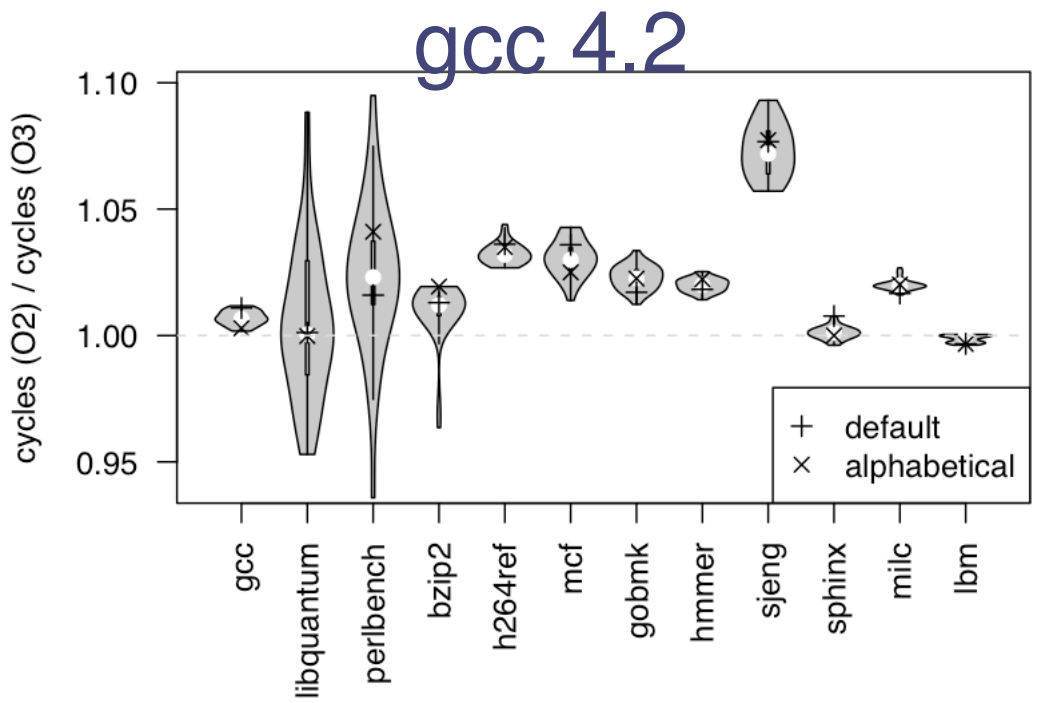
Are we just showing you corner cases?

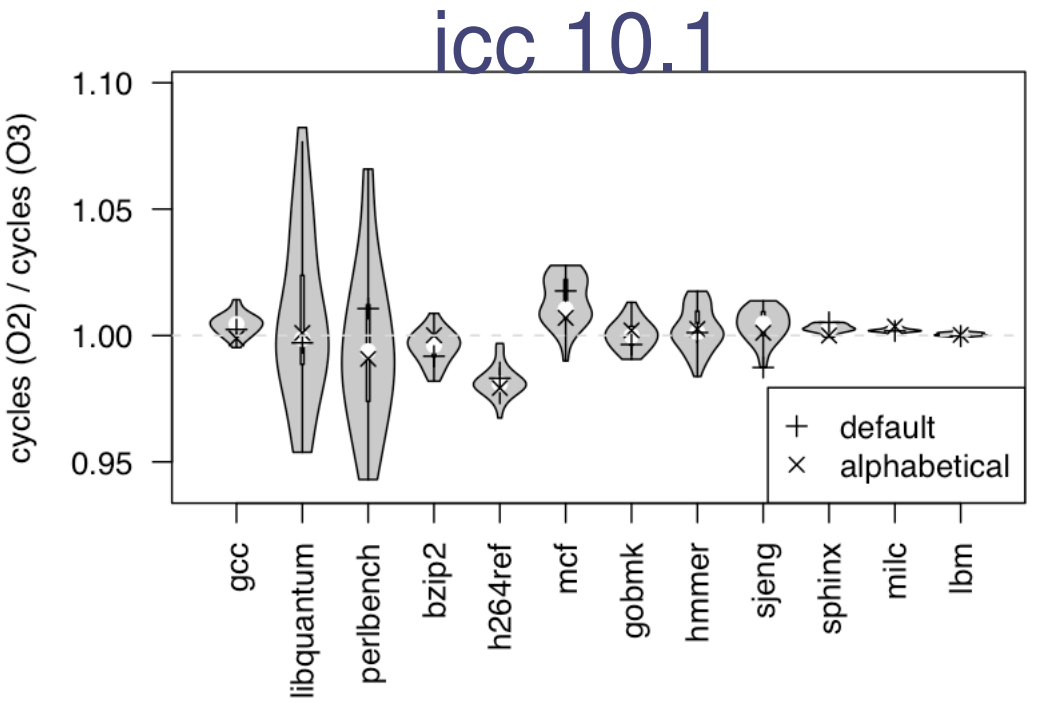
No: Bias occurs on multiple compilers



Are we just showing you corner cases?

No: Bias occurs on multiple compilers



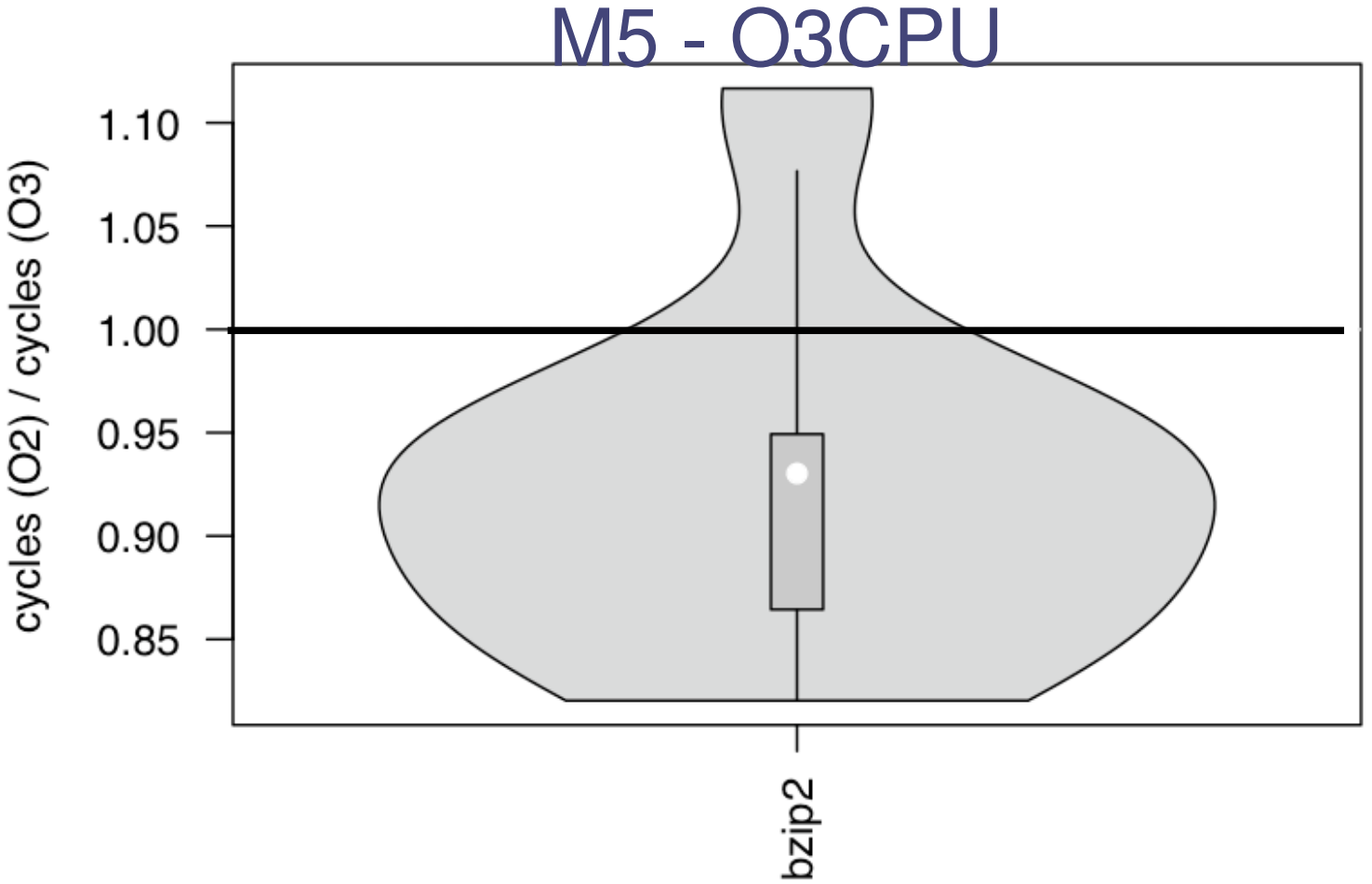


Are we just showing you corner cases?

No: Bias occurs in simulation

Are we just showing you corner cases?

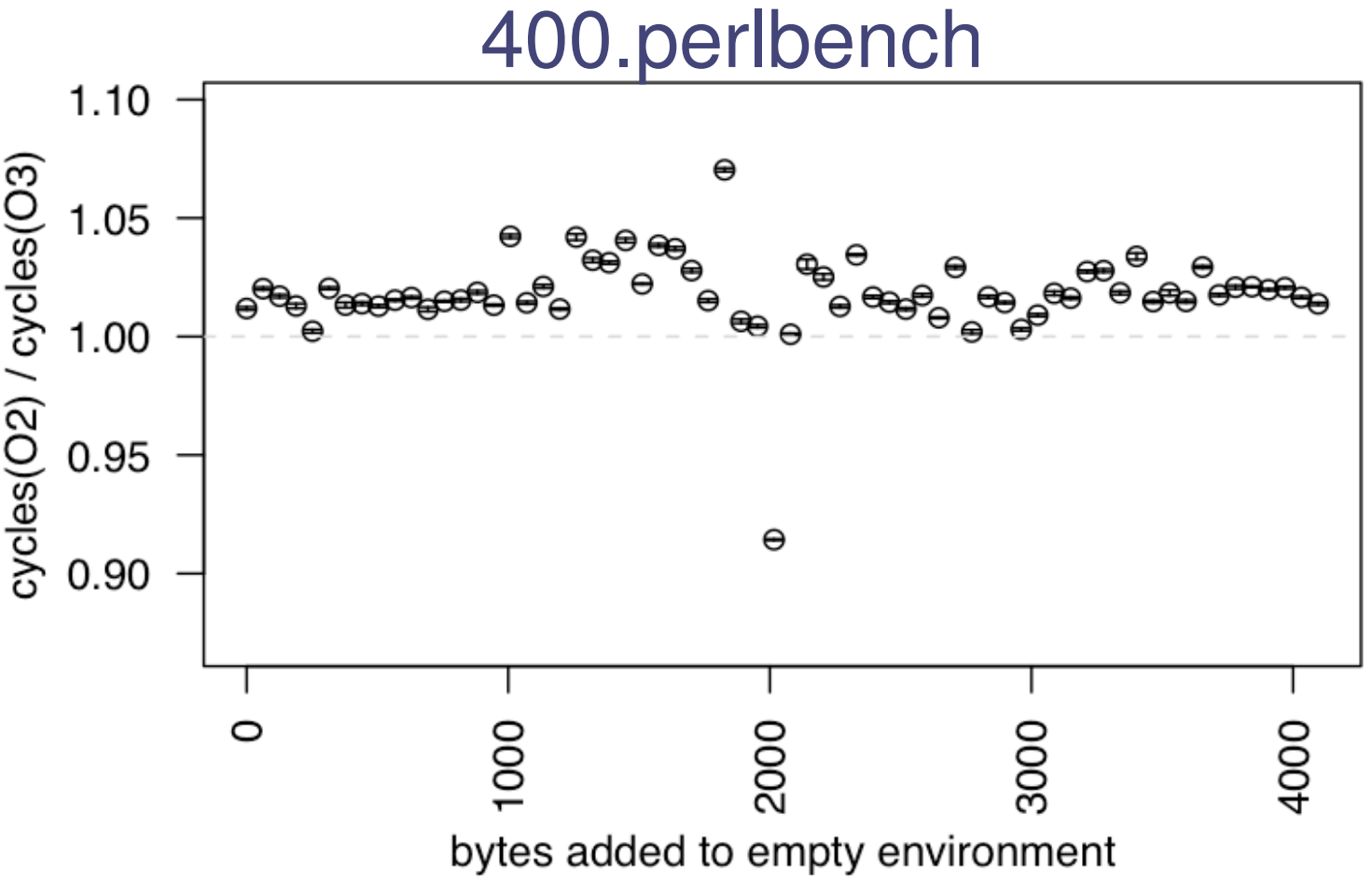
No: Bias occurs in simulation



Can we **easily** avoid bias?

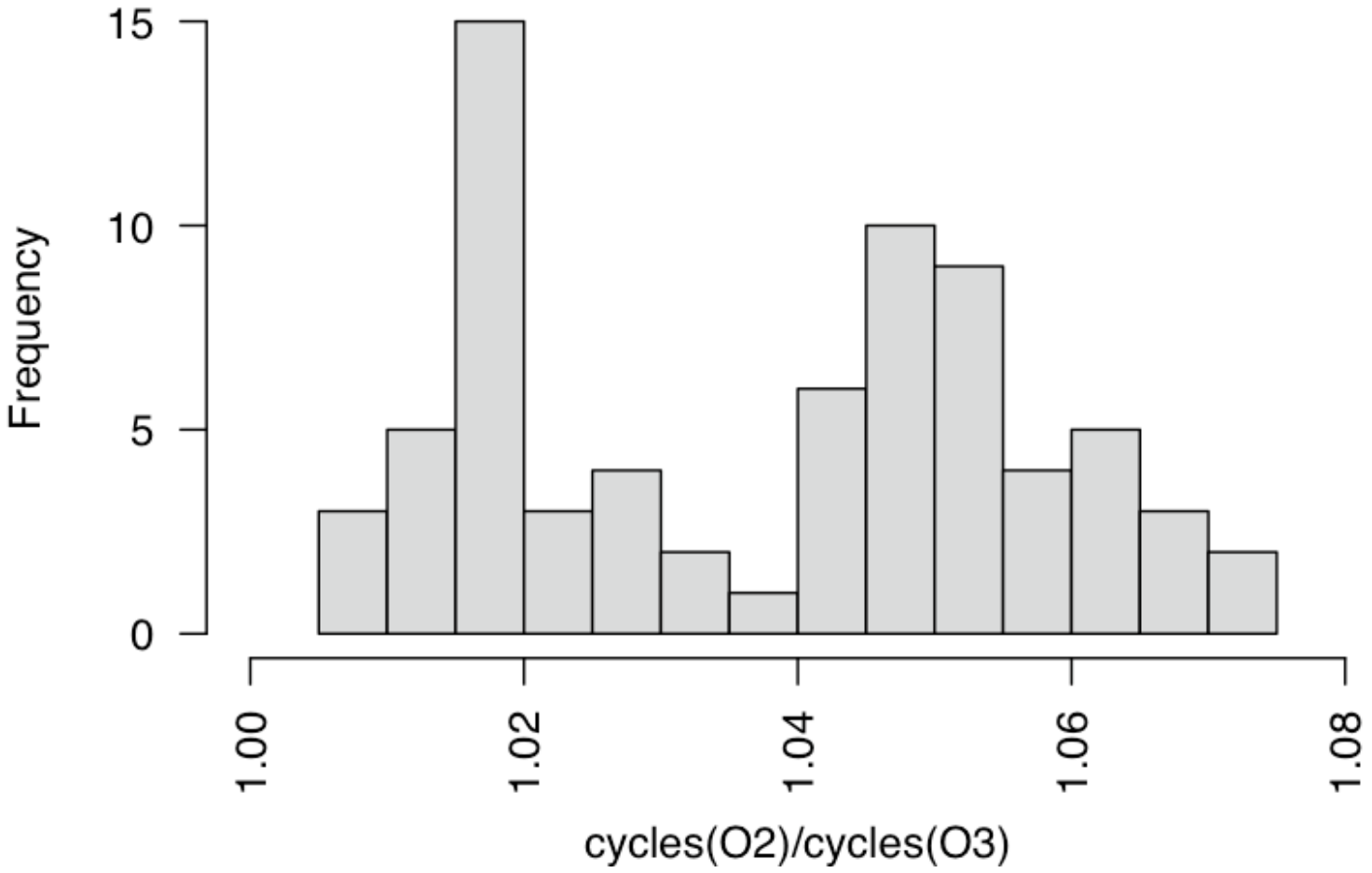
Can we easily avoid bias?

No: Bias is not predictible



Can we easily avoid bias?

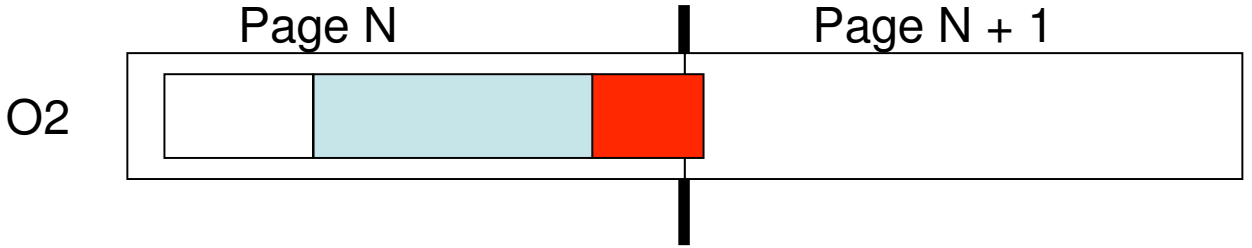
No: Averaging across benchmarks does not cancel it out



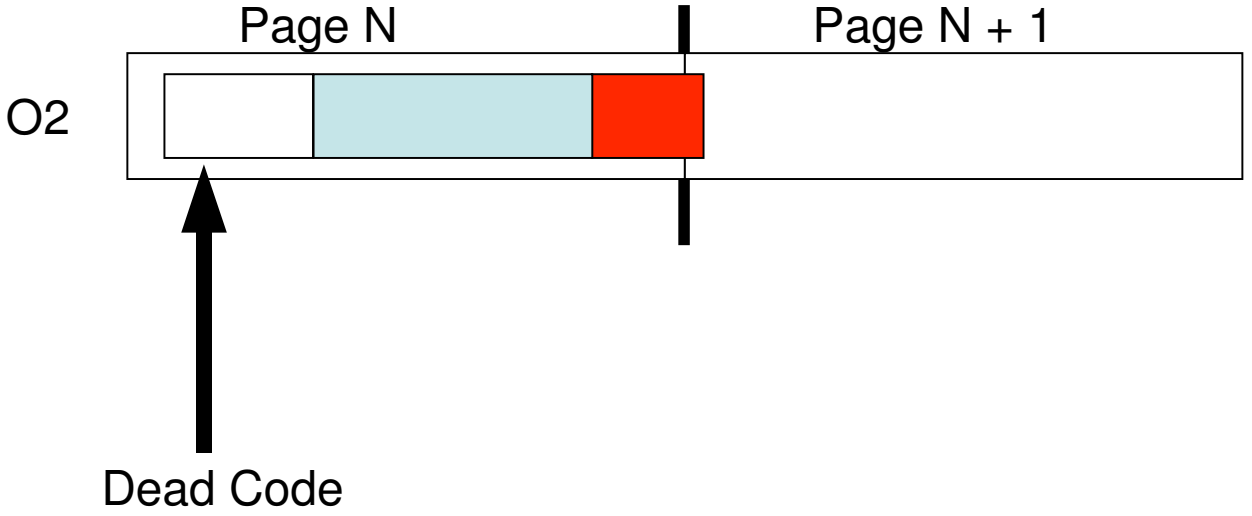
Where does bias come from?

Where does bias come from?

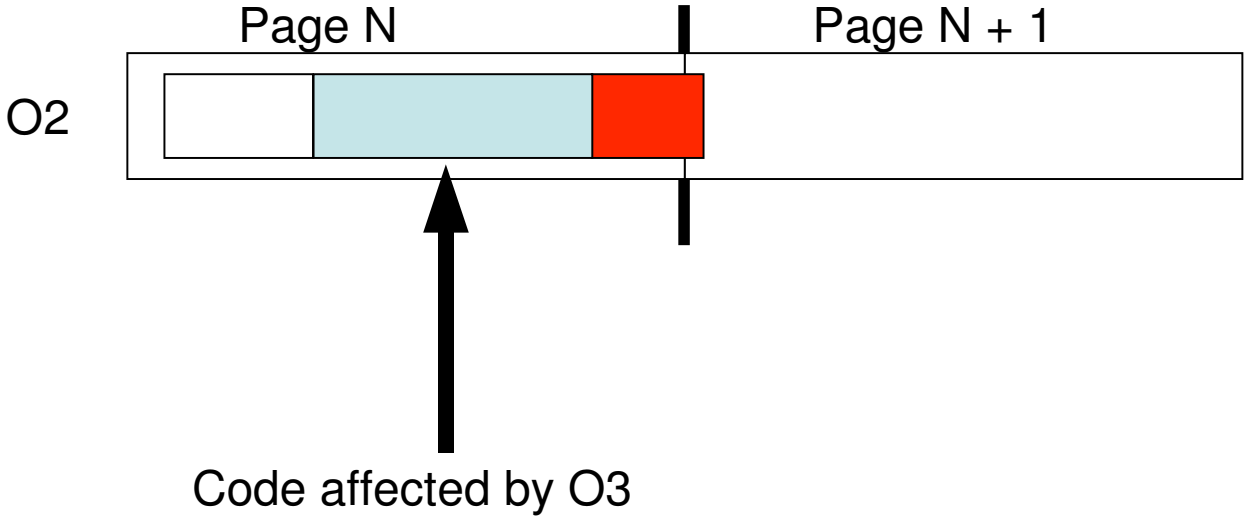
Interactions with hardware buffers



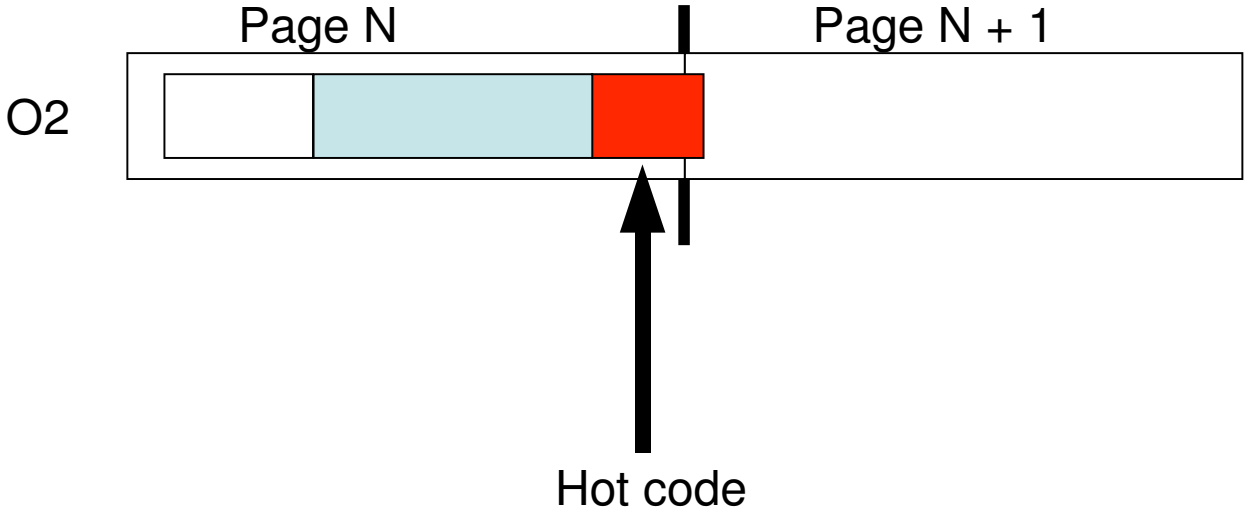
Interactions with hardware buffers



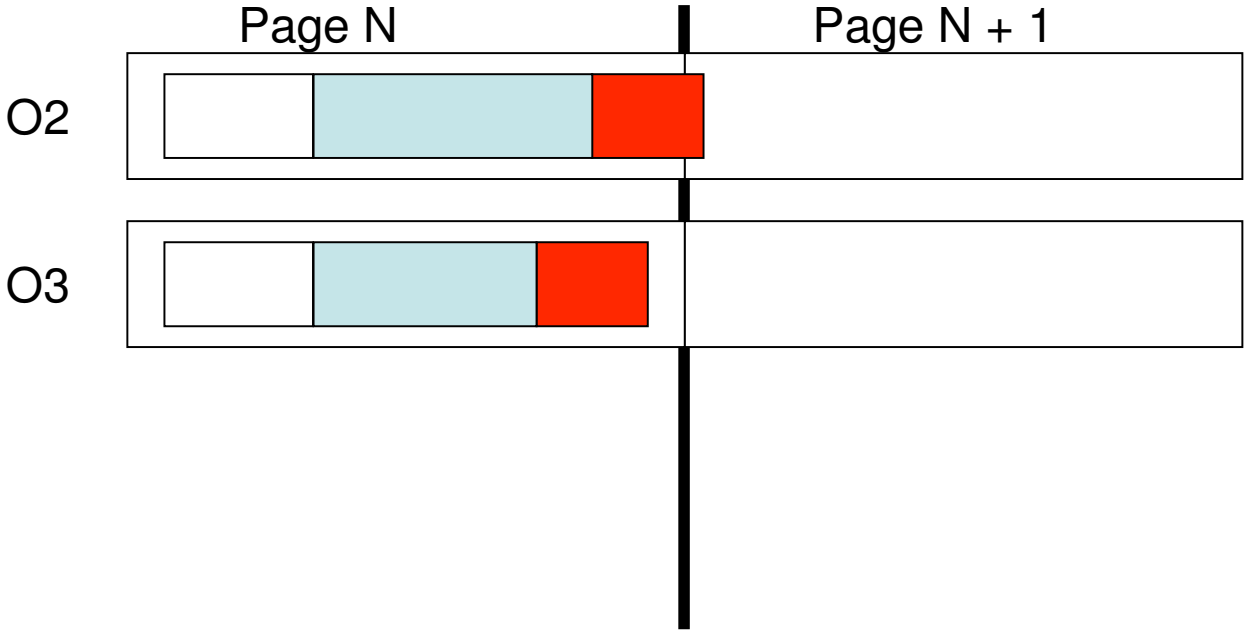
Interactions with hardware buffers



Interactions with hardware buffers



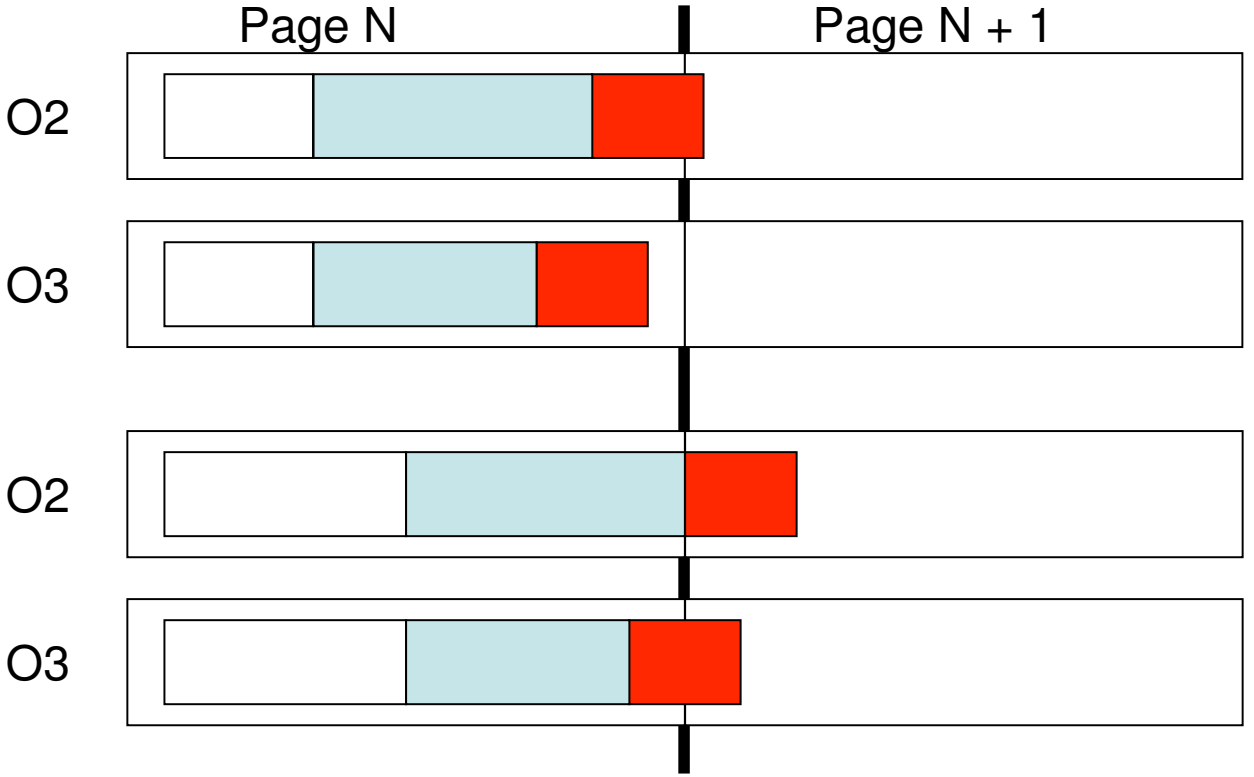
Interactions with hardware buffers



O3 better than O2

Where does bias come from?

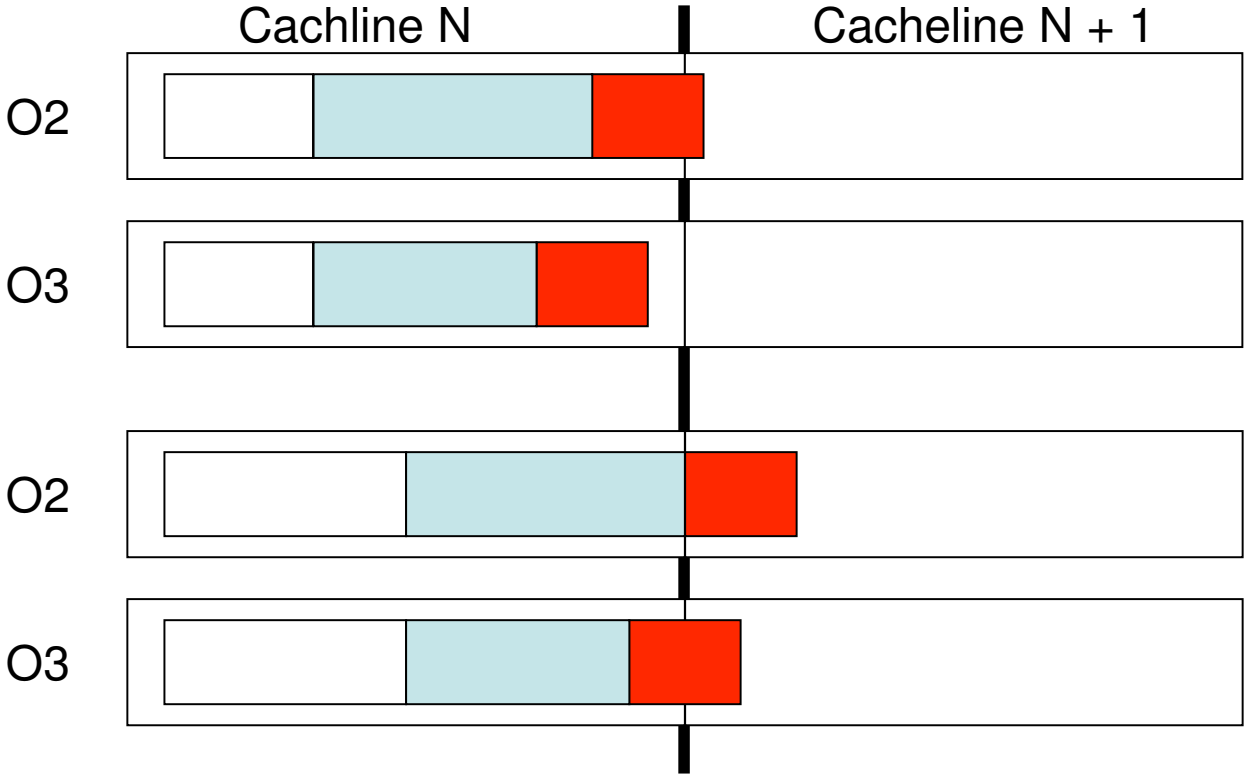
Interactions with hardware buffers



O3 better than O2

O2 better than O3

Interactions with hardware buffers



O3 better than O2

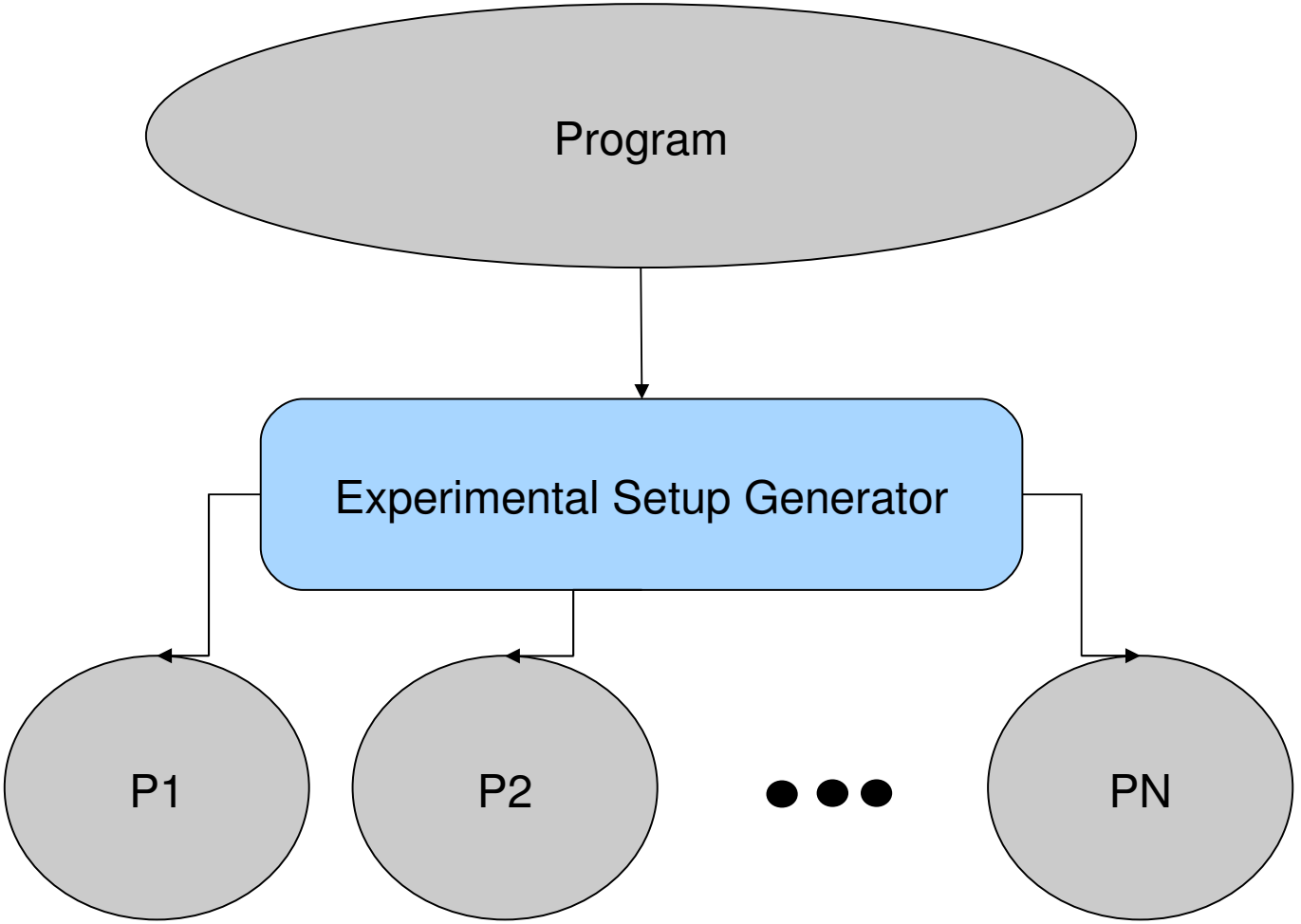
O2 better than O3

Hardware buffers abound in modern systems

What can we do about bias?

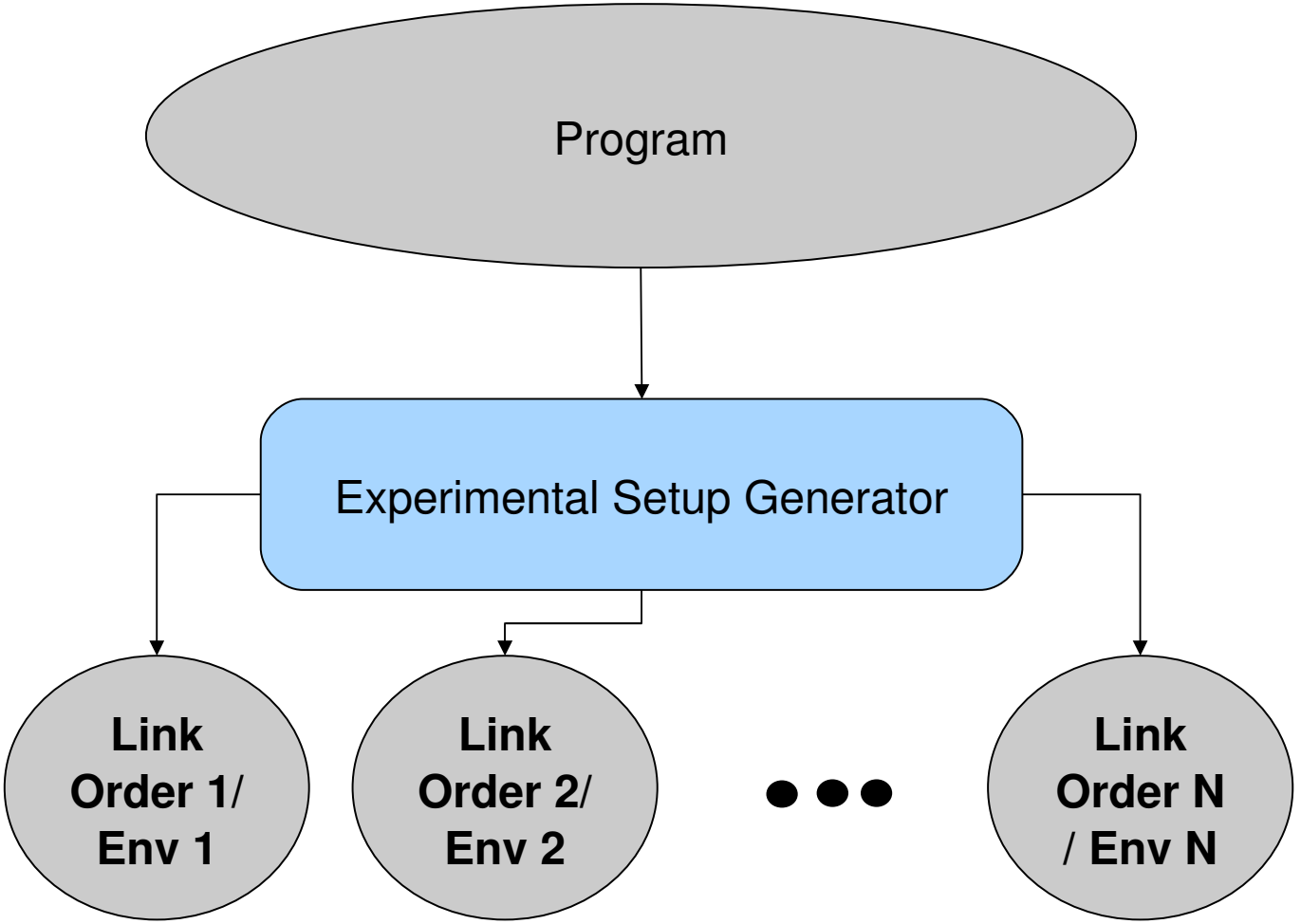
What can we do about bias?

Randomized Trials



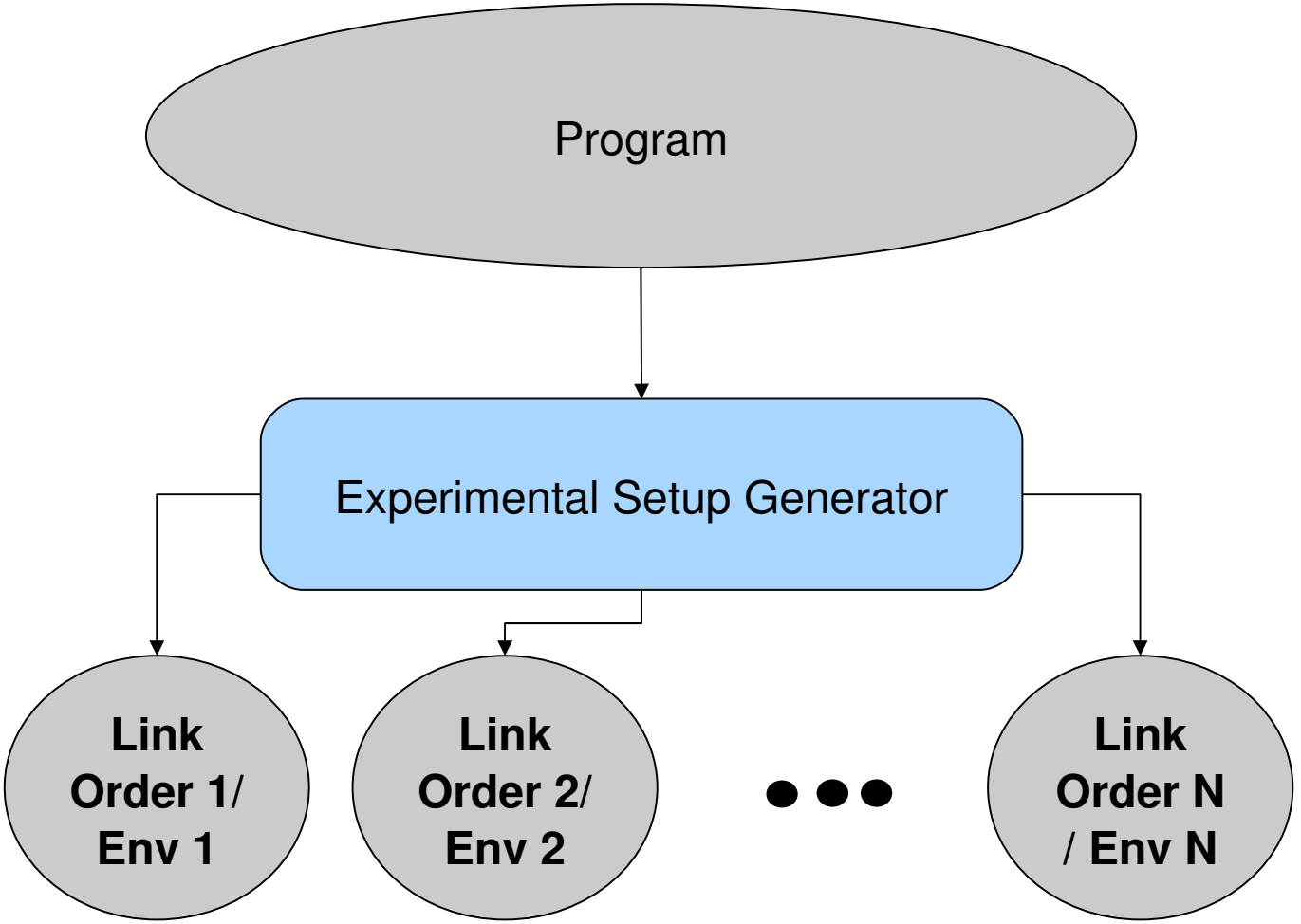
What can we do about bias?

Randomized Trials



What can we do about bias?

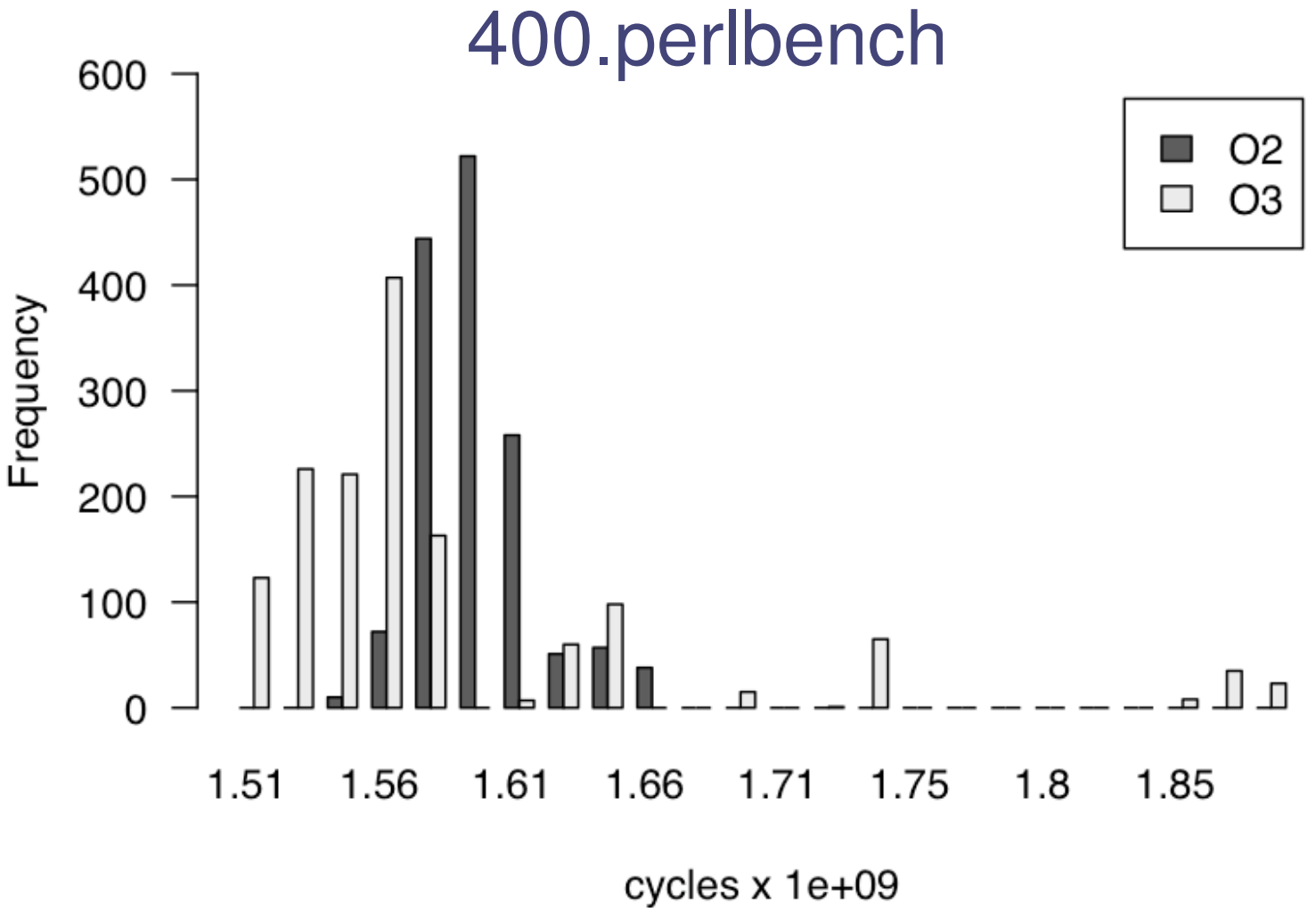
Randomized Trials



Effectiveness depends upon representativeness of setups

What can we do about bias?

Randomized Trials



Causality Analysis

(1) Analyze data to arrive at a hypothesis

Causality Analysis

(1) Analyze data to arrive at a hypothesis

(2) Perform intervention to test hypothesis

Causality Analysis

(1) Analyze data to arrive at a hypothesis

(2) Perform intervention to test hypothesis

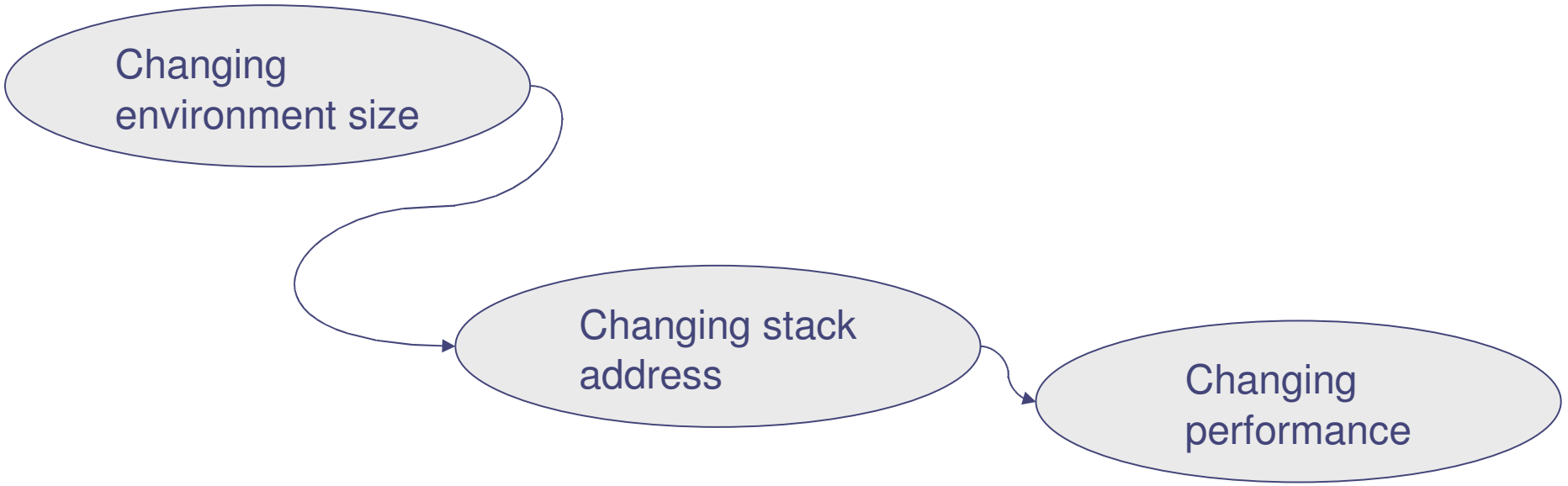
(3) Validate effects of hypothesis

Causality Analysis

(1) Analyze data to arrive at a hypothesis

(2) Perform intervention to test hypothesis

(3) Validate effects of hypothesis

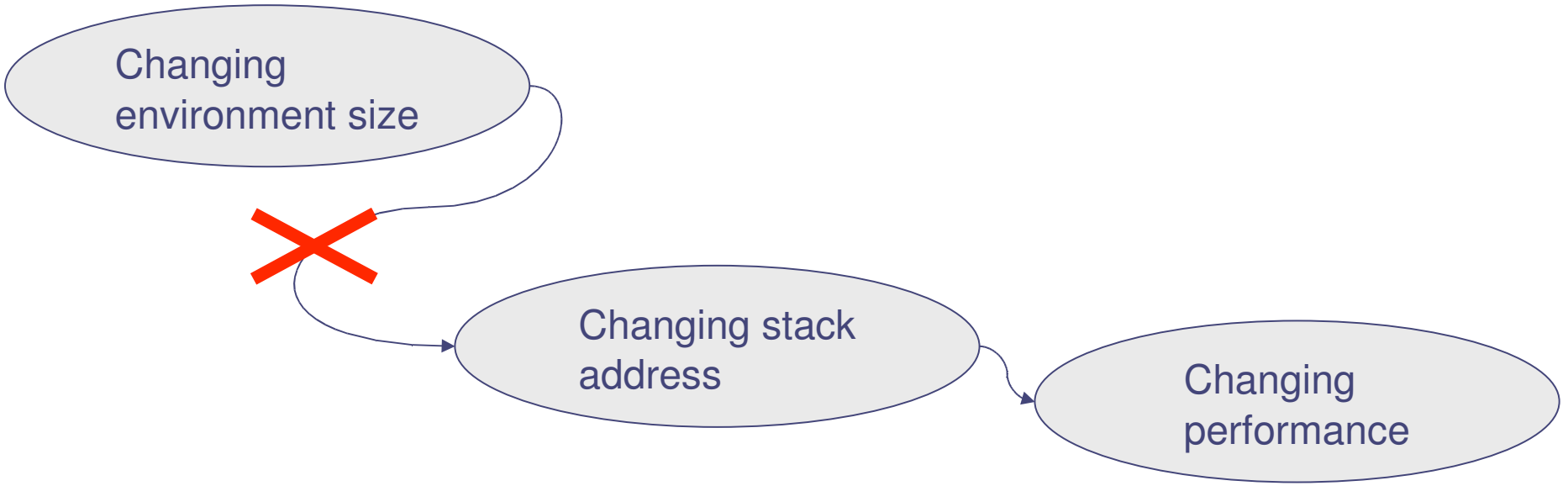


Causality Analysis

(1) Analyze data to arrive at a hypothesis

(2) Perform intervention to test hypothesis

(3) Validate effects of hypothesis

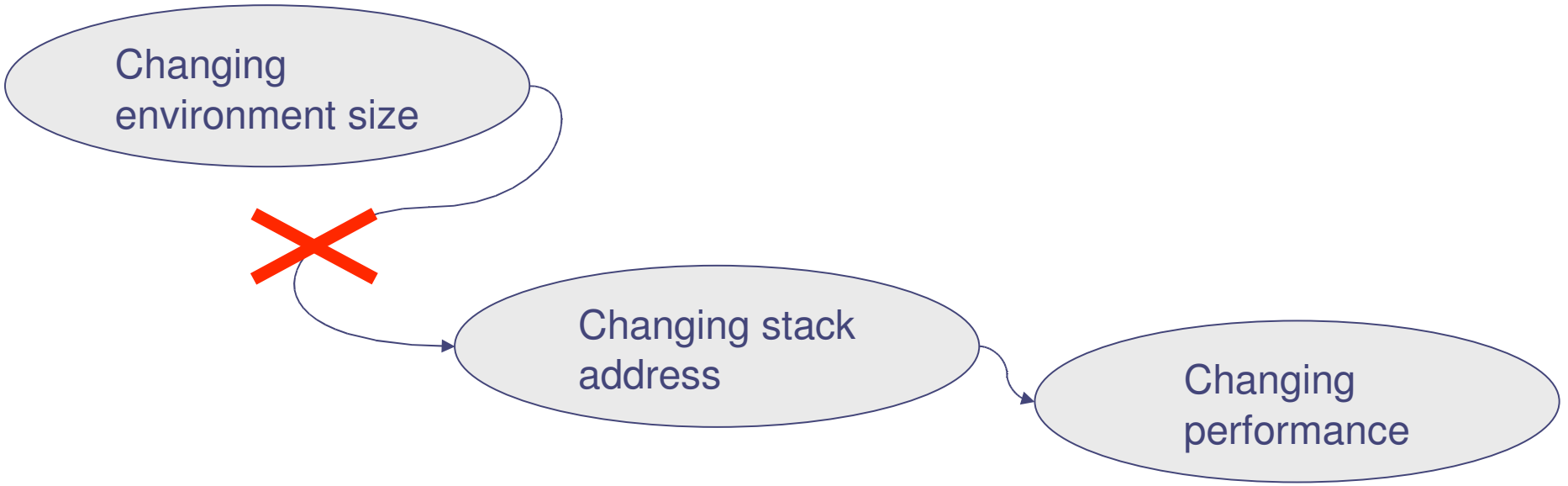


Causality Analysis

(1) Analyze data to arrive at a hypothesis

(2) Perform intervention to test hypothesis

(3) Validate effects of hypothesis



Popular in sciences but difficult and manual

Related Work

- “Correctness” via microkernels
  - [Korn et al: IPCCC '01], [Maxell et al: LASCI '02] and [Moore: ICCS '02]
- Other sources of bias
  - Heap Size for GC [Blackburn et al: OOPSLA '06]
  - Variability in multi-threaded simulation [Alameldeen and Wood: HPCA '03]
  - Input Shaking [Tsafrir et al: MASCOTS '07]
- Statistical Rigor in performance evaluations
  - [Georges et al: OOPSLA '07]

Summary

Would you believe a U.S. Census conducted in **one** small town?

Summary

Would you believe a U.S. Census conducted in **one** small town?

Would you believe a systems experiment conducted in **one** experimental setup?

Summary

Would you believe a U.S. Census conducted in **one** small town?

Would you believe a systems experiment conducted in **one** experimental setup?

We show bias is pervasive, unpredictible and significant

Microkernel example

```
static int i, j, k, inc;
int main() {
  int g;
  i = j = k = 0;
  inc = 1;
  for (g = 0; g < 65536; g++) {
    i += inc;
    j += inc;
    k += inc;
  }
  return (0);
}
```

Microkernel example

stack allocated variable

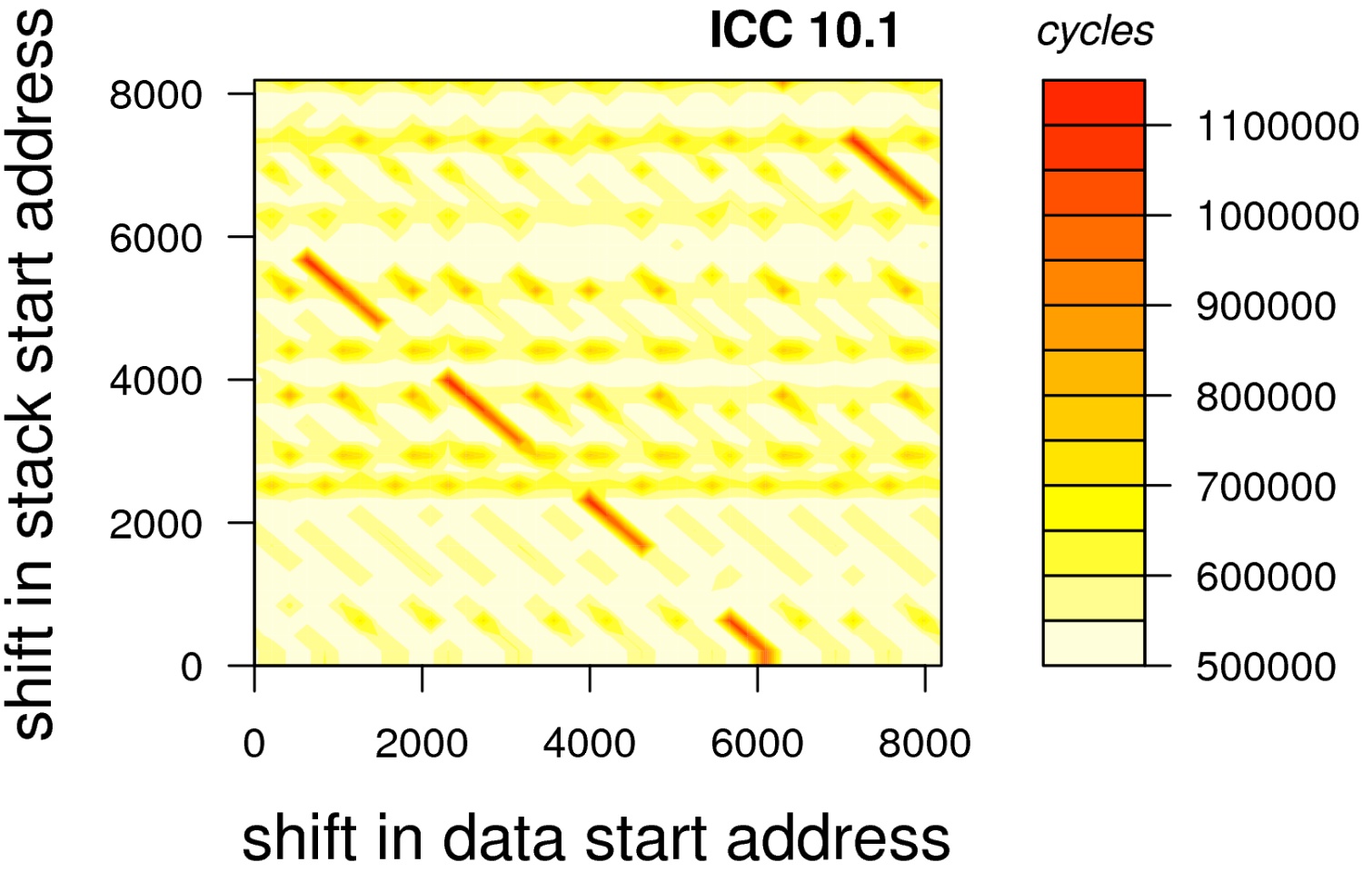
```
static int i, j, k, inc;
int main() {
  int g;
  i = j = k = 0;
  inc = 1;
  for (g = 0; g < 65536; g++) {
    i += inc;
    j += inc;
    k += inc;
  }
  return (0);
}
```

Microkernel example

global variable

```
static int i, j, k, inc;
int main() {
  int g;
  i = j = k = 0;
  inc = 1;
  for (g = 0; g < 65536; g++) {
    i += inc;
    j += inc;
    k += inc;
  }
  return (0);
}
```

Microkernel example



Bias comes from sensitivity of program behavior to an experimental setup