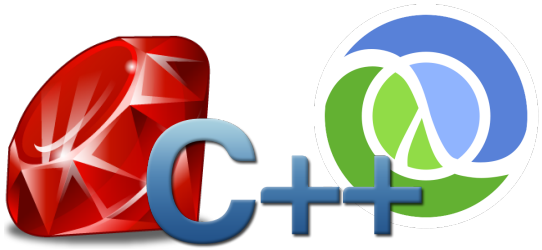
Benchmarking: You're Doing It Wrong

Aysylu Greenberg
@aysylu22

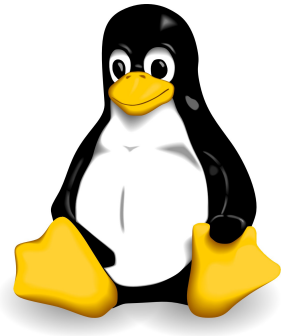
Google

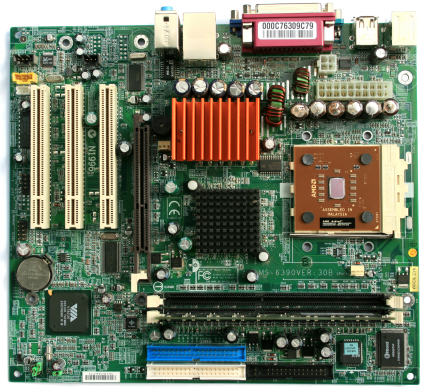
In Memes…

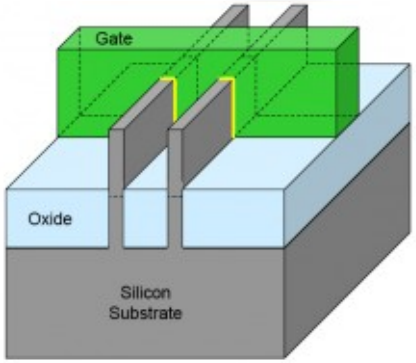
To Write Good Benchmarks…



Need to be Full Stack







What's a Benchmark

How fast?

Your process vs Goal

Your process vs Best Practices

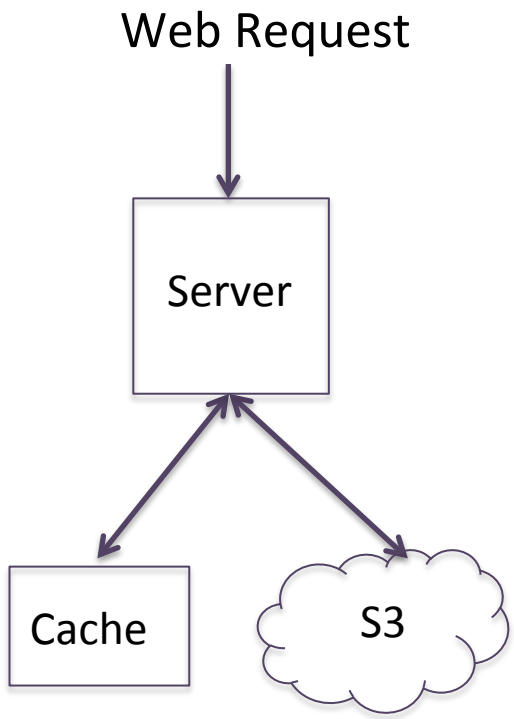
Today

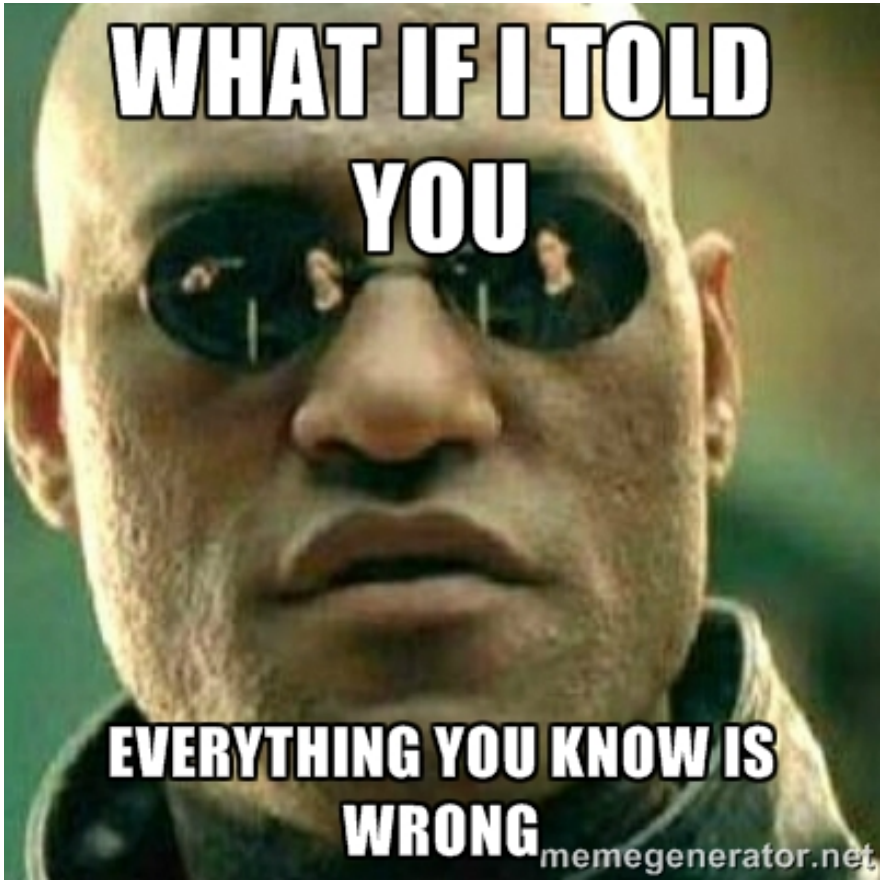
- How Not to Write Benchmarks
- Benchmark Setup & Results:
  - Wrong about the machine
  - Wrong about stats
  - Wrong about what matters
- Becoming Less Wrong

HOW NOT TO WRITE BENCHMARKS

Website Serving Images

- Access 1 image 1000 times
- Latency measured for each access
- Start measuring immediately
- 3 runs
- Find mean
- Dev machine





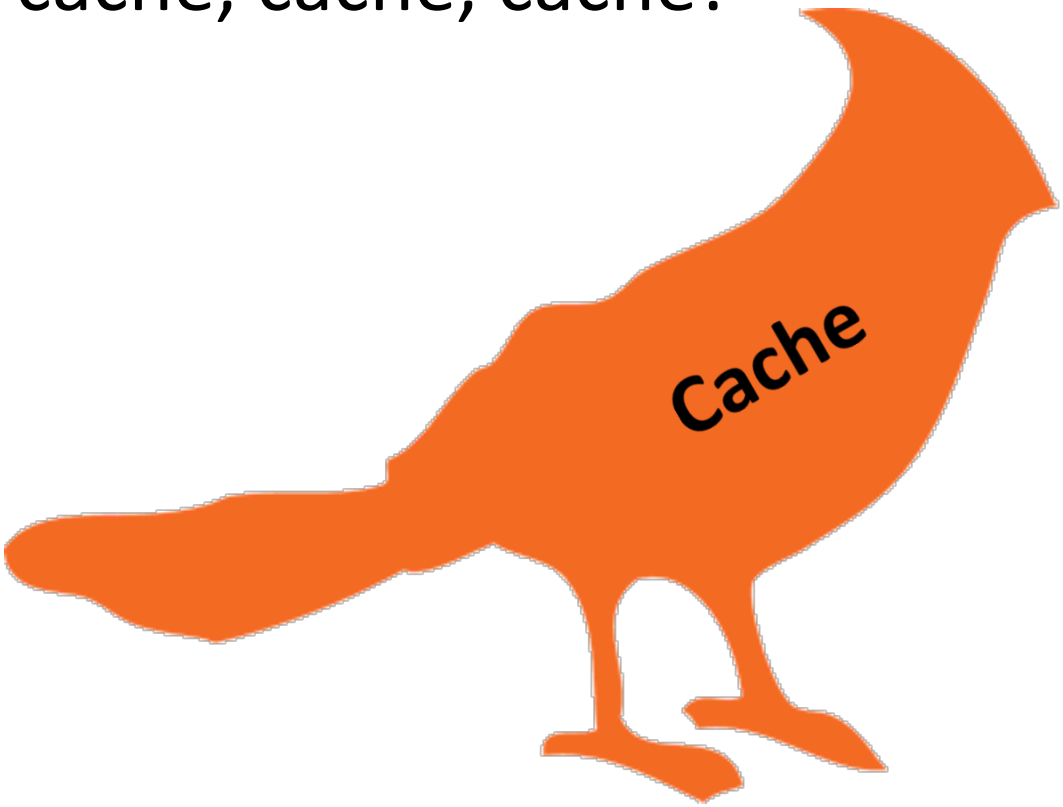
WHAT'S WRONG WITH THIS BENCHMARK?

You're wrong about the machine

BENCHMARK SETUP & RESULTS: COMMON PITFALLS

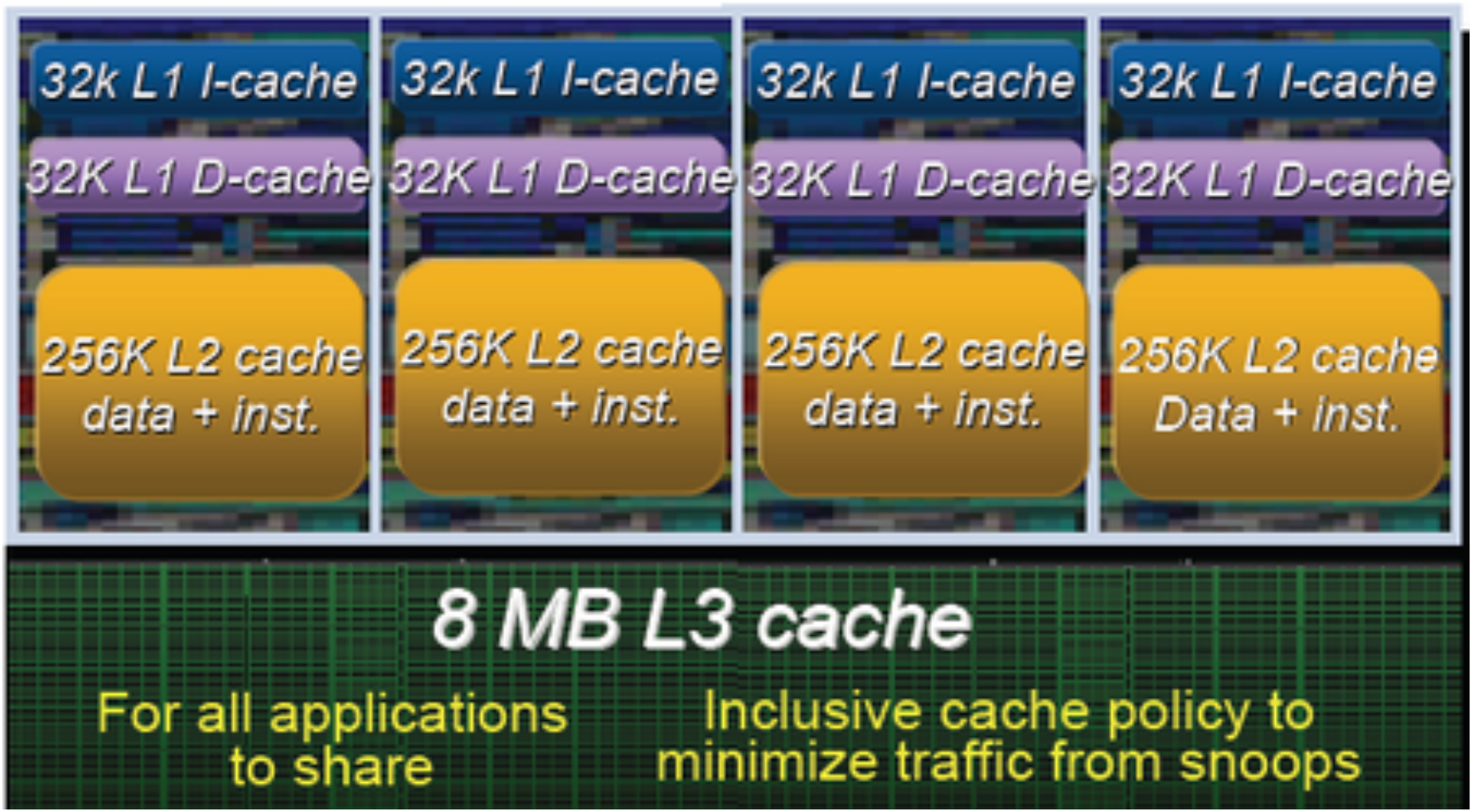
Wrong About the Machine

- Cache, cache, cache, cache!



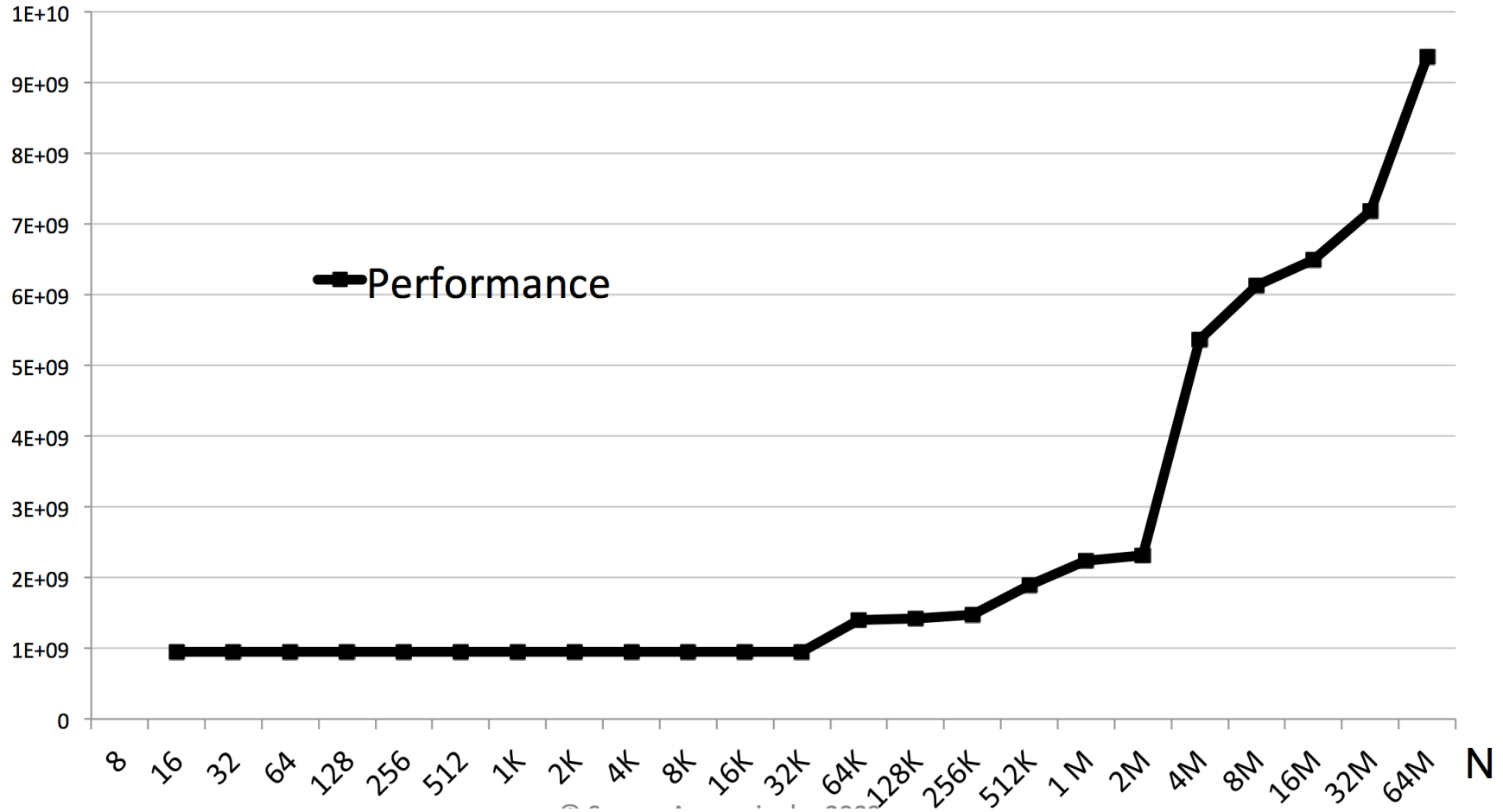
It’s Caches All The Way Down

It’s Caches All The Way Down



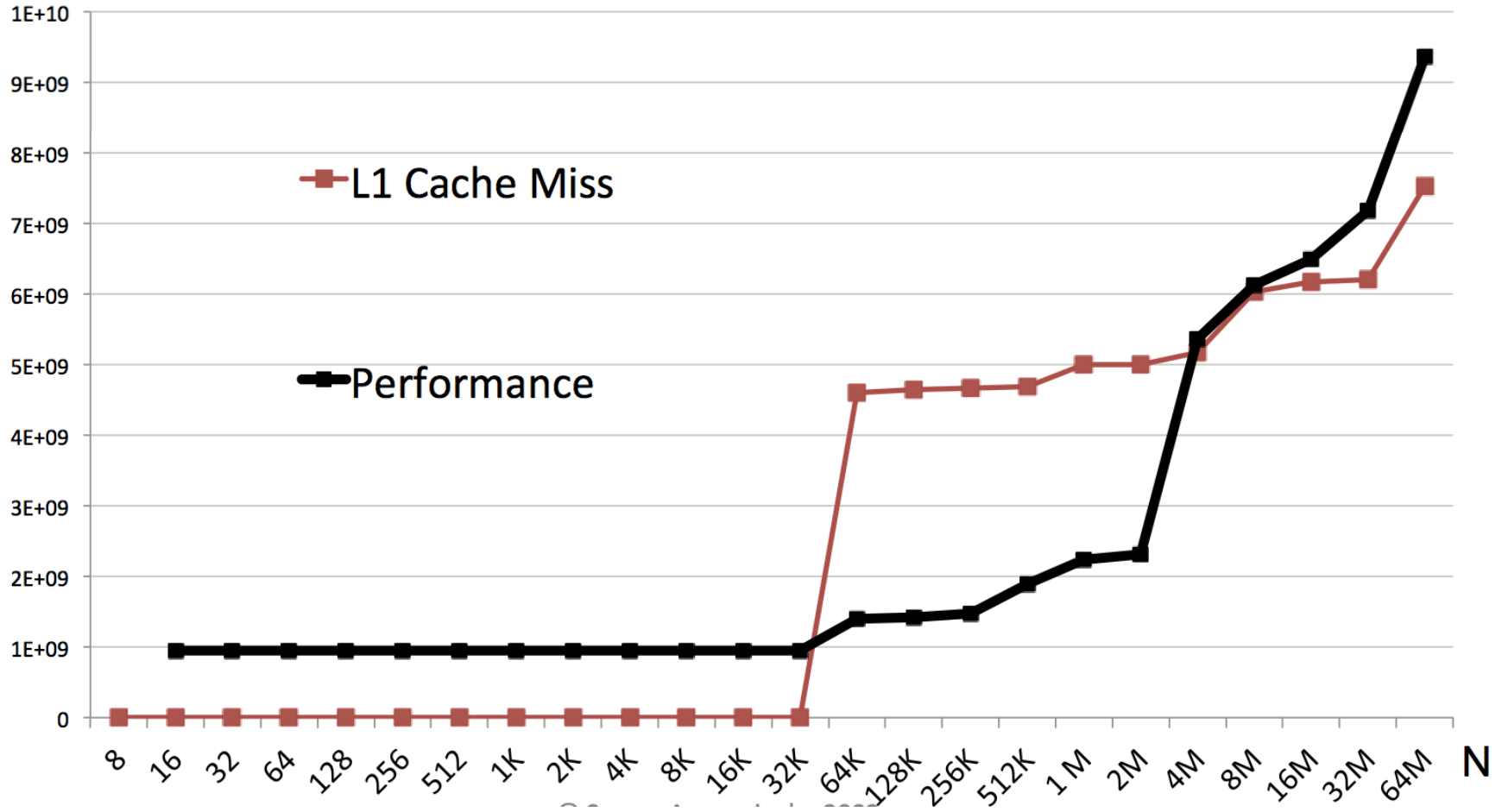
Caches in Benchmarks

Prof. Saman Amarasinghe, MIT 2009



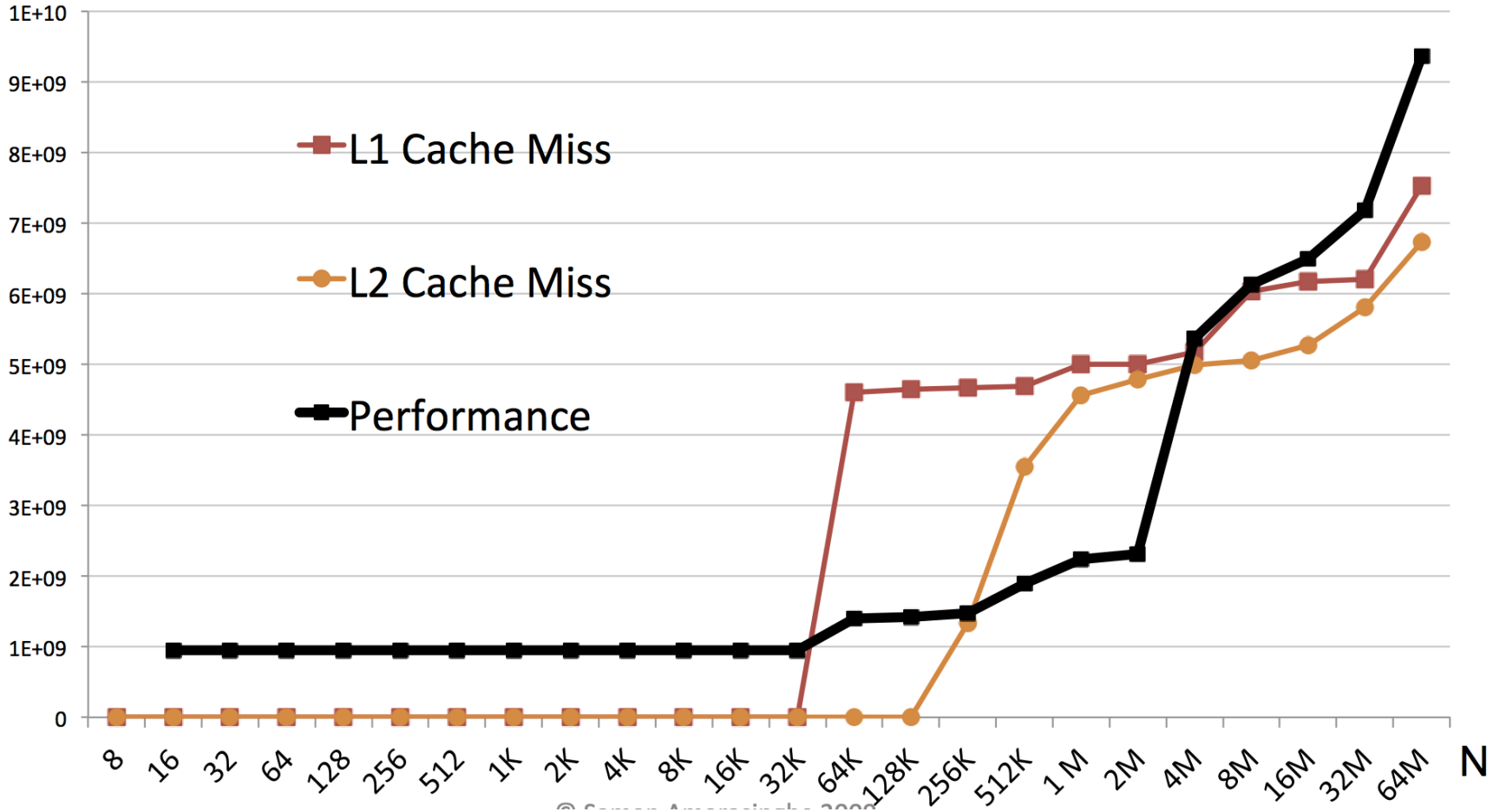
Caches in Benchmarks

Prof. Saman Amarasinghe, MIT 2009



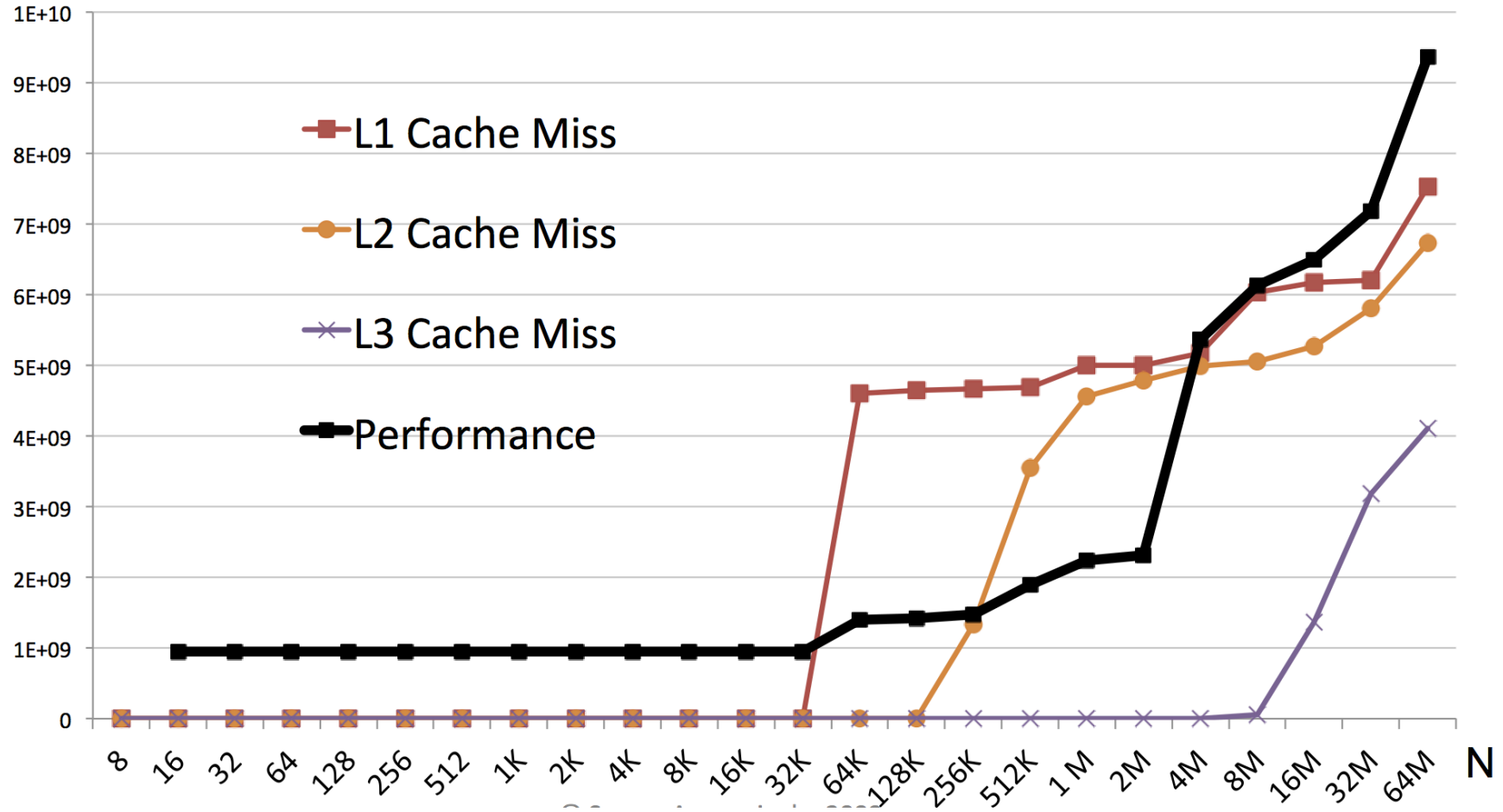
Caches in Benchmarks

Prof. Saman Amarasinghe, MIT 2009



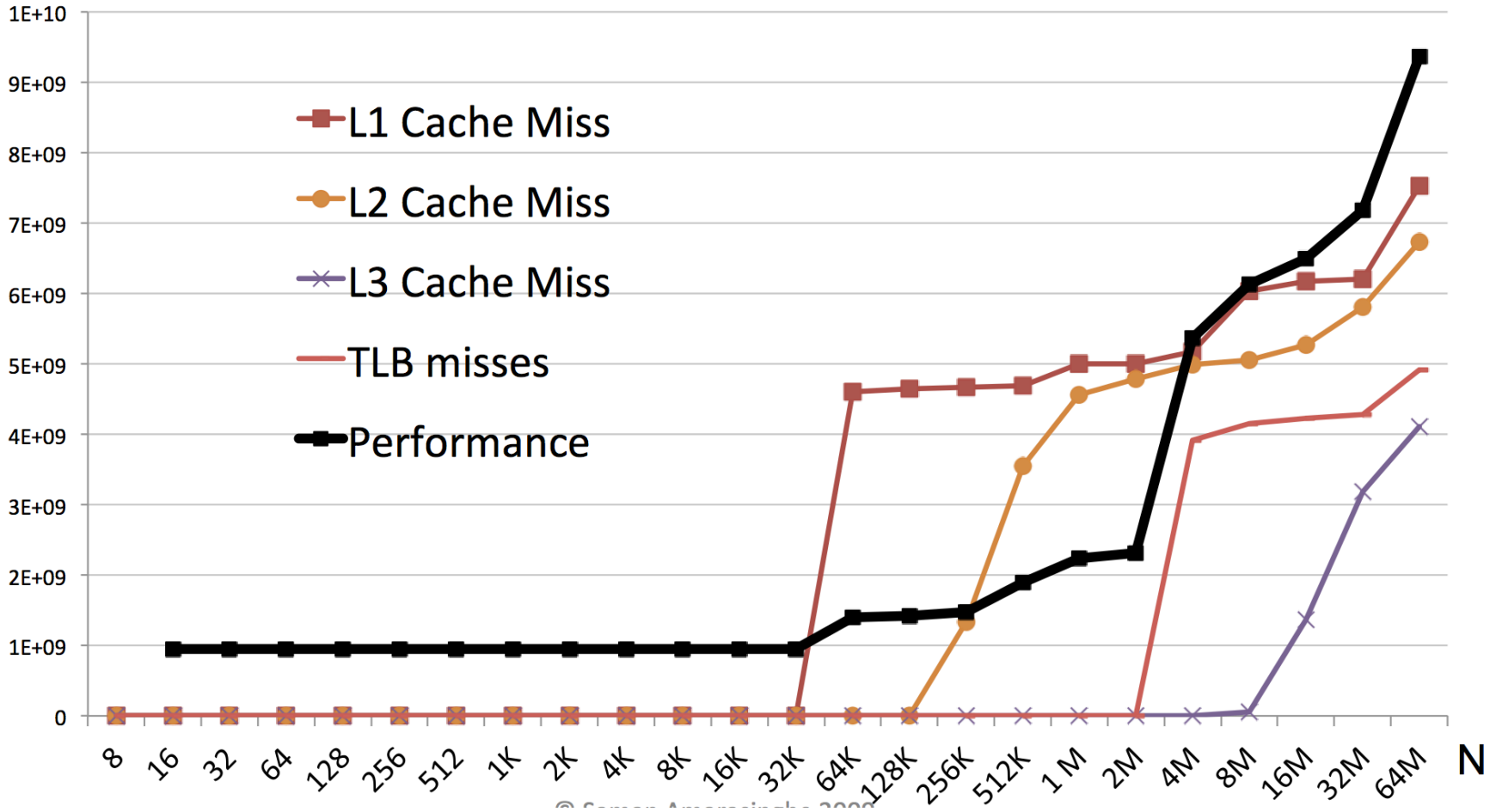
Caches in Benchmarks

Prof. Saman Amarasinghe, MIT 2009



Caches in Benchmarks

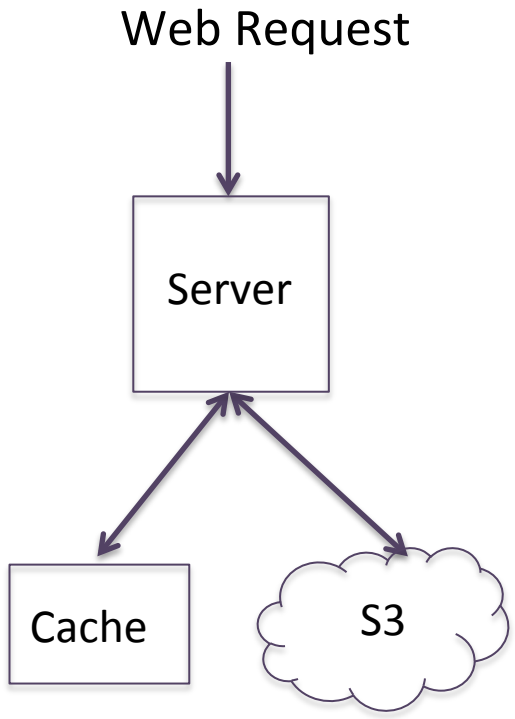
Prof. Saman Amarasinghe, MIT 2009



Website Serving Images

Access 1 image 1000 times

- Latency measured for each access
- Start measuring immediately
- 3 runs
- Find mean
- Dev machine

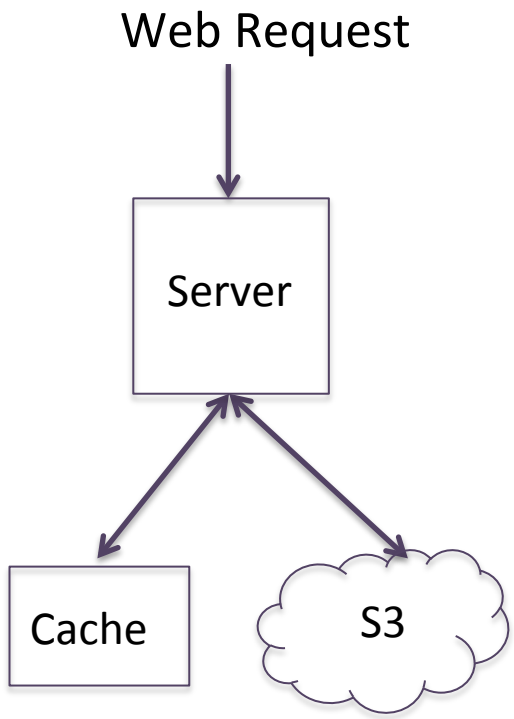


Wrong About the Machine

- Cache, cache, cache, cache!
- Warmup & Timing

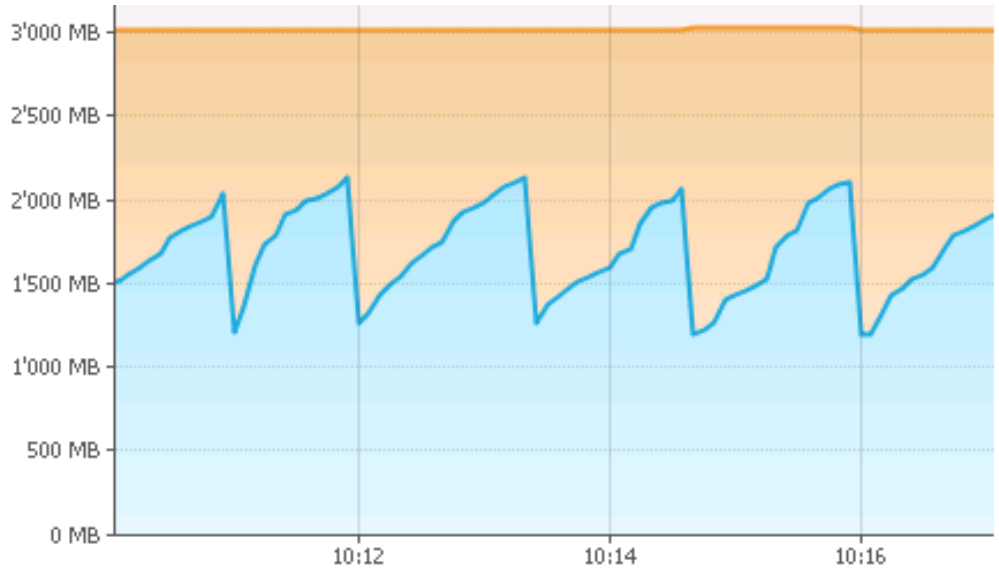
Website Serving Images

- ✗ Access 1 image 1000 times
  - Latency measured for each access
- ✗ Start measuring immediately
  - 3 runs
  - Find mean
  - Dev machine



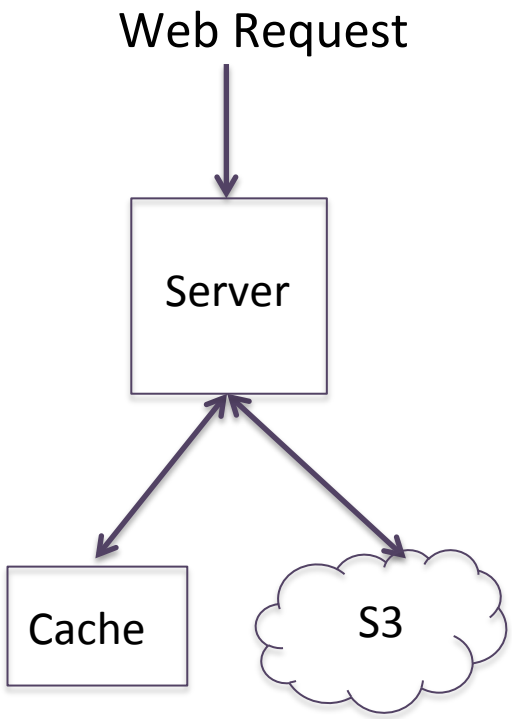
Wrong About the Machine

- Cache, cache, cache, cache!
- Warmup & Timing
- Periodic interference



Website Serving Images

- ~~Access 1 image 1000 times~~
- ~~Latency measured for each access~~
- ~~Start measuring immediately~~
- 3 runs
- Find mean
- Dev machine

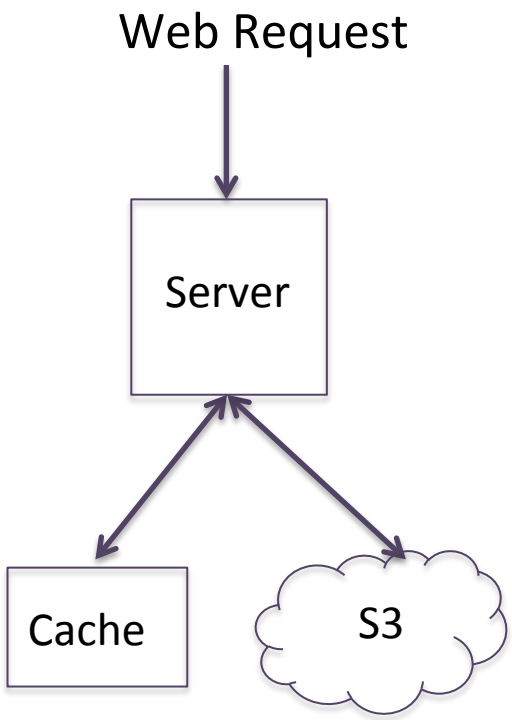


Wrong About the Machine

- Cache, cache, cache, cache!
- Warmup & Timing
- Periodic interference
- Different specs in test vs prod machines

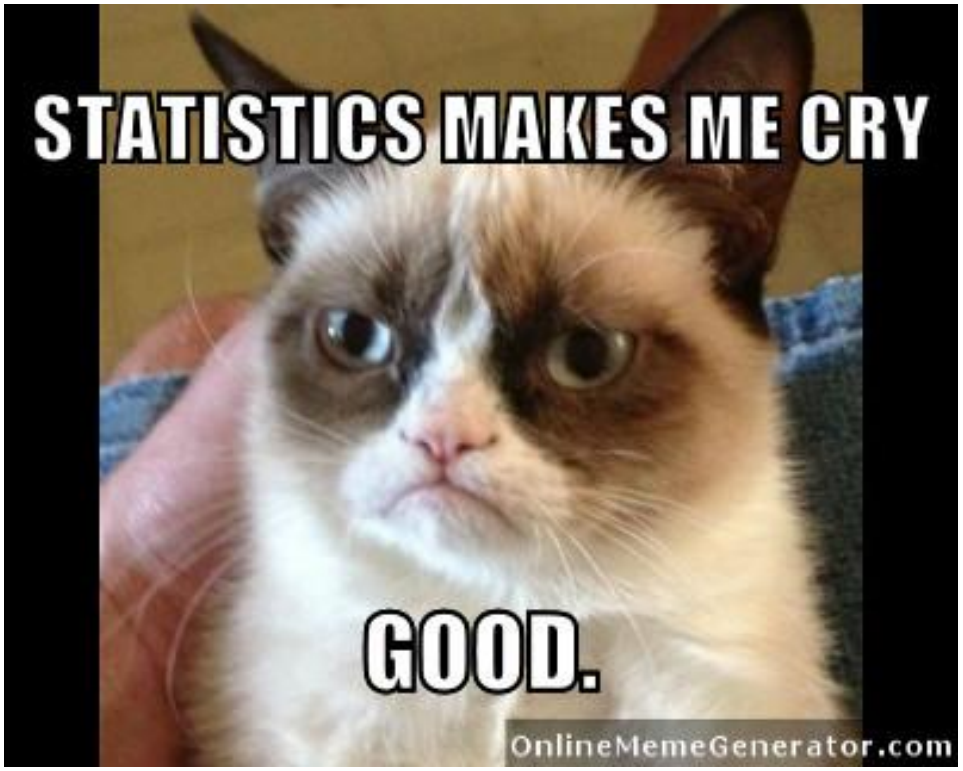
Website Serving Images

- ✗ Access 1 image 1000 times
- ✗ Latency measured for each access
- ✗ Start measuring immediately
  - 3 runs
  - Find mean
- ✗ Dev machine



Wrong About the Machine

- Cache, cache, cache, cache!
- Warmup & Timing
- Periodic interference
- Different specs in test vs prod machines
- Power mode changes



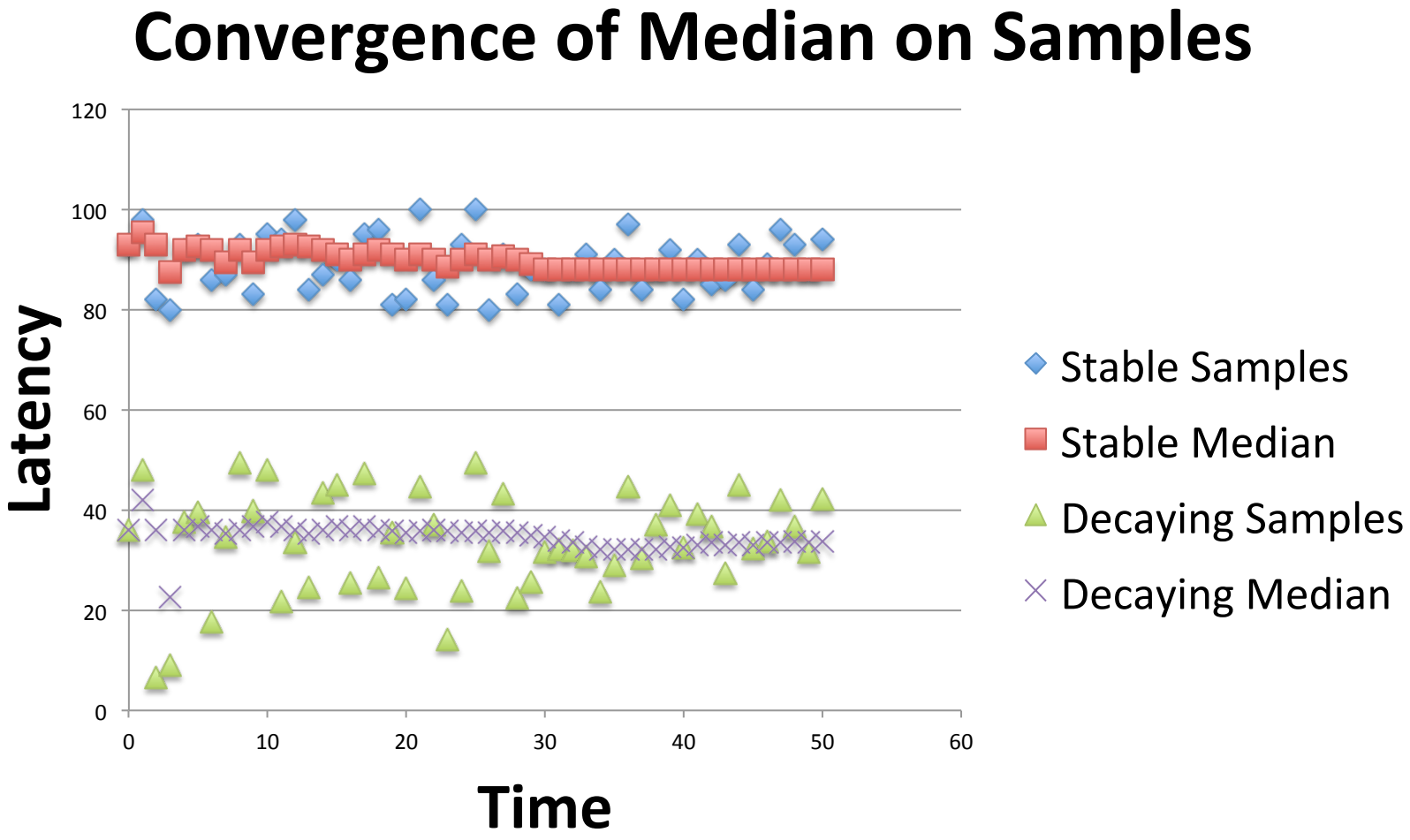
You're wrong about the stats

BENCHMARK SETUP & RESULTS: COMMON PITFALLS

Wrong About Stats

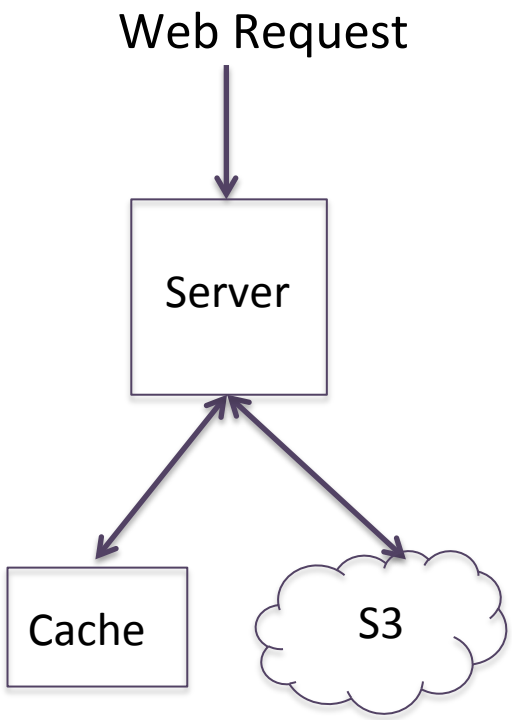
- Too few samples

Wrong About Stats



Website Serving Images

- ~~Access 1 image 1000 times~~
- ~~Latency measured for each access~~
- ~~Start measuring immediately~~
- ~~3 runs~~
- Find mean
- ~~Dev machine~~

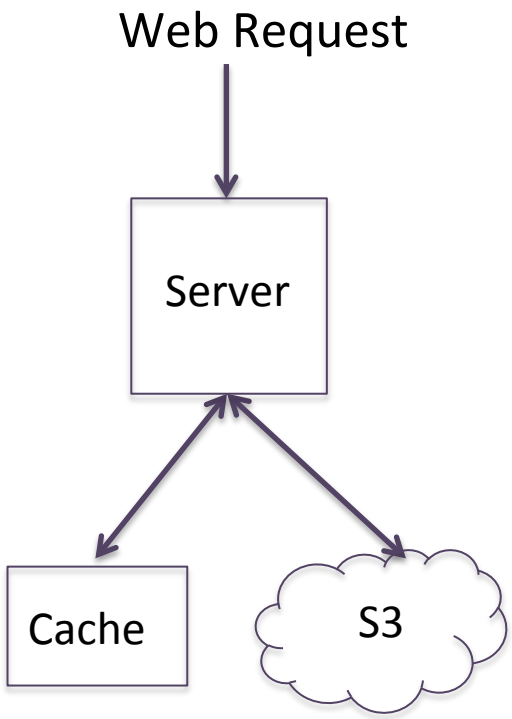


Wrong About Stats

- Too few samples
- Non-Gaussian

Website Serving Images

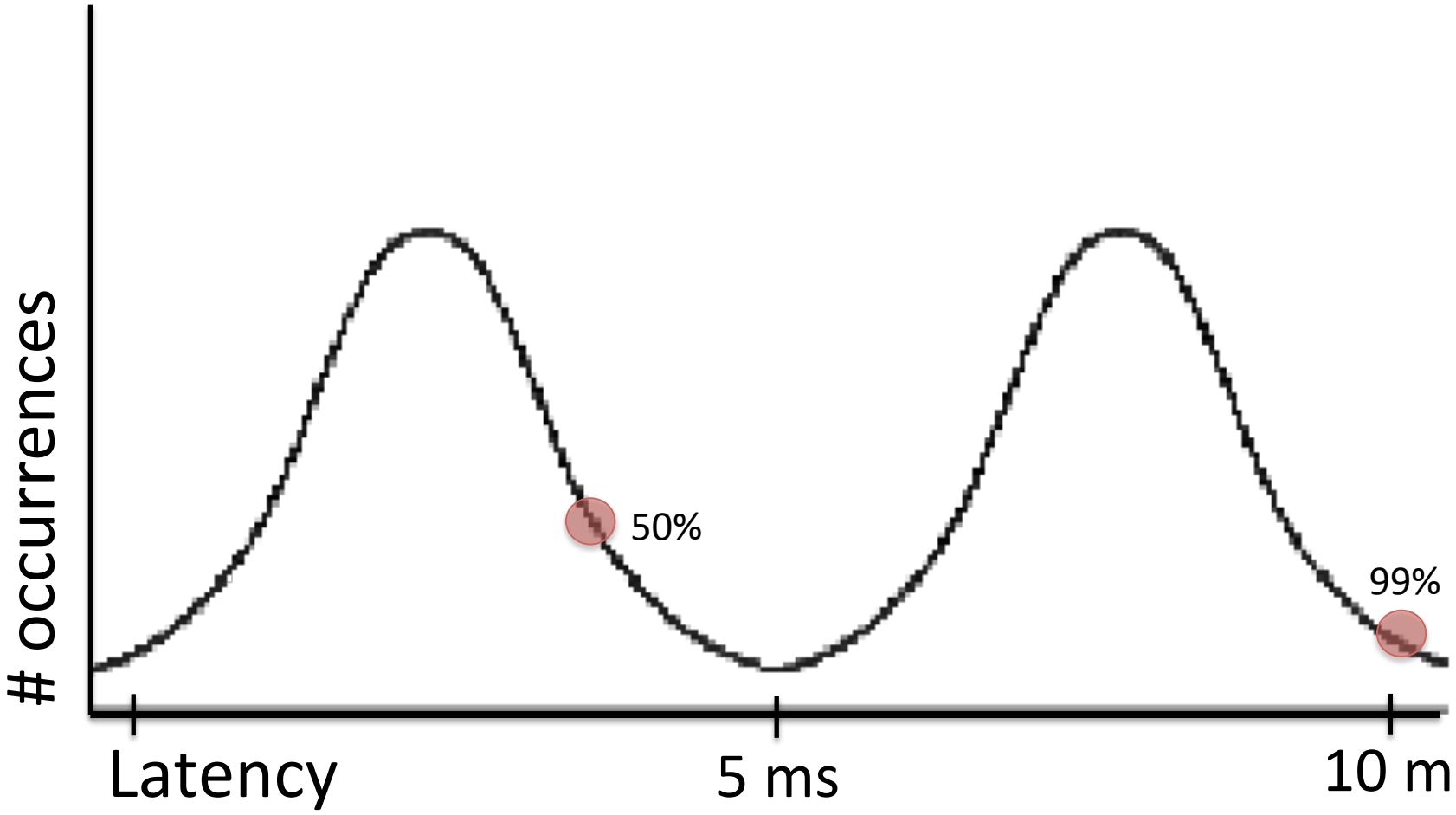
- Access 1 image 1000 times
- Latency measured for each access
- Start measuring immediately
- 3 runs
- Find mean
- Dev machine



Wrong About Stats

- Too few samples
- Non-Gaussian
- Multimodal distribution

Multimodal Distribution



Wrong About Stats

- Too few samples
- Non-Gaussian
- Multimodal distribution
- Outliers

You're wrong about what matters

BENCHMARK SETUP & RESULTS: COMMON PITFALLS

Wrong About What Matters

- Premature optimization

“Programmers waste enormous amounts of time thinking about … the speed of noncritical parts of their programs … Forget about small efficiencies …97% of the time: **premature optimization is the root of all evil**. Yet we should not pass up our opportunities in that critical 3%.”

-- Donald Knuth

Wrong About What Matters

- Premature optimization
- Unrepresentative Workloads

Wrong About What Matters

- Premature optimization
- Unrepresentative Workloads
- Memory pressure

The How

BECOMING LESS WRONG

Becoming Less Wrong

User Actions Matter

X > Y for workload Z

with trade offs A, B, and C

- http://www.toomuchcode.org/

Becoming Less Wrong

Profiling

Code Instrumentation

Aggregate Over Logs

Traces

Microbenchmarking: Blessing & Curse

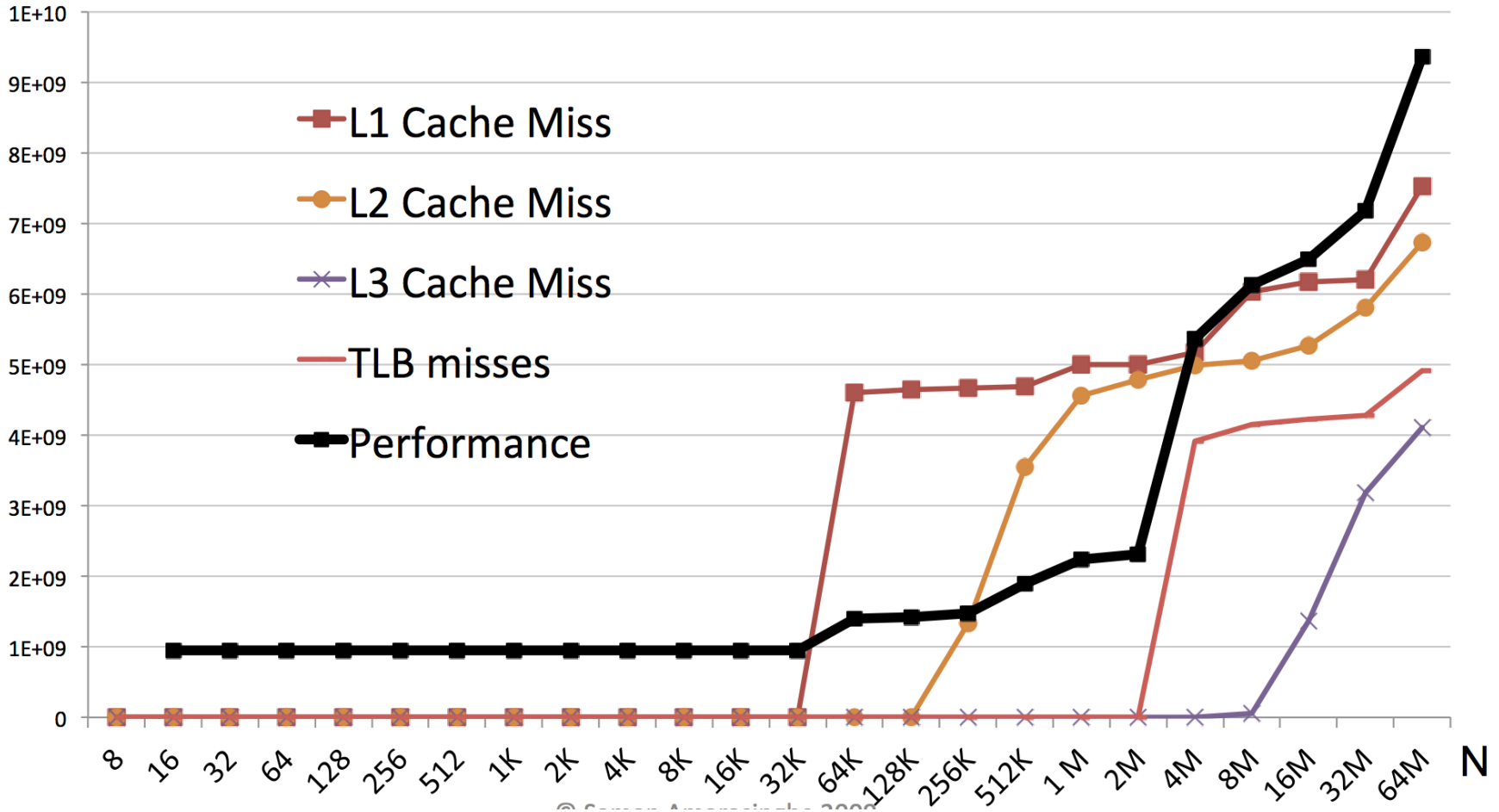
+ Quick & cheap
+ Answers narrow ?s well
– Often misleading results
– Not representative of the program

Microbenchmarking: Blessing & Curse

- Choose your N wisely

Choose Your N Wisely

Prof. Saman Amarasinghe, MIT 2009



Microbenchmarking: Blessing & Curse

- Choose your N wisely
- Measure side effects

Microbenchmarking: Blessing & Curse

- Choose your N wisely
- Measure side effects
- Beware of clock resolution

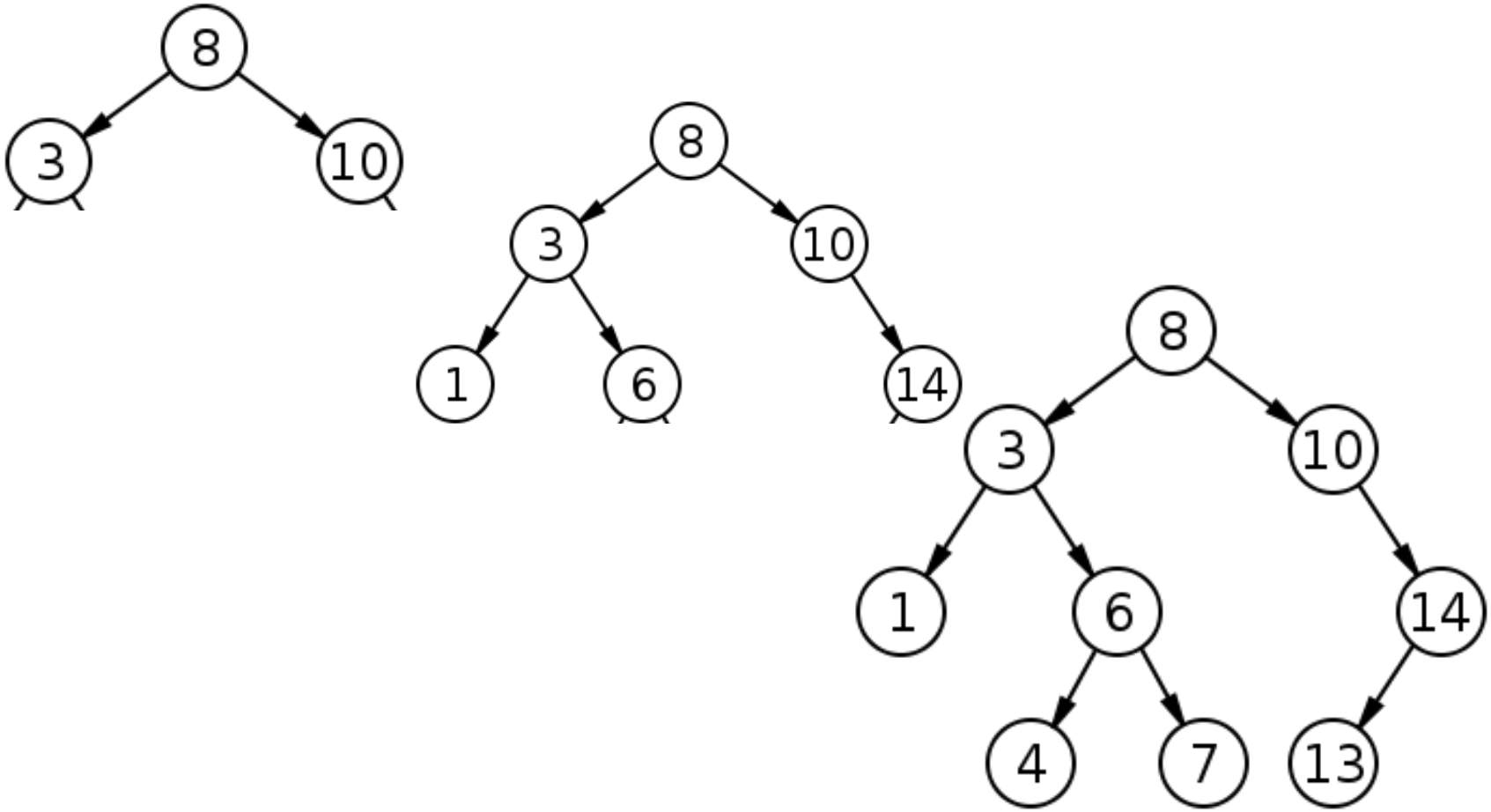
Microbenchmarking: Blessing & Curse

- Choose your N wisely
- Measure side effects
- Beware of clock resolution
- Dead Code Elimination

Microbenchmarking: Blessing & Curse

- Choose your N wisely
- Measure side effects
- Beware of clock resolution
- Dead Code Elimination
- Constant work per iteration

Non-Constant Work Per Iteration

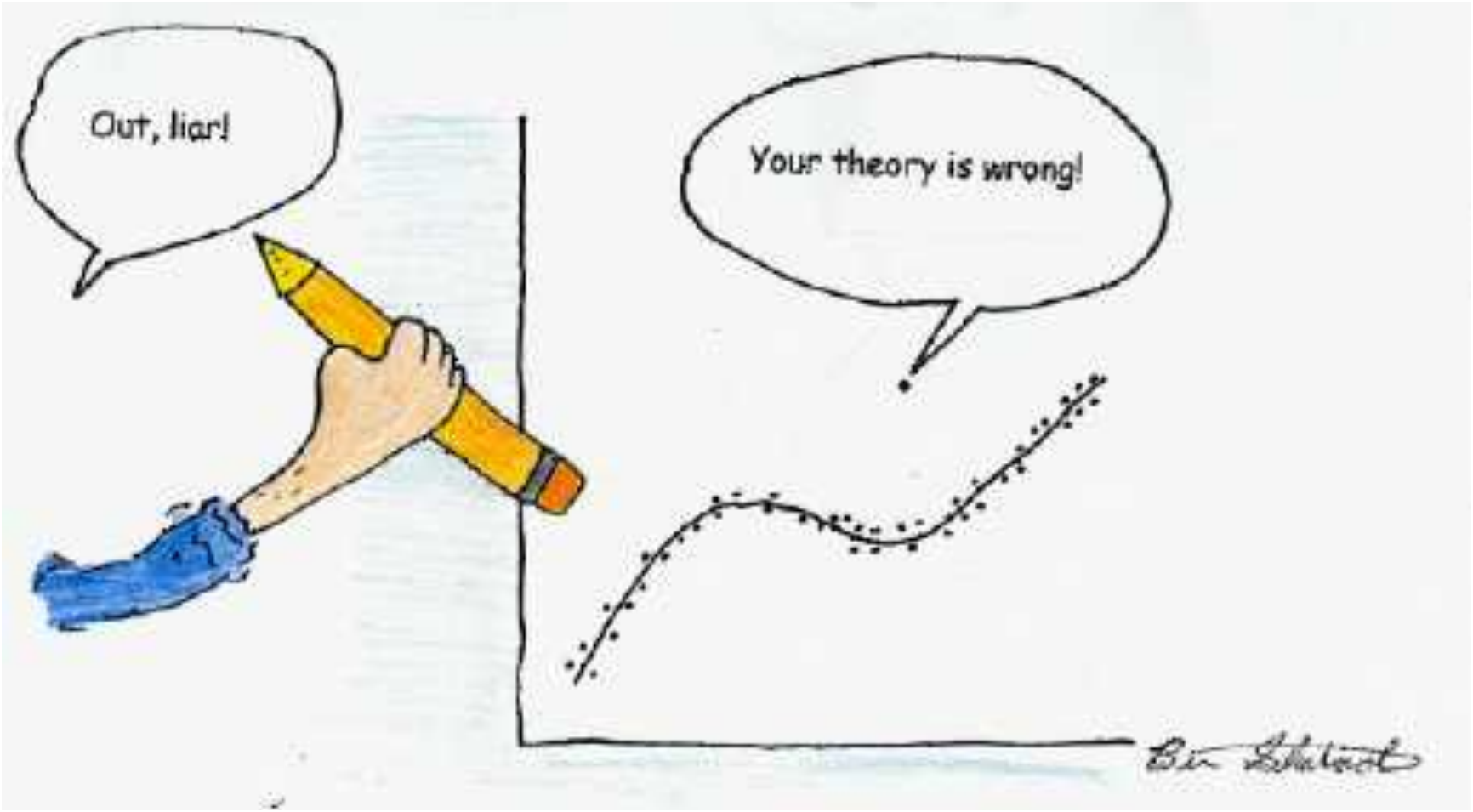


Follow-up Material

- *How NOT to Measure Latency* by Gil Tene
- *Taming the Long Latency Tail* on highscalability.com
- *Performance Analysis Methodology* by Brendan Gregg
- *Robust Java benchmarking* 1 & 2 by Brent Boyer
- Benchmarking articles by Aleksey Shipilëv

Takeaway #1: Cache

Takeaway #2: Outliers



Takeaway #3: Workload

Benchmarking: You're Doing It Wrong

Aysylu Greenberg
@aysylu22

Google