IBM Research

How to cheat at benchmarking

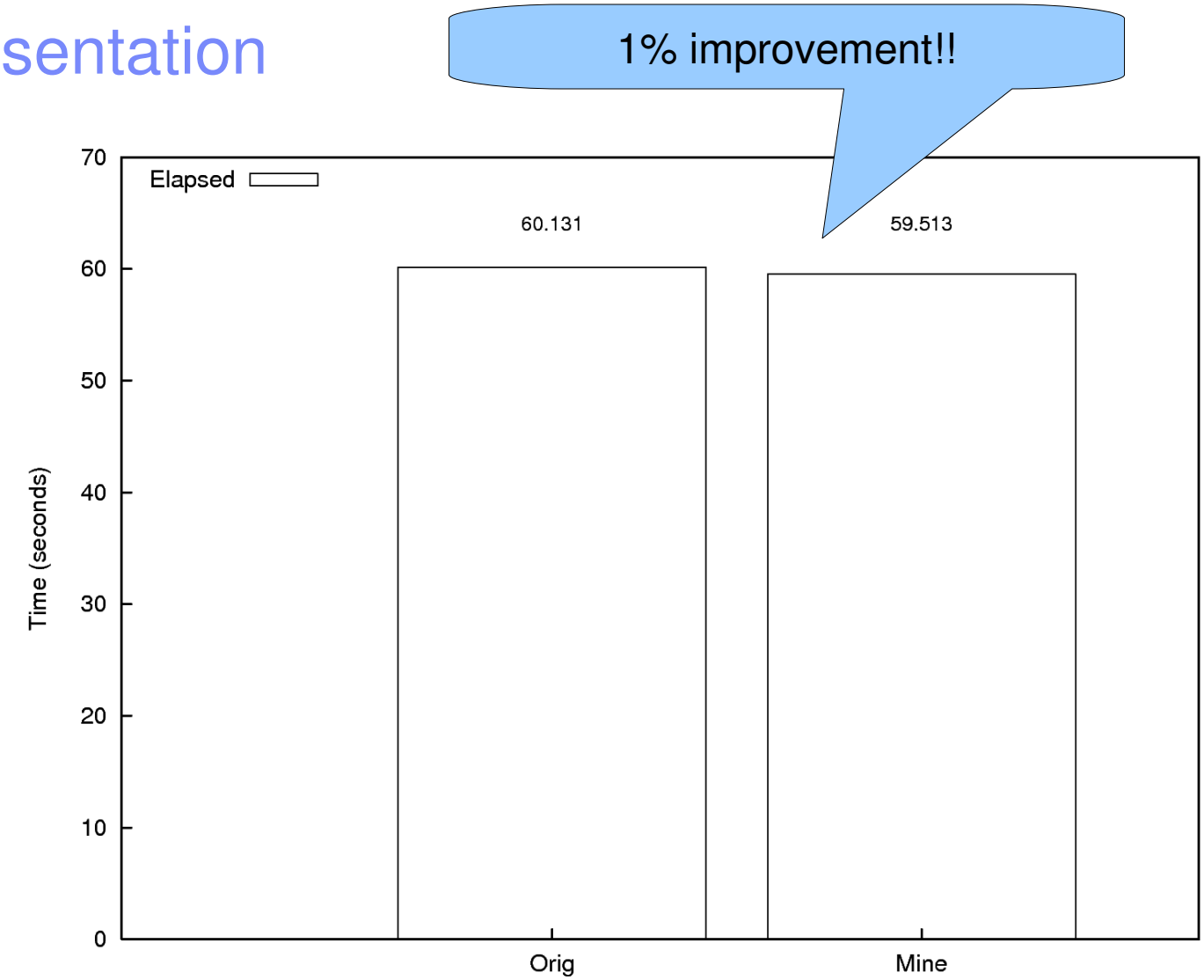
Avishay Traeger (IBM HRL)

Erez Zadok (Stony Brook University)

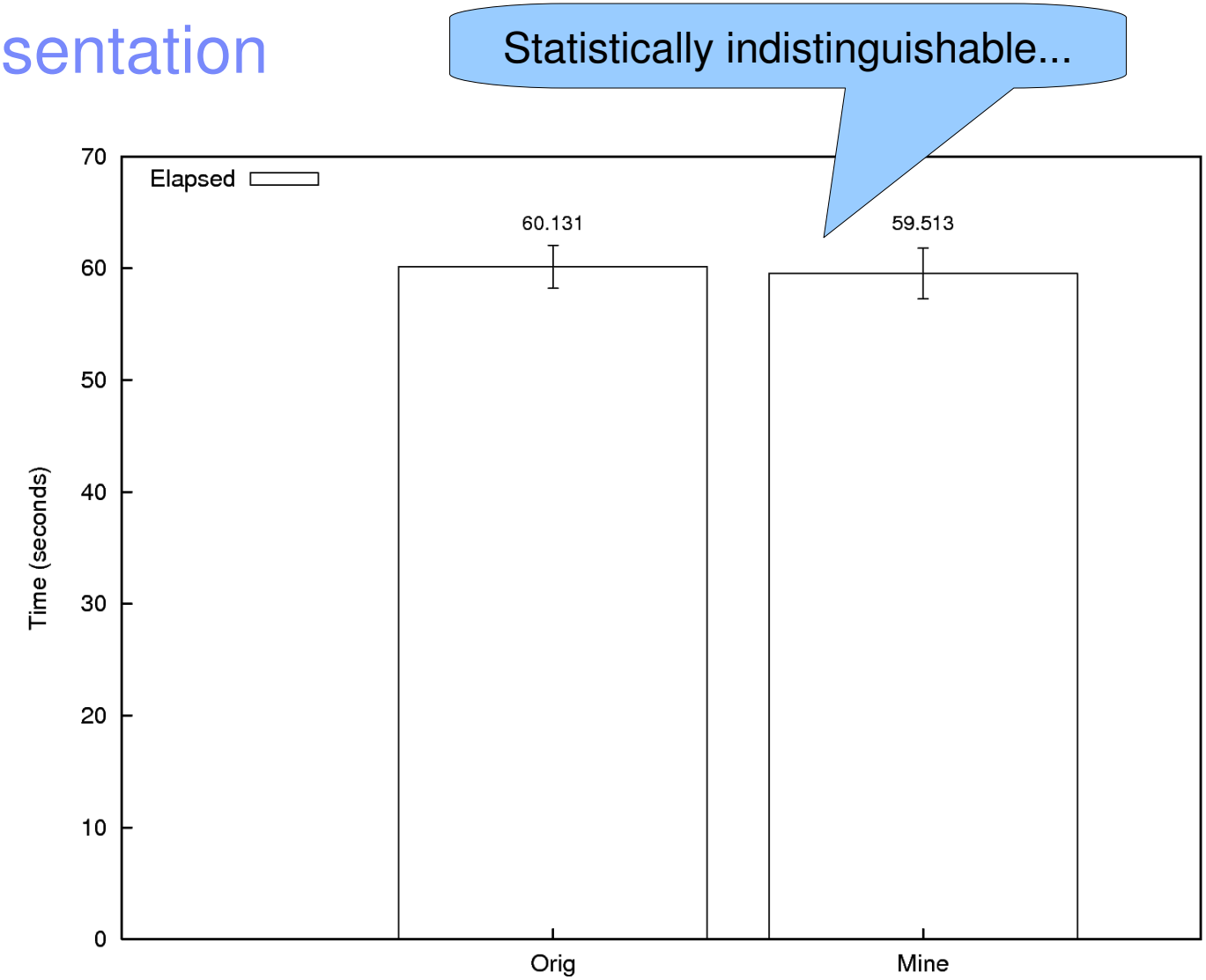
Welcome!

- Thanks for coming to our BoF on how to cheat at benchmarking
- You should all be ashamed of yourselves! ☺
- (Everyone who didn't show up already knows)

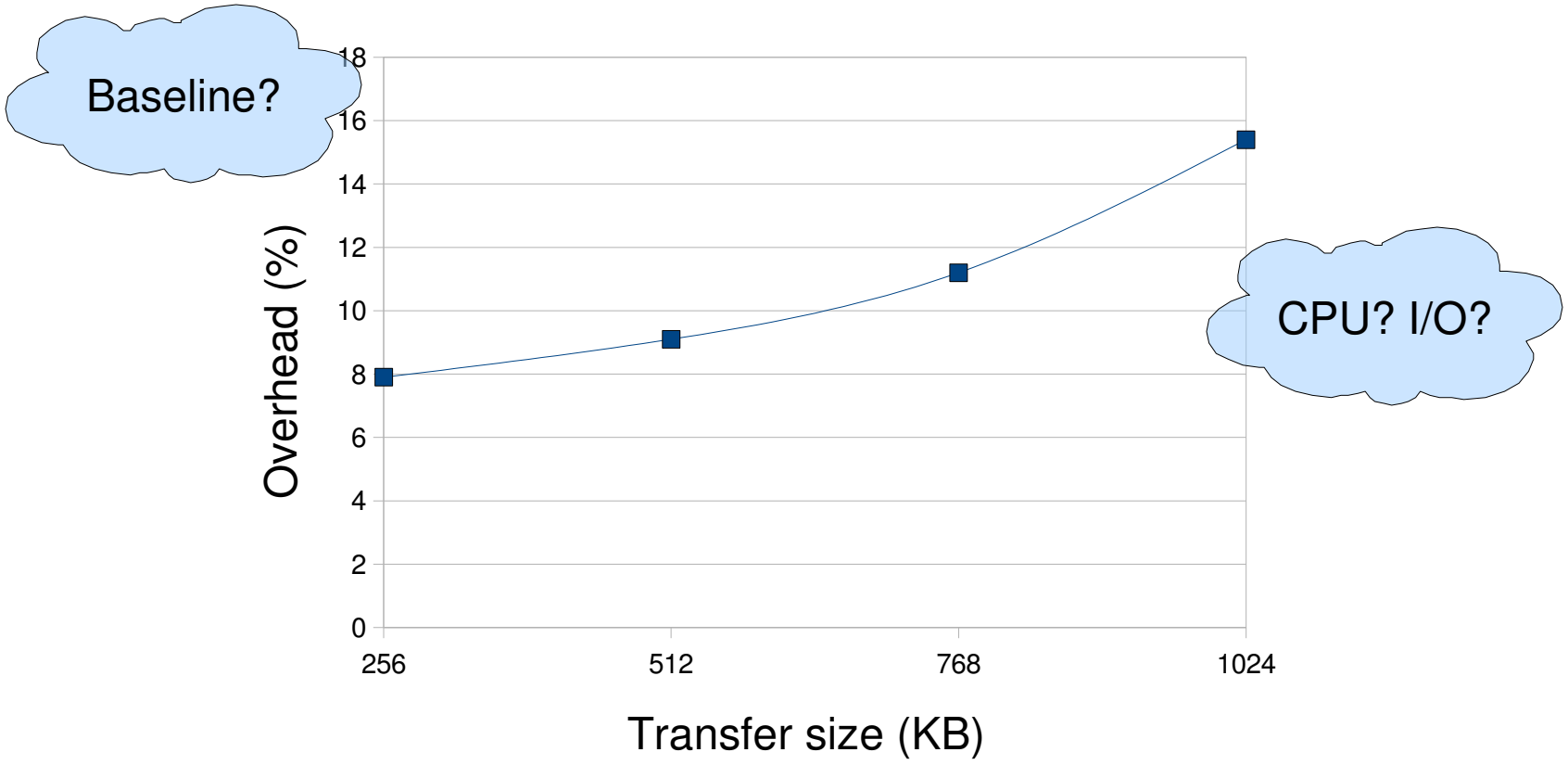
1. Presentation



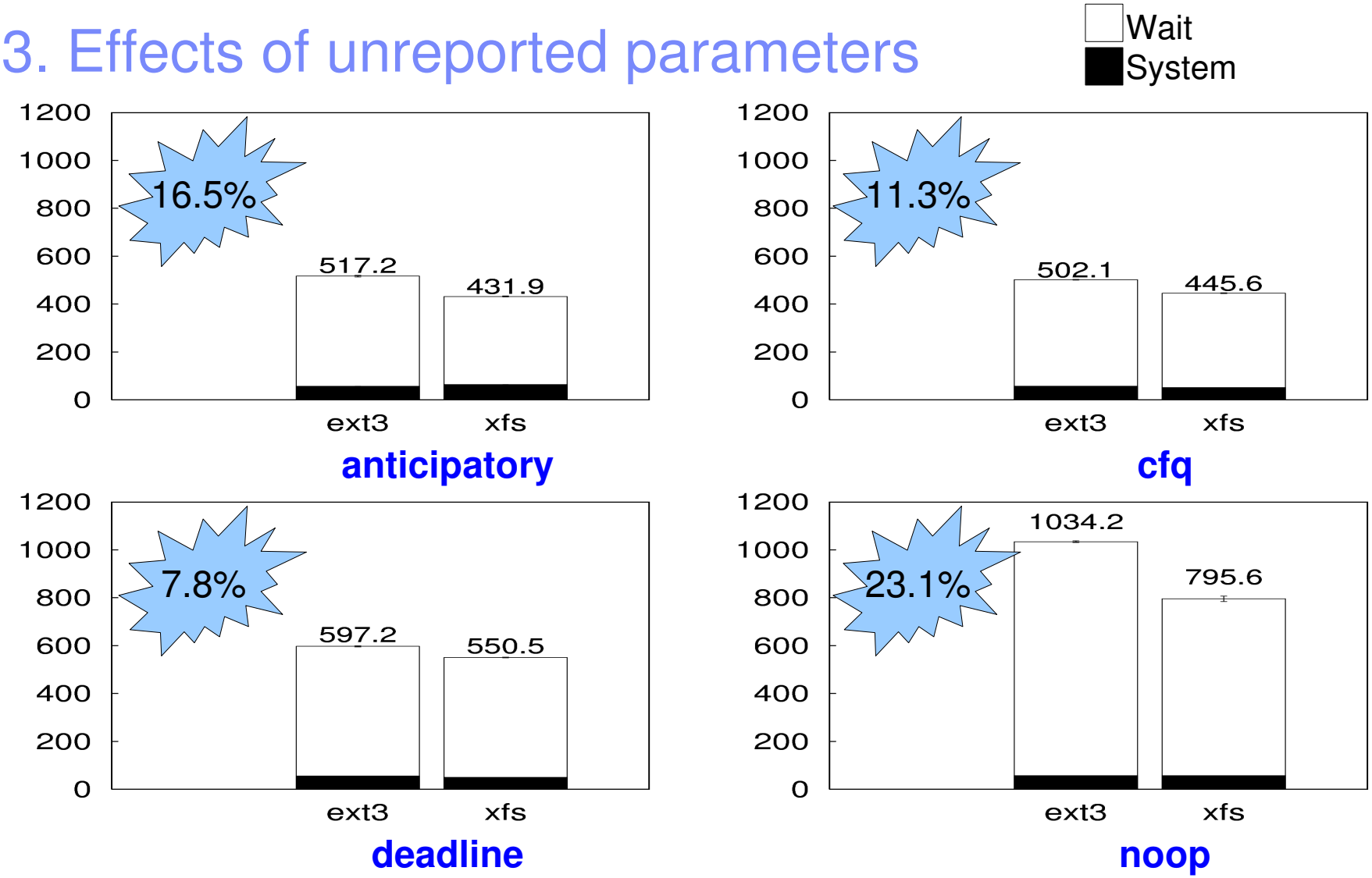
1. Presentation



2. Presentation



3. Effects of unreported parameters



iozone random read/write, 4MB chunks, 10 processes

Some other examples of “cheating”

4. “My new feature uses lots of CPU, so I'll benchmark it on a machine with a quad-core 3.5 GHz CPU and 5400K RPM disks. Now both the *before* and *after* systems will have disk bottlenecks, and the CPU overhead won't show too much!”

   What would happen with a 2GHz CPU and 15000K RPM disks?

Nobody will question your evaluation if...

- Your CPU is anywhere from 1.7—3.5GHz
- You used 1 core vs. 4 cores
- Your kernel version age from 2.6.0—2.6.28
- You used a slow vs. fast disk
- You used ext2, ext3, xfs, or reiserfs

Some other examples of “cheating”

5. “My file system's performance is awful when it comes to I/O, but the overheads don't look too bad with a compile benchmark...”

   (Or the other way around)

6. “We ran Postmark with 10,000 files, 200,000 transactions, and left the other parameters as defaults.”

   Who's going to realize that the default maximum file size is 10KB and that the whole working set easily fits in RAM?

7. “We ran mkfs on the file system and flushed caches between runs to ensure workload consistency.”

   This makes the workload consistently unrealistic (see talk on *Impressions* this morning)

What's wrong with systems benchmarking today?

- Examples 1 and 2 (presentation):
  - Not enough scientific methodology or statistical rigor
  - I also messed up here – the calculation for standard deviation that I used assumes a normal distribution, which I didn't check
  - Possible answers: Education, raising standards

What's wrong with systems benchmarking today?

- Example 3 (I/O schedulers), Example 4 (CPU vs. disk bottleneck):
  - Full descriptions of systems aren't provided
    - Obvious answer: provide full descriptions
  - Even if they were, systems are too complex
    - Possible answer: Modeling?
  - Results are extremely difficult to reproduce
    - Possible answers: VMs? Modeling?

What's wrong with systems benchmarking today?

- Example 5 (compile benchmark), Example 6 (Postmark):
  - Popular benchmarks don't scale
  - Popular benchmarks may not always be appropriate
  - Possible answers: Guidelines, standardized benchmarks?

  [What impact do benchmarks have on system design in industry?]

  [How do we create good general-purpose benchmarks?]

What's wrong with systems benchmarking today?

- Example 7 (clean system between runs):
  - We usually benchmark in a sterile, unrealistic environment
  - Fragmentation? File system contents? Cache contents?
  - Work has been done here, but nobody uses it

What we've done so far...

- BoF at FAST '05
- Survey article with guidelines in TOS (Vol 4, Issue 2 – May '08)
- Workshop at UCSC in May '08 (summary in *;login:*)
- Website: http://fsbench.filesystems.org/
- Mailing list

- We hope to encourage more research and activity in this area

Discussion

- Improving education
- Raising standards
- Modeling for to help understand systems/workloads
- VMs to help reproduce results
- Benchmarking guidelines
- Standardized benchmarks
- Aging
- What does 35% overhead really mean?

- How do we make things better?
- Action items?

Thanks!

How to cheat at benchmarking

http://fsbench.filesystems.org/

Avishay Traeger (IBM HRL)

Erez Zadok (Stony Brook University)