File System Tracing, Replaying, Profiling, and Analysis on HEC Systems

Erez Zadok

Klaus Mueller

Stony Brook University

www.cs.stonybrook.edu

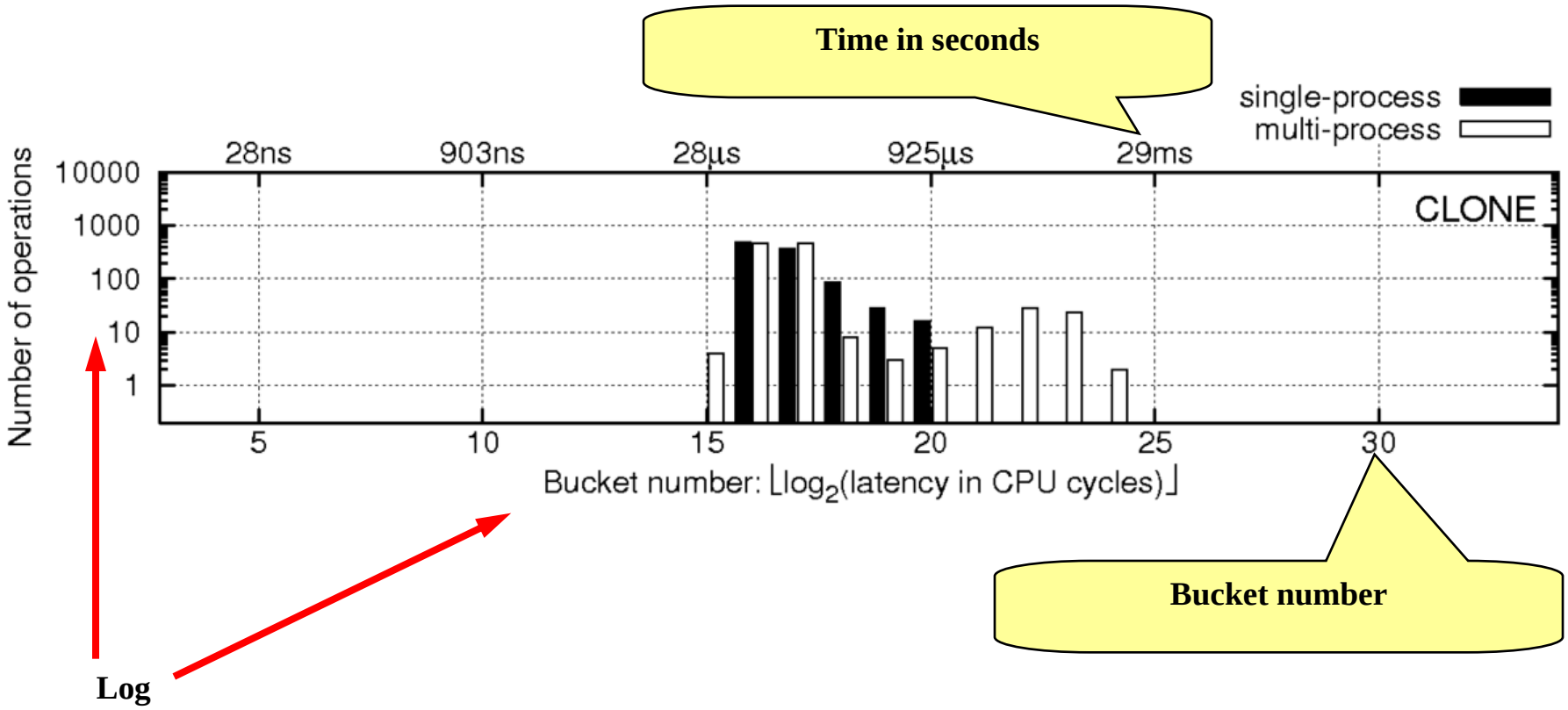
Ethan L. Miller

UC Santa Cruz

www.cs.ucsc.edu

Example OSprof Profile

[OSDI 2006]

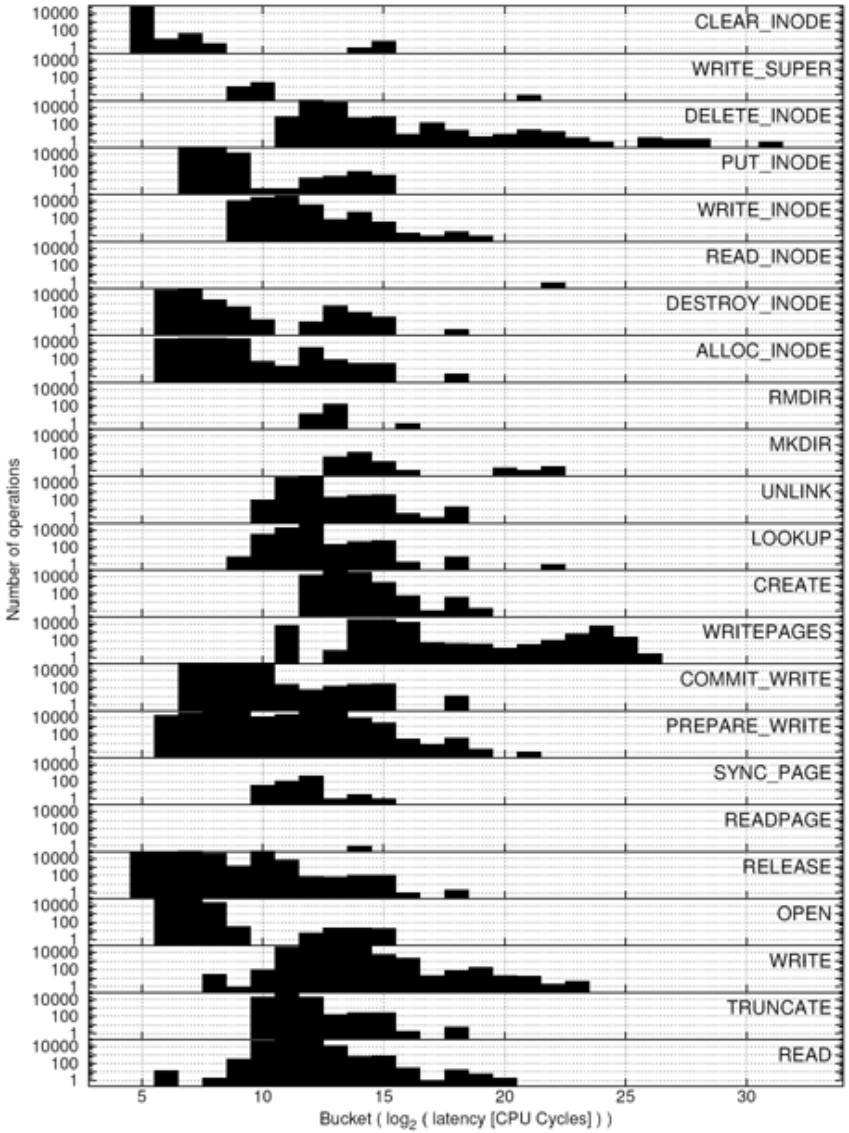


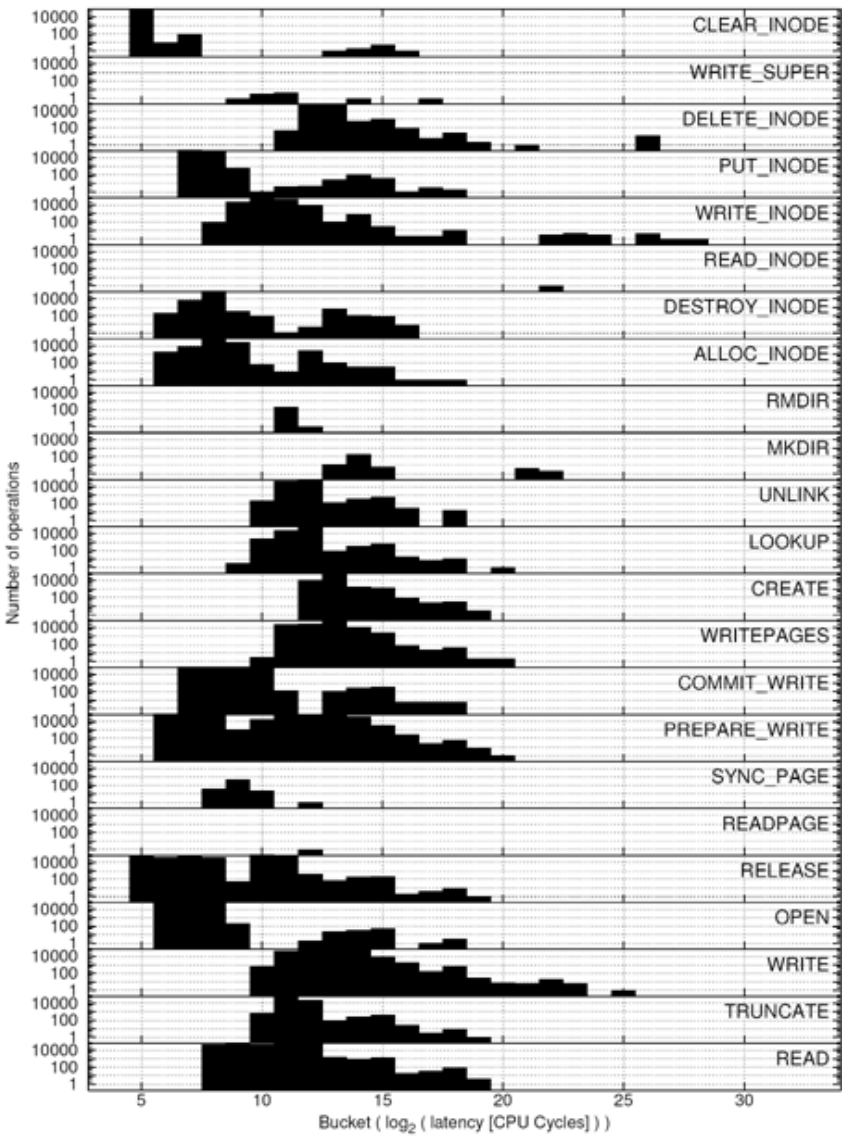
`clone(2)` called in a loop on FreeBSD 6.0

OSprof: Pros & Cons

- ☑ Powerful profiling methodology
  - ☑ Time-lapse profiles
- ☑ Highly portable (Linux, FreeBSD, Windows, user-space, kernel-space)
- ☑ Profile CPU time, I/O, locks, semaphores, scheduler, interrupts, networking protocols
- ☑ Minimal effect on instruction & data caches
  - ☑ Minimal behavior perturbations
  - ☑ Under 4% overhead
- ☒ Manual investigation

Lots of Profiles…





Erez Zadok and Klaus Mueller (Stony Brook University)
Ethan L. Miller (UC Santa Cruz)

Outline

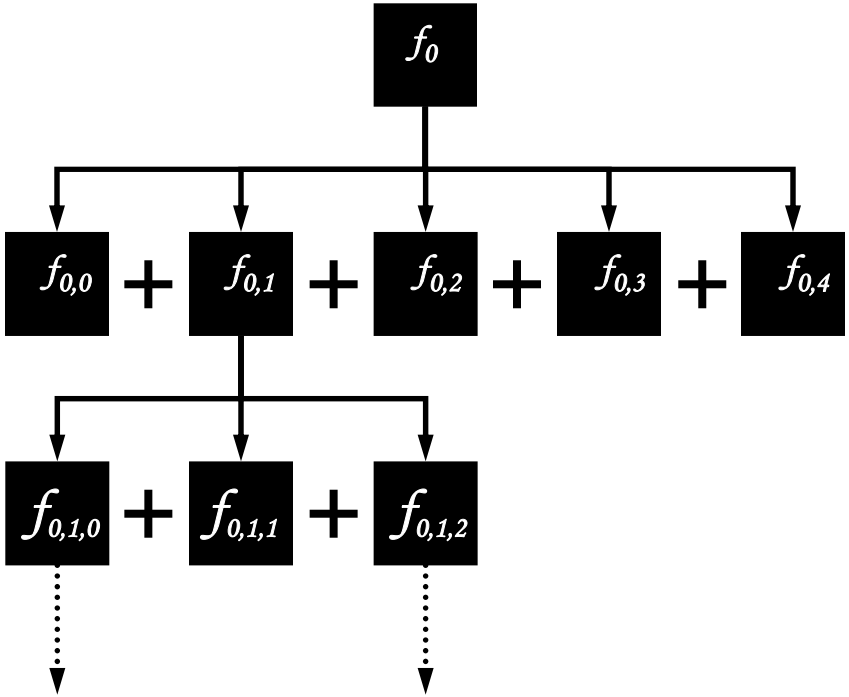
- **Analyzing Root Causes of Latency Distributions**
- Visual Analytics

DARC: Dynamic Analysis of Root Causes

[SIGMETRICS 2008]

- **Goal**: To find the causes for high-level behaviors seen by users
  - Seen as peaks in OSprof profiles
- *Root cause functions*
  - Cause of high-level behaviors
  - Major latency contributors to a given peak
- Dynamically instrument running code
- DARC outputs *root cause paths*

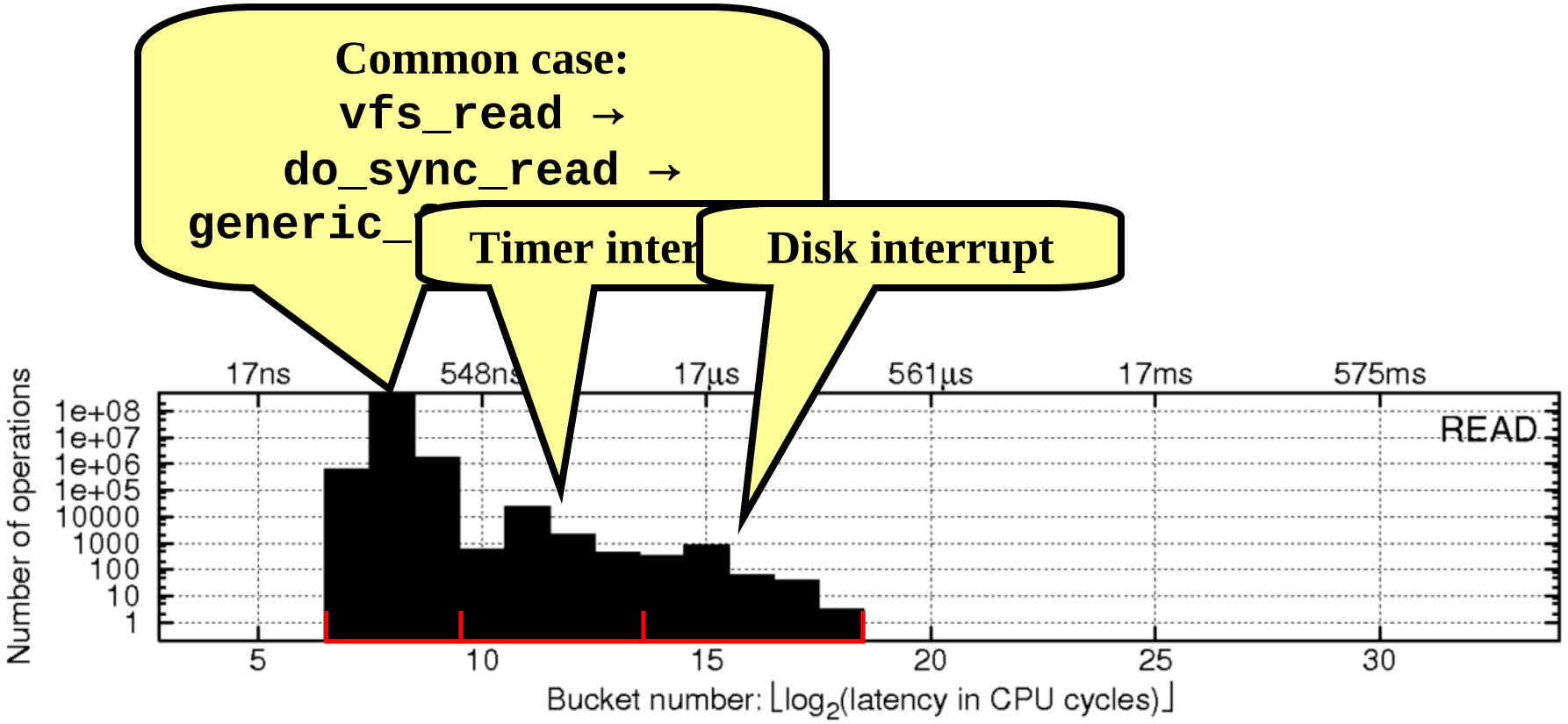
Latencies are Additive



DARC Features

- PID filtering
- Call-path filtering
- Resuming DARC
  - Analysis crosses user-kernel boundary
- Recursion handling, loops
- Indirect call handling
- Preemptive behavior analysis
- Asynchronous path analysis
  - Instrument interrupt handlers

Analyzing Interrupts



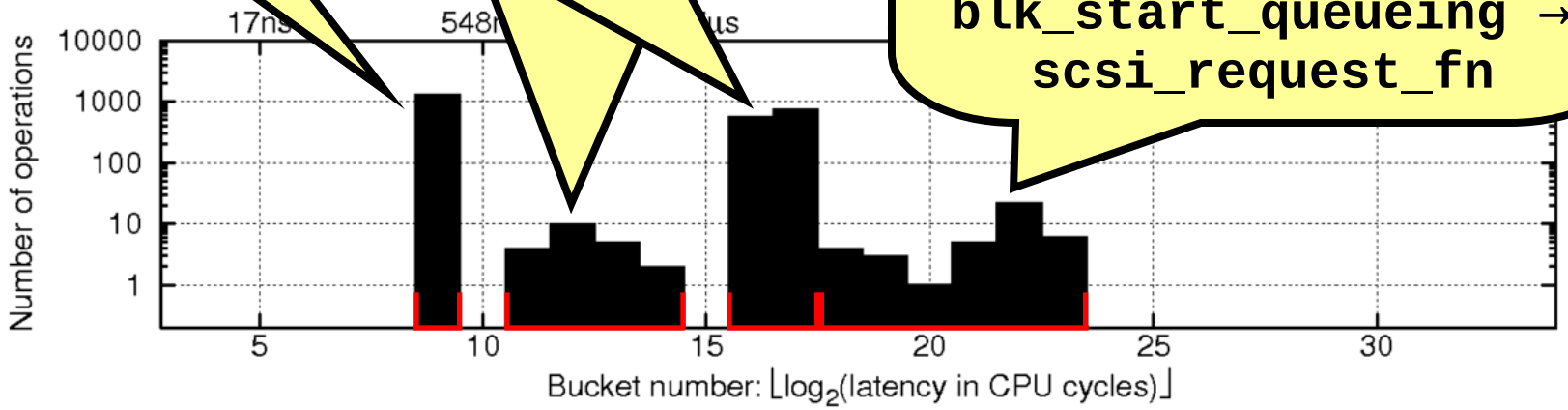
Two processes repeatedly reading zero bytes from a file (ext2)

Analyzing Asynchronous Paths

```
ext2_readdir →
ext2_get_page →
read_cache_page →
read_cache_page_async →
readpage →
readpage →
io_submit →
mit_bio
```

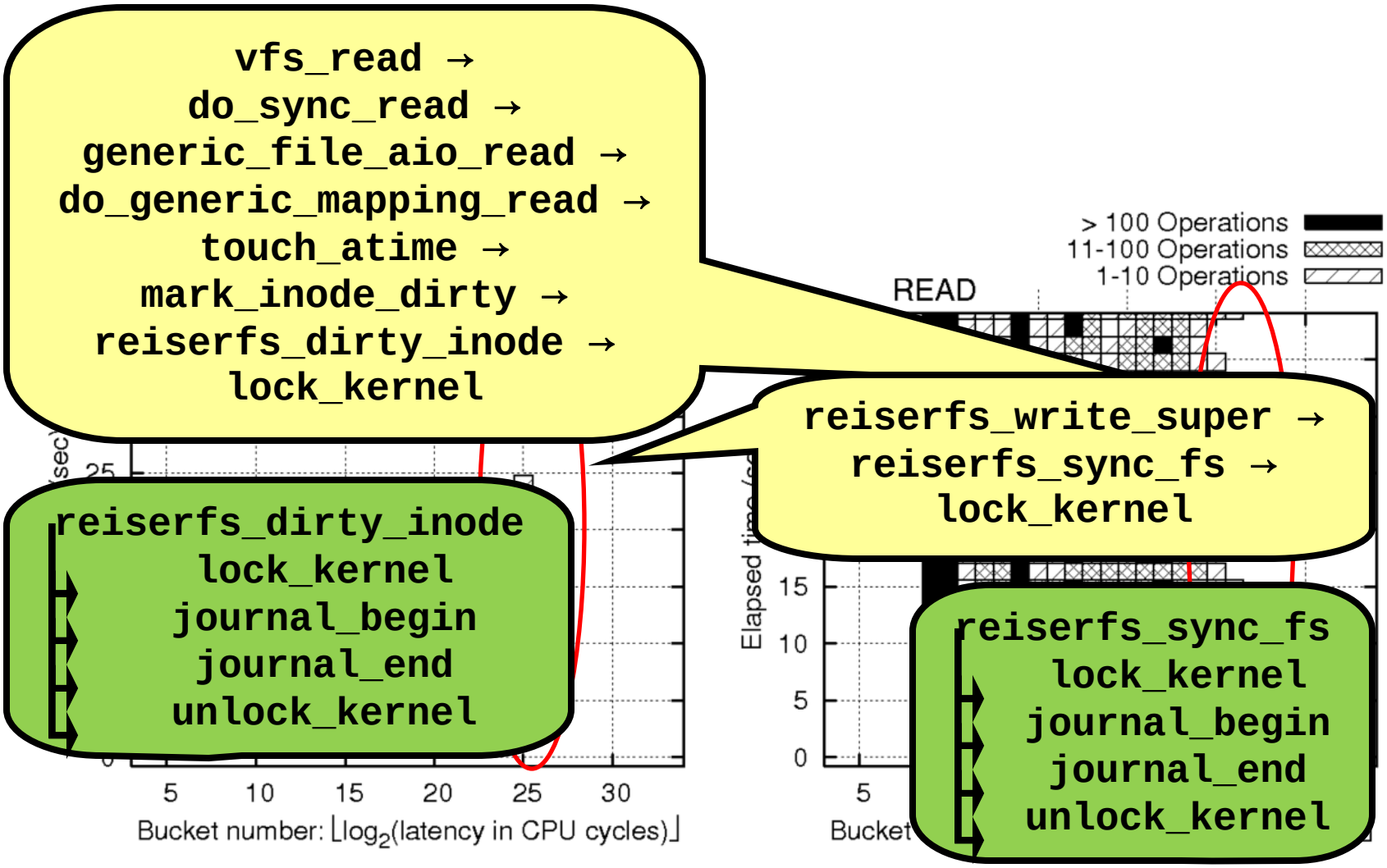
Top-level function (Read past EOF)*

```
make_request →
elv_add_request →
elv_insert →
cfq_insert_request →
blk_start_queueing →
scsi_request_fn
```



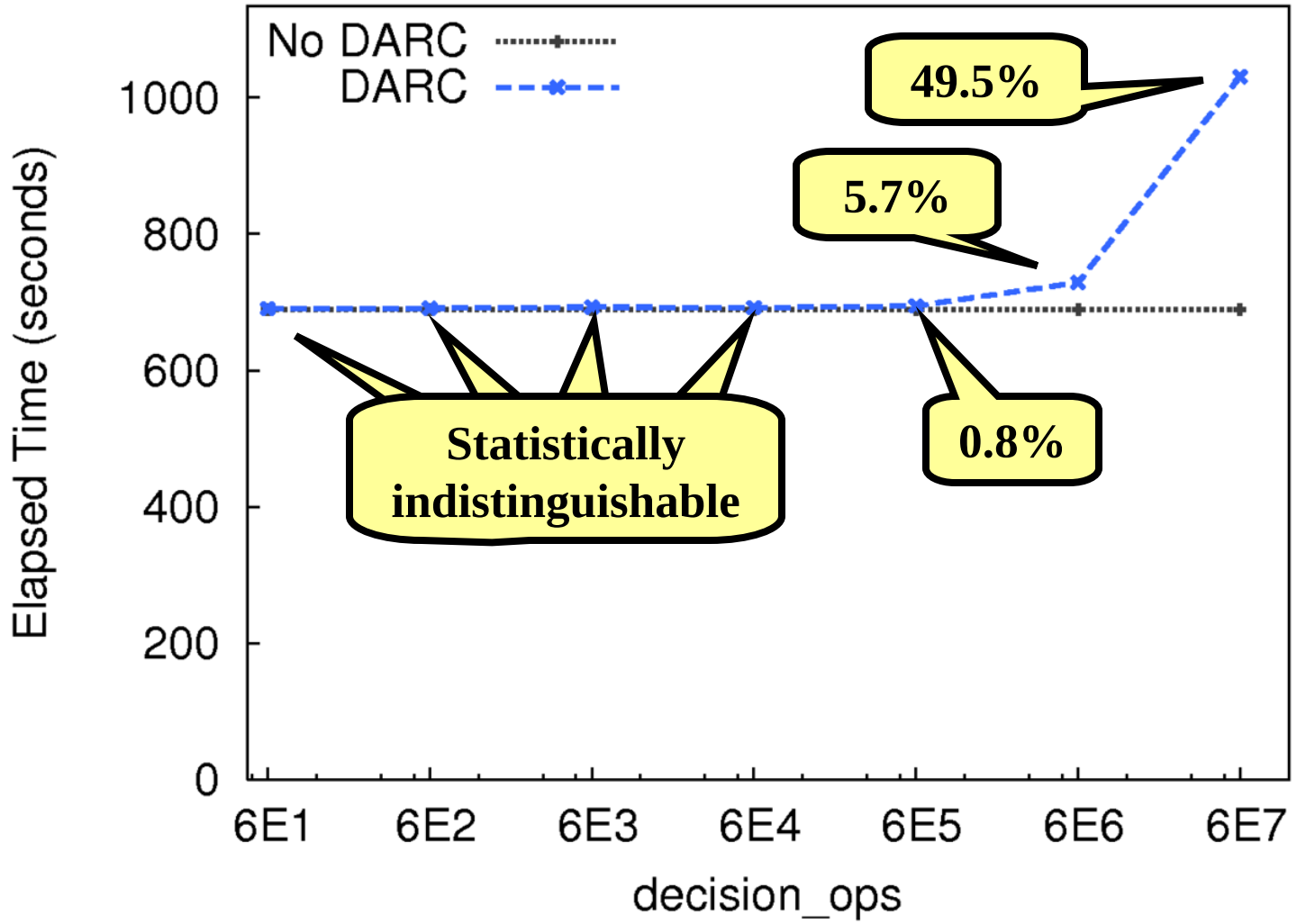
Running `grep` on the Linux kernel source files (ext2)

Analyzing Intermittent Behavior

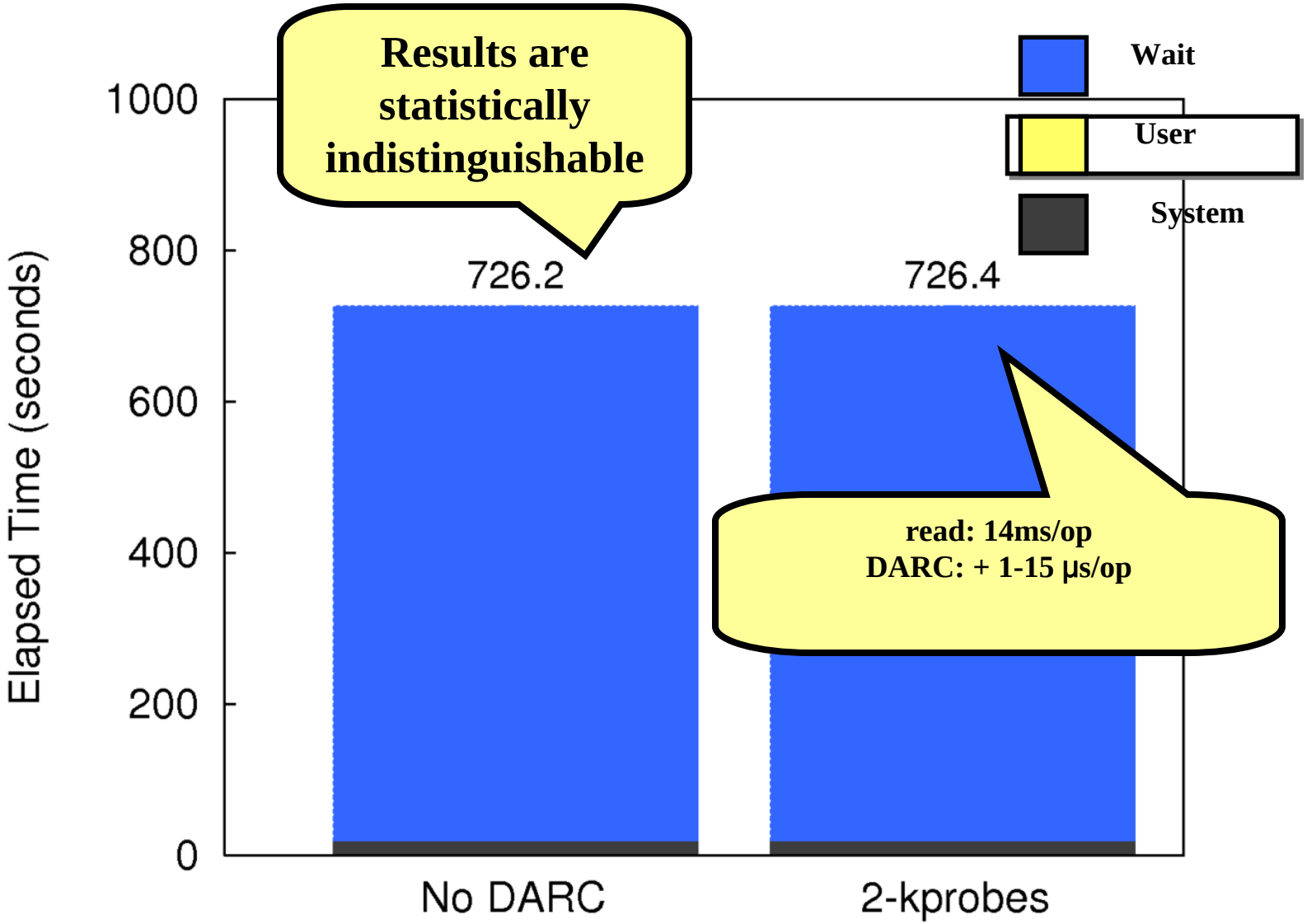


Running `grep` on the Linux kernel source files (Reiserfs 3.6)

Stat Results: Varying decision_ops



Random Read Results



Conclusions

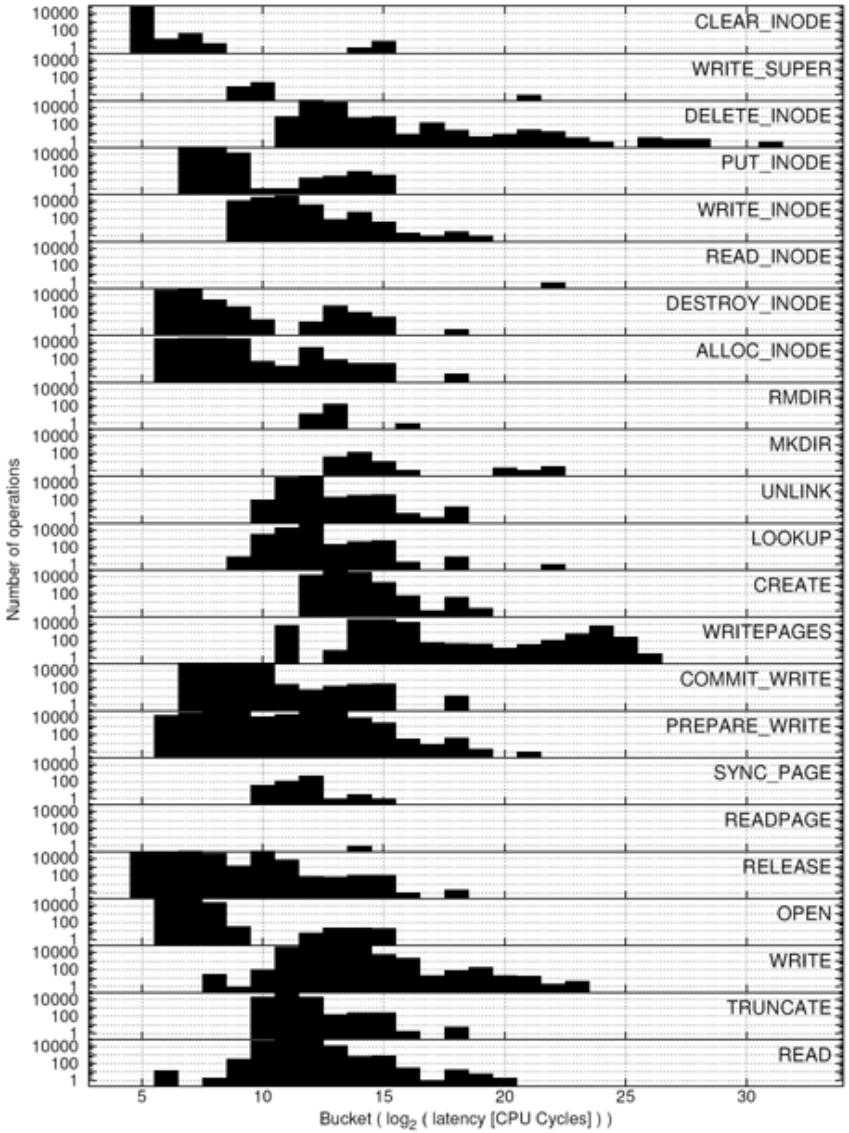
- A new performance-analysis method that finds causes of high-level behaviors
- Versatile
  - PID and call-path filtering
  - Recursion and indirect calls
  - Resume searches
  - Preemptive behavior
  - Asynchronous paths
  - Etc.
- Negligible overheads

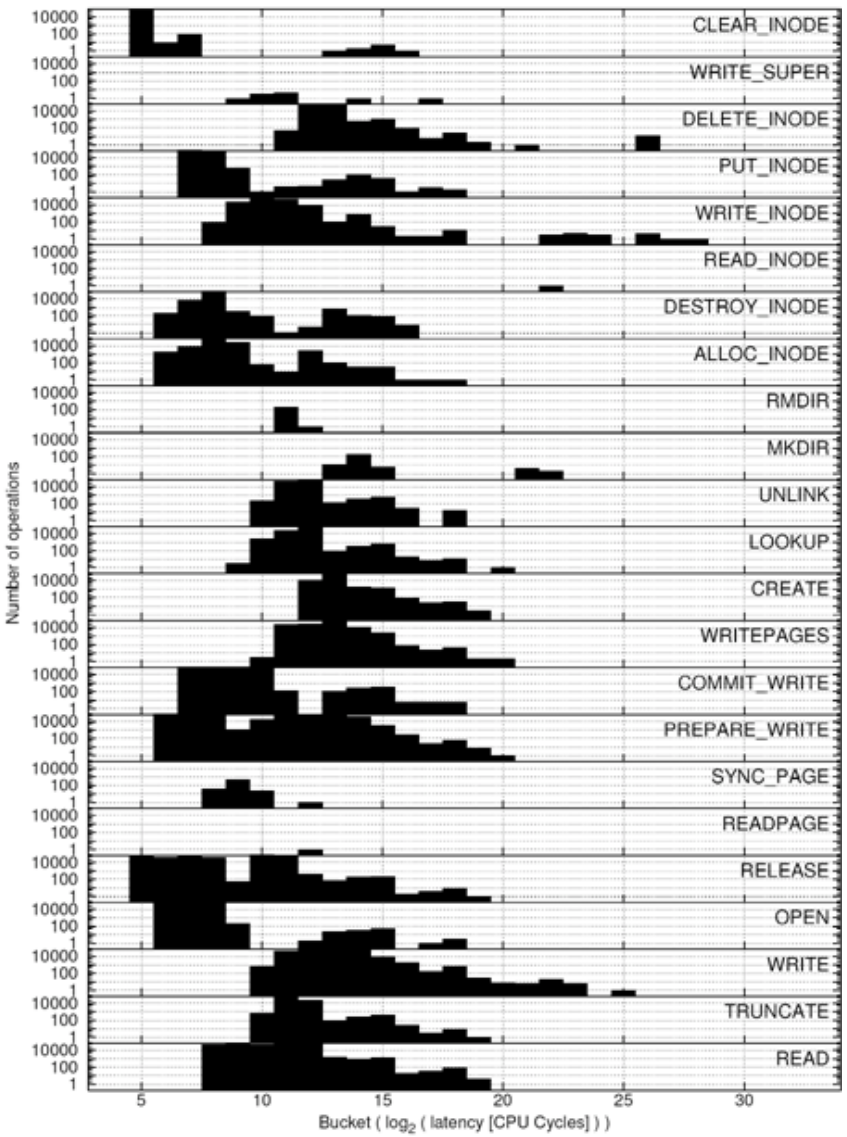
➔ Provides more concrete evidence while requiring less time, expertise, and intuition

Outline

- Analyzing Root Causes of Latency Distributions
- **Visual Analytics**

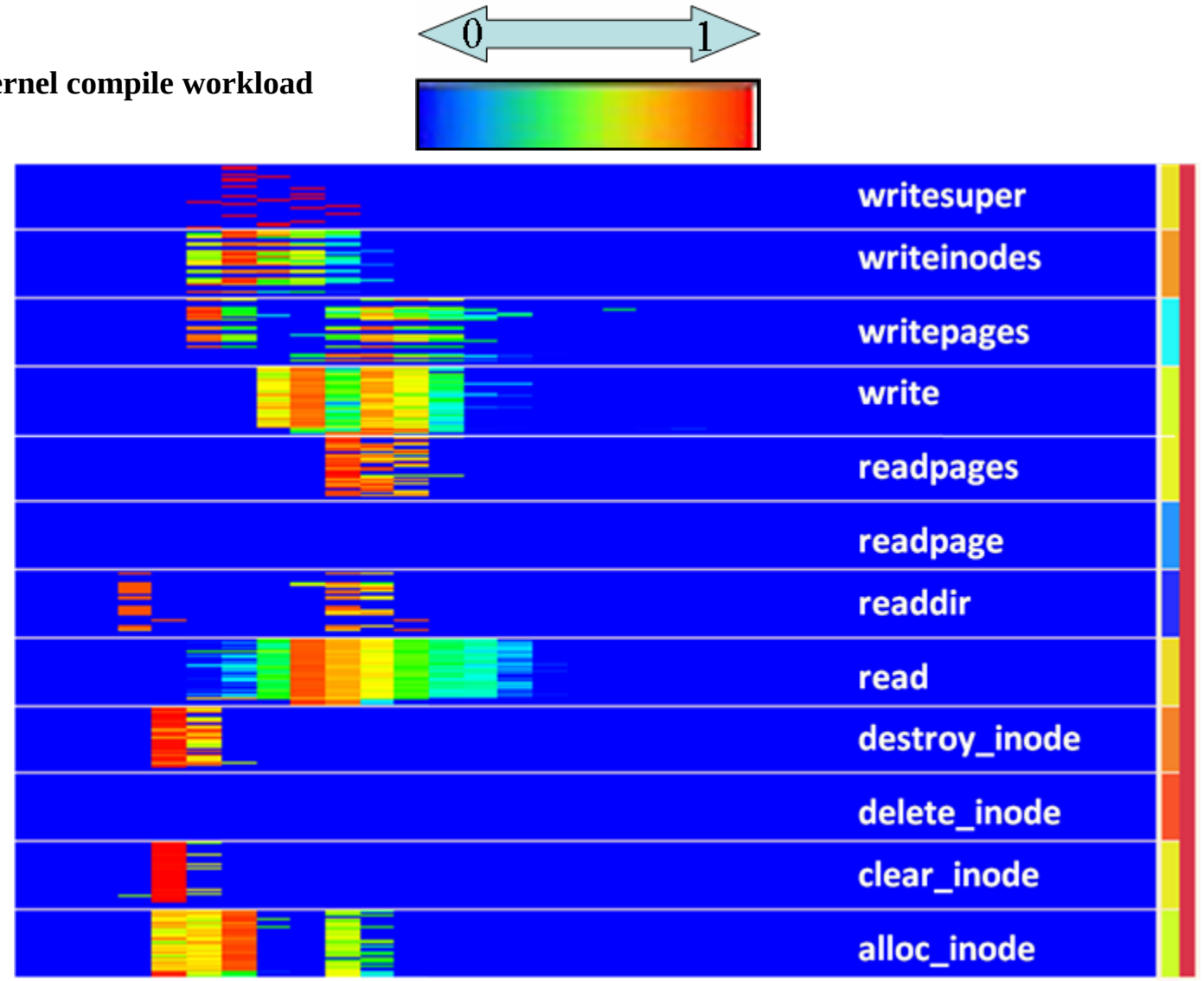
Lots of Profiles…





N-D visualization of 12 ops over time

Linux kernel compile workload



(Dis)similarity map

First Filesystem & Storage Benchmarking Workshop

- Held in May 2008 at UCSC/SSRC
- Goal: discuss status of benchmarking in the field
  - Educate the community, learn
  - Produce a report
- Presentations from industry, academia, tool developers, standards, traces, statistics, …

http://fsbench.filesystems.org/

Q&A

File System Tracing, Replaying, Profiling, and Analysis on HEC Systems

http://www.fsl.cs.sunysb.edu/hpcvperf/

Erez Zadok

Klaus Mueller

Stony Brook University

www.cs.stonybrook.edu

Ethan L. Miller

UC Santa Cruz

www.cs.ucsc.edu