Scaling OS Storage Stack Performance Using NVRAM Technologies

Erez Zadok

Dept. of Computer Science

Stony Brook University

http://www.cs.stonybrook.edu/~ezk

STONY BROOK
STATE UNIVERSITY OF NEW YORK

50+ Years of Hard Disks

- Magnetic spinning media
- Affect software algorithms
  - Data structures, optimizations
  - Minimize head movements & rotational latencies

Modern NVRAM Devices

- NVRAM devices are different
  - Erase cycles, TRIM, LBA indirection
  - Different read/write latencies
  - Costs: $$/GB, energy
- Dozens of new NVRAM devices proposed in coming years
- HDD industry isn't giving up
  - Shingled Magnetic Recording (SMR) are coming

50 Years of Software Development

- Increased storage stack complexity
  - More software layers (RAID, VM, …)
  - App-to-storage access patterns randomized
- Old optimizations no longer hold
- Storage hybrids and multi-tier storage for decades to come
- Even more storage stack complexity!

Software Complexity Impact

- Hard
  - Properly measure performance and energy
- Harder
  - Analyze and understand performance and energy trends
- Hardest
  - Replace existing storage stacks
  - Control both performance and energy

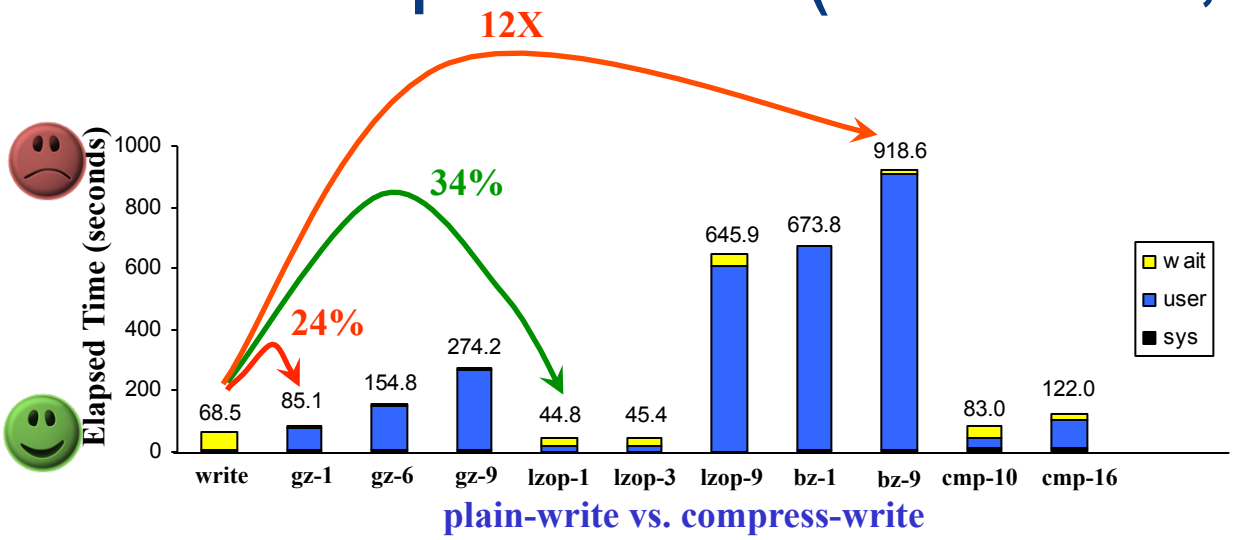
Outline

- ~~Motivation~~
- **Software is wasteful**
- Understanding where the waste is
- Rewriting storage software
- Controlling complexity
- Conclusions

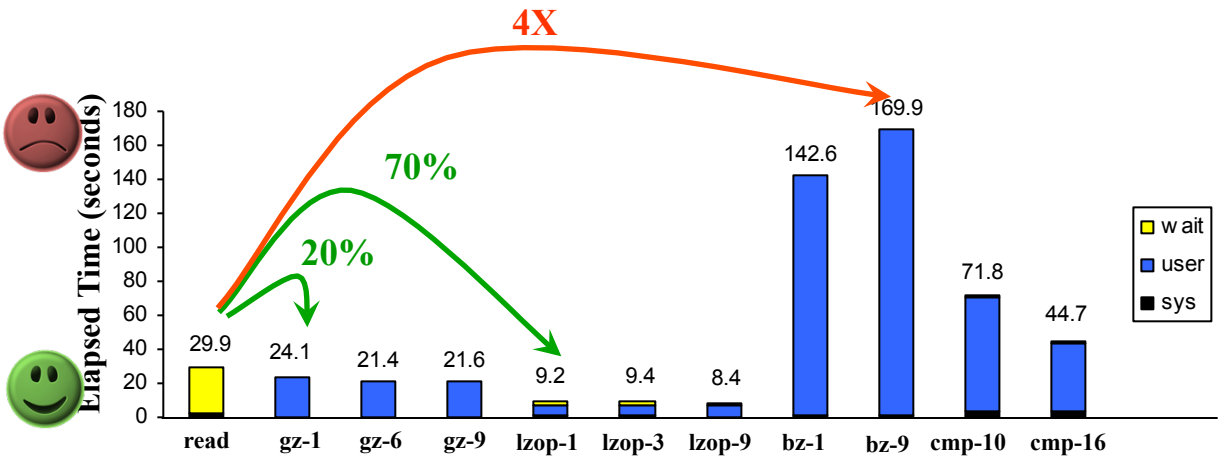
Compression Study

- Can compression help?
  - Spend CPU to compress
  - Save on I/O later on
- Comprehensive study
  - Compression algorithms
  - Different hardware
  - Different file types
  - Different storage devices

Compression (Server 1, Text File)



plain-write vs. compress-write



plain-read vs. decompress-read

lzop-1,3	∀
bzip	✗
gzip	**n reads/writes**

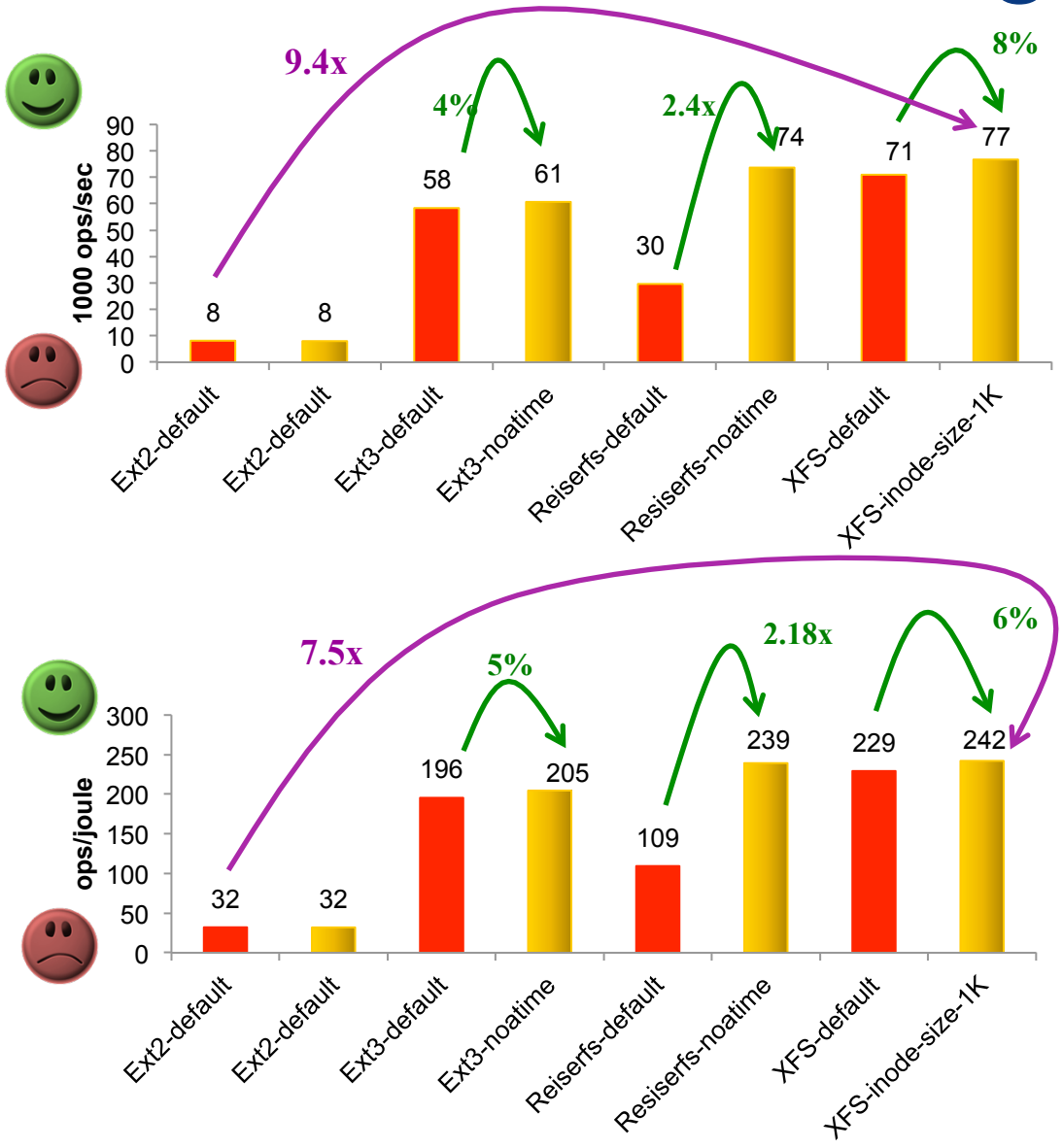
[SYSTOR 2009]

STONY BROOK STATE UNIVERSITY OF NEW YORK

Server Workload Study

- Internet servers run for years
  - Web, database, email, etc.
- Study performance and energy
  - Different workloads
  - Different file systems
    - Vary mount and format options
  - Different hardware (storage, servers)
  - Several Linux systems

Web Server Configurations



Default FS Config.

Best FS Config.

Performance

Is XFS the best for all workloads? (no)

Energy Efficiency

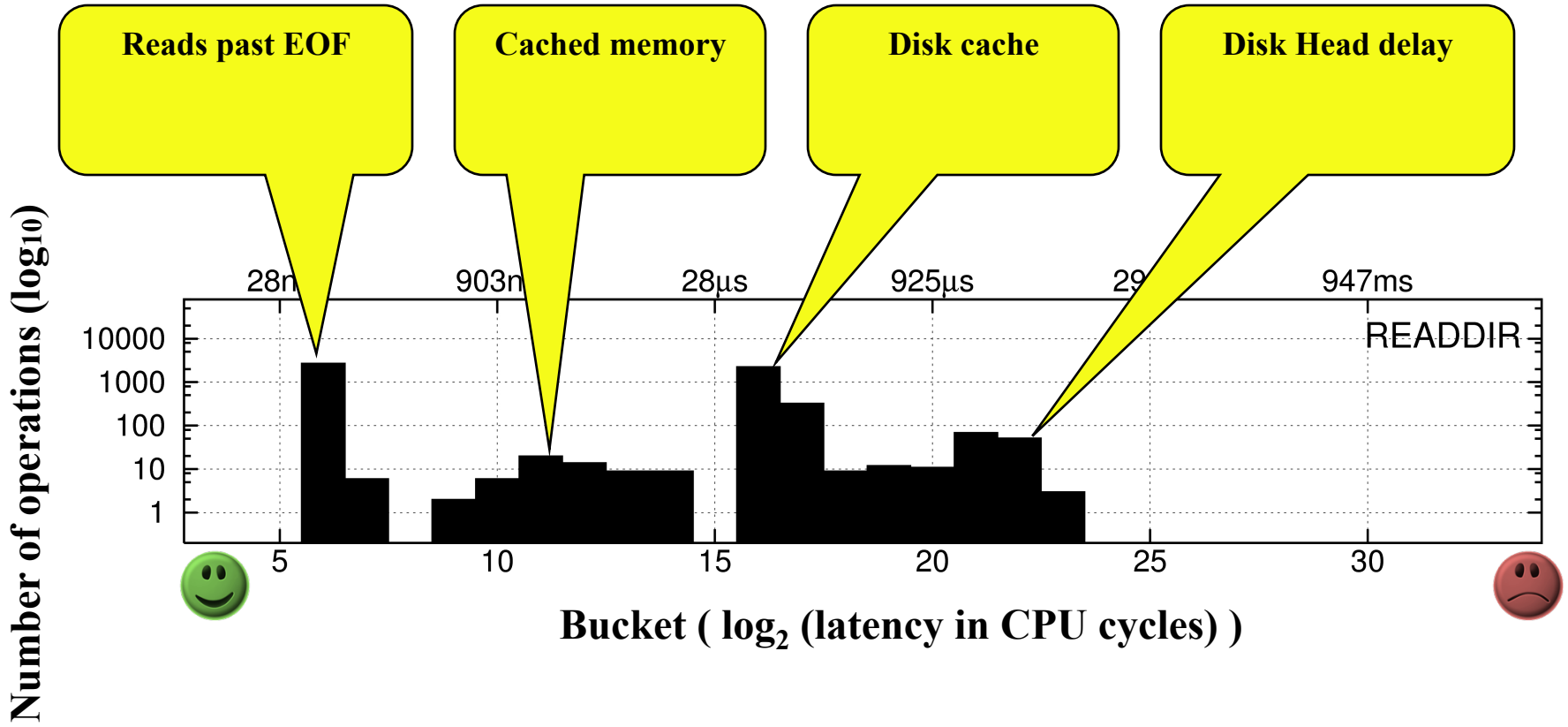
[FAST 2010, ACM TOS 2010]

STONY BROOK STATE UNIVERSITY OF NEW YORK

Outline

- ~~Motivation~~
- ~~Software is wasteful~~
- **Understanding where the waste is**
- Rewriting storage software
- Controlling complexity
- Conclusions

Multi-Modal Behavior



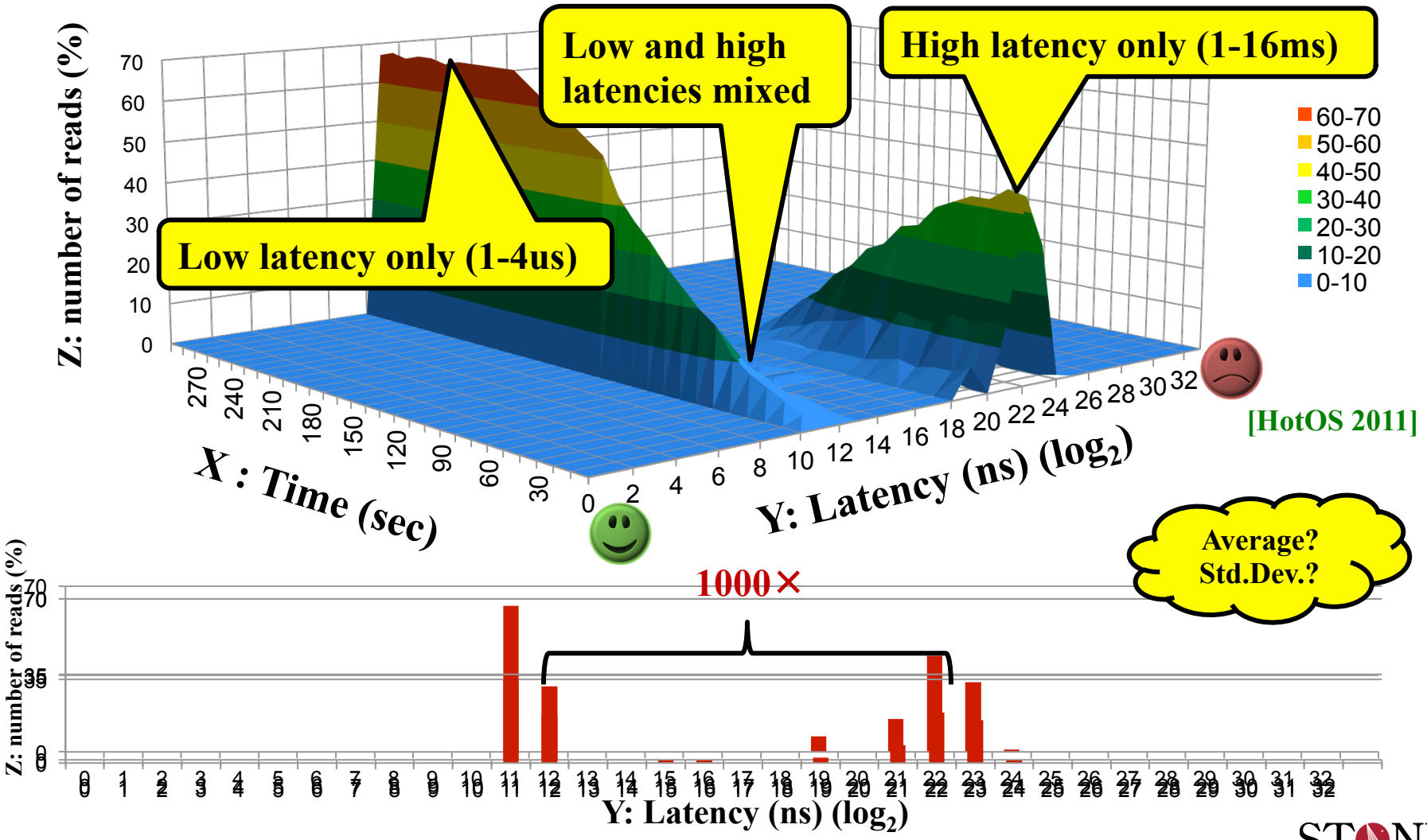
Linux 2.6.11, Ext3, `grep -r` on source tree

[OSDI 2006]

STONY BROOK STATE UNIVERSITY OF NEW YORK

Temporal Modality

Filebench 1.4.8 (modif.): Single Thread, Single File (256MB), Random Read (2KB), Ext2



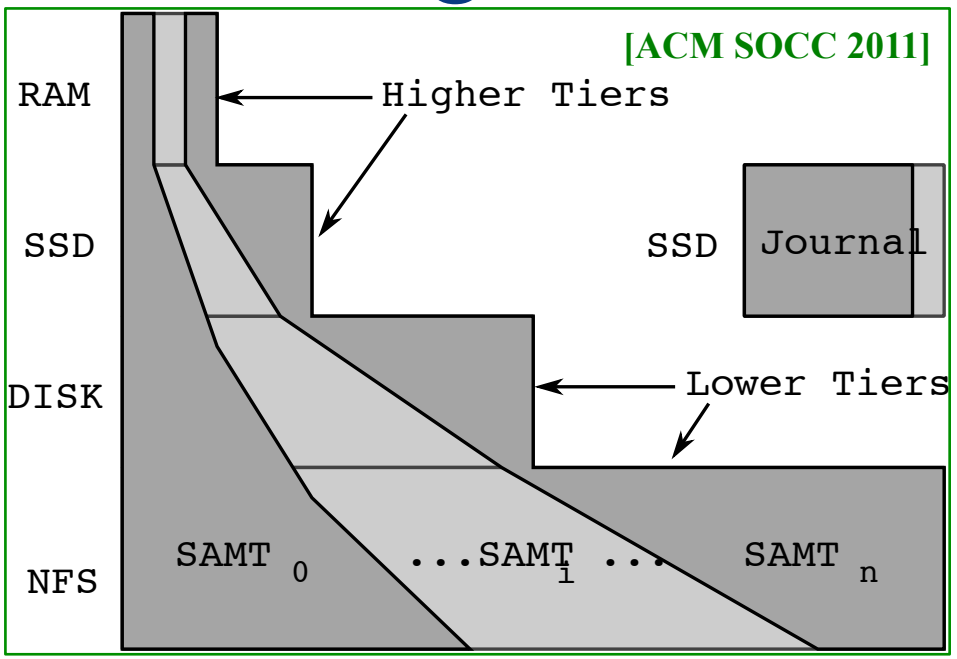
STONY BROOK
STATE UNIVERSITY OF NEW YORK

Outline

- ~~Motivation~~
- ~~Software is wasteful~~
- ~~Understanding where the waste is~~
- **Rewriting storage software**
- Controlling complexity
- Conclusions

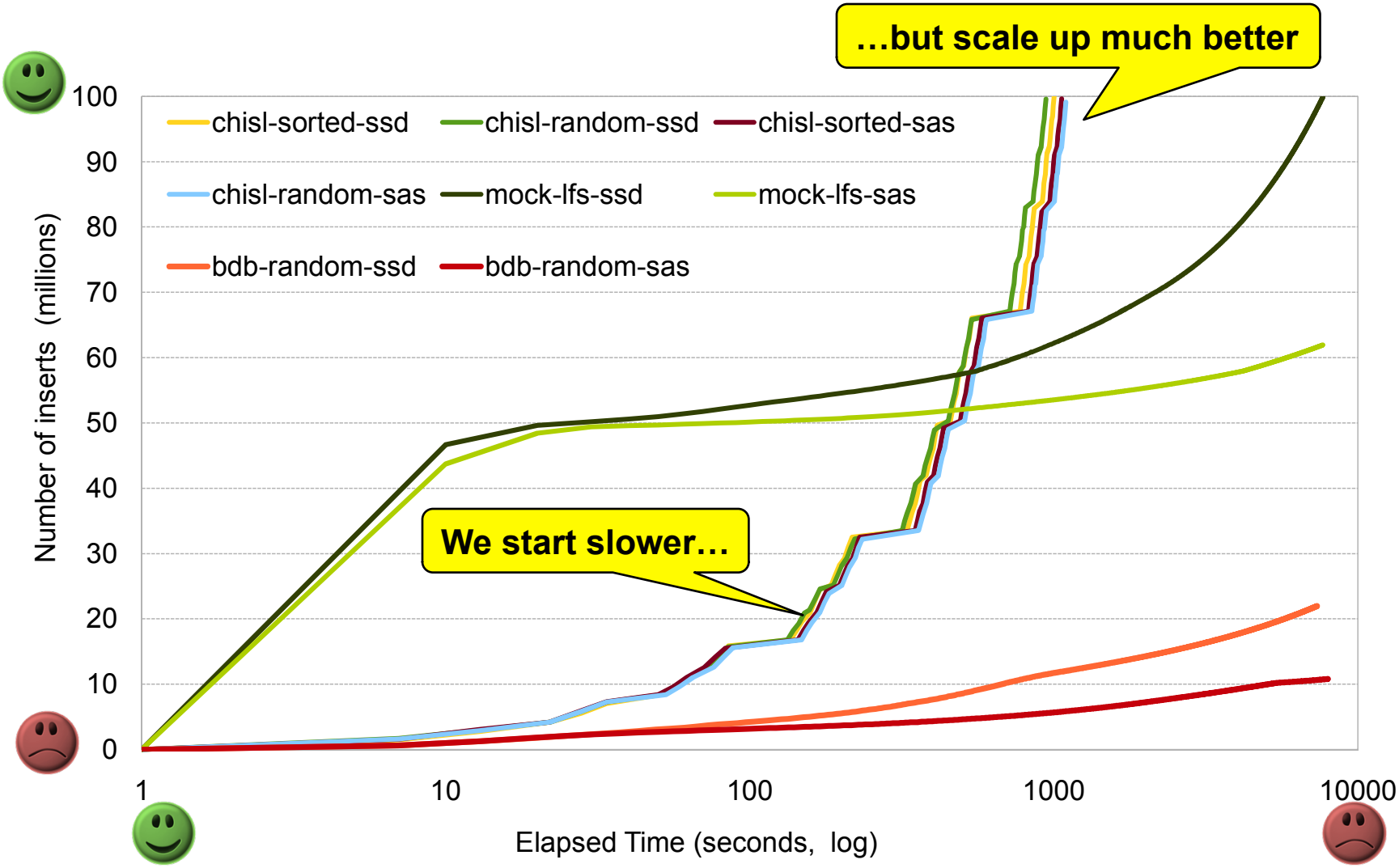
Multi-Tier Storage

- Future: heavy indexing workloads
  - Random read/writes
  - Massive data
- SAMT: Sorted Array Merge Trees
  - Can merge efficiently
- Transactional KV store
- Hot items percolate up
  - Colder ones trickle down



- Scales better than B-trees
- 10−1000x better than BDB, MySQL, XFS/Ext3, best log-structured index

Insertion Performance

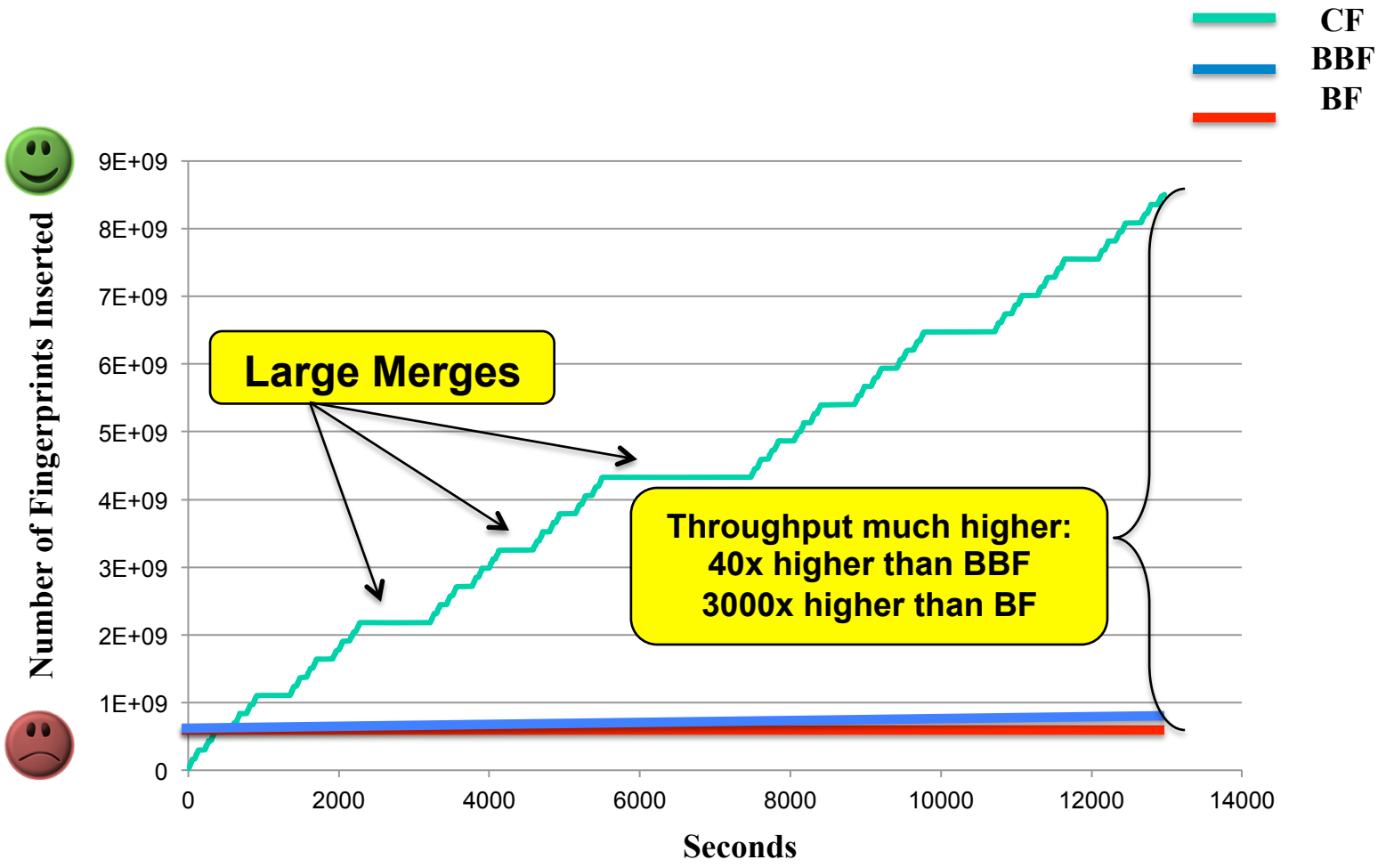


New Membership Filters

- **Cascade Filter** (CF), a Bloom Filter replacement optimized for fast inserts on Flash/SSD
- Our performance
  - We do 670,000 inserts/sec (40x of other variants)
  - We do 530 lookups/sec (1/3x of other variants)
- We use **Quotient Filters** (QF) instead of Bloom Filters
  - They have better access locality
  - You can efficiently merge two QFs into a larger QF (w/ same FP rate)
- We use **merging techniques** to compose multiple QFs into a CF

[HotStorage 2011]

Insertion Throughput



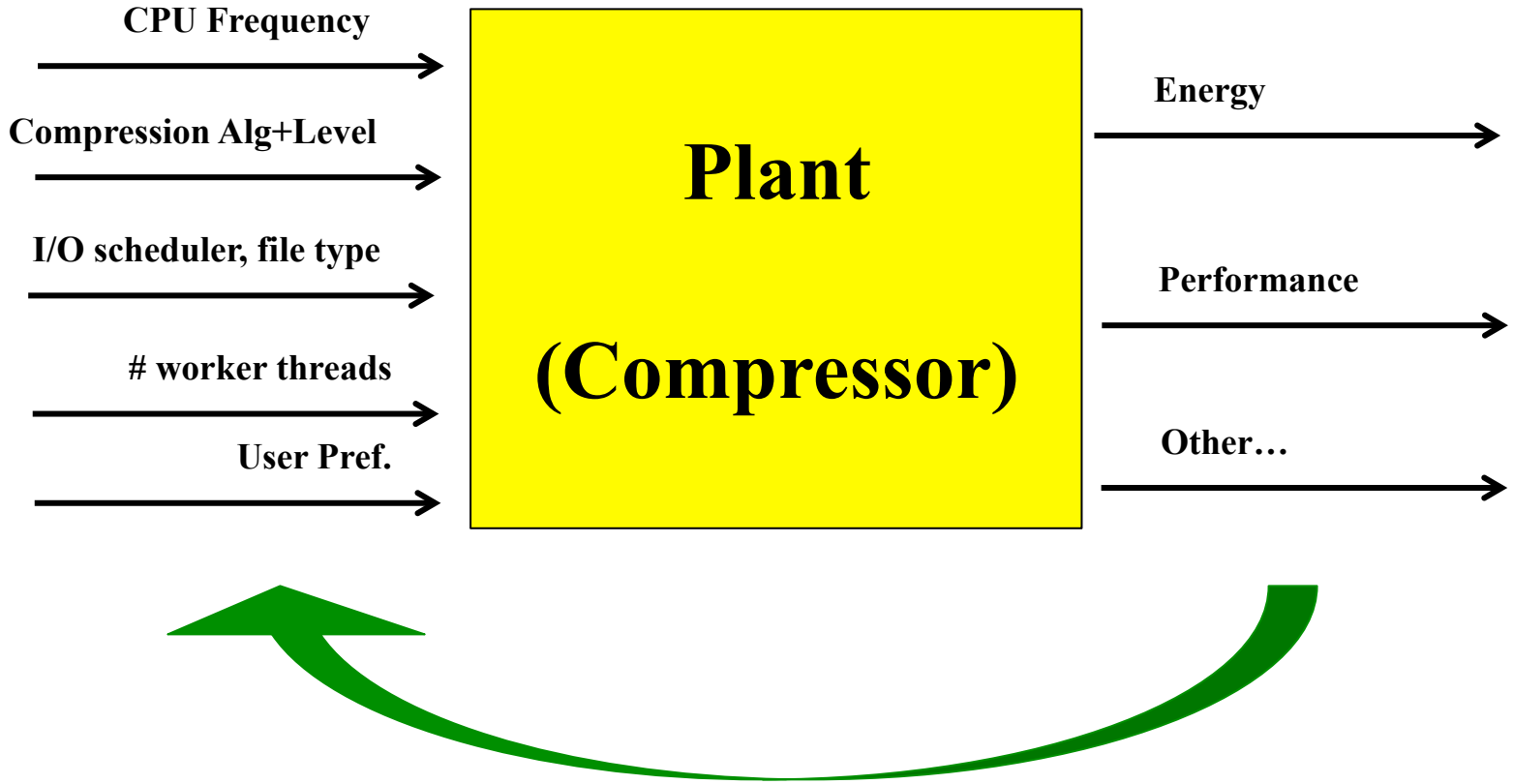
Outline

- ~~Motivation~~
- ~~Software is wasteful~~
- ~~Understanding where the waste is~~
- ~~Rewriting storage software~~
- **Controlling complexity**
- Conclusions

Intelligent Compression?

- Balance Energy, Performance, Reliability, $$$, etc.
  - User tradeoff inputs
- Use Control Theory
  - Self-adaptation
- Problems with traditional controllers
  - Assumes linear, stable behavior, low st.dev
  - Assumes single inputs/outputs
  - Assume numeric, meaningful inputs

Desired Plant Model



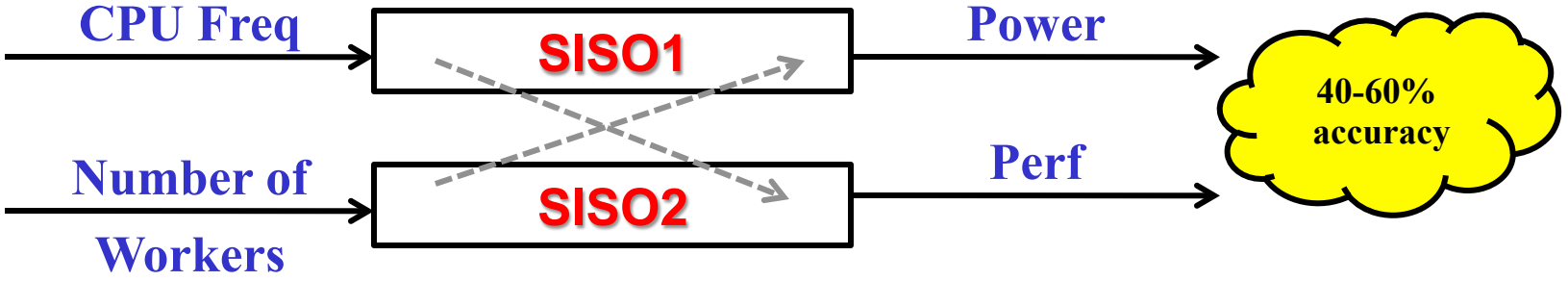
System Identification Problems

- Nonlinearity
- Instability
- Multi-dimensionality
  - CPU Frequency
  - I/O Schedulers
  - File Types
  - Disk Types (SSD vs. SATA vs. SAS)
  - Compression Algorithm + Level
- Non-numeric labels

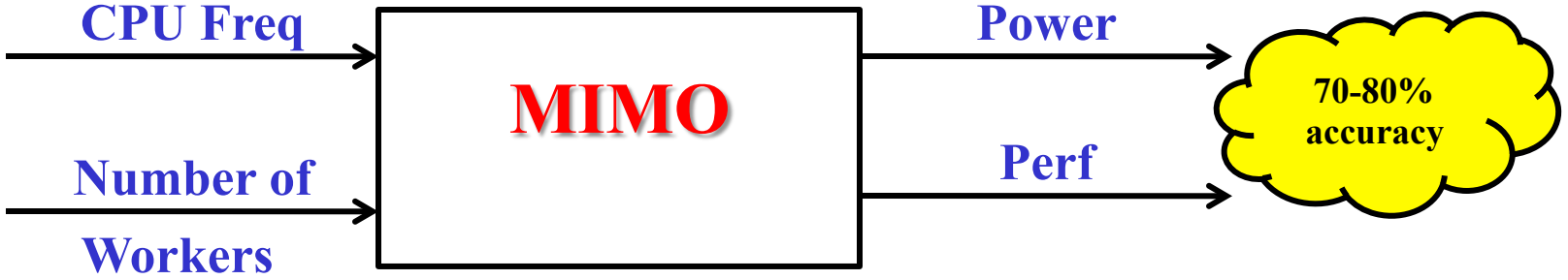
Techniques Being Investigated

- Data Mining, HMMs
- Visual Analytics
- Multi-Dimensional Scaling
- Hierarchical controllers
- Segmentation
- Multiple-Input Multiple-Output models
- …

Models



[SYSTOR '11, ERSS '11]



MIMO model and two SISO models

Outline

- ~~Motivation~~
- ~~Software is wasteful~~
- ~~Understanding where the waste is~~
- ~~Rewriting storage software~~
- ~~Controlling complexity~~
- **Conclusions**

Conclusions

- Software getting more complex
- More hardware hybrids & combos
- Software is wasteful
  - Need new algorithms & data structures
  - Techniques to control complexity
- Faster hardware alone is not enough
  - Better cooperation with software developers
  - Hinting, active storage, profiling, control

Scaling OS Storage Stack Performance Using NVRAM Technologies

Q&A

Erez Zadok

Dept. of Computer Science

Stony Brook University

http://www.cs.stonybrook.edu/~ezk

STONY BROOK
STATE UNIVERSITY OF NEW YORK