# Evaluating Performance and Energy in File System Server Workloads

Priya Sehgal, Vasily Tarasov,
and Erez Zadok

***File systems and Storage Lab***

***Dept. of Computer Science***

***Stony Brook University***

**http://green.filesystems.org/**






STONY BROOK
STATE UNIVERSITY OF NEW YORK

# Motivation

- For every $1 spent on hardware $0.50 spent on power and cooling [IDC 2007]
- Energy use in U.S. data centers = 1–2% of total energy in U.S. [EPA 2007]
- Even more outside the data center [Forrester 2008]

**Build performance- and energy-efficient systems**

**Evaluate the efficacy of file system in achieving this goal**

# Overview

- Motivation
- **Related Work**
- Experimental Methodology
- Evaluation Results
  - Machine 1 (M1) Results
  - Machine 2 (M2) Results


- Conclusion and Future Work

# Techniques

**Complimentary**

**Right Sizing** | **Work Reduction**

- CPU DVFS
- Machine ACPI states
  - standby, hibernate, off, etc.
- Opportunistic spin-down
- DRPM
- Virtualization

- Aggregation, Localization
- Compression, DeDUP
- Reconfiguration
  - Application/Services
  - File Systems
  - RAID Levels, etc.

# Related Work - 1

- Right Sizing
  - Redirect the request elsewhere
  - PDC, MAID, GreenFS, Write-offloading, EAVFS, etc.
- Work Reduction
  - Improve locality
  - FS2, EEFS, Predictive Data Grouping, etc.
- Others
  - FAWN
  - quFiles, etc.

# Related Work - 2

- Benchmarks
  - SPECPower
    - Metric: operations/second/watt
  - JouleSort
    - Metric: sortedrecs/joule
- Benchmark Studies
  - Compression evaluation [Kothiyal 2009]
  - RAID evaluation [Gurumurthi 2003]

# Overview

- Motivation
- Related Work
- **Experimental Methodology**
- Evaluation Results
- Conclusion and Future Work

# Experimental Methodology


- Workloads (4)
  - Web server, Database server, File server, Mail server
  - FileBench emulated workloads
- File Systems (4)
  - Type: Ext2, Ext3, ReiserFS, XFS
  - Mount Options: **noatime**, **notail**, **journal=<*modes*>**
  - Format Options: inode size, blocksize, allocation/block group count.
- Hardware (2)

**We ran a total of 248 benchmarks → 414 clock hours!**

# FileBench

- Sun Microsystems, 2005
  - Used for performance analysis of Solaris OS
  - Other studies: [Macko '10, Zhang '10, Gulati '10], etc.
- Rich language to emulate complex workloads
- Provide with a few emulated workloads
  - Application traces
  - Recommend parameters for server workloads
- Superior to few other benchmarks
  - E.g., Bonnie, Postmark, Andrew Benchmark, etc.

**We ported FileBench to different platforms (FreeBSD, Linux)**

# FileBench Workloads

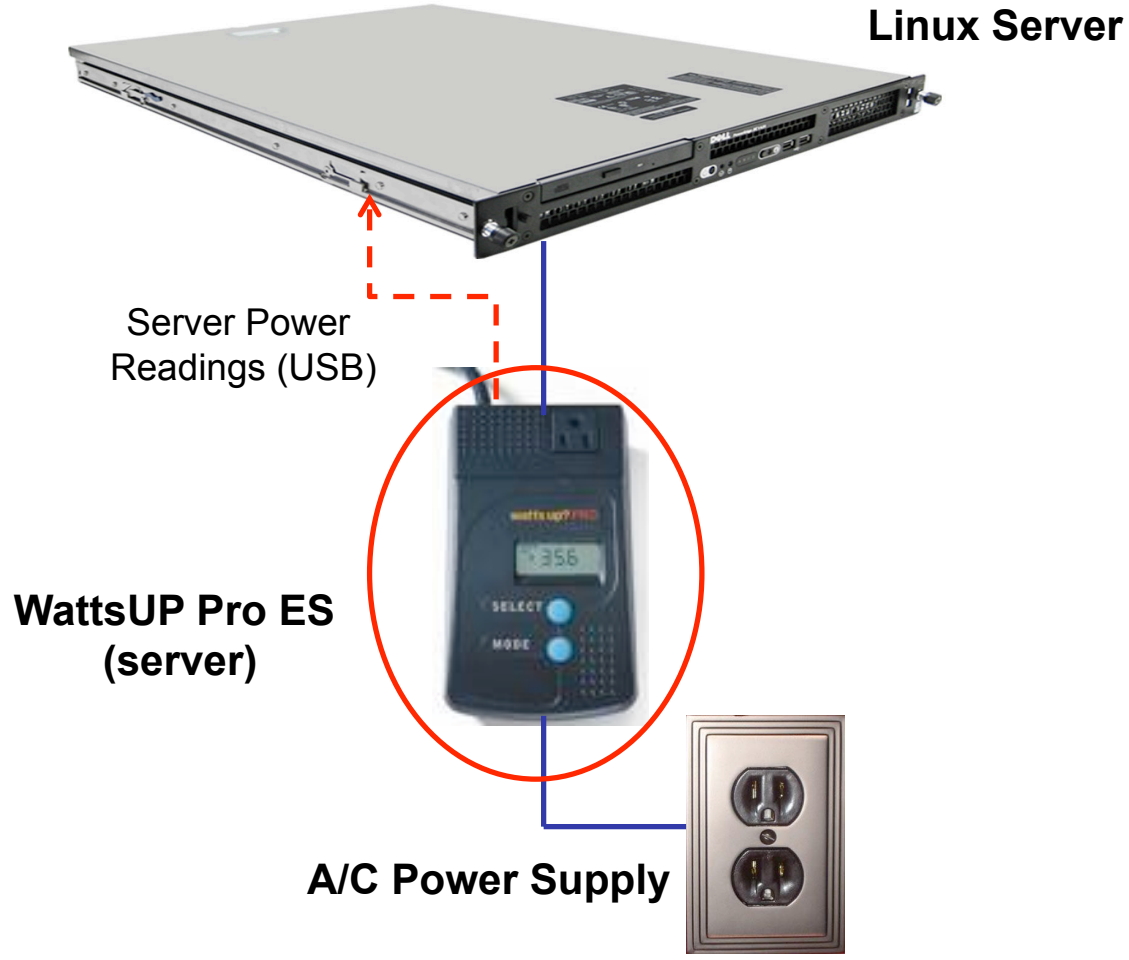
| Server workload | Avg. file size | Avg. directory depth | No. of files | I/O size (R/W) | No. of threads | R/W ratio |
|---|---|---|---|---|---|---|
| **Web** | 32KB | 3.3 | 20,000 | 1MB/16KB | 100 | 10:1 |
| **File** | 256KB | 3.6 | 50,000 | 1MB/16KB | 100 | 1:2 |
| **Mail** | 16KB | FLAT | 50,000 | 1MB/16KB | 100 | 1:1 |
| **Database** | 0.5GB | FLAT | 10 | 2KB/2KB | 200+10 | 20:1 |

# File System Properties

| Features | Ext2 | Ext3 | ReiserFS | XFS |
|---|---|---|---|---|
| **Disk Layout** | Linear | Linear | S+ Tree | B+ Tree |
| **Allocation unit / strategy** | Fixed-sized blocks | Fixed-sized blocks | Fixed-sized blocks | Variable-sized extents (Delayed allocation) |
| **No. of Files** | Fixed | Fixed | Variable | Variable |
| **Journaling modes** | None | Ordered, writeback, data | Ordered, writeback, data, none | Writeback |
| **Special Feature** | Block groups | Block groups | Tail Packing | Allocation groups |

**We used CentOS 5.3 Linux 2.6.18-128.1.16.el5.centos.plus**

# Hardware Setup



# Machine Configurations

| | M1 (Reported in paper) | M2 |
|---|---|---|
| **Machine Age** | 3 years | < 1 year |
| **CPU Model** | Intel Xeon | Intel Nehalem (E5530) |
| **CPU Speed** | 2.8GHz | 2.4GHz |
| **# of CPUs** | 2 dual core | 1 quad core |
| **DVFS** | No | Yes |
| **L1 cache size** | 16KB | 128KB |
| **L2 cache size** | 2MB | 1MB |
| **L3 cache size** | No | 8MB |
| **FSB speed** | 800 MHz | 1066 MHz |
| **RAM size** | 2048 MB | 24GB **(used 2GB)** |
| **RAM type** | DIMM | DIMM |
| **Disk RPM** | 15K RPM | 7.2K RPM |
| **Type of Disk** | SCSI | SATA |
| **Average Seek Time (ms)** | 3.2/3.6 ms | 10.5/12.5 ms |
| **Disk Cache** | 8MB | 16MB |

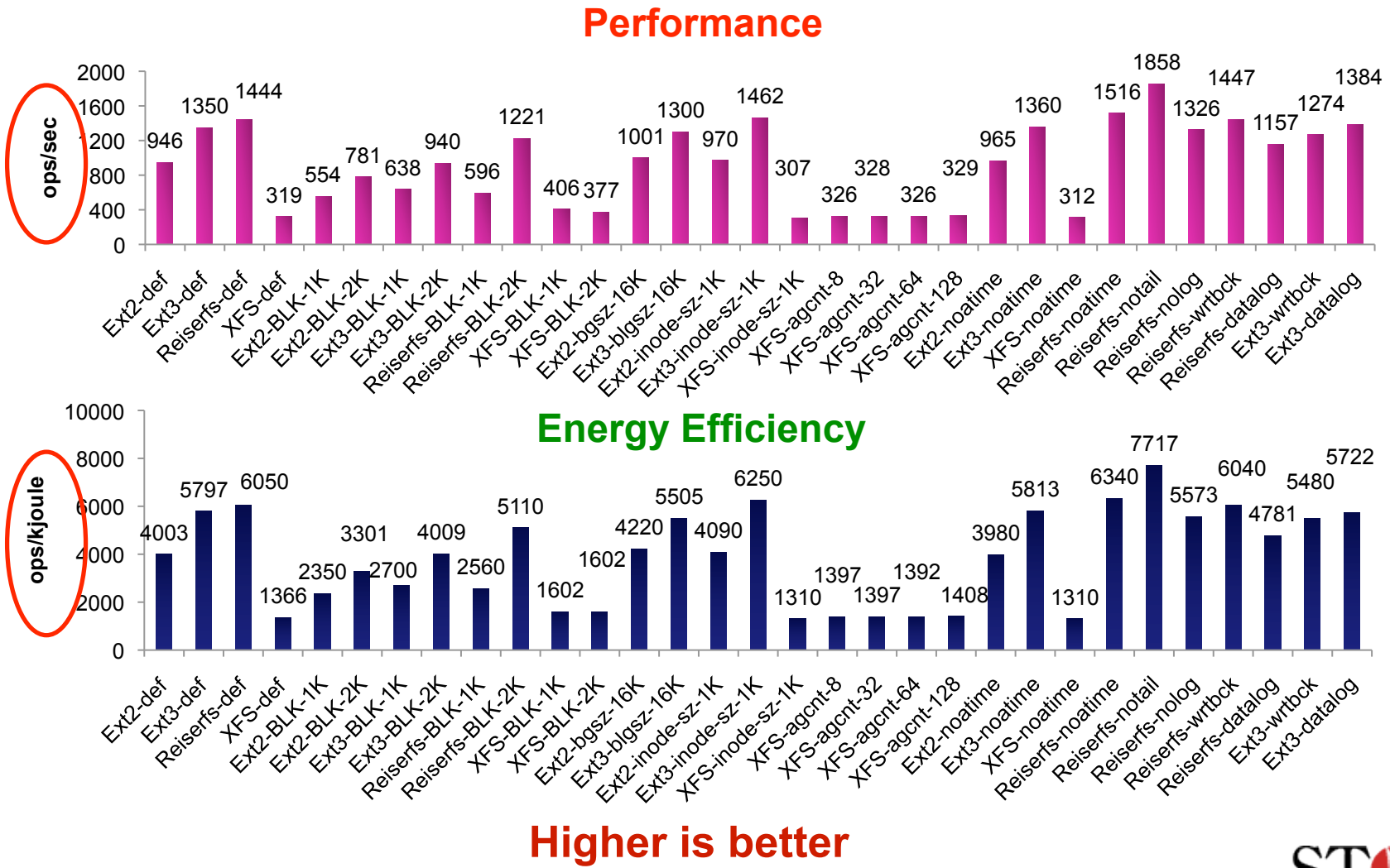
# Overview

- Motivation
- Related Work
- Experimental Methodology
- **Evaluation Results**
  - **Machine 1 (M1) Results**
  - Machine 2 (M2) Results
- Conclusion and Future Work




STONY BROOK STATE UNIVERSITY OF NEW YORK

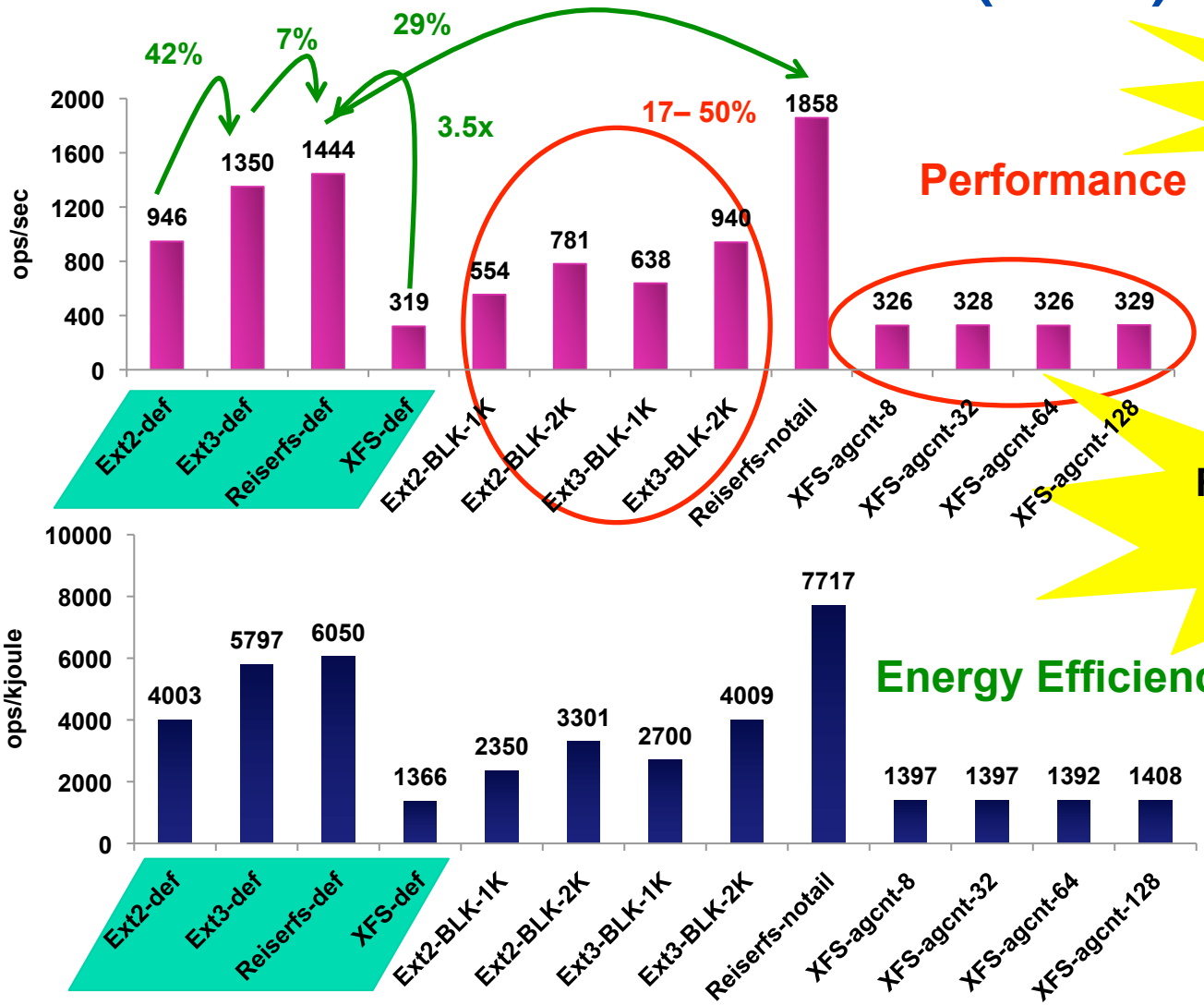
# Mail Server (M1)






STONY BROOK STATE UNIVERSITY OF NEW YORK

# Mail Server (M1)




Tail packing on by default

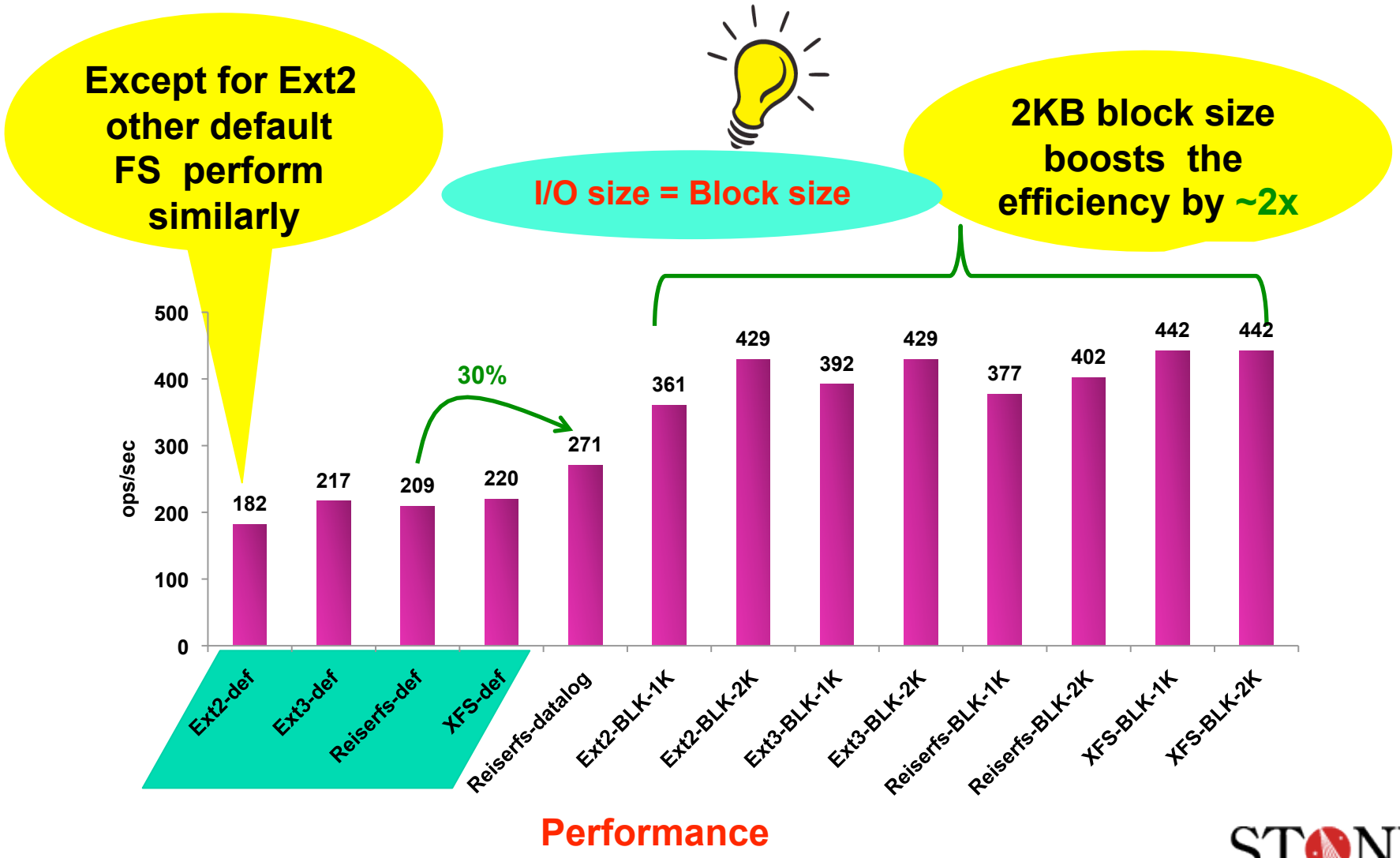
ReiserFS-notail best for this configuration

**Linearity between Performance and Energy Efficiency**



STONY BROOK STATE UNIVERSITY OF NEW YORK

# Database Server (M1)






STONY BROOK STATE UNIVERSITY OF NEW YORK

# File System Selection Matrix (M1)

- Newer hardware → Different results

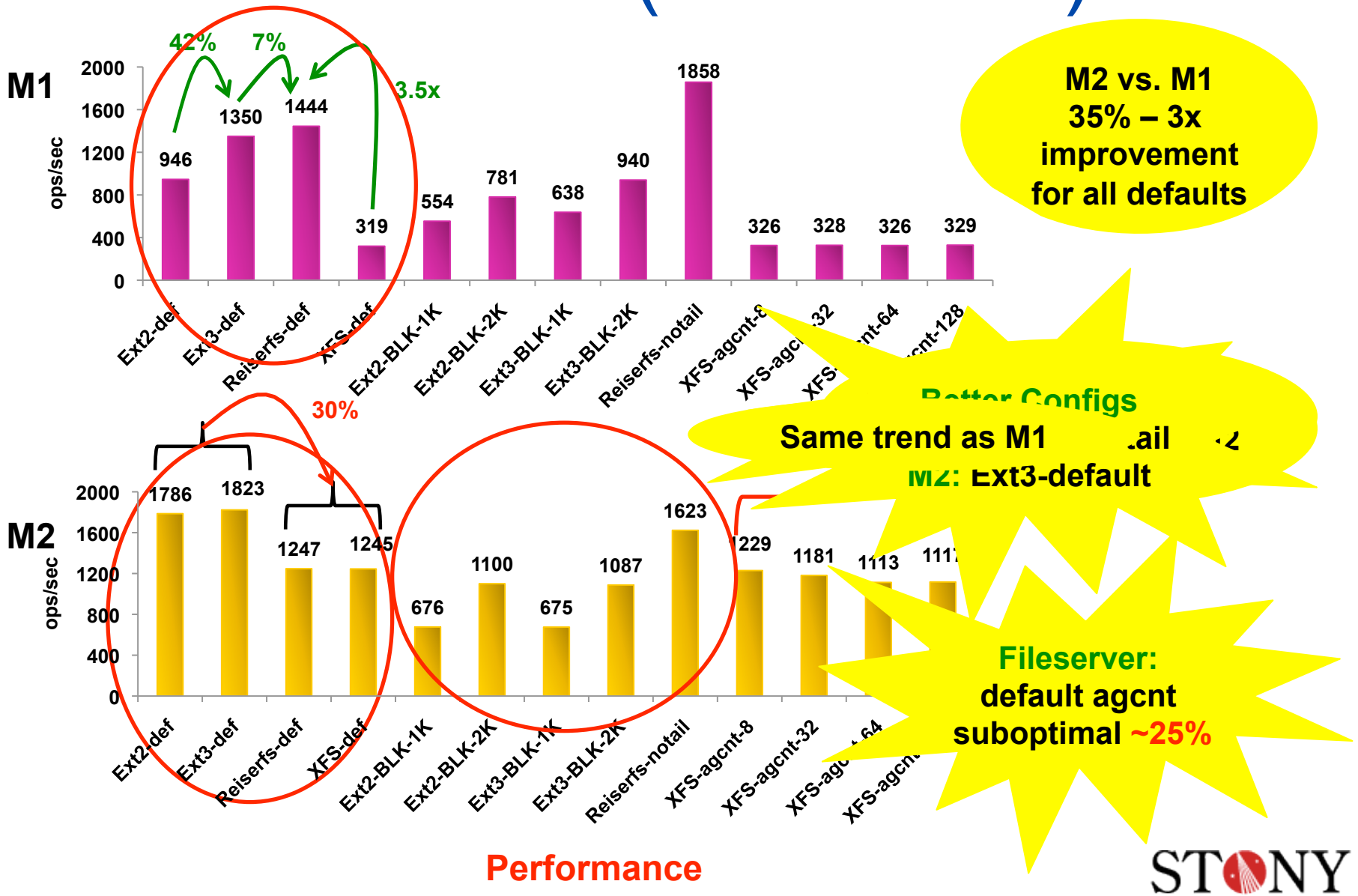
| Workload | Best File System (Combination) | Improvement Range (compared to all default FS) | |
|---|---|---|---|
| | | Ops/sec | Ops/joule |
| Web Server | XFS (inode-size-1K) | 8% – 9.4x | 6% – 7.5x |
| File Server | ReiserFS (default) | 0% – 1.9x | 0% – 2.0x |
| Mail Server | ReiserFS (notail) | 29% – 5.8X | 28% – 5.7x |
| Database Server | XFS/Ext3 (BLK-2K) | 2.0 – 2.4x | 2.0 – 2.4x |

**This recommendation matters but …**

# Overview

- Motivation
- Related Work
- Experimental Methodology
- **Evaluation Results**
  - **Machine 1 (M1) Results**
  - **Machine 2 (M2) Results**
- Conclusion and Future Work

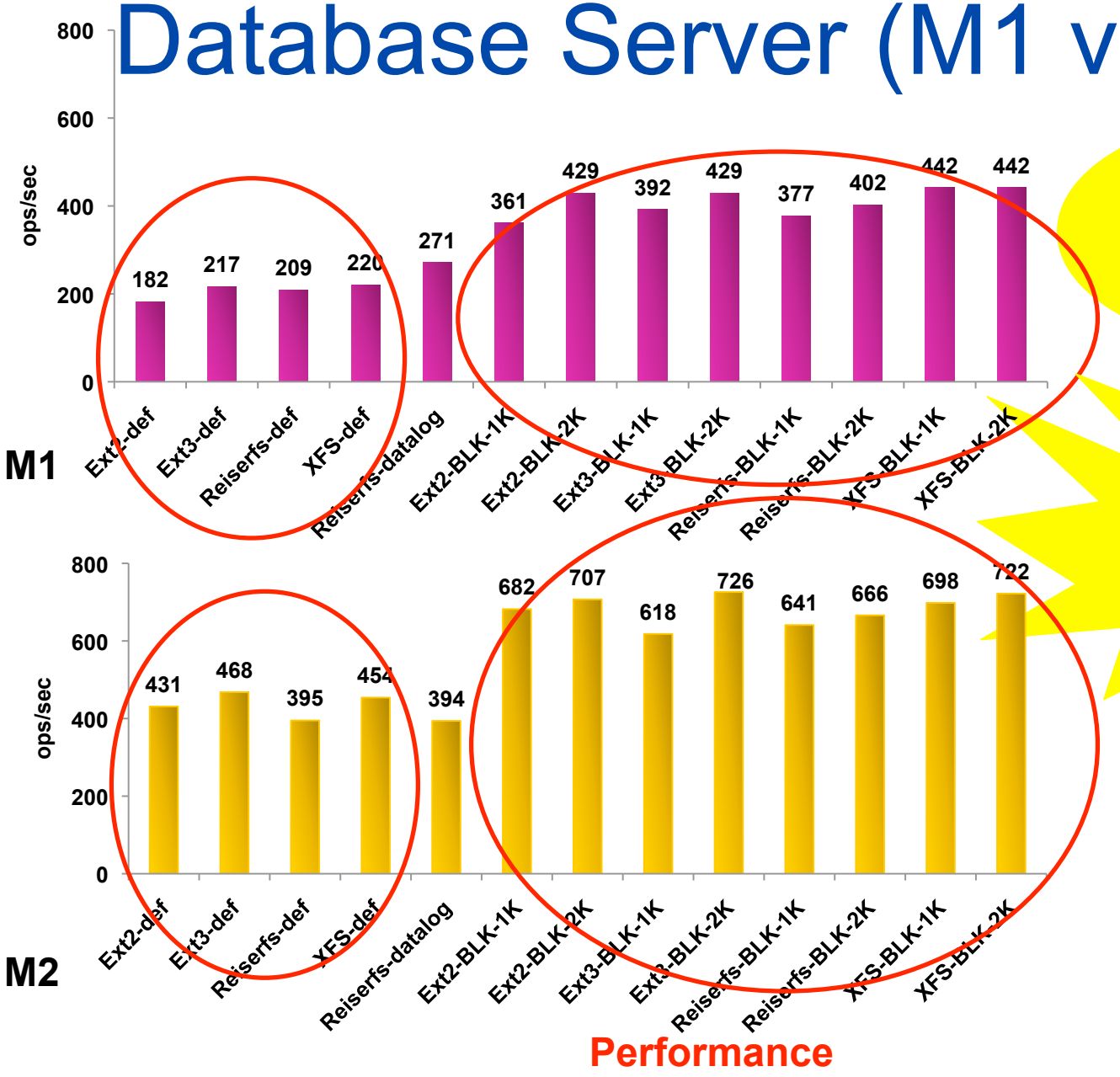
# Mail Server (M1 vs. M2)






STONY BROOK
STATE UNIVERSITY OF NEW YORK

# Database Server (M1 vs. M2)




**Performance trend remains the same across M1 and M2**

**2K block size increases performance by ~1.5x**



STONY BROOK
STATE UNIVERSITY OF NEW YORK

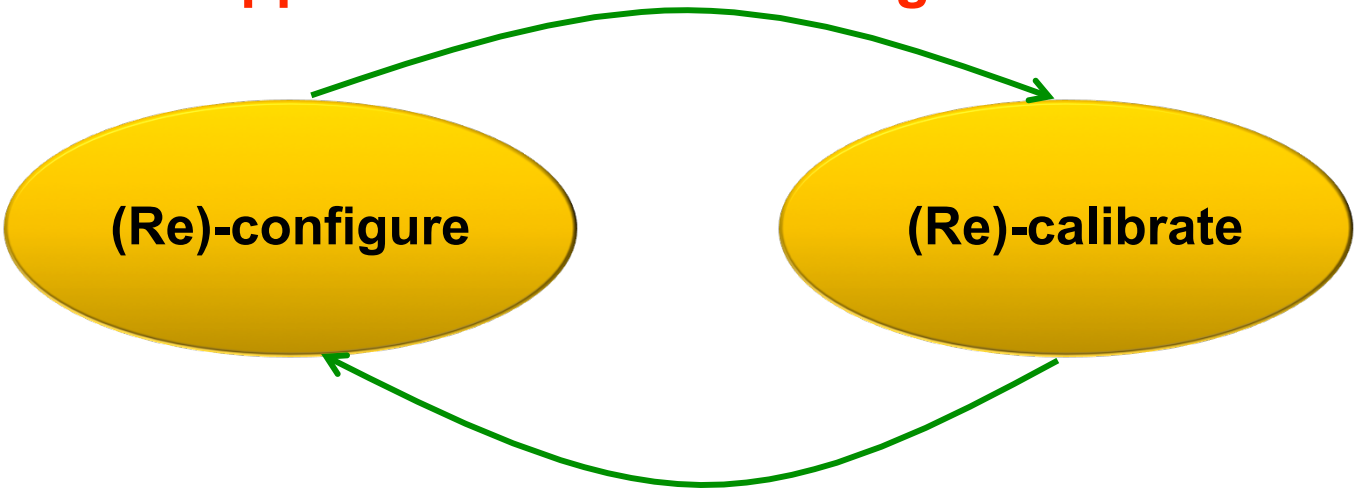
# Overview

- Motivation
- Related Work
- Experimental Methodology
- Evaluation Results
- **Conclusion and Future Work**

# Conclusions

- Workloads drive performance-energy
  - Depend also on hardware, software, config
  - Significant savings possible
- Recipe to improve work done per dollar

**Applicable to entire storage/software stack**




**It is expensive and time consuming but …**

**Small savings matter over the long run !**



STONY BROOK STATE UNIVERSITY OF NEW YORK

# Ongoing/Future Work

- Study multiple dimensions
  - New FS, Disk Scheduler, RAID, LVM, etc.
  - Client/Server Systems
    - Poster on NFSv4 at Poster Session in FAST 2010
  - Disk Types: SAS, SSD, etc.
  - Cluster Storage, SANs, OS
- Develop auto-configuration tools
- Develop workload specific storage stack

# Evaluating Performance and Energy in File System Server Workloads

Priya Sehgal, Vasily Tarasov, and Erez Zadok

# Q&A

***File systems and Storage Lab***

***Dept. of Computer Science***

***Stony Brook University***

**http://green.filesystems.org/**