Server-Class Energy and Performance Evaluations

Erez Zadok

ezk@fsl.cs.sunysb.edu

File systems and Storage Lab

Stony Brook University

http://green.filesystems.org/

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]
  - Growth Rate of 2x per 5 years
- 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

Reduce P_{idle}

Reduce $P_{dynamic}$

Complementary

Right Sizing ⟷ **Work Reduction**

Right Sizing

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

Work Reduction

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

Right Sizing Techniques

- Techniques to increase disk sleep time
  - Massive Array of Idle disks (MAID) [Colarelli 2002]
  - Popular Data Concentration (PDC) [Pinheiro 2004]
  - Write off-loading [Narayanan 2008]
  - GreenFS [Joukov 2008]
  - Scale down Hadoop clusters [Leverich 2009]

Work Reduction Techniques

- Grouping/replication and prediction
  - FS2 [Huang 2005]
  - EEFS [Li 2006]
  - Predictive Data Grouping [Essary 2008]
- Energy-aware prefetching
  - [Manzanares 2006]
- Hybrid: Low-powered hardware with intelligent data-structure
  - FAWN [Andersen 2009]

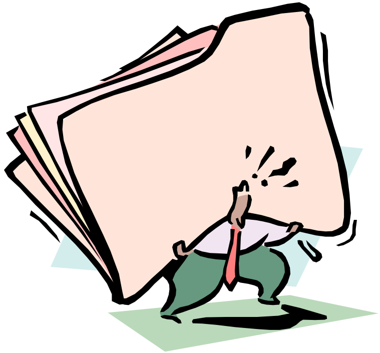
Benchmarking Studies

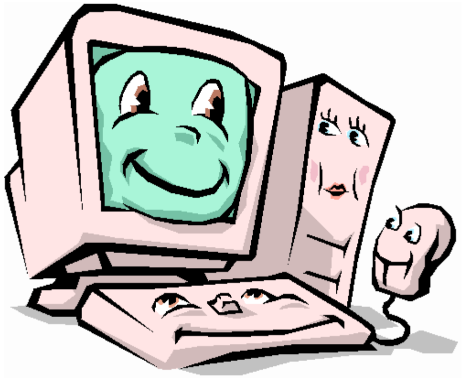
- Benchmarks
  - SPECPower
    - Metric: operations/second/watt
  - JouleSort
    - Metric: sortedrecs/joule
- Benchmark Studies
  - RAID evaluation [Gurumurthi 2003]
  - Compression evaluation [Kothiyal 2009]

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
- 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 maintain/release new version

FileBench Workloads

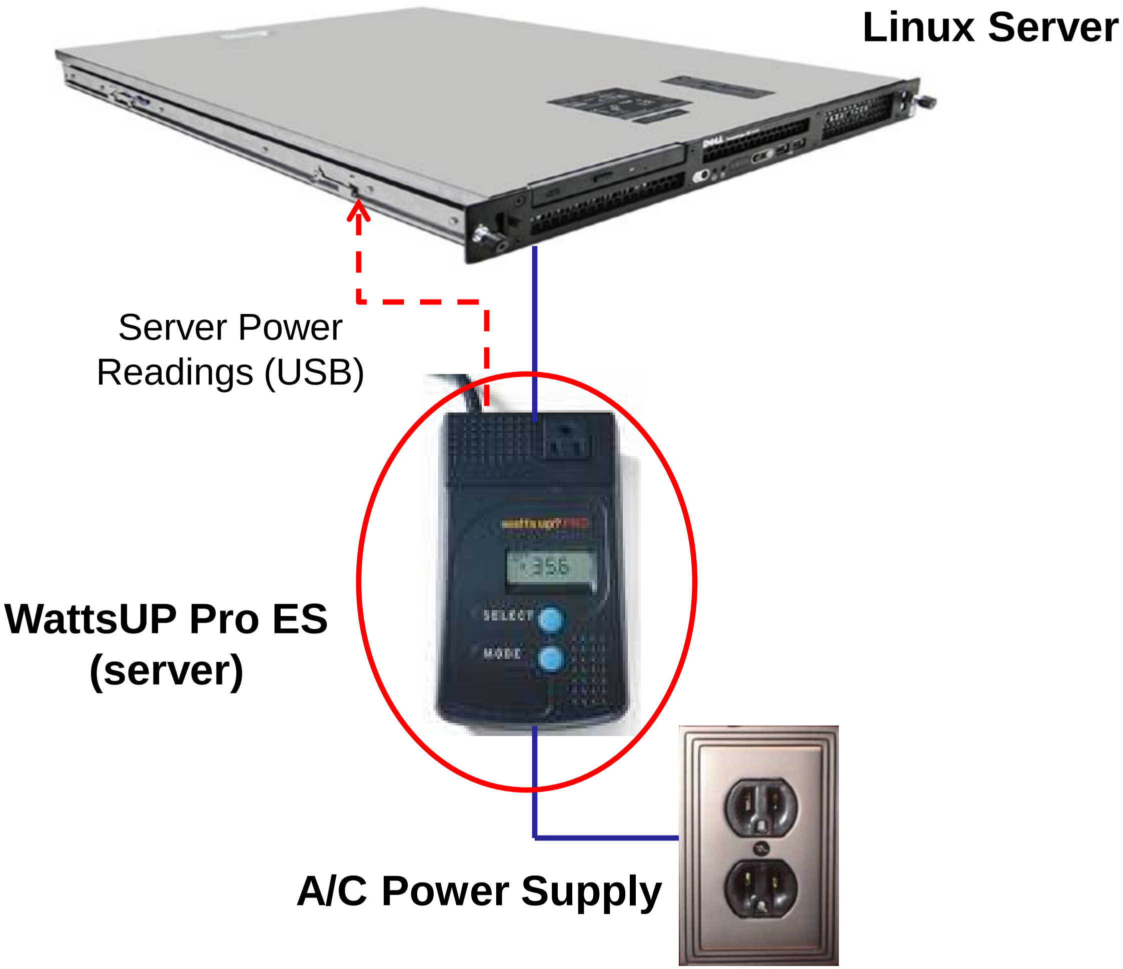
Server workload	Avg. file size	Avg. directory depth	No. of files	I/O size (R/W)	No. of threads	R/W ratio
Mail	16KB	FLAT	50,000	1MB/16KB	100	1:1
Database	0.5GB	FLAT	10	2KB/2KB	200+10	20:1
Web	32KB	3.3	20,000	1MB/16KB	100	10:1
File	256KB	3.6	50,000	1MB/16KB	100	1:2

File System Properties

Features	Ext2	Ext3	ReiserFS	XFS
Disk Layout	Linear	Linear	B+ 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

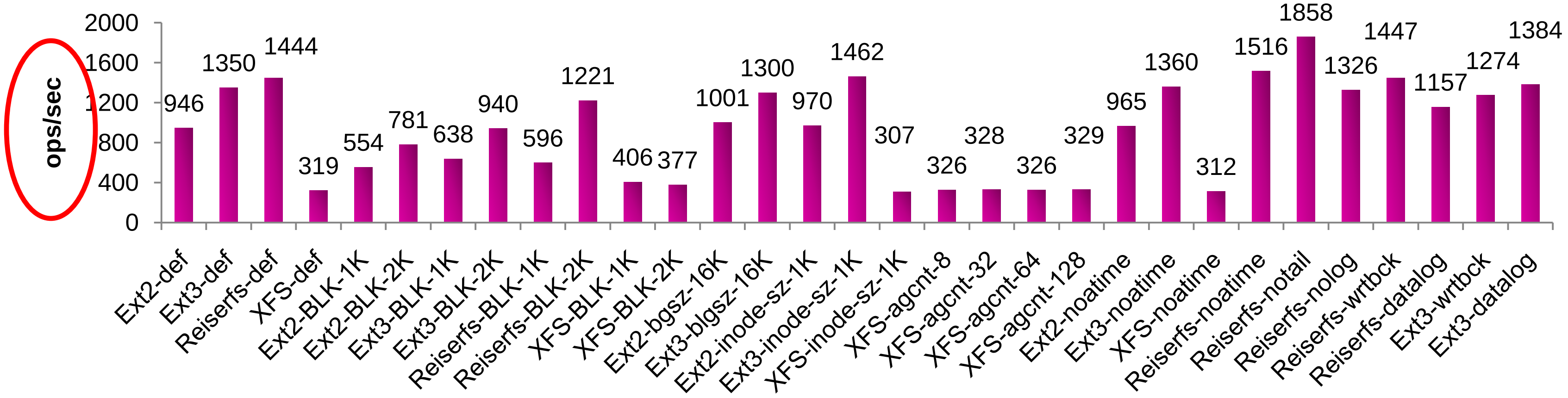
	M1	M2
Machine Age	3+ years (2007)	< 1 year (2009)
CPU Model	Intel Xeon	Intel Nehalem (E5530)
CPU Speed	2.8GHz	2.4GHz
# of CPUs	2 dual core	1 quad core
DVFS	No	Yes
L1 cache size	16KB	128KB
L2 cache size	2MB	1MB
L3 cache size	No	8MB
FSB speed	800 MHz	1066 MHz
RAM size	2048 MB	24GB **(used 2GB)**
RAM type	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

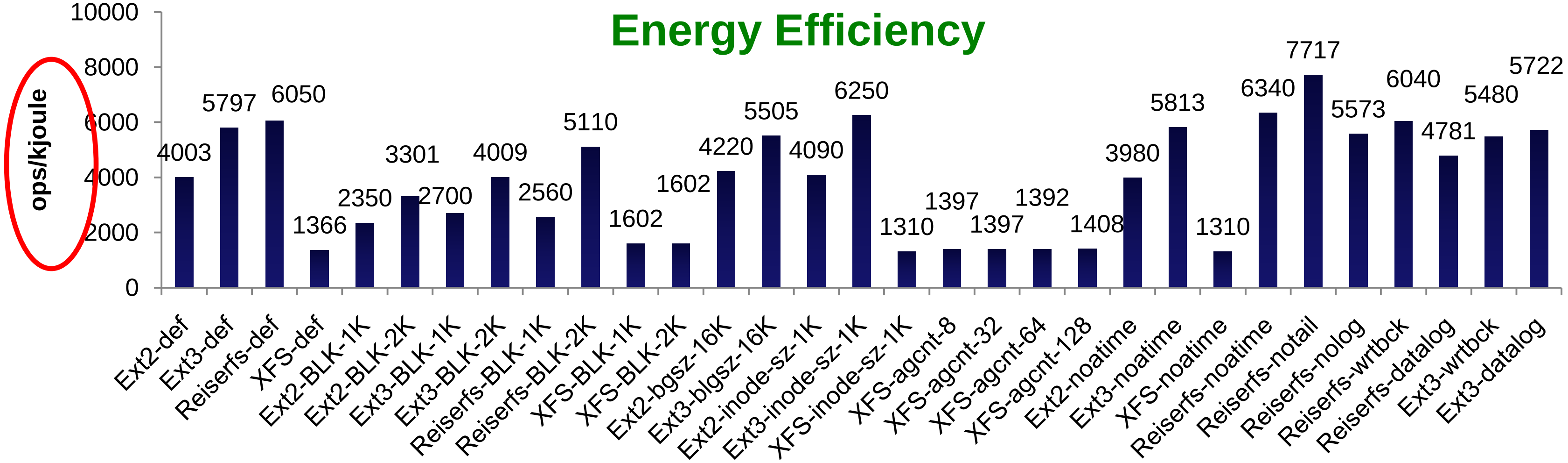
- Motivation
- Related Work
- Experimental Methodology
- **Evaluation Results**
  - **Machine 1 (M1) Results**
  - Machine 2 (M2) Results
- Conclusion and Future Work

Mail Server (M1)

Performance

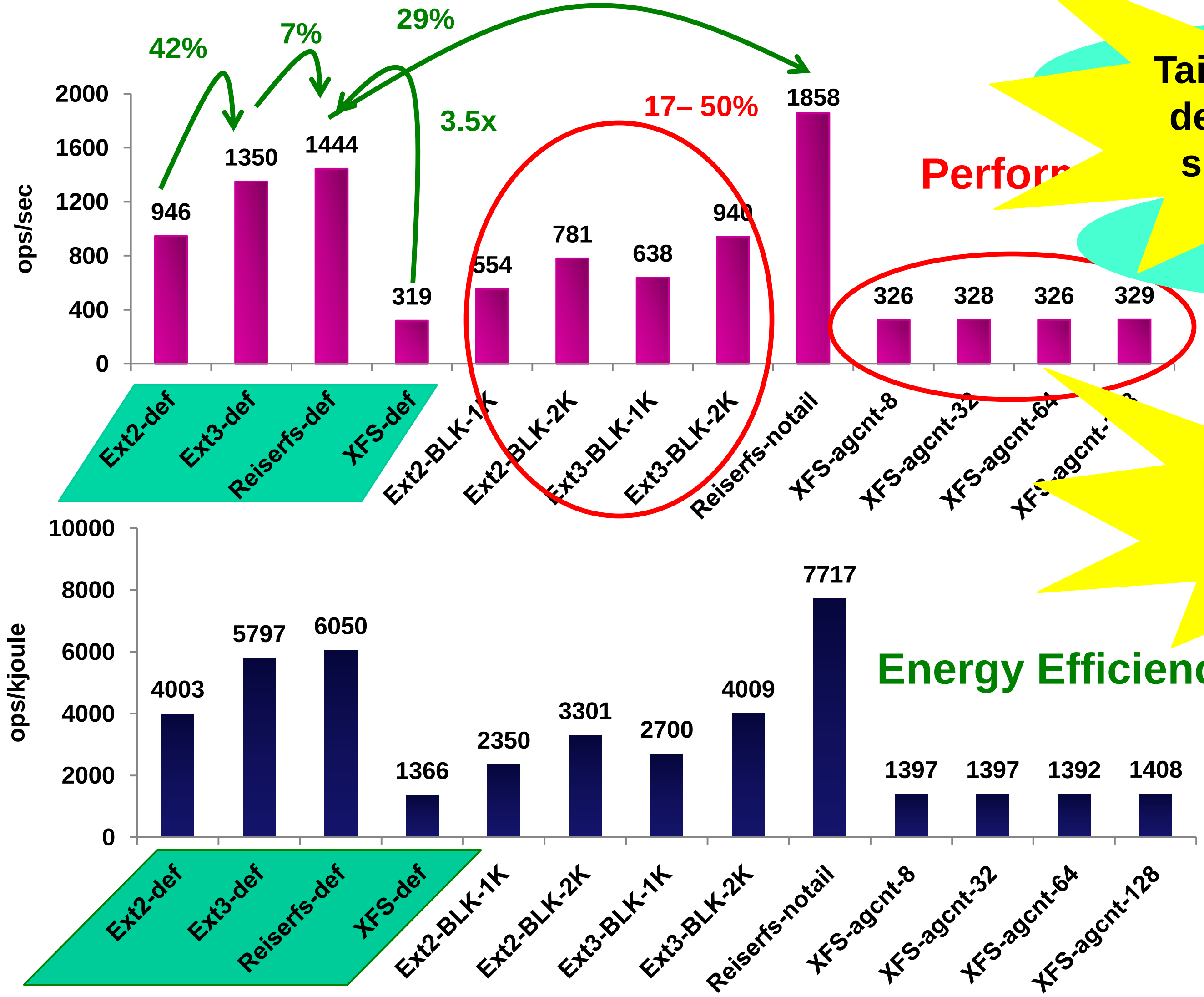


Energy Efficiency



Higher is better

Mail Server (M1)



Perforn

b nec
fsy c

Tail packing on by default – hurting small file reads

ReiserFS-notail best for this configuration

Energy Efficiency

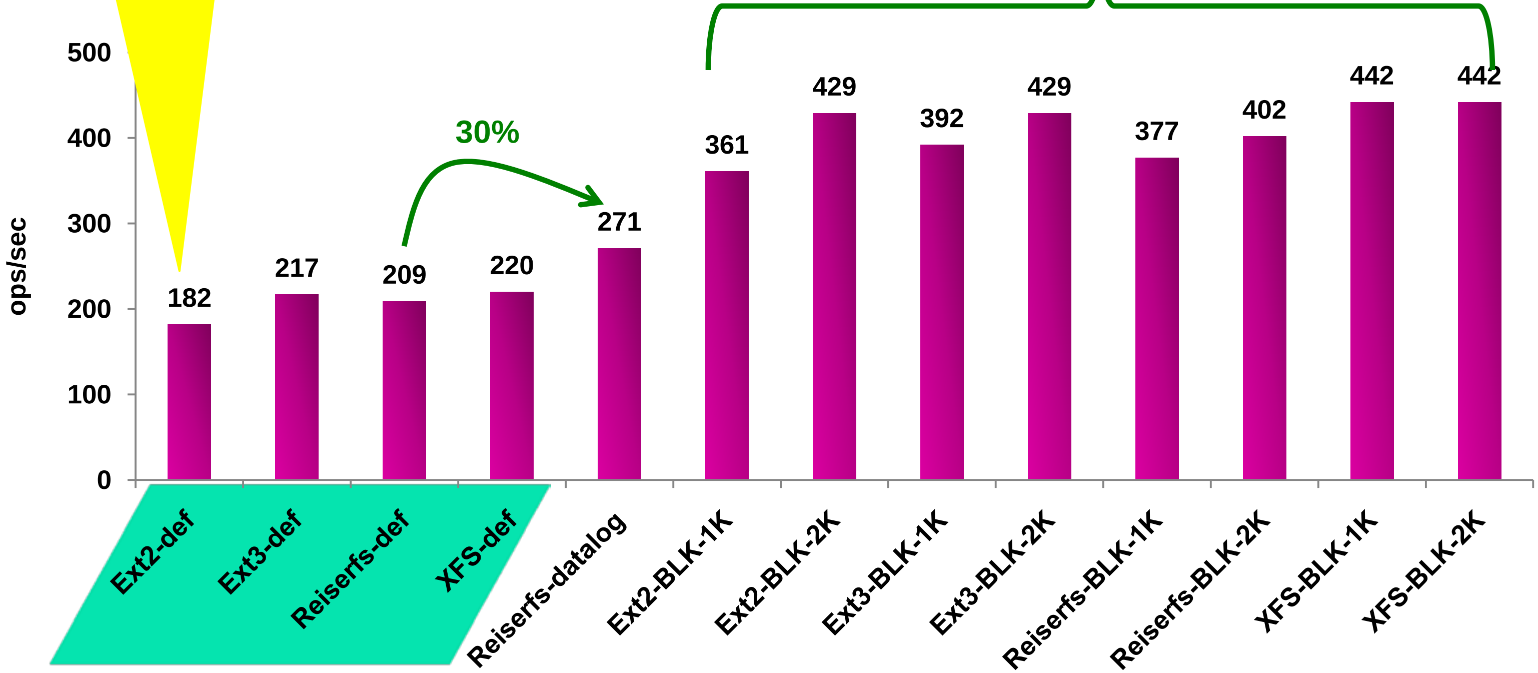
Linearity between Performance and Energy Efficiency

Database Server (M1)

Except for Ext2 other default FS perform similarly

I/O size = Block size

2KB block size boosts the efficiency by ~2x



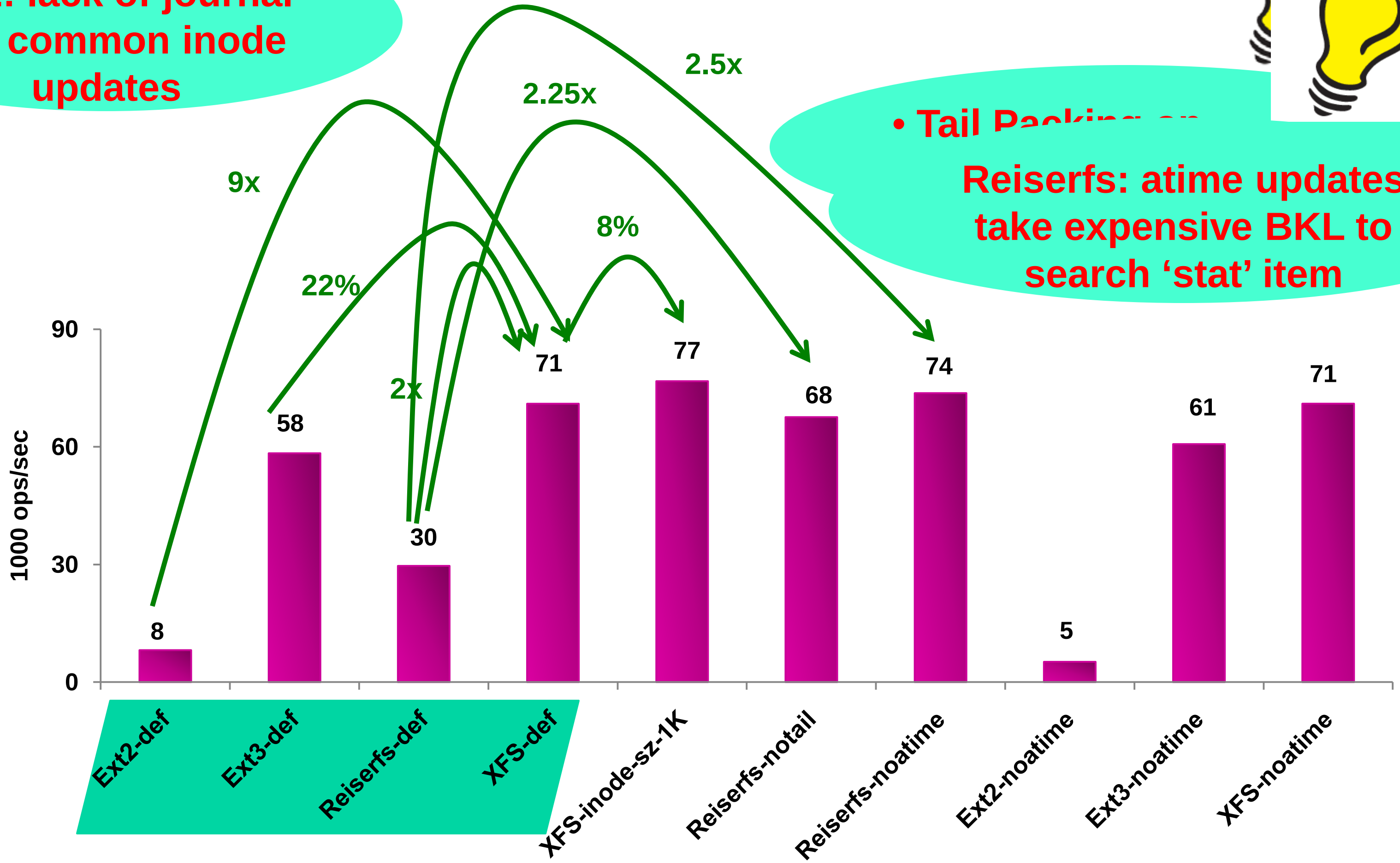
Performance

Web Server (M1)

Ext2: lack of journal and common inode updates

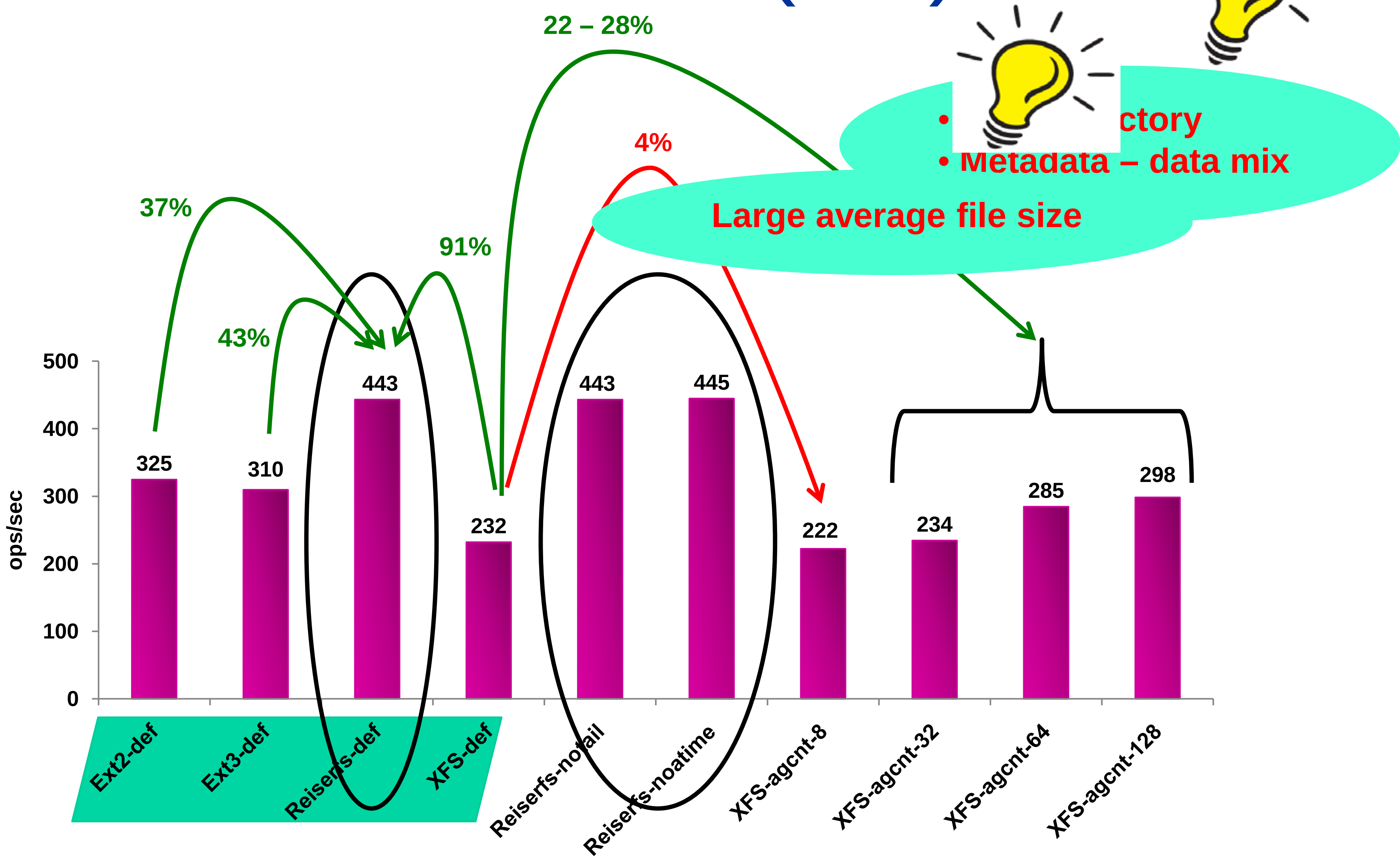
• Tail Packing on

Reiserfs: atime updates take expensive BKL to search ‘stat’ item



Performance

File Server (M1)



Performance

File System Selection Matrix (M1)

- Newer hardware → Different results

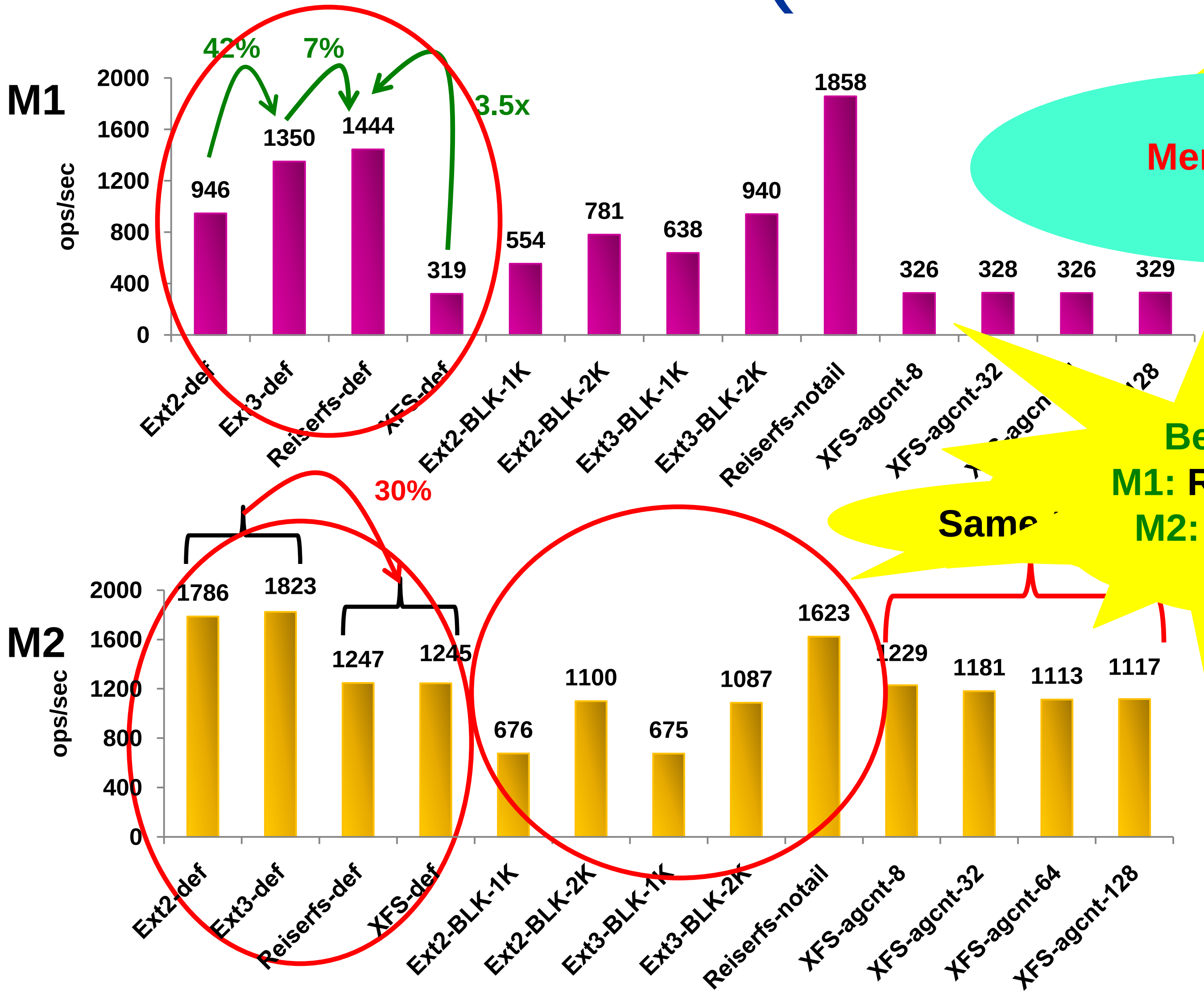
Workload	Best File System (Combination)	Improvement Range (compared to all default FS)	
		Ops/sec	Ops/joule
Web Server	XFS (inode-size-1K)	8% – 9.4x	6% – 7.5x
File Server	ReiserFS (default)	0% – 1.9x	0% – 2.0x
Mail Server	ReiserFS (notail)	29% – 5.8X	28% – 5.7x
Database Server	XFS/Ext3 (BLK-2K)	2.0 –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



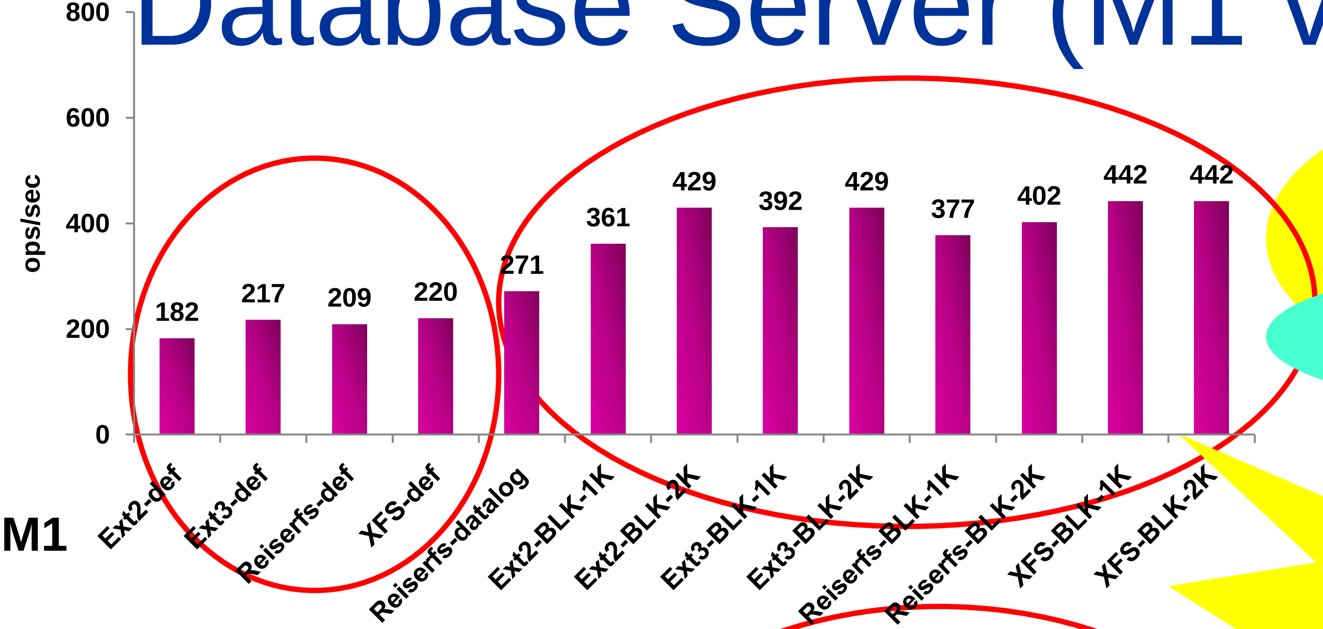
Memory intensive workload

Best Configs
M1: Reiserfs-notail
M2: Ext3-default

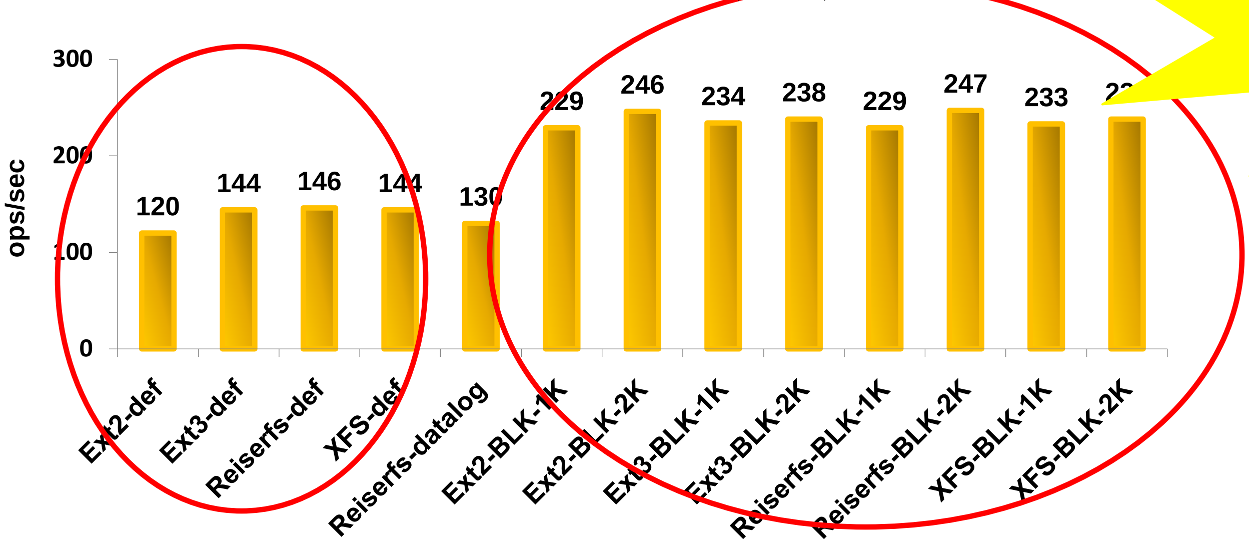
Same

Performance

Database Server (M1 vs. M2)



M1



M2

Perform
trend re
the across

Disk intensive workload

Best Configs for M1 and M2 Ext3 and XFS w/ BLK-2K

2K block size increases performance by ~1.5x

Performance

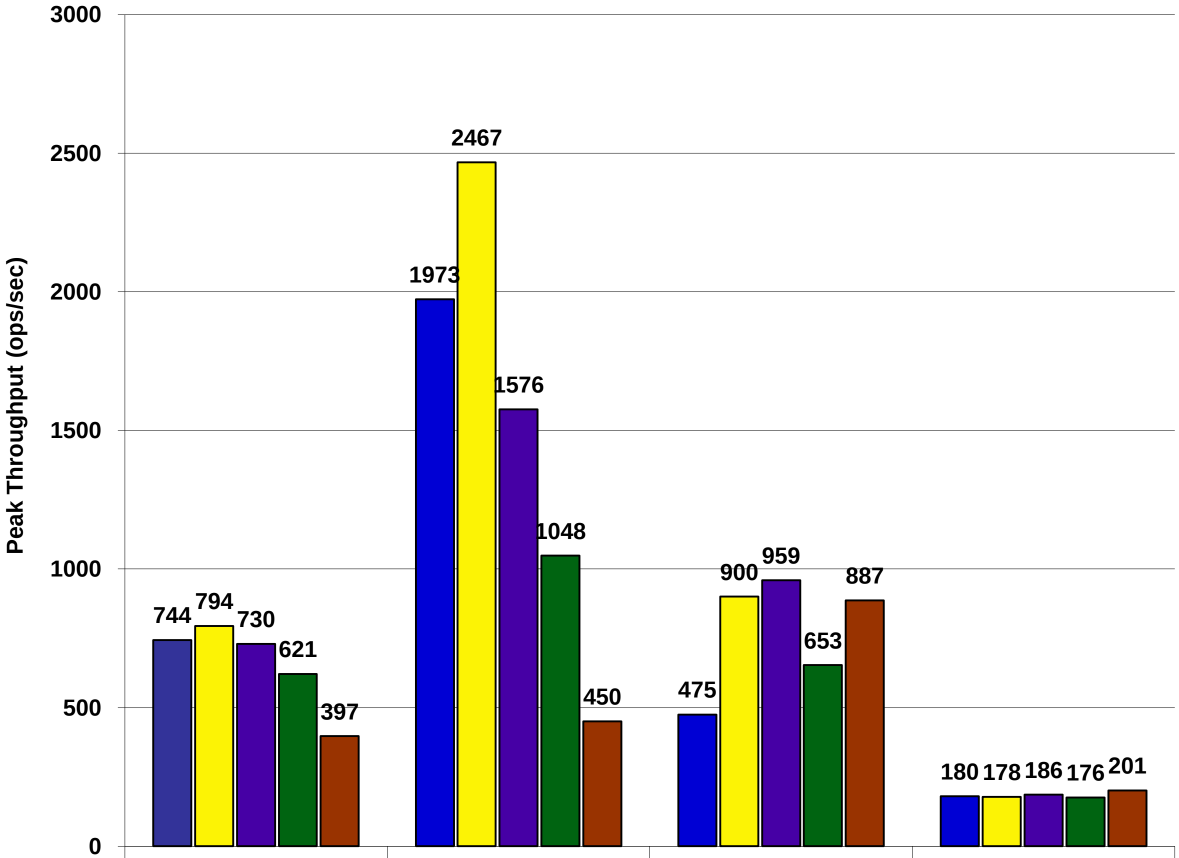
Overview

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

Ongoing Work

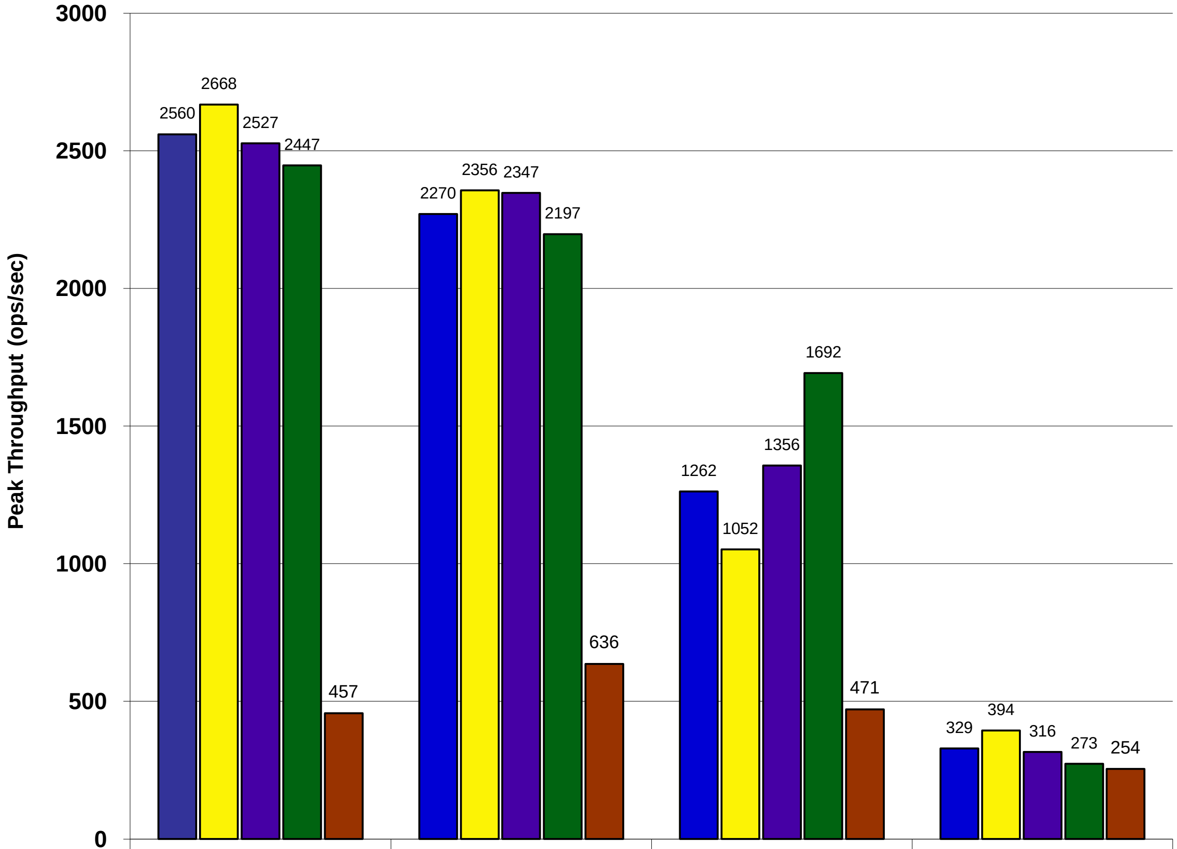
- We are evaluating end-to-end impact of workloads on NFSv4 servers
- Several workloads
- Mix clients and servers
  - Same hardware
  - Linux (Ubuntu, CentOS), FreeBSD, OpenSolaris

Results: Web Server, Server-wise



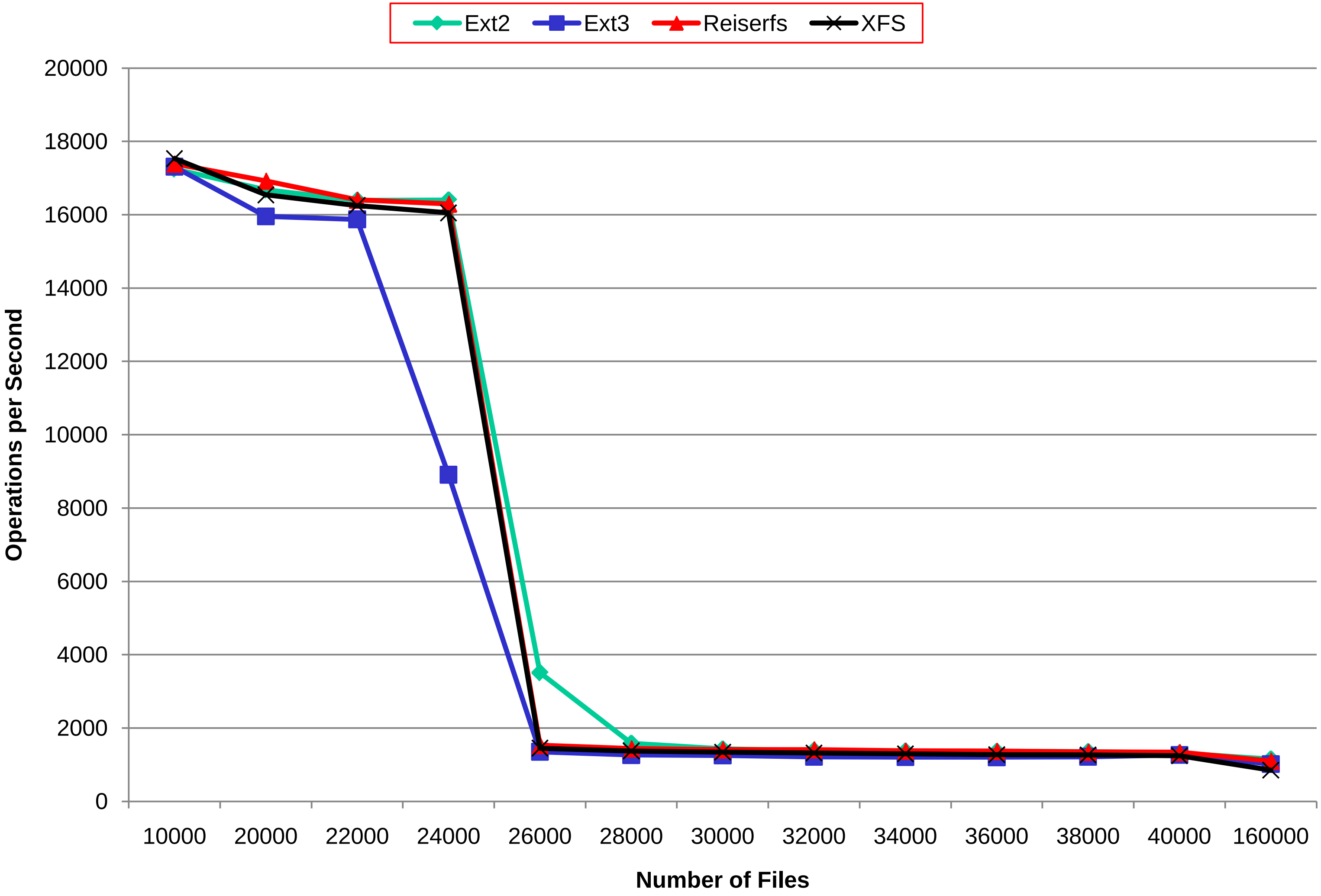
	CentOS-server	Ubuntu-server	FreeBSD-server	OpenSolaris-server
🟦 CentOS-client	744	1973	475	180
🟨 Ubuntu-client	794	2467	900	178
🟪 FreeBSD-client	730	1576	959	186
🟩 OpenSolaris-client	621	1048	653	176
🟫 LocalFS-client	397	450	887	201

Results: Mail Server, Server-wise



	CentOS-server	Ubuntu-server	FreeBSD-server	OpenSolaris-server
🟦 CentOS-client	2560	2270	1262	329
🟨 Ubuntu-client	2668	2356	1052	394
🟪 FreeBSD-client	2527	2347	1356	316
🟩 OpenSolaris-client	2447	2197	1692	273
🟫 LocalFS-client	457	636	471	254

Scaling Web Server Performance



Conclusions

- The Bad
  - Software had gotten too complex
  - Workloads drive performance-energy
  - Depend also on hardware, software, configurations
- The Good
  - Significant savings possible
    - Small savings accumulate over long run
  - Commercial & Research opportunities
- The Ugly
  - Need workload-specific software

Ongoing/Future Work

- Study multiple dimensions
  - New FS, Disk Scheduler, RAID, LVM, etc.
  - Client/Server Systems
  - Disk Types: SAS, SSD, etc.
  - Cluster Storage, SANs, OS
- Develop auto-configuration tools
- Develop workload-specific storage stacks
  - I/O schedulers, file systems, caching

Server-Class Energy and Performance Evaluations

Q&A

Erez Zadok

ezk@fsl.cs.sunysb.edu

File systems and Storage Lab

Stony Brook University

http://green.filesystems.org/