Exploiting superpages in a nonvolatile memory file system

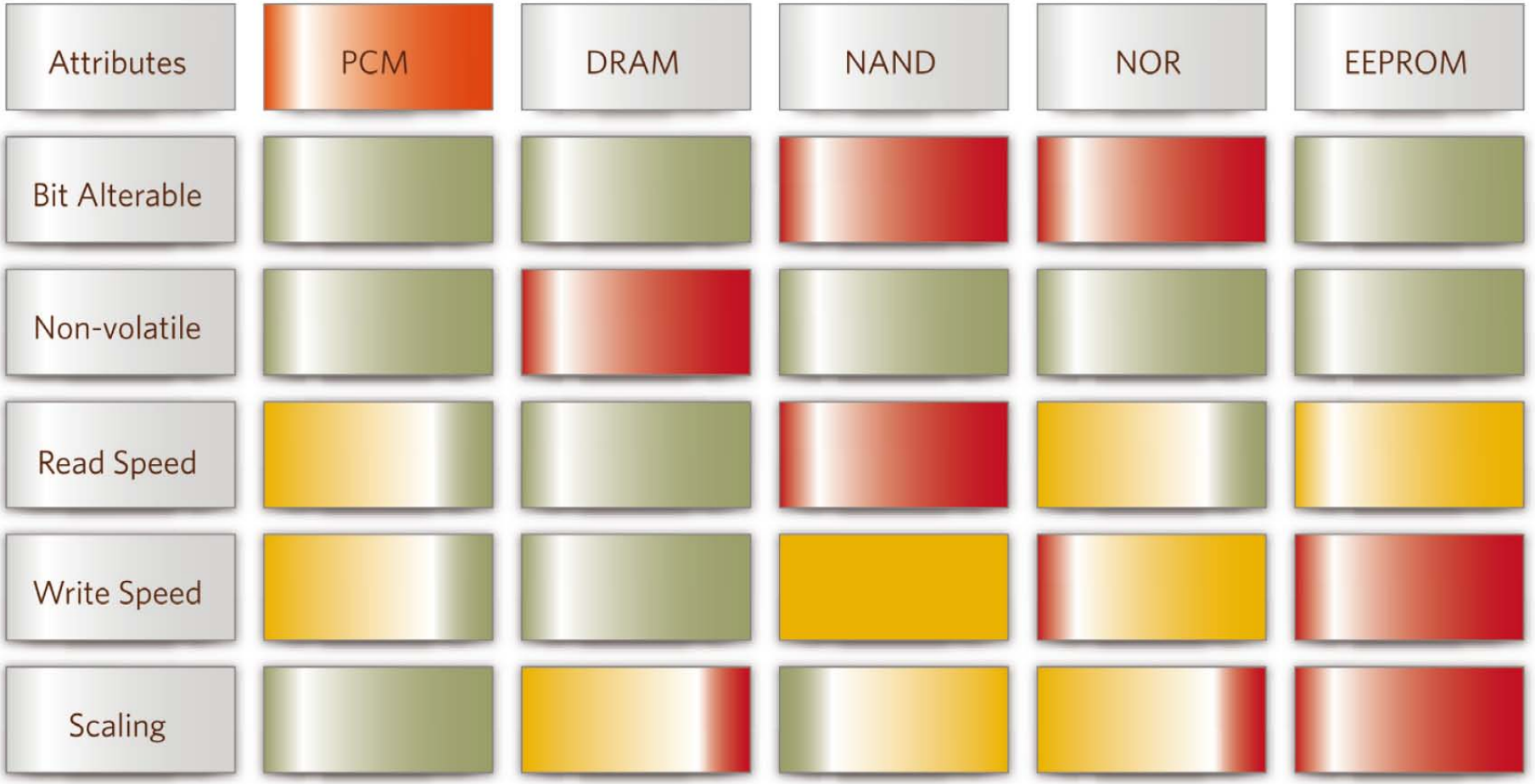
Sheng Qiu, Narasimha Reddy

Texas A&M University

What is SCM?

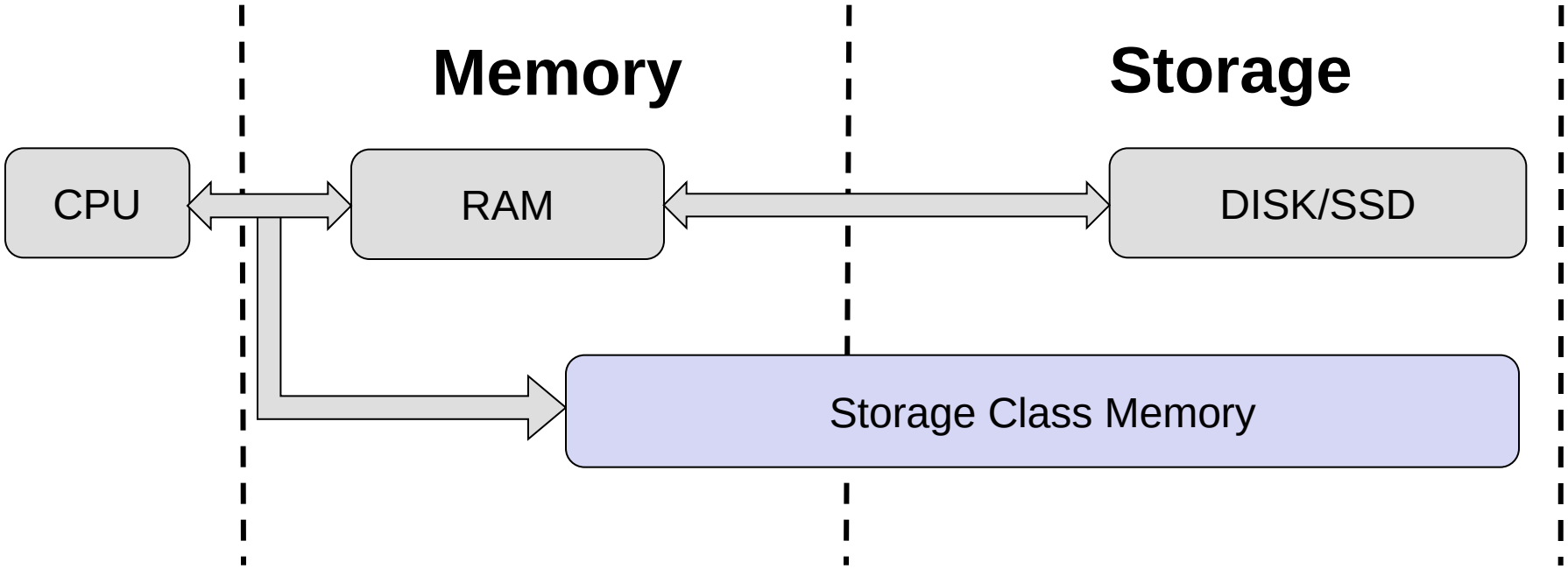
- Storage Class Memory
  - Byte-addressable, like DRAM
  - Non-volatile, persistent storage
- Example: Phase Change Memory

PCM Attributes



From Numonyx

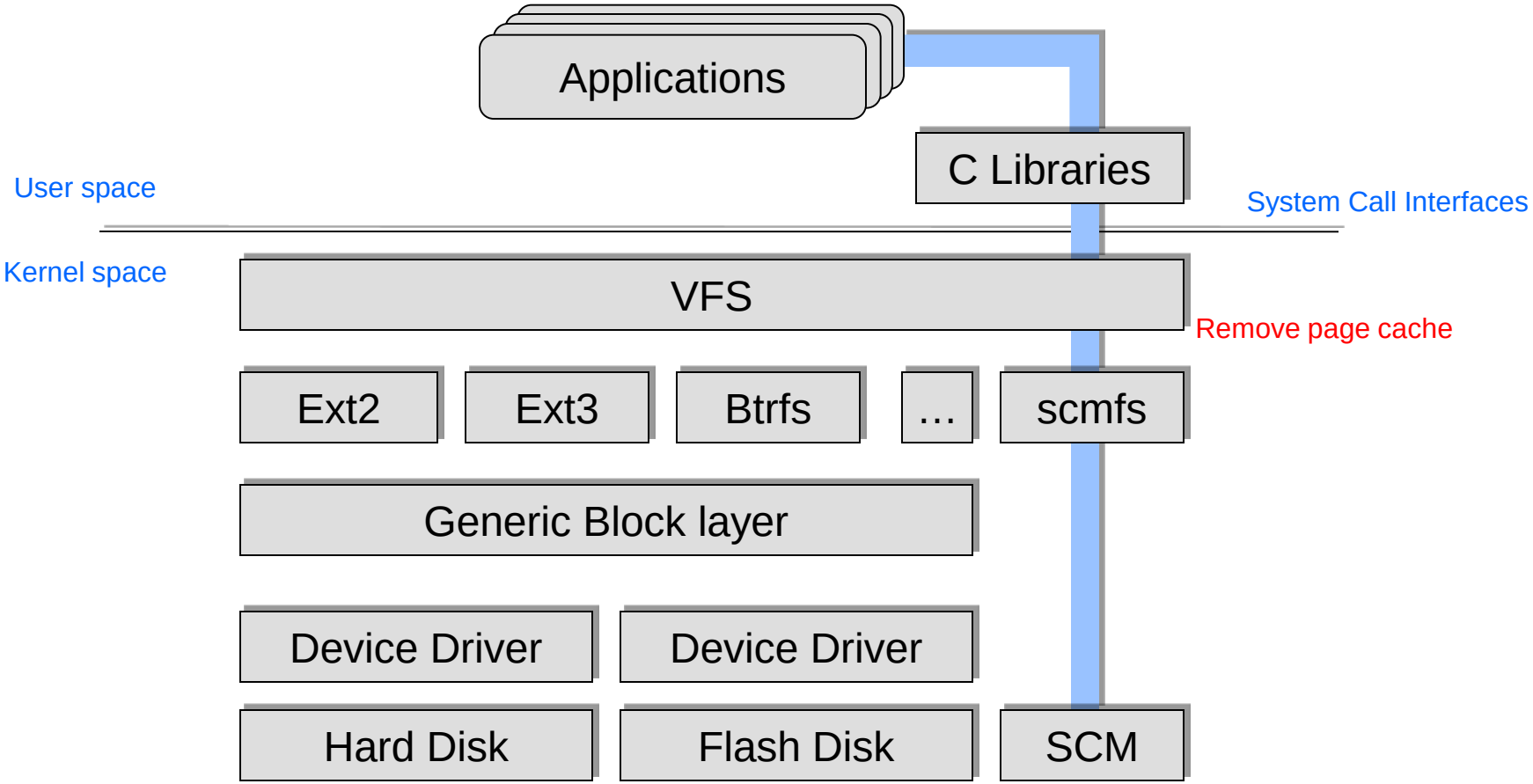
Hardware hierarchy



How to use SCM as storage?

- Device level
  - Use existing file system on RamDisk
- File system level
  - Design a new file system

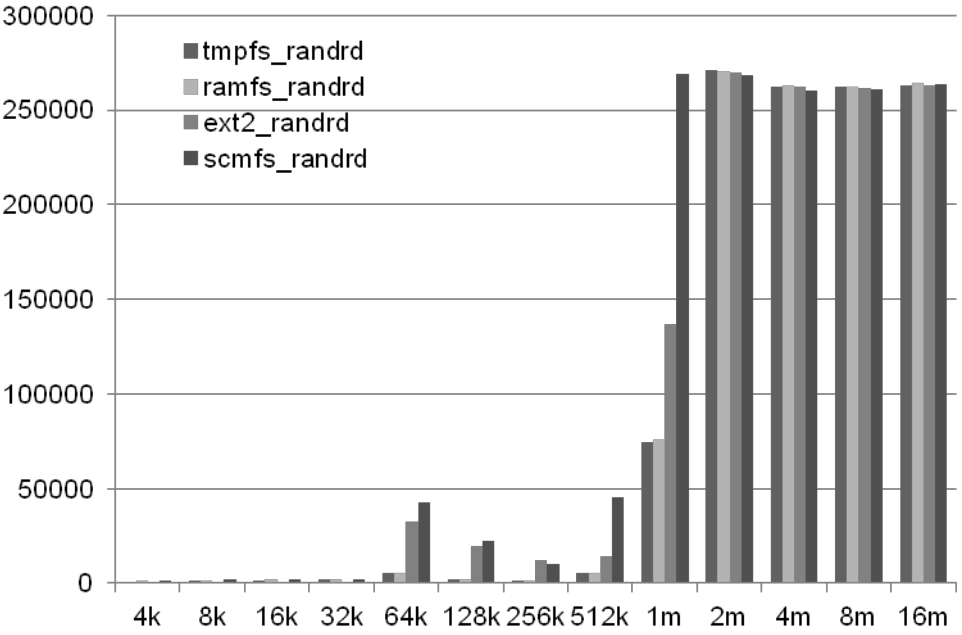
New FS on SCM


Applications
C Libraries
User space
System Call Interfaces
Kernel space
VFS
Remove page cache
Ext2
Ext3
Btrfs
…
scmfs
Generic Block layer
Device Driver
Device Driver
Hard Disk
Flash Disk
SCM

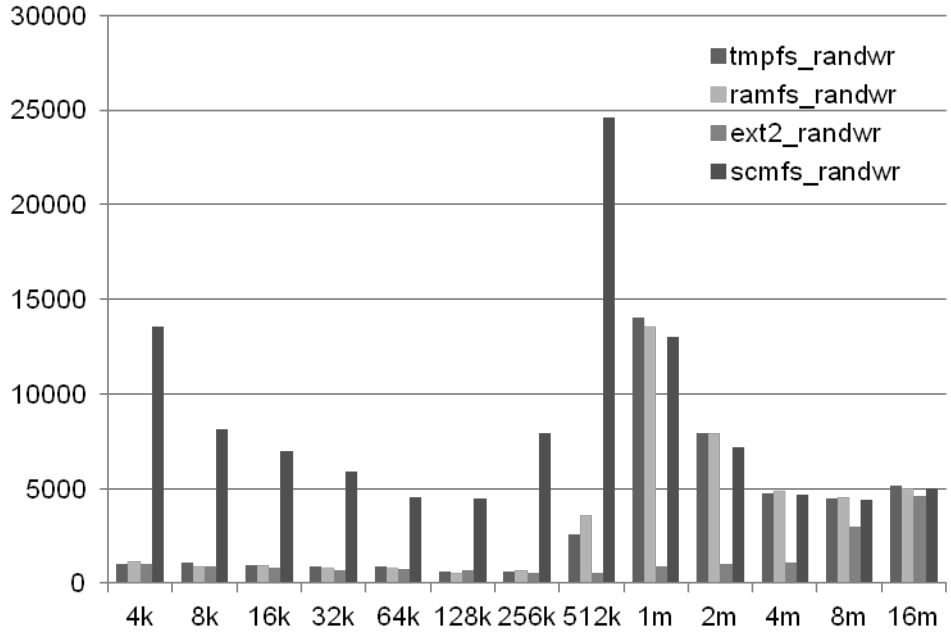
Wrap up

- Re-utilize Memory Management (MM) module in O/S to do block management
- Implement the file system in Virtual Address Space
- Keep all the files contiguous in Virtual Address Space
- **Cause more Data TLB Misses**

Data TLB Misses (IoZone Random)



Read



Write

SCMFS suffers from TLB misses.

Why higher TLB misses in SCMFS?

Scmfs works here, use small page size (4K)

RamDisk works here, use big page size(2M)

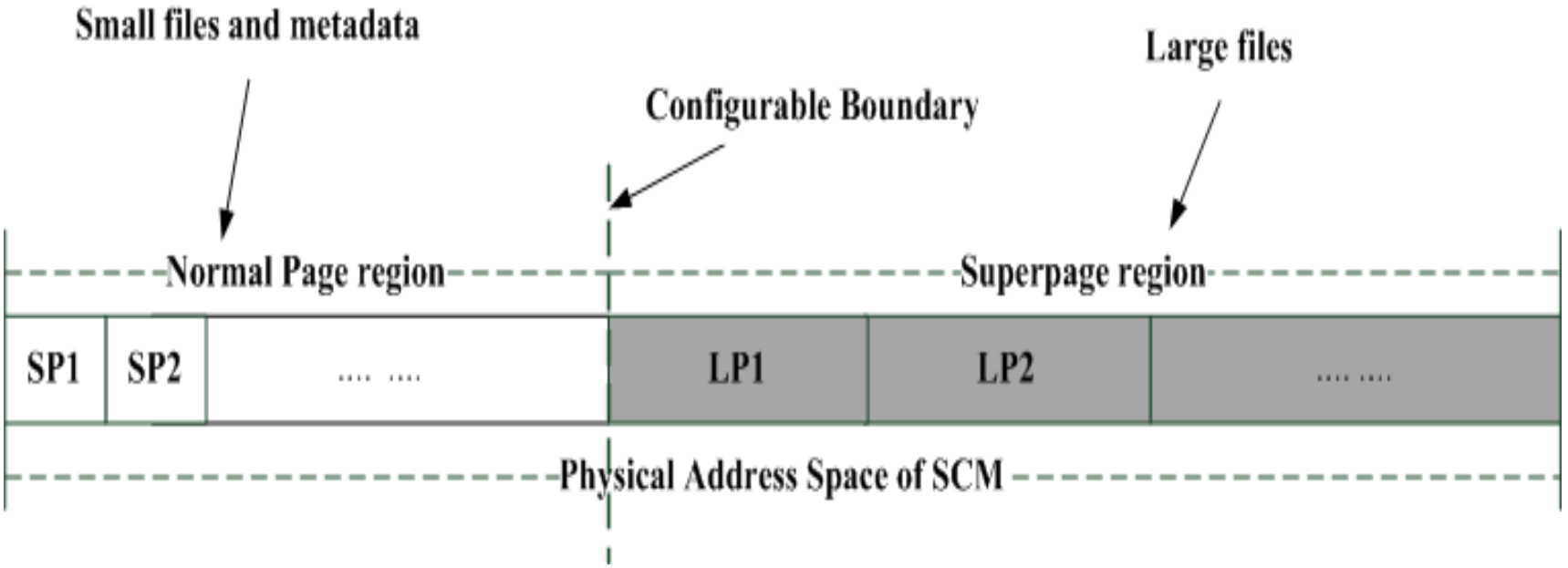
Memory Map (x86_64)

(=47 bits) user space	0000000000000000 - 00007fffffffffff
hole caused by [48:63] sign extension	
(=47 bits)nvmalloc space	**ffff000000000000 – ffff7fffffffffff**
(=40 bits) guard hole	ffff800000000000 - ffff80ffffffffff
(=64 TB) direct mapping of all phys.	ffff880000000000 - ffffc7ffffffffff
(=40 bits) hole	ffffc80000000000 - ffffc8ffffffffff
(=45 bits) vmalloc/ioremap space	**ffffc90000000000** - ffffe8ffffffffff
(=40 bits) hole	ffffe90000000000 - ffffe9ffffffffff
(=40 bits) virtual memory map (1TB)	ffffea0000000000 - ffffeaffffffffff
(=512 MB) kernel text mapping, from phys 0	**ffffffff80000000** - ffffffffa0000000
(=1536 MB) module mapping space	ffffffffa0000000 - ffffffffff00000

How to reduce Data TLB Misses?

- Utilize two types of memory pages (i.e. 4K and 2M Bytes on x86_64)
- Use regular 4kb page for small files and metadata
- Pre-allocate superpages and allocate large files on them

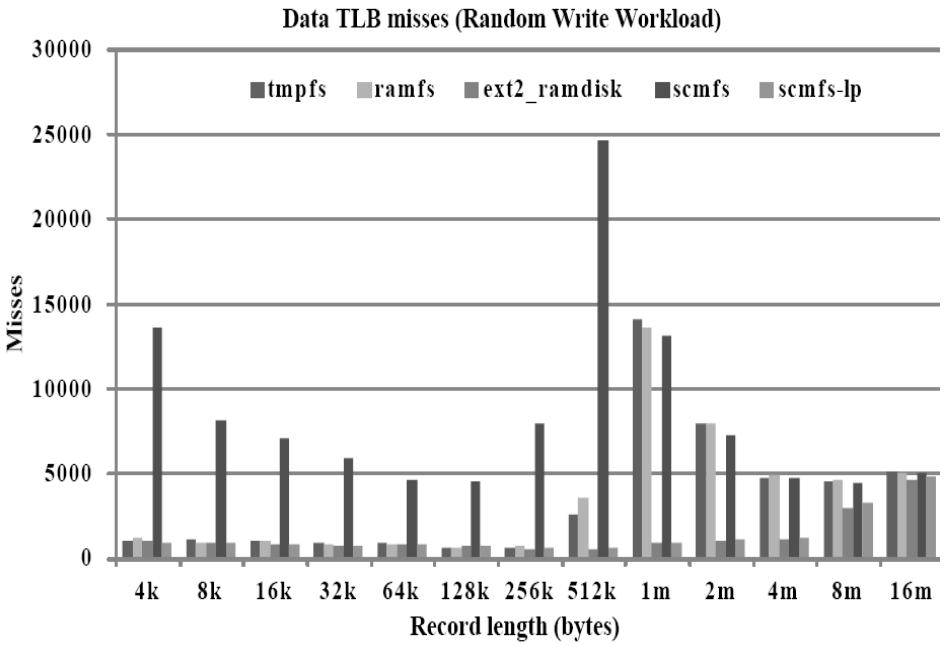
Management of two types of memory pages



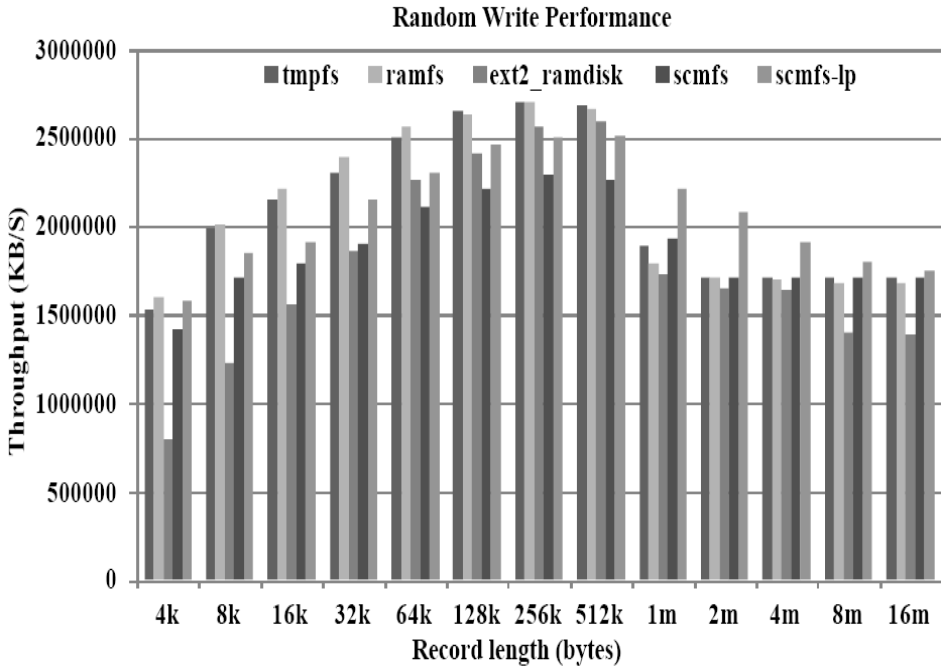
Evaluation

- Environment
  - 2.33GHz Intel Core2 Quad Processor Q8200
  - 8GB RAM, 4GB is used as SCM.
  - Linux 2.6.33
- Benchmarks
  - IoZone
  - Postmark
- Use Performance Counter to analyze the results

IOZONE Random Write

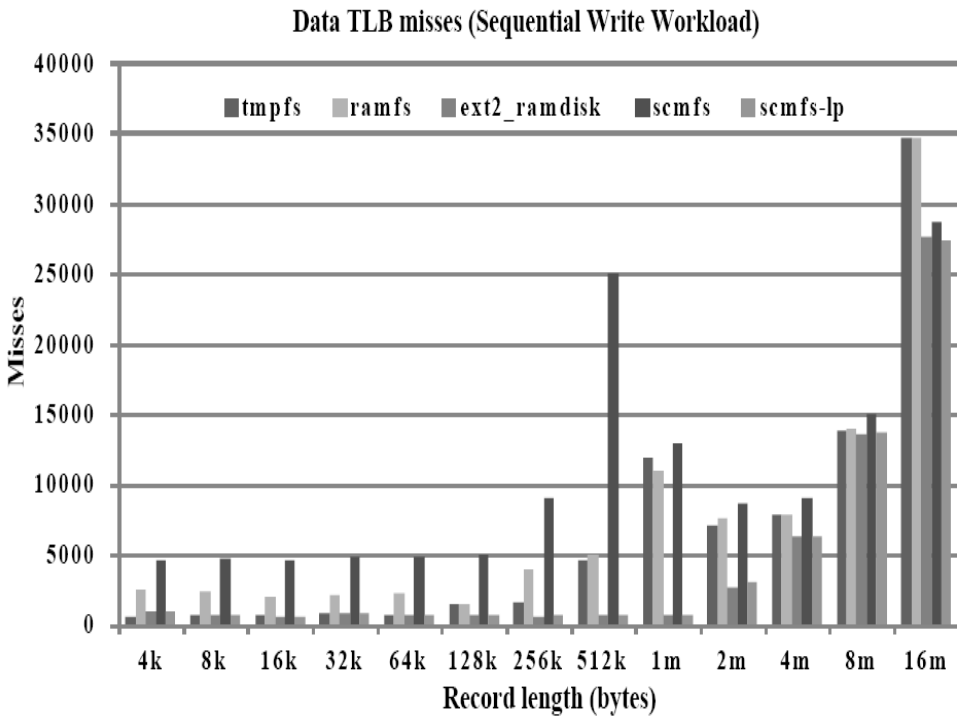


Data TLB Misses

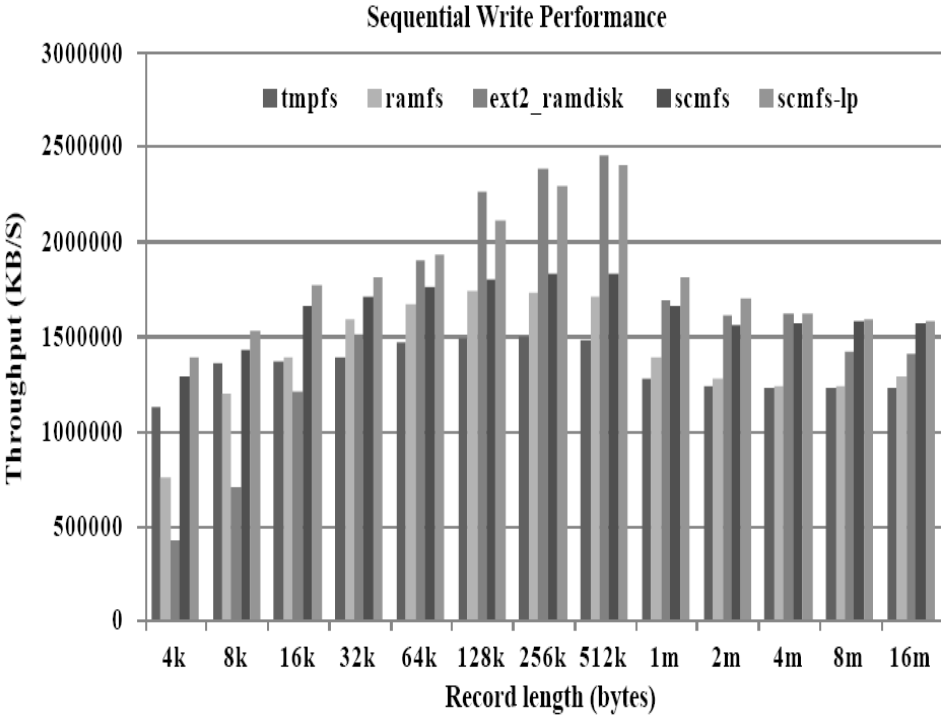


Throughput(kbytes/s)

IOZONE Sequential Write

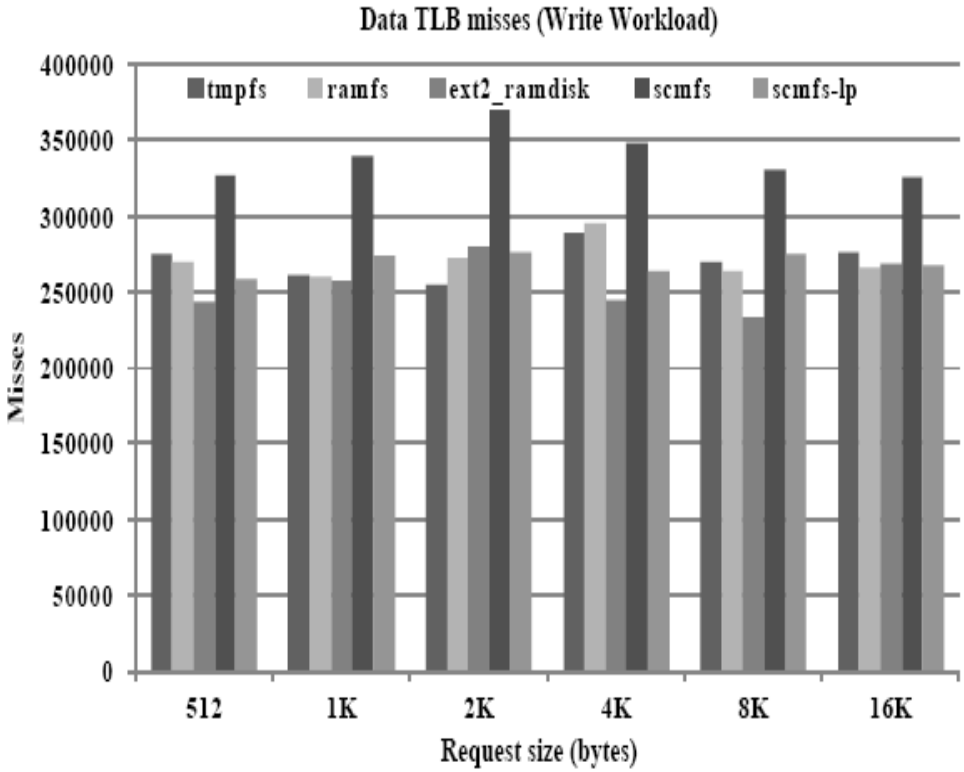


Data TLB misses

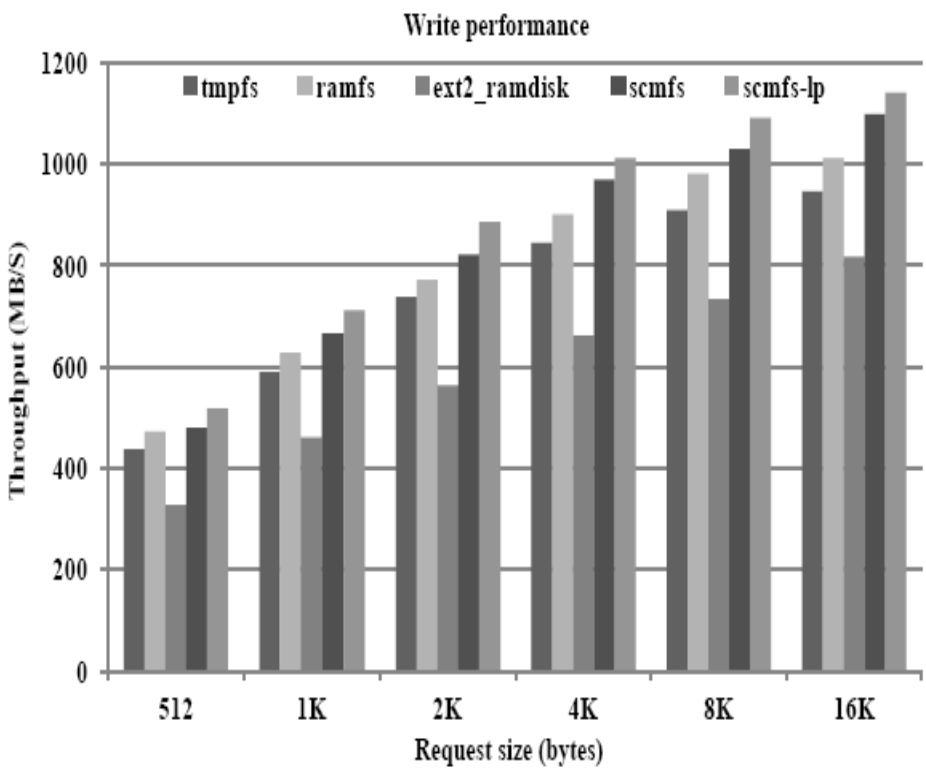


Throughput (KB/S)

Postmark Write

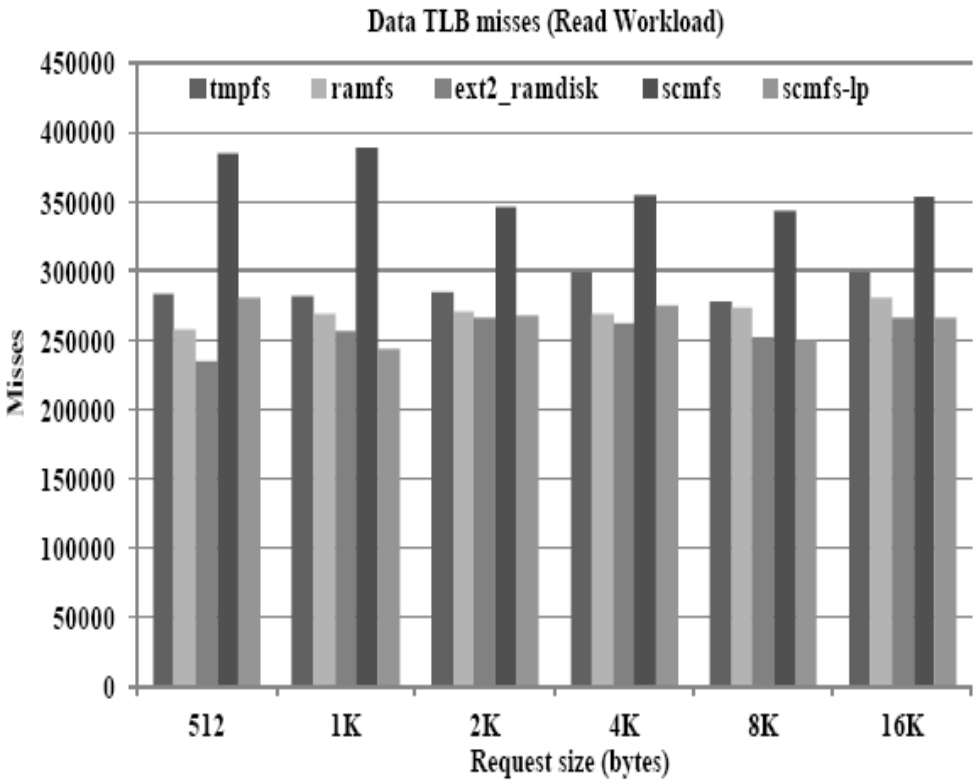


Data TLB misses

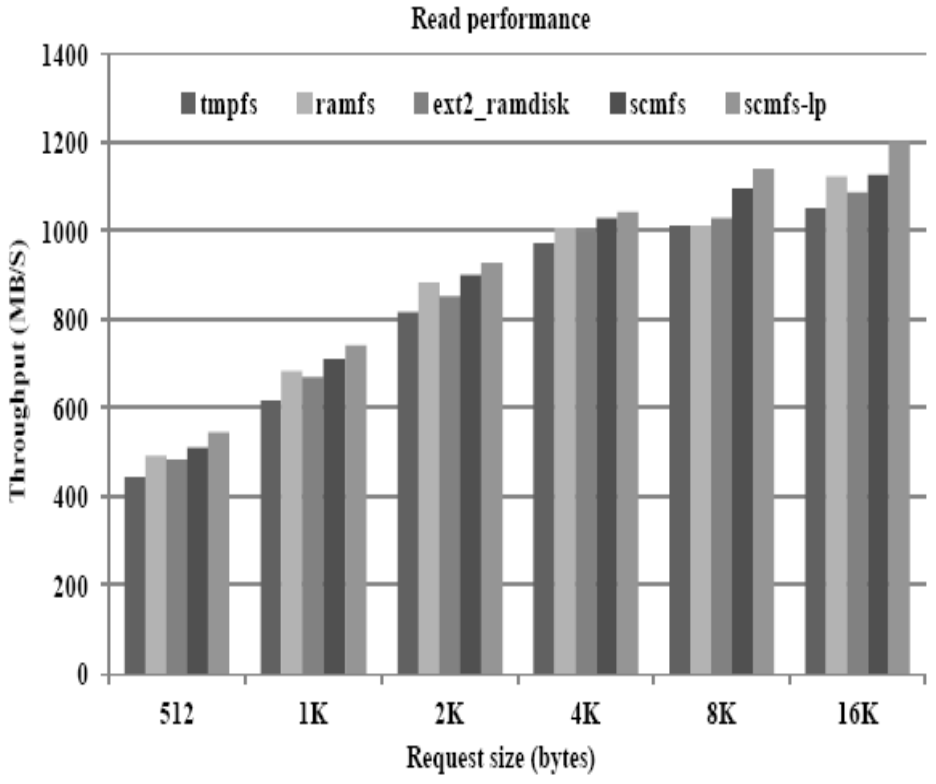


Throughput (MB/S)

Postmark Read



Data TLB misses



Throughput (MB/S)

Conclusion

- We effectively reduced Data TLB misses by utilizing both regular and super pages.
- The FS's performance is further improved.
- Design of File System should adapt to the change of hardware hierarchy.
- Performance depends on more factors than ever.

Thanks