Moneta-Direct:
Providing Safe, User Space Access to Fast, Solid State Disks

Adrian Caulfield, Todor Mollov,

Louis Eisner, Arup De, Joel Coburn, Steven Swanson

Non-volatile Systems Laboratory
Department of Computer Science and Engineering
University of California, San Diego

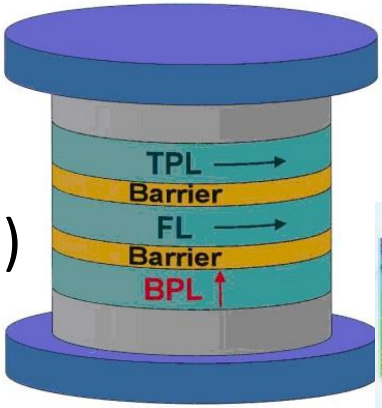
Welcome to the Data Age

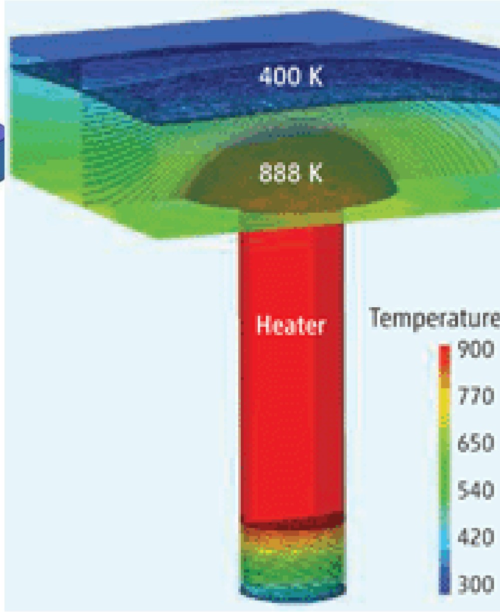
- The world processed 9 Zettabytes of data in 2008*
- Acquiring data is easy
- Extracting knowledge is hard
  - Storage performance is major bottleneck
  - Solid-state storage can help

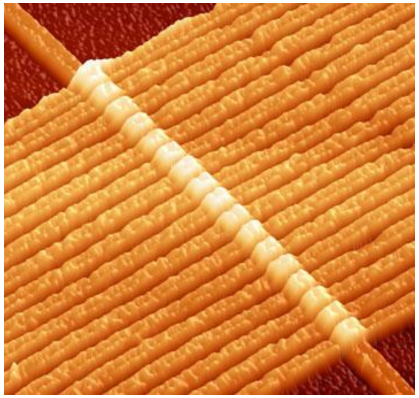
*http://hmi.ucsd.edu

Faster-than-flash Non-volatile Memories

- Necessary characteristics
  - As fast as DRAM (or nearly so)
  - As dense as flash (or nearly so)
  - Non-volatile
  - Reliable
- Candidates
  - Phase change memory
  - Spin-torque MRAMs
  - Memristor memories







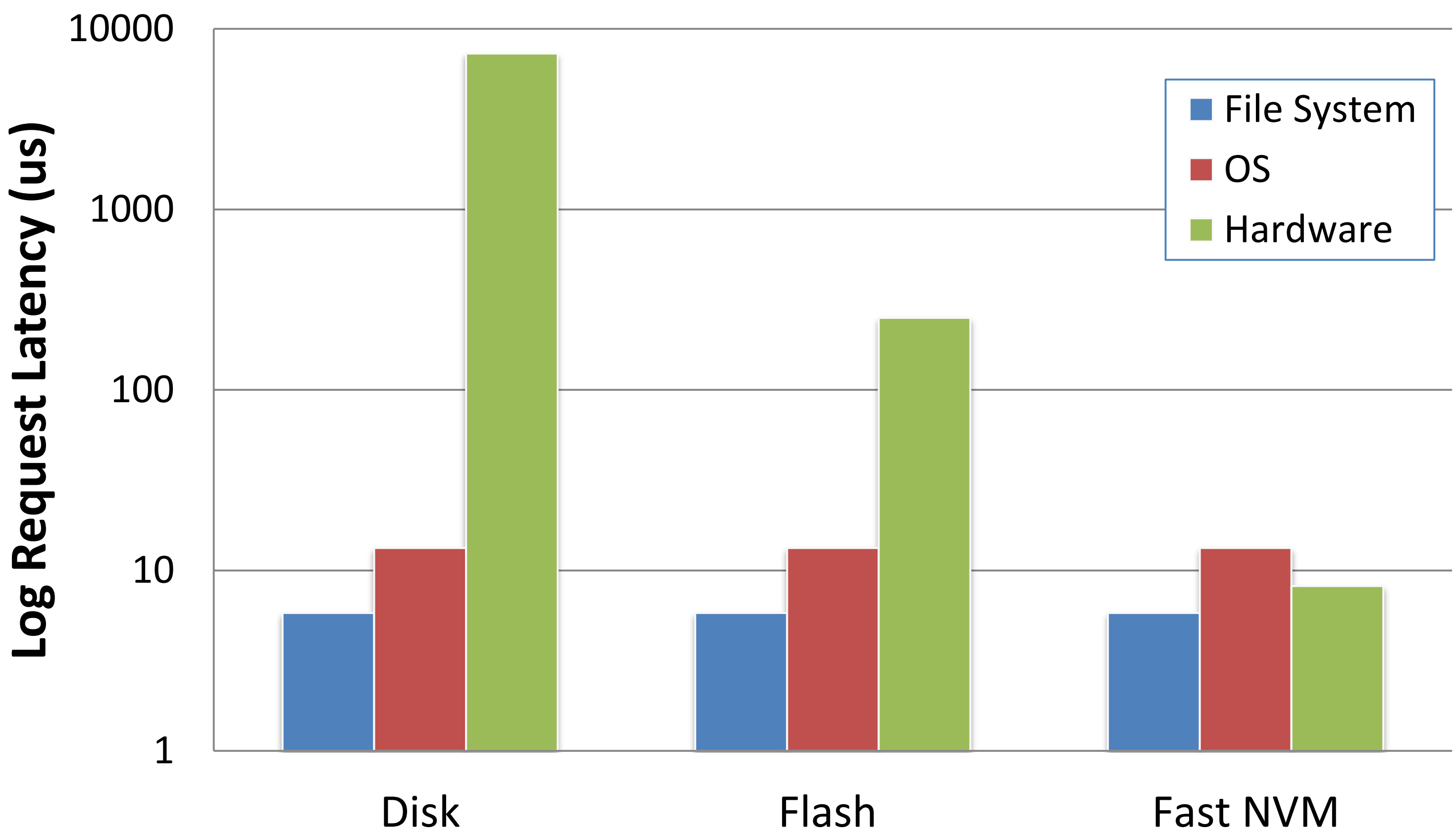
The Future Storage Performance: More than Moore's Law

Hard Drives	PCIe-Flash 2007	PCIe-NVM 2013?	
			
Lat.: 7.1ms	68us	12us	
BW: 2.6MB/s	250MB/s	1.7GB/s	
1x	104x	591x	= 2.89x/yr
1x	96x	669x	= 2.95x/yr

*Random 4KB Reads from user space

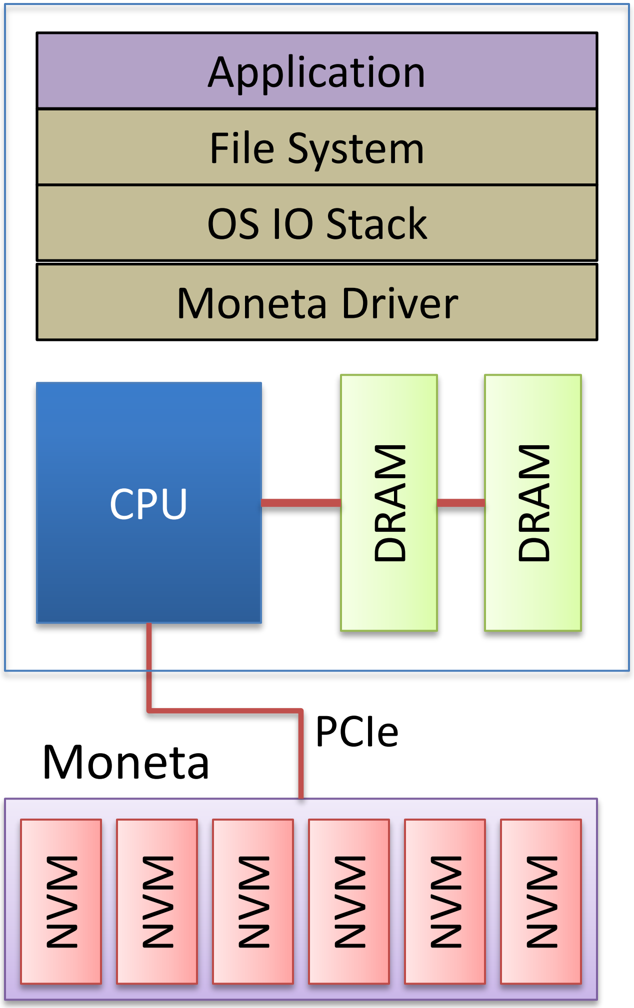
NVSL
Non-volatile Systems Laboratory

Software Overheads



NVSL
Non-volatile Systems Laboratory

Baseline Moneta: An SSD for Fast NVMs



[SC 2010, MICRO 2010]

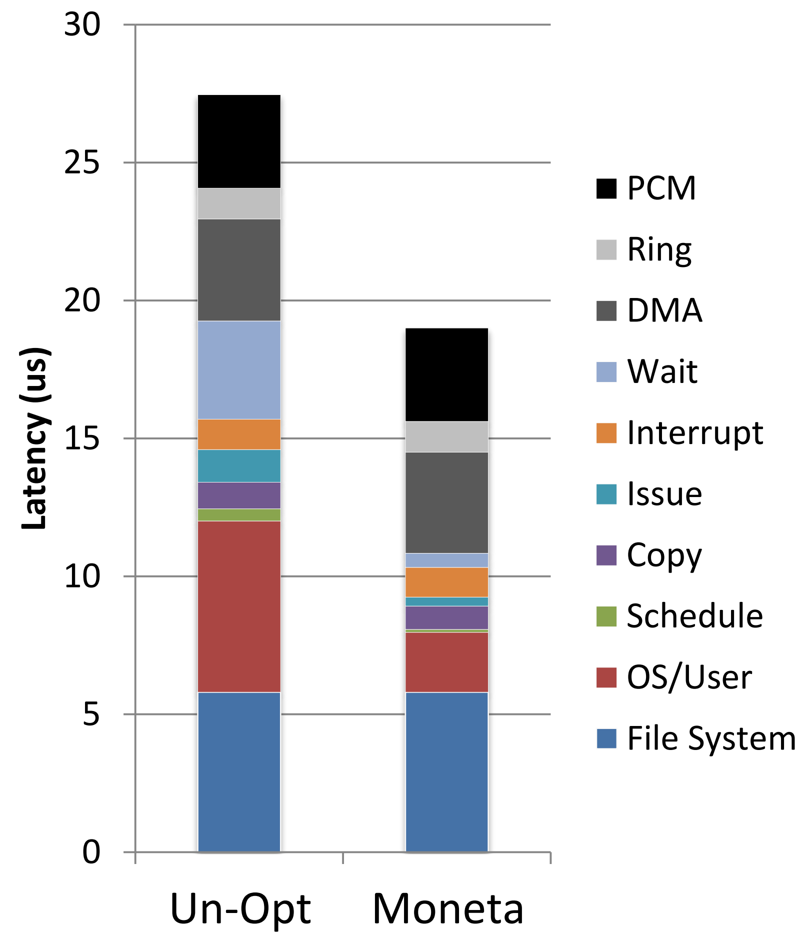
NVSL
Non-volatile Systems Laboratory

The Moneta Prototype

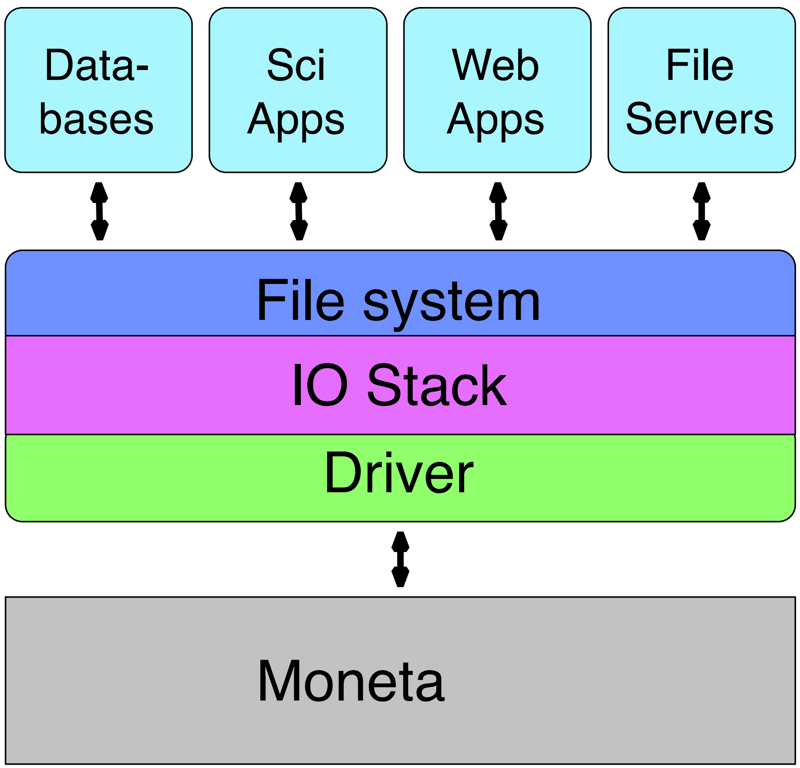
- FPGA-based implementation
- DDR2 DRAM emulates PCM
  - Configurable memory latency
  - 48 ns reads, 150 ns writes
  - 64GB across 8 controllers
- PCIe: 2 GB/s, full duplex

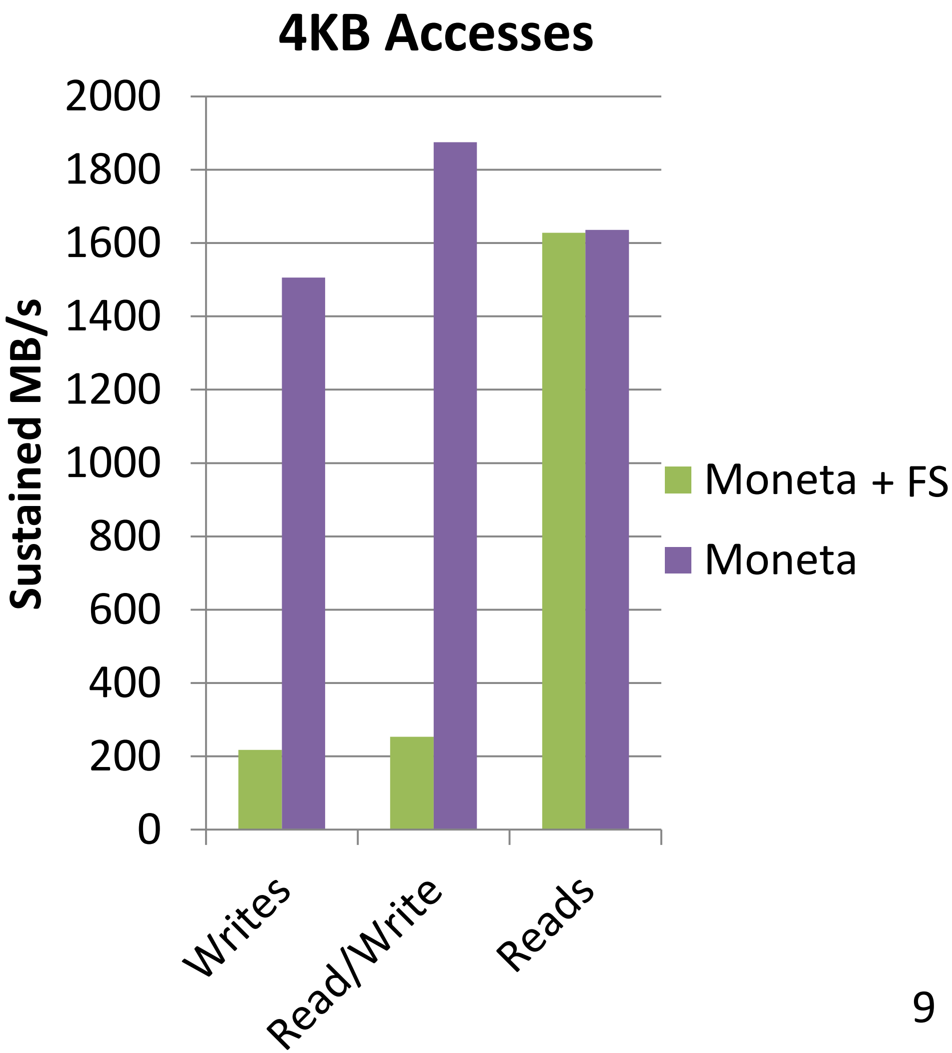
Optimizing Moneta Latency

- Optimizations
  - Remove IO Scheduler
  - Atomic, Lock-free Structures
  - Codesigned HW/SW
- Results
  - 62% less SW overhead (w/o FS)
  - 940K 512B IOPS
- What's left?
  - 5us of OS/driver latency
  - 5us of FS overhead



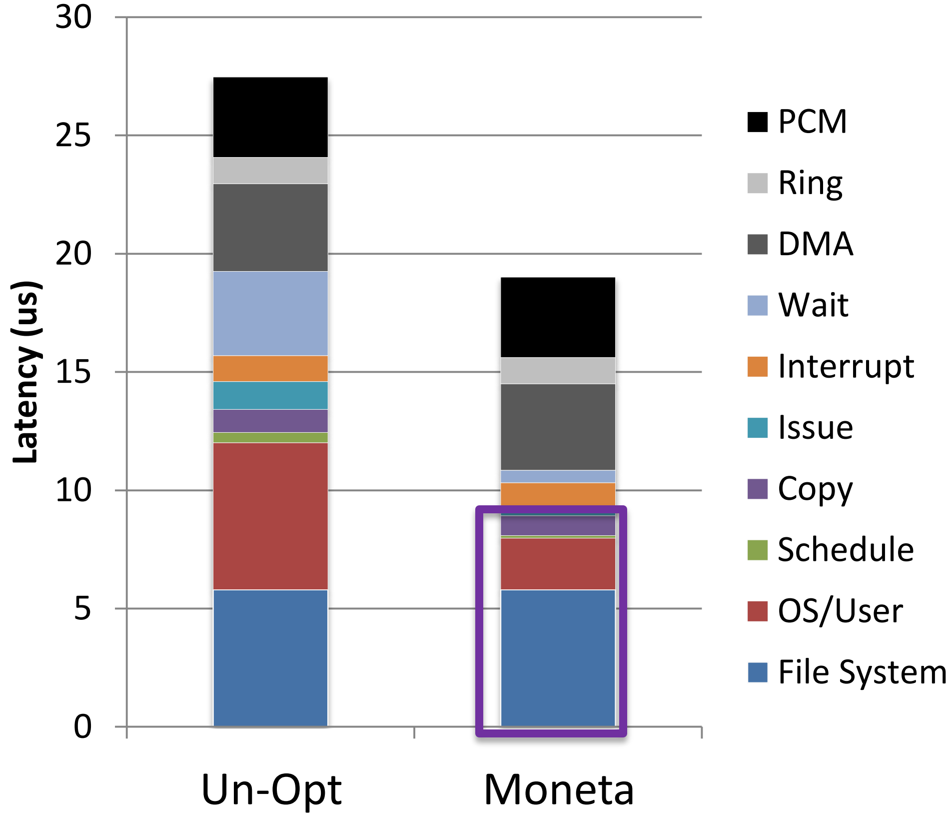
File System Impact

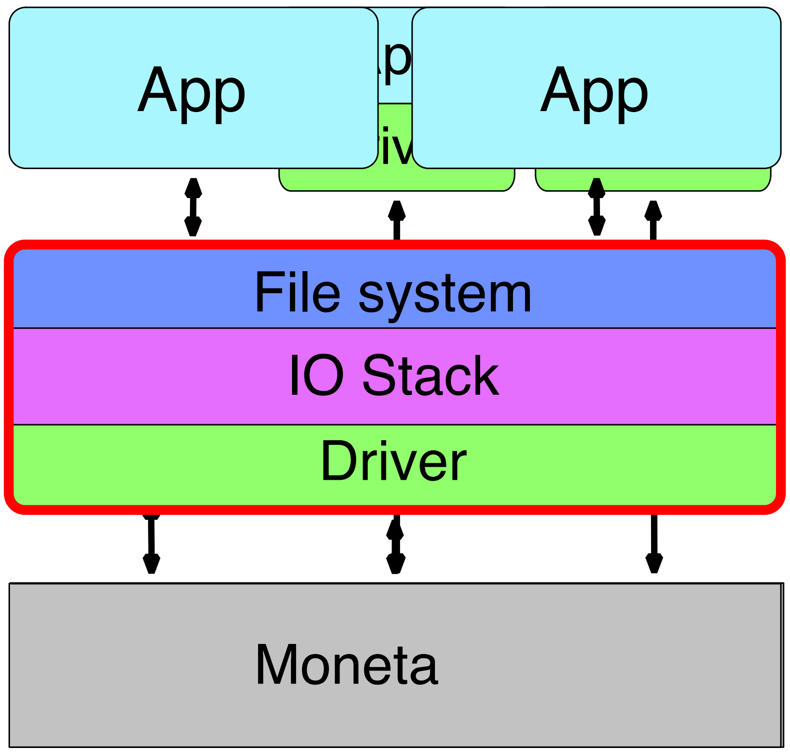




NVSL
Non-volatile Systems Laboratory

Eliminating FS and OS overheads



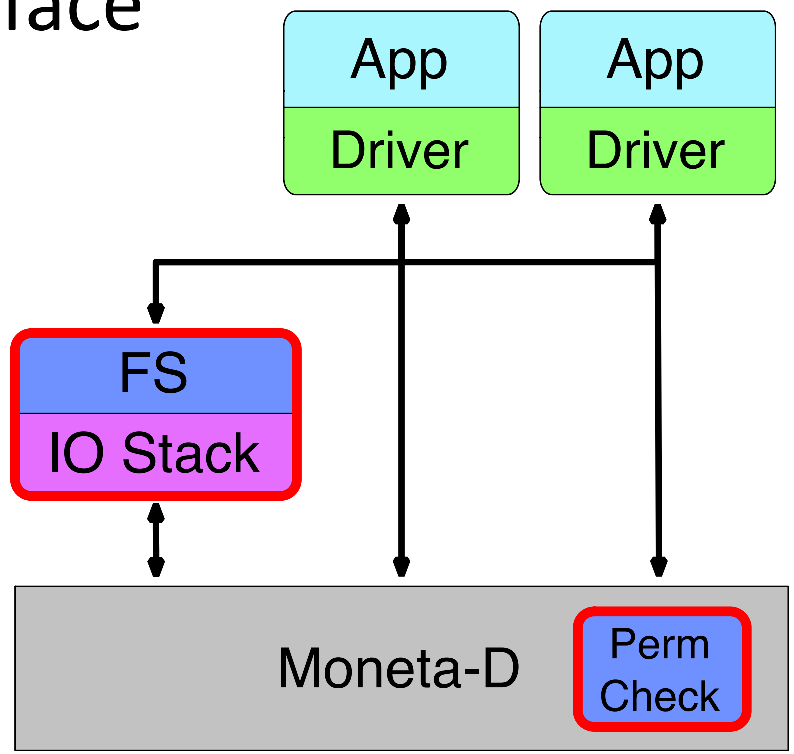


41% total latency, 73% software is due to support protection and sharing

- Separate protection mechanism from policy
  - Move protection checks to hardware
  - Allow applications to access Moneta directly

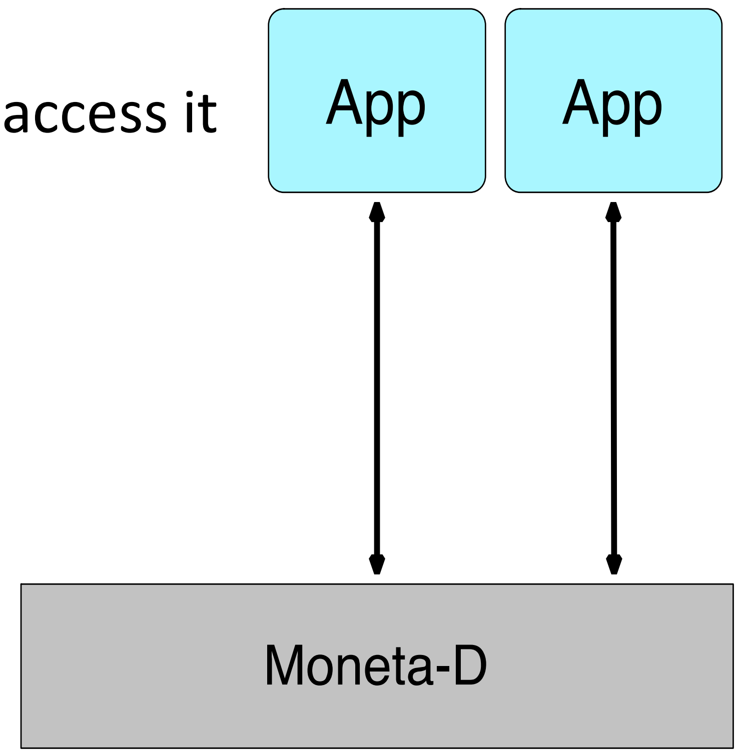
Removing Protection Overheads

1. Virtualized Moneta interface
2. User space library
3. Protection enforcement
4. Changes to the OS



Moneta-D's Virtualized Interface

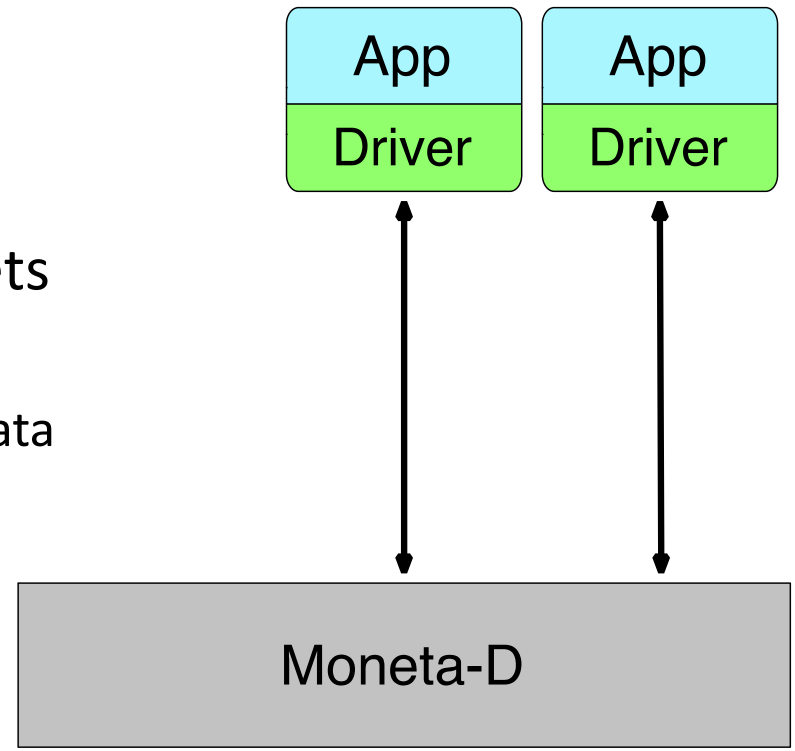
- Virtualize the *interface*, not the device
  - Only one device
  - Many, independent "channels" to access it
- Channel components
  - Unique PCIe address mapping
  - Control registers
  - Request tags
  - Interrupts
  - DMA buffers
- Support 1000 channels
  - This is not a "boutique" interface



NVSL
Non-volatile Systems Laboratory

The User Space Library: LibMoneta

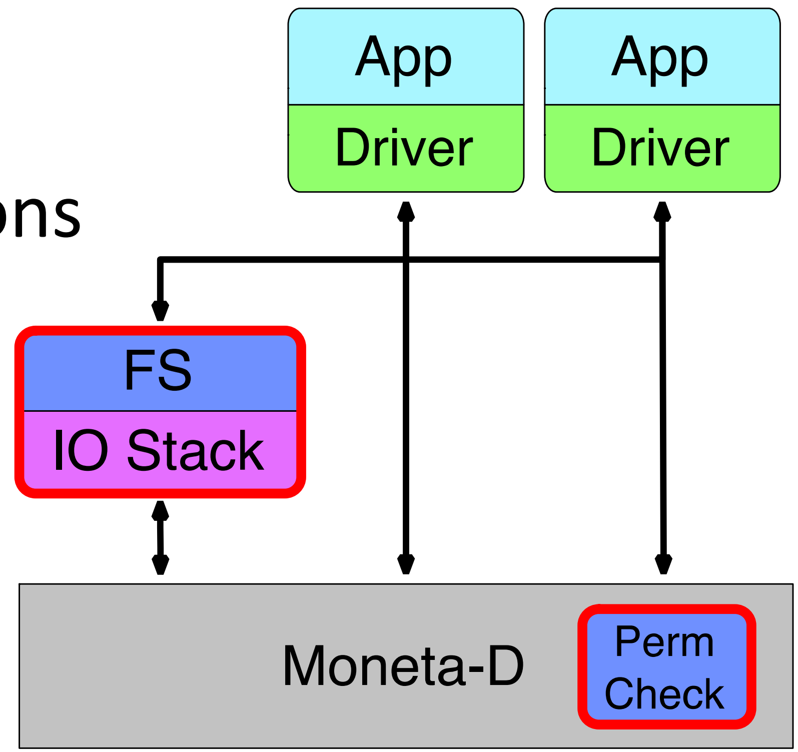
- Transparently intercept FS calls
  - No application changes
- Provides OS functionality
  - **File system:** Translate file offsets to physical storage locations
    - Retrieve and cache translation data via a system call
    - Retry if hardware signals failure
  - **OS:** POSIX compatibility
  - **Driver:** Issue and complete hardware requests



NVSL
Non-volatile Systems Laboratory

Enforcing Protection

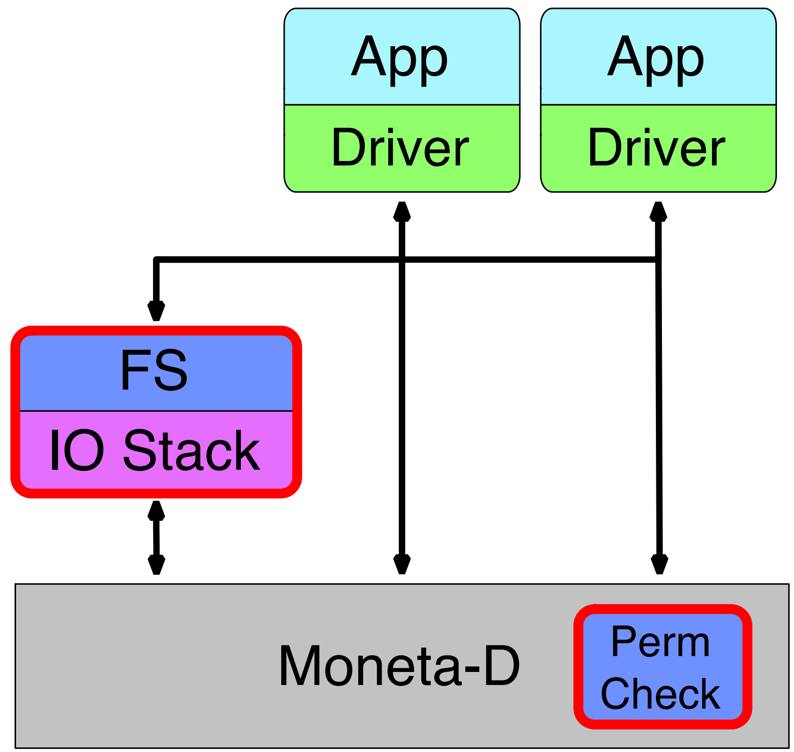
- File system still sets policy
  - User space asks OS driver to update permissions table
- Hardware caches permissions
  - Moneta checks on access
- The permission table
  - Extents based
  - Per channel mappings
  - 16K entries shared between channels



NVSL
Non-volatile Systems Laboratory

Operating System Changes

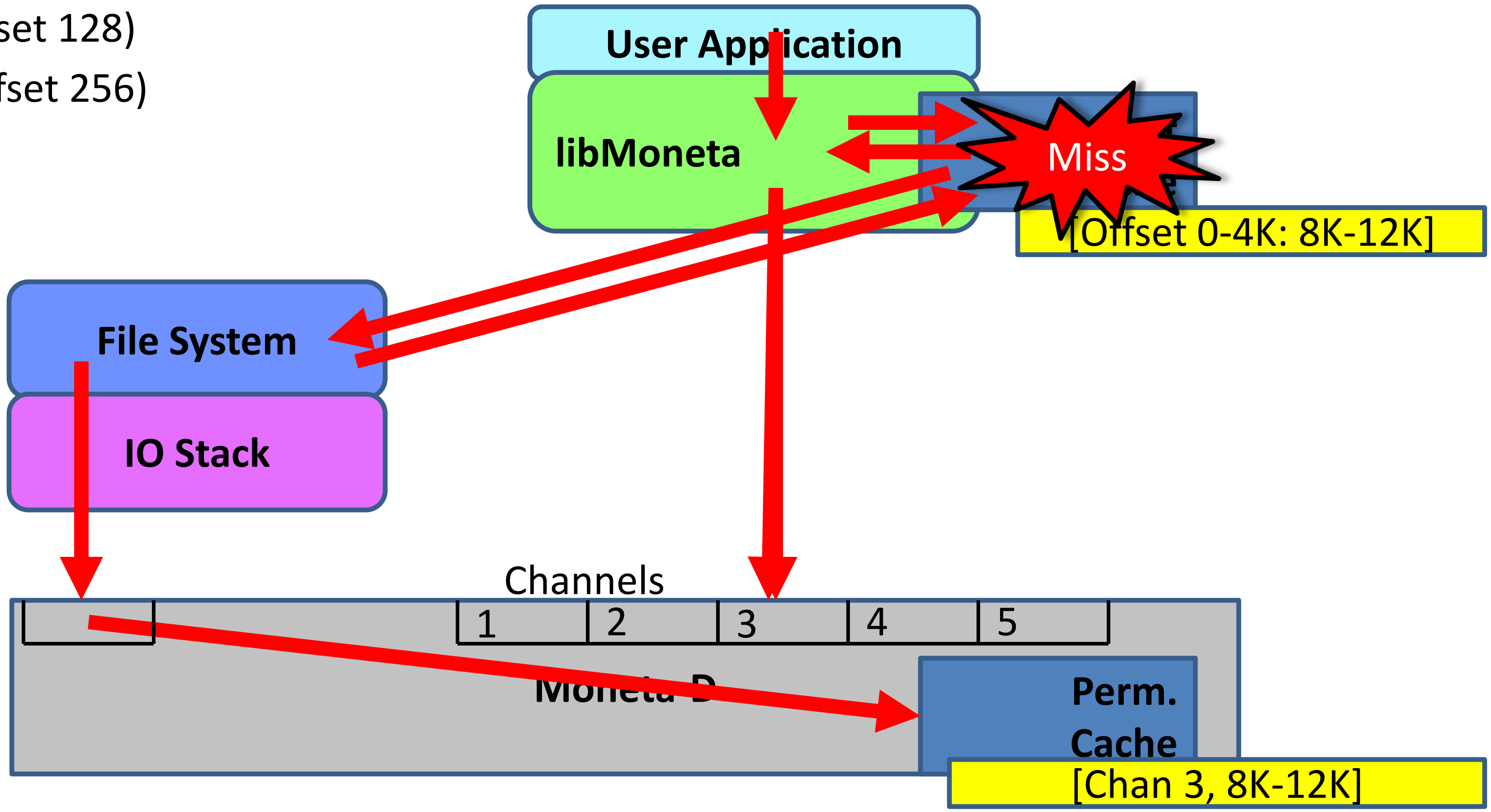
- Small changes to XFS (194 lines)
  - To extract extent details
- Some open questions
  - LibMoneta and the block cache can't see each other
  - File fragmentation



Request Example

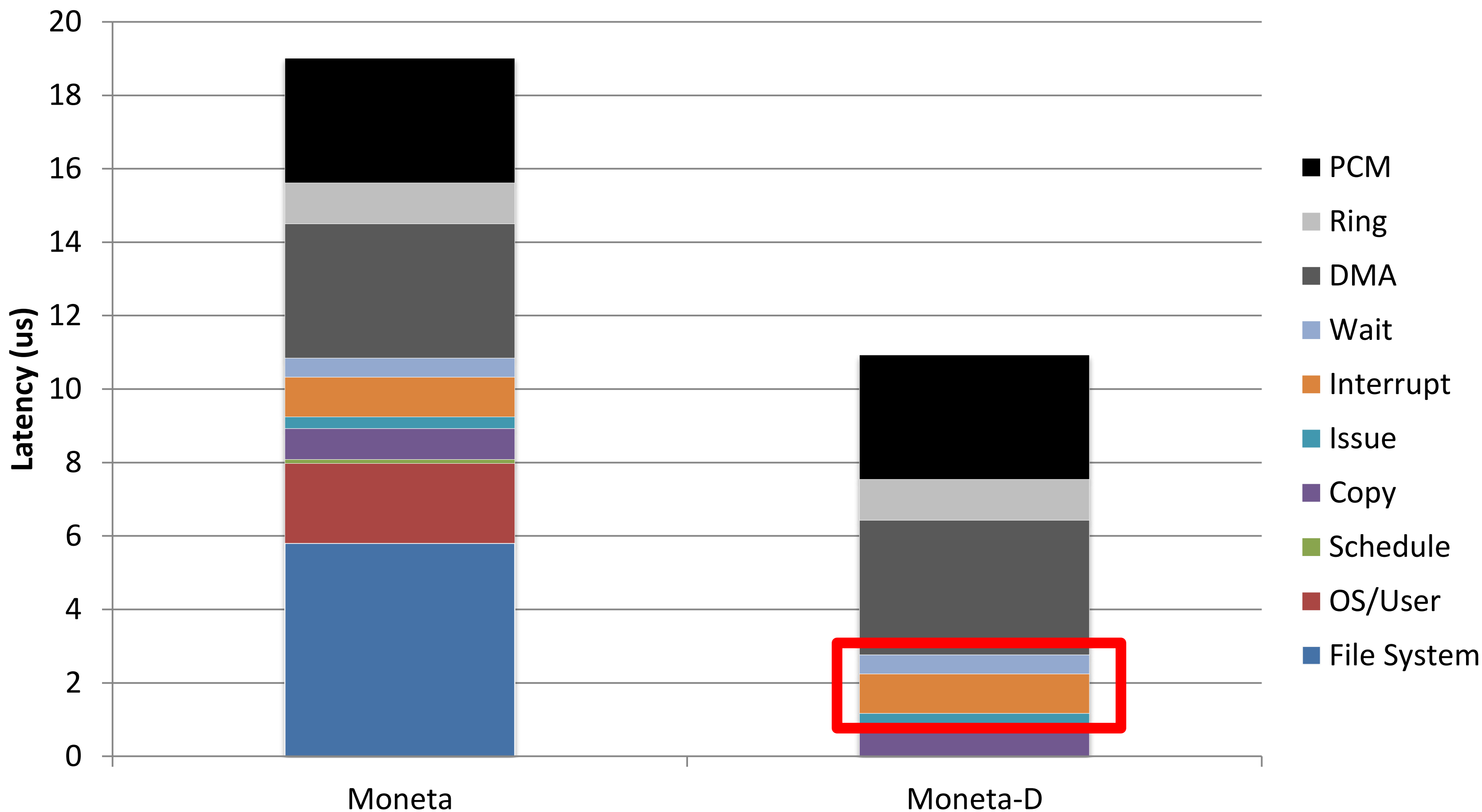
read(offset 128)

write(offset 256)



NVSL
Non-volatile Systems Laboratory

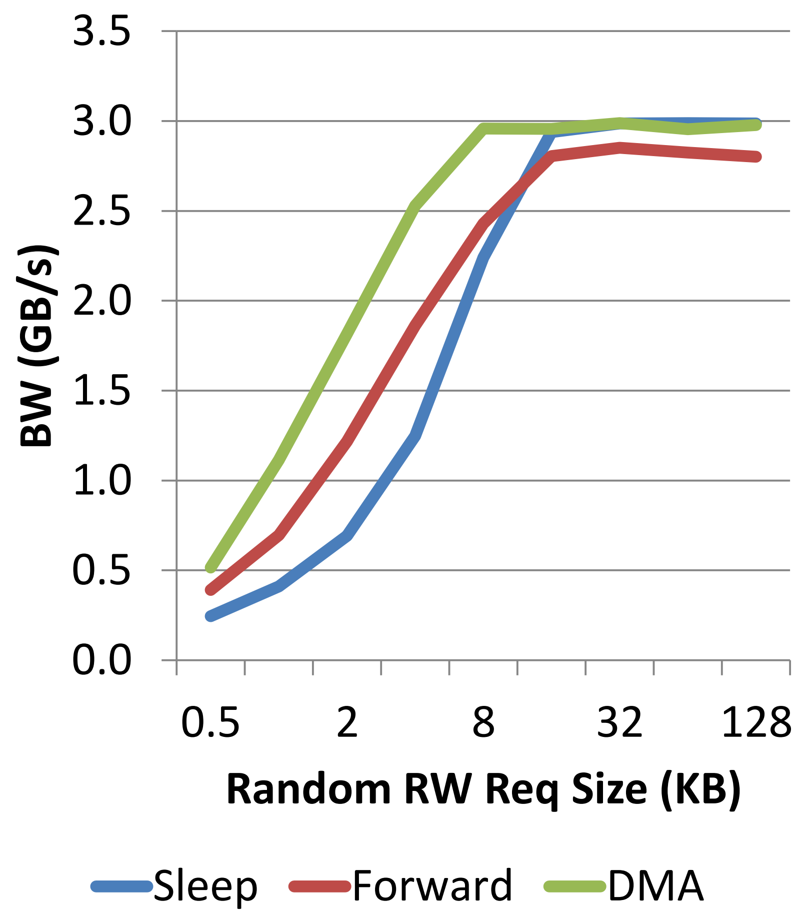
Latency Improvements



NVSL
Non-volatile Systems Laboratory

Completing Requests

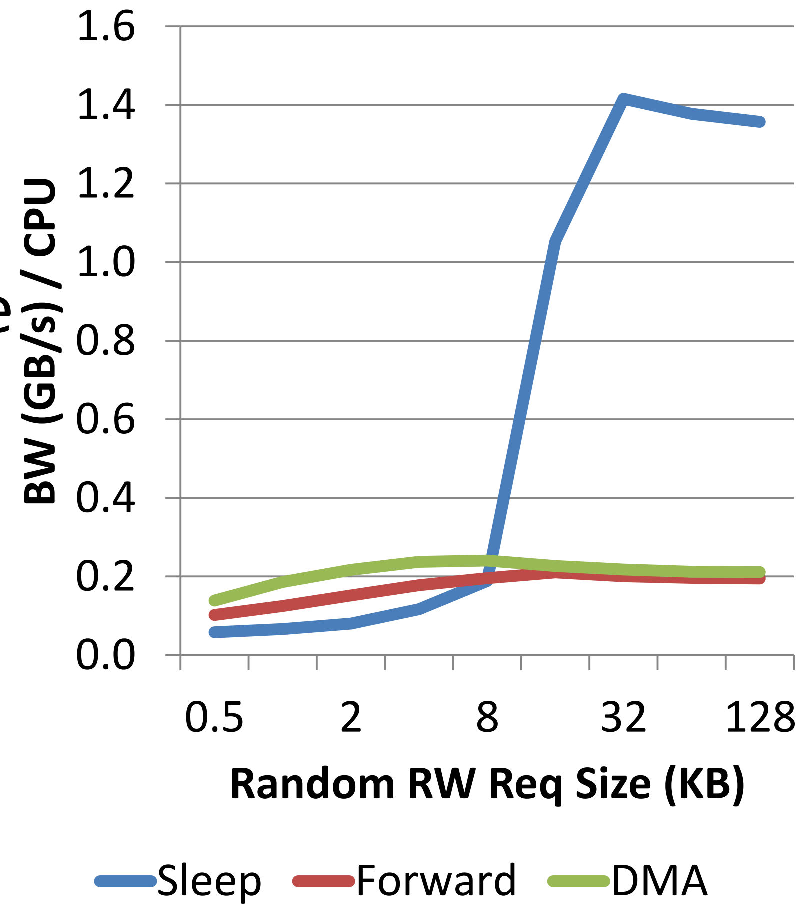
- DMA
  - Signal completion directly to spinning thread via DMA
- Forward
  - Kernel interrupt, Kernel signals spinning thread through mapped page
- Sleep
  - Sleeps after request issue
  - Kernel interrupt, kernel wakes up sleeping thread
  - **7x more latency than DMA or Forward**



NVSL
Non-volatile Systems Laboratory

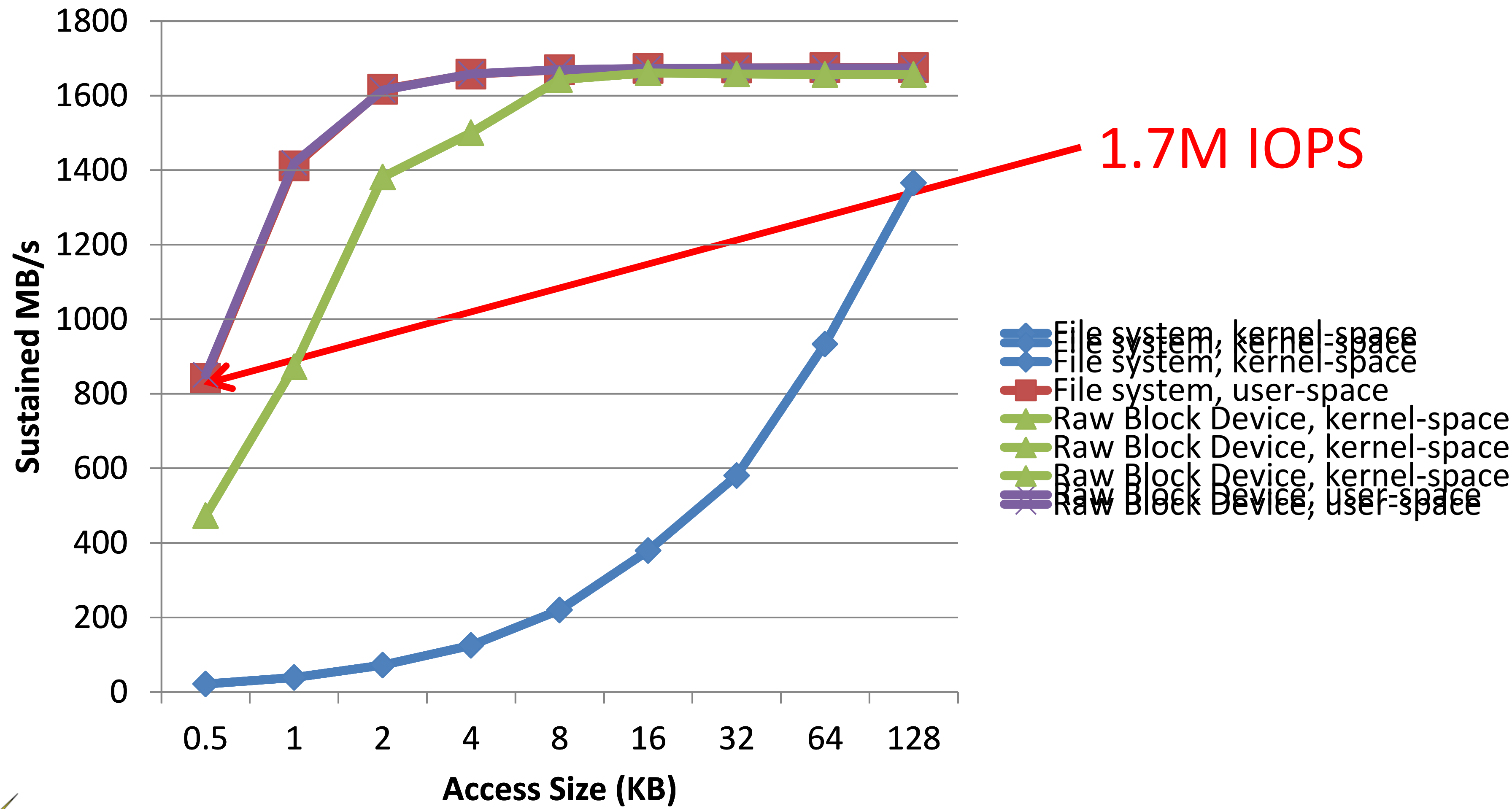
Completion Efficiency

- Desirable to handle more load per CPU core
- DMA and Forward spin while waiting on request
- Sleep: primitive async.
  - Large context switch overhead, only good for large requests
  - Need sufficient number of threads



NVSL
Non-volatile Systems Laboratory

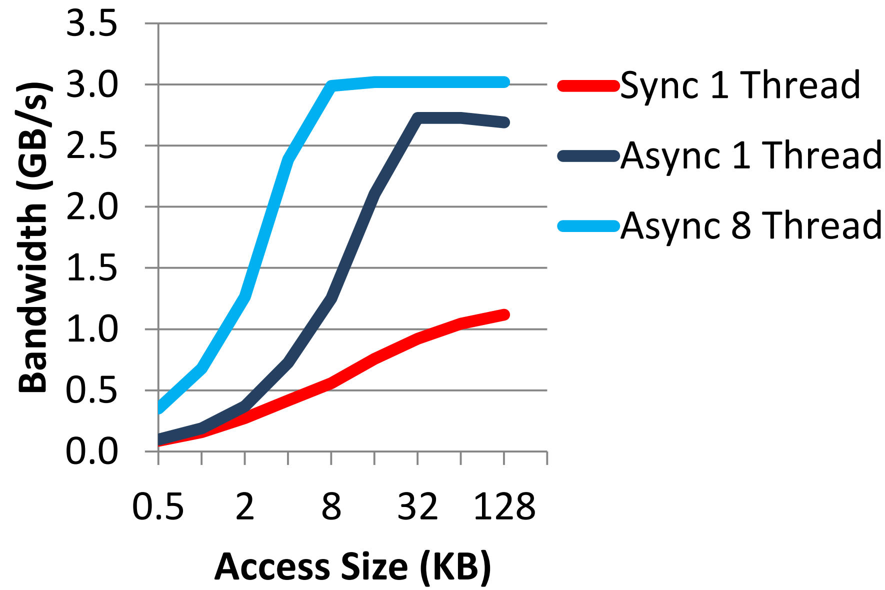
Raw Performance Impact (Writes)



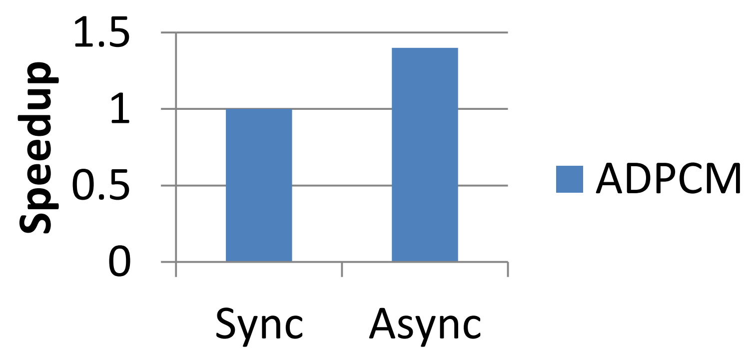
NVSL
Non-volatile Systems Laboratory

Asynchronous Interface

- libMoneta provides Async. Interface also
- Overlap requests for better parallelism
- Requires app changes



- 3.0x with 1 thread, 32 KB
- 1.4x gain in ADPCM decode from MediaBench



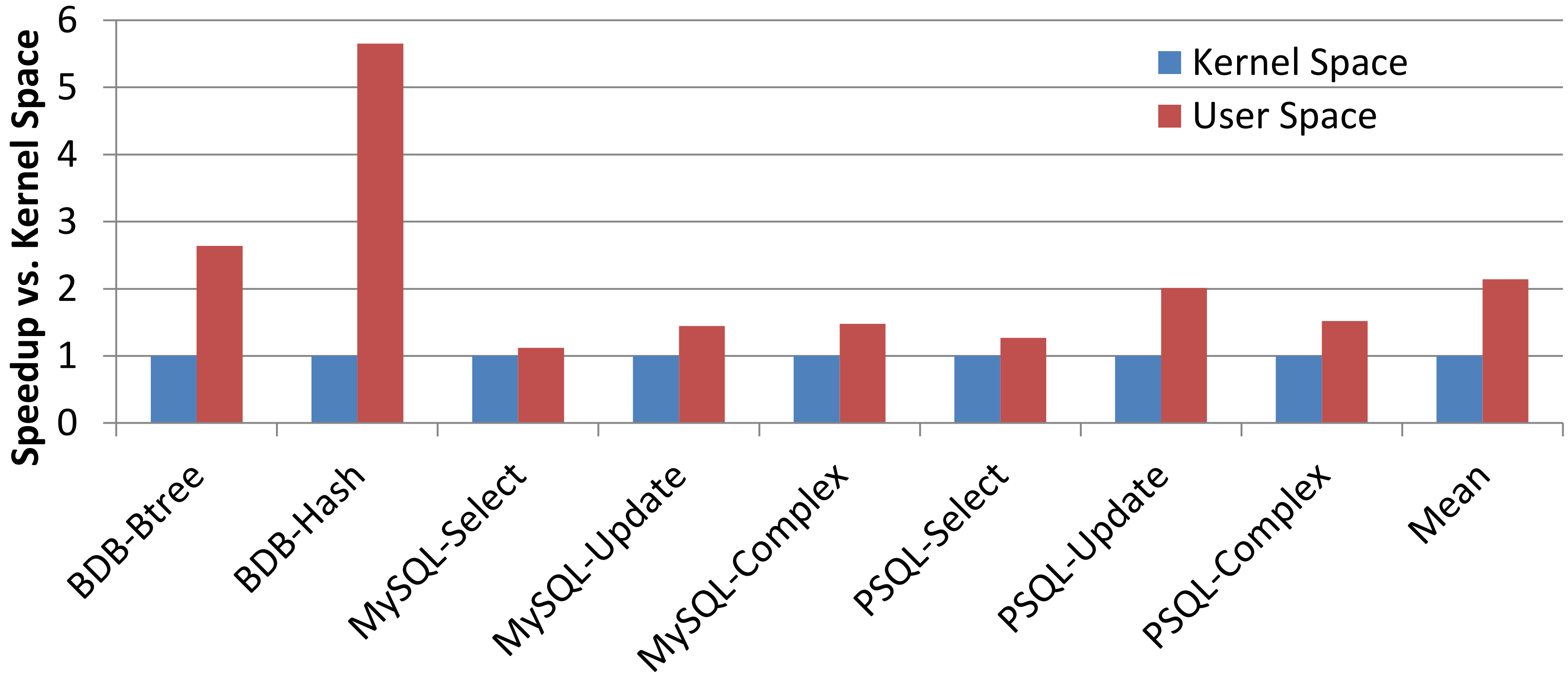
NVSL
Non-volatile Systems Laboratory

Workloads

Name	Footprint	Description
Berkeley-DB Btree	45 GB	Transactional updates to btree key/value store
Berkeley-DB HashTable	41 GB	Transactional updates to hash table key/value store
MySQL-*	46 GB	Random select, update, and complex transaction queries to MySQL database
PGSQL-*	55 GB	Random select, update, and complex transaction queries to Postgres database

NVSL
Non-volatile Systems Laboratory

Application Level Gains

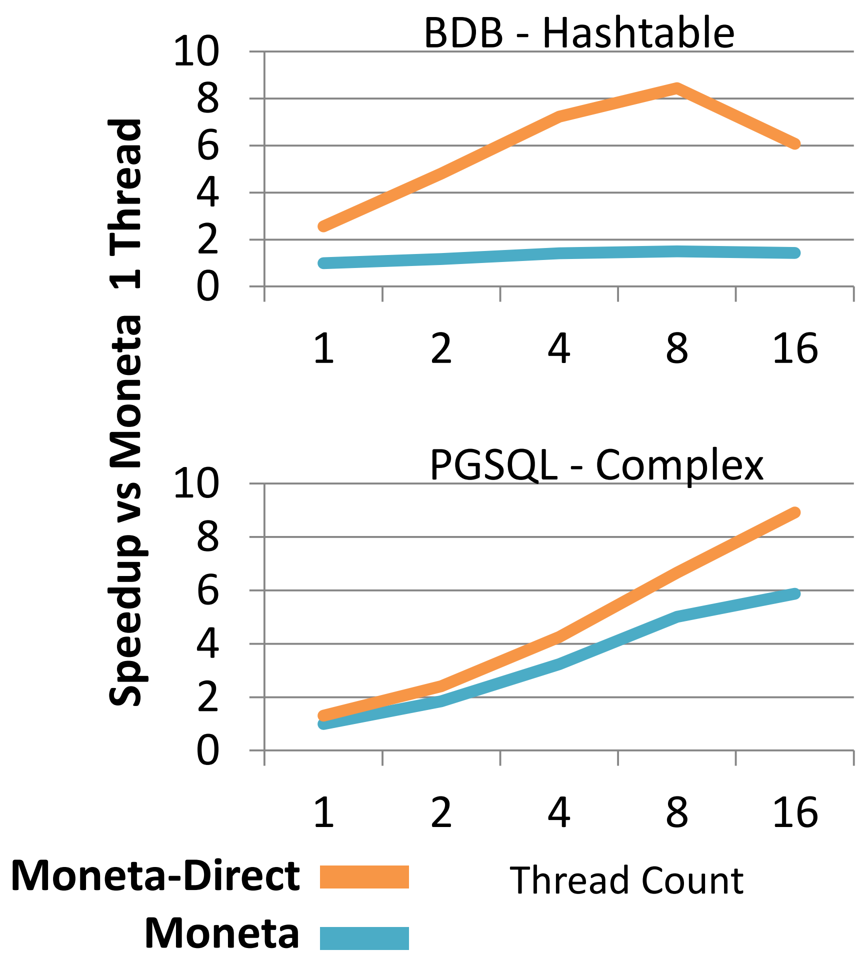


- **No Application Changes**
- Heavy optimization for disks hurts performance in SQL Apps
  - Application optimization should address this

NVSL
Non-volatile Systems Laboratory

Increased (MB/s)/CPU

- 50% less Compute/IO
- Reduced IO power
- Improved Scaling



NVSL
Non-volatile Systems Laboratory

Conclusion: Moneta-Direct

- Virtualized storage interface
  - Direct, user-space access
- Separate protection policy from checking
- Eliminates FS/OS overhead for most accesses
- Improves application performance
  - Up to 5.5x application level performance gain
  - 50% Compute/IO savings

Thank You!

Any Questions?

NVSL
Non-volatile Systems Laboratory

UCSDCSE
Computer Science and Engineering