Session 302-A: PCIe Storage - 2 (PCIe Storage Track)

Flash Drives: Block Storage or Memory?

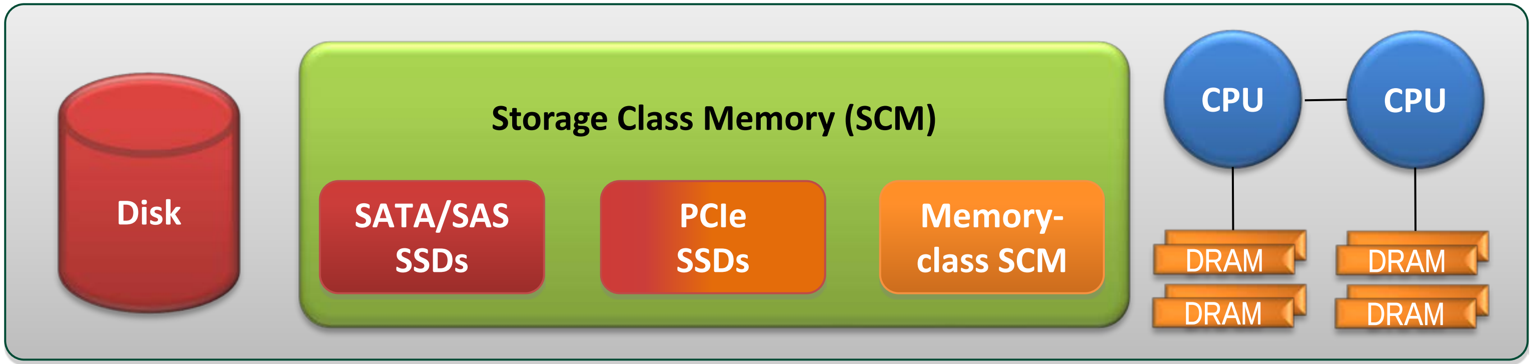
Tony Roug, Virident

Director Solution Architecture

Agenda

- Storage Class Memory: The Bridge
- Access Models: The Architecture
- Access Capabilities: The Value

Storage Class Memory



Disk

Storage Class Memory (SCM)

SATA/SAS SSDs

PCIe SSDs

Memory-class SCM

CPU

CPU

DRAM

DRAM

DRAM

DRAM

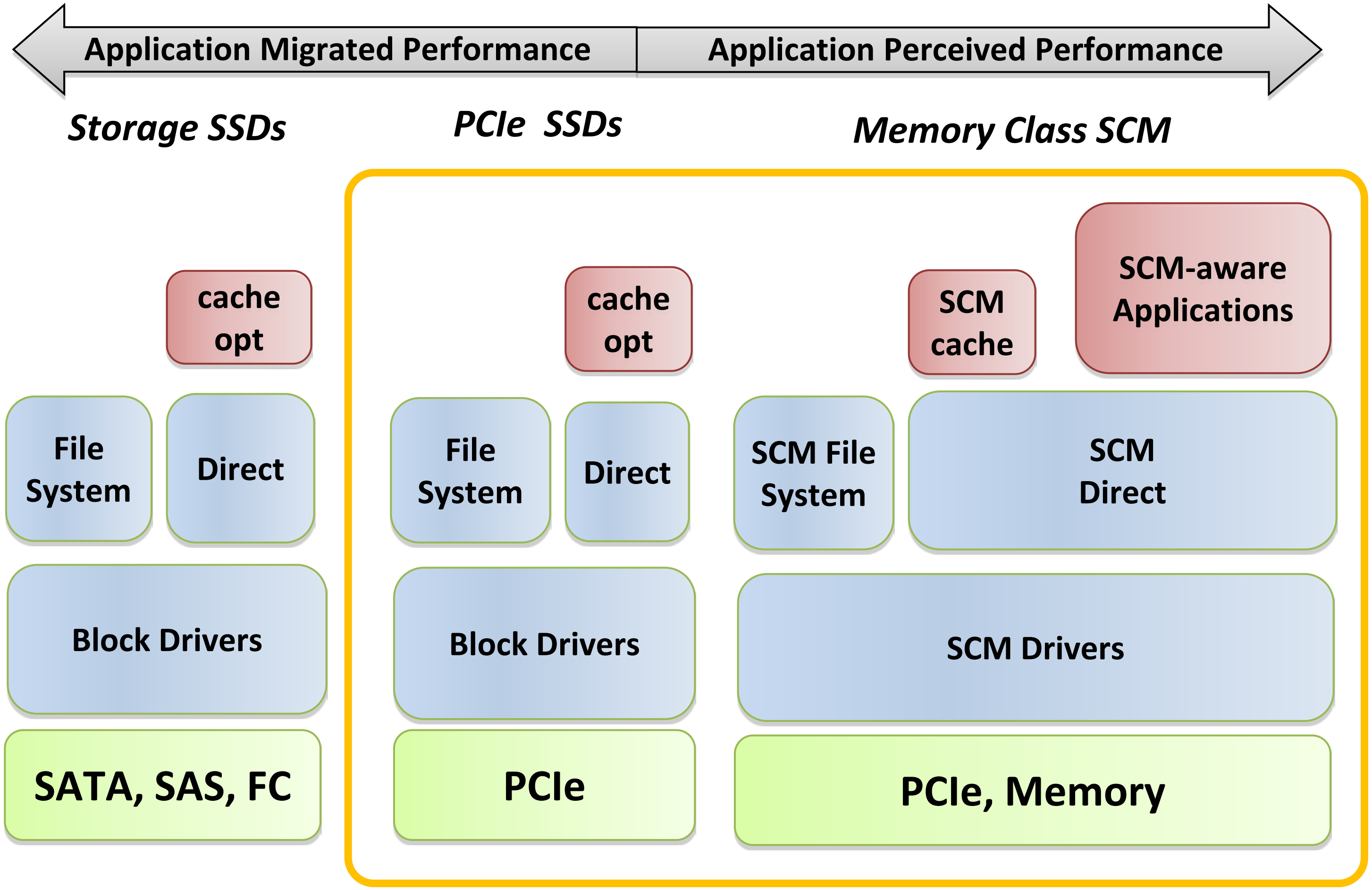
Attribute	Disk	SATA/SAS SSDs	PCIe SSD	*Memory-class* SCM	DRAM
Capacity (GB)	100's-1000's	100's	100's-1000's	100's-1000's	10's-100's
Read performance	10's ms, ~100 MB/s	100's us, ~100's MB/s	**50 – 75 us, 2 – 4 GB/s**	**200 – 400 ns, 2 – 5 GB/s**	~100 ns, 10's GB/s
Write performance	10's ms ~100 MB/s	100's us ~100 MB/s	~25 – 50 us, ~0.5 – 1+ GB/s	~5 us, 2-5 GB/s	

PCIe SSDs established, Memory-class emerging

Agenda

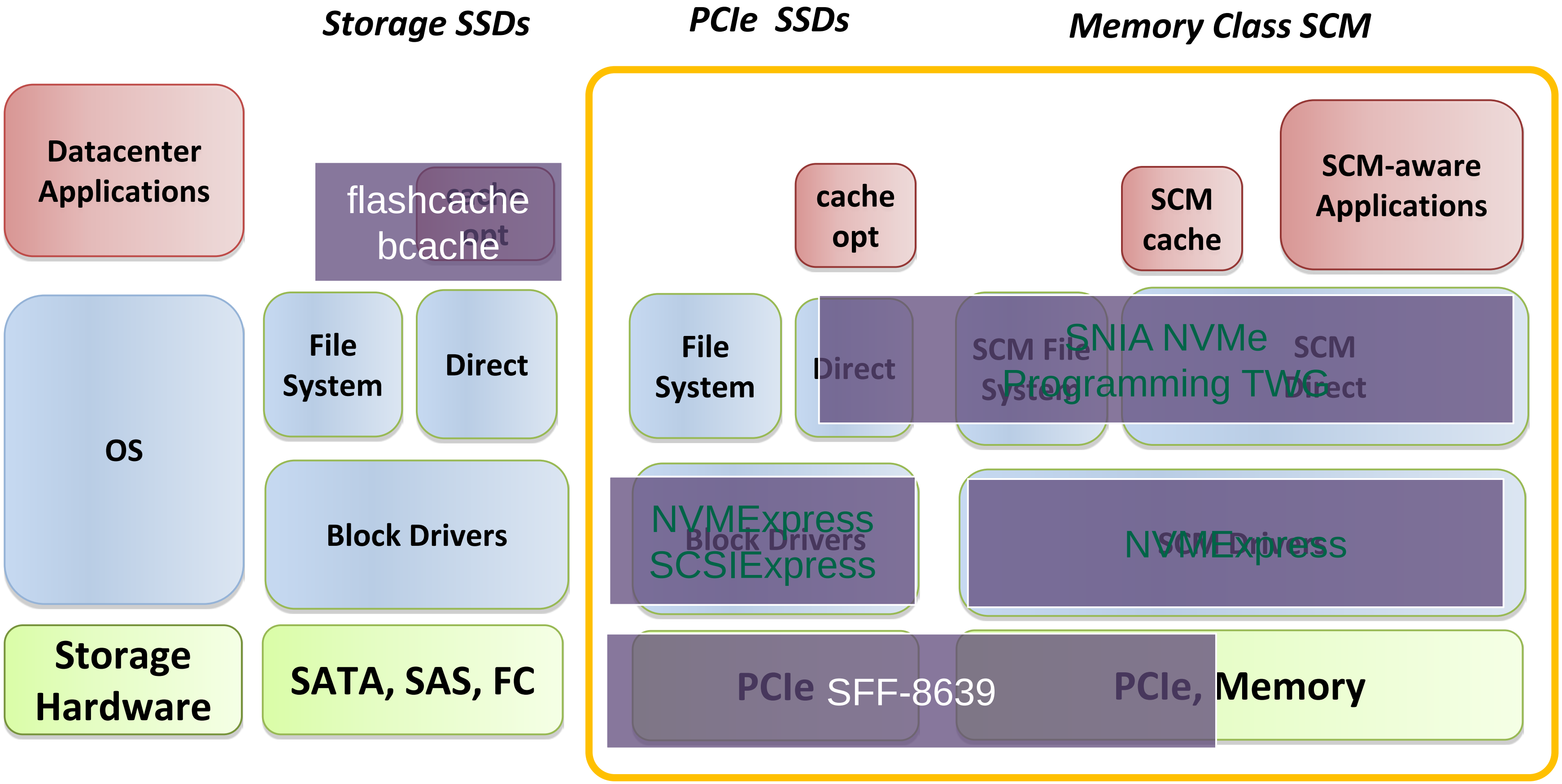
- Storage Class Memory: The Bridge
- Access Models: The Architecture
- Access Capabilities: The Value

Block versus Memory access



Optimization required for applications to realize memory class benefit

Open industry directions



Storage SSDs

PCIe SSDs

Memory Class SCM

Datacenter Applications

flashcache
bcache

cache opt

cache opt

SCM cache

SCM-aware Applications

OS

File System

Direct

File System

Direct

SCM File System

SNIA NVMe Programming TWG

SCM Direct

Block Drivers

NVMExpress
SCSIExpress

Block Drivers

NVMExpress

SCM Drivers

Storage Hardware

SATA, SAS, FC

PCIe

SFF-8639

PCIe, Memory

Industry agreement for optimized architecture in place

Agenda

- Storage Class Memory: The Bridge
- Access Models: The Architecture
- Access Capabilities: The Value

Example PCIe Block Capabilities: Virident FlashMAX

User Level

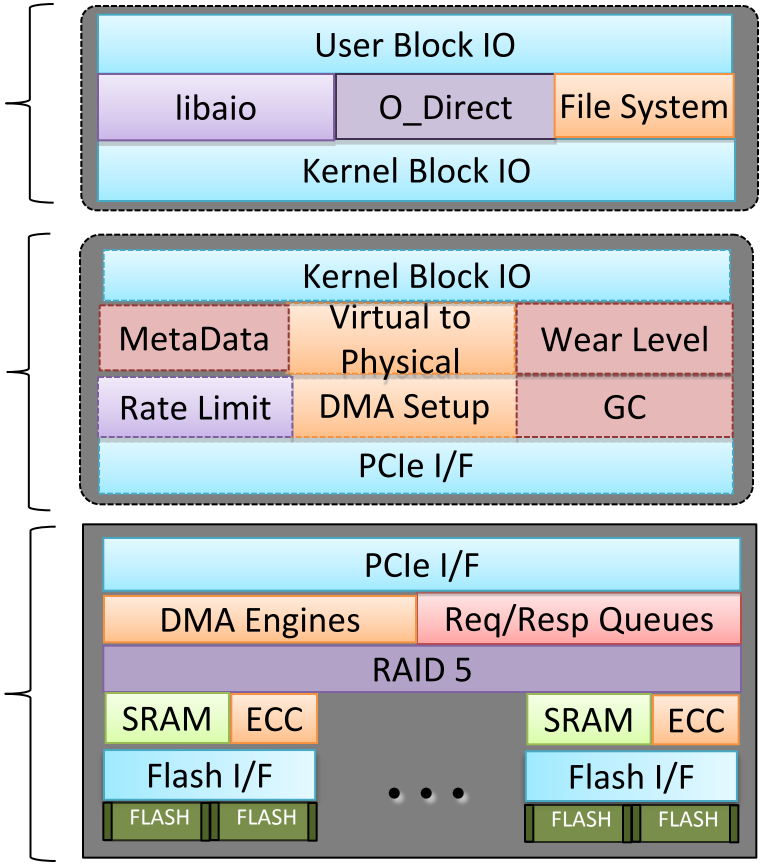
- **File open** and **close, file read** and **write, raw read** and **write**
- Multiple threads through **libaio**
- Direct DMA from user to board/board to user through **o_direct**
- CLI: **monitor, diag, beacon, secure erase**

Kernel Level

- block register: **register, unregister, open, release, ioctrl**
- block transfer: **request, queue**
- DMA scatter/gather: **blk_rq_map_sg**
- rate limit: guaranteed minimum performance at 100% capacity
- wear leveling: identification/adaptation to app access patterns

Board Level

- partition options: **512 and 4K sector size**
- write acceleration mode: **maxperformance, maxcapacity**
- reliability option: **raid5 enabled/disabled**



PCIe SSD optimizations within a block level interface

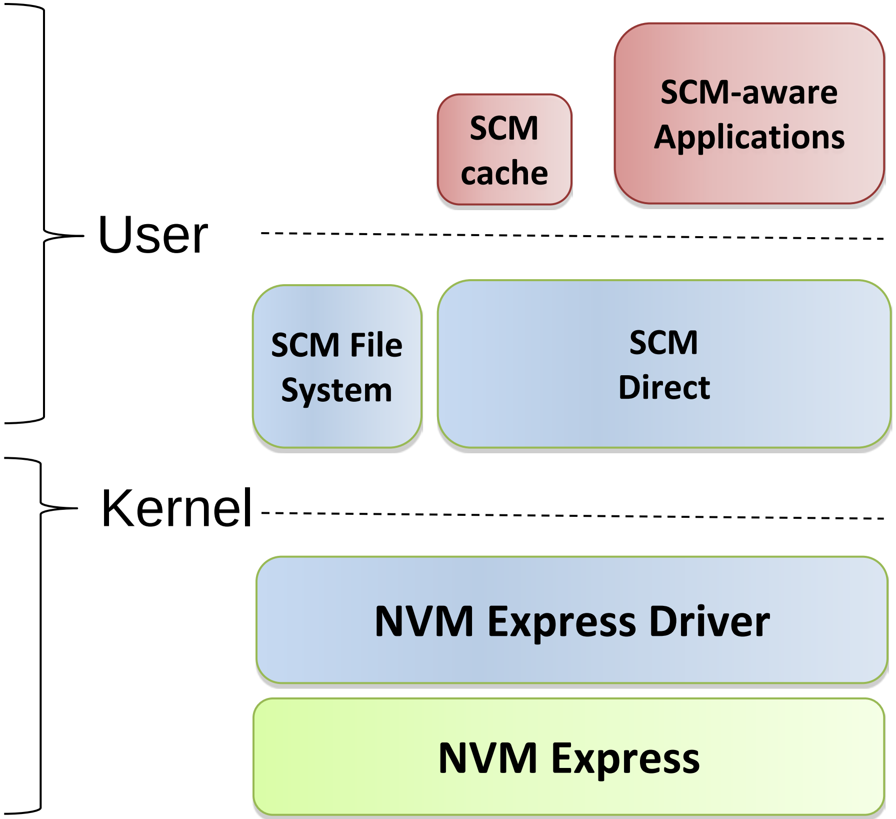
NVM Express: “memory class” capabilities?

Function	NVMExpress 1.0 Capability	Block Compatibility
Basic Features	Read, Write	Read, Write
	Flush	Sync Cache
	Format	Format
ACID Operations	Fused Operations	Compare and Write
	Atomic Write Unit	New ACID Capabilities
	Write Atomicity	
Performance, Endurance, and Efficiency Features	Queues/PRP	New Performance Capabilities
	Data Set Management: Incompressible, Access Size, Write Prepared, Sequential Read/Write, Access Freq, Access Latency	
	Compare	
	Namespaces	
	Deallocate	TRIM/UNMAP
	Power Management	Start Stop
Data protection	End-to-end Protection + Metadata	DIF/DIX
	Format Secure Erase	Secure Erase
	Write uncorrectable	Write Long
	Write protect	WRPROTECT

New capabilities make “memory class” access possible, but not easy

Enabling SCM: Making it easy

1. **streamline the application access path**: e.g. RDMA access
2. **offer improved semantics**: e.g. atomic-write
3. **simplify caching implementations**: e.g. key-value store,
4. **construct cluster-level solutions**: e.g. synchronize cross-server activities and orchestrate server to SAN data movement
5. **emulate memory**: e.g. mmap and malloc
6. **expose internals of flash management**: e.g. optimized file system with snapshot, cross-card data transfers, trading capacity vs performance



Create industry standard capabilities enabling memory class access

Summary

- Storage Class Memory:
  - First level value enabled by block mode
- Access Models:
  - NVM Express enables memory class usage models
- Access Capabilities:
  - Industry agreement on application enabling technology launched
  - SNIA Programming TWG

Thanks

Tony Roug
Phone: 4085829647
Email: tonyr@virident.com