STORAGE INDUSTRY
SUMMIT

Convergence of Storage and Memory Developing the Needed Ecosystem

JANUARY 20, 2016, SAN JOSE, CA

Neal Christiansen

Microsoft

Principal Development Lead

Storage Class Memory in Windows

SNIA SSSI | SOLID STATE STORAGE

What is "Storage Class Memory"?

- Non-volatile storage medium with RAM-like performance
  - Low latency/high bandwidth.
- Resides on the memory bus
- Several different terms in use:
  - Storage Class Memory (**SCM**)
  - Direct Access Storage (**DAS**)
  - Byte Addressable Storage (BAS)
  - Persistent Memory (PM)
  - Non-Volatile Memory (NVM)

File Systems and Storage Class Memory

- **SCM is a disruptive technology**

- Customers want the fastest performance
- System software is in the way!
- Customers want application compatibility
- Conflicting goals

Windows Goals for Storage Class Memory

- Support zero-copy access to persistent memory
- Most existing user-mode applications will run without modification
- Provide an option to support 100% backward compatibility
  - Does Introduce new types of failure modes
- Make available sector granular failure modes for application compatibility

Introducing a New Class of Volume

- Direct Access Storage (DAS) Volume
  - Memory mapped files will provide applications with direct access to byte-addressable SCM
    - Maximizes performance
  - DAS mode is chosen at volume format time
    - Why: compatibility issues with various components, examples:
      - File system filters
      - Bitlocker
      - Volsnap
  - Some existing functionality is lost
  - DAS Volumes will be supported by both the NTFS and ReFS file systems

SCM Storage Drivers

- New type of volume requires a new driver model
  - SCM Bus Driver
    - Enumerates the physical and logical SCM devices on the system
    - Not part of the IO Path
  - SCM Disk Drivers
    - Driver for logical SCM devices
    - Storage abstraction layer to rest of the OS
    - Hardware-specific
      - Supports both in-box or vendor-specific drivers
    - Windows will use a native 4K sector size
- Introduces new interfaces
  - Expose byte addressable storage functionality
  - Supports management of SCM hardware

Memory Mapped IO in DAS mode

- On DAS formatted volumes memory mapped sections map directly to SCM hardware
  - No change to existing memory mapping APIs
- When an application creates a memory mapped section:
  - The memory manager (MM) asks the File System if the section should be created in DAS mode (Direct Access Storage)
  - The file system returns YES when:
    - The volume resides on SCM hardware
    - The volume has been formatted for byte addressable mode

Memory Mapped IO in DAS mode

- When a section is created in DAS mode
  - MM asks the file system for the physical memory ranges for a given range of the file
  - The file system translates the range into one or more volume relative extents (sector offset and length)
  - The file system then asks the storage stack to translate these extents into physical memory ranges
  - MM then updates its paging tables for the section to map directly to the persistent storage

Memory Mapped IO in DAS mode

- This is true zero-copy access to storage
  - An application has direct access to persistent memory

- Important → No paging reads or paging writes will be generated

Cached IO in DAS mode

- When cached IO is requested on a DAS enabled volume the cache manager creates a cache map that maps directly to SCM hardware
- Cache manager copies directly between the user's buffer and persistent memory
  - Cached IO has one-copy access to persistent storage
- Cached IO is coherent with memory mapped IO
- As in memory mapped IO, no paging reads or paging writes are generated
  - No Cache Manager Lazy Writer thread

Non-cached IO in DAS Mode

- Sends IO operations down the storage stack to the SCM storage driver
  - Maintains existing failure semantics for application compatibility
  - Is coherent with cached and memory mapped IO

File System Metadata in DAS Mode

- File system metadata will not use DAS mode sections
  - Meaning paging reads/writes will be generated for all file system metadata operations
  - Needed to maintain existing ordered write guarantees for write-ahead logging
- One or more metadata files may use DAS mode in the future

Impacts to File System Functionality in DAS Mode

- Direct access to persistent memory by applications eliminates the traditional hook points that file systems use to implement various features
- File System functionality that can not be supported on DAS enabled volumes:
  - No NTFS encryption support (EFS)
  - No NTFS compression support
  - No NTFS TxF support
  - No NTFS USN range tracking of memory mapped files
  - No NTFS resident file support
  - No ReFS integrity stream support
  - No ReFS cluster band support
  - No ReFS block cloning support

Impacts to File System Functionality in DAS Mode

- The file system no longer knows when a writeable memory mapped sections are modified
  - The following file system features are now updated at the time a writeable mapped section is created
    - File’s modification and access times
    - Marking the file as modified in the USN Journal (change journal)
    - Signaling directory change notification

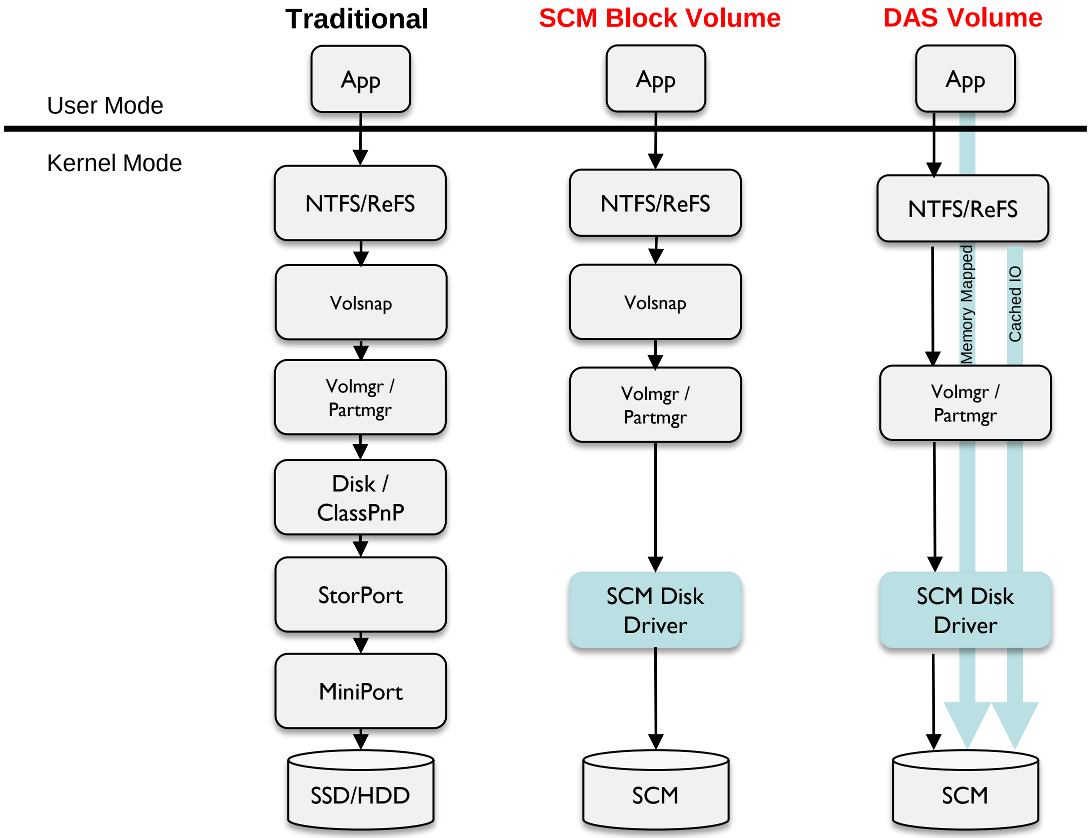
Backward Compatibility with SCM Hardware

- Block Mode Volumes
  - Maintains existing storage semantics
    - All IO operations traverse the storage stack to the SCM driver
    - Has a shortened path length through the storage stack
      - No storport or miniport drivers (too much latency)
      - No SCSI translations
  - Fully compatible with existing applications
  - Supported by all Windows file systems
  - Works with existing file system and storage filters
  - Block mode vs. DAS mode is chosen at format time

New Volume Device Class (ScmVolume)

- New byte addressable partition type
  - Set at format time
- Why: Prevents non-DAS aware components from attaching to this new volume class
  - VOLSNAP – no support for volume snapshots
  - BITLOCKER – no support for software encryption
  - 3rd Party volume stack filters
  - Improves performance by removing non-DAS aware drivers

IO Stack Comparisons



What is a File System Filter

- A driver that layers above the file system
- Augments file system functionality
  - May interact with all operations as they come into and out of the file system
- Example classes of filters:
  - Anti-virus
  - Replication
  - Hierarchical Storage Management (HSM)
  - Security Enhancer
  - Encryption
  - Compression
  - Quota
  - Activity monitor

File System Filters in DAS Mode

To minimize compatibility issues:

- No existing filters will receive notification when a DAS volume is mounted
- At filter registration time filters will indicate via a new registration flag that they understand DAS mode semantics

Compatibility Issues with Filters in DAS Mode

- Reason: No paging IO
- Data transformation filters (ex: encryption and compression)
  - There is no opportunity for these filters to do their work for memory mapped files
- Anti-virus filters
  - Minimally impacted because scanning is performed at file open and close time
  - Detecting when a file is modified will need to be updated
    - Watch for creation of writeable mapped sections
- Replication filters
  - Difficult to detect when a file has changed
  - Difficult to efficiently track what ranges of a file have been modified

Sector Atomicity

- BTT – Block Translation Table
  - Algorithm created by Intel
  - Provides efficient sector level atomicity of writes
    - Eliminates sub-sector torn writes
    - On power loss either see contents of old sector or new sector
    - Provides compatibility for existing applications that have built-in assumptions around storage failure patterns
    - Minimal performance impact
  - Implemented by remapping the physical SCM address of a given LBA
    - Not compatible with DAS mode memory mapped sections where physical SCM addresses are given to MM

BTT Usage

- Uses small portion of SCM space for mapping tables and control structures
  - BTT structures are not visible outside of the SCM driver
  - BTT control structures are always allocated
- File system controls if a given write should use BTT or not
  - A new per-IO flag has been added to indicate if the given LBA may be remapped or not
    - If CLEAR the given LBA may be remapped (use BTT)
    - If SET the given LBA must **not** be remapped (do not use BTT)

BTT Usage

- Block mode volumes will always indicate that an LBA may be remapped (use BTT)
- NTFS DAS volumes
  - File system metadata writes may be remapped (use BTT)
    - Can not detect sub-sector torn writes
  - All other writes must not be remapped (do not use BTT)
- ReFS DAS Volumes
  - All writes must not be remapped (do not use BTT)
  - Because ReFS uses a copy-on-write model and its metadata is checksummed, ReFS can detect and recover from sub-sector torn writes without BTT

Application use of SCM

- Intel NVML Library
  - Open source library implemented by Intel
    - Available for Linux via GitHub
    - https://github.com/pmem/nvml/
  - Defines a set of application API's for efficient use of SCM hardware
    - Abstracts out OS specific dependencies
    - Underlying implementation uses memory mapped files
      - All access via API calls
        - No direct access to underlying memory mapped files
    - Has its own BTT implementation for atomicity guarantees
    - Works in both SCM and non-SCM hardware environments
  - Microsoft is working with Intel on a Windows port
    - Most functionality is up and running

Overview of NVML Libraries

- **libpmemobj** – transactional object store
- **libpmemblk** – provides arrays of atomically updated fixed size blocks
- **libpmemlog** – atomic append to log
- **libpmem** – low level support for rest of libraries
- **libvmmalloc** – persistent heap

- http://pmem.io/

Conclusions

- SCM is an exciting new technology
- SCM is a disruptive technology
- Performance tradeoffs
  - Significant storage performance improvement without application modification
  - Even better performance improvements possible with application modification
- System software is a barrier to performance

Solid State Storage Initiative

Questions