New Memories

Dongkun Shin (dongkun@skku.edu)
Embedded Software Laboratory
Sungkyunkwan University
http://nyx.skku.ac.kr/

Emerging New Memories

- Non-Volatile ≈ NAND flash
- Byte-addressable ≈ DRAM
- Erase-less Programming

- PCM (Phase Change Memory) – PRAM
- MRAM (Magnetic or Magnetoresistive RAM)
- FeRAM (Ferroelectric RAM) – FRAM
- ReRAM (Resistive RAM)

How to Use

- *As a Storage System*
  - Integrated in the traditional storage software stack
  - e.g., *PCIe-based storage, Hybrid storage*
- *As a Main Memory*
  - DRAM can be regarded as a cache for the NVM to reduce the access latency to the primary memory
- *As a Processor Cache*
  - for the last-level cache

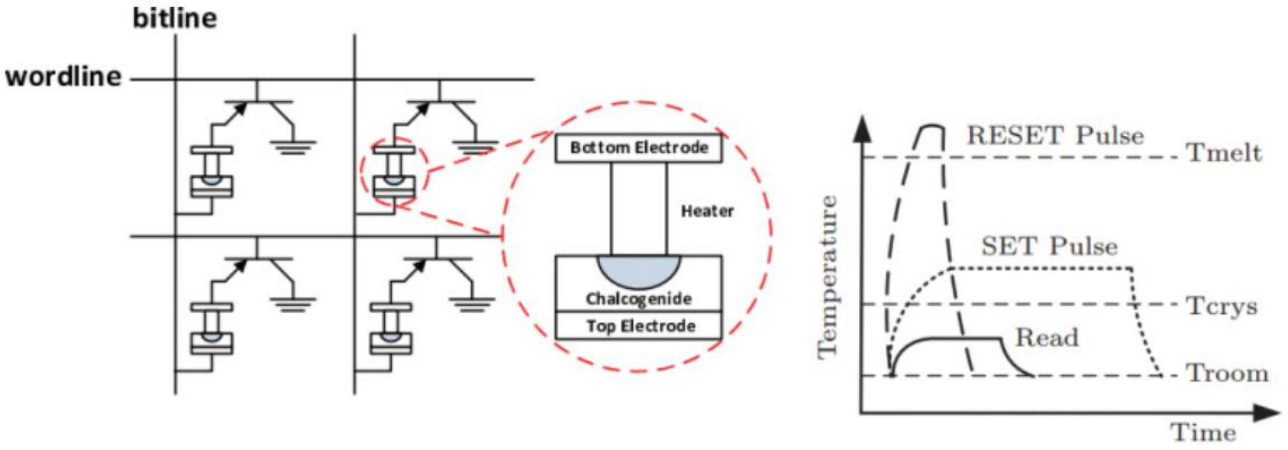
Characteristics of NVMs

	SRAM	DRAM	HDD	NAND flash	STT-RAM	ReRAM	PCM	FeRAM
Cell size (F^2)	120–200	60–100	N/A	4–6	6–50	4–10	4–12	6–40
Write Endurance	10^{16}	$>10^{15}$	$>10^{15}$ (pb: mechanical parts)	10^4–10^5	10^{12}–10^{15}	10^8–10^{11}	10^8–10^9	10^{14}–10^{15}
Read Latency	~0.2–2ns	~10ns	3–5ms	15–35μs	2–35ns	~10ns	20–60ns	20–80ns
Write Latency	~0.2–2ns	~10ns	3–5ms	200–500μs	3–50ns	~50ns	20–150ns	50–75ns
Leakage Power	High	Medium	(Mechanical parts)	Low	Low	Low	Low	Low
Dynamic Energy (R/W)	Low	Medium	(Mechanical parts)	Low	Low/High	Low/High	Medium/ High	Low/High
Maturity	Mature	Mature	Mature	Mature	Test chips	Test chips	Test chips	Manufactured

Emerging NVM: A Survey on Architectural Integration and Research Challenges, TDAES, 2017

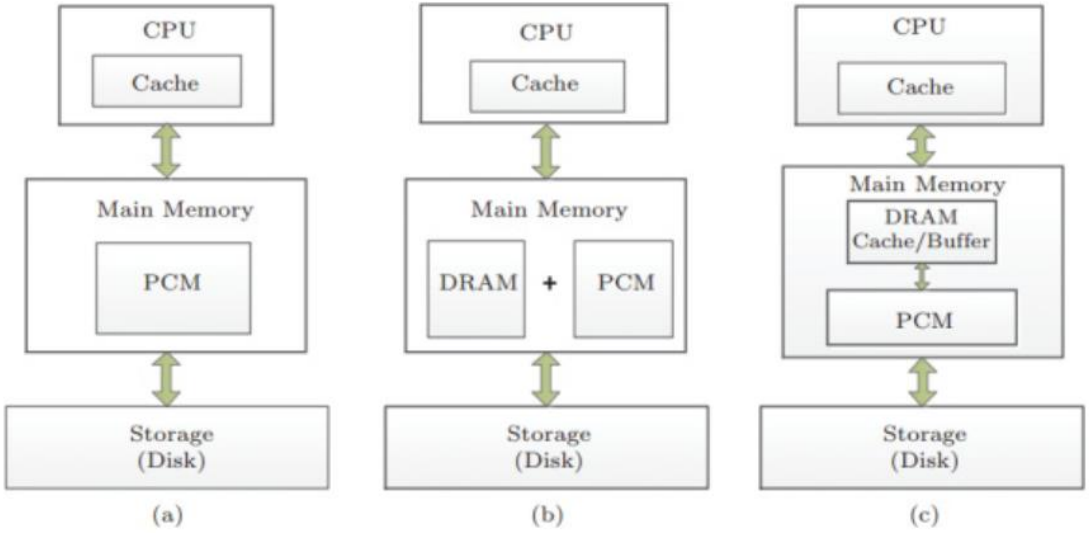
PCM

- A cell uses a thin layer of chalcogenide such as Ge2Sb2Te5 (GST) and two electrodes that wrap the chalcogenide on both sides, in addition to a heater
- Electrically-initiated amorphous-to-crystalline phase-change process
- A short but high voltage pulse ➔ amorphous state (RESET, bit to 0)
  - GST is heated above the melting temperature (Tmelt).
- A long and low voltage ➔ crystalline state (SET, bit to 1).
  - GST is heated above the crystallization temperature (Tcrys) but below Tmelt.
- Write latency > Read latency



PCM

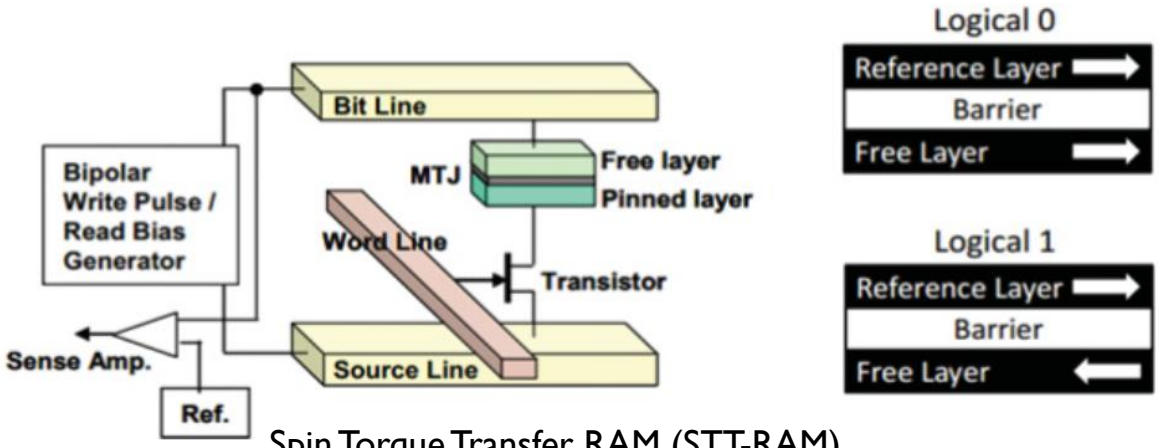
- Possible to store multiple bits (Multi-Level Cells) by making use of intermediate resistive values
- Endurance
  - flash memory (10^5 cycles) < PCM < DRAM (10^{15})
- Can be used in Main memory or Storage



Three Possible Memory Usage

MRAM

- The information carrier is the Magnetic Tunnel Junction (MTJ) instead of electrical charges
- MTJ consists of two ferromagnetic layers separated by an oxide (tunnel) barrier layer.
  - *Reference (or pinned) layer* has a fixed direction of magnetization
  - *Free layer* has a variable direction
  - Parallel state (logical 0): low resistance when the two layers have the same direction of magnetization
  - Anti-parallel state (logical 1): high resistance when magnetization directions are opposite to one another



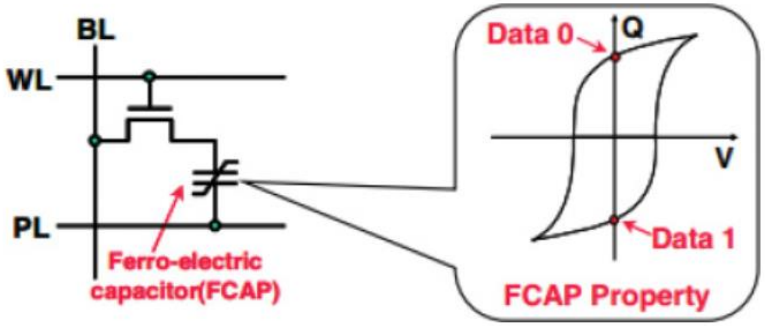
Spin Torque Transfer RAM (STT-RAM)

MRAM

- Write operation
  - The magnetization direction of the free layer changes according to the voltage that is applied between the source line (SL) and the bit line (BL)
- Read operation
  - A very small voltage is applied between SL and BL, which causes a current to flow through the MTJ.
  - The value of this current is relative to the resistance of the MTJ.
- The read performance of STT-RAM is comparable to that of SRAM and in some cases better than that of DRAM
  - Can be used in on-chip cache or main memory

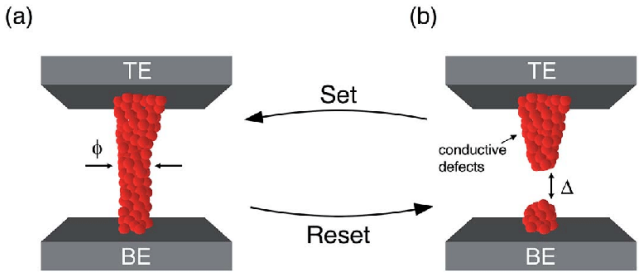
FeRAM

- ITIC structure: one transistor and one Ferro-electric capacitor (FCAP)
- The charge of the FCAP retains its polarity without the need for power, and not require refreshing operations.
- has made it to the stage of mass production earlier than other NVMs and used in many products today, mainly in the embedded domain
  - Automobile equipment such as data recorders and wireless cards,
  - Good candidate for the non-volatile storage for wearable electronics
- Contrary to NAND flash memory, FeRAM does not need asymmetric voltage and/or special high voltage for write operations
- Difficult to make scalable



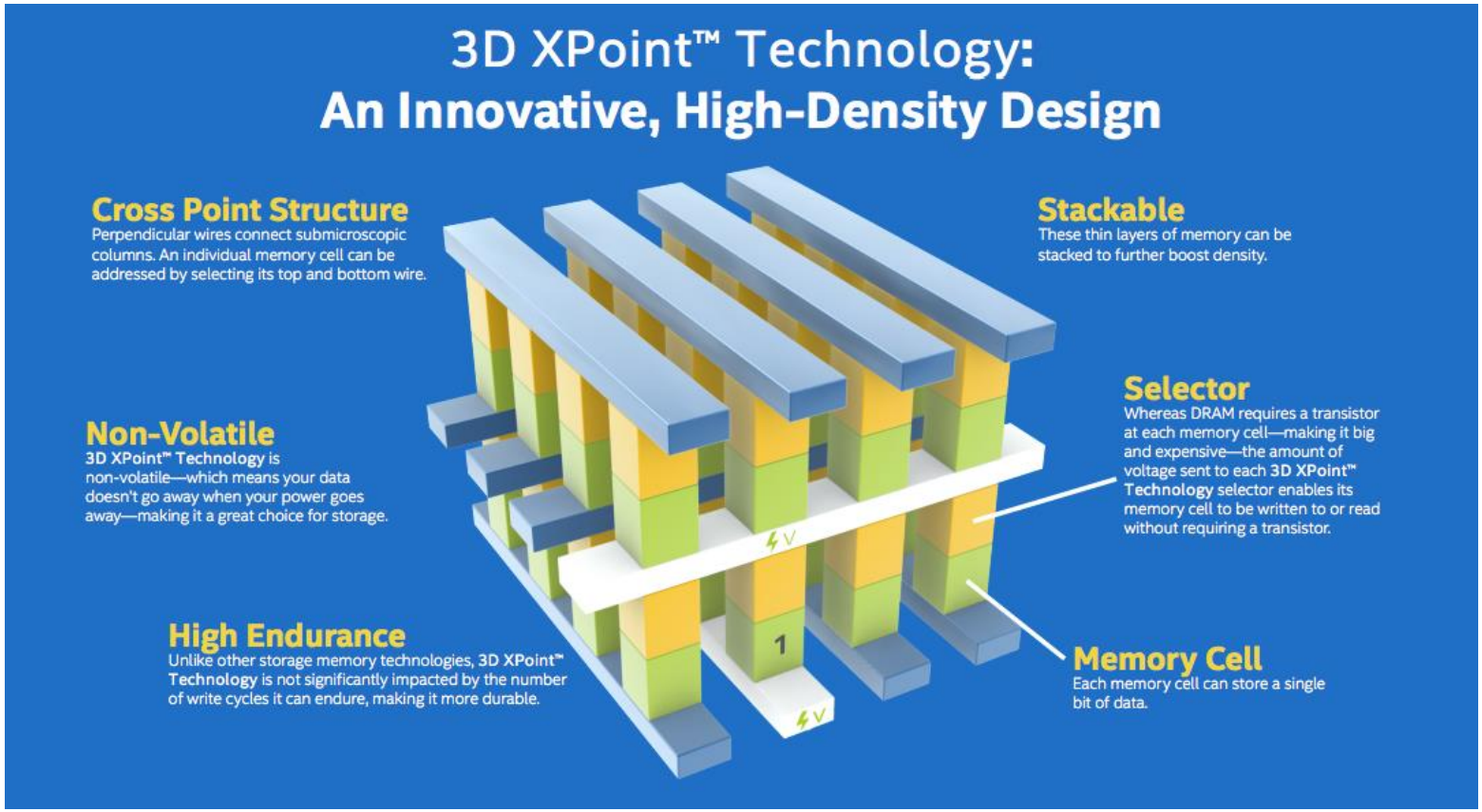
ReRAM

- Based on memristor
  - Two-terminal resistor device with varying resistance
  - The resistance value does not change when the power is turned off
  - A low resistance value is a logical 1, a high resistance value is a logical 0.
- A ReRAM cell is a two-terminal Metal-Insulator-Metal (MIM)
- The formation (Low Resistive State) and disruption (High Resistive State) of a conductive filament
- Lower power consumption than PCM and a higher density than MRAM



3D Xpoint (Optane)

- Intel and Micron



Optane SSD

INTEL® OPTANE™ SSD 800P

KEY SPECIFICATIONS

Specifications
3D XPoint™ Memory Media
PCIe* 3.0x2 with NVMe* Interface
M.2 2280 Single Sided (2280-S3-B-M)
0-85C operating temperature
Power L1.2: 8mW (idle)
Intel controller and firmware
365 TBW Endurance (Lifetime Writes)
Sequential Read (at QD4): Up to 1450 MB/s
Sequential Write (at QD4): Up to 640 MB/s
4KB Random Read (at QD4): 250K IOPs
4KB Random Write (at QD4): 140K IOPs

Breakthrough Performance
vs. competitive PCIe* Gen3x4 based NAND drives

Sysmark 2014 SE*	
Overall Score	Up to 15% better[1]
Office Productivity	Up to 20% better[2]
Responsiveness	Up to 38% better[3]
PCMark 10*	
Application Startup	Up to 12% better[4]
PCMark Vantage*	
PC Mark Vantage* HDD Score	Up to 44% better[5]

Consistent Responsiveness
vs. competitive PCIe* Gen3x4 based NAND drives

Random Read IOPS @QD1	Up to 4x better[6]
PCMark 8*: Extended Storage Test & SNIA Steady State Test	Consistent Performance

1,2,3,4,5,6 see appendix for footnotes. See Appendix II for System Configuration and testing procedures. All testing done internally by Intel; *Other names and brands may be claimed as the property of others. Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown". Implementation of these updates may make these results inapplicable to your device or system. Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

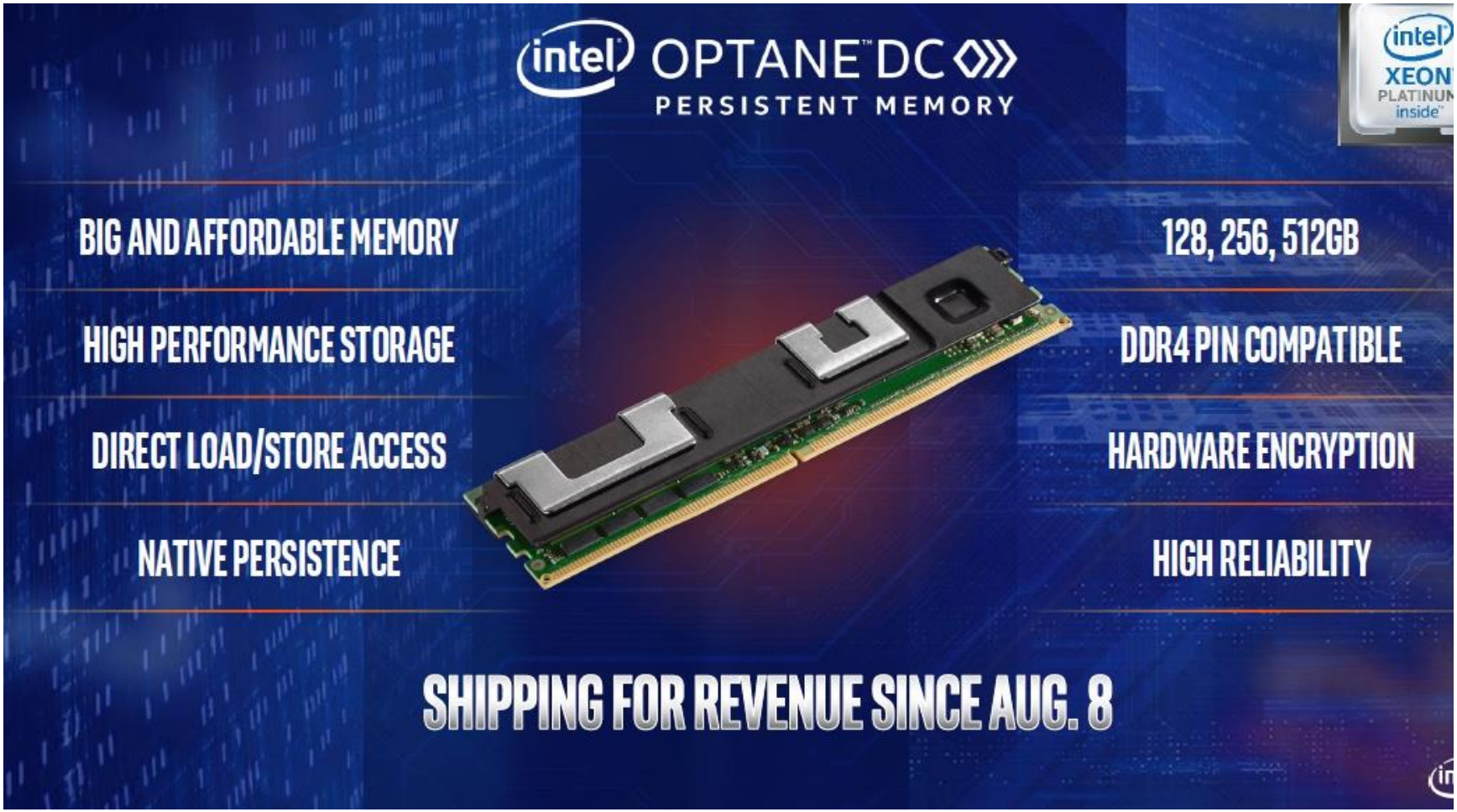
NVM SOLUTIONS GROUP

Under Press Embargo Until March 8, 9:30am PT

intel

7

Optane Memory



NVM File Systems

- BPFS (Microsoft, SOSP'09)
  - Uses an optimized shadow-paging technique for crash consistency
- PMFS (Intel, EuroSys'14)
  - Optimized memory-mapped IO, Light-weight file system
- Aerie (WISC, EuroSys'14)
  - provides direct access for file data IO, using a user-level lease for NVM updates.
- EXT4-DAX
  - extends the Linux EXT4 file system to allow direct mapping of NVM, bypassing the buffer cache
- NOVA (UCSD, FAST'16)
  - Log-Structured, Scalability by employing per-CPU metadata

NVM Memory Systems

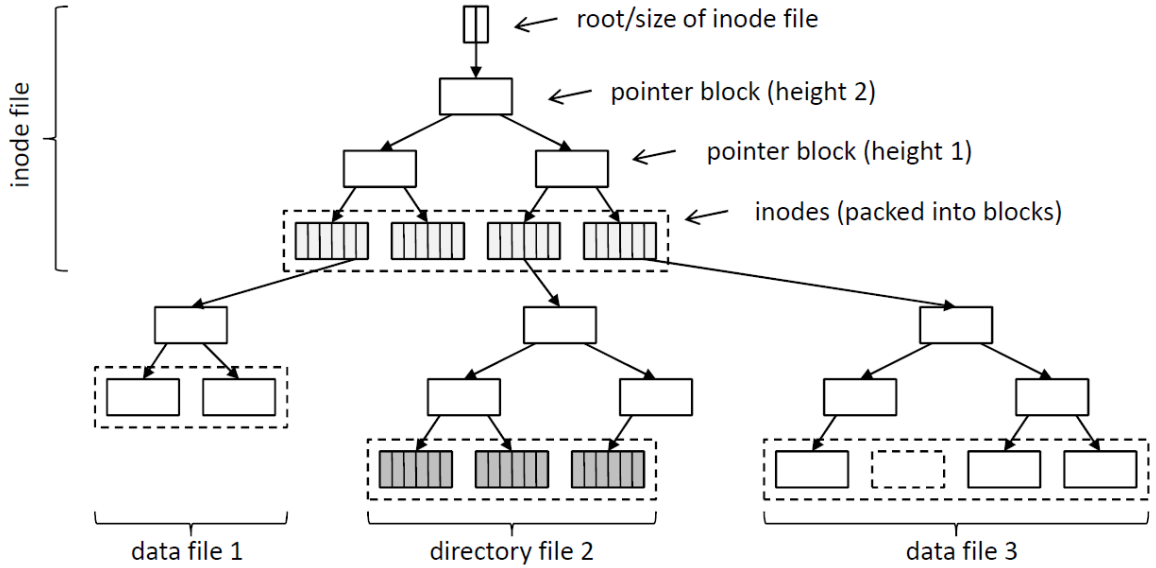
- NV-Heaps (UCSD, ASPLOS'11)
  - Persistent object store
- Mnemosyne (WISC, ASPLOS'11)
  - Exposing persistent memory to user-mode

BPFS

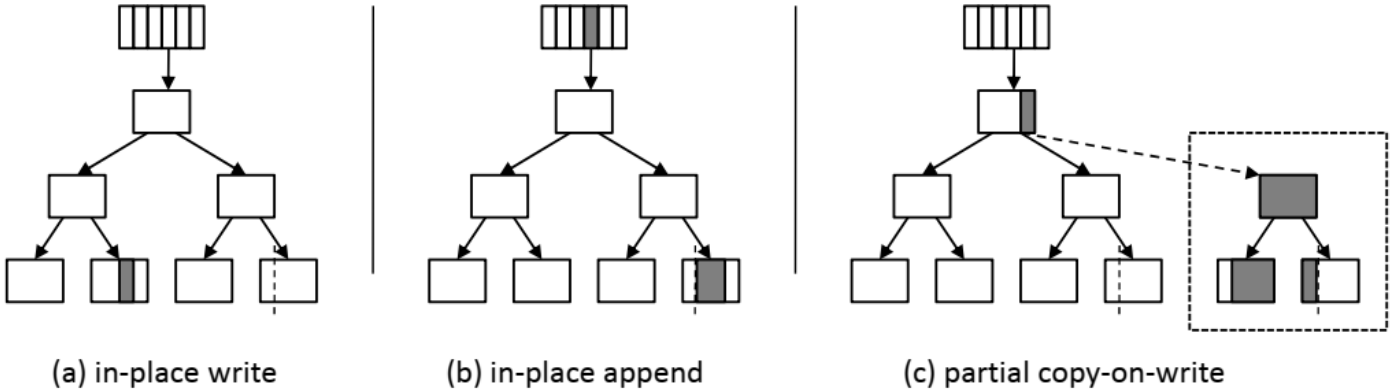
- Expose BPRAM Directly to the CPU on the memory bus
- Enforce Ordering and Atomicity in Hardware
  - Epoch barriers: reordering data in epoch is ok, but reordering epoch is not allowed
  - Atomic 8-byte writes
- Use Short-Circuit Shadow Paging
  - Updates are committed either in-place or using a localized copy-on-write
  - Updates are committed to the file system by performing an atomic write at an appropriate point in the tree

BPFS

File system layout



Data update



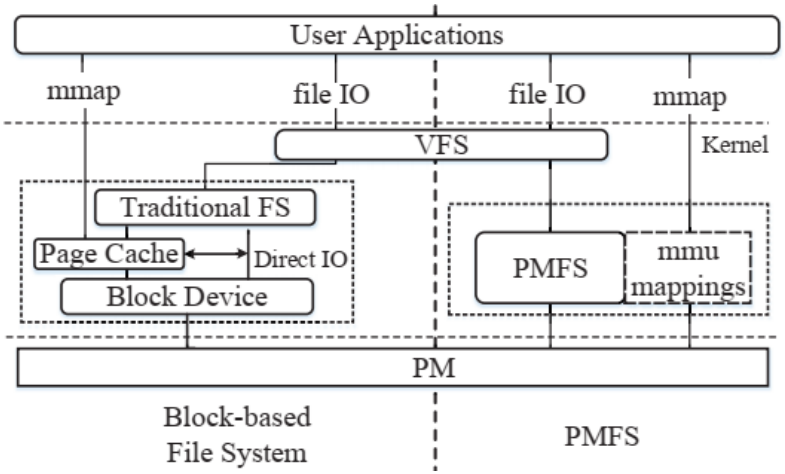
64-bits or less
Atomic write

Modify inode data
(File size)

Issue epoch barriers before and after the atomic write that commits the operation

PMFS

- Support for a light-weight file system
  - Avoid block layer → reduce copy between storage and DRAM
  - Consistency using atomic in-place updates
  - Fine-grained undo journaling, CoW
- Optimized memory-mapped I/O
  - PMFS maps PM pages directly into an application's address space
  - Transparent large page support (efficient use of TLB)
- pm_wbarrier
  - Enforce the durability of a cacheline
  - clflush → sfence → pm_wbarrier



Aerie

- Direct access
  - Reduces kernel overhead and optimizes using app semantic
- Decentralized Architecture
  - Untrusted user-mode lib (libFS)
    - Provides functionality to find and access data
    - Virtual memory protection hardware enforces access control
  - Trusted file-system service (TFS)
    - Integrity for metadata updates and concurrency
- SCM Manager (Kernel)
  - OS kernel provides only coarse-grained allocation and protection

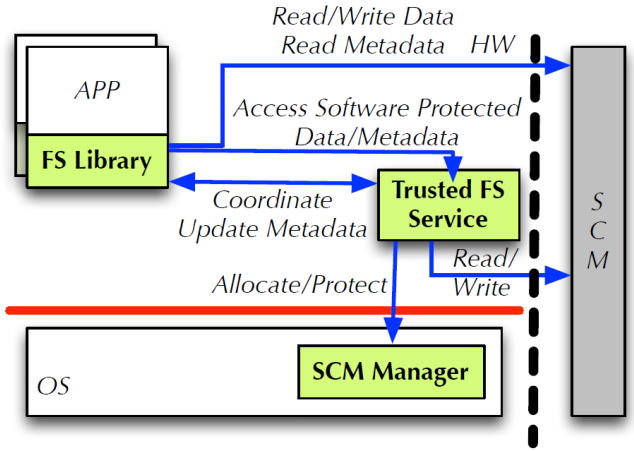
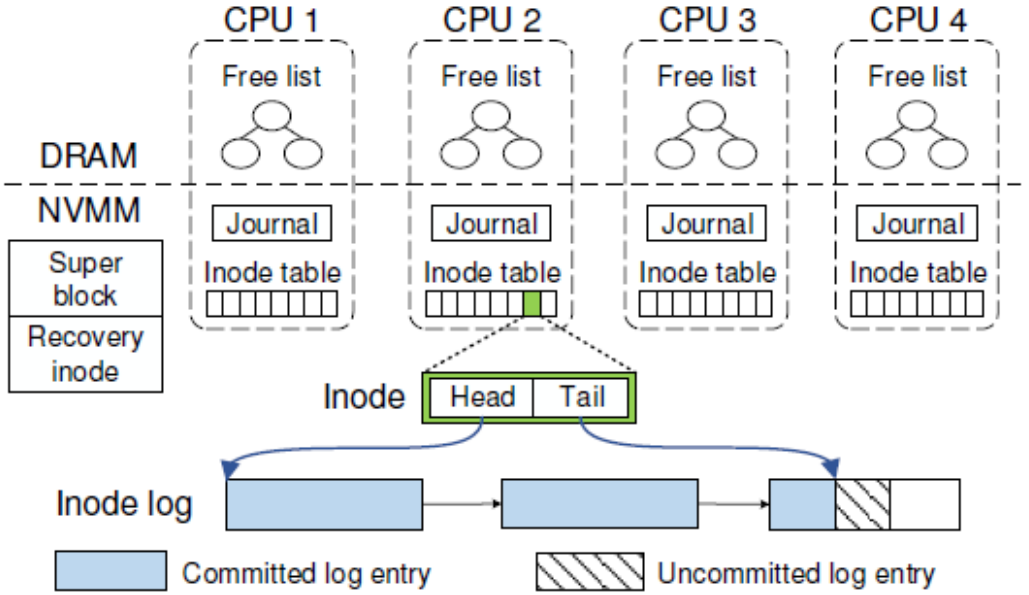


Figure 2: Decentralized Architecture. *Functionality is split between a user-mode library, a trusted service, and the kernel.*

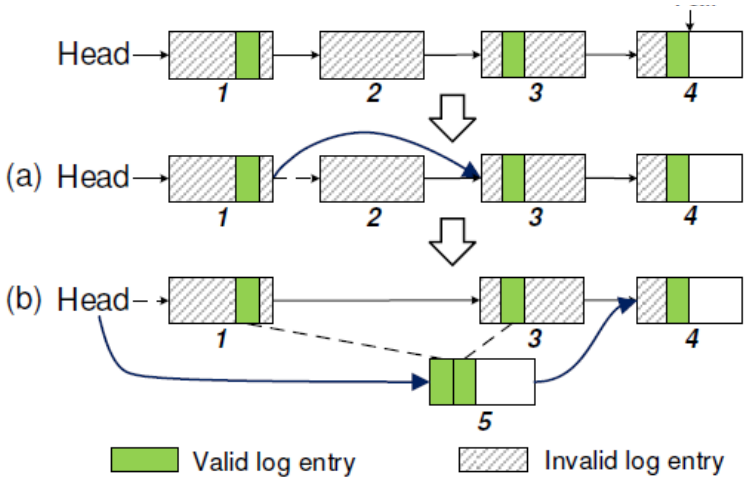
Nova

- 64-bit atomic updates
  - To modify metadata (ex. log pointer)
- Log structuring for single log update
  - Per-inode logging
- Lightweight journaling for update across logs
  - Operations that require changes to multiple inodes
  - Journaling log tails
- Per-CPU NVMM free list, journal and inode table
  - Concurrent transactions

Nova



Per-CPU data structure



Garbage Collection