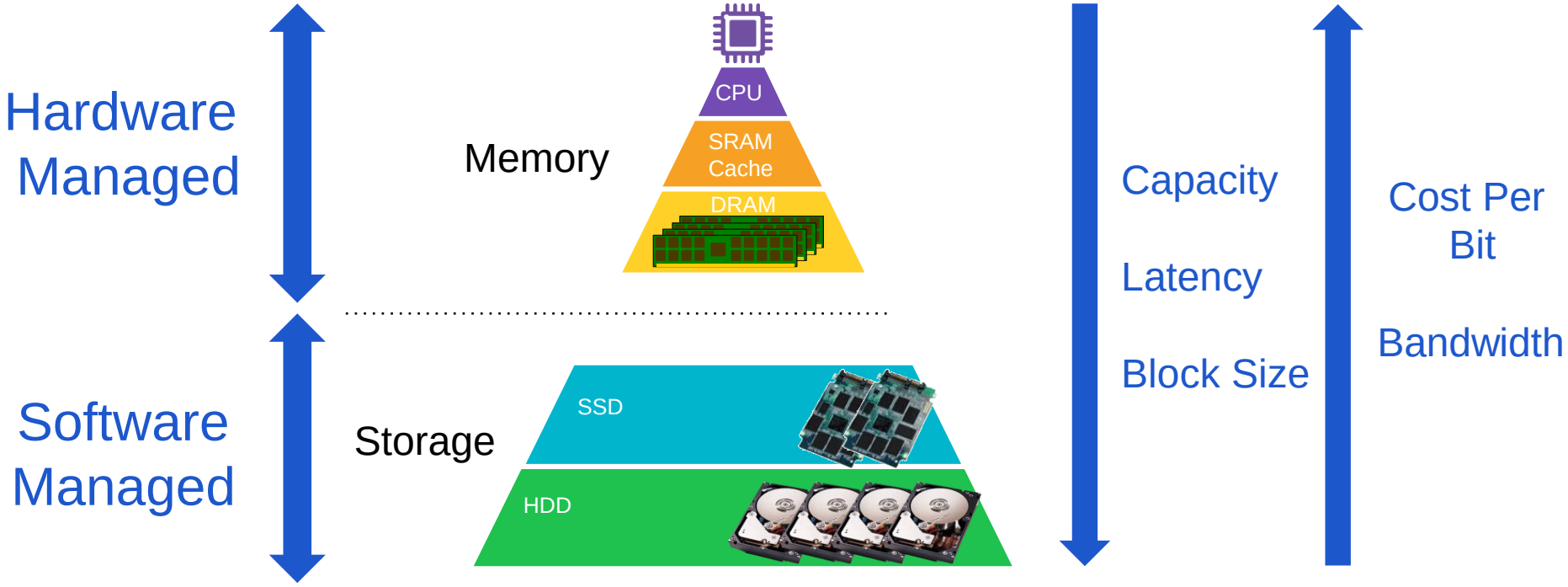
2016 VLSI TECHNOLOGY SYMPOSIUM HONOLULU

The Confluence of Memory and Storage Technologies

Craig Hampel
Chief Scientist, Rambus

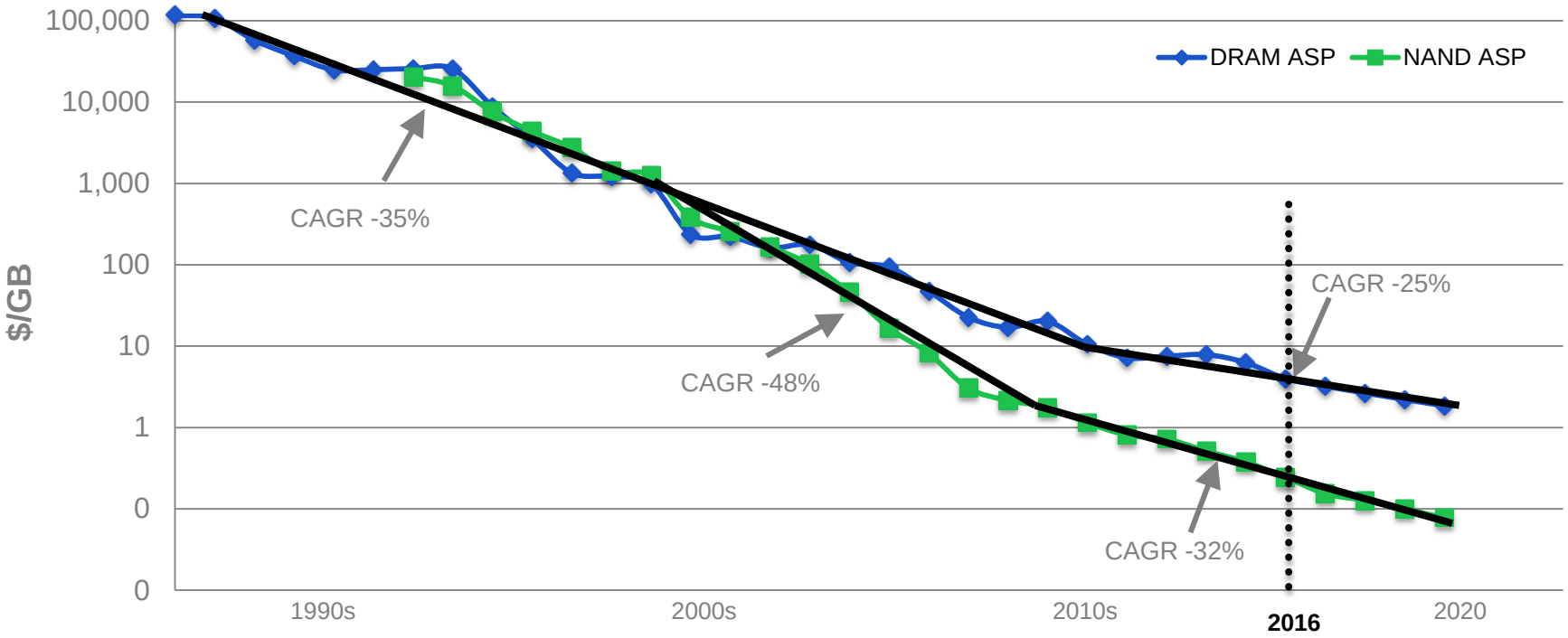
June 13, 2016

Memory and Storage Hierarchy



Historically, the distinction between storage and memory have corresponded to their location in the hierarchy.

DRAM and Flash Prices

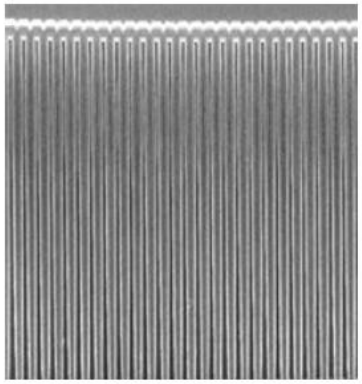


Sources: IDC

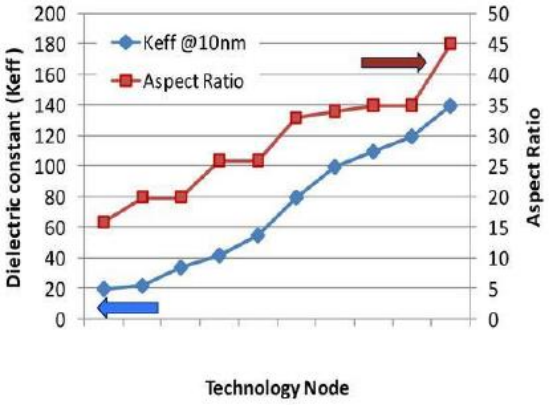
While both NAND Flash and DRAM have followed semiconductor scaling trends, NAND now costs 1/16th as much as DRAM.

DRAM Scaling to Slow

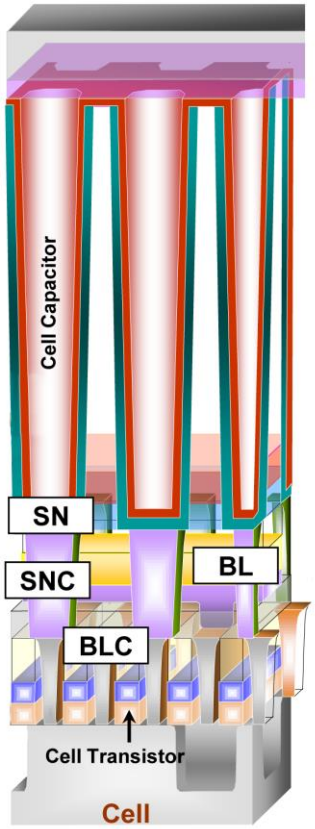
- DRAM capacitor no longer scaling
  - 3D capacitor and insulator near physical limits
- Process complexity growing
- DRAM cost reduction slowing
- No clear path to cell stacking



Capacitor Structuring



Ref: E. Byers (Micron) - Lithovision 2015



Ref: S.Y. Cha (Hynix) - IEDM Short Course 2011

2016 VLSI TECHNOLOGY SYMPOSIUM HONOLULU

Differences at the Device Level

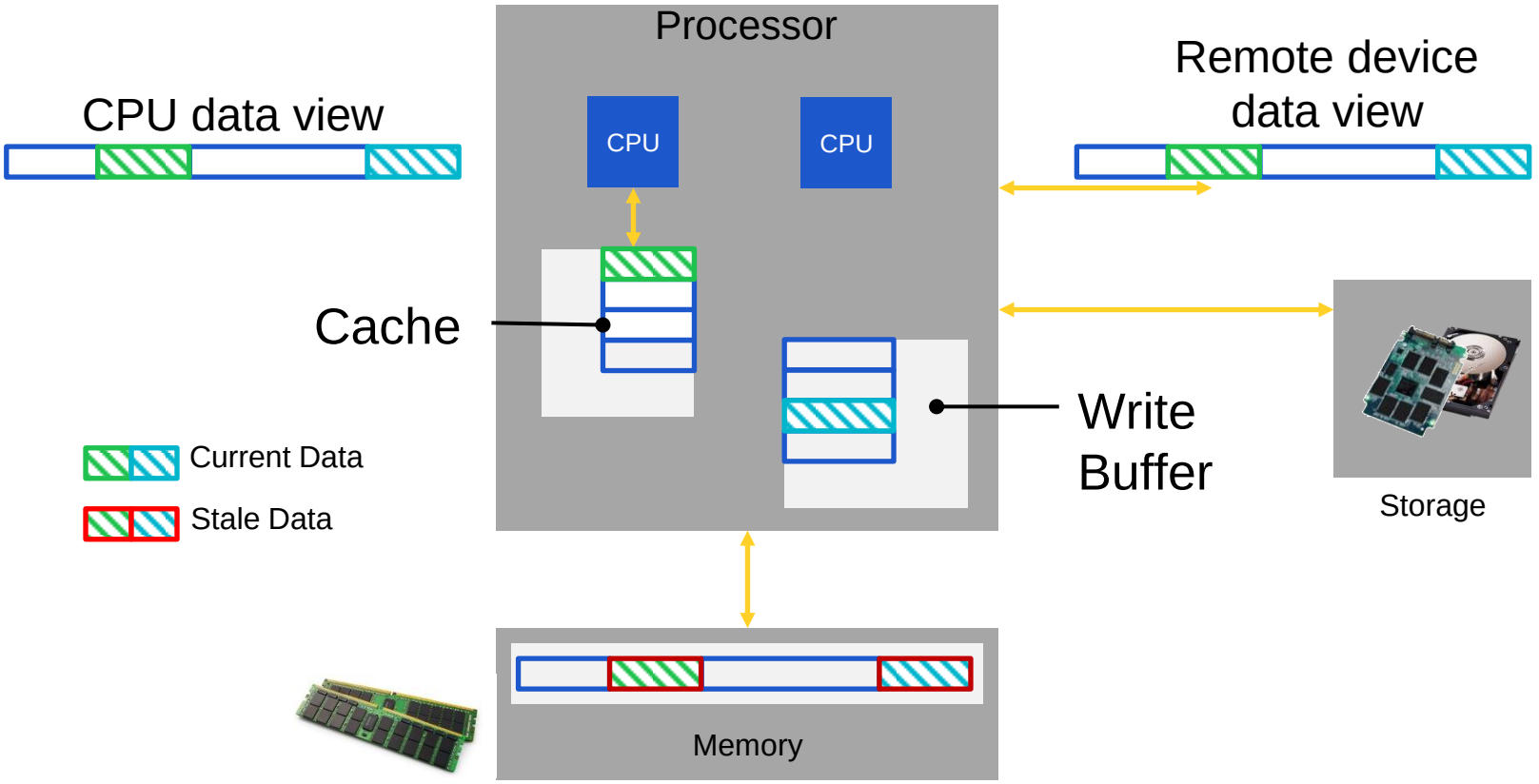
Memory	Storage
Volatile	Non-volatile
Byte or line addressable	Block addressable
Directly writeable	Erase required
Deterministic and uniform latency	Non-deterministic and varied latency
Endurance > 10^{15}	Endurance > 10^{6}
Latency in ns	Latency in µs to ms

System Integration Distinctions

Memory	Storage
CPU waits	CPU context switches
Hardware controlled state	Software controlled state
Contextually unaware	Context present
Not power coherent	Power coherent

System integration has largely driven by latency and block size.

Making Memory Persistent



In most systems today data in DRAM is volatile and not coherent

Basic Interface Differences

Memory	Storage
CPU direct interface	Abstracted interface
100s of GB/s per CPU	10s of GB/s per CPU
10s of outstanding transactions	100-1000s of outstanding transactions
Fixed scheduling	Split transaction and dynamic scheduling

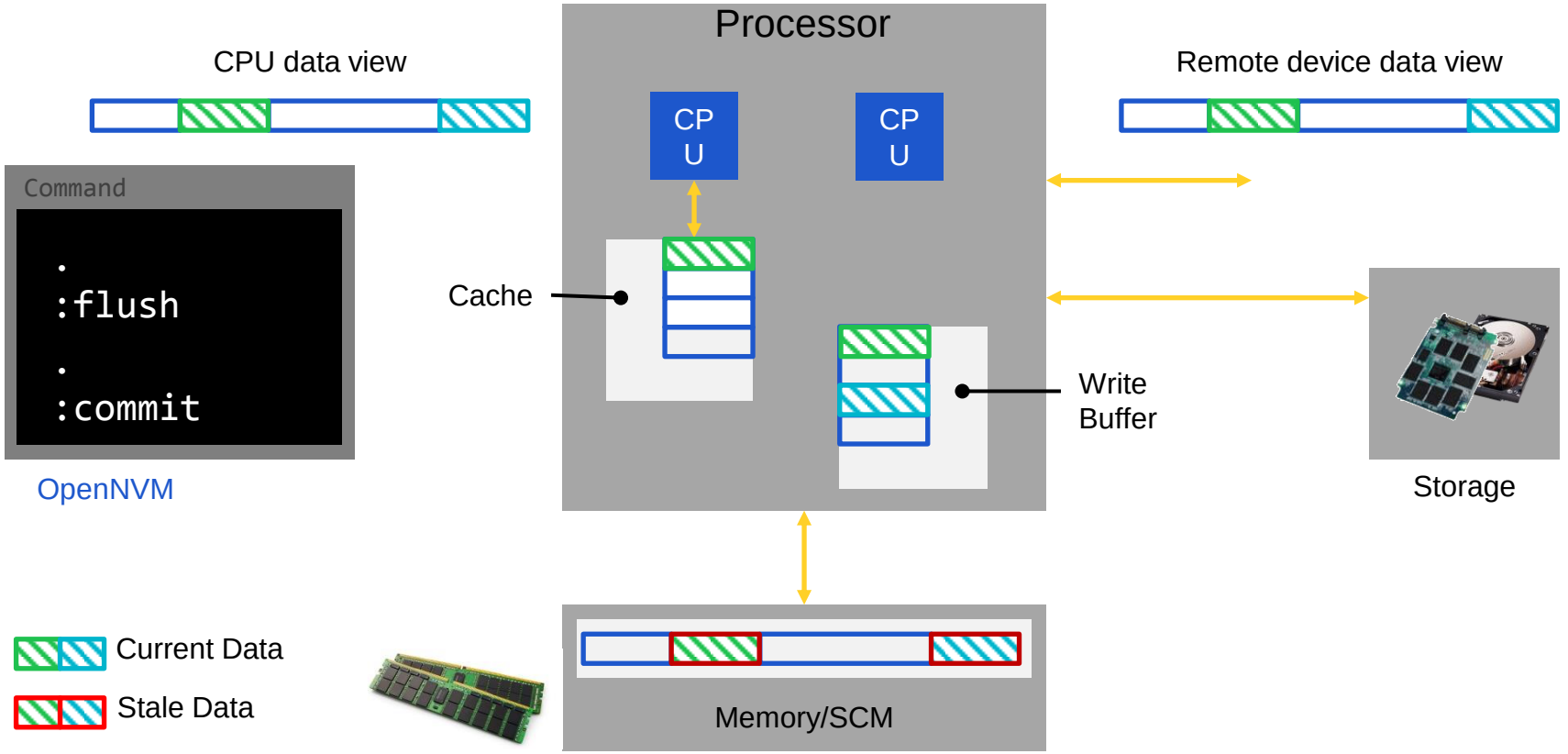
Memory and storage interfaces have largely been defined to accommodate the device characteristics.

Software View

Memory	Storage
Partial and intermediate data	Complete and final data
Unsaved Data	Saved and persistent data

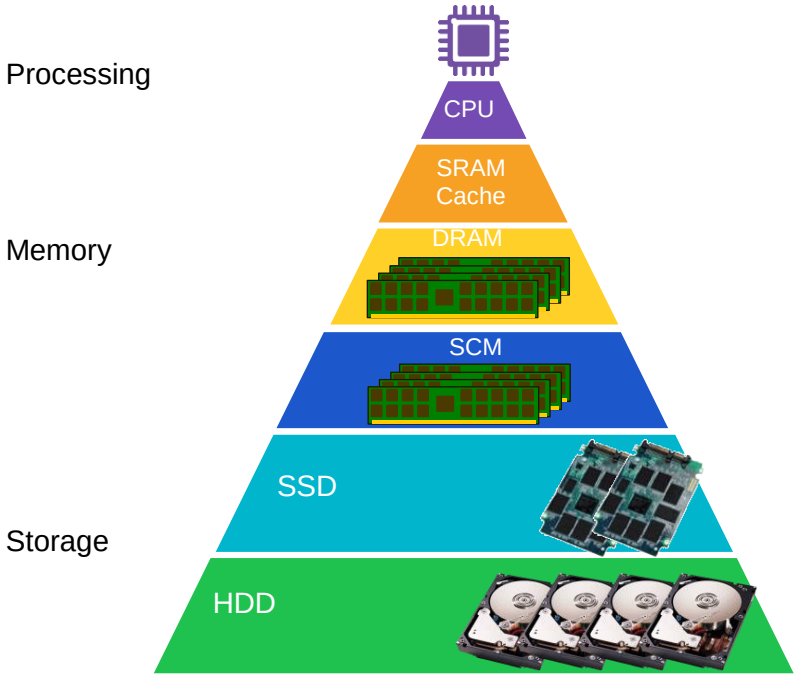
Emerging applications need to support the software and users view of data.

Making Memory Persistent



New processor hardware is needed to allow "DRAM" interfaces to be coherent and persistent.

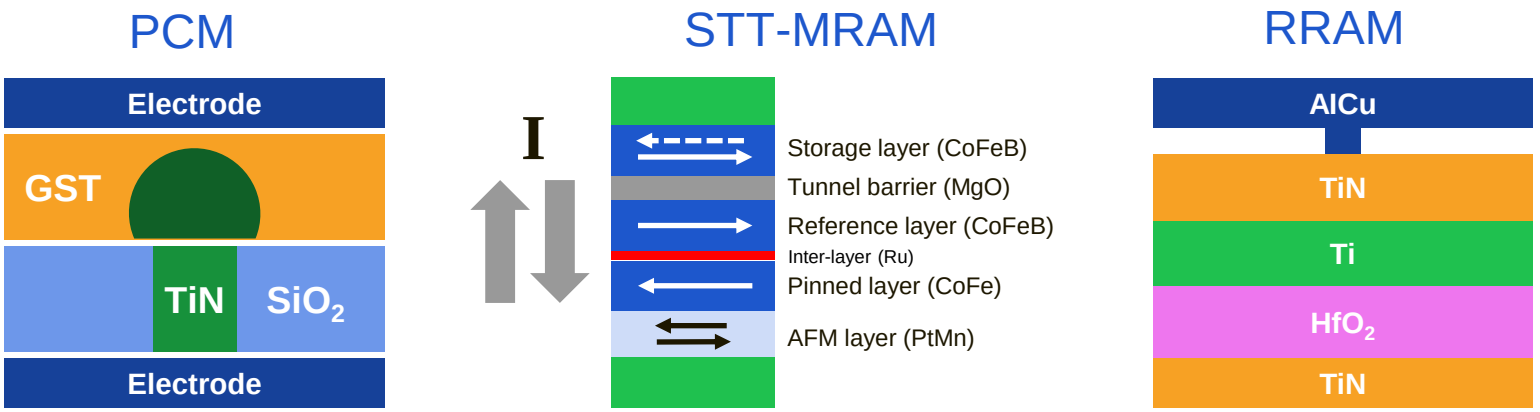
Storage Class Memory (SCM)



- Storage devices on DDR channel
  - NAND-based (slow)
    - NV-DIMM, UltraDIMM^{tm}
    - NAND-based future
  - Byte-addressable SCM
    - 3D Xpoint^{tm}
    - ReRAM
- SCM likely not as fast as DRAM, requires management and caching, but …
  - Lower cost per bit
  - Provides application with direct access to large data with minimal OS overhead
  - Enables big data application speed improvements (with optimized software)

Non volatile memory with lower latency and reduced block size facilitate storage on a DDR channel

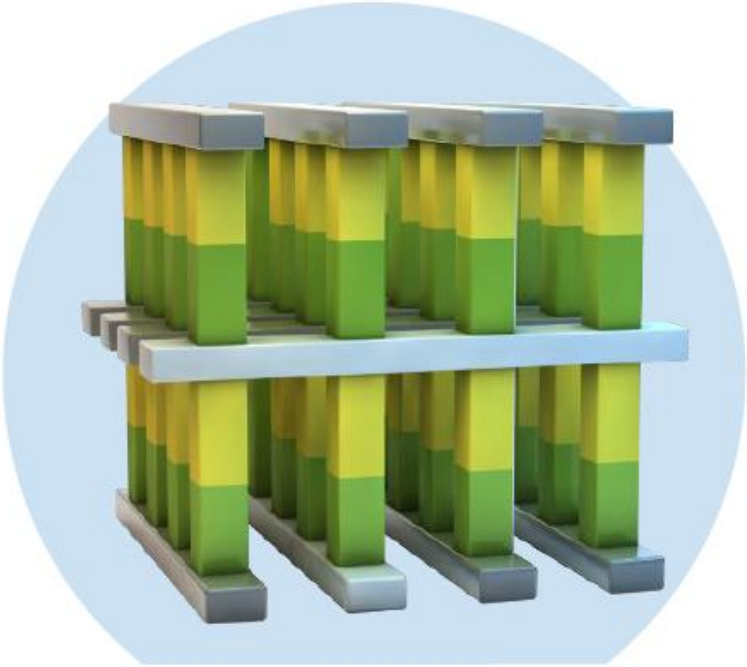
Candidate Storage Class Memories



The goal should be non-volatility, fast access time, and small block size at “better-than-DRAM” cost

3D X-Point (PCM with selector)

- Announced by Intel/Micron in August 2015
- “10 times the latency of DRAM”
- Cache line write and read addressable
- Direct write (no erase)
- Target cost 2-4x lower than DRAM
- Initial technology 2 layers with roadmap to 4 layers
- First products – fast SSDs
- Later direct access DIMMs



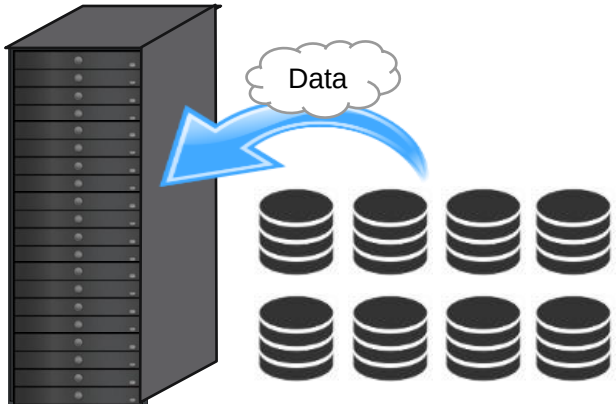
Source: Intel

SCM Bit Cell Emerging Options

- STT-MRAM
  - Often considered as DRAM replacement
  - “Infinite endurance” and non-volatile - but very challenging to make
  - Cost not low enough to replace DRAM
  - Embedded SRAM replacement more likely
- PCM / 3D-Xpoint
  - Performance between DRAM and Flash
  - Cost and power seem high today – can they improve?
  - Most mature
- Filamentary Memory (metal filament [CBRAM] or oxygen vacancy [RRAM])
  - Less mature than STT-MRAM or PCM
  - Opportunity for better cost – similar to 3D NAND
  - Lower power
  - First application as embedded NVM

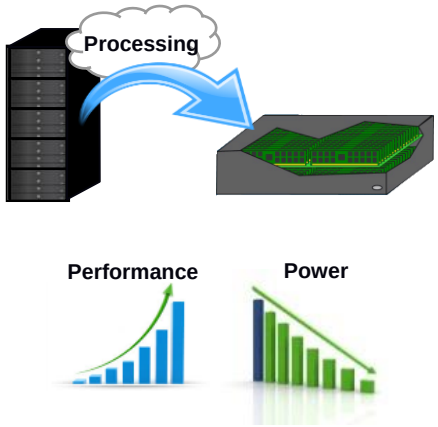
Future SCM Integration

1. Locate more data closer to processing



SCM on memory channels facilitate more low latency data closer to processing.

2. Locate data centric processing closer to data

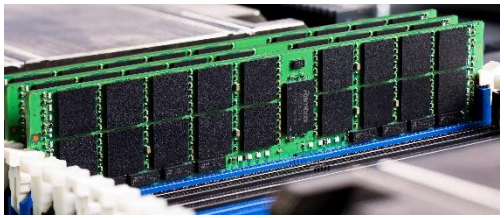


Compute offload near direct access storage

New SCM opportunities improve performance, power efficiency, and TCO which are increasingly limited by data movement.

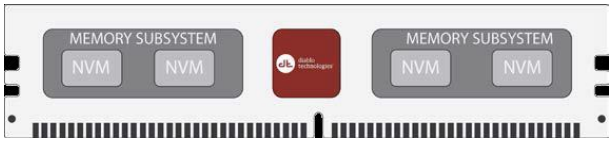
Expanding Memory Near Processing

LRDIMM



Increase local DRAM memory capacity through buffering

NVDIMM-F, Memory1



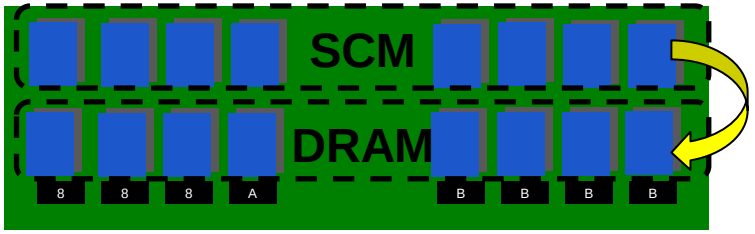
Source: Diablo

Block NV memory to extend local capacity, DRAM as page cache and write buffer

3D XPoint DIMM

Byte (line) read/writeable NV with processor managed caching in DRAM

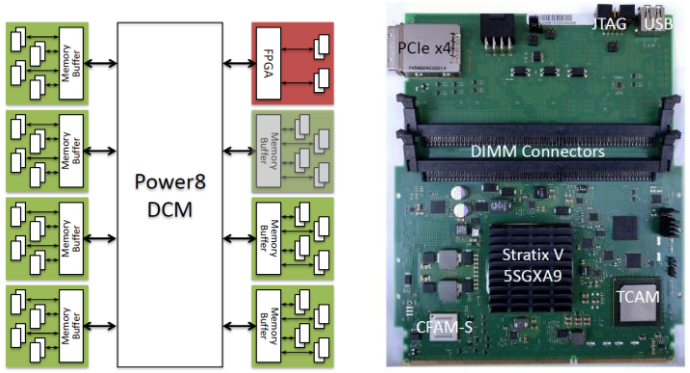
Future - Caching DIMMs



Byte (line) read/writeable NV with local cache management

Memory Processing Closer to Memory

IBM ConTutto



ConTutto is an FPGA-based memory card that connects as IBM Power8 memory

Different memory technologies can be used in Power system

Highest bandwidth & lowest latency FPGA attach point in any computer system

Enables Near-memory acceleration

Rambus Smart Data Appliance

Offload application libraries and functions to an FPGA connected directly to DRAM (24DIMMs)

Reconfigurable hardware acceleration enables dynamic provisioning and repurposing

Minimize data movement to and from SDA engines by chaining commands together

Converged Memory Interface

1. Data Rate – deliver up to 6.4+Gbps (anticipate add'l storage bandwidth)

2. 2 DIMMs Per Channel – must support both DRAM & SCM modules on same channel

3. Allow efficient allocation and scheduling of SCM & DRAM bandwidth

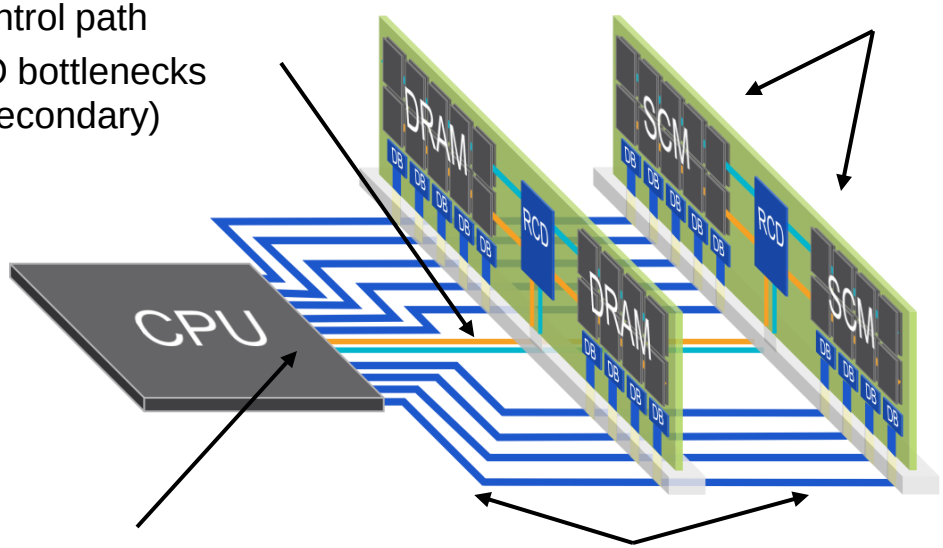
   - Non deterministic
   - Out of order
   - Minimal (no) latency impact on DRAM access

4. Maintain similar economics, infrastructure for low-risk industry adoption (DDR roadmap compatible)

DRAM/SCM Interface Potential

Revamp control buses

- Provide more general purpose control path
- Remove I/O bottlenecks (Primary, Secondary)



Extend LRDIMM architecture

- Support higher data and control rates
- Address and data buffers to abstract memory types
- Support caching and mixed DRAM/Storage-class Memory (SCM) module types

Improved protocol

- Support storage class memory over memory type bus
- Pipeline mixed minimum latency with non-deterministic and varied latencies

Improved Data Bus

- Lower swing, power efficient, single-ended signaling
- New data bus topologies to improve data rates

2016 VLSI TECHNOLOGY SYMPOSIUM HONOLULU

Conclusion: Storage and Memory Convergence

- The distinctions between memory and storage will blur
  - Memory and storage will begin to share numerous characteristics
- DRAM scaling is likely to slow with NAND and storage class memories continuing to scale
- Low Latency SCM will be located on DIMMs to expand processor memory with an eveolution from paging to caching to on-DIMM Caching
- With small block size, and lower latency SCM will facilitate data processing offload near data at rest
- Memory interfaces (like DDR) with extensions for storage class memory support will facilitate more data to be stored near processing

For Further Reading

K. Suzuki and S. Swanson. The Non-Volatile Memory Technology Database (NVMDB). Technical Report CS2015-1011, Department of Computer Science & Engineering, UCSD, http://nvmdb.ucsd.edu.

M. J. Breitwisch. Phase change memory. Interconnect Tech*nology Conference, 2008. IITC 2008. International*, pages 219–221, June 2008.

B. C. Lee, E. Ipek, O. Mutlu, and D. Burger. Architecting phase change memory as a scalable DRAM alternative. In ISCA ’09: Proceedings pages 2–13.

R. Fackenthal, etal, . A 16Gb ReRAM with 200MB/s write and 1GB/s read in 27nm technology. In Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International, pages 338–339, Feb 2014.

Y. Zhang and S. Swanson. A Study of Application Performance with Non-Volatile Main Memory. In *Proceedings of the 2015 IEEE Symposium on Mass Storage Systems and* Technologies (MSST’15), 2015.

Persistent Memory Programming. http://pmem.io

Maher Amer, CTO, Diablo Technologies Flash in the Memory Channel: Enabling Truly Converged Memory Flash Memory Summit 2015 Proceedings

ConTutto is an IBM Research Configurable Platform for Innovation in the Memory Subsystem of an OpenPOWER Node http://openpowerfoundation.org/presentations/contutto/