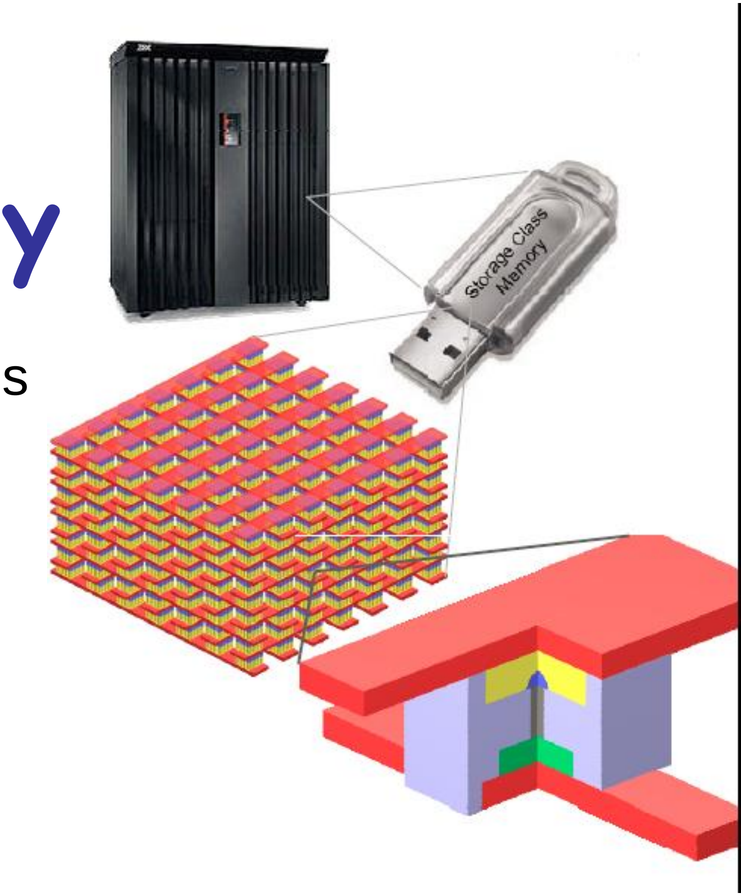
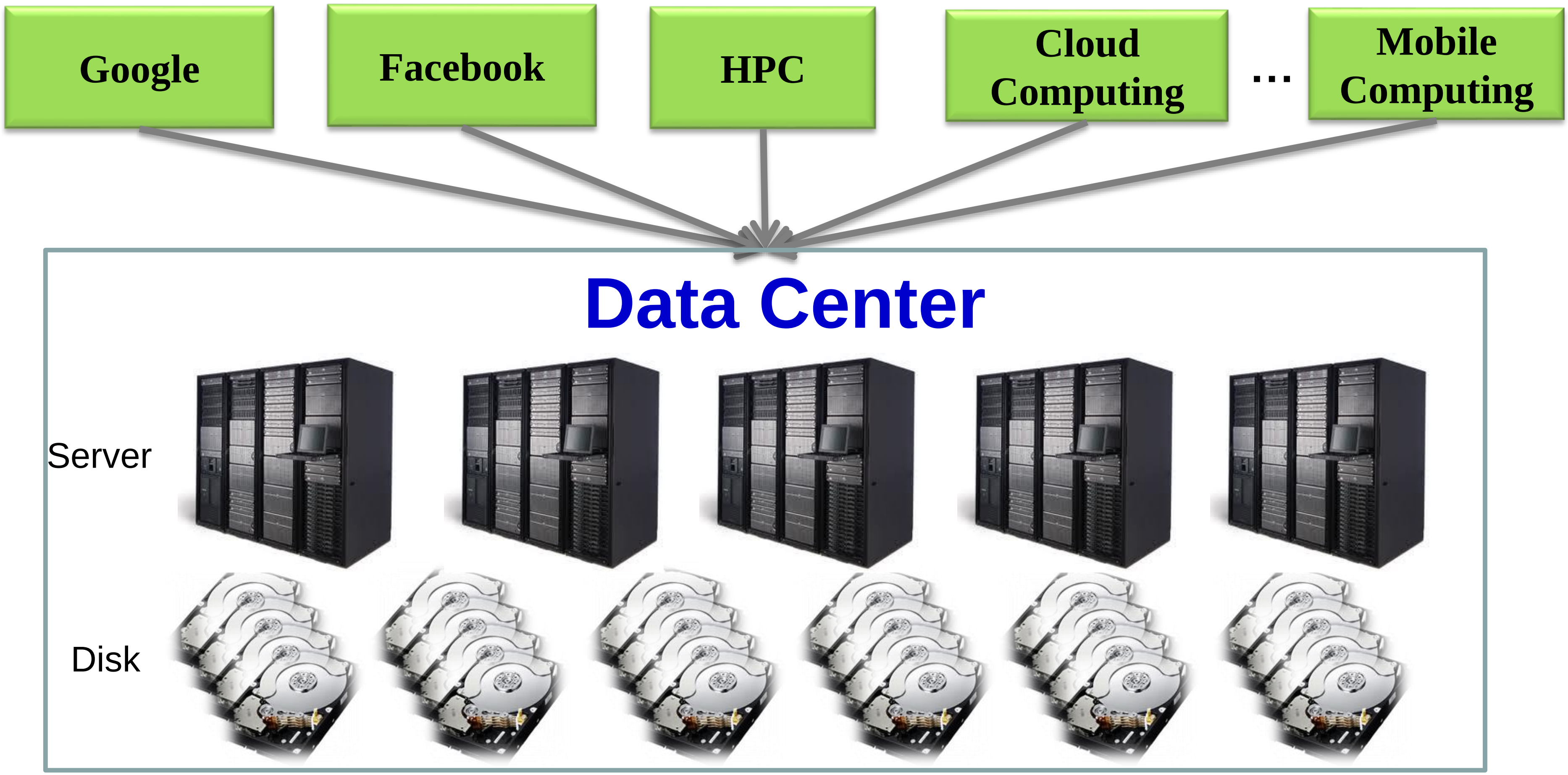
Storage Class Memory

A solid-state memory that blurs the boundaries between storage and memory by being low-cost, fast, and non-volatile.



Current Challenges



Google

Facebook

HPC

Cloud Computing

…

Mobile Computing

Data Center

Server

Disk

Challenges

Performance

Scalability

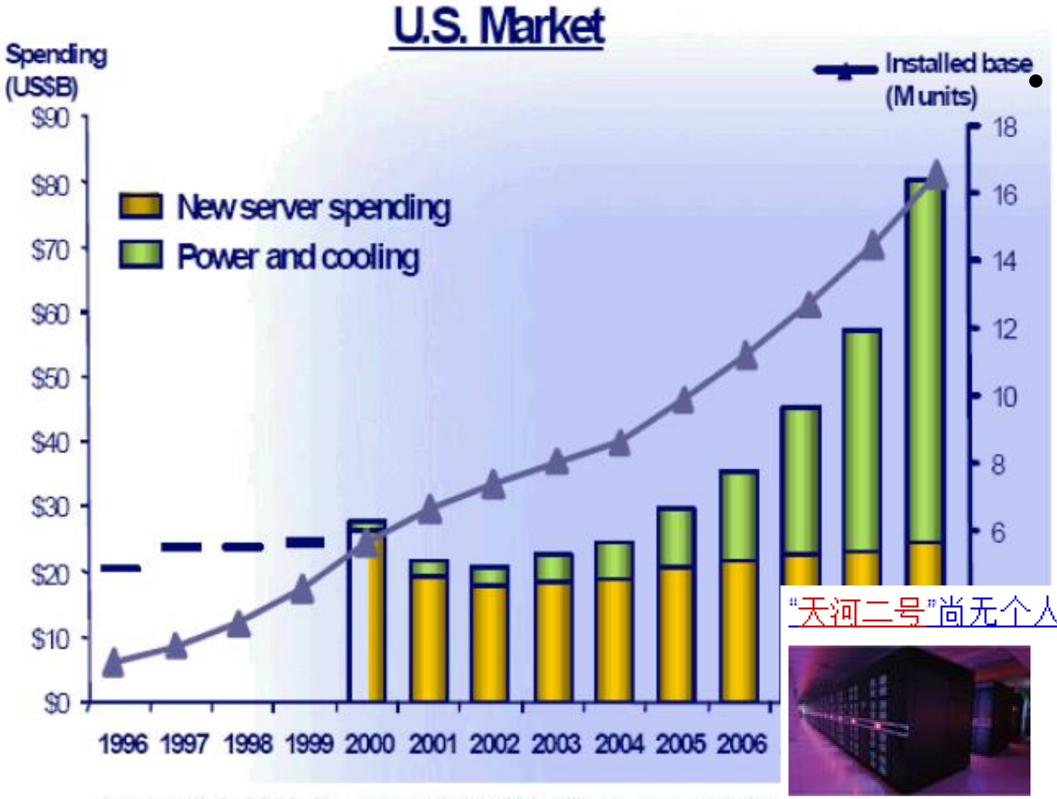
Energy Consumption

Space

Operation Cost

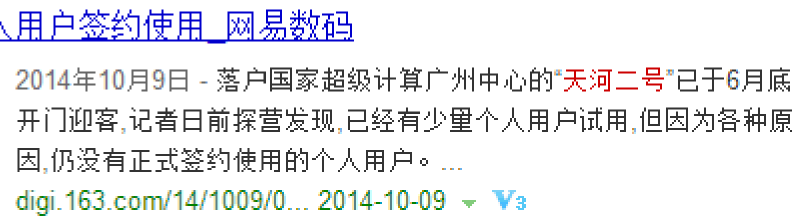
Power & space in the server room

- The **cache/memory/storage hierarchy** is rapidly becoming the **bottleneck for large systems.**



Source IDC: 2006, Document # 201722, "The Impact Of Power and Cooling On Data Center Infrastructure", John Humphreys, Jed Scaramella

- We know how to create MIPS & MFLOPS cheaply and in abundance, but **feeding them with data has become** the performance-limiting and most-expensive part of a system (in **both $ and Watts).**



“天河二号”尚无个人用户签约使用_网易数码

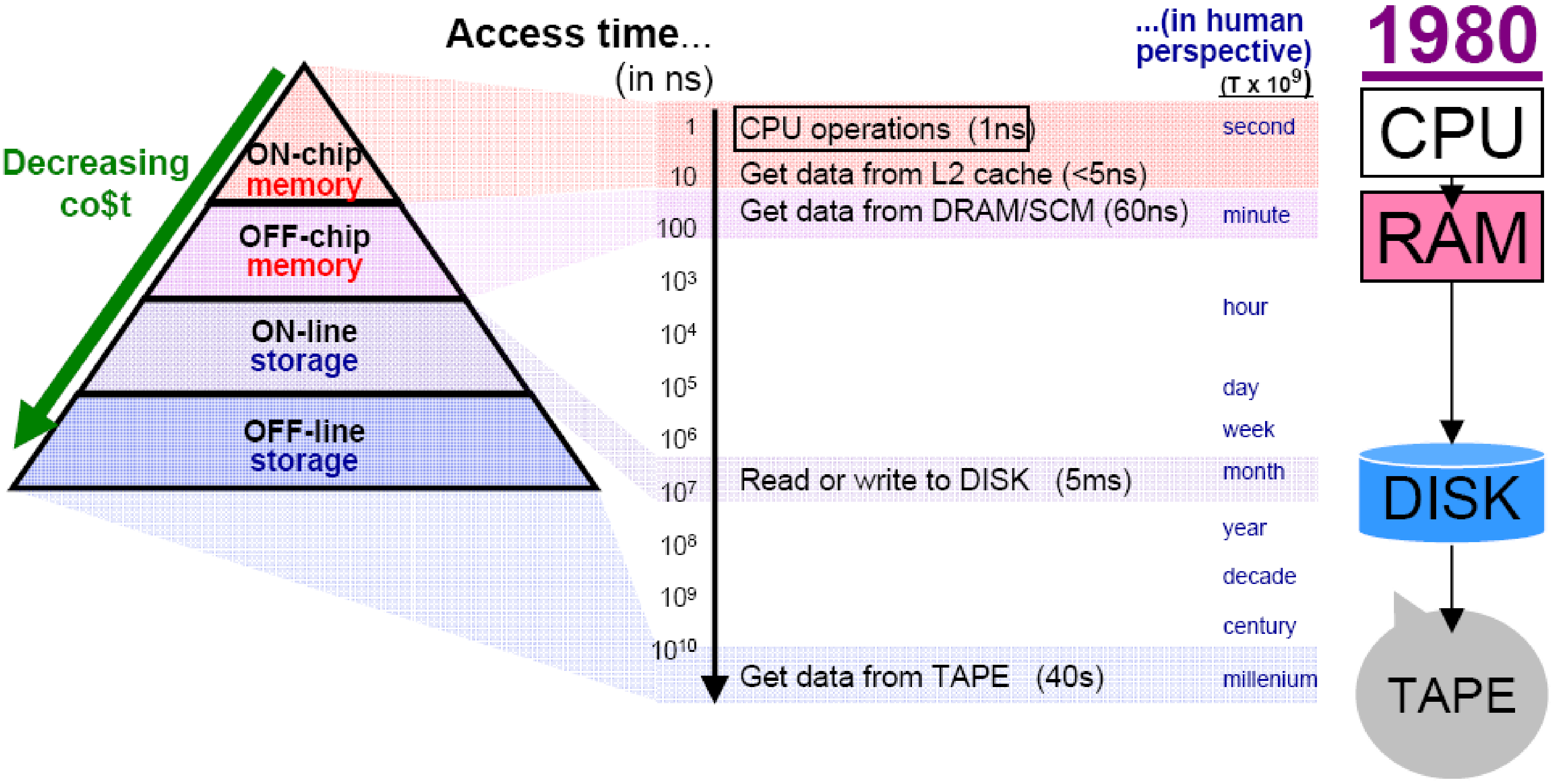
2014年10月9日 - 落户国家超级计算广州中心的“天河二号”已于6月底开门迎客,记者日前探营发现,已经有少量个人用户试用,但因为各种原因,仍没有正式签约使用的个人用户。...

digi.163.com/14/1009/0... 2014-10-09 V3

Problem & Opportunity

The access-time gap between memory & storage

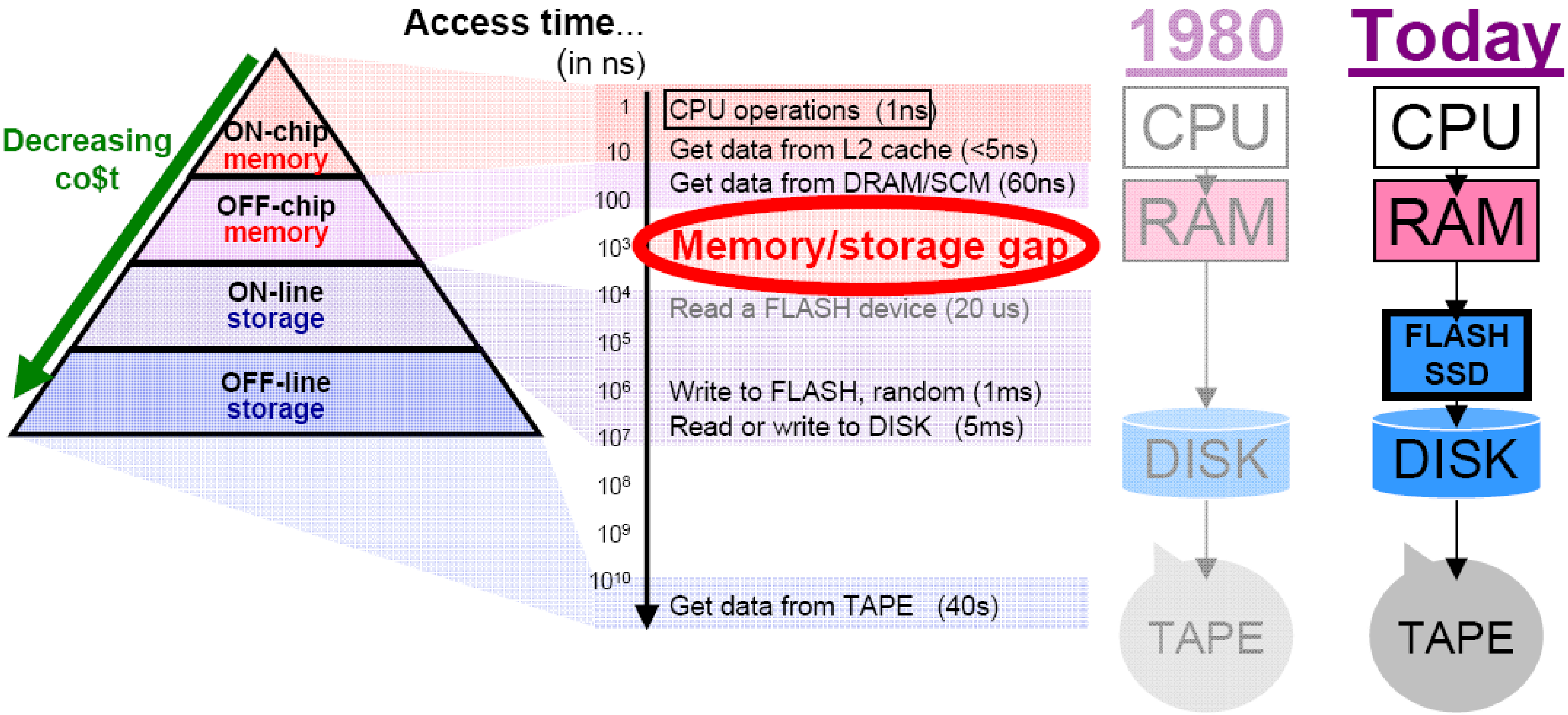
- Modern computer systems have long had to be designed around **hiding the access gap** between **memory and storage → caching, threads, predictive branching, etc.**
- “Human perspective” – if a CPU instruction is analogous to a 1-second decision by a human, retrieval of data from off-line tape represents an analogous delay of 1250 years



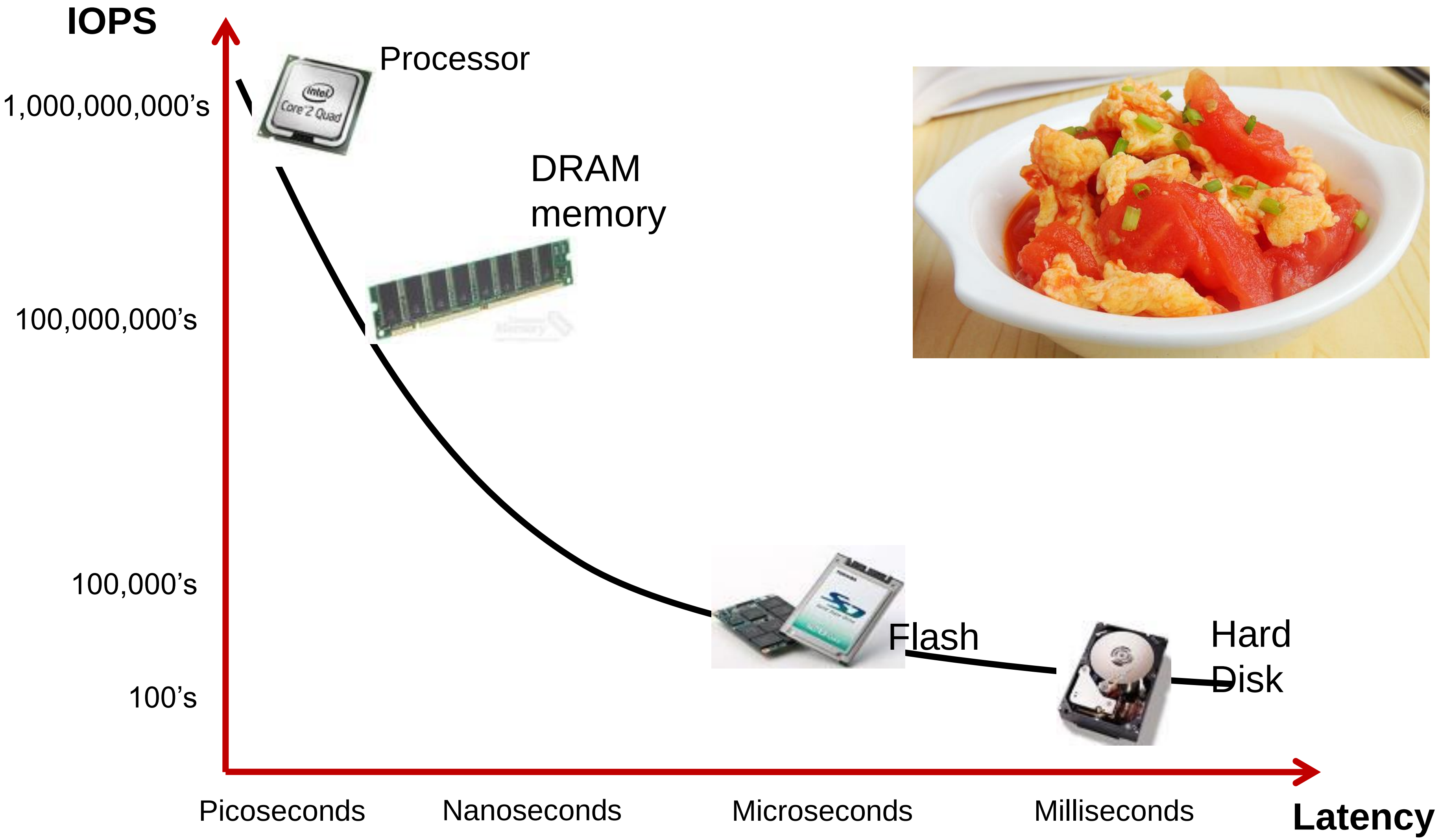
Problem & Opportunity

The access-time gap between memory & storage

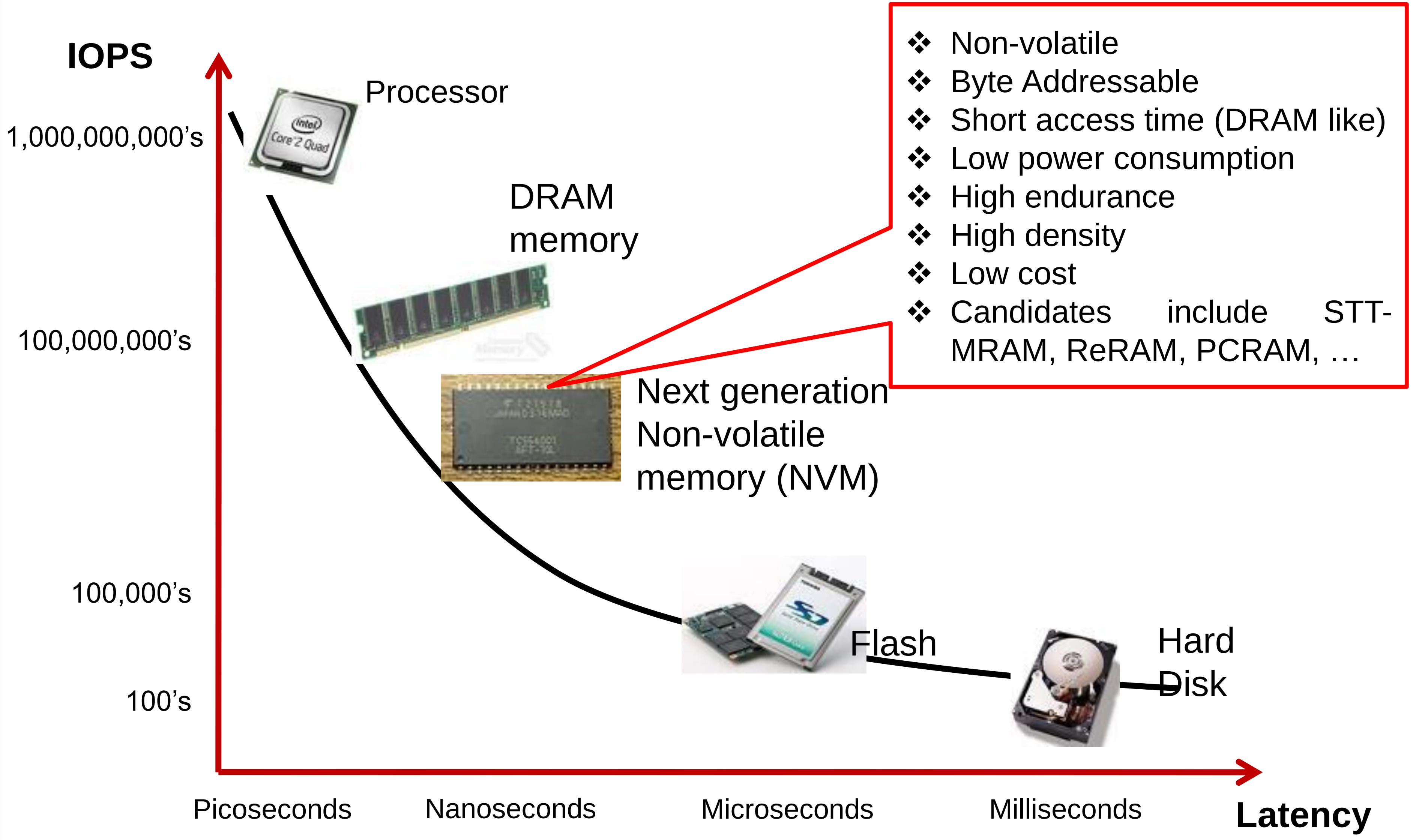
- Today, Solid-State Disks based on NAND Flash can offer fast ON-line storage, and storage capacities are increasing as devices scale down to smaller dimensions…
- but while prices are dropping, the performance gap between memory and storage remains significant, and the already-poor device endurance of Flash is getting worse.



Storage Class Memory

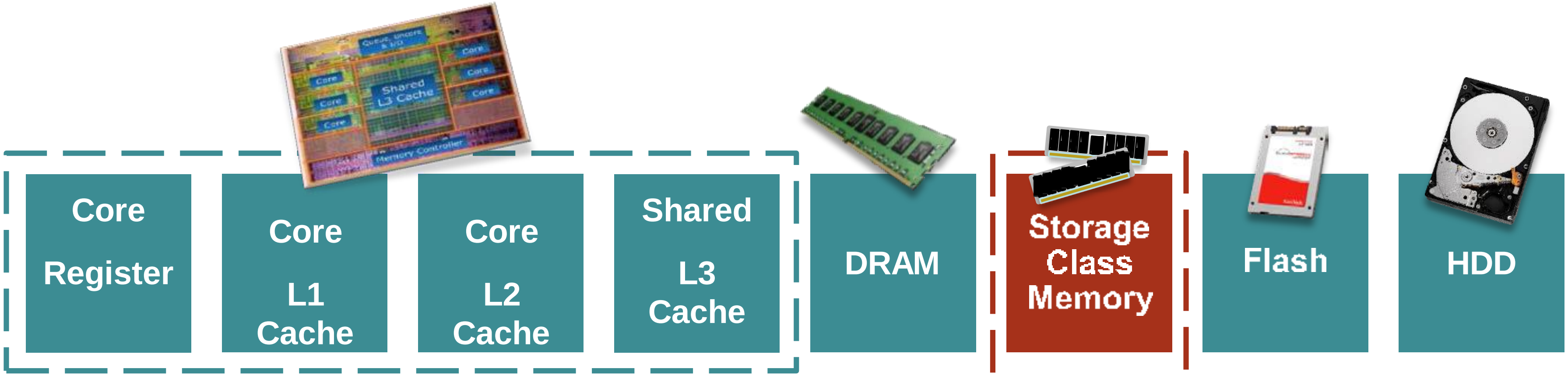


Storage Class Memory



- ❖ Non-volatile
- ❖ Byte Addressable
- ❖ Short access time (DRAM like)
- ❖ Low power consumption
- ❖ High endurance
- ❖ High density
- ❖ Low cost
- ❖ Candidates include STT-MRAM, ReRAM, PCRAM, …

Storage Class Memory



Size	64KB	256KB	2-4MB	16-128GB	128GB-1TB	512GB-4TB	4-16TB
Speed	1ns	3-10ns	10-20ns	50-100ns	250-5,000ns	100,000ns-2,000,000ns	5-10,000,000ns
Cost				100x	20-25x	5x	1x

Source: Western Digital estimates

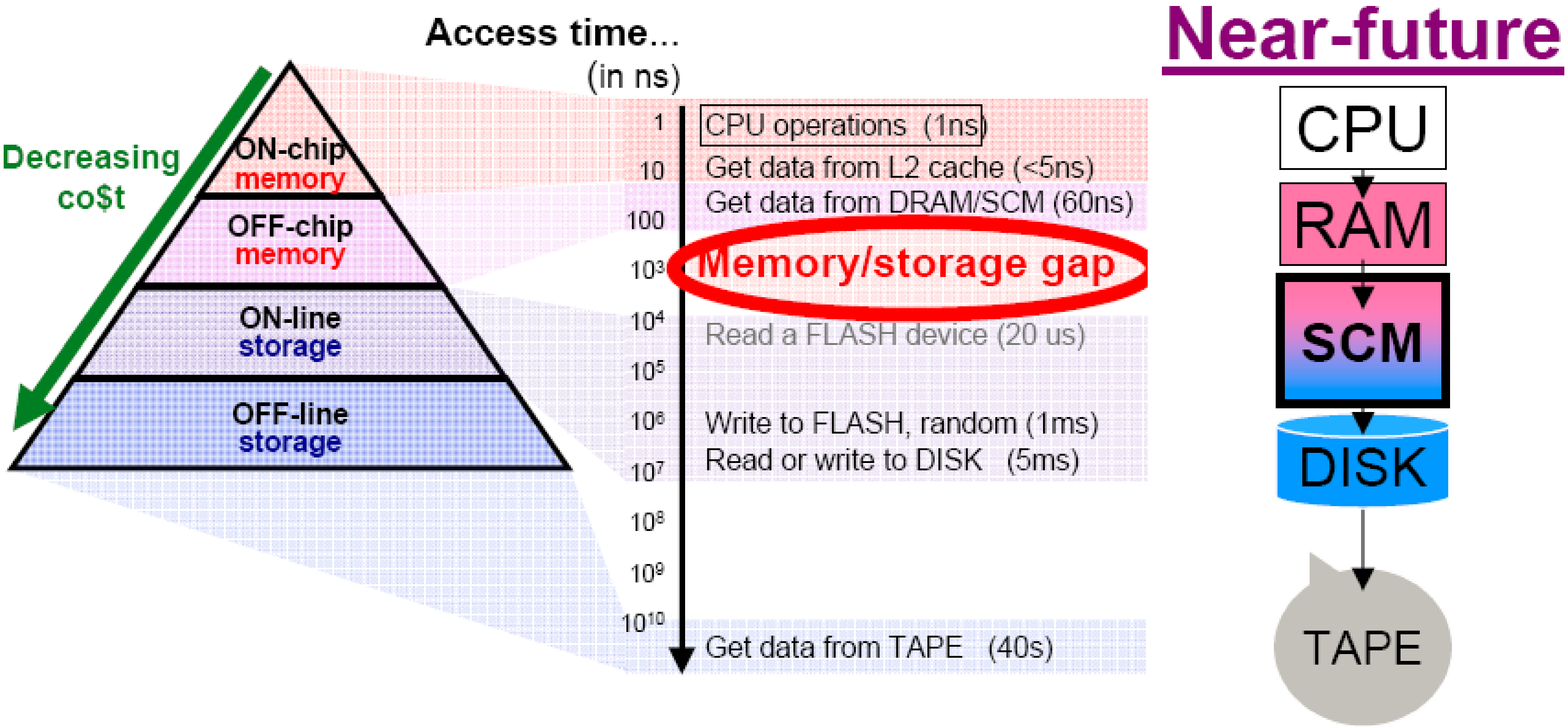
Problem & Opportunity

The access-time gap between memory & storage

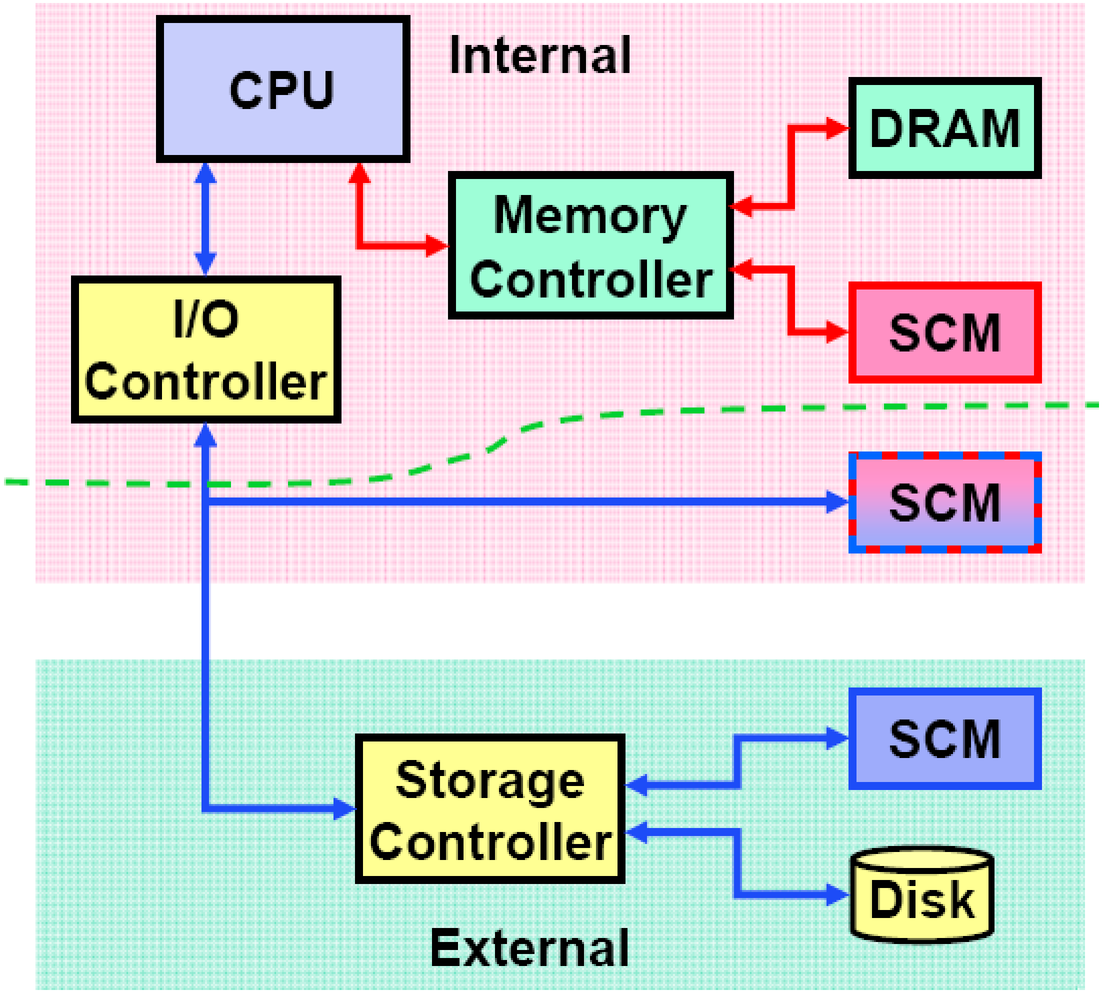
- **Several interesting ways to change the memory/storage hierarchy**

  1) **M-type Storage Class Memory** – high-density, fast OFF- (or ON*)-chip NVM

  2) **S-type Storage Class Memory** – high-density, very-near-ON-line storage



S-type vs. M-type SCM



M-type: Synchronous

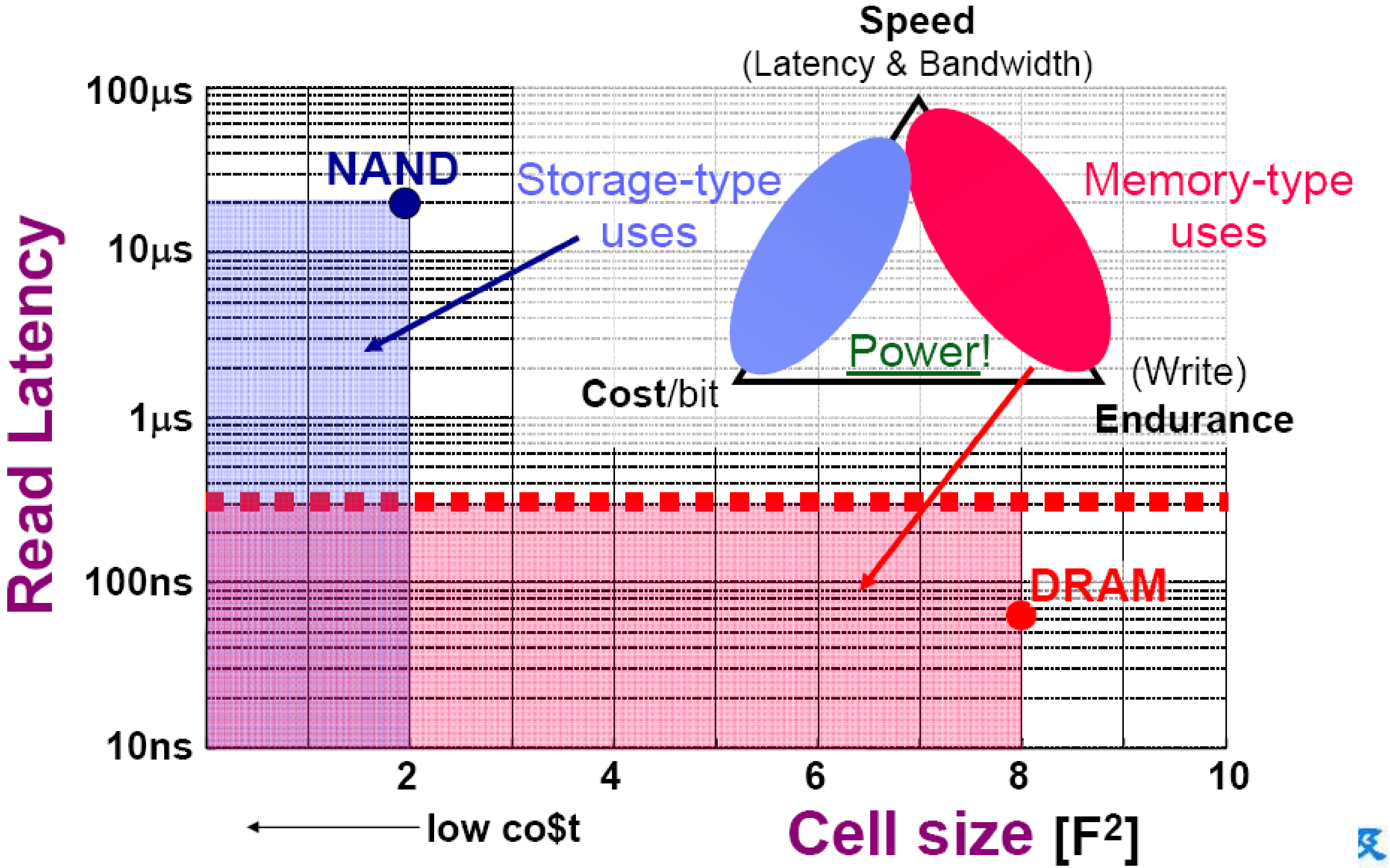
- Hardware managed
- Low overhead
- Processor waits
- New NVM → **not Flash**
- Cached or pooled memory
- Persistence (data survives despite component failure or loss of power) requires redundancy in system architecture

~1us read latency

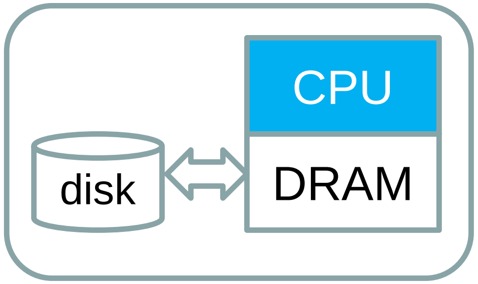
S-type: Asynchronous

- Software managed
- High overhead
- Processor doesn't wait, (process-, thread-switching)
- Flash or new NVM
- Paging or storage
- Persistence → RAID

Storage-type vs. memory-type Storage Class Memory

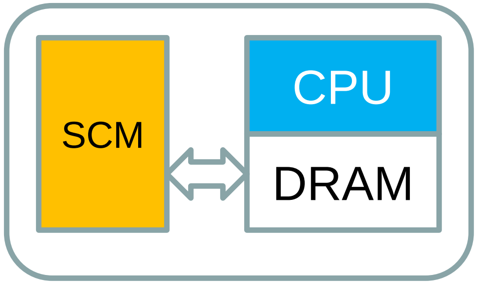


SCM System Integration

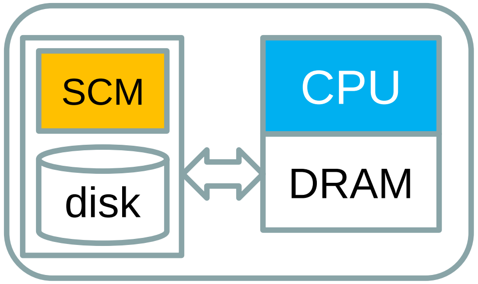


a. Current system

SCM as Block Device

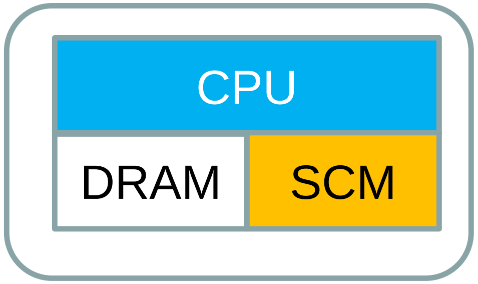


b. Replace disk

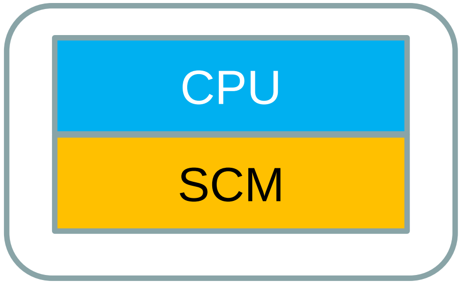


c. Hybrid disk

SCM as Memory Device



d. Hybrid Memory



e. Entire SCRAM

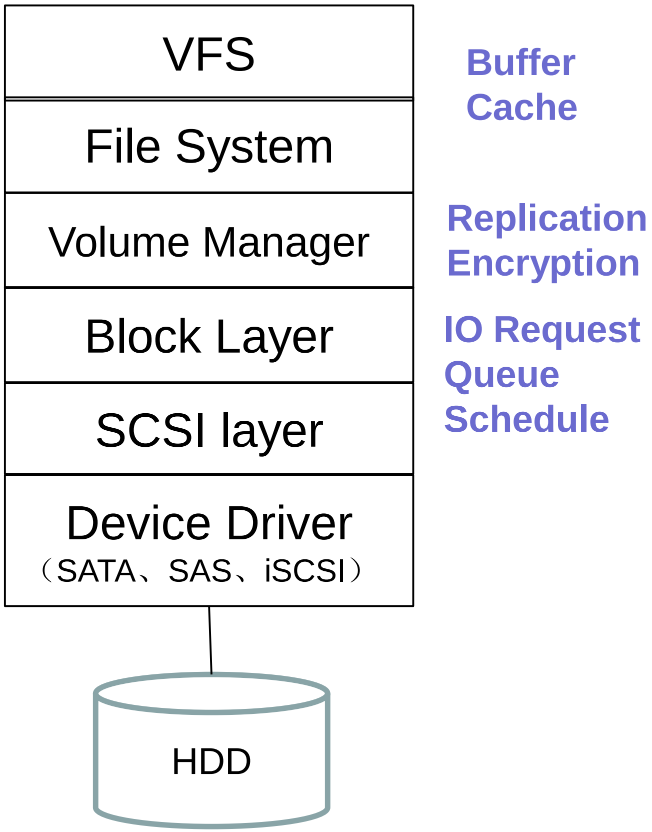
Disk-based Storage Software Systems

- Software plays an important role in improving system **performance**

  I/O scheduling，Buffer cache et al.

- Software can also provide useful services like replication, encryption, compression, provenance tracking, et al. (**Reliability** & **Security**)
- For a 4-Kbyte access to a commodity disk, the stock Linux software stack accounts for just 0.3% of the latency and 0.4% of the energy

Linux IO stack



VFS

File System

Volume Manager

Block Layer

SCSI layer

Device Driver
（SATA、SAS、iSCSI）

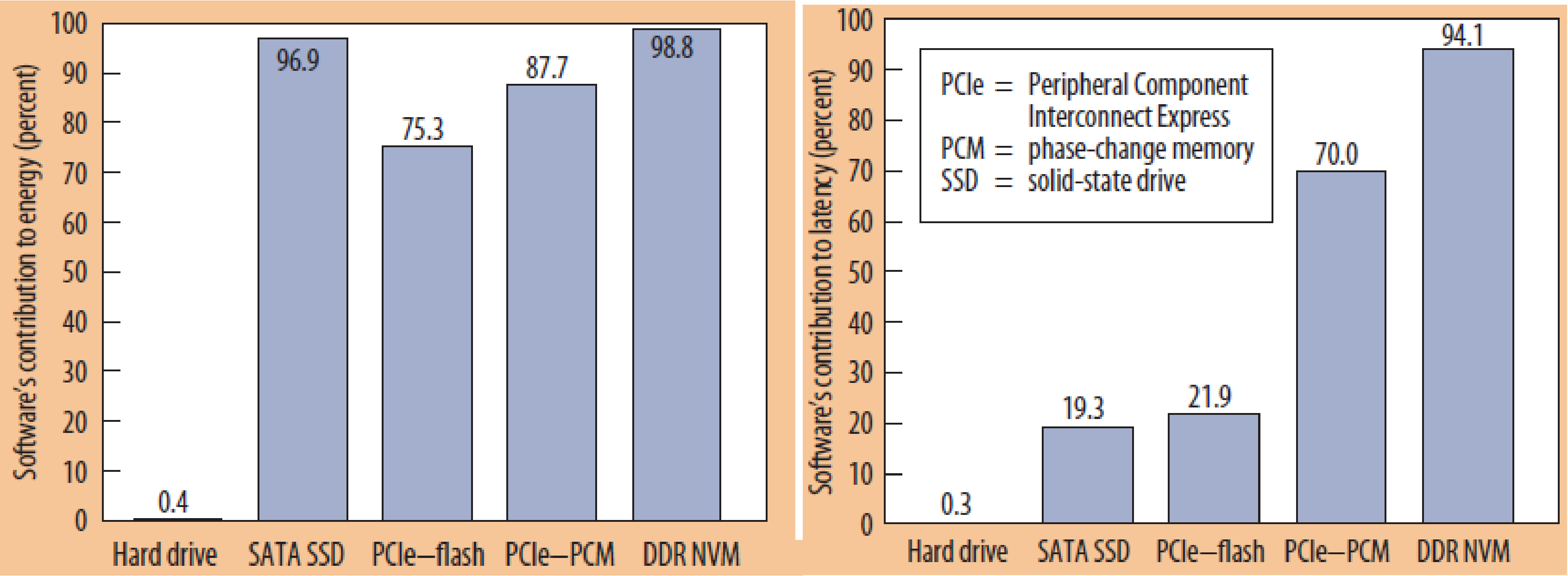
HDD

Buffer Cache

Replication
Encryption

IO Request Queue Schedule

Disks replaced by SCM directly



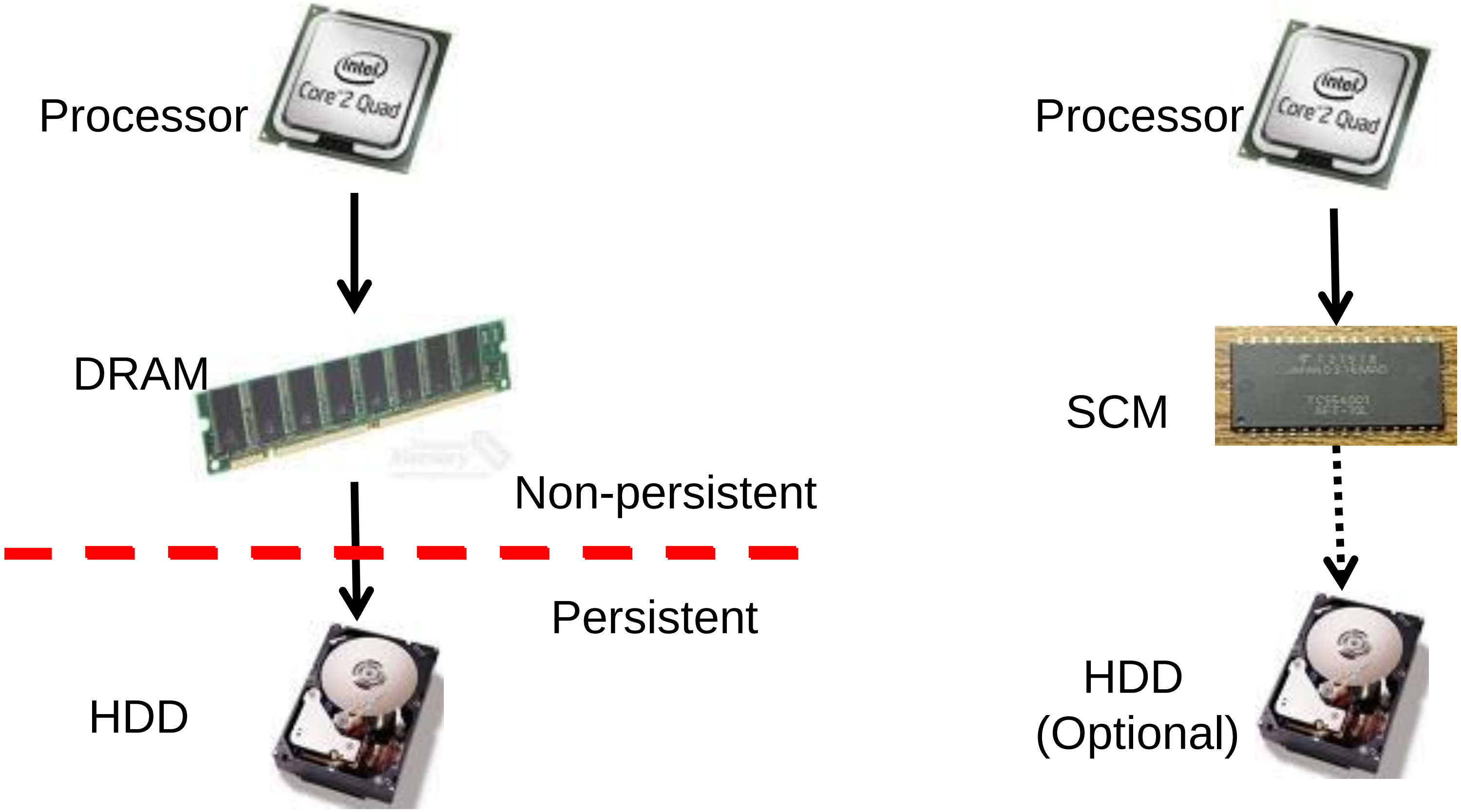
Rethinking the role and structure of software and hardware in storage systems

Steven Swanson and Adrian M. Caulfield. *Refactor, Reduce, Recycle: Restructuring the I/O Stack for the Future of Storage*. Computer, IEEE, 2013.

Persistency Moves Up to Memory Layer

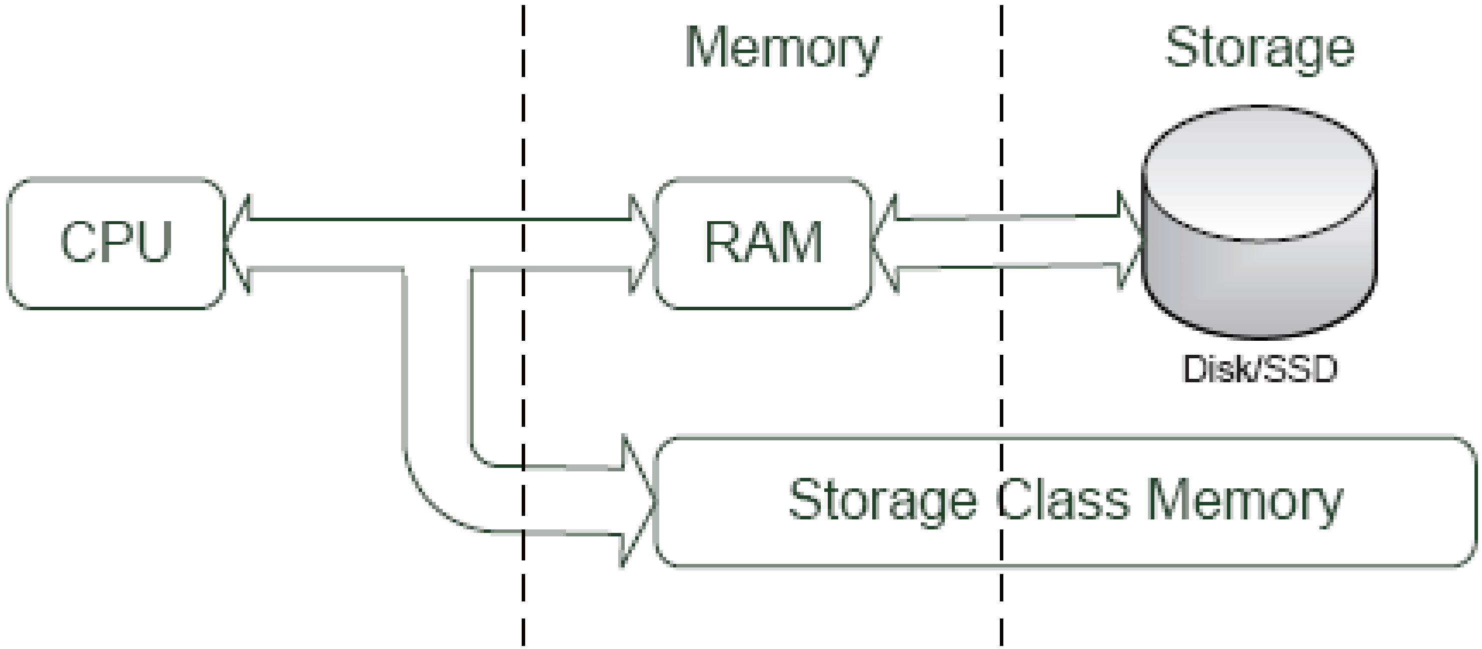
Current System

System with NVM



Memory is becoming update in-place. Current operating system is not aware of SCM. How can we leverage memory persistency?

How to utilize the SCM (Hardware)



- **Hardware:** SCM is attached to memory bus directly

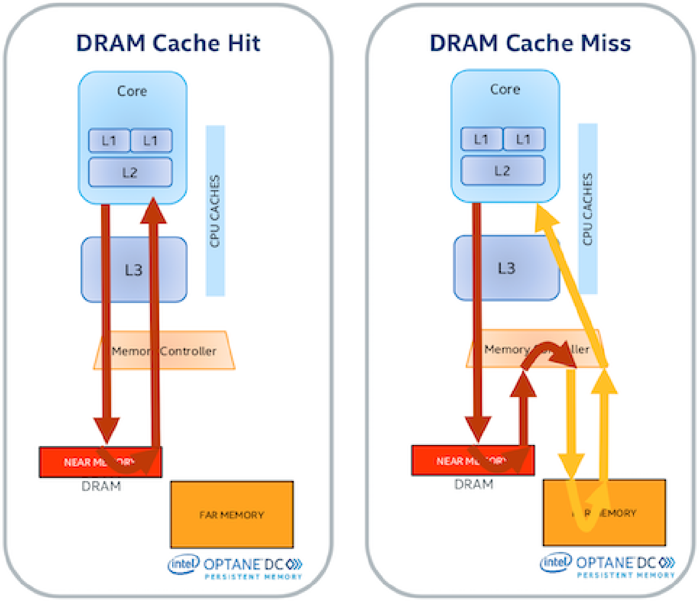
How to utilize the SCM (Software)

- **Software:**
  - 1. RamDisk mode, use a regular FS, nothing need to be changed. (Generic block layer overhead will affect the performance of whole system)
  - 2. Modify the existing memory based file system, such as tmpfs, ramfs. (not for persistent storage device, metadata: in-memory data structure. Modifying ≈ redesigning)
  - 3. Design **new FS for SCM**

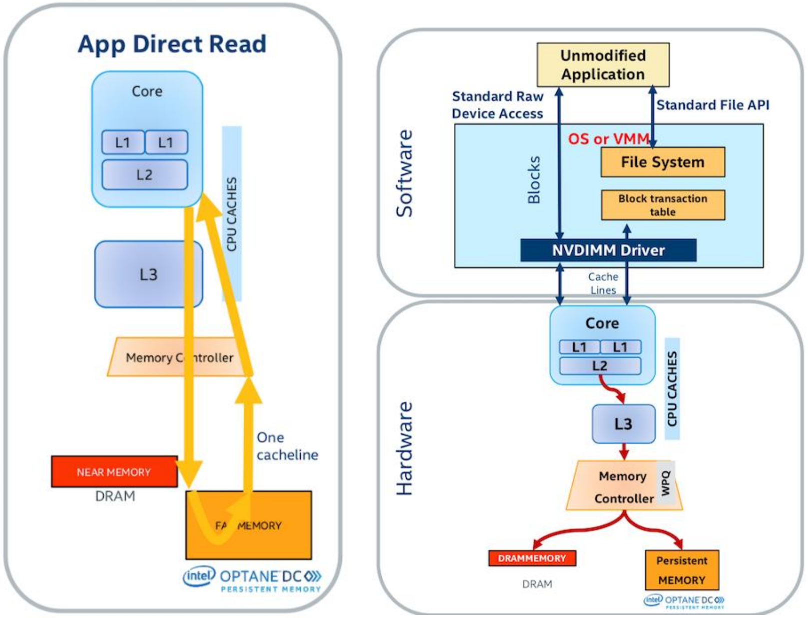
Intel X3D

- 英特尔已经推出Optane DC Persistent Memory模块，单条最大容量可达512GB Intel Xeon(Cascade Lake) 单处理器支持6条

- Memory Mode 和 App Direct Mode



Memory Mode

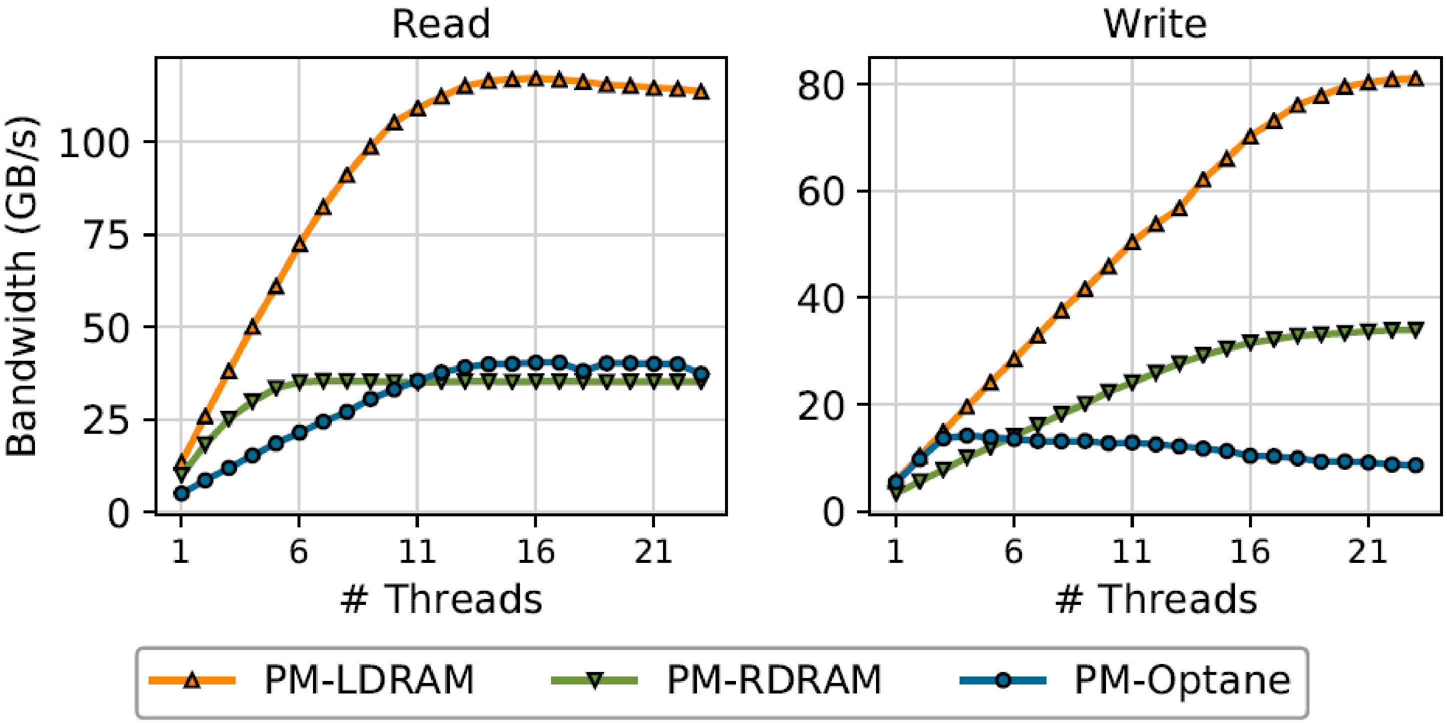


App Direct Mode

Intel X3D

英特尔Optane DC Persistent Memory基本性能

- ➢ 延迟：100~300 ns
- ➢ 带宽：读：39.4GB/s(单条6.6GB/s)
  写：13.9GB/s(单条2.3GB/s)

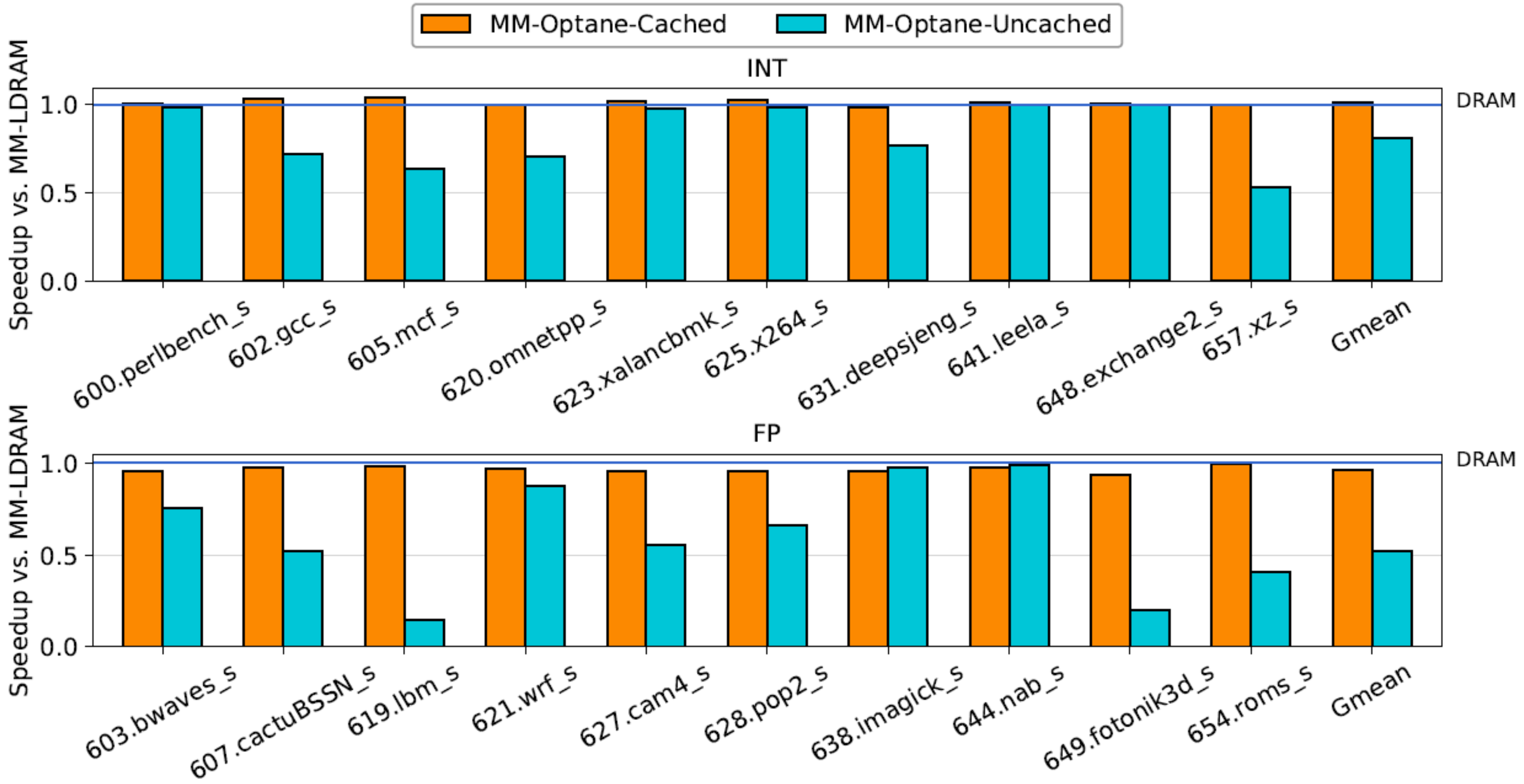


From NVSL of UCSD, 2019-08-09

Intel X3D

英特尔Optane DC Persistent Memory性能

- ➢ Optane DC as Main Memory

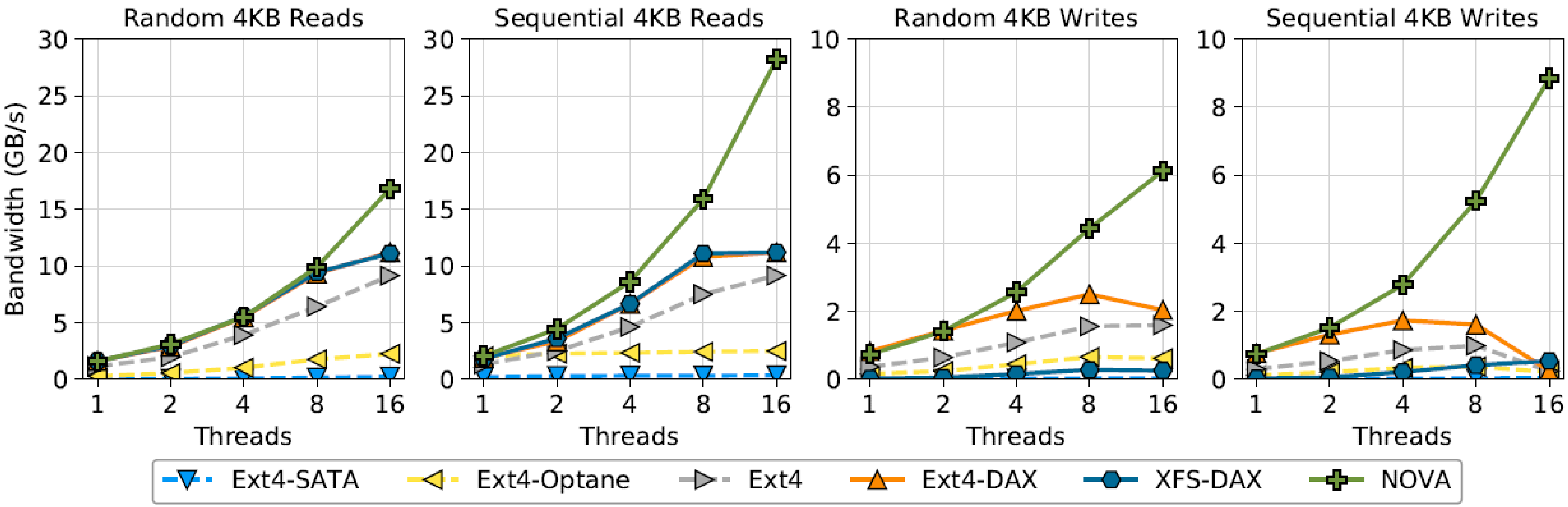


From NVSL of UCSD, 2019-08-09

Intel X3D

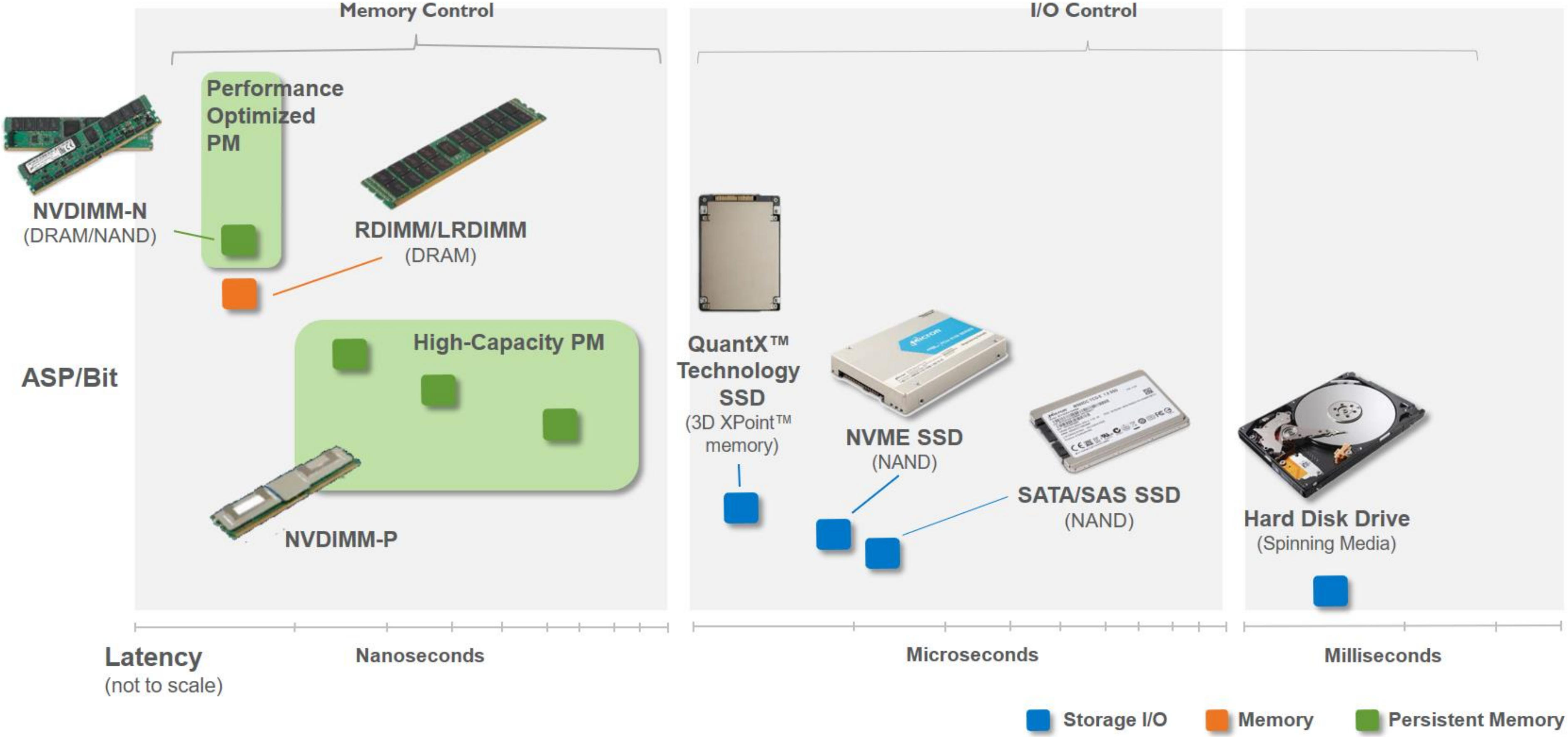
英特尔Optane DC Persistent Memory性能

- Optane DC as Persistent Storage



From NVSL of UCSD, 2019-08-09

Closing the Latency Gap



Throughput easy, latency hard

Throughput is easy

Latency is hard

Throughput is an engineering problem, latency is a physics problem!

Where Are We?

Thanks!