THE QUEST FOR LOW LATENCY STORAGE

Rick Coulson

Senior Fellow, Intel NVM Solutions Group

Outline

- The history of storage latency and where we stand today
- The promise of Storage Class Memory (SCM) and 3D Xpoint™ Memory
- Extensive compute platform changes driven by the quest for lower storage latency with SCM / 3D Xpoint™ Memory
  - As traditional storage
  - As persistent memory
- Innovation opportunities abound

1956: IBM RAMAC 350

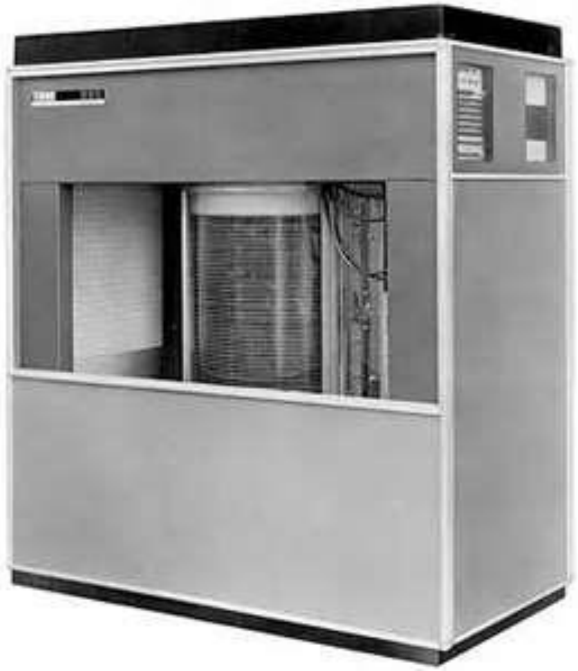
5 MBytes

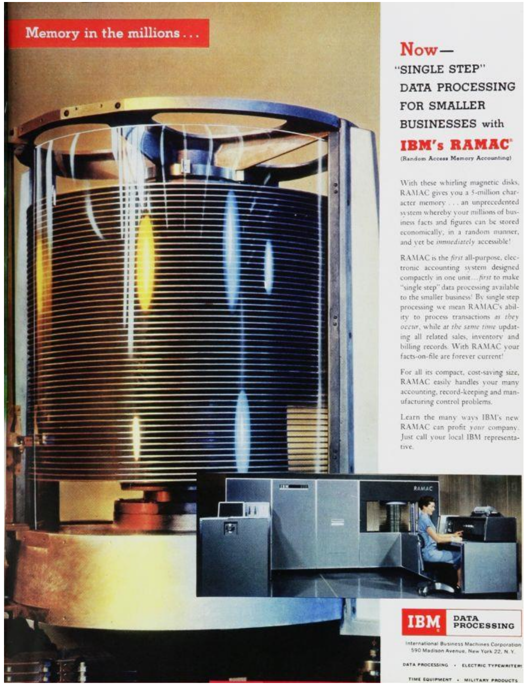
$57,000

$15200/Mbyte

~1.5 Random IOPs*

600ms latency





Memory in the millions . . .

Now—

"SINGLE STEP"
DATA PROCESSING
FOR SMALLER
BUSINESSES with
IBM's RAMAC®
(Random Access Memory Accounting)

With these whirling magnetic disks, RAMAC gives you a 5-million character memory . . . an unprecedented system whereby your millions of business facts and figures can be stored economically, in a random manner, and yet be *immediately* accessible!

RAMAC is the *first* all-purpose, electronic accounting system designed compactly in one unit . . . *first* to make "single step" data processing available to the smaller business! By single step processing we mean RAMAC's ability to process transactions *as they occur*, while *at the same time* updating all related sales, inventory and billing records. With RAMAC your facts-on-file are forever current!

For all its compact, cost-saving size, RAMAC easily handles your many accounting, record-keeping and manufacturing control problems.

Learn the many ways IBM's new RAMAC can profit *your* company. Just call your local IBM representative.

IBM® DATA PROCESSING

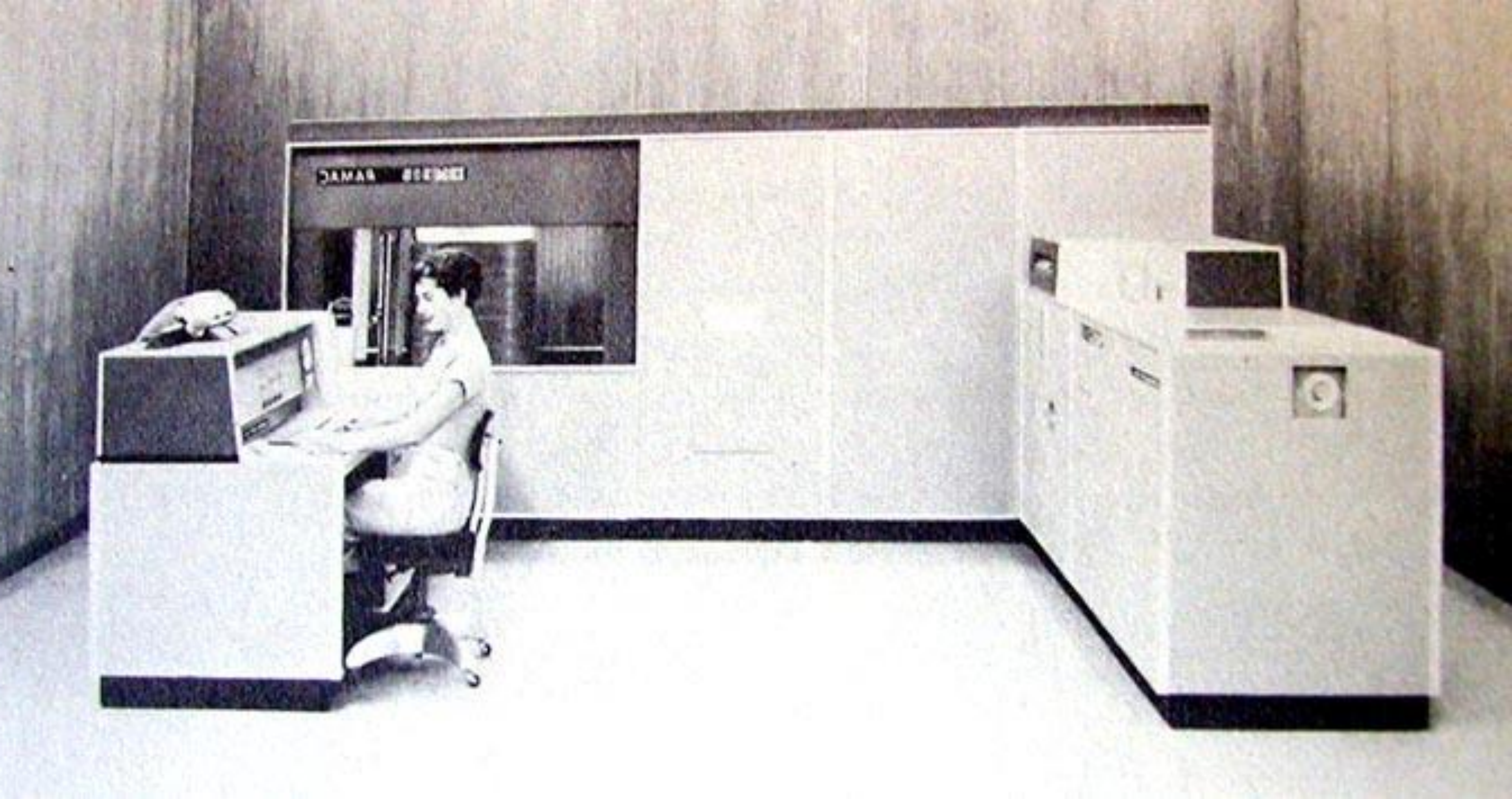
International Business Machines Corporation
590 Madison Avenue, New York 22, N. Y.

DATA PROCESSING • ELECTRIC TYPEWRITERS
TIME EQUIPMENT • MILITARY PRODUCTS

* IOPs depends on the workload and is a range

*Other names and brands may be claimed as the property of others.

PAA
IBM



1980: IBM 3350

300 Mbytes

$60,000

$200/MByte

30 random IOPs

33ms latency

1983: IBM3380

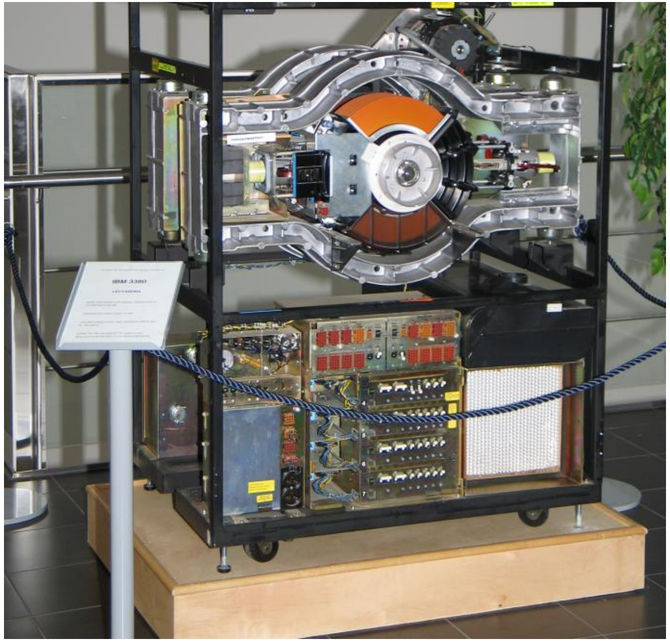
2.52 GBytes

$82,000

$36/MByte

~160 IOPS total

25ms latency



Two hard disk assemblies each with two independent actuators each accessing 630 MB gigabyte within one chassis

2007: 15K RPM HDD

15K RPM HDD

About 200 random IOPs*

~5ms latency

IOPS scaling problem was addressed through HDDs in parallel in Enterprise

* IOPs depends on the workload and is a range

2016: 10K RPM HDD

1.8TB

~150 IOPs

6.6ms latency

2016 NVMe NAND SSD

2TB

500,000+ IOPs

~60 usec latency

The Continuing Need For Lower Latency

RAMAC 350
600ms

~100x
Access time reduction

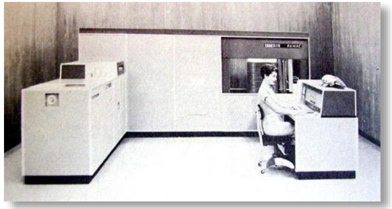
10K RPM
~6ms access

~10,000x
Access time reduction

NAND SSD
~60us access

59 Years

RAMAC 305
100 Hz best
case “clock”



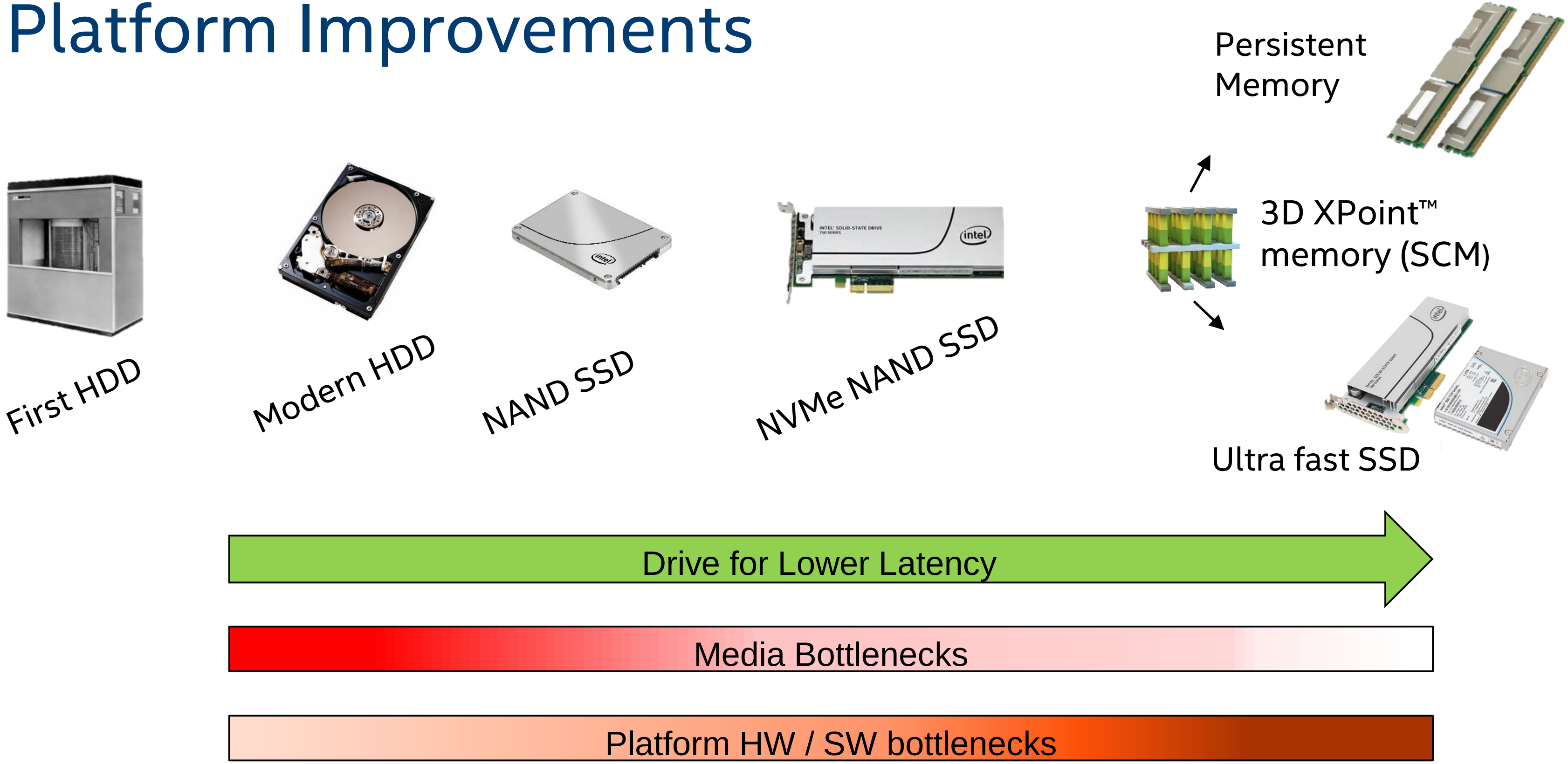
~40,000,000x
Clock speed increase

Core™ i7
~4 Ghz clock

Source: Wikipedia

*Other names and brands may be claimed as the property of others.

Lower Storage Latency Requires Media and Platform Improvements



First HDD

Modern HDD

NAND SSD

NVMe NAND SSD

3D XPoint™ memory (SCM)

Persistent Memory

Ultra fast SSD

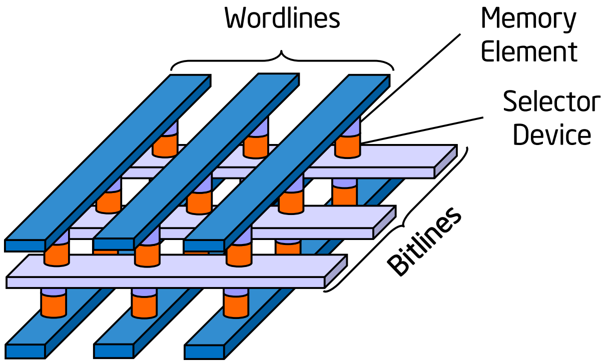
Drive for Lower Latency

Media Bottlenecks

Platform HW / SW bottlenecks

Addressing Media Latency: Next Gen NVM / SCM

Scalable Resistive Memory Element



Cross Point Array in Backend Layers ~4λ^2 Cell

Resistive RAM NVM Options

Family	Defining Switching Characteristics
Phase Change Memory	Energy (heat) converts material between crystalline (conductive) and amorphous (resistive) phases
Magnetic Tunnel Junction (MTJ)	Switching of magnetic resistive layer by spin-polarized electrons
Electrochemical Cells (ECM)	Formation / dissolution of “nano-bridge” by electrochemistry
Binary Oxide Filament Cells	Reversible filament formation by Oxidation-Reduction
Interfacial Switching	Oxygen vacancy drift diffusion induced barrier modulation

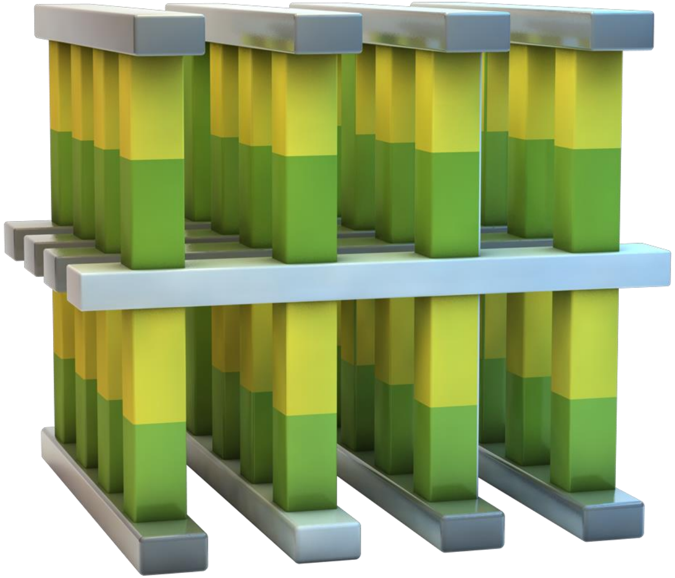
Scalable, with potential for near DRAM access times

3D XPoint™ Technology

Crosspoint Structure
Selectors allow dense packing and individual access to bits

↓

Scalable
Memory layers can be stacked in a 3D manner



Breakthrough Material Advances
Compatible switch and memory cell materials

↓

High Performance
Cell and array architecture that can switch states 1000x faster than NAND

A NEW CLASS OF NON-VOLATILE MEMORY

1000X
FASTER
THAN NAND

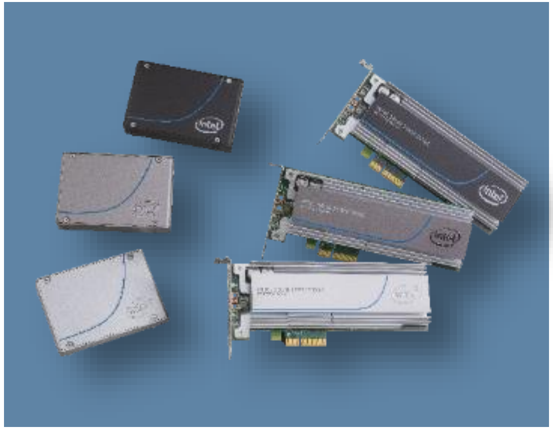
1000X
ENDURANCE
OF NAND

10X
DENSER
THAN DRAM

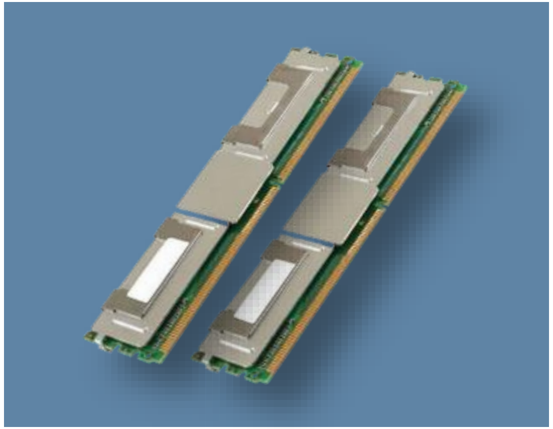
*Results have been estimated or simulated using internal analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance

3D XPoint™ Technology Instantiation

INTEL® OPTANE™ SSDS



DIMMS BASED ON 3D XPOINT™

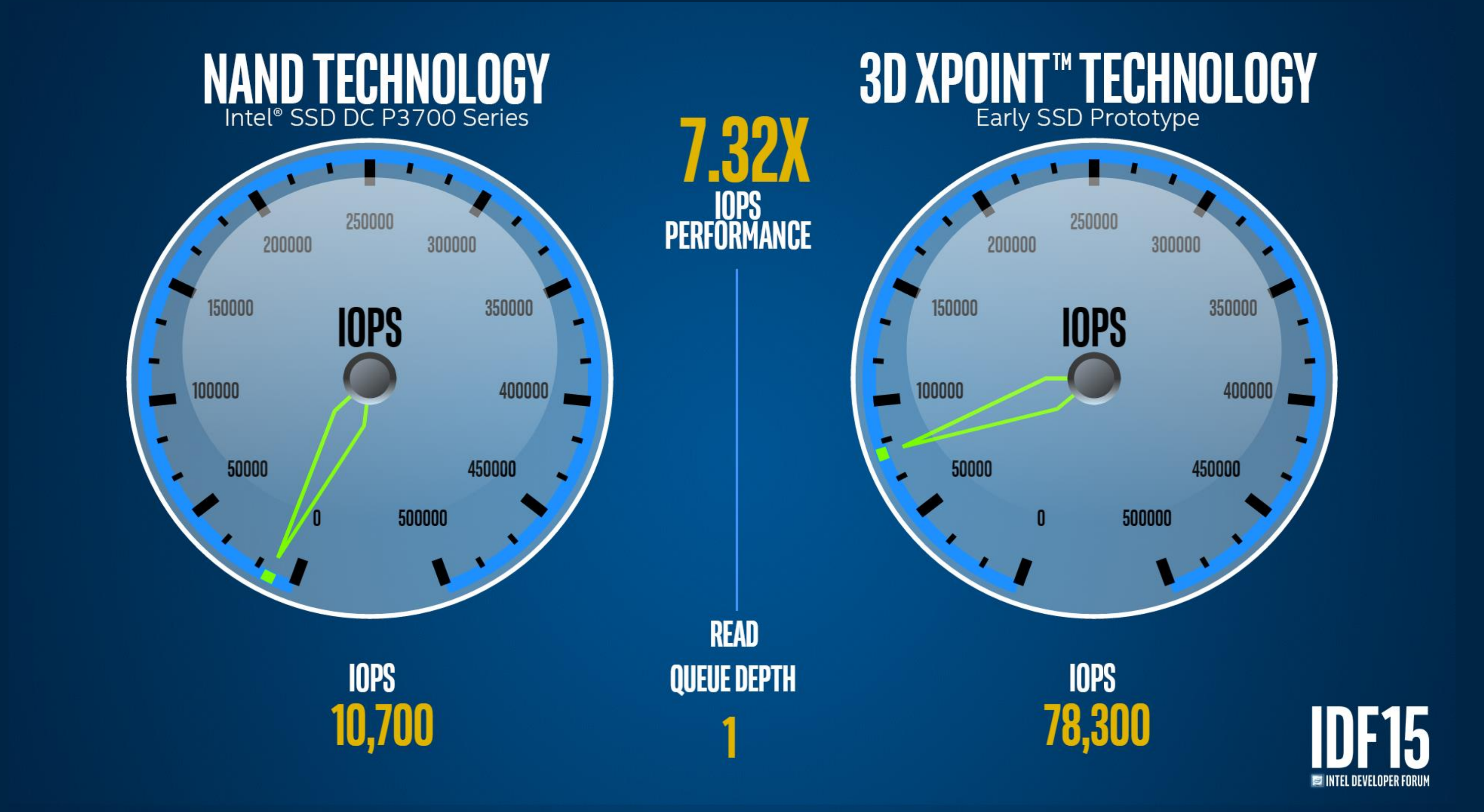


3D Xpoint™ Technology Video

Please excuse the marketing

3D Xpoint™ Technology Video

Demonstration of 3D Xpoint™ SSD Prototype



NAND TECHNOLOGY
Intel® SSD DC P3700 Series

3D XPOINT™ TECHNOLOGY
Early SSD Prototype

7.32X
IOPS PERFORMANCE

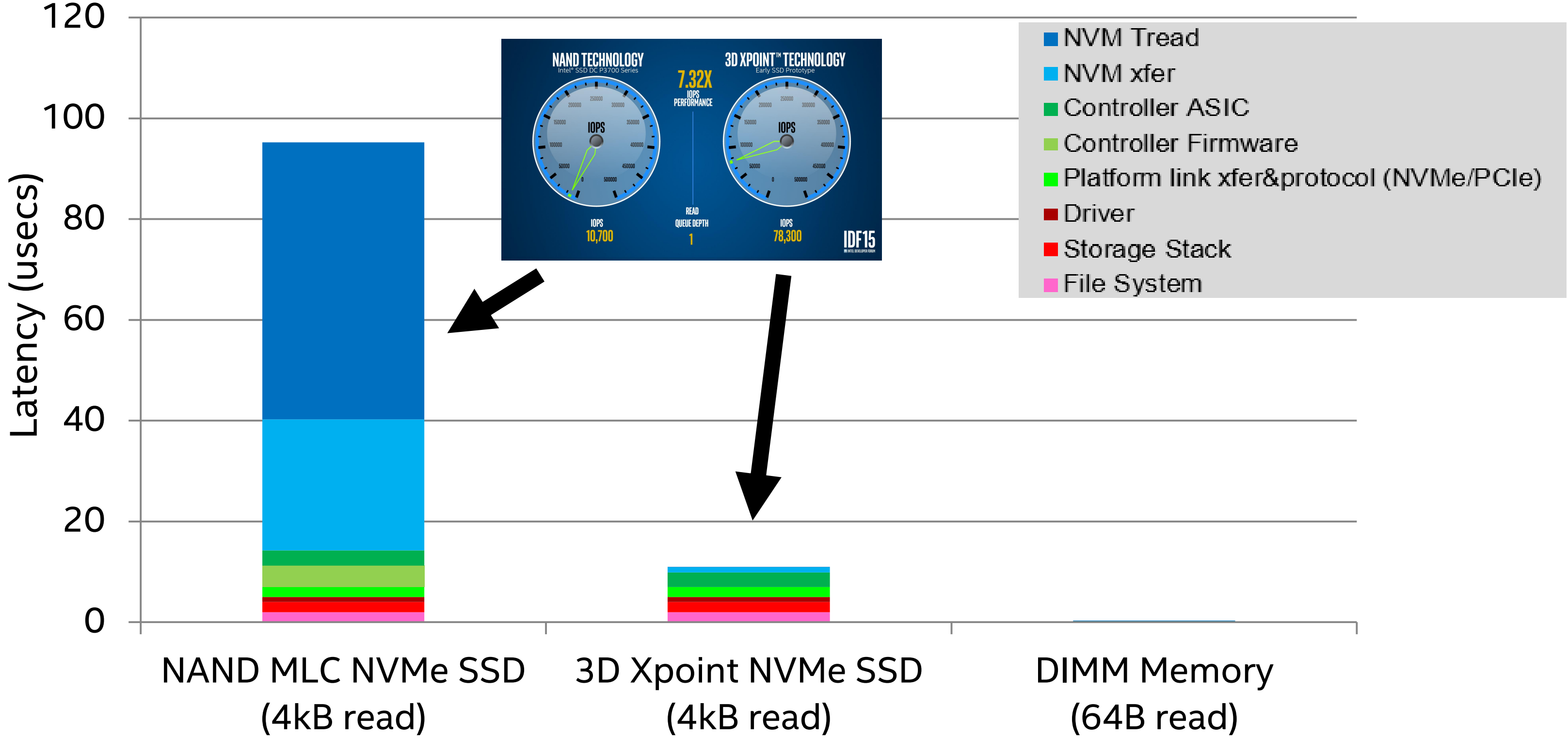
IOPS
10,700

READ
QUEUE DEPTH
1

IOPS
78,300

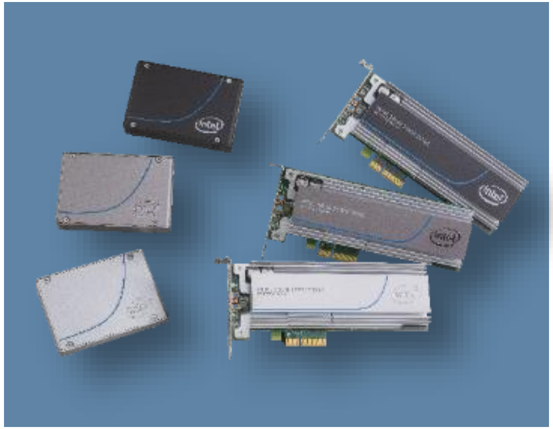
IDF15
INTEL DEVELOPER FORUM

Need to Address System Architecture To Go Lower

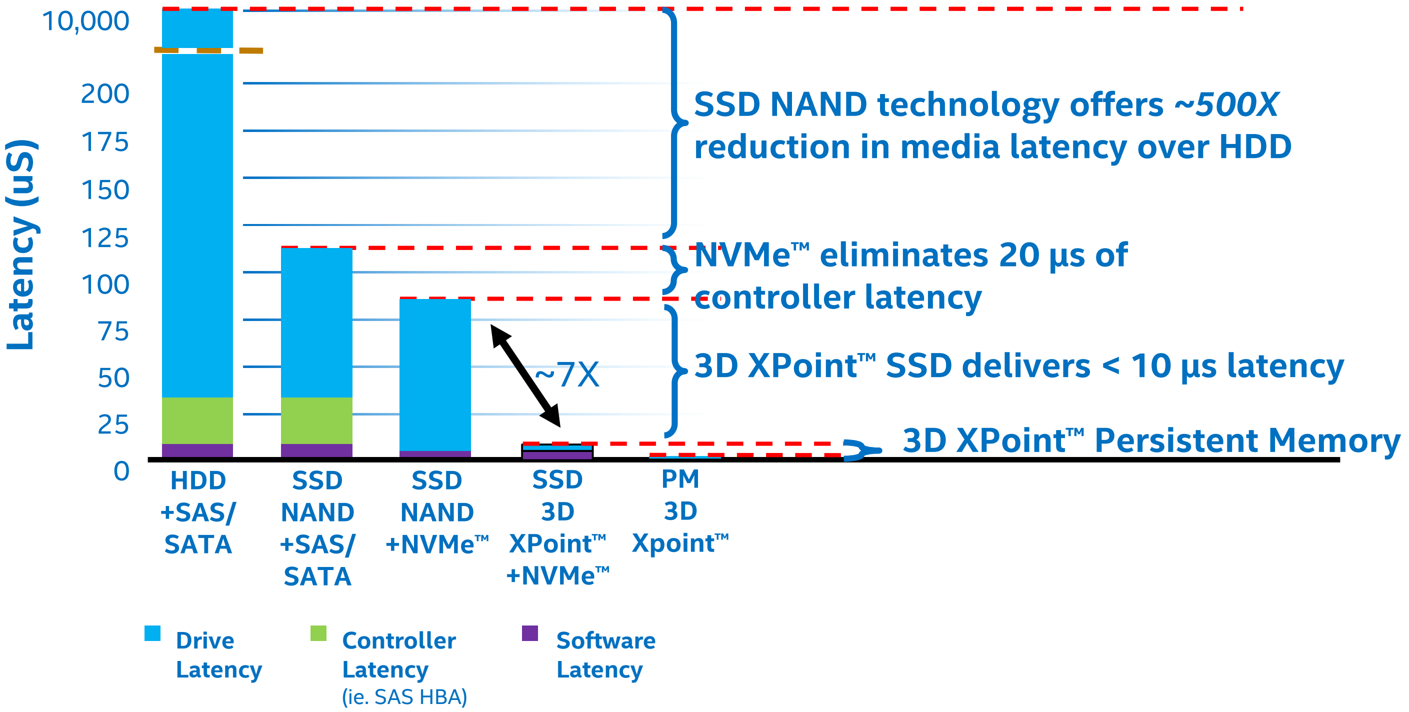


Block Storage Platform Changes

INTEL® OPTANE™ SSDS



Addressing Interface Efficiency With NVMe / PCI



SSD NAND technology offers ~*500X* reduction in media latency over HDD

NVMe™ eliminates 20 µs of controller latency

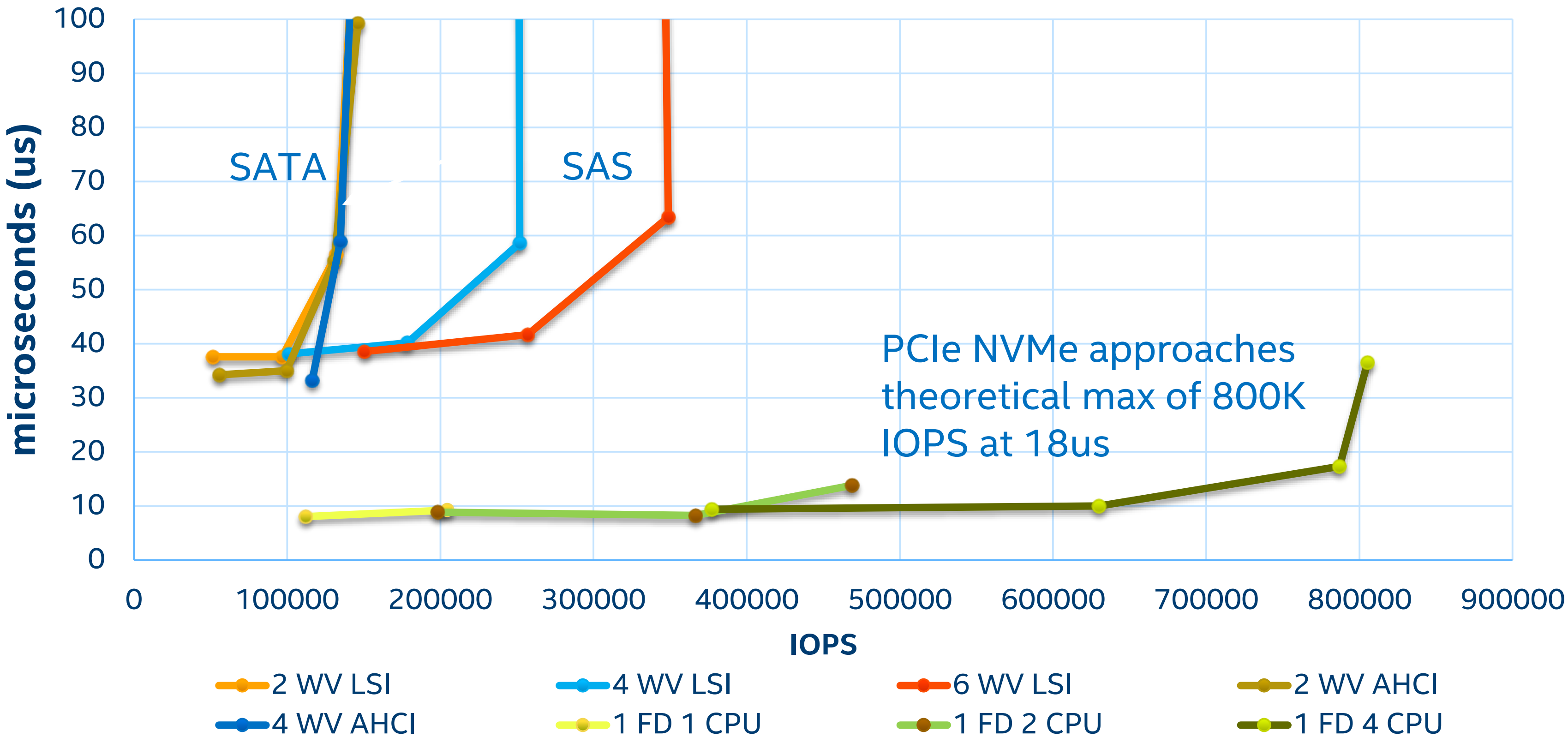
3D XPoint™ SSD delivers < 10 µs latency

3D XPoint™ Persistent Memory

Source: Storage Technologies Group, Intel

NVMe Delivers Superior Latency

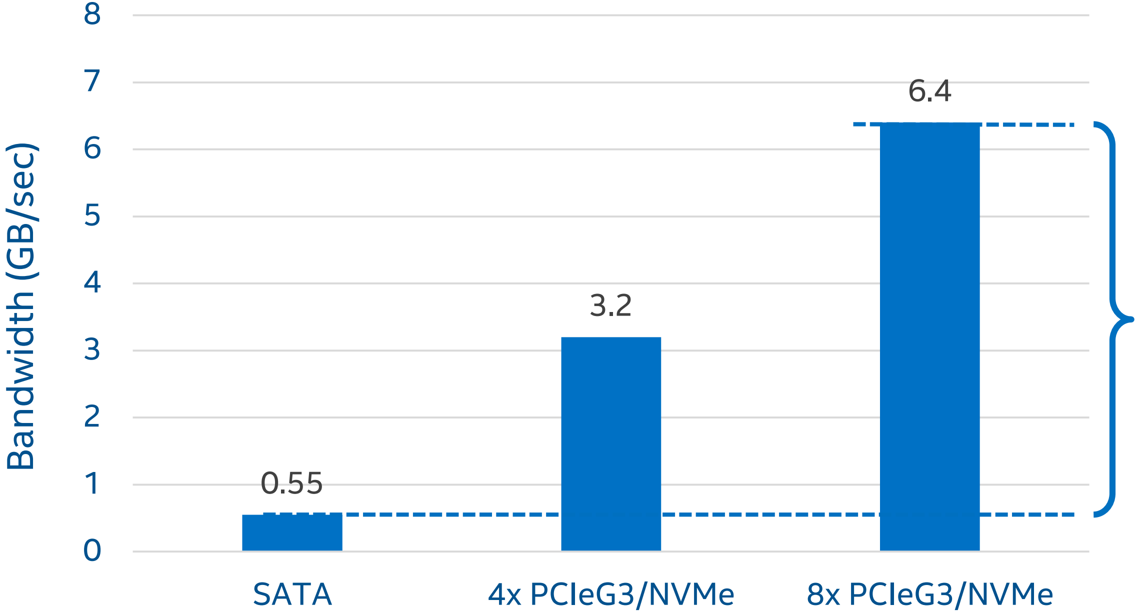
Platform HW/SW Average Latency Excluding Media 4KB



PCIe NVMe approaches theoretical max of 800K IOPS at 18us

Source: Storage Technologies Group, Intel

NVMe/PCIe Provides More Bandwidth



PCIe/NVMe provides more than 10X the Bandwidth of SATA. Even More with Gen 4

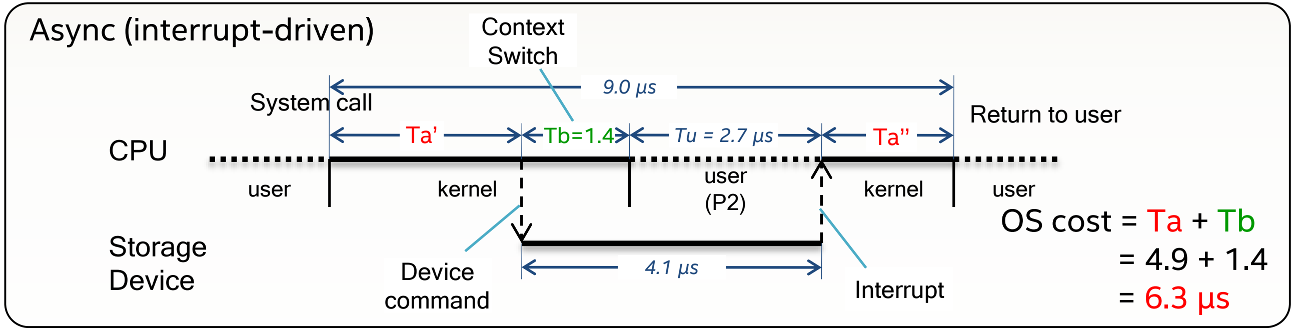
Source: Storage Technologies Group, Intel

Storage SW Stack Optimizations

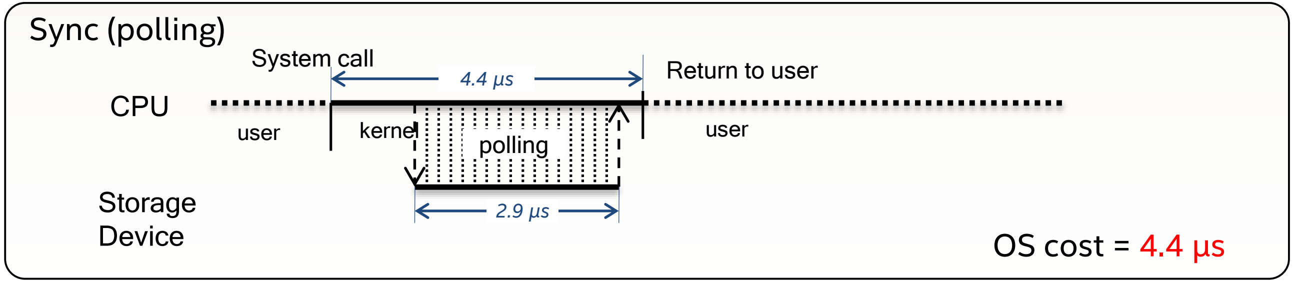
Much of the storage stack designed with HDDs latencies in mind

- No point in optimizing until now
- Example: Paging algorithms with seek optimization and grouping

Synchronous Completion for Queue Depth 1?



OS cost = Ta + Tb
= 4.9 + 1.4
= 6.3 µs



OS cost = 4.4 µs

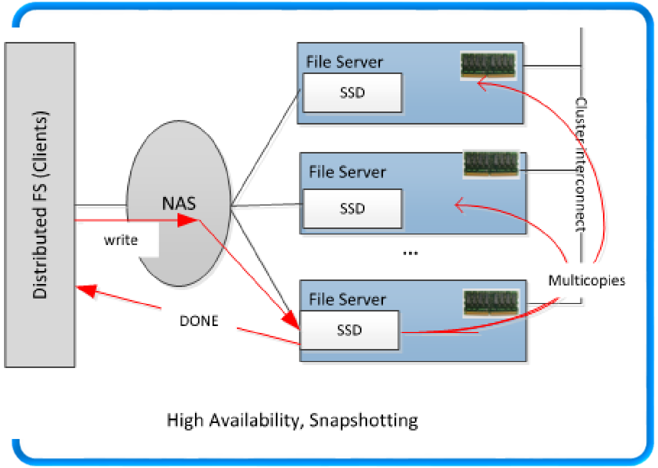
From Yang: FAST '12 -10th USENIX Conference on File and Storage Technologies

Standards for Low Latency Replication

In most Datacenter usage models, a storage write does not “count” until replicated

High replication overhead diminishes the performance differentiation of 3D XPoint™ technology

NVMe over Fabrics is a developing SNIA specification for low overhead replication



High Availability, Snapshotting

Summary: Block Storage Platform Changes

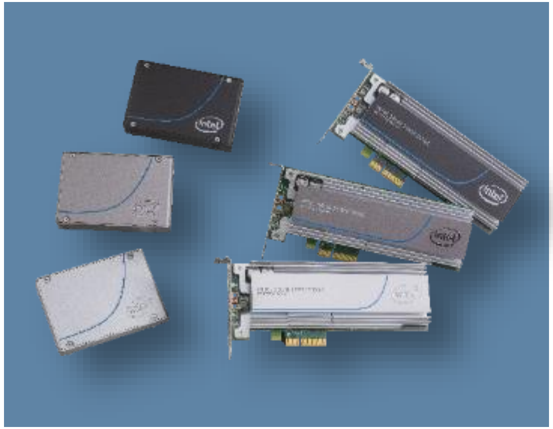
Move to PCIe based storage

Streamlined command set NVMexpress

OS / SW stack optimizations

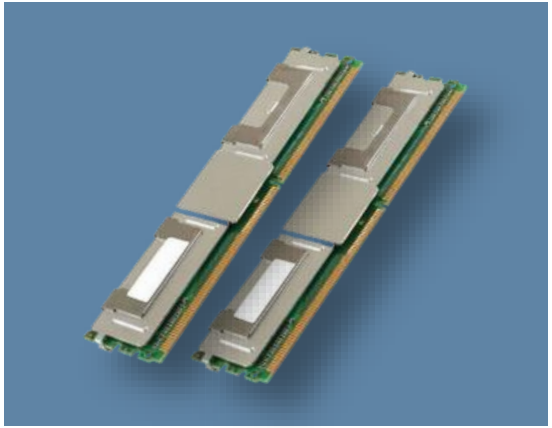
Fast replication standards

INTEL® OPTANE™ SSDS



Persistent Memory Oriented Platform Changes

DIMMS BASED ON 3D XPOINT™



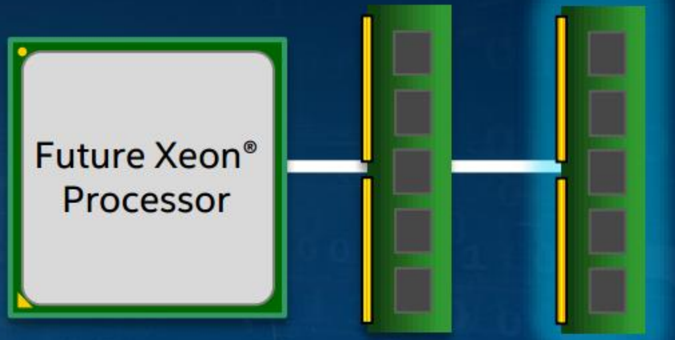
INTEL DIMMs

Based on 3D XPoint™ Technology

- DDR4 electrical & physical compatible
- Required support delivered by next generation Intel® Xeon® platform
- Up to 4X system memory capacity, at significantly lower cost than DRAM
- Can deliver big memory benefits without modifications to OS or applications

INTEL DIMM

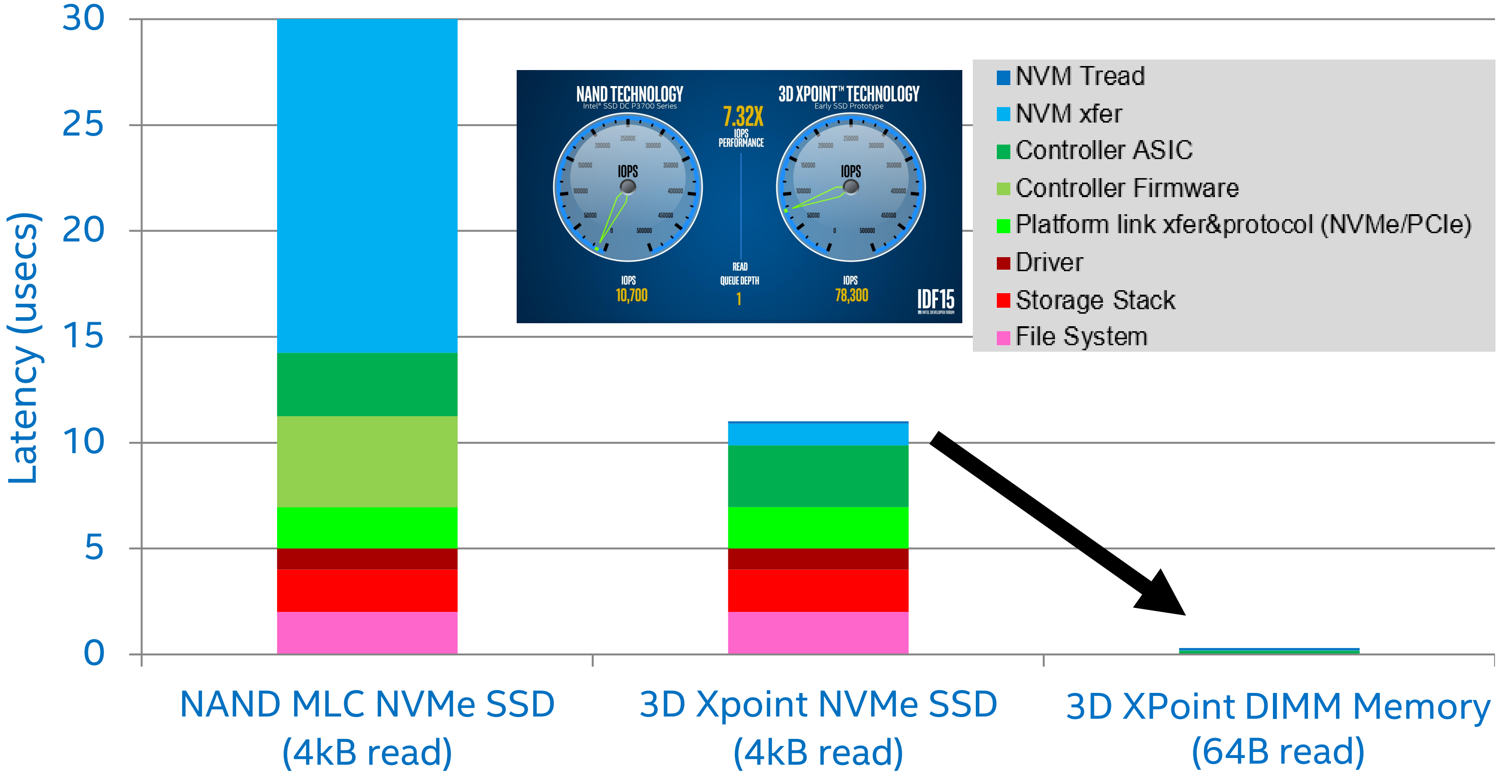
(based on 3D XPOINT™ Technology)



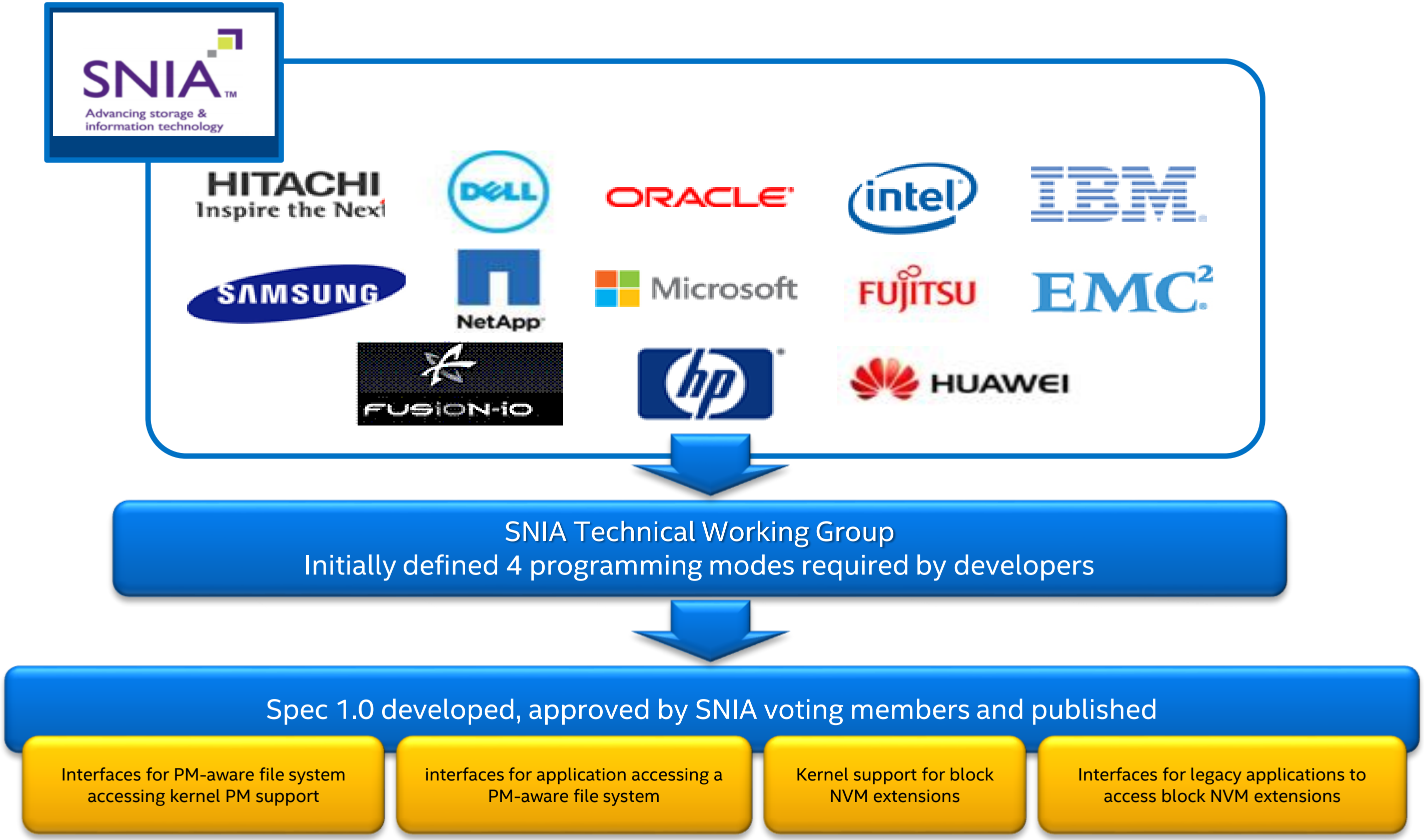
DDR4 DIMM

(acts as write-back cache)

Why Persistent Memory?



Open NVM Programming Model



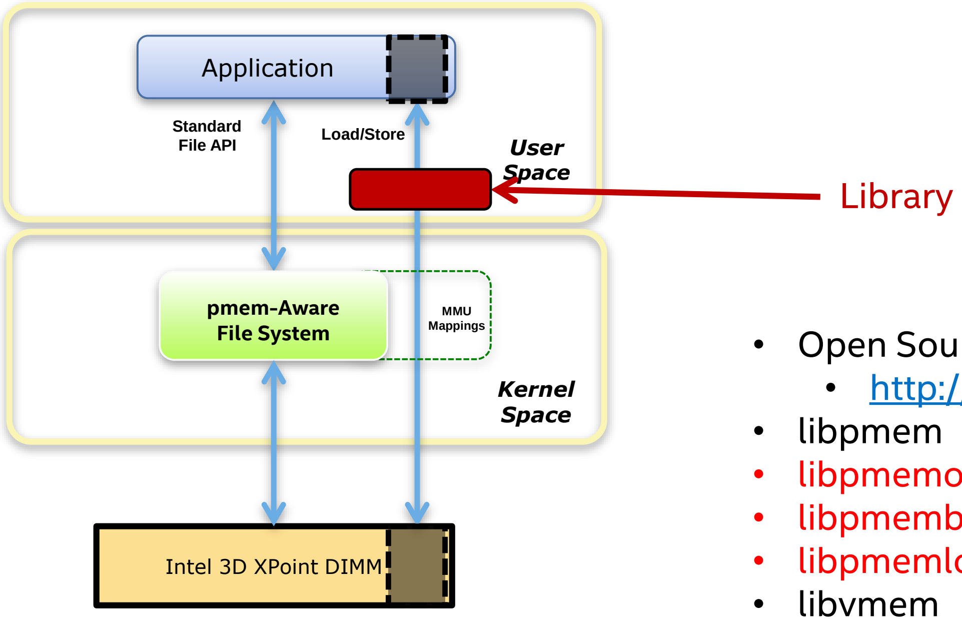
SNIA Technical Working Group
Initially defined 4 programming modes required by developers

Spec 1.0 developed, approved by SNIA voting members and published

- Interfaces for PM-aware file system accessing kernel PM support
- interfaces for application accessing a PM-aware file system
- Kernel support for block NVM extensions
- Interfaces for legacy applications to access block NVM extensions

NVM Library: pmem.io

64-bit Linux Initially

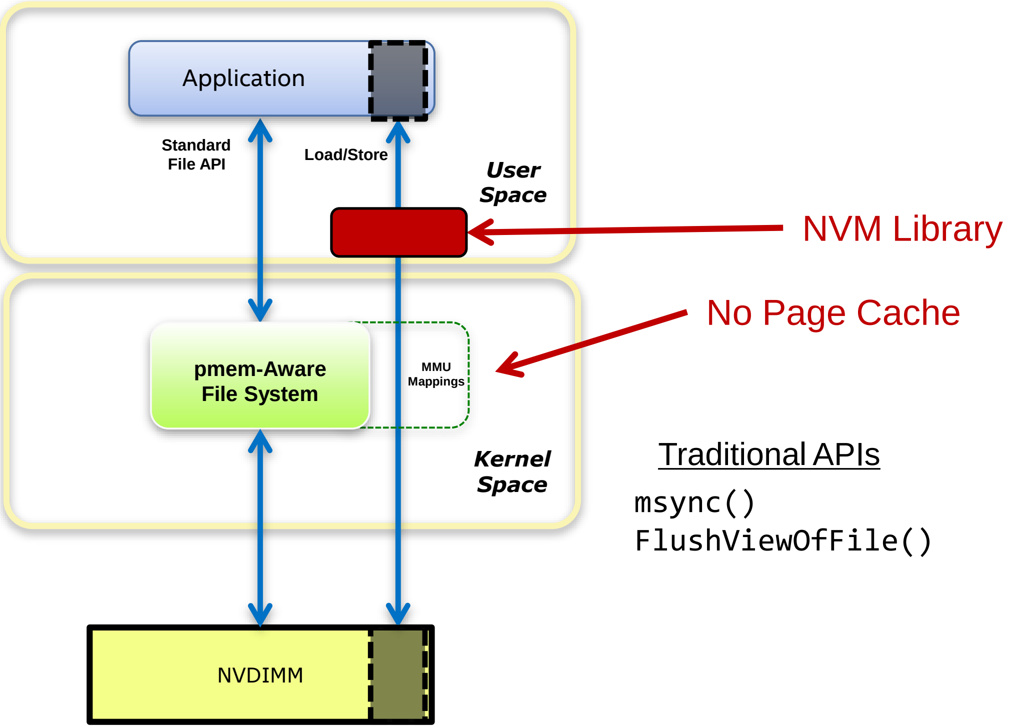


33

- Open Source
  - http://pmem.io
- libpmem
- libpmemobj
- libpmemblk
- libpmemlog
- libvmem
- libvmmalloc

Transactional

Write I/O Replaced with *Persist Points*



Traditional APIs

```
msync()
FlushViewOfFile()
```

NVML API

```
pmem_persist()
```

Operating System Support for Persistent Memory

SDC

STORAGE DEVELOPER CONFERENCE

SNIA ■ SANTA CLARA, 2015

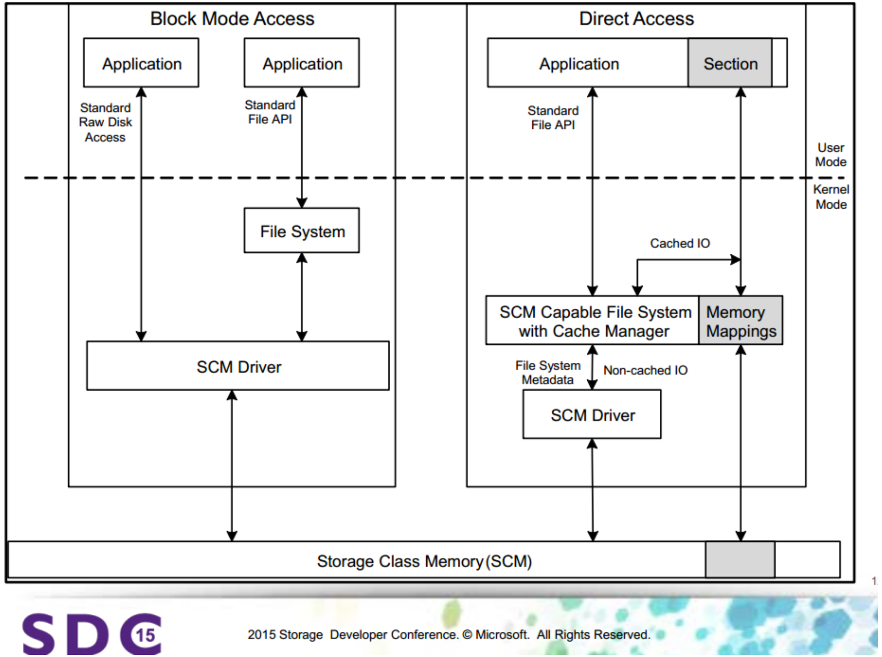
Storage Class Memory Support in the Windows Operating System

Neal Christiansen

Principal Development Lead

Microsoft

nealch@microsoft.com

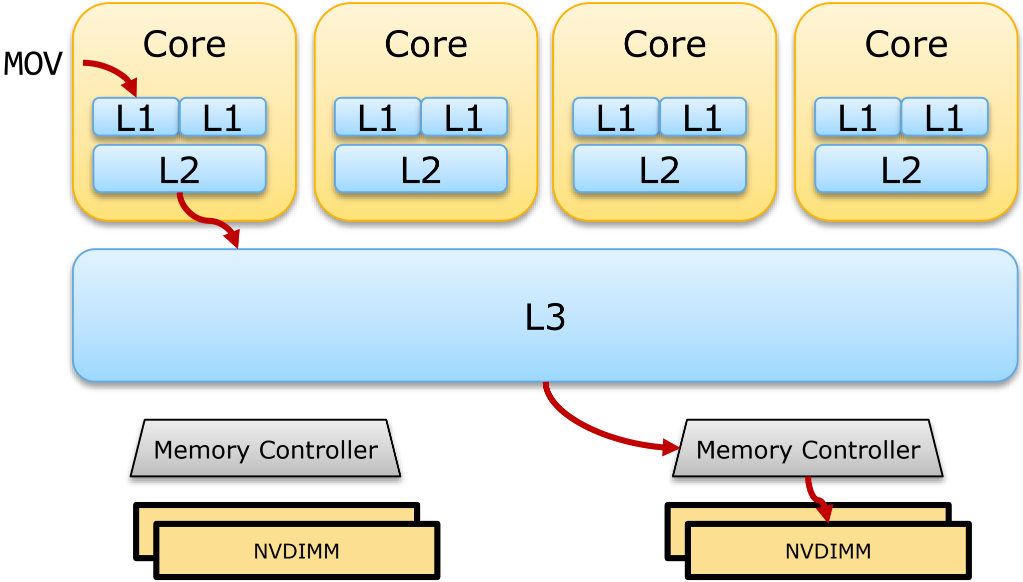


1

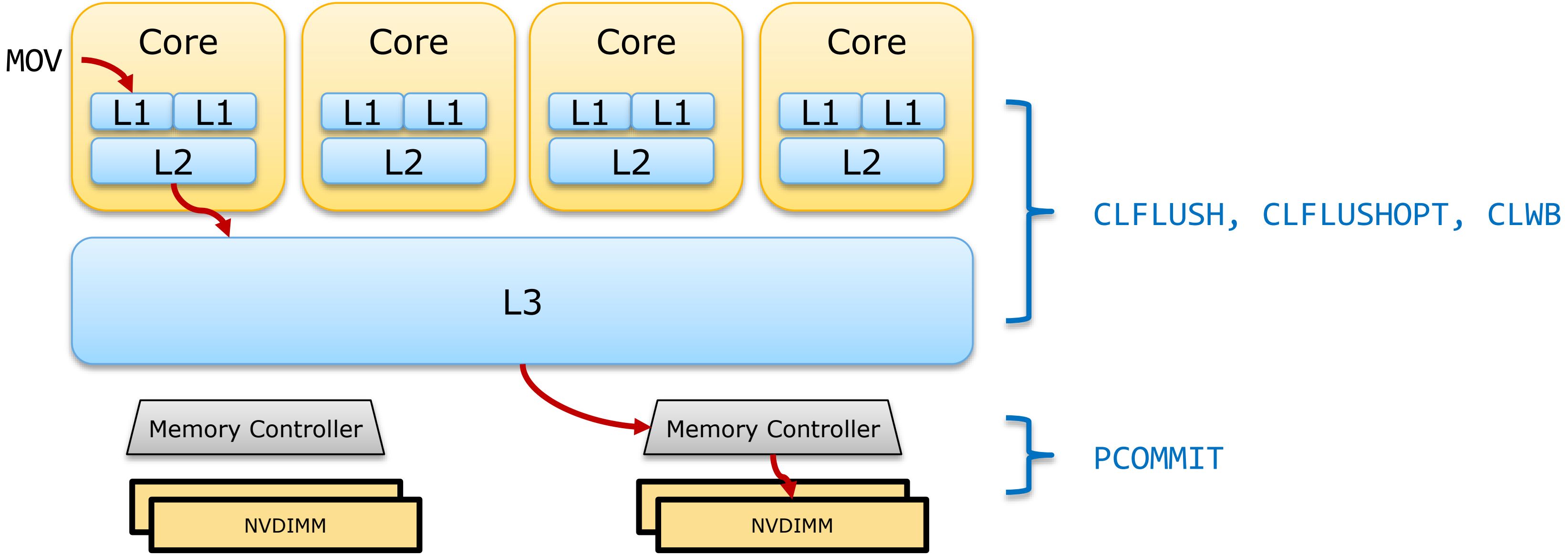
SDC 15

2015 Storage Developer Conference. © Microsoft. All Rights Reserved.

The Data Path



New Instructions For Flushing Writes



Flushing Writes from Caches

Instruction	Meaning
`CLFLUSH addr`	Cache Line Flush: Available for a long time
`CLFLUSHOPT addr`	Optimized Cache Line Flush: New to allow concurrency
`CLWB addr`	Cache Line Write Back: Leave value in cache for performance of next access

Flushing Writes from Memory Controller

Instruction	Meaning
`PCOMMIT`	Persistent Commit: Flush stores accepted by memory subsystem
`Asynchronous DRAM Refresh`	Flush outstanding writes on power failure Platform-Specific Feature

Example Code

```
MOV X1, 10
MOV X2, 20
.
MOV R1, X1
.
.
.
CLFLUSHOPT X1
CLFLUSHOPT X2
.
.
SFENCE
PCOMMIT
.
.
SFENCE
```

Comments

X2,X1 are in pmem

Stores to X1 and X2 are globally visible, but **may not** be persistent

X1 and X2 moved from caches to memory

Ensures **PCOMMIT** has completed

Join the Discussion about Persistent Memory

Learn about the Persistent Memory programming model

- http://www.snia.org/forums/sssi/nvmp

Join the pmem NVM Libraries Open Source project

- http://pmem.io

Read the documents and code supporting ACPI 6.0 and Linux NFIT drivers

- http://www.uefi.org/sites/default/files/resources/ACPI_6.0.pdf
- https://git.kernel.org/cgit/linux/kernel/git/djbw/nvdimm.git/log/?h=nd
- https://github.com/pmem/ndctl
- http://pmem.io/documents/
- https://github.com/01org/prd

Intel Architecture Instruction Set Extensions Programming Reference

- https://software.intel.com/en-us/intel-isa-extensions

Intel 3D XPoint^{TM} Memory

- http://www.intel.com/content/www/us/en/architecture-and-technology/non-volatile-memory.html

Persistent Memory Summary

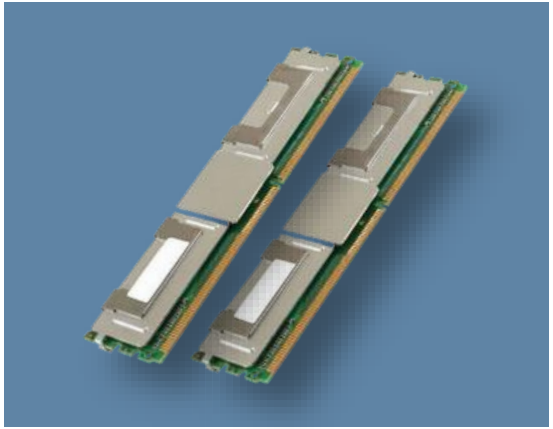
New storage model for low latency

New instructions to support persistence

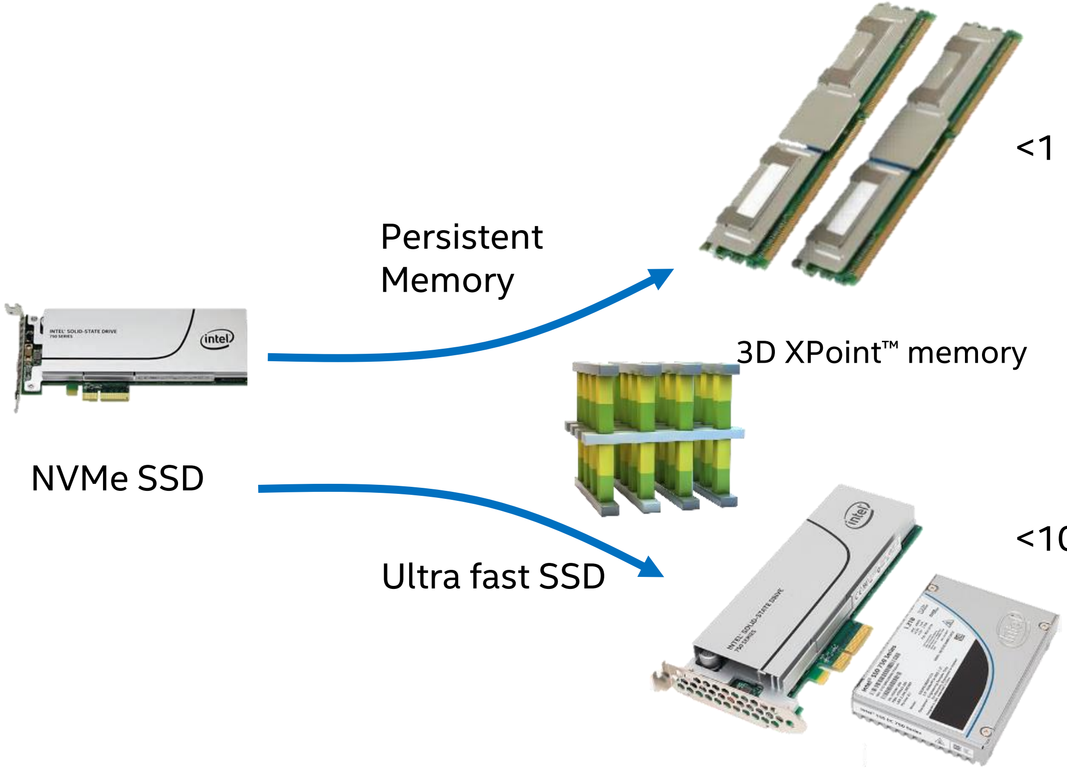
OS support

Lots of innovation opportunity

DIMMS BASED ON 3D XPOINT™



Low Latency Ahead



Persistent Memory

<1 usec

3D XPoint™ memory

NVMe SSD

Ultra fast SSD

<10 usec

intel®