# RAID on CPU

## RAID for NVMe SSDs without a RAID Controller Card

# Today's Presenters

**Paul Talbut**
**SNIA EMEA General Manager**

**Fausto Vaninetti**
**Senior Technical Solution Architect, Cisco Systems & Board Advisor SNIA EMEA**

**Igor Konopko**
**Software Engineer**
**Intel**

SNIA EMEA

# SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

# SNIA-At-A-Glance

**185**
industry leading organizations

**2,000**
active contributing members

**50,000**
IT end users & storage pros worldwide

SNIA EMEA

# Agenda

1. NVMe SSDs: Opportunity and Challenge
2. Back to Basics: RAID Levels and Write Penalty
3. Intel VROC: an Overview
4. Practical use cases
5. What's Next

# NVMe SSDs: Opportunity and Challenge

# High Impact Technology Ingredients: NVMe Drives

## Unlocking the drive bottleneck

Similarly priced SATA and NVMe drives

**Actual SATA SSD vs NVMe SSD at Similar $/TB**

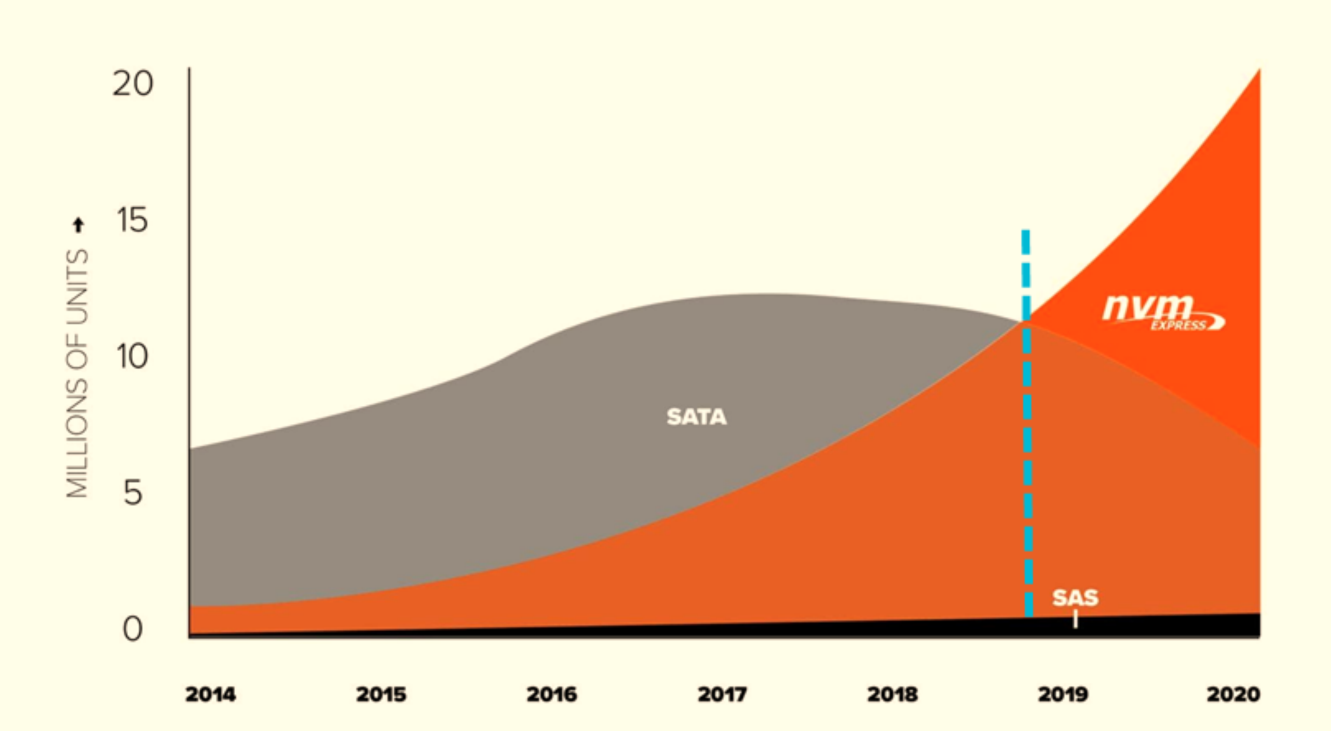
| | Vendor A<br>SATA SSD<br>**3.8TB** | Vendor B<br>NVMe SSD<br>**4TB** | |
|---|---|---|---|
| Max Random Read (KIOPS) | 97 | 361 | ~4X |
| Max Random Write (KIOPS) | 24 | 47 | ~2X |
| Max Sequential Read (MB/s) | 520 | 3100 | ~6X |
| Max Sequential Write (MB/s) | 480 | 1200 | 2.5X |

- Higher Perf Means Workload Acceleration
- Higher Perf and Higher Capacity NVMe Drives Mean Higher Workload Consolidation

# NVMe SSD Sales Have Surpassed SATA/SAS HDD

- higher performance
- new form factors
- shrinking price difference

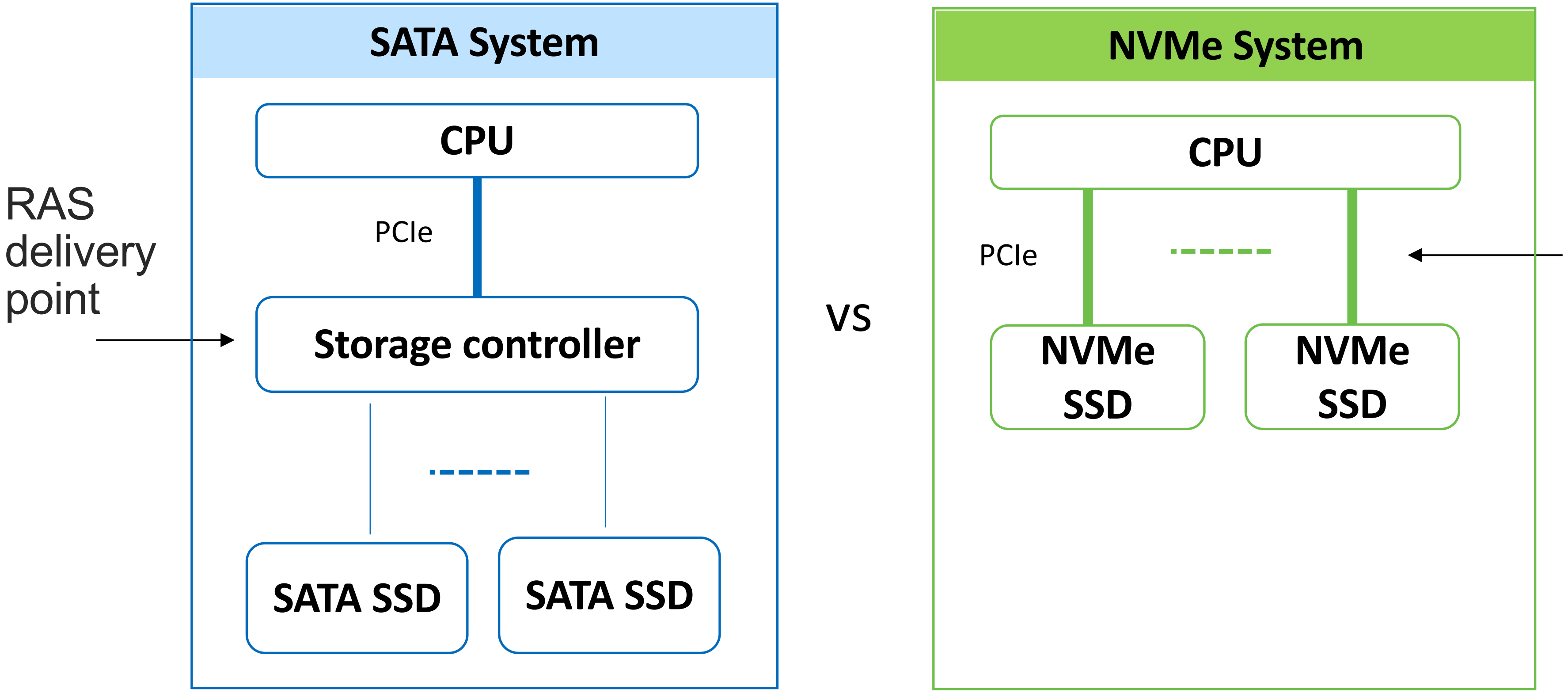



SNIA EMEA

# SATA vs NVMe Architecture: What About RAS?

## Hot Plug, Surprise Removal, LED management, Data Protection




- Where is the storage controller?
- Where is the RAS delivery point?
- Is surprise removal of NVMe drives possible?
- What about location LED on NVMe drives?
- NVMe specification for Hot Plug is not there yet, should we give up?
- How to implement RAID for NVMe SSDs?

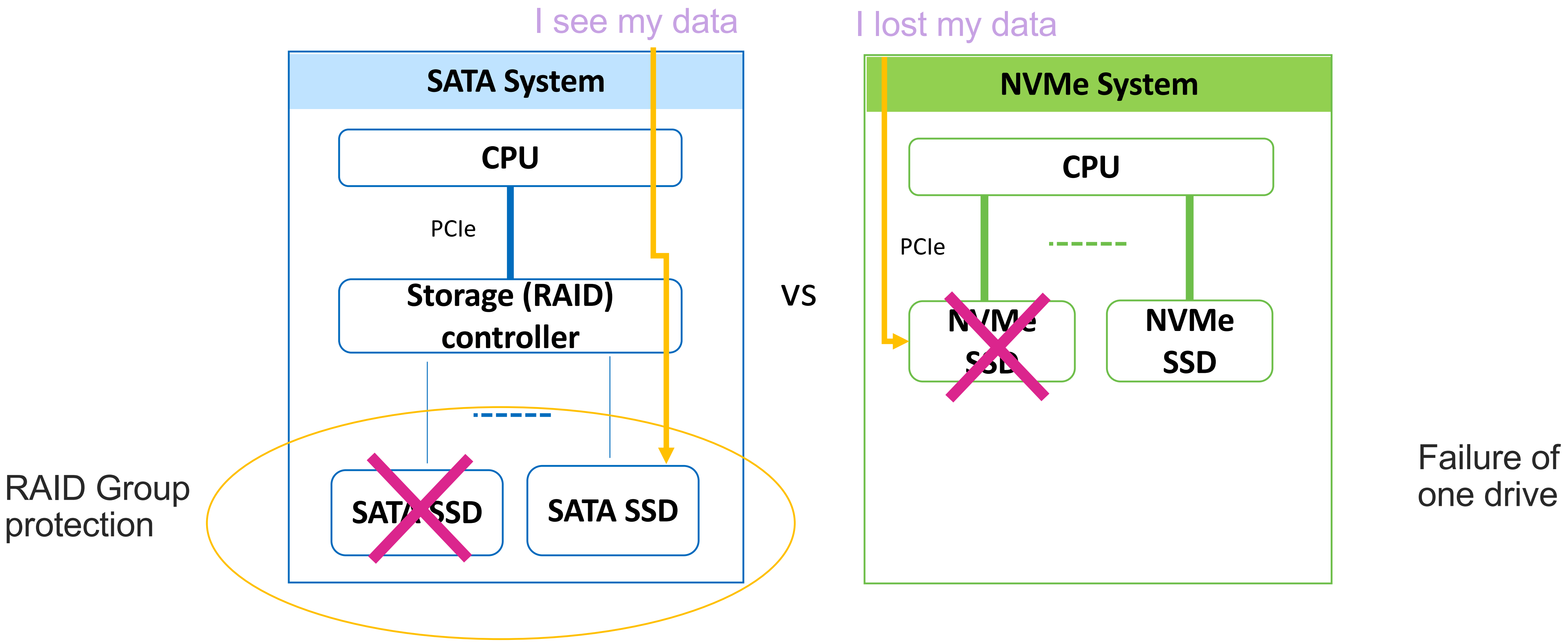
RAS = Reliability, Availability, Serviceability

SNIA EMEA

# Absence of RAS Means...

Drive Failure = Data Loss



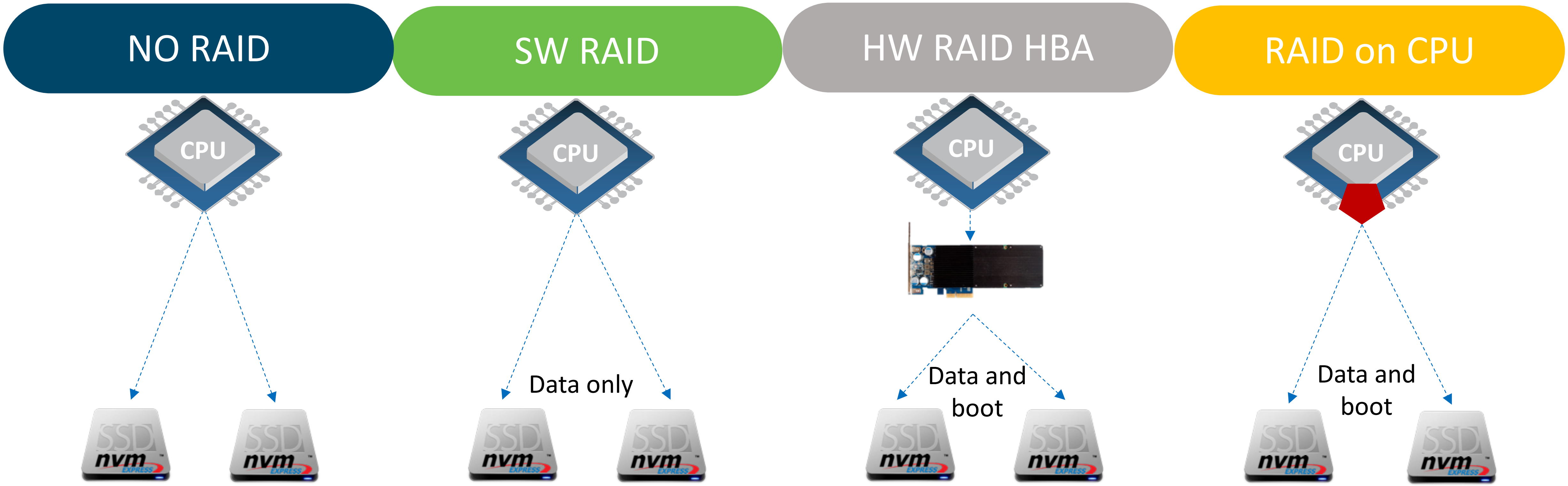
# RAID Implementations: Concepts

| RAID features | HW RAID | SW RAID | Hybrid RAID |
|---|---|---|---|
| SSD/bus errors isolation from OS | ✔ | ✖ | ✔ |
| RAID5 write hole closure | ✔ | ✖ | ✔ |
| Boot support | ✔ | ✖ | ✔ |
| Avoids use of CPU cycles for RAID | ✔ | ✖ | ✖ |
| Less HW required | ✖ | ✔ | ✔ |

SNIA EMEA

# Data Protection for NVMe PCIe Storage Devices




Application protection (SDS, erasure coding, replication factor)

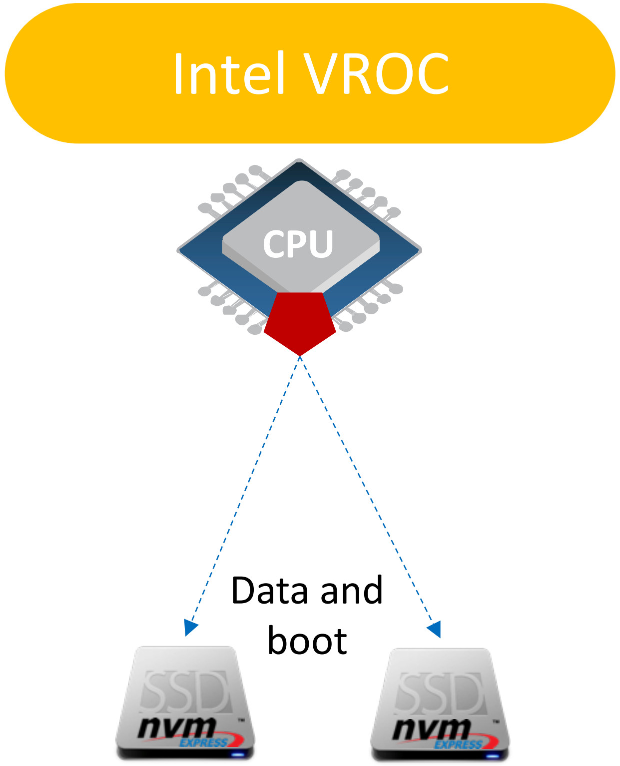
LVM/MD for Linux, inbox SW-RAID for Windows

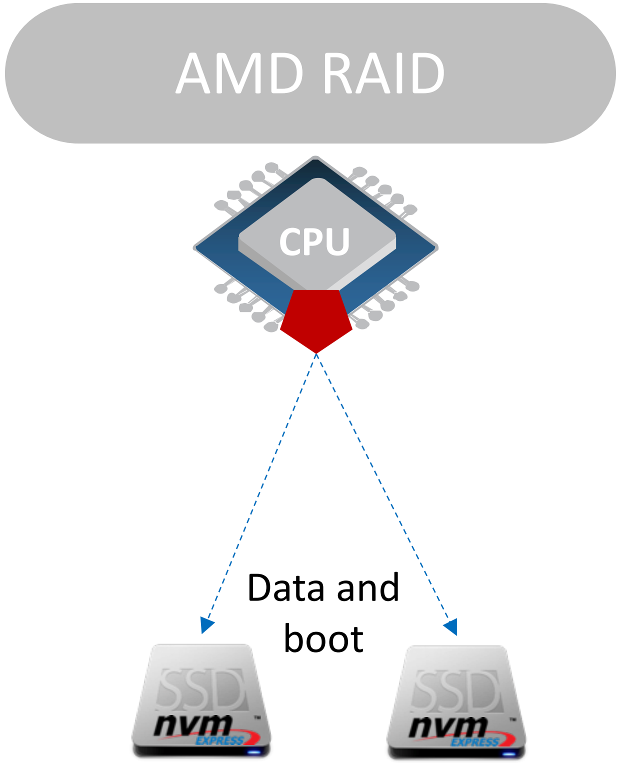
Tri-mode RAID controller (SATA/SAS/NVMe)

Vendor specific implementations

# RAID on CPU: A New Arrow in the Quiver




Could support SATA SSDs as well but focus is on NVMe SSDs
UEFI mode required

# Back to Basics: RAID Levels and Write Penalty

SNIA EMEA

# What is RAID?

- **Definition**: ”Redundant Array of ~~Inexpensive~~ Independent Disks”
  - Ability to read and write to multiple disks as a single entity, increasing performance and availability over a single, large, expensive disk
- **Performance**: increase the # of targets for write I/O, decreasing queuing and latency; does nothing for individual small reads, since data only written to a single disk, but scales performance for parallel I/Os
- **Availability**: Add in redundancy to provide superior error checking and tolerate hardware failure
- **Cost**: Do so with standard cheap disks


 
SNIA EMEA

# RAID Levels

- RAID is ***k*** data and ***p*** parity (***k,p***) disks
- Parity is an important concept:
  - determines tolerance to drive failures
- Striping over ***k*** disks makes serial read/write actions became parallel actions (similar concept to memory interleaving):
  - at the block-level for commercial implementations (not bytes or bits)
- Common RAID levels: 0, 1, 5, 0 + 1, 1 + 0, 5 + 1, 6
- Standardized by the Storage Networking Industry Association (SNIA) in the “Common RAID Disk Drive Format (DDF) “standard

https://www.snia.org/tech_activities/standards/curr_standards/ddf

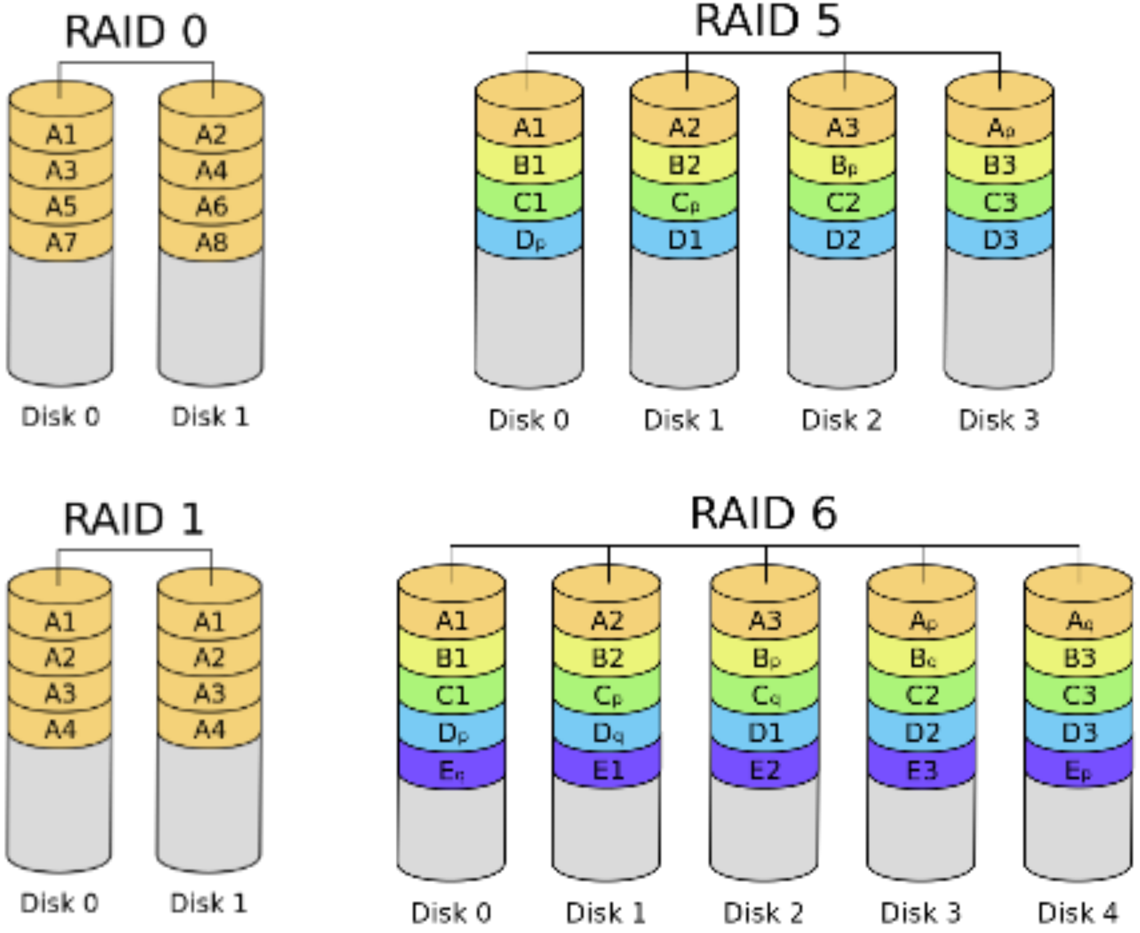


SNIA EMEA

Simplified View

# RAID Levels: Few Examples

- RAID-0 (*k*,0)
  - Block-level striping
  - No data protection
- RAID-5 (*k*,1)
  - Block-level striping with distributed parity
  - Parity is XOR across drives, tolerates 1 drive failure
- RAID-1 (1,1)
  - Mirroring, like RAID-5 with 1 data 1 parity; except parity is an exact copy
  - Tolerates 1 drive failure
- RAID-6 (*k*,2)
  - Block-level striping with distributed double parity
  - Two parities calculated, tolerates 2 drive failures
- Combos possible
  - E.g. RAID-10 is RAID-1 and RAID-0



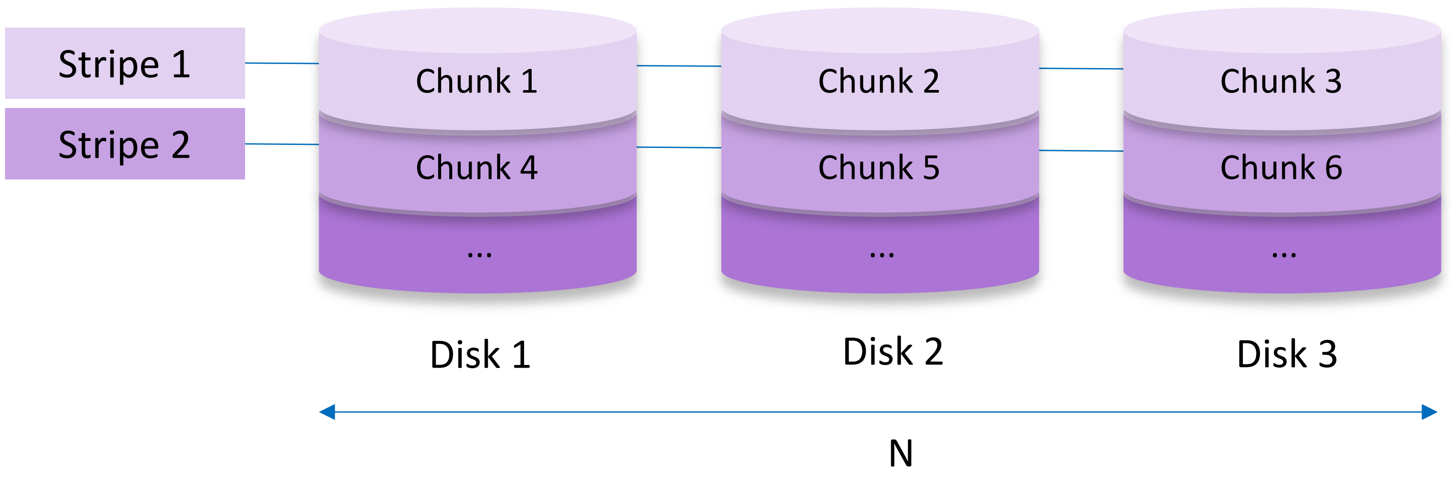

 

SNIA EMEA

# Terminology: Chunks and Stripes

CS = Chunk (Strip) size
N = number of disks

Stripe size = $\sum_{k=1}^{n} Chunk\ size_k$




- The choice of CS value is a compromise between storage efficiency and performance
  - Larger chunk size favors sequential access patterns but can waste storage space for data smaller than the chunk size and reduce performance for smaller random access.
  - Adjusting the chunk size to your workload is especially crucial for RAID 5/6 performance. When the value is chosen correctly, some of the requests can be handled as full stripe writes, which is significantly better than handling them as partial stripe writes.
  - Typically different chunk size settings does not have a significant impact on RAID 0, 1, 10 performance. Setting too small value can lead to multiple IO splits, so larger stipe sizes are typically more universal ones.

SNIA EMEA

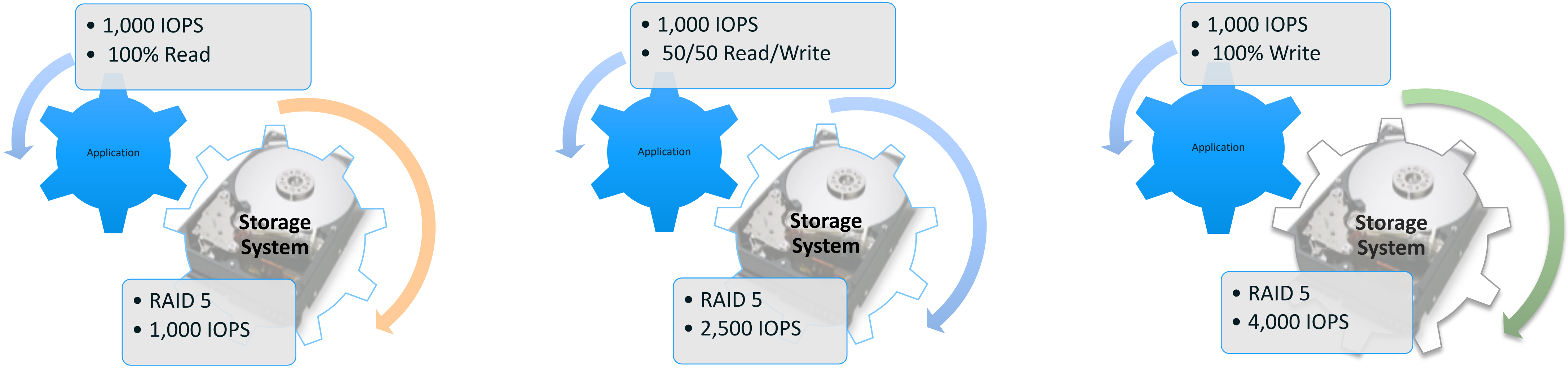
# What is The Right RAID Level?

- Failure tolerance
- Performance (write penalty)
- Disk number
- Disk capacity and rebuild time
- Storage efficiency
- Silent data corruption with double failure (write hole challenge)
- Workload needs

Lots of wrinkles: performance, capacity and data protection are all compromises. Choice essentially related to workload requirements

# Understanding The RAID Effect
# Write Penalty




- The backend storage system must produce enough IOPS to meet the application's IO requests and accommodate RAID protection requests
- RAID5 and RAID6 are impacted the most
- Solutions exist to alleviate the Write Penalty as seen by applications (i.e. caching)

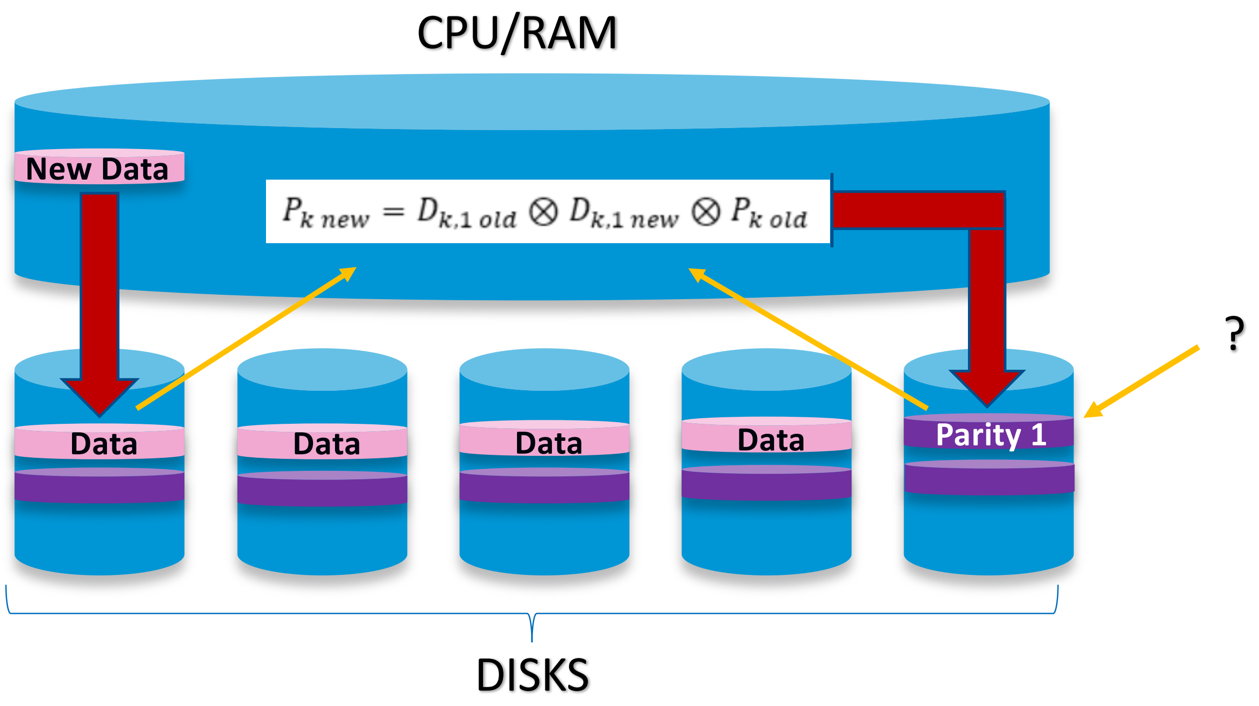
# RAID 5 Write Penalty Explained

- Read-modify-write method
- RAID 5 has a Write Penalty of 4
  - 4 IO operations for every Write IO
- Sequence of actions:

1. Read the old data strip
2. Read the old parity strip

----execute calculation, adds latency not IOPS

3. Write the new data
4. Write the new parity




- Solutions exist to alleviate the Write Penalty as seen by applications (i.e. caching)

# RAID Levels, Write Penalty & IOPS

| RAID Level | READ Penalty | WRITE Penalty (*) | Capacity Impact |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 & 10 | 1 | 2 | #Disks/2 |
| 5 | 1 | 4 | #Disks-1 Disk |
| 6 | 1 | 6 | #Disks-2 Disks |

| Disk Type | IOPS |
|---|---|
| 7,200 RPM | 75-100 |
| 10,000 RPM | 125-150 |
| 15,000 RPM | 175-210 |
| NVMe Flash | 220,000 |
| NVMe Optane | 500,000 |

Assumption for each disk: read IOPS = write IOPS
Assumption: single sector write (not full stripe)

**(*) Solutions exist to alleviate the Write Penalty as seen by applications (i.e. caching)**

BE IOPS Required = [(FE IOPS x %READ)+(FE IOPS x %Write) x RAID Write Penalty]
Example: Application requires 100000 IOPS with 50% Read and 50% Write and you're using RAID5 & 15K rpm drives with 200 IOPS each. How many disks are required?

✗ 100000/200=500 Disks required

✓ [(100000x50%)+(100000x50%)x4]=250000 BE IOPS Required
250000/200=1250 Disks Required

# RAID5 Write Hole (WH) Challenge

- Write operation not completed due to drive failure and power loss (double fault) happening at the same time during the write operation
- Leads to silent data corruption
- Cacheless RAID 5 is affected the most
- HW RAID solves the issue with dedicated resources (persistent cache, battery/supercap protected local RAM)
- SW RAID needs alternative and more complex approaches
  - Distributed Partial Parity Log: distributes recovery info among members of RAID group
  - Journaling: requires an additional disk to journal recovery information
  - Local disk cache power loss protection (or local disk cache disabled) required

# Intel VROC: an Overview

SNIA EMEA

# A New Approach: Intel VMD and Intel VROC

**Intel Volume Management Device**

- Enterprise-grade serviceability features for NVMe SSD units:
  - Surprise hot insertion and removal
  - LED management
  - Error isolation

Considerations:

- Intel Xeon Scalable CPUs
- Supports non Intel NVMe SSDs
- Compatible BIOS and drivers required
- Free

**Intel Virtual RAID On CPU**

- Enterprise-grade data availability for NVMe SSD units:
  - Bootable RAID, UEFI mode only
  - RAID 0, 1, 5, 10 levels and R5WH closure
  - Linux/Windows support

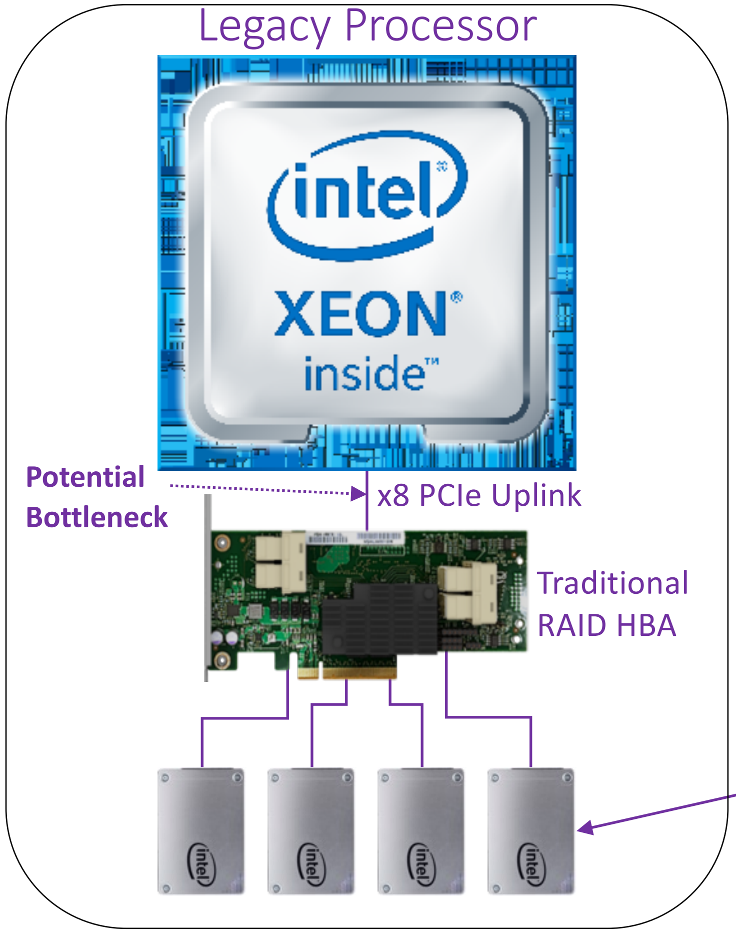
Considerations:

- Intel Xeon Scalable CPUs
- Intel VMD is a prerequisite
- Support for non Intel NVMe SSDs (*)
- Licensed feature (* )

(*) depends on server vendor

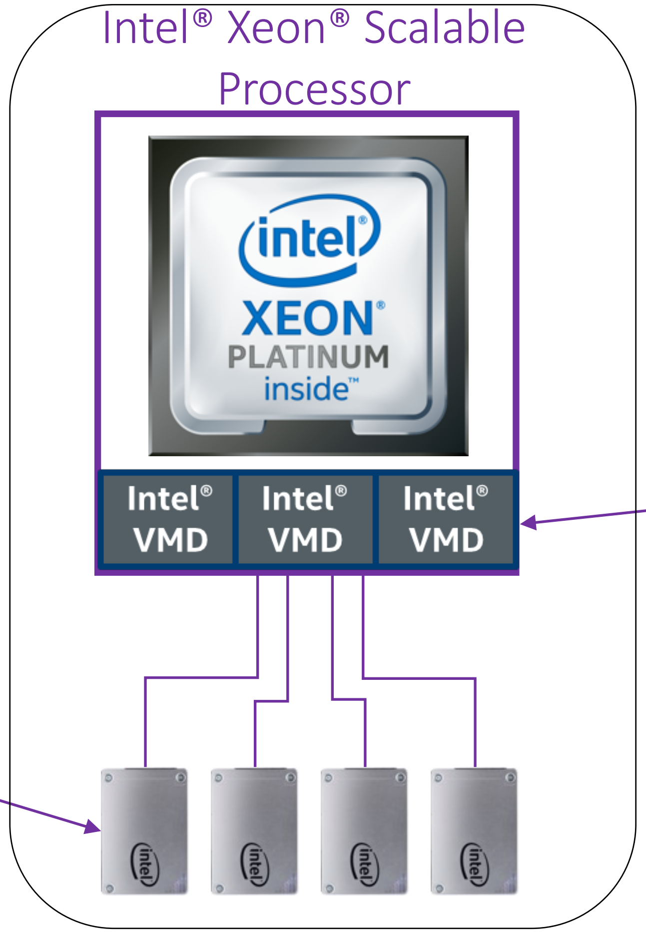
# Intel® Virtual RAID on CPU (Intel® VROC)




Simpler Design
Excellent Performance

Less BOM
Less Power

NVMe SSDs
x4 PCIe Uplink per SSD



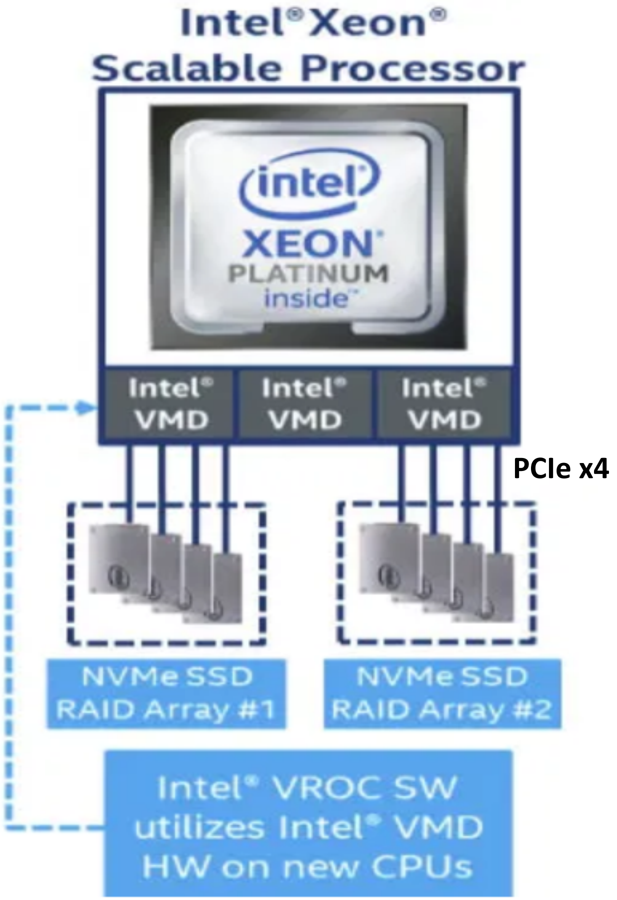

- Intel® VROC uses Intel® VMD as the RAID controller to RAID NVMe SSDs to the CPU directly
- x16 PCIe lanes per VMD
- Ability to create data RAID levels which spans VMD domains (virtual HBAs)

**Intel® VROC provides compelling RAID solution for NVMe SSDs**

# Intel VROC Feature: At A Glance

- Enterprise-grade RAID solution for NVMe SSD’s
- Leverages Intel VMD for hot swap and LED management
- Intel VROC is a hybrid RAID solution
- Intel VROC supports data volumes and boot volumes
- RAID options are 0,1, 10, 5 with Write Hole closure
- High performance, no HBA card
- Supported on Linux and Windows (ESXi only supports VMD)




Each drive connected to Intel VMD by PCIe x4

SNIA EMEA

# Intel VROC: Supported RAID Levels

RAID settings are configurable via BIOS (pre OS) or CLI or GUI or RESTful agent (post OS)

- RAID0: 2+ drives (striping)
- RAID1: 2 drives (mirroring)
- RAID5: 3+ drives (striping with parity), R5WH Closure options
- RAID10: 4 drives (nested RAID)

Data RAID arrays can be built within a single VMD domain, across domains, and even across CPU’s: performance is not the same though

Bootable RAID arrays must be within a single VMD domain

Chunk size: 4K - 128K (default chunk size depends on RAID level and number of member drives)

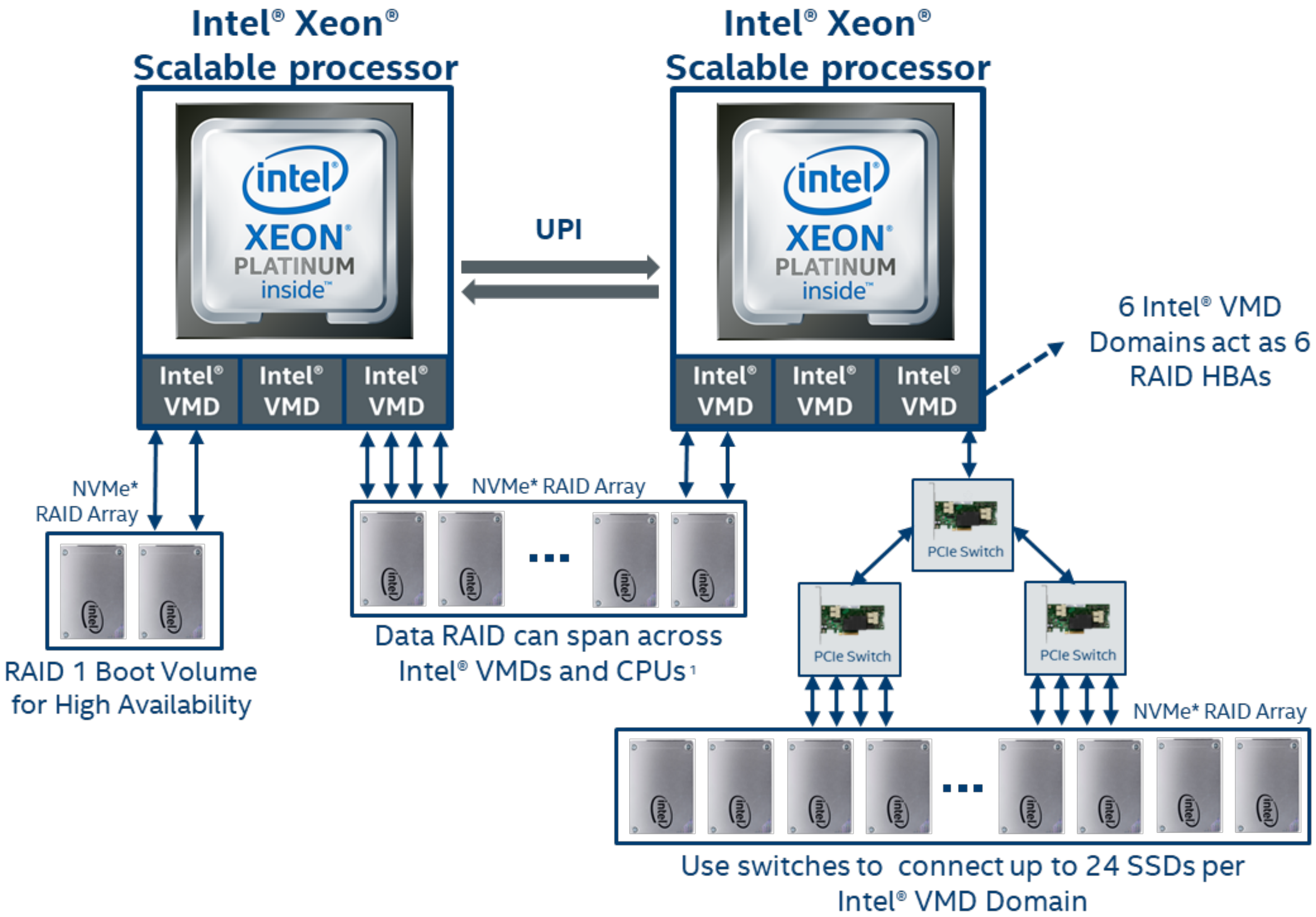
Spare drives, auto-rebuild, RAID volume roaming, volume expansion, volume type migration

Matrix RAID: Multiple RAID levels configurable on common disks, if space available

SNIA EMEA

# Intel VROC: Data and Boot RAID Arrays




- Data RAID array spannable across VMD domains and even across CPUs
- Boot RAID array must be within a single VMD domain
- A server with dual Xeon Scalable CPUs could theoretically support up to 24 NVMe direct attached (full speed) drives
- PCIe switches on the motherboard can be used to expand the number of NVMe SSDs in the server (up to 48)

# Intel VROC: Double Fault Protection

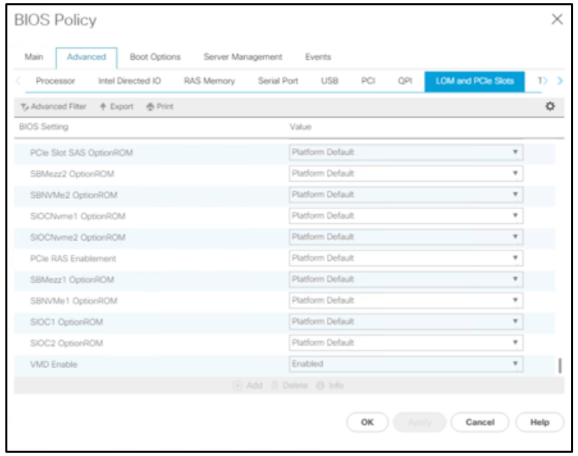
- RAID Write Hole challenge: write operation not completed due to drive failure and power loss happening at the same time, silent data corruption
- HW RAID solves the issue with dedicated resources
- SW RAID needs alternative approaches to achieve reliable RAID 5 data protection
- Hybrid RAID of Intel VROC can provide R5WHC with a combination of techniques:
  - OS dependent
  - RAID5 Write Hole Closure is disabled by default: it has a performance impact

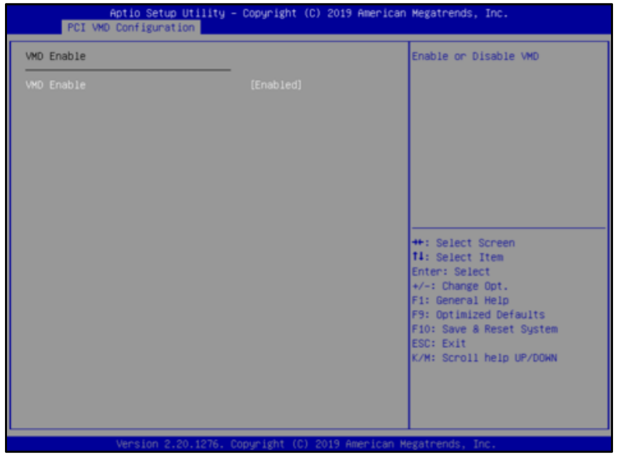
 

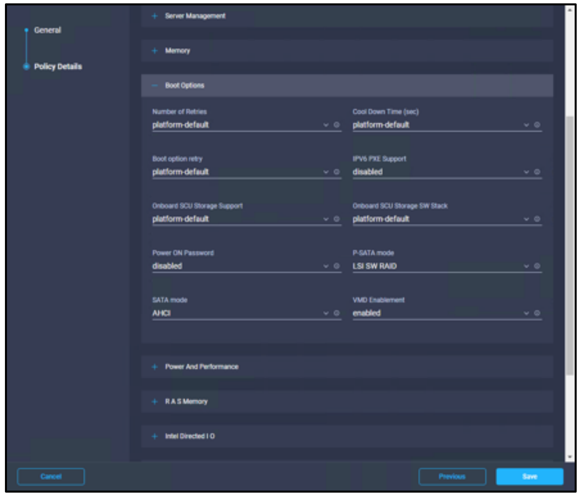
SNIA EMEA

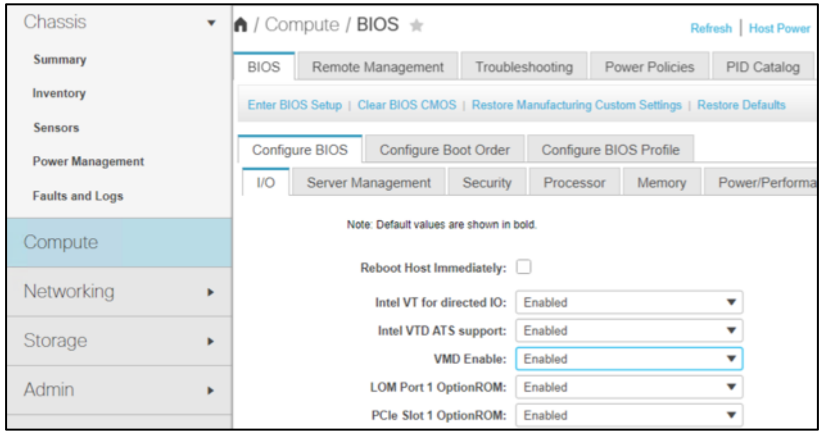
# VROC Configuration and Management Examples

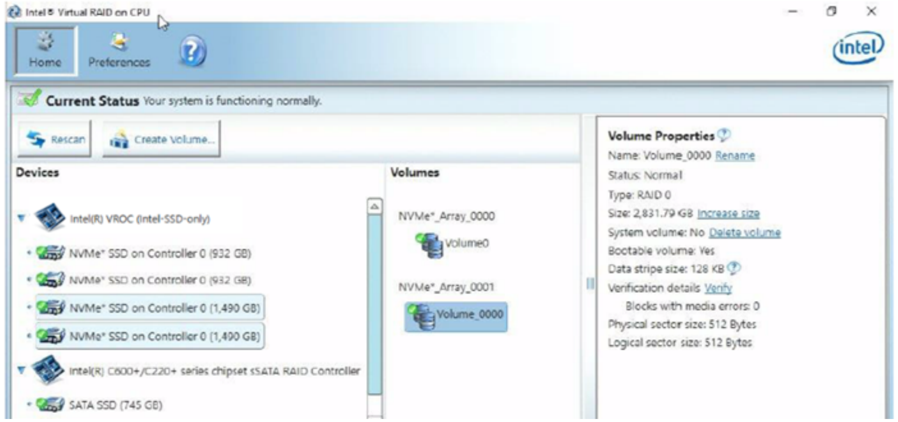
- Configuring VROC RAID:
  - Intel VROC UEFI HII
  - Intel VROC GUI (Windows)
  - Intel VROC CLI Tool
  - Integrated support for some vendor management tools



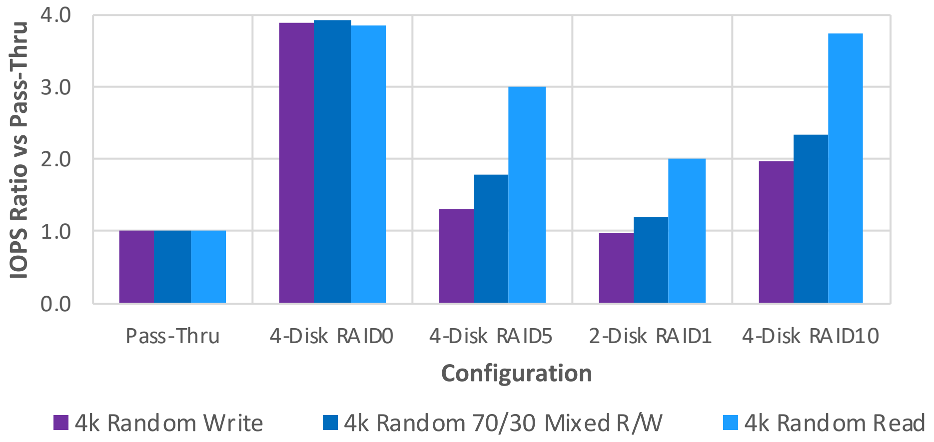



SNIA EMEA

# Performance – RAID vs Pass-thru

## Windows/Linux with Intel® SSD DC P4510, 4k Random I/O profile

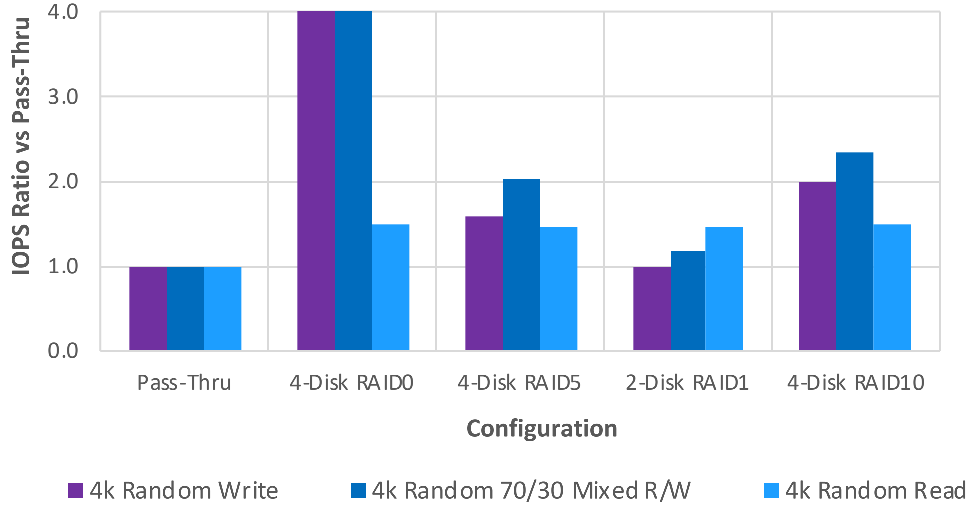



- Pass-thru raw data:
  - 4k Rand Write: 84k IOPS
  - 4k Rand Mixed: 183k IOPS
  - 4k Rand Read: 645k IOPS
- Physical CPU Cores Used:
  - 4-Disk RAID0 Read: 4.7 Cores
  - 4-Disk RAID5 Write: 1.2 Cores
- 4-Disk RAID0 Read: 2.5M IOPS

RHEL 7.4

Windows 2016

- Pass-thru raw data:
  - 4k Rand Write: 80k IOPS
  - 4k Rand Mixed: 179k IOPS
  - 4k Rand Read: 634k IOPS
- Physical CPU Cores Used:
  - 4-Disk RAID0 Read: 17 Cores
  - 4-Disk RAID5 Write: 6.3 Cores
- 4-Disk RAID0 Read: 952k IOPS
- Up to 1.4M IOPS with multiple RAID volumes

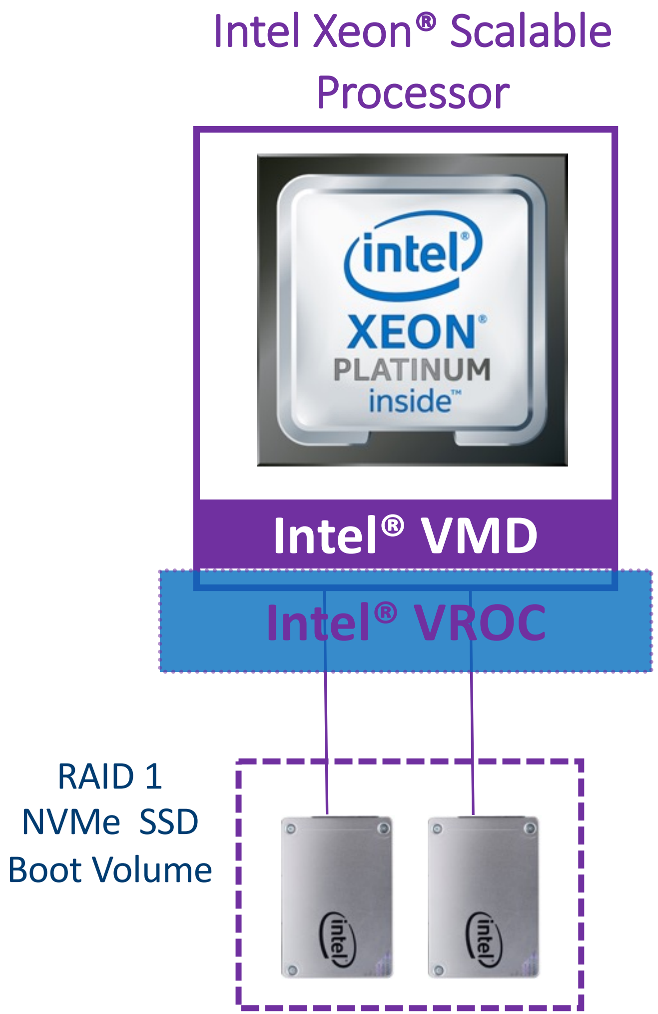


# Practical Use Cases

# High Availablity Boot




Boot Requirement Include:

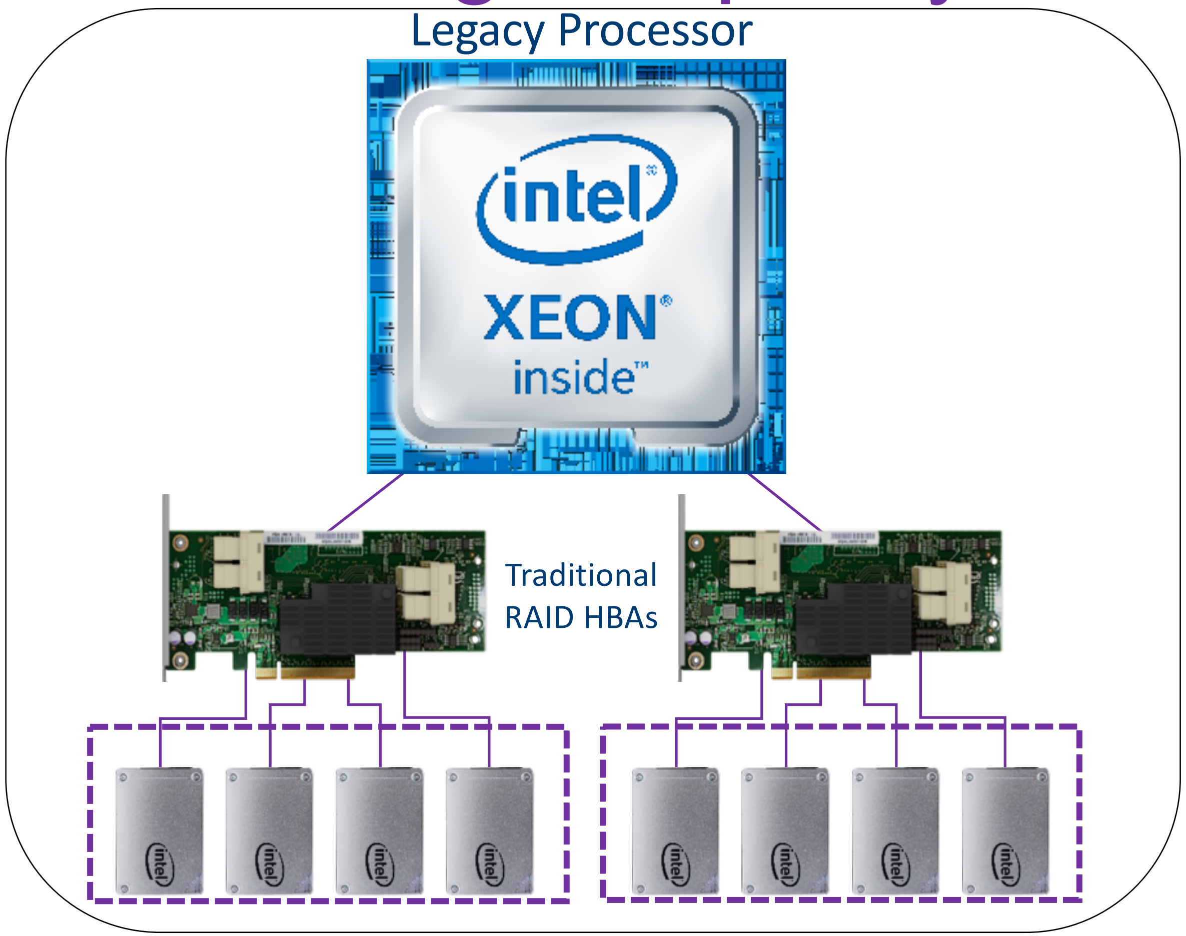
1) 2x Intel® Boot SSDs
2) Intel® SSD Only VROC HW Key

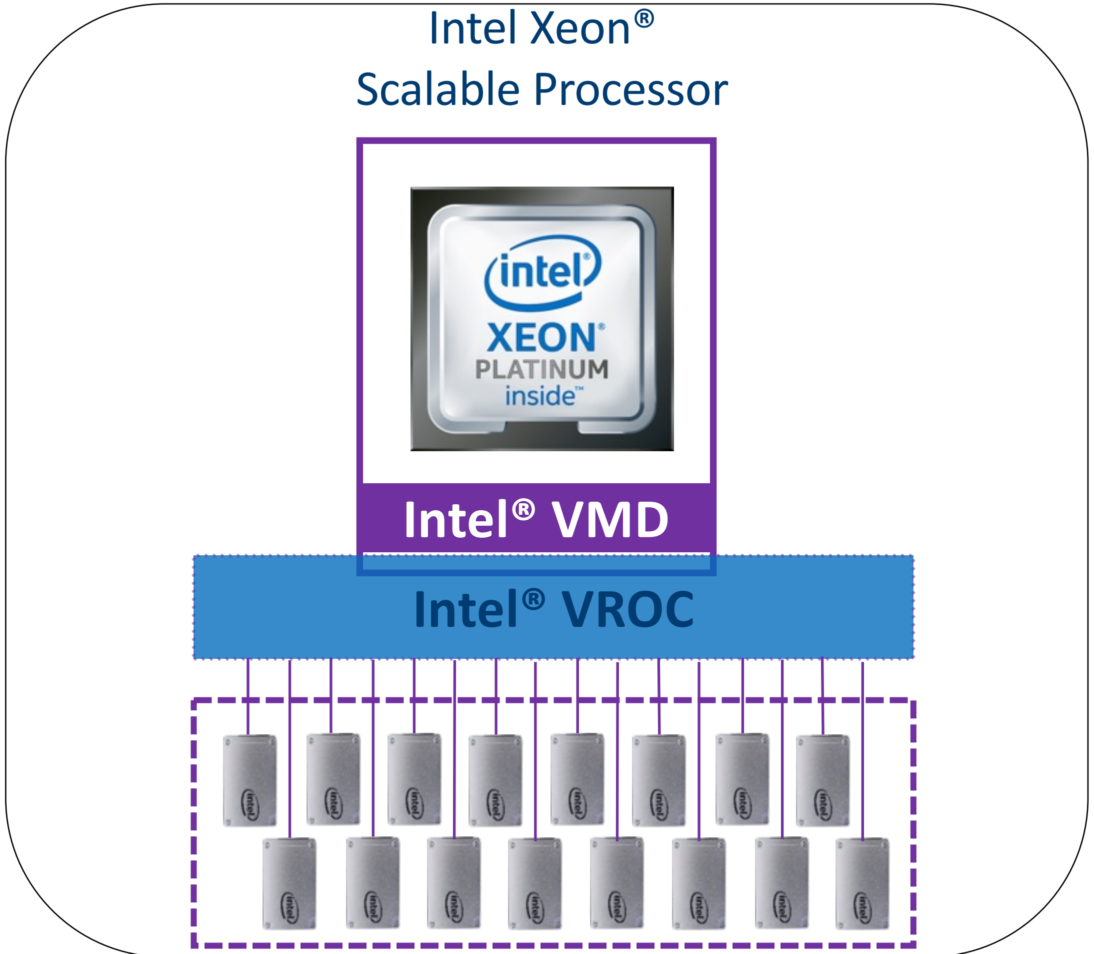
- **High Performance** boot for quick power on
- SATA RAID card is **no longer needed**

SNIA EMEA

# Scalable High Capacity Data RAID




In case of traditional HBAs you need to choose:

- Limited performance data RAID with higher capacity and lower cost
- High performance data RAID with limited capacity and higher cost

In case of VROC you can have both: high capacity (up to 384 TB) and high performance. You can also scale up your RAID any time without need to purchase additional VROC license.

SNIA EMEA

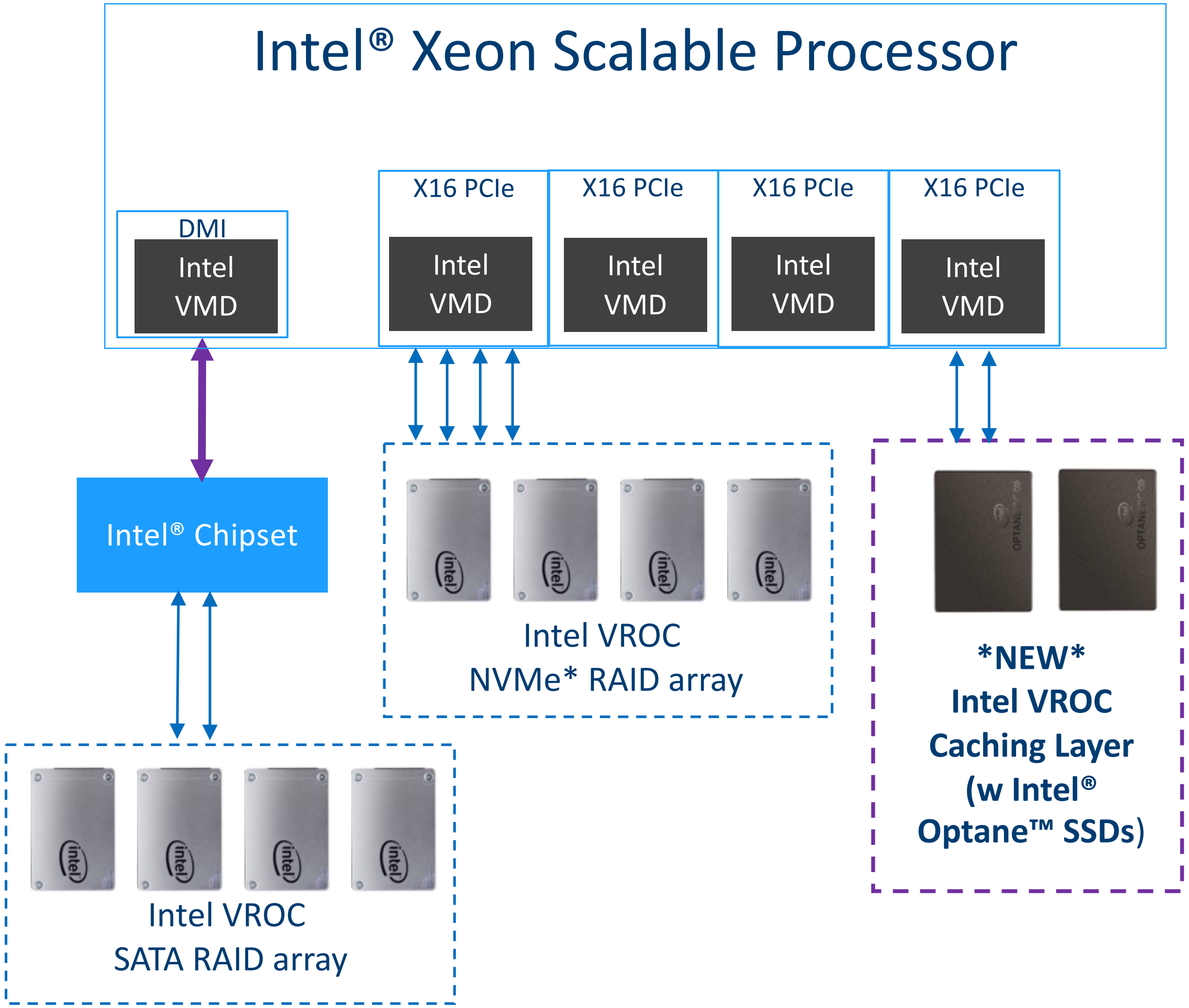
# What’s Next

SNIA EMEA

# Intel® VROC Integrated Caching

- A new Intel VROC capability to add an Intel Optane SSD cache layer in front of storage volumes
  - An improved WB Cache
    - Replace DRAM Cache used by RAID HBAs today
  - Open Source
    - Linux Only (to start) and powered by Open CAS
  - Enterprise Supported and Validated
    - Just like VROC RAID model
  - Platform Integrated
    - Designed into OEM platforms with VROC
  - Flexible Usage Models:
    - Caching for SATA or NVMe SSDs
    - Sophisticated Caching policies
  - Eliminate Single Point of Failure:
    - Use Intel VROC RAID1 for a redundant cache



# Intel® VROC IC Acceleration Details

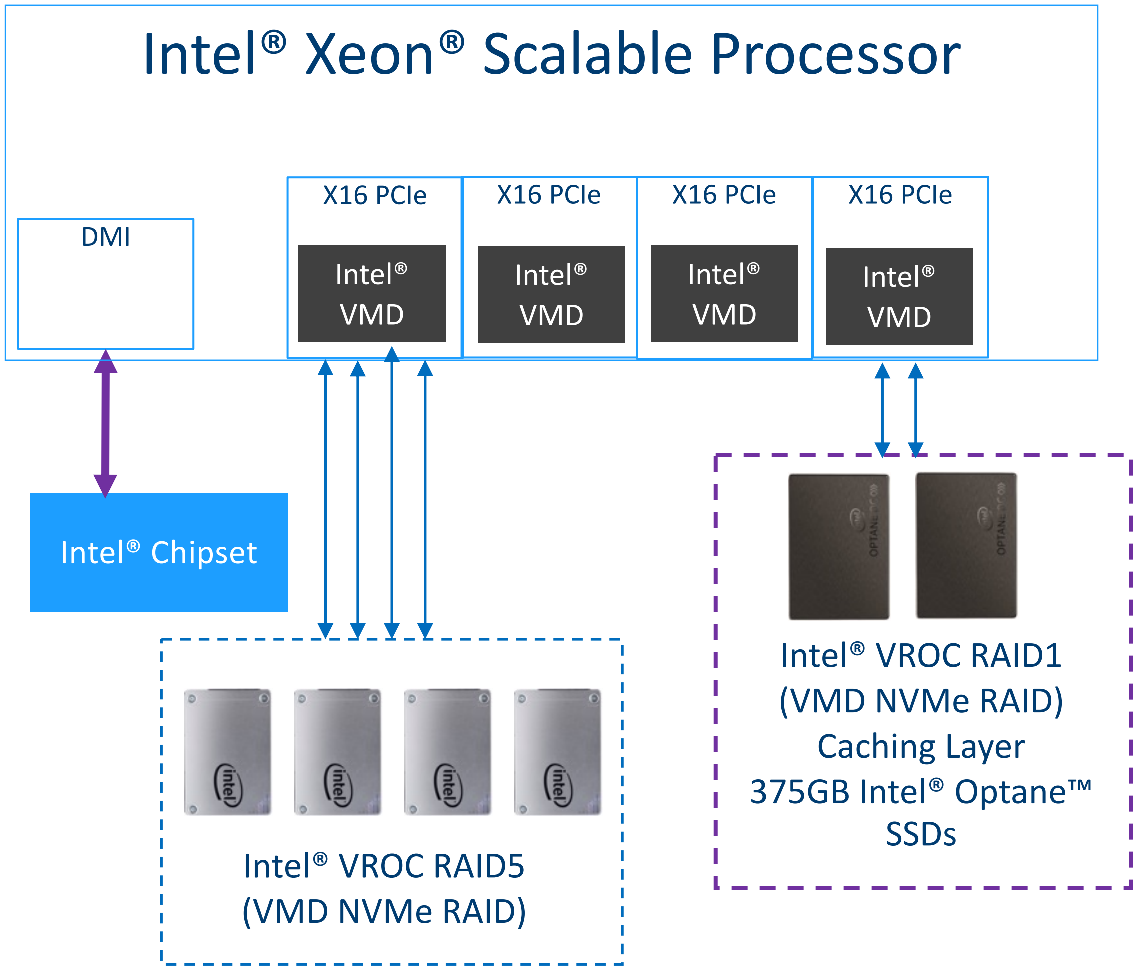
- Intel® VROC IC provides an attach point to leverage Intel® Optane™ SSDs to improve 3 critical server performance and cost metrics:
  - Total Storage Bandwidth
  - Application Latency
  - Aggregate Storage Subsystem Endurance
- To achieve desired results, recommended caching policies are designed to redirect write IO that are at least one of the following:
  - Invalidated often (short lifetime)
  - Overwritten frequently
  - Accessed Often (“Hot Data”)
- Intel® Optane™ SSDs are effective to absorb the thrash these write IO can cause on a storage subsystem

# Intel® VROC IC MySQL Proof Point (Optane + TLC NVMe)

- Use-case: MySQL Database (MySQL 8.0.2.1)
- Benchmark: Sysbench OLTP_Read_Write
  - 1 hr. test, 128 threads, 120GB Database
- Cache policy: Everything but DB blocks (16k)

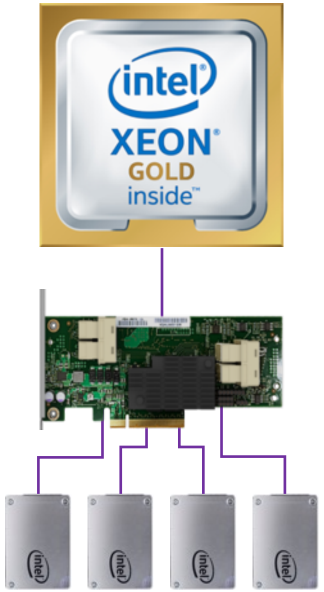
| | TLC NVMe RAID Only | Intel® VROC IC w/ Intel® Optane™ SSDs | % |
|---|---|---|---|
| Performance (tps)<br>Higher is Better | 6,750 | 10,201 | ↑ 51% |
| Avg Latency (ms)<br>Lower is Better | 18.96 | 12.55 | ↓ 34% |
| P99 Latency (ms)<br>Lower is Better | 36.24 | 20.00 | ↓ 45% |
| Endurance<br>(storage lifetime transactions)<br>Higher is Better | 23.88B | 64.26B | ↑ 169% |



# Intel® VROC IC MongoDB Proof Point (Optane + SATA)

Legacy RAID HBA Solution




4x S4610 RAID5 behind RAID HBA

- Benchmark: YCSB Workload-A
  - 32 threads, 200M operations, 2TB database
- Cache policy: Write only mode

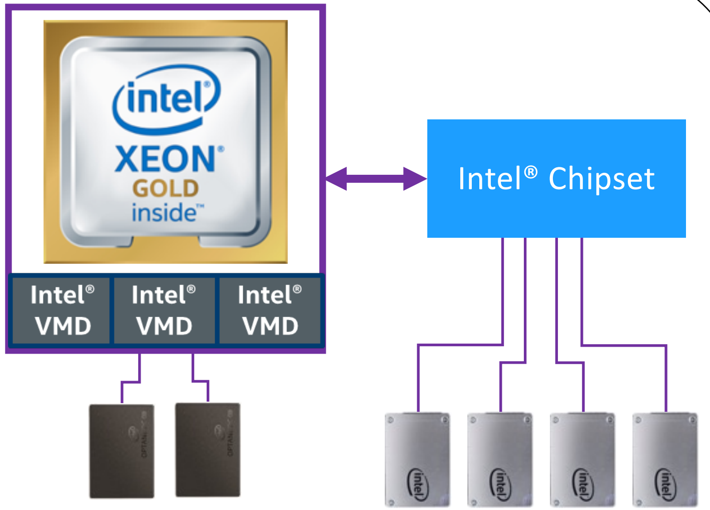
| | |
|---|---|
| Ops/s | 9,892 ops |
| Avg. Update Latency | 5,701 us |
| Storage Lifetime Ops. | 2,389B |

| | |
|---|---|
| Ops/s | 11,912 ops |
| Avg. Update Latency | 4,425 us |
| Storage Lifetime Ops. | 3,443B |

**Intel VROC IC with Intel Optane SSDs delivers:**

- **20% ↑ Performance**
- **29% ↓ Avg. Latency**
- **44% ↑ Storage Lifetime Operations**

Intel Optane SSD Solution w/ Intel VROC Integrated Caching




Intel VROC IC Intel Optane SSD RAID1 Caching Layer

4x S4510 RAID5 with VROC (SATA RAID)

SNIA EMEA

Thank you for attending