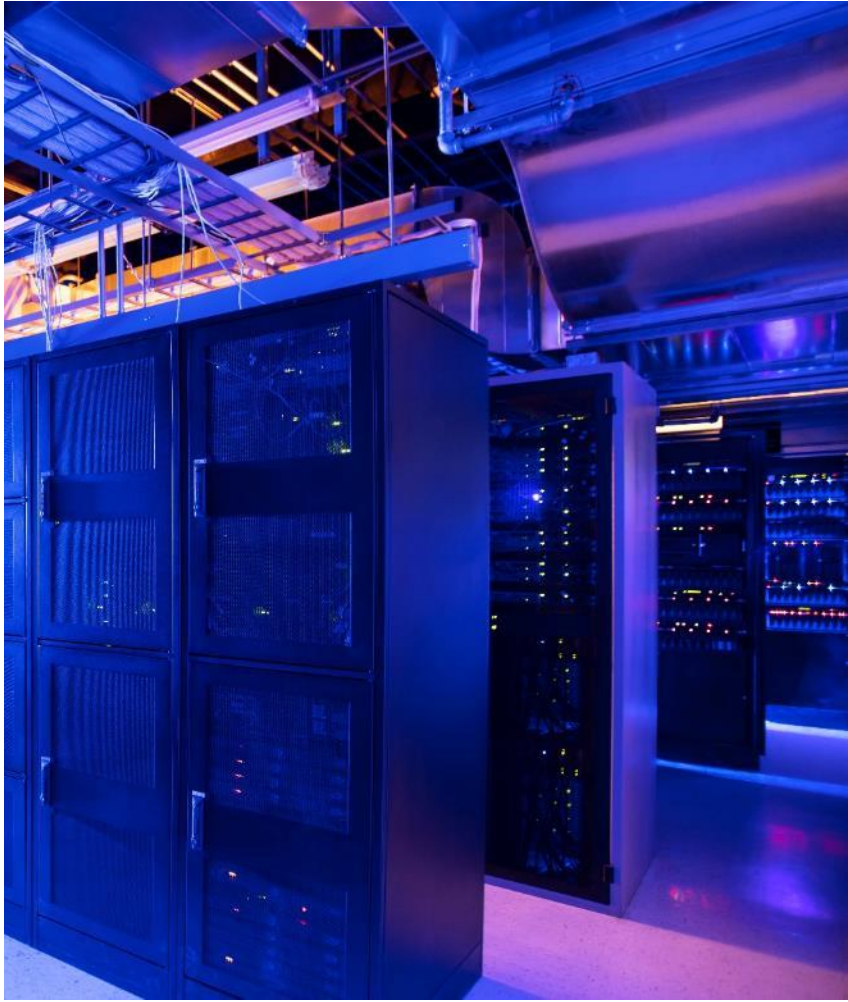
HIGH PERFORMANCE POOLED STORAGE FOR RSD ARCHITECTURES

Steve Miller, Siddhartha Panda, & Sujoy Sen

AGENDA

- Overview of RSD
- Composable Block Storage CBS
  - Feature Requirements
  - Why SPDK
  - Performance



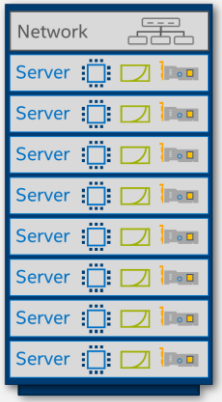
Data Center Agility, Built on Open Standards

TODAY'S DATA CENTER CHALLENGES

45% UTILIZATION OF EQUIPMENT[1]

50% EFFICIENCY IT OPERATIONS[1]

35 PEOPLE HOURS PER RACK UPDATE[2]



Current Infrastructure

- **Fixed ratio** of compute, storage, and accelerator resources
- **Expensive** refresh & scale out
- **Outdated** software interface
- **Cumbersome** hardware provisioning process

Intel® Rack Scale Design

*"an **industry-aligned architecture** for composable, disaggregated infrastructure **built on modern, open standards.**"*

- Disaggregated
- Composable
- Interoperable

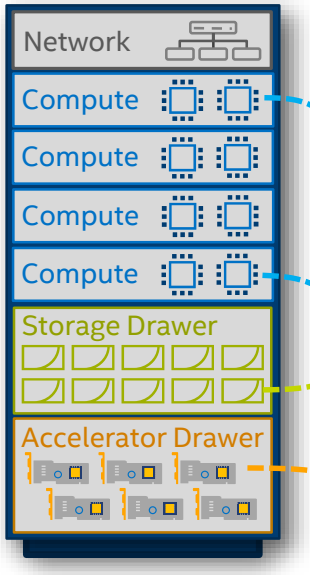
INCREASE AGILITY

DECREASE COSTS

1. Source: Quantifying Datacenter Inefficiency: Making the Case for Composable Infrastructure, IDC, Document #US42318917, 2017.
2. Source: Disaggregated Server Architecture Drives Data Center Efficiency and Innovation, Shesha Krishnapura, Intel Fellow and Intel IT CTO, 2017

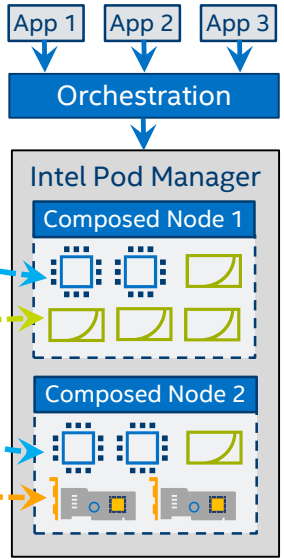
Intel® RSD Key Attributes

DISAGGREGATED



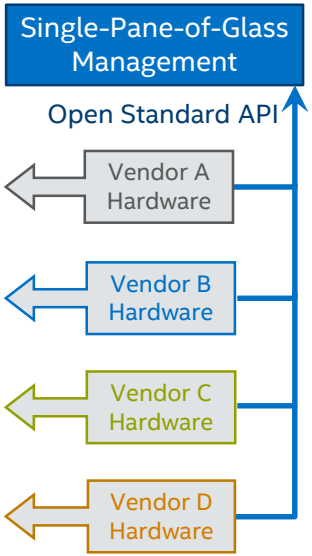
Buy less up front and
Save money over time

COMPOSABLE



Compose hardware
resources “on the fly”

INTEROPERABLE

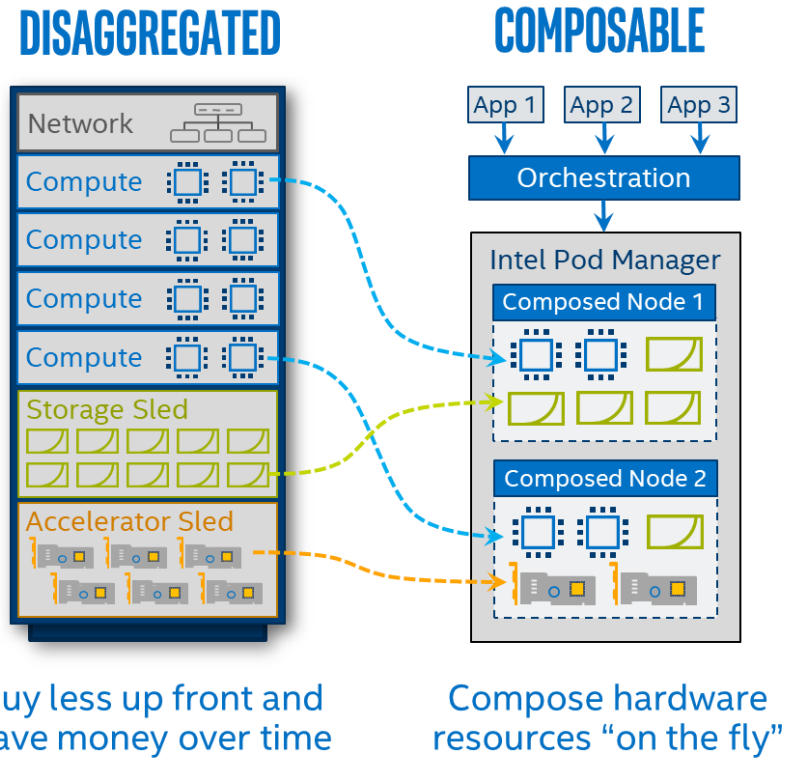


Choose the best now
without vendor lock-in

Intel® Rack Scale Design

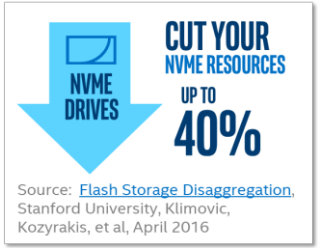
OEMs with solutions
based on Intel RSD

Benefits of Disaggregation and Composability



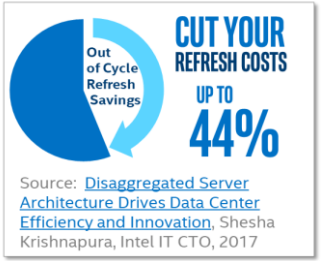
Resource Pooling

Maximize utilization of high-value assets and improve agility with dynamic composability



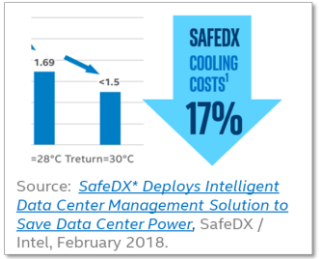
Modular Refresh

Independently scale and upgrade resources with better lifecycle management



Operational Costs

Lower Power Usage Effectiveness (PUE) and streamline operations and HW management

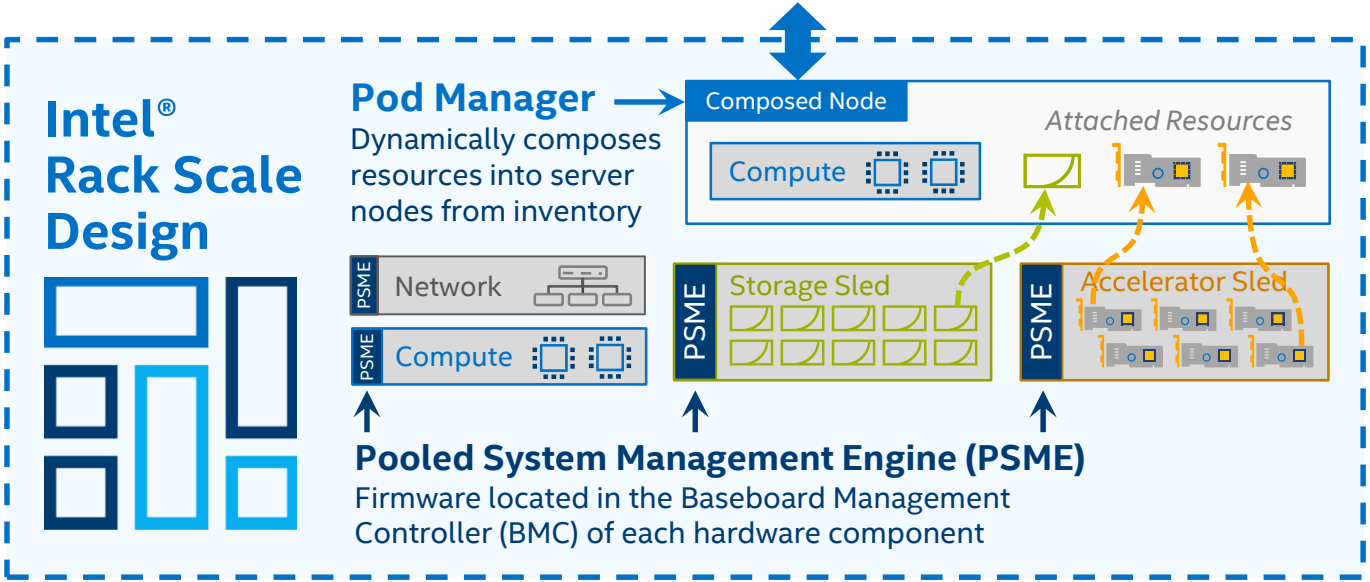


Intel® RSD – Composability

Compose hardware resources “on the fly” for specific workloads

App 1 | App 2 | App 3

Orchestration Software



Intel® RSD software functions include:

Resource Discovery

Automatically discover and store hardware characteristics and location for all your resources

Node Composition

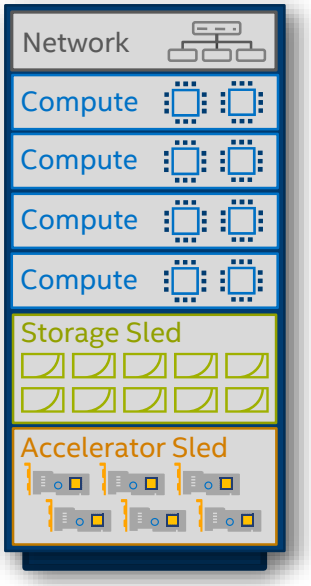
Dynamically compose compute, storage, and other resources to meet workload specific demands

Telemetry Data

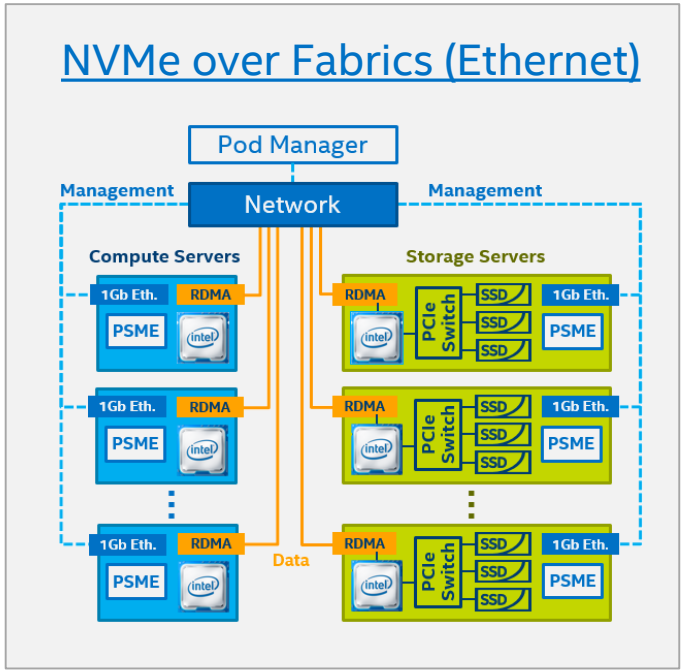
Monitor data center efficiency and detect, diagnose, and help predict resource failures.

Intel® RSD – Storage Disaggregation

Disaggregation



Save $$ over time with modular refresh



Feature Set for Storage Target

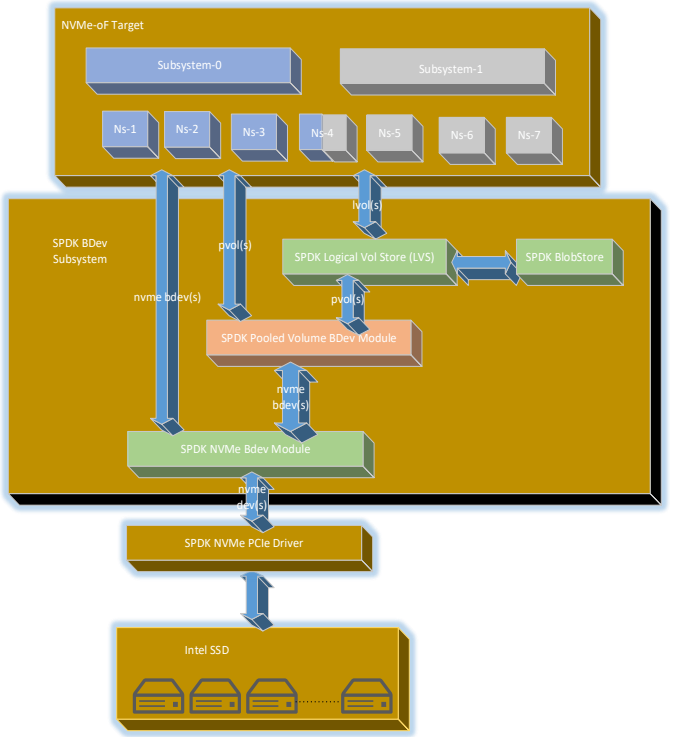
- High Performance
- Low Latency
- RAID 0 (Stripes across SSD)
- Clones
- Thin Provisioning
- Snap Shots
- Quality of Service per logical volume
- Efficient use of CPU and memory resources
- Standards Based Management
- Simplicity of use

CBS – Key Features

- Disaggregated high speed NVME-o-F target over RDMA (powered by Intel Networking)
- Pooling of Intel SSDs for capacity and performance
- Volume Management functionalities: Thin Provisioning, Snapshots, Clones
- Data at rest security and multi-tenant volume level security
- Smart volume level Quality of service. (Rate limiting, guaranteed BW)
- Centralized Storage manager at PoD level. Policy based provisioning and automatic deployment
- Online Storage Pool capacity Expansion (both stripe width expansion & drive concatenation)
- Proactive detection of drive failure and seamless drive replacement
- Scale and Performance (Network saturation up to 400Gbs, 8K volume support)
- Support Industry standard management Interfaces by DMTF/SNIA (Redfish/Swordfish)
- Integration across flavors of cloud OS's through various plugins.
- Event and Telemetry Support to trigger AI for PoD level management

Why SPDK ?

- CBS leverages the Modularized bdev abstraction framework of SPDK
  - Most of CBS functionalities are developed as new SPDK bdev modules or uses the SPDK framework extensively.
- SPDK being an user space application, it is easy to develop, debug and maintain the data path application like CBS
  - No system call overhead in IOPATH
- SPDK designed to be a Async Poll mode driver with no interrupts
  - currently most of the IO path drivers are limited in interrupt mode
  - Poll mode enables IOPs scalability and avoid complex MSIX interrupt overhead across cores.



Why SPDK cont..

- Lockless architecture
  - SPDK framework enforces the developer minimize or eliminate locks. This reduces the complexity and increases efficiency
  - Also if lock is required, SPDK provides a good infrastructure for inter core messaging which minimizes processor cache thrashing and avoids traditional expensive locks like mutex/semaphore etc.
- Core affinity
  - Significantly lower overhead in synchronization with per core data structure
  - End to end IO request processing in same core enables faster and more efficient IO processing
  - HPSP uses the SPDK for efficient Core distribution based on functionality and workload. Enabling feature SLAs, for example, HPSP uses separate cores for telemetry collection and analysis, dedicated cores for expansion, QOS management, and non-IO path config commands.
- Error handling
  - Using SPDK's error propagation feature across different bdev hierarchy, handling error in each module is very systematic and can be kept module specific.

CBS with SPDK (Scale & Performance)

- CBS provides disaggregated storage for big datacenters, where IOPs and scale are critical.
    - SPDK is designed for 8-10x IOPs/core vs kernel NVMe and fabric drivers.
    - SPDK scales easily to 8-10M IOPs (both RW) with just a few cores.
- Intel Optane NVMe SSD require a lightweight, low latency stack, SPDK provides that infrastructure.
    - Optimized tail latencies with per core IO processing.
- With SPDK volume management layer, number of volumes can scale well to 8K, and volume size can go up to TBs.

Performance Test Configuration

Hardware

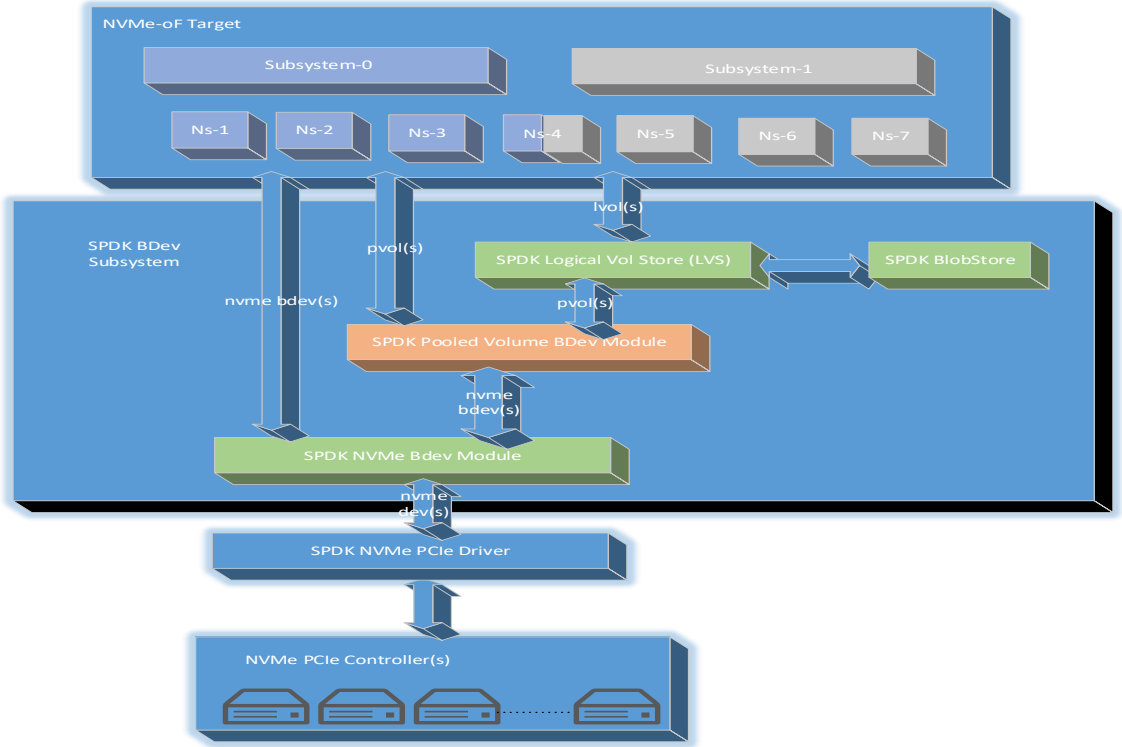
Wolfpass storage sled

- Skylake-Dual socket 36 cores
- Mellanox ConnectX-5 100Gbps NIC
- 8x Intel P4500 SSDs

Software

- Fedora Core 26
- SPDK base version 18.07 with new pooled volume module
- OFED 4.2 with RDMA support

SPDK stack with new Pooled Volume BDEV Module

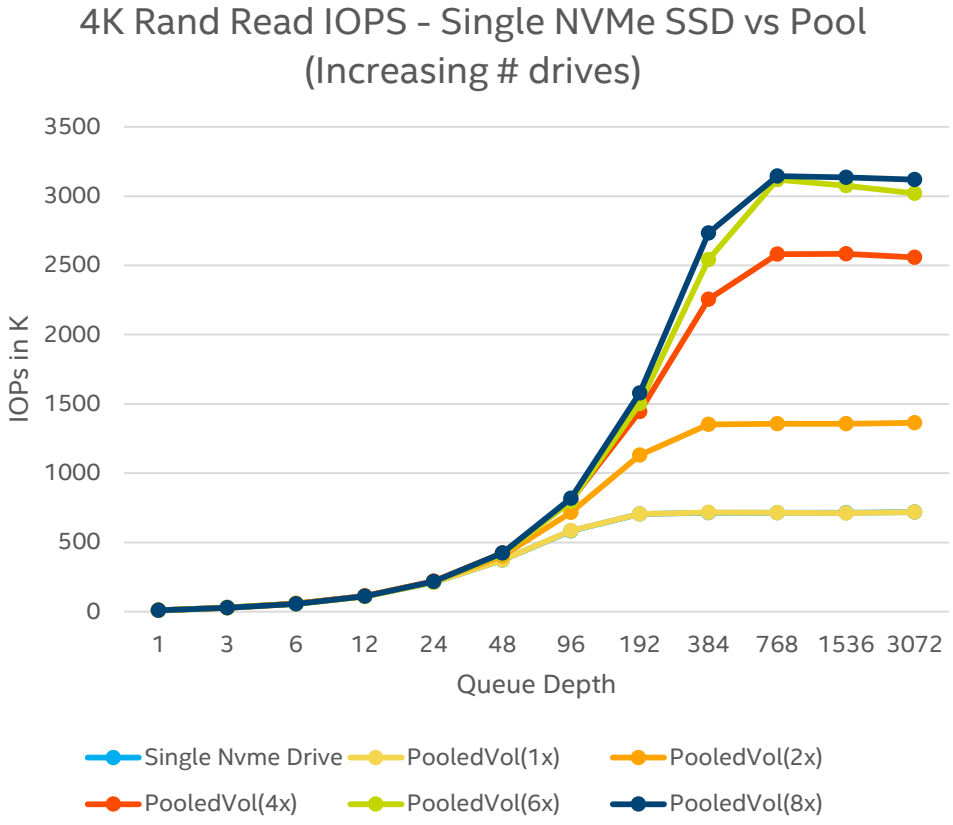


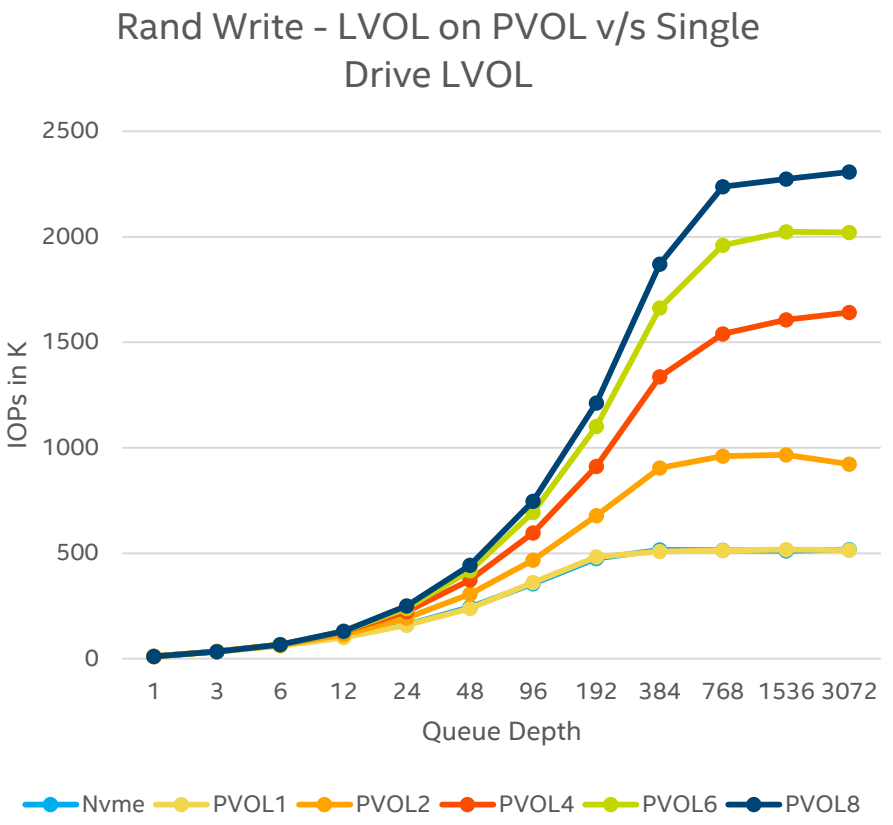
Pool Configuration - Performance

4 Corner Performance 100Gb/s RDMA network

Workload	Measured Performance
4K RND 100% RD	3.14 MIOPS
4K RND 100% WR	2.3 MIOPS
4K RND 70/30 R/W	2.23 MIOPS
128k SEQ 100% RD	12.5 GBps
128k SEQ 100% WR	12.5 GBps
1 QD RND RD	101 us
1 QD RND WR	31 us

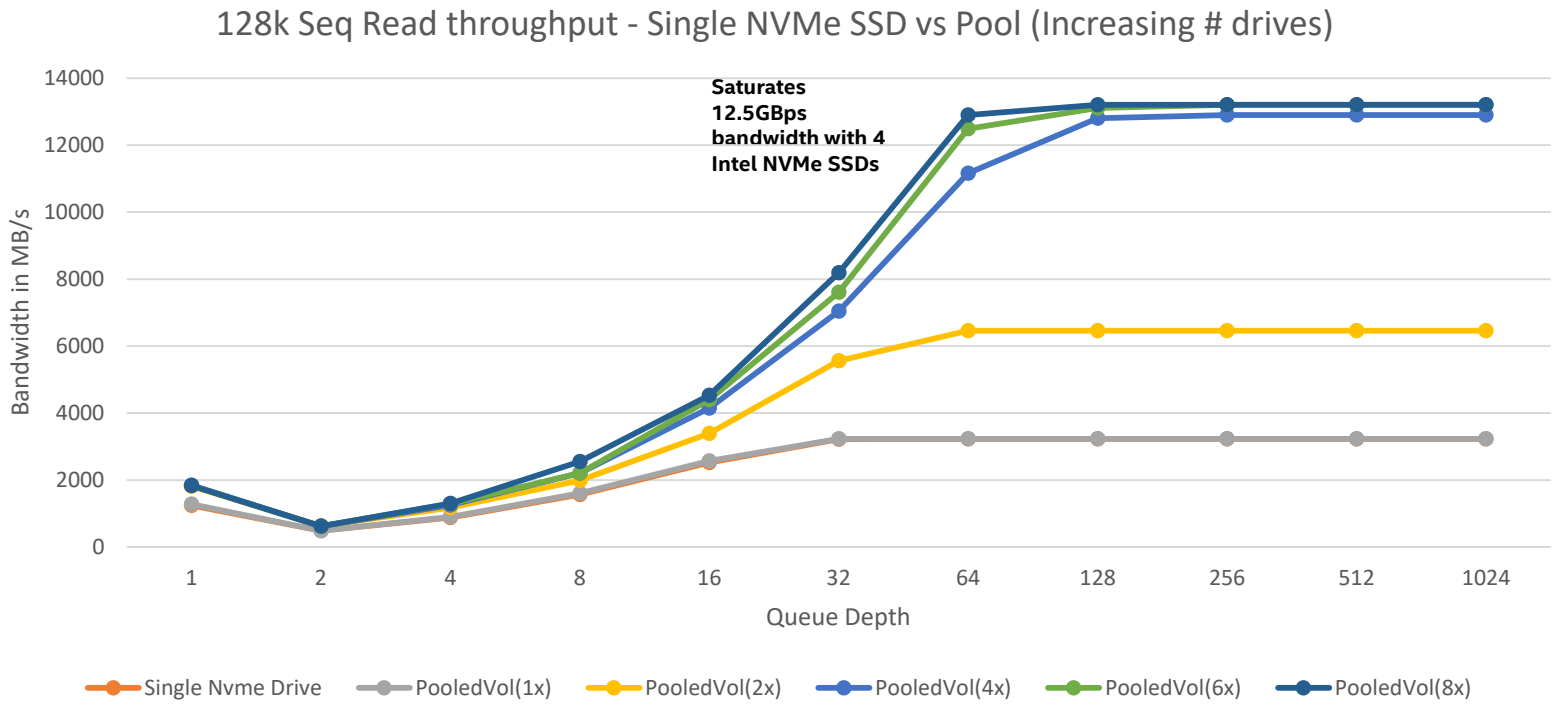
CBS NVMeoF SSD performance data





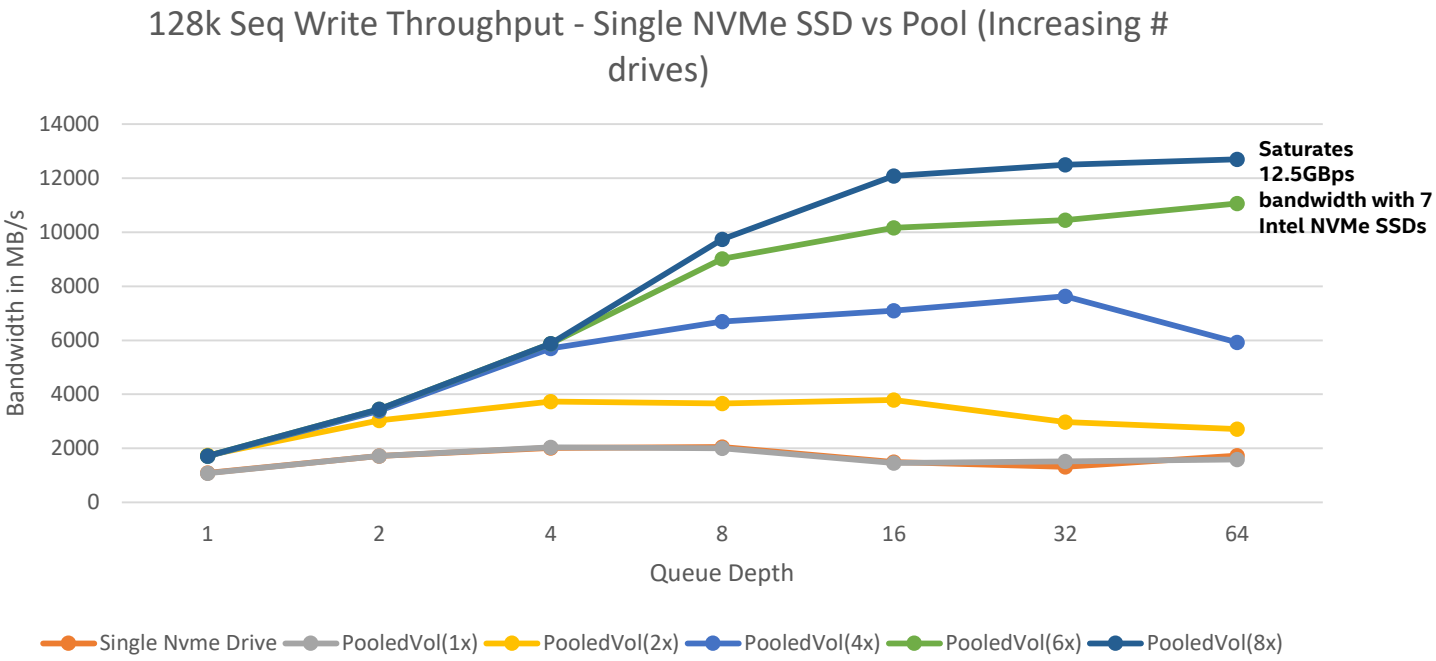
Pool Configuration - Performance

Throughput scales as pool width scales



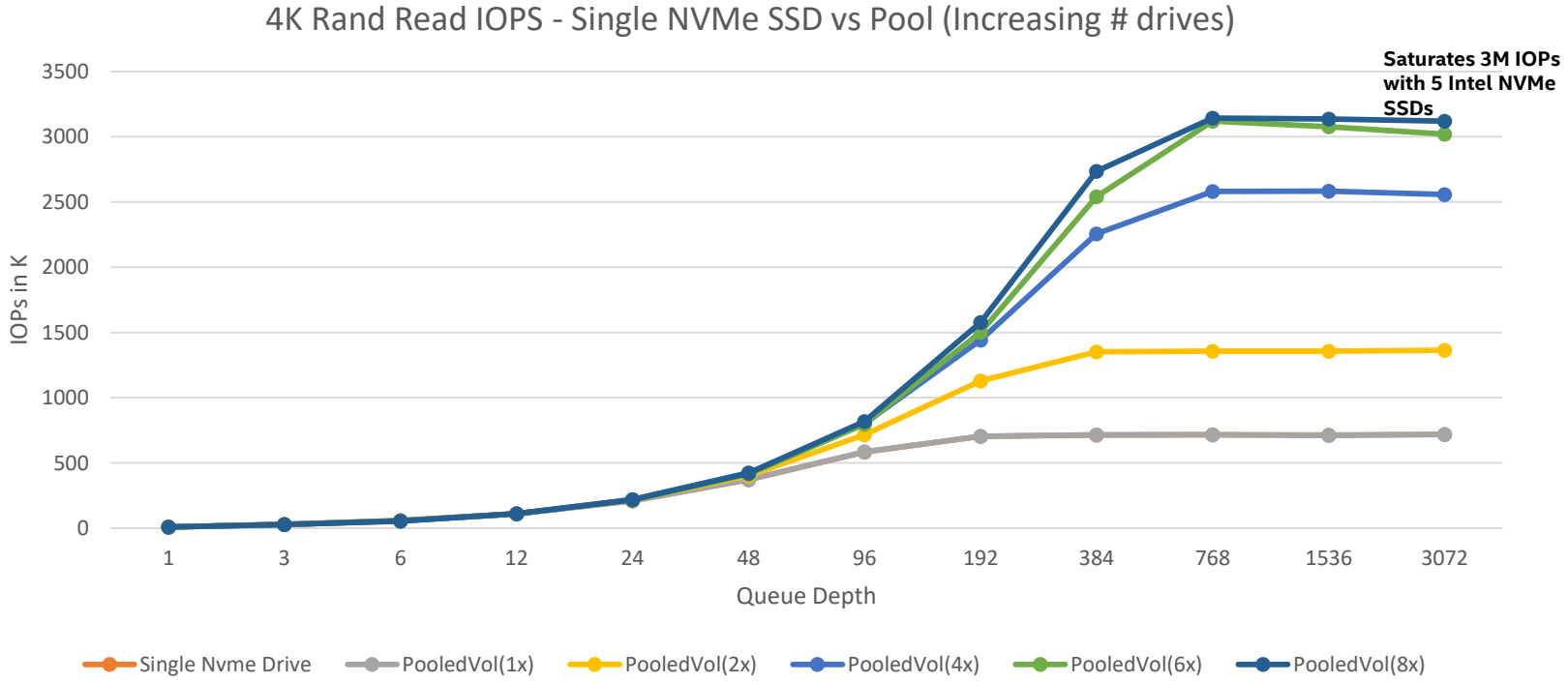
Pool Configuration - Performance

Throughput scales as pool width scales



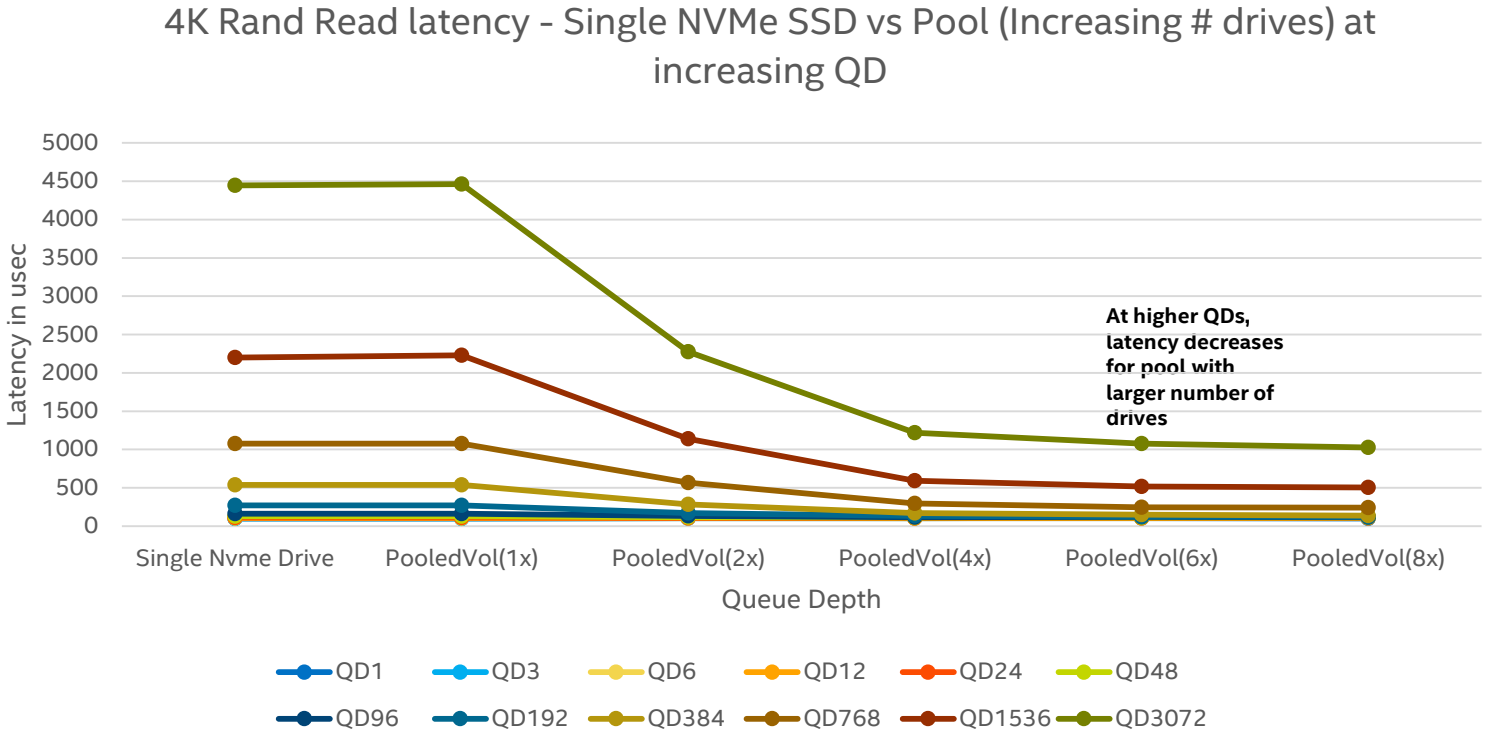
Pool Configuration - Performance

IOPs scale as pool width increases



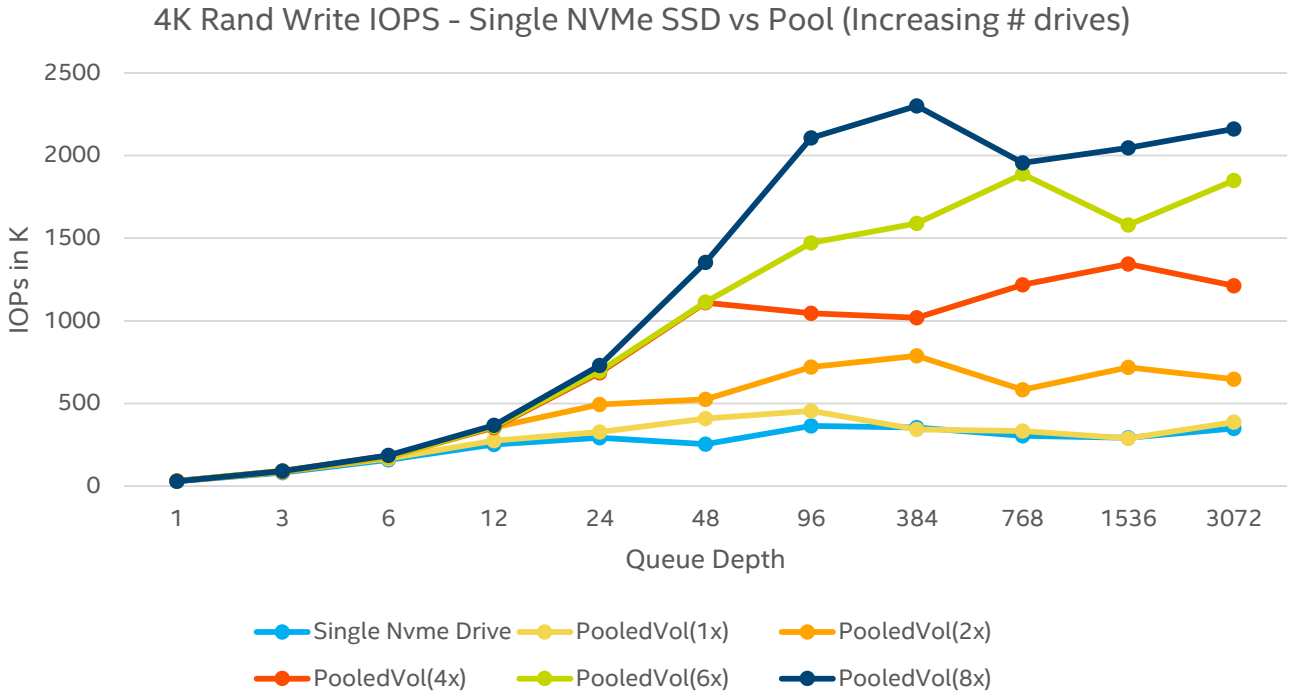
Pool Configuration - Performance

Latency decreases as pool width increases



Pool Configuration - Performance

IOPs scale as pool width increases



- **Random writes scales with increasing number of drives.**
- **Variations seen due to underlying Intel NVMe SSDs**


intel

CBS with SPDK & intel RSD Architecture

- Intel CBS solution uses the SPDK stack as its base platform and leverages the Intel RSD architecture for management and cloud OS integrations.
- CBS software delivers the best from Intel components like Xeon Processors, SSDs, Networking, Optane Persistent Memory with optimized smart algorithms implemented using SPDK stack.
- Apart from new features, CBS also leverages many of SPDK features like NVMeoF target, volume management , Poll mode driver etc.