apeiron
DATA SYSTEMS

NVMe Storage Networking

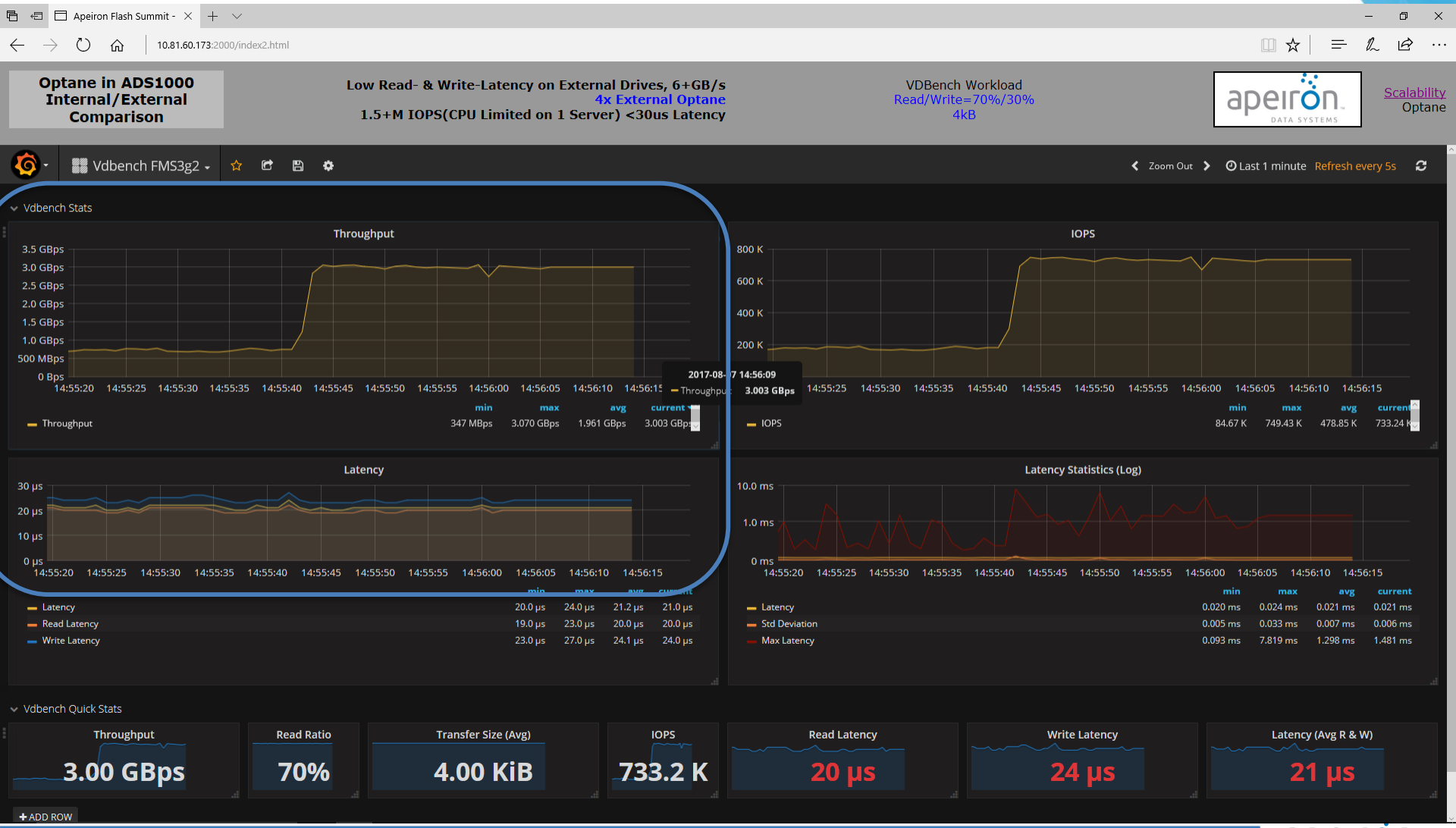
Bob Hansen – V.P. System Architecture - FMS '17

Agenda –
The Ideal Storage Network Solution

- Business and Technology Drivers
  - The 3rd Platform – Scale-out
  - Storage Aware Applications
  - NVMe
  - Flash / Storage Class Memory Performance
  - Storage Network Requirements
- Storage Network Architecture
  - Requirements
  - HW / SW Architecture
- Results

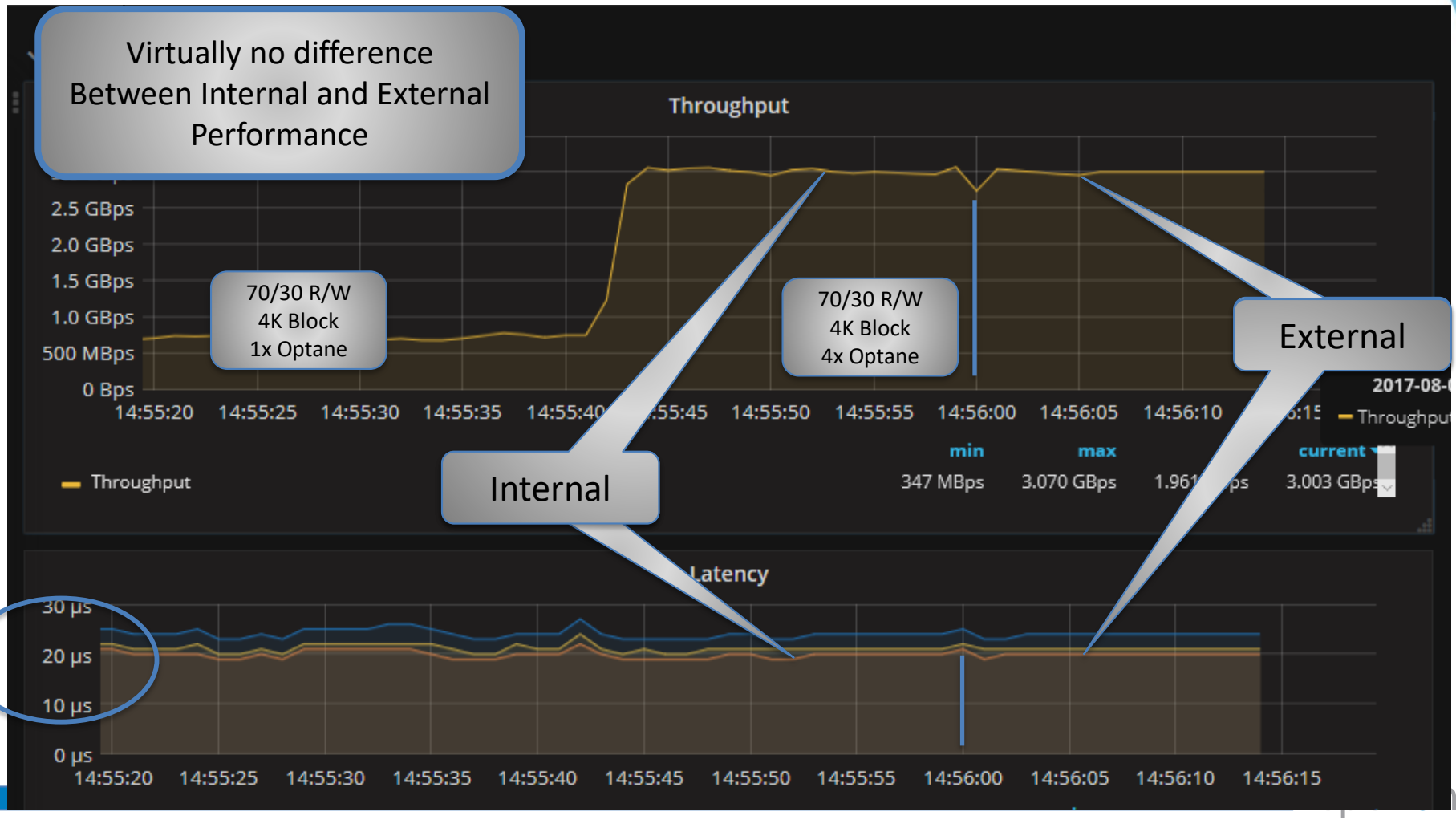
Storage Networking Performance

See it live in Booth 422



Storage Networking Performance With 3D-XPoint™ (Optane)

See it live in Booth 422



Why Not Captive Storage

- Captive (direct attached) Storage
  - Limited total capacity and performance
  - No dynamic scaling
  - No SSD virtualization
  - No high performance data sharing / tiering across cluster
  - A severe management challenge
  - Inefficient power, cooling, rack space
  - ***Storage provisioning is tied to CPU scale out***
  - AIC solutions are worse!

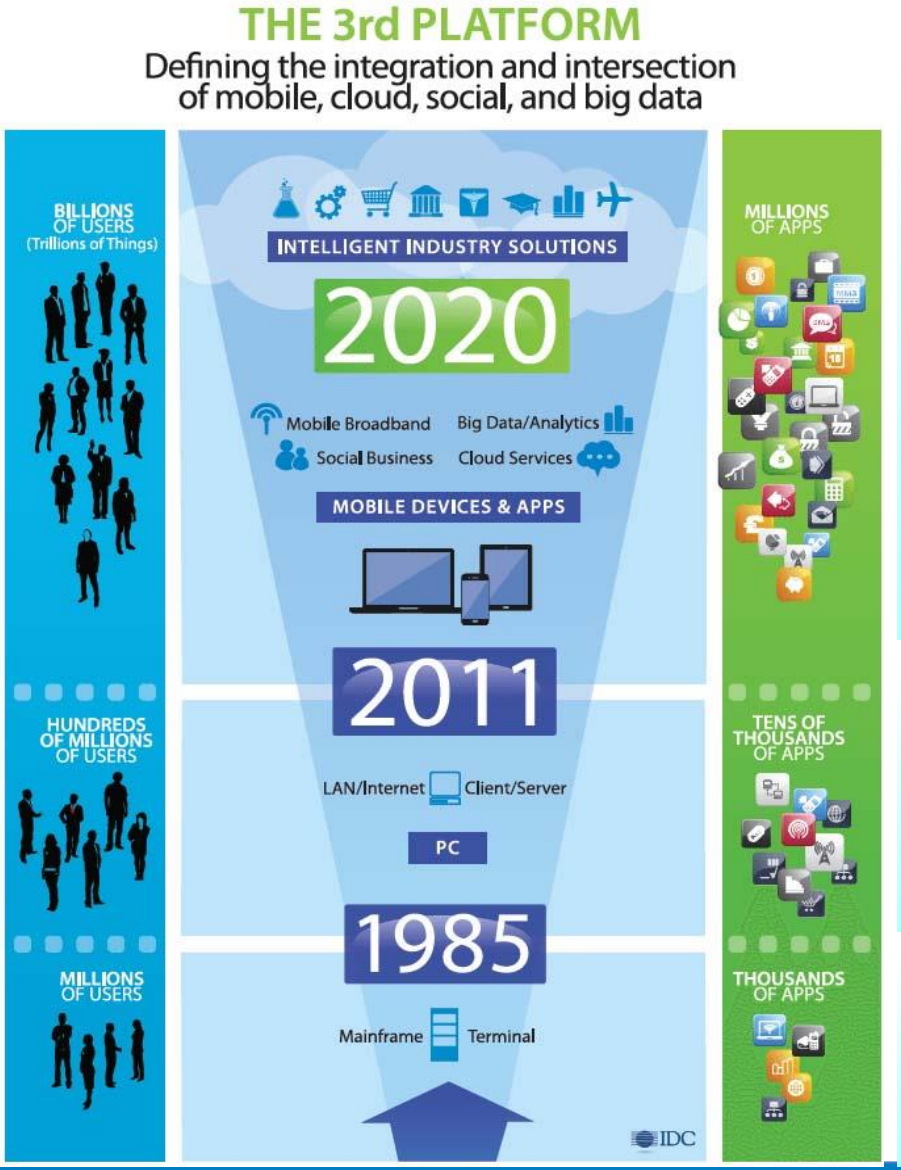
Get the Storage Out of the Server!! (again)

From the Session Description

"~~Flash~~ NVMe Storage Networking"

- Advantages
  - Higher utilization
  - Simpler expansion
  - Simpler Scalability
  - *High Availability*
  - *Much better time to recovery*
  - *Multiple Storage Tiers on Single Network*
- Disadvantages – *of "classic" storage networks*
  - Greater complexity
  - Difficult Maintenance
  - Higher Latency

The Migration of Storage Intelligence

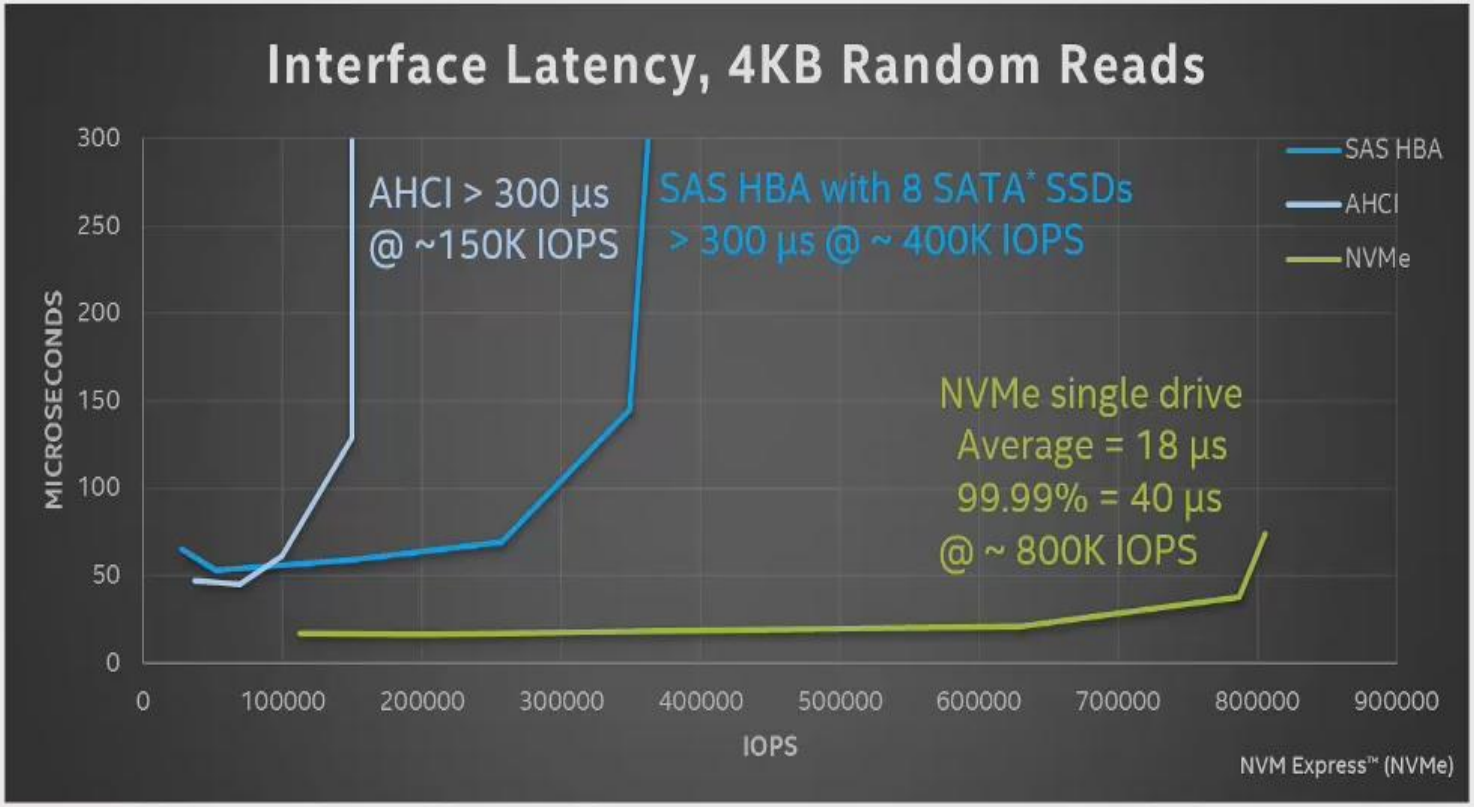


- 3rd Platform Storage
  - Millions of developers (open source)
  - Storage aware applications (and OS)
    - Architected for Scale out – not scale up
    - In-memory data base, native tiering
    - Sever is the critical component
  - Very High Performance persistent storage
    - NVMe
    - Flash now, Storage Class Memory soon
  - Very intelligent storage devices
    - > 500K lines of code in an NVMe controller
  - Direct attached storage >= networked
- 2nd Platform Storage
  - Simple storage drivers, SCSI, smarter devices
  - Network attached storage (SAN)
  - Array Controller centric intelligence
    - > 25M lines of code in storage controller SW release
- 1st Platform Storage
  - Direct Connect, Dumb Storage Hardware
  - Software (OS) centric storage management and control

3rd Platform Storage Drivers

- Millions of developers (open source)
- Storage aware applications and SDS
  - Architected for Scale out – not scale up
  - In-memory data base, native tiering
  - Sever is the critical component
- Very High Density 3D NAND
  - 30TB SSDs this year – 720TB in 2 Rack Units (standard SFF SSD)
  - New High density form factors - > 1TB in 1 Rack Unity
- Very High Performance persistent storage
  - NVMe
  - Flash and ***Storage Class Memory***
- Very intelligent storage devices
  - > 500K lines of code in an NVMe controller
- Very High Performance Storage Network

NVMe™ Delivers Higher IOPs and Better QoS



NVMe™ delivers 18 µs average and 40 µs 99.99% interface latency. Other interfaces have outliers in 100s of µs as interface reaches saturation.

Results measured by Intel based on the following configurations. Intel Server Board S2600WTT with 28 E5-2695 CPUs, 2 sockets, 2.3 GHz clock speed per CPU, Ubuntu* 14.04.1 LTS (GNU/Linux* 3.16.0-rc7tickles x86_64), idle=poll kernel settings, SAS HBA is LSI SAS9207-4i4e with controller LSI SAS 2308. SATA SSDs are Intel® SSD DC 3500 at 800 GB. NVMe SSD is Intel SSD P3700 at 1.6 TB. Workload details are Workload: 4K Random Reads using FIO – 4 + threads. Drives tested empty to test interface only (no NVM access.)

IDF15 INTEL DEVELOPER FORUM

apeiron

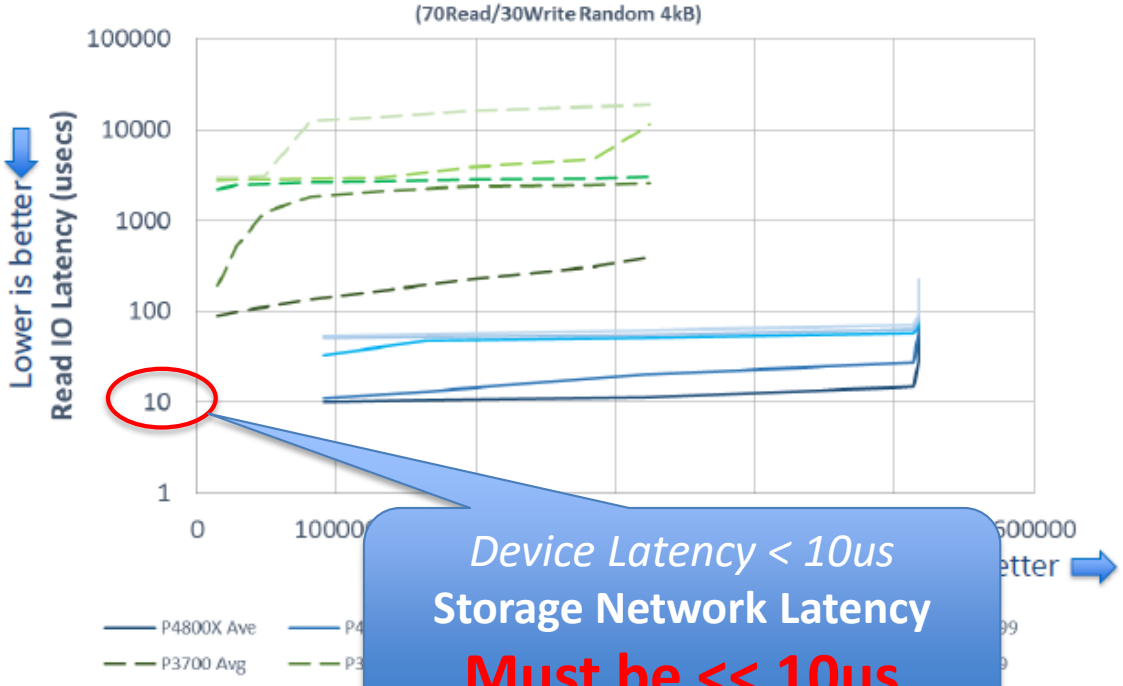
NVMe delivers

- Performance
- Managability
- Robust ecosystem
- Well defined standard SSD form factor
- Steep innovation and healthy competition
  - Performance, durability, capacity and cost

Storage Network Latency Requirements

STORAGE PERFORMANCE CHARACTERIZATION

Latency vs. Load: NAND SSD vs. Intel® Optane™ SSD (Intel® DC P3700 vs. Intel® Optane® P4800X)



Device Latency < 10us

Storage Network Latency

Must be << 10us

10x latency reduction

- < 10usec latency †

100x QoS improvement

- < 200usec 99.999th r/w †

†vs. NAND based SSD

NVM Solutions Group | Intel® Optane™ Technology Workshop | intel

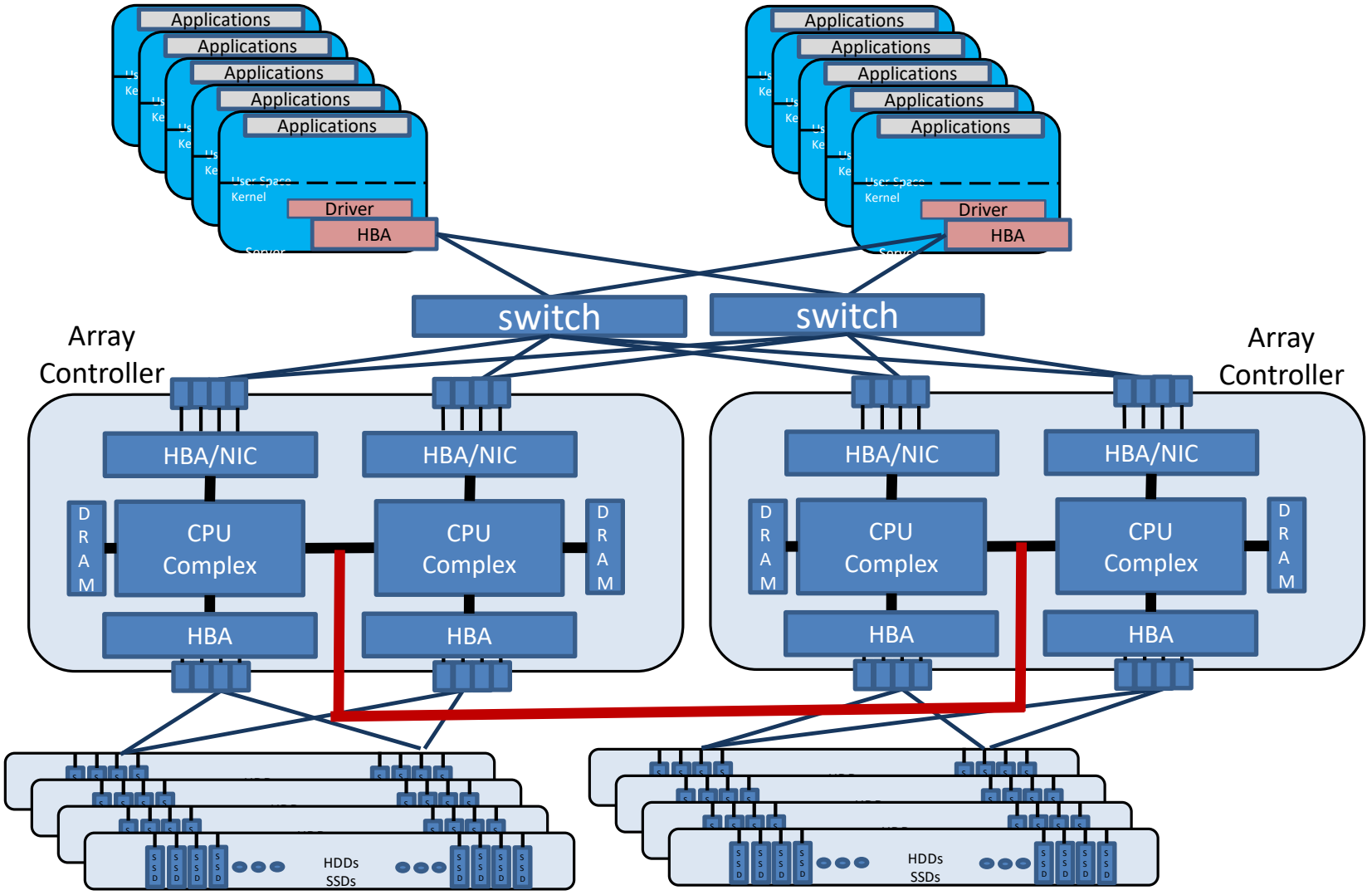
The Ideal NVMe Storage Network

- High Throughput, Very Low Latency
  - Latency must be << Storage Class Memory
- Supports any NVMe Device
  - Performance, Endurance, Cost
  - High Performance Cache to Archive (HDD replacement)
- High Availability
- Simple, standards based
- Low Cost
- *Not required*
  - Rich data management feature set

Deliver **all** of the performance of the NVMe Device to the Application

2nd Platform "Classic" Storage Array

including all flash arrays



Array Controllers are DEAD! (well, dying)

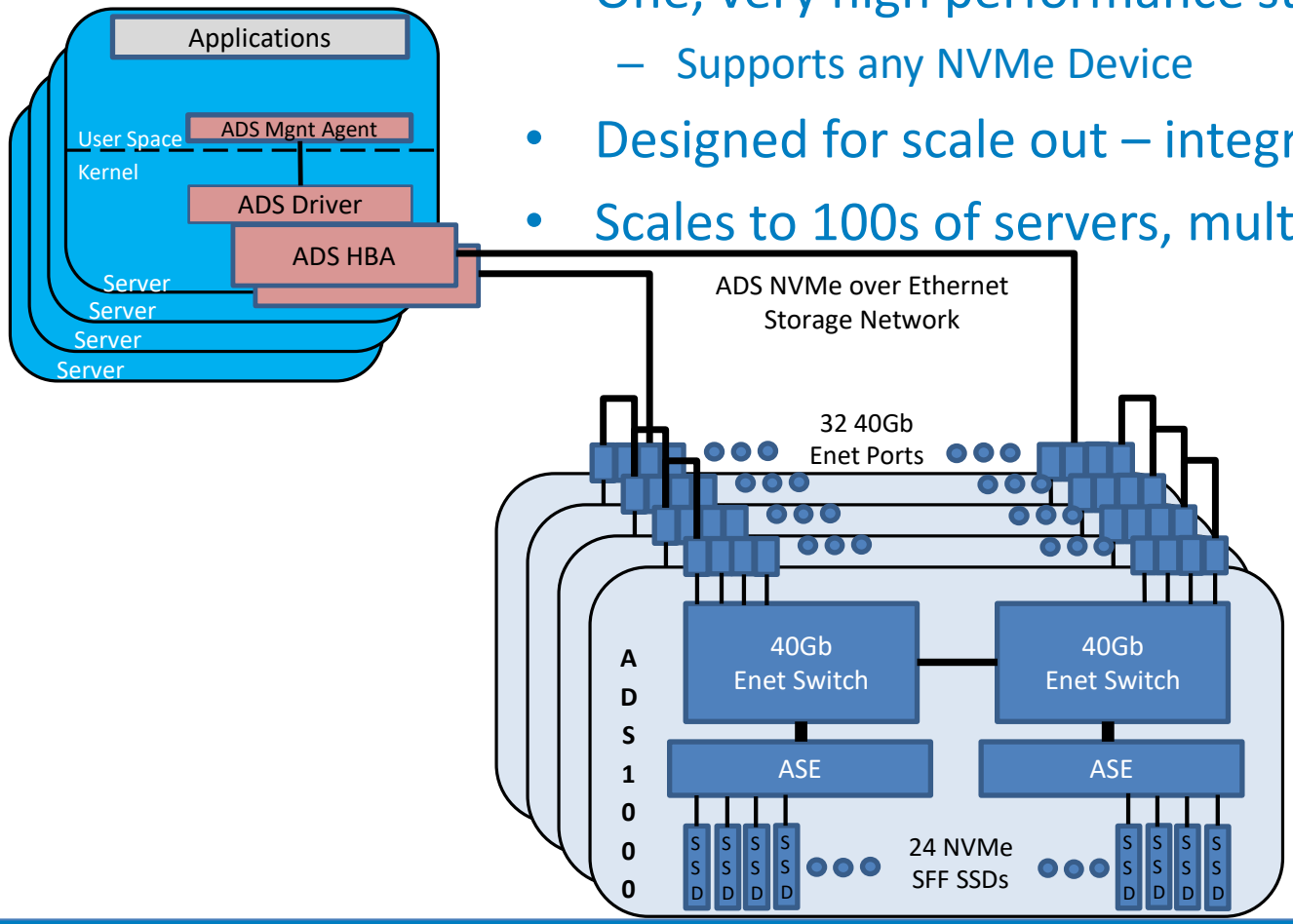
- Storage aware, scale-out, real time analytics
  - Application manages data placement across compute cluster
  - Server is now the critical component – HA / failover strategies required
  - Application manages HA, data tiering and migration
  - Complex, array centric data management = slow perf – new apps just don't use it
  - Multiple tiers of storage are a given and managed by the app.
    - DRAM, High Performance NVMe, , flash, Archive then maybe HDD
- NVMe storage devices
  - Solid state drives are more reliable than HDDs
    - HDD MTBF drove the development of Storage Array Controllers (NetApp, EMC, HP . . .)
  - But flash wears out – requires complex management code including loads of data movement "behind the curtains" in the device
  - NVMe standard was written for flash controllers with ample processing power
    - Excellent device management and monitoring
  - NVMe controller handles data movement
    - Turns the SCSI model upside down

Array Controllers Just Get In The Way of Scale-out Apps on NVMe

3rd Platform External Storage

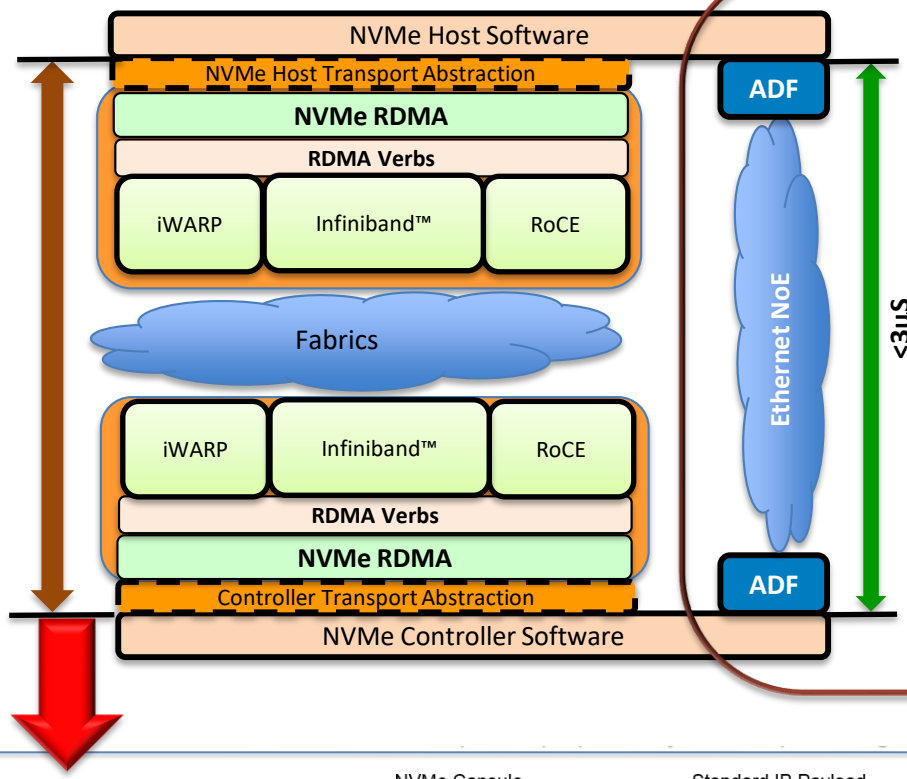
Direct Attach scale-out storage

- Software / application defined storage
- One, very high performance storage network
  - Supports any NVMe Device
- Designed for scale out – integrated switches
- Scales to 100s of servers, multiple petabytes



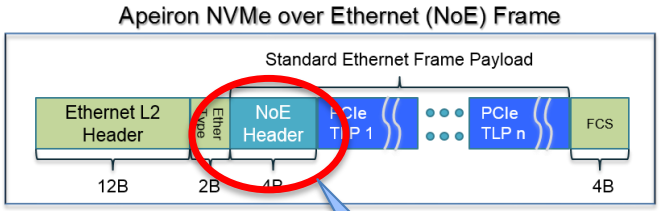
apeiron

NVMeoF (RDMA) / NoE Comparison



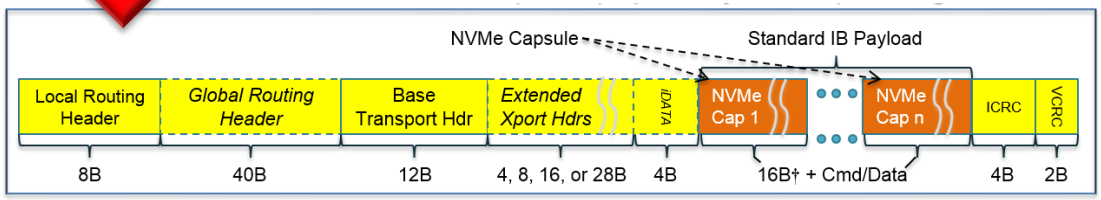
Apeiron Data Fabric™

- Simple, Robust
- Optimized for Ethernet and NVMe
- ***light weight, hardened*** Layer 2



22B Fixed Overhead

L2 Ethernet Headers + 4B NoE H der/encapsulated PCIe TLPs



42B ~
> 112B
Overhead

Optional in some configurations

Only 4 Bytes added Overhead

The standard is not tied to any particular physical layer

RDMA approach adds between 26B and 96B of headers, in addition to NVMe Encapsulation

Flexible but adds complexity, link consumption and latency !

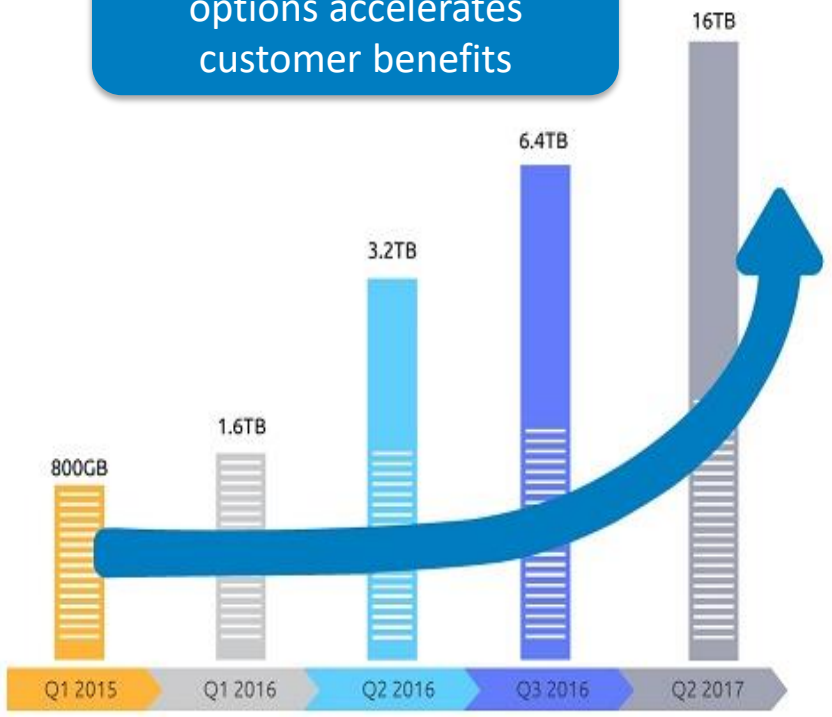
Storage Management

- SSD Virtualization
  - Many to one, one to many
  - SSD slicing
- Mirroring (2 or 3 way)
  - 2x read performance when mirrored
  - Policy based rebuild
- High Availability / Scale Out
  - Hot spares, Hot provisioning, Hot FRU replacement
- Storage Network Management
  - Single point of management for entire Storage Network including all SSD enclosures and HBAs
  - REST, CLI, SNMP

The World's Only Universal NVMe Platform

- Unlike captive storage, Apeiron enables independent scaling of servers and storage
- Compatible with ANY commercial NVMe drive-Data resides on appropriate SSD type for its value
  (Including *3D XPoint™ technology*)
- Adoption of NVMe SSD's is rapidly increasing; Only Apeiron can provide compatibility with all suppliers and drive profiles

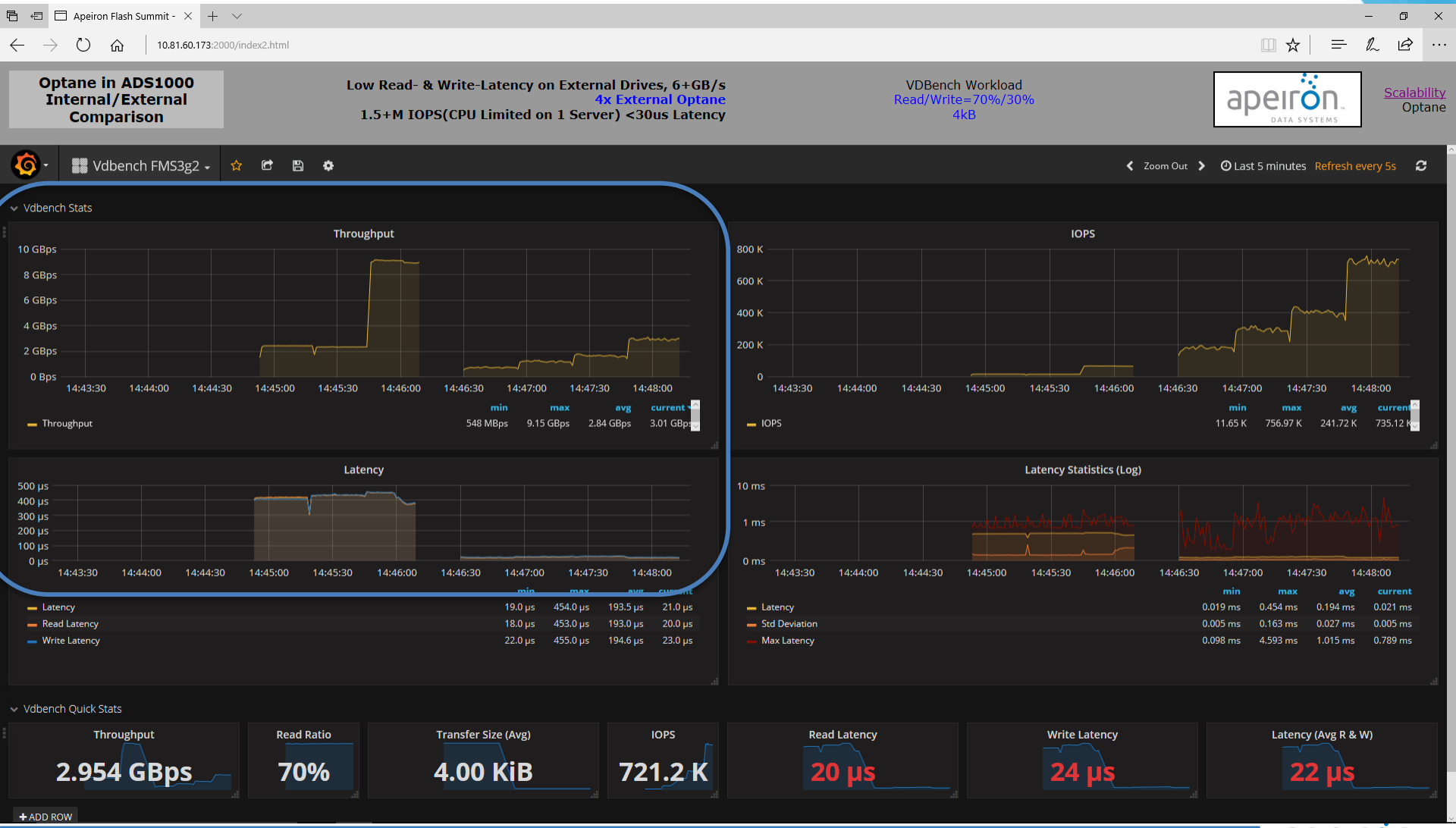
Offering multiple NVMe options accelerates customer benefits



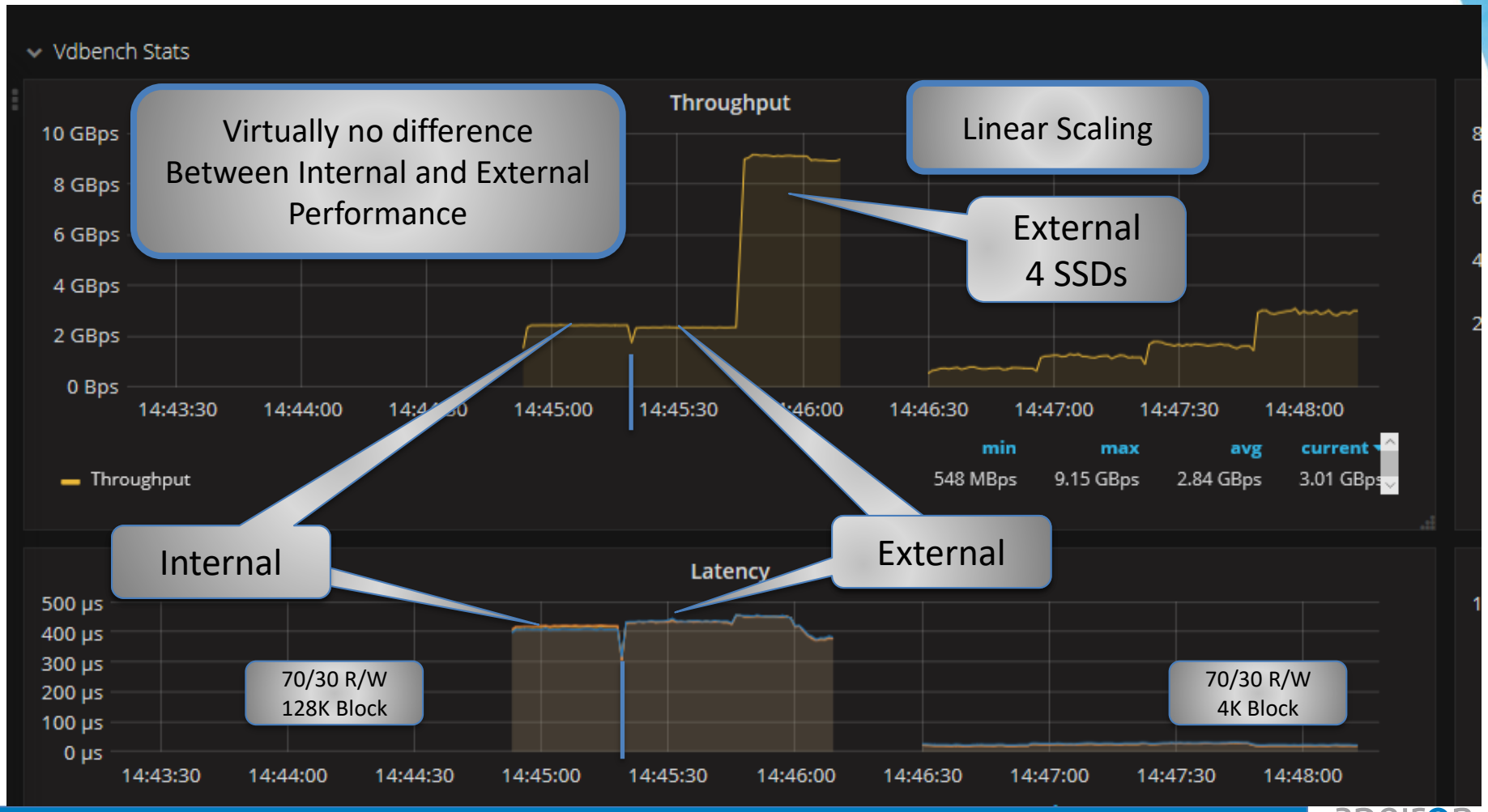
The roadmap for density and performance of NVMe SSD's is accelerating; Apeiron passes this advantage to the customer

Storage Networking Performance

See it live in Booth 422



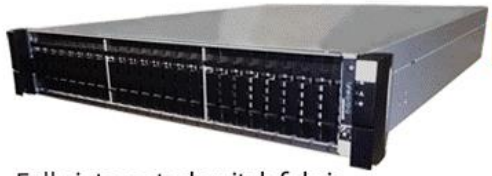
Ideal Storage Network Performance with Optane 3D-XPoint™



ADS1000 Scale-out NVMe Solution Unmatched Performance, Scalability and Efficiency

720TB Sep '17

24 NVMe 2.5" SSD



Fully integrated switch fabric

32 Apeiron Data Fabric Ports | Field Serviceability | Redundant PS & Cooling Modules

ADS1000 Performance (2U)	
Capacity	38/76/154/184/360TB
Latency (NAND LIMITATION)	100µs
Protocol Overhead	<3µs (roundtrip)
Bandwidth sustained	72 GB/s (drive limited)
Random 4K reads	**18.4 M IOPS**

x2 ADS40G-HBA | 40 GbE Data Fabric ports | Dual port 10 GBaseT

apeiron™
DATA SYSTEMS

Thank You
Come See Us – Booth 422

bob@apeirondata.com