Accelerating Real-Time Big Data

Breaking the limitations of captive NVMe storage – 18M IOPs in 2u

Agenda

- Everything related to storage is changing!
  - The 3rd Platform
  - NVM Express – architected for solid state storage
- Captive storage limitations
- Apeiron architecture – NVMe over Ethernet
  - Noe / NVMe over Fabrics comparison
- The Apeiron Storage Solution
  - Captive storage vs. external storage performance
  - 18.4M IOPs in 2 rack units
  - The universal NVMe storage platform
  - Advanced features

The 3rd Platform:
A Fundamental Shift in the IT Industry

3nd Platform = 12.7% CAGR

2nd Platform = -5.1% CAGR

Digital Transformation

Augmented & Virtual Reality

Internet of Things

Cognitive Systems

Robotics

Next Gen Security

INNOVATION ACCELERATORS

3D Printing

Cloud

Big Data/ Analytics

Social Business

Mobility

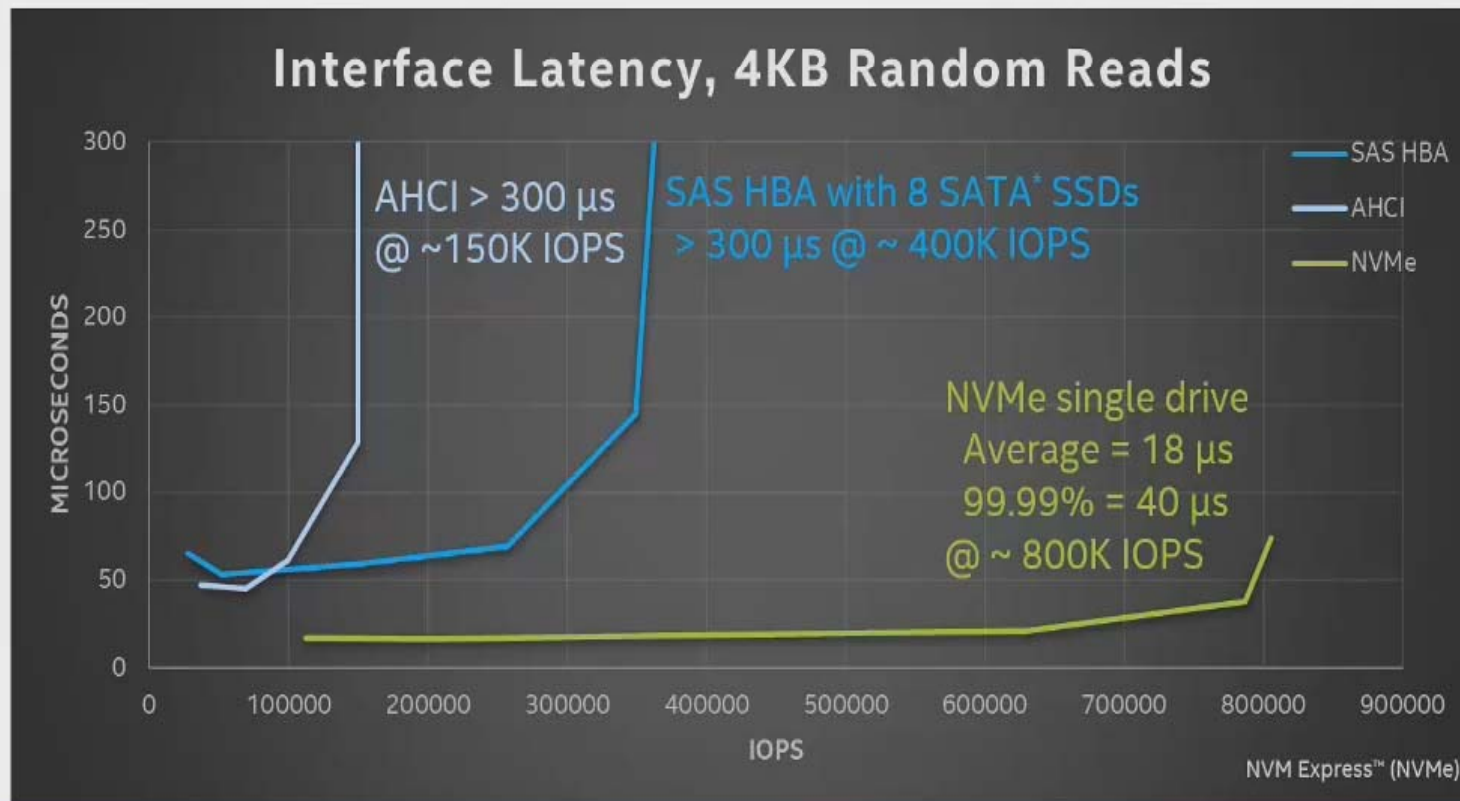3rd Platform

- Cloud, big data/analytics, mobility, social media collectively define the 3rd computing platform

- Underneath the 3 - 5% overall IT market growth projection there is a significant opportunity
  - Certain markets growing at an excess of 20%

- Businesses are Changing –
  - How they engage with customers
  - The speed at which they deliver products and services
  - How they innovate
  - The reliability of their operations
  - Their overall resiliency

Driving the move to high performance, scale-out applications

Source: IDC Worldwide Black Book, additional modeling

NVMe™ Delivers Higher IOPs and Better QoS

Interface Latency, 4KB Random Reads

AHCI > 300 µs @ ~150K IOPS

SAS HBA with 8 SATA* SSDs > 300 µs @ ~ 400K IOPS

NVMe single drive
Average = 18 µs
99.99% = 40 µs
@ ~ 800K IOPS

Legend: SAS HBA, AHCI, NVMe

MICROSECONDS (y-axis): 0, 50, 100, 150, 200, 250, 300
IOPS (x-axis): 0, 100000, 200000, 300000, 400000, 500000, 600000, 700000, 800000, 900000

NVM Express™ (NVMe)

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.)

9

IDF15
INTEL DEVELOPER FORUM

Proof Point – Database Analytics

A Lenovo* ThinkServer* RD650 with four NVM Express™ (NVMe) SSDs transformed the performance of the SQL database workload.

	NVMe SSDs	SATA SSDs	Benefit
Total Database Performance Queries/Hr Across All Instances	25,062.1	9,524.8	**2.63 X**
Single Database Performance Queries/Hr Across One Instance	6,265.5	4,762.4	**1.31 X**
Time to Answer Average Query Time	27.9 min	38.0 min	**10 MINUTES**

NVMe enables one server to replace four legacy servers

Detailed whitepaper at http://www.principledtechnologies.com/Lenovo/RD650_storage_performance_0415.pdf.

NVMe delivers

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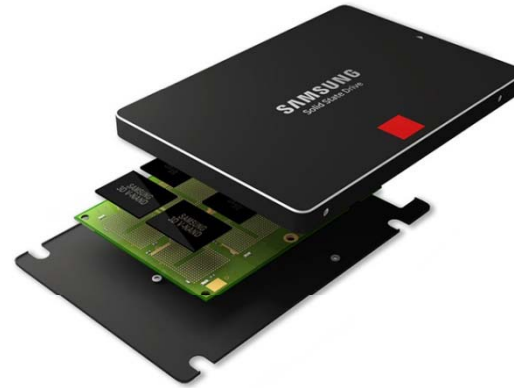
Proof Point – Database Analytics

A Lenovo* ThinkServer* RD650 with four NVM Express™ (NVMe) SSDs transformed the performance of the SQL database workload

	NVMe SSDs	SATA SSDs	Benefit
Total Database Performance Queries/Hr Across All			2.63 X
Single Database Performance Queries/Hr Across One instance			1.31 X
Time to Answer Average Query Time	27.9 min	38.0 min	10 MINUTES

Captive Storage is a huge problem

NVMe enables one server to replace four legacy servers

Detailed whitepaper at http://www.principledtechnologies.com/Lenovo/RD650_storage_performance_0415.pdf.

Captive Storage Limitations

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

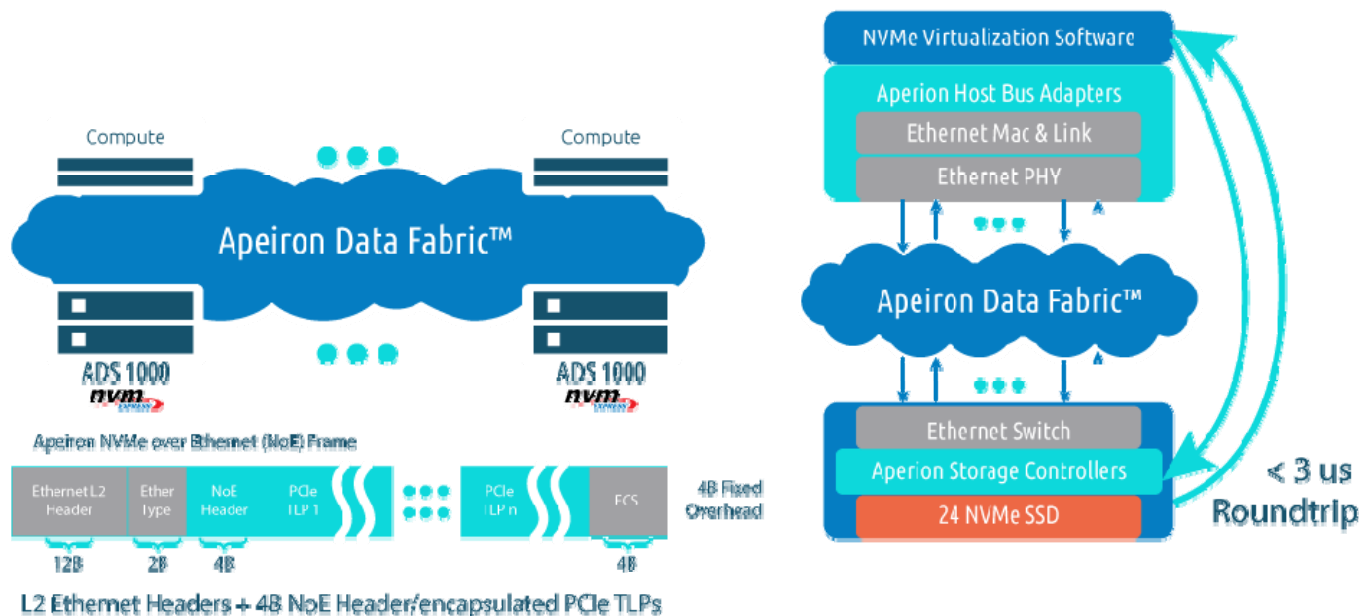**Get the Storage Out of the Server!!**

Apeiron Addresses

- Apeiron was founded to address three inflection points in the storage market:

    1. Move to server-managed scale-out storage and storage aware apps
    2. Demand for real-time queries on massive data sets
    3. The rapid adoption of NVMe as the standard SSD interface

- Apeiron delivers the highest performance density of any SSD platform on the market

- Ethernet and FPGA based networking acceleration enables ultra-low latency switching to 1000's of NVMe Drives
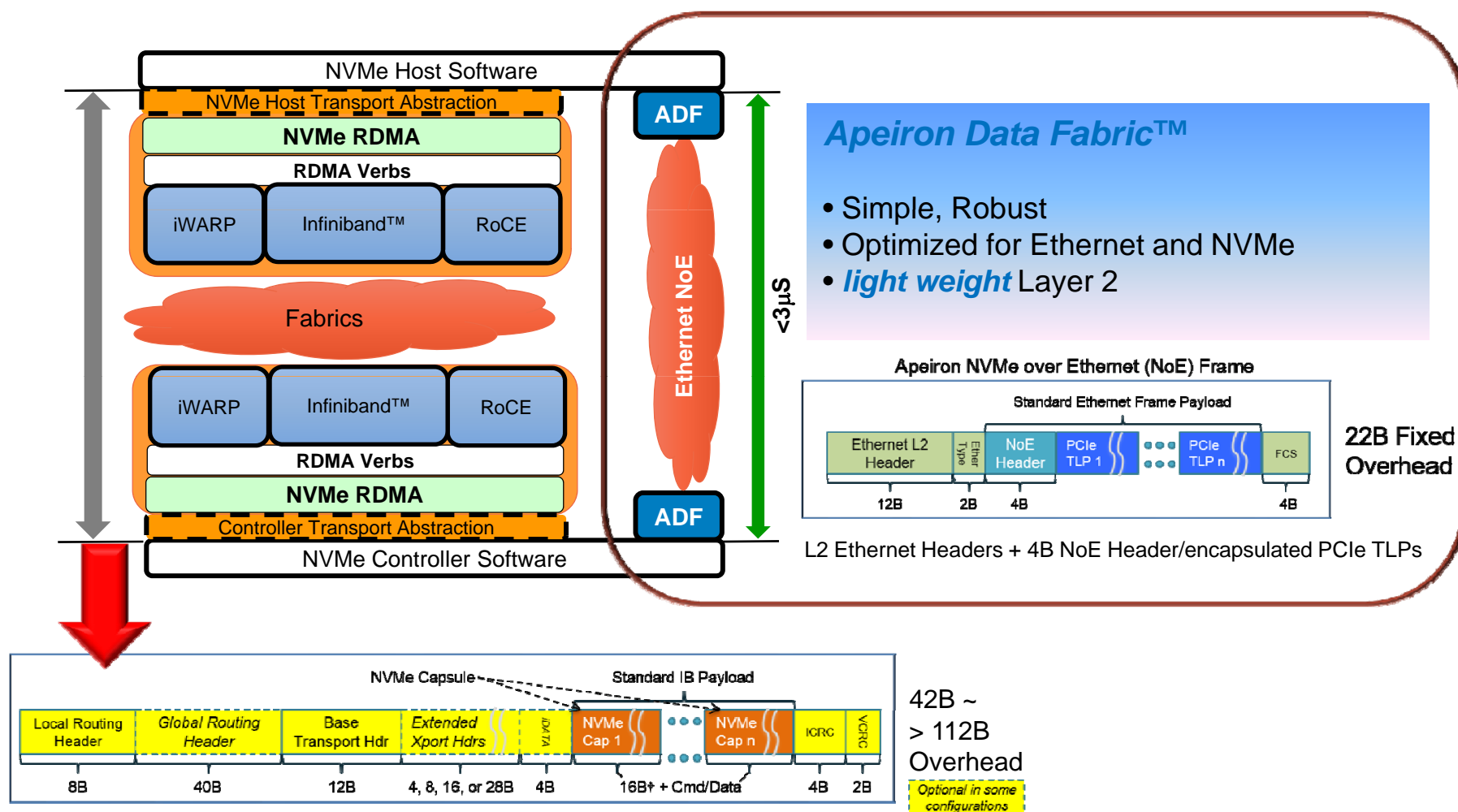
NVMe over Ethernet (NoE)

- Hardware accelerated Layer 2 Ethernet fabric
  - Layer 3 robustness without the overhead and latency

- Fully integrated NVMe fabric (no external switching)

- The industry's lowest latency transport protocol delivers predictable performance at scale

NVMeoF (RDMA) / NoE Comparison

Apeiron Data Fabric™

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

Apeiron NVMe over Ethernet (NoE) Frame

22B Fixed Overhead

L2 Ethernet Headers + 4B NoE Header/encapsulated PCIe TLPs

42B ~ > 112B 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 !

NoE / NVMe over Fabrics comparison

NVMe over Ethernet (NoE)	NVMe over Fabrics
Transports NVMe commands	Transports Data (RDMA)
Optimized for Ethernet (minimizes overhead)	Transport independent (more complexity)
4 Byte per packet added overhead	>> overhead (depends on implementation)
Optimized for scale-out clusters	Architected for traditional storage arrays
Supports ANY standard NVMe SSD	????
Next gen (3D XPoint) ready	???? (demonstrated latency is a problem)
Shipping today	NVMeoF standard out for approval

Apeiron is in production *today*
shipping the highest performance scale out
NVMe storage solution
in the world

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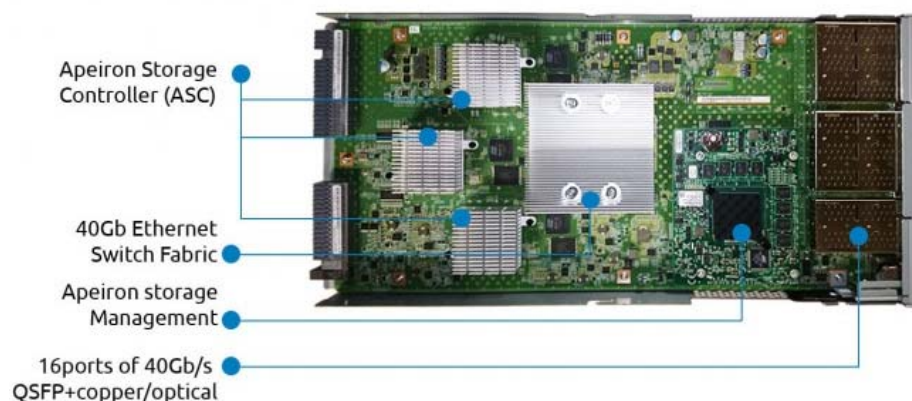

nvm EXPRESS

24 NVMe 2.5" SSD

Front

Fully integrated switch fabric

Back

32 Apeiron Data Fabric Ports | Field Serviceability | Redundant PS & Cooling Modules
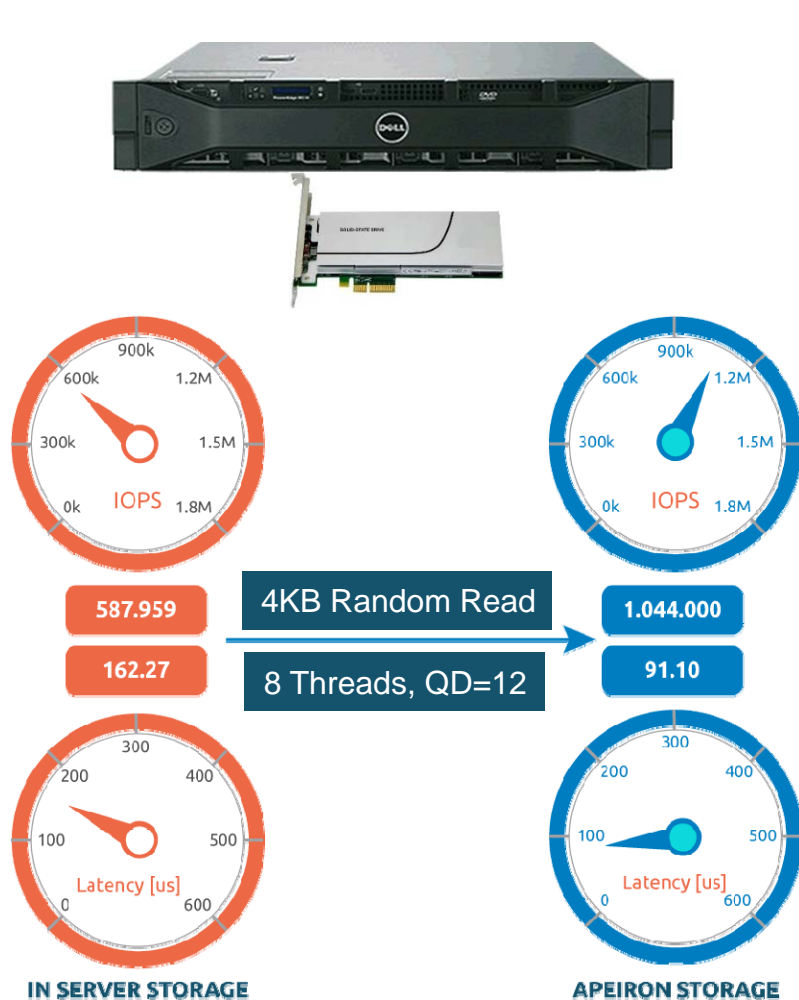
ADS1000 Performance (2U)	
Capacity	38/76/154/192TB
Latency (NAND LIMITATION)	100us
Protocol Overhead	<3us (roundtrip)
Bandwidth sustained	72 GB/s
Random 4K reads	**18.4 M IOPS**

Apeiron Storage Controller (ASC)

40Gb Ethernet Switch Fabric

Apeiron storage Management

16ports of 40Gb/s QSFP+copper/optical

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

A New Standard in Storage Networking Performance

24 NVMe 2.5" SSD

IOPS

587.959

162.27

4KB Random Read

8 Threads, QD=12

1.044.000

91.10

IOPS

Latency [us]

Latency [us]

IN SERVER STORAGE

APEIRON STORAGE

The ultra low latency Apeiron network technology is 100% transparent to the servers*

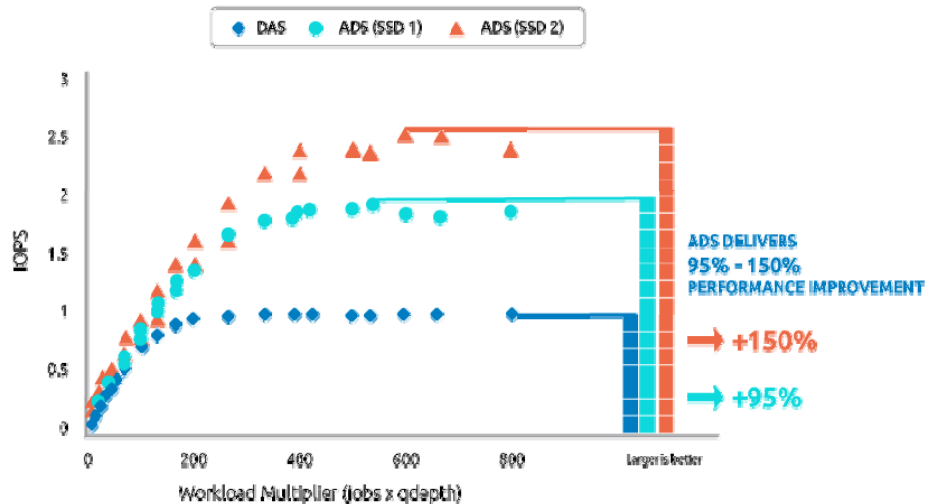Apeiron vLUN's enable workload optimization across multiple NVMe drives

FIO Performance Benchmarks

RANDOM READ, 4K, IOPS VS MULTIPLIER

◆ DAS ● ADS (SSD 1) ▲ ADS (SSD 2)

IOPS (y-axis)
Workload Multiplier (jobs x qdepth) (x-axis)

ADS DELIVERS
95% - 150%
PERFORMANCE IMPROVEMENT
→ +150%
→ +95%

Vendor B
770k per SSD

18.4M IOPS – 2U

18,400,000 **Industry Leading Performance Density**

- The balanced architecture of the ADS1000 improves server utilization up to 3x (limited by SSD type)

- Apeiron's driver technology moves the bottleneck from the network to the NAND architecture itself
  (18.4M IOPs single ADS1000)

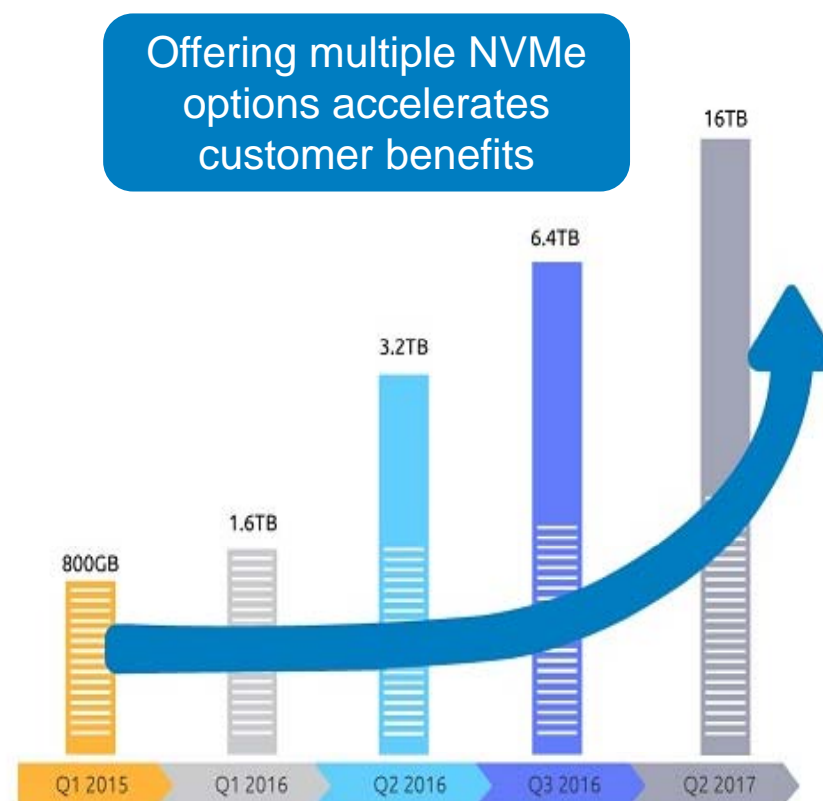- Performance scales linearly to 100's of millions of IOPs as the system solution grows

M IOPS (y-axis)
RACK UNITS (x-axis)
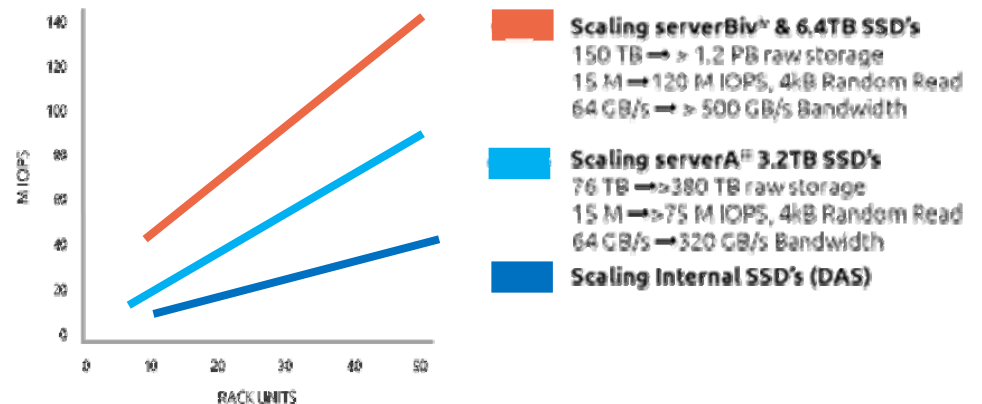
Scaling serverBiv & 6.4TB SSD's
150 TB → > 1.2 PB raw storage
15 M → 120 M IOPS, 4kB Random Read
64 GB/s → > 500 GB/s Bandwidth

Scaling serverA 3.2TB SSD's
76 TB → >380 TB raw storage
15 M → >75 M IOPS, 4kB Random Read
64 GB/s → 320 GB/s Bandwidth

Scaling Internal SSD's (DAS)

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

16TB

6.4TB

3.2TB

1.6TB

800GB

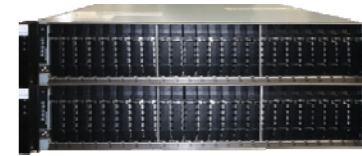| Q1 2015 | Q1 2016 | Q2 2016 | Q3 2016 | Q2 2017 |

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

All mentioned brand names are registered trademarks and property of their respective owners.
"3D XPoint is a trademark of Intel Corporation in the U.S. and/or other countries"

NVMe Solution Comparison

	System A	ADS1000
Rack Units	5U	**4U (2x 2U)**
Bandwidth	100GB/s	**144GB/s**
IOPs	10M IOPs	**37M IOPs**
SSD	Proprietary	***Any* SFF NVMe SSD**
Latency	100us (avg)	**100us**
Interconnect	PCIe 3.0	**40Gb Ethernet**
Maximum Capacity	144 TB	**4.6 PB, 9.2 PB Q3'16** **(60 enclosures)**
Intel 3D XPoint™	No	**Yes** **3D Xpoint = 7us latency***
Entry Level List Price	3x Apeiron	**1/3 System A**

*Note: Publically Disclosed By Intel

Performance, Scalability and Simplification

"All the simplicity and promise of direct attached storage with the capabilities of network attached storage."

-Ahmed Shihab, VP Engineering, Amazon Web Services