Accelerating Real-Time Big Data

Breaking Through Performance and Scale Out Barriers –
A Storage Solution for Today's Hot Scale Out Applications
Agenda

• Everything related to storage is changing!
  • The 3rd Platform
  • NVM Express – architected for solid state storage

• The Migration of Storage Intelligence
  • Storage Aware, Scale-Out Applications
  • NVMe controllers

• Storage Array Controllers are DEAD!

• 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

3rd Platform = 12.7% CAGR

2nd Platform = -5.1% CAGR

- 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**

![Graph showing interface latency comparison between AHCI, SAS HBA, AHCI, and NVMe](image)

**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

---

**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 ES-2695 CPUs, 2 sockets, 2.3 GHz clock speed per CPU, Ubuntu* 14.04.1 LTS (GNU/Linux* 3.16.0-rc7itudes x86_64), idle=poll kernel settings, SAS HBA is LSI SAS9207-8i4e 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).
NVMe delivers

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

**THE 3rd PLATFORM**

Defining the integration and intersection of mobile, cloud, social, and big data

---

**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
Array Controllers are DEAD! (well, dying)

- Storage aware, scale-out, real time analytics
  - Examples: Ad Tech, fraud detection, facial recognition, personalized user experience, enterprise big data analysis, etc.
  - 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 storage SW = slow perf – new apps just don’t use it
  - Multiple tiers of storage are a given and managed by the app.
    - DRAM, captive NVMe, captive SATA, external flash then maybe HDD

- NVMe storage devices
  - Solid state drives are much 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 add no value for Scale-out Apps
2nd Platform Storage
including all flash arrays
3rd Platform External Storage
Direct Attach scale-out storage

- Software / application defined storage
- One, very high performance storage network
- Designed for scale out – integrated switches
- Scales to 100s of servers, multiple petabytes
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
  • PCIe board solutions are worse!

Get the Storage Out of the Server!! (again)
Storage Network Protocol
NVMe over Ethernet (NoE)

- Hardware accelerated, hardened 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
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

<table>
<thead>
<tr>
<th>NVMe over Ethernet (NoE)</th>
<th>NVMe over Fabrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transports NVMe commands</td>
<td>Transports Data (RDMA)</td>
</tr>
<tr>
<td>Optimized for Ethernet (minimizes overhead)</td>
<td>Transport independent (more complexity)</td>
</tr>
<tr>
<td>4 Byte per packet added overhead</td>
<td>&gt;&gt; overhead (depends on implementation)</td>
</tr>
<tr>
<td>Optimized for scale-out clusters</td>
<td>Architected for traditional storage arrays</td>
</tr>
<tr>
<td>Supports ANY standard NVMe SSD</td>
<td>????</td>
</tr>
<tr>
<td>Next gen (3D XPoint) ready</td>
<td>???? (demonstrated latency is a problem)</td>
</tr>
<tr>
<td>Shipping today</td>
<td>NVMeoF standard now approved</td>
</tr>
</tbody>
</table>

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

---

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

**ADS1000 Performance (2U)**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>38/76/154/192TB</td>
</tr>
<tr>
<td>Latency (NAND LIMITATION)</td>
<td>100us</td>
</tr>
<tr>
<td>Protocol Overhead</td>
<td>&lt;3us (roundtrip)</td>
</tr>
<tr>
<td>Bandwidth sustained</td>
<td>72 GB/s</td>
</tr>
<tr>
<td>Random 4K reads</td>
<td>18.4 M IOPS</td>
</tr>
</tbody>
</table>

24 NVMe 2.5" SSD

Fully integrated switch fabric

32 Apeiron Data Fabric Ports

Field Serviceability

Redundant PS & Cooling Modules

Apeiron Storage Controller (ASC)

40Gb Ethernet Switch Fabric

Apeiron storage Management

16 ports 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

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

Apeiron vLUN’s enable workload optimization across multiple NVMe drives

* Please see the March 2016 ESG Whitepaper-“Validation of Apeiron Performance” at apeirondata.com
FIO Performance Benchmarks

- 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

Vendor B
770k per SSD
18.4M IOPS – 2U

Industry Leading Performance Density
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

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

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

*Note: Publically Disclosed By Intel*
“All the simplicity and promise of direct attached storage with the capabilities of network attached storage.”

-Ahmed Shihab, VP Engineering, Amazon Web Services