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

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

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

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-rc7, x86_64), idles=4, kernel settings: SAS HBA is LSI SAS2007-4i4e 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 x threads. Drives tested empty to test interface only (no NVMe access).
Proof Point – Database Analytics

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

<table>
<thead>
<tr>
<th></th>
<th>NVMe SSDs</th>
<th>SATA SSDs</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Database Performance</strong></td>
<td>25,062.1</td>
<td>9,524.8</td>
<td>2.63 X</td>
</tr>
<tr>
<td>Queries/Hr Across All Instances</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Single Database Performance</strong></td>
<td>6,265.5</td>
<td>4,762.4</td>
<td>1.31 X</td>
</tr>
<tr>
<td>Queries/Hr Across One Instance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Time to Answer</strong></td>
<td>27.9 min</td>
<td>38.0 min</td>
<td>10 MINUTES</td>
</tr>
<tr>
<td>Average Query Time</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NVMe enables one server to replace four legacy servers


Intel is a sponsor and member of the BenchmarkXPRT Development Community, and was the major developer of the XPRT family of benchmarks. Principled Technologies is the publisher of the XPRT family of benchmarks. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases.
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.

<table>
<thead>
<tr>
<th></th>
<th>NVMe SSDs</th>
<th>SATA SSDs</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Database Performance Queries/Hr Across All Instances</td>
<td>420 QPH</td>
<td>203 QPH</td>
<td>2.63 X</td>
</tr>
<tr>
<td>Single Database Performance Queries/Hr Across One Instance</td>
<td>39 QPH</td>
<td>22 QPH</td>
<td>1.31 X</td>
</tr>
<tr>
<td>Time to Answer Average Query Time</td>
<td>27.9 min</td>
<td>38.0 min</td>
<td>10 MINUTES</td>
</tr>
</tbody>
</table>

Captive Storage is a huge problem

NVMe enables one server to replace four legacy servers


Intel is a sponsor and member of the BenchmarkXPRT Development Community, and was the major developer of the XPRT family of benchmarks. Principled Technologies is the publisher of the XPRT family of benchmarks. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases.
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

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>
</tbody>
</table>
| 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

---

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>Specification</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

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

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

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
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
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 (60 enclosures)</td>
</tr>
<tr>
<td>Intel 3D XPoint™</td>
<td>No</td>
<td>Yes 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