NVMe® Technology in Cloud Applications

Sponsored by NVM Express, the owner of NVMe®, NVMe-oF™ and NVMe-MI™ standards
Moderator

Mark Carlson

KIOXIA
Agenda

• NVMe Technology at Scale – Lee Prewitt
• NVMe Technology and Flash SSDs in Cloud Applications – Kamaljit Singh
• NVIDIA NVMe® Technology in the Cloud – John Kim
• Deploying NVMe Flash at Facebook – A Journey – Wei Zhang
NVMe® Technology at Scale
– or “Oh the places your data will go!”

Lee Prewitt, Principle Program Manager Lead, Microsoft
Microsoft mission

Empower every person and every organization on the planet to achieve more
Datacenter is about 40’ long and 12’ wide, about the size of a shipping container.

12 racks of 864 standard Microsoft datacenter servers with FPGA acceleration and 27.6 petabytes of disk.

It took less than 90 days to deploy from factory ship to operation.

100% locally-produced, renewable electricity from solar, off-shore, tide, and wave energy.

They expect it to run for 5 years.

PROJECT NATICK
Server component failure rates were 1/8 that of a dry land data center.

Pulled up after 2 years in 117 feet of water.
Remote Debuggability

• Telemetry Command
• Device Self Test Command
• Error Injection
• Cooperative Error Recovery
• Out of Band debug via SMBus
• No Vendor Unique commands or tools

Reduces Cost!
# HDDs Versus SSDs – A Quick Comparison

<table>
<thead>
<tr>
<th>HDDs</th>
<th>SSDs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros</strong></td>
<td><strong>Pros</strong></td>
</tr>
<tr>
<td>• Cost</td>
<td>• Everything Else</td>
</tr>
<tr>
<td><strong>Cons</strong></td>
<td><strong>Cons</strong></td>
</tr>
<tr>
<td>• Everything Else</td>
<td>• Cost</td>
</tr>
</tbody>
</table>
Zoned Namespaces (ZNS)

- Allows for radical reduction in on device DRAM
  - By as much as 90%
- Can use minimal overprovisioning
  - As low as 1%
- Enforcement of large sequential writes reduces WAF
  - Allows for use of QLC NAND

Reduces Cost!
NVMe® Cloud SSD Specification

- OCP work that builds on NVMe
- Over 433 Requirement IDs covering 70+ pages
- Allows for a common firmware base
- Benefits system makers and SSD vendors
- Enables broad collaboration between hyperscale and industry

Reduces Cost!
NVMe® Technology and Flash SSDs in Cloud Applications

Kamaljit Singh, Technologist, Western Digital Technologies
Data Center SSDs: Key Trends & Transitions

- Continuing strength in Cloud deployments driving PB growth this year
  - PB CAGR ’19–’23: 40.6%
- Expansion of NVMe® technology, driven primarily by Cloud customers and spec standardizations
  - Performance, Mainstream (value), and Capacity segments emerge for NVMe SSDs expected to displace SATA and Dual Ported SAS in servers/storage at higher rates
- TLC to remain primary NAND for performance consistency and endurance
  - ZNS accelerates transition in 2022 as a QLC enabler. Continued development to enable QLC in very read intensive workloads including content delivery, streaming services and read-intensive AI;
- U.2/U.3 are going to be dominant FFs
  - Expect EDSFF (E1.L) shipments in 2020; M.2 transitioning to E1.S in 2021
- NVMe-oF™ specification can deliver latencies on par with NVMe SSDs inside servers
  - NVMe-oF attached SSDs can be shared amongst many application servers resulting in higher utilization and lower TCO
World Wide Cloud Flash TAM continues to grow at a ~43% CAGR.

NVMe® standard will be the interface of choice for majority of the deployments.

Increasingly dynamic workloads driving next-generation data infrastructures

Composable Disaggregated Infrastructure TAM.

Source: IDC Worldwide Composable/Disaggregated Infrastructure Forecast, August 2018.

* 2020 and forward: WDC long range TAM – September 20
NVMe® Technology Differentiation

Business benefits

- **Low Latency**
  - 99.999% Latency
  - QoS driven applications
- **Zone Named Spaces**
  - Lower Cost ($/GB) / Capacity Gain
  - Improves Quality of Service & Performance
  - Leveraged Software Stack & Standardized Protocol
- **High-Density**
  - ESSD Capacities moving up to 4TB and above, expected increases with QLC NAND adoption
  - Expanding use cases to include cooler storage

- **Traditional applications:** average latency of SSDs is a performance criterion for making purchase decisions
- **QoS-driven applications** (e.g. Web apps): Five 9’s latency is critical
- **ZNS** reduces write amplification and enables QLC adoption
- **Eliminates GC**, reduces long tail latencies/consistency
- **50%** (or higher) average latency improvement
Ultrastar® SN640: a case study

Capabilities:
- TLC based NVMe® SSD
- Capacities: 0.96 – 7.68 TB
- Consistent Performance
  - 75R/25W random mixed I/O
  - Coefficient of variance <1, (benefits real-world applications)
  - QoS latency of 5 nines, at higher queue depths, (benefits large-scale workloads w/ many concurrent users)
  - Latency similar or better by 2x than most drives. makes it more cost-effective than the competition.

Optimized for:
- Web Search engines
  - Cost-effective fast storage and caching layers for warm data
- AI-enabled search and contextual analytics
  - Composable infrastructure with NVMe-oF™ Flash storage for mixed AI workloads, like training and inference
- Data warehouses
  - Read-only databases with minimal writes; execute ad-hoc queries for analytics
- Data Hubs
  - Data stores serving various application domains involving mixed I/Os with higher % reads, including Big data, Fast data, AI/ML, backups/object storage
- Hybrid Flash/HDD back-up storage for on-demand access to data
  - User logs and models
- AI/Deep learning for image/video analytics
  - NLP for text analytics
SN640: Performance Highlights

**QoS of Random Mixed**

- Similar or 2x better QoS latency of 5 nines than similar drives at highest capacity
- Consistent QoS latency across the product family
NVMe® over Fabrics Specification

Business benefits

**LOW LATENCY**
NVMe-oF™ technology can deliver latencies on par with NVMe® SSDs inside servers

**HIGH-PERFORMANCE SHARING**
NVMe-oF attached SSDs can be shared amongst many of application servers resulting in higher utilization and lower TCO

**DATA ACCESS & MOBILITY**
Fabric-attached data enables Cloud-like dynamic access and workload mobility
NVIDIA NVMe® Technology in the Cloud

John F. Kim, Director of Marketing for Storage Networking, NVIDIA
NVIDIA NVMe® Technology in the Cloud

Private Cloud– EGX/HGX/DGX
- Often use internal NVMe® SSDs
- Larger systems use GPUDirect for faster storage access

NVIDIA DGX A100 Storage
- 2 M.2 NVMe SSDs for OS
- 4 U.2 NVMe SSDs for compute
- 4x 200Gb/s ports for shared storage

Public: NVIDIA Quadro vWS
- Avail. in GCP and AWS
- GPU workstation anywhere--for engineering and creative apps

Public: GPU Compute Instances
- AWS EC2 (P3 & G4 instances)
- GCP GPUs on Compute Engine
- IBM, Alibaba, Azure, Oracle
NVIDIA DGX A100 Design
Deploying NVMe® Flash at Facebook – A Journey

Wei Zhang, Software Engineer, Facebook
The Beginning – Flash Add-In-Card

• Facebook started flash journey in 2010

• DB apps required higher IOPS and lower latency
  • HDD storage cannot meet the requirements cost-effectively

• Flash AICs
  • Pros: Superb performance
  • Cons: proprietary technology (hw + sw), super expensive
The Cheaper Alternative – SATA SSD

- More applications were moving to flash
- SAS/SATA SSDs were the mainstream
  - NVMe® specification was still in embryonic state
- Leverage OCP Knox designed for HDD
  - Pros: cheaper flash, standard hw + sw
  - Cons: perf bottleneck

x4 SAS
~400K IOPS

OCP Knox

SATA
Flash Performance Unleashed – NVMe® JBOF

- Flash applications are performance demanding
- The Lightning JBOF
  - NVMe® flash pooling
  - Allows optimal compute to storage ratio
  - End-to-end PCIe connection
- Complex system design due to technology limitations
Technology Matured - NVMe® Flash Server

- Technology advancement has allowed us to design integrate flash server
- CPU
  - Abundant PCIe lanes
  - Root port PCIe error containment
- SSD
  - Density increased
  - EDSFF E.1S form factor
The Benefit of NVMe® Flash

NVMe® technology enables a cost-effective way to leverage NAND flash performance at Facebook scale!
Driving and Working with Industry

EDSFF E1.S Form Factor

- Performance scaling
- Better thermal characteristics
- Hot plug support

Cloud SSD Spec

NVMe Cloud SSD Specification

Version 1.0 (03182020)
Panel Discussion