Challenges & Opportunities with Hyper-Scale Boot Drives

Hyper-Scale Boot Drives

Karthik Shivaram, Storage Engineer, Facebook Inc.
Introduction

Hyper-Scale Boot Options

SSD Boot Drive: Challenges

SSD Boot Drive: Solution
Facebook Infrastructure’s goal is to build the most innovative and highly efficient Data Centers on earth
Facebook Data Center Locations

Boot Drives are deployed all over the world!
Background

- Data Center Server’s typically contain two forms SSDs:
  - Data Drives
    - Use Case: Generally used as a data-store e.g., Database, Cache
    - Capacity: 2 to 8TB
    - Power: 8.5W to 20W
    - Form-Factor: E1.S, M.2 (22x110), U.2
    - Power Loss Protection: Required
  - Boot Drives
    - Use Case: Host OS, Logs, Scratchpad
    - Capacity: <512GB
    - Power: <5W
    - Form-Factors: M.2 (22x80)
    - Power Loss Protection: Not Supported

- Hyper-Scaler’s deploy boot drives, but the requirements for these boot drives are not public or understood in the industry
Where do Hyper-Scalers use Boot Drives?

Network Switches

Compute Servers

Storage Nodes
  - JBOD
  - JBOF
Hyper-Scale Boot Drive Options
Boot Options

- HDD Boot Drive
- Client SSD Boot Drive
- Network Boot
HDD Boot Drive

- Contains mechanical components
  - Reduces reliability which increases operational complexity
- Poor random performance
- High active power (>5W)
- Significantly larger in capacity than what’s needed
- No side-band access (I2C)
- Physically occupies more space which doesn’t fit in high density designs

HDD as a Boot Drive is undesirable in high density server designs
Network Boot

- Consumes critical network bandwidth
- Reduces reliability of the system and rack
  - Blast radius on a single failure can be very high
- Increases boot time
- Increases I/O latency
- Disaster recovery can be very challenging

• Network Boot is complex and challenging to implement @Scale
  • Introduces many risks into Data Center Infrastructure Reliability
Client SSD Boot Drive

- High random performance
- Capable of supporting Hyper-Scaler needs:
  - Consumes lower power
  - Supports security features such as Secure Boot
  - Increased reliability as there are no moving parts
  - Reduces blast radius due to being local to the system
- Physically small (M.2: 2280/2230)
- Widely available

- Client SSDs are more aligned to be used as a Boot Drive in a Hyper-Scaler environment, compared to other options
- But it comes with some challenges ...
SSD Boot Drive: Challenges
Capacity Trends

Exabytes, Average Capacity

Source: TRENDFOCUS

- Data Drive capacity keeps increasing
- Boot Drive capacity needs 512GB or smaller
- Increasing capacity = Increasing expense
Boot Drive Utilization

- Typical Disk Utilization Remains low (<50%)
- Mostly of capacity used by user-space applications.
  - OS + Swap doesn’t occupy a lot of footprint

Hyper-Scaler desire is to have support for low-capacity Boot Drives
Differences in Client and Hyper-Scale Usage

- Client SSDs are designed typically around a laptop usage model
- Client vs Hyper-Scale feature comparison:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Client</th>
<th>Hyper-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle Time</td>
<td>Plenty</td>
<td>Almost none</td>
</tr>
<tr>
<td>Power Saving Features</td>
<td>Required</td>
<td>Not Required</td>
</tr>
<tr>
<td>On-board PLP</td>
<td>Not Required</td>
<td>Not Required</td>
</tr>
<tr>
<td>Performance</td>
<td>Fresh out-of-box</td>
<td>Sustained</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Not important</td>
<td>Very important</td>
</tr>
<tr>
<td>Endurance</td>
<td>Low</td>
<td>Very high</td>
</tr>
</tbody>
</table>

Client and Hyper-Scale SSDs have different requirements
An example Hyper-Scale Boot I/O Profile

- Boot Drives experiences constant traffic with no idle time
- TRIM rate on Boot Drive is very high
  - Latency stalls due to TRIM are not desirable
Boot Drive I/O Traffic Breakdown

- Majority of the traffic is random in nature
  - Increases Background Activity
  - Increases Write Amplification
- Majority of the traffic is low queue-depth:
  - Services are sensitive to drive latency
  - Latency stalls lead to poor user experience

- Majority of traffic is random in nature
- Workloads have low queue depth
- User experience is sensitive to latency
Managing variable performant devices @Scale

- Performance methodology for Boot Drives is not clear
  - No minimum bar (or performance target) defined
- No open benchmarks for Boot Drives
  - Leads to huge drive-to-drive performance variation

Hyper-Scalers must deal with huge variation in drive performance due to lack of industry standards
Hyper-Scale Endurance & Monitoring Requirements

- Monitoring at scale is important
  - Boot Drives are deployed all over the world
- Monitoring helps predict & detect failing drives
- Boot SSDs need higher Endurance to prevent early wear out
  - Reliability is extremely important as repair at-scale is extremely challenging

Hyper-Scaler Boot SSD require higher endurance and enhanced monitoring
Summary of Challenges with Boot SSDs

- Capacity of SSDs are increasing
  - Boot Drive capacity needs remains constant
- Client SSDs are designed with a focus on Client use-cases
- Hyper-Scalers require higher endurance and enhanced monitoring compared to Client SSDs
- Hyper-Scalers have confidential Boot SSD specifications which doesn’t encourage industry collaboration
SSD Boot Drive: Solution

How do we solve these challenges?
Path to solving the problem...

Facebook & Google are collaborating and combining requirements to create a OCP Hyper-Scale Boot SSD Specification.
Benefits of an Open Boot Drive Specification

- Facebook & Google have merged their SSD boot drive requirements into a single document enabling the following benefits:
  - Allows the market to understand what features Hyper-Scalers need to manage an SSD at-scale.
  - Allows the market to understand and use the SSD’s that Hyper-Scalers are using.
  - Reduces SSD market fragmentation.
  - Enables open-source tools like NVMe-CLI to manage & monitor SSDs at-scale.
  - Allows 3rd parties to create test-suites which simplifies the drive qualification process.

Opening requirements helps increase industry collaboration and reduces SSD market fragmentation.
Key Focus Sections of the Specifications

- Specifies requirements needed to build & manage a Hyper-Scale Boot SSD
- This includes requirements around:
  - NVM Express
  - PCI Express
  - SMART Logs
  - Reliability
  - Thermal
  - Power
  - Performance
  - Security
  - Side-Band/SMBus
  - Monitoring & Tooling

Everything needed to build a Hyper-Scale Boot SSD!
Conclusion: Roadmap to a brighter future

Today

Lack of Industry Standards for Hyper-Scale Boot Drives

• SSD Boot Drives are customized but there is no Industry Standard to capture all the requirements.

Future

OCP Hyper-Scale Boot Drive Specification

• Benefits system makers and SSD providers.
• Enables additional collaboration between Hyper-Scaler’s and industry.
Thank you!
Please take a moment to rate this session.

Your feedback is important to us.