Does Gen6x4 Make Sense for SSDs Claiming 25W Due to Form Factor Recommendations?

Steven Wells – Fellow
Suresh Rajgopal - Distinguished Engineer
Trends to Gen6 suggest > 25W

- Power efficiency trends at each PCIe® Generation are not keeping up with the ~2x speed of each generation
  - Gen5 20-25W → Gen6 > 25W
- EDSFF informatively suggested E1.S and E3.S 1T target a maximum of 25W
- Are there options to benefit from Gen6 without moving to >25W FF such as E3.S 2T, E1.L or E3.L.
What Options do we have?

- Abandon harder to cool form factors overall as we move to Gen6
- Keep the form factors but limit to a maximum 25W power state
- Higher operating temperatures and/or higher airflows with higher power states
- Other “out of box” thinking

This presentation will hopefully offer insights to how to rethink power and thermals mitigations at both the Host and SSD while remaining aligned to NVMe™.
Summary of NVMe™ SSD Standards for Power and Thermals

**Power**
- Drive reports a table of possible active power states
  - PS0 = highest power state
  - PS1-n = lower power states
- “Host may dynamically modify the power states” using Features Command, optionally persistent
- PCIe® slot power limit needs to be honored

**Thermals**
- Composite Temperature
- Host Controlled Thermal Management (HCTM)
- Set feature offers
  - TMT1 – temperature (K) to start throttling
  - TMT2 – “heavy throttling”
- Drive can select VU thermal actions or can transition power states
- Warning and critical thermal notifications.
  (WCTEMP/CCTEMP)
Other Relevant Standards Impacting Power/Thermals

- EDSFF – the well known 25W and 40W “recommended maximum sustained power”

- OCP – Sets a paradigm that thermal throttling is only for failure conditions
OCP Power Measurement Guidance

- Sub 100uS peaks are covered by filter capacitors

- Peak Power (100us window) is beyond what on-board capacitors can filter - required on platform regulators to track noise; IR drop and brown-out conditions

- Max Average is typically considered thermally relevant (1 second or greater)
How fast can temp change in E1.S and E3.S?

- Composite temperature change during operation*
  - E1.S 15mm - smaller FF
    - 0.5-0.75 degrees per sec from Idle
  - E3.S 2T - larger FF
    - Less than 0.25 degrees per sec from Idle

Depending on FF and system airflow capabilities, the temperature gradient is between 0.25-1C/s
Max Average Power from a real workload

ML Perf Storage (UNET 3D) – Medical Image Segmentation

Real workloads are bursty

8W

2W

Steady State Sequential Write
SSD Internal Power and Thermal Throttling – One Possible Conceptual Implementation

1. **Host Cmd Ingress**
   - **Cmd Decode**
   - **Schedule “Back End”**
     - GC
     - Refresh
   - **NAND Command Ingress**

2. **Temperature-based derating (d)**
   - CTemp
   - d

3. **Multiplier (d x PS)**
   - NVMe™ Power State Target (PS)

4. **Power Needed > Max Power Allowed?**
   - **Power Needed**
   - **Max Power Allowed**
   - **Y**
   - **N**

5. **Stalling induces latencies**
   - **Stall**
   - **Execute**

**Composite SSD temp**

- 40W, 25W, 20W, 15W, etc.
Latency Impacts to Throttling

Mixed workload read latencies during 2 NVMe™ power states

- Cases with throttling will have extended NVMe™ command completion latencies (avg and/or tails)
- The goal is to minimize residency in throttling
Workload Managed by Host Initiated Power States

Rate of host polling and power state policy is host dependent

~60% throttling residency

Conceptual Example

Oscillating Power States Creates Oscillating Latency Distributions
HCTM Throttling

Example Latency Distributions across power states

HCTM enables stable latency distribution on stable workloads
## Three Options for Power Management

<table>
<thead>
<tr>
<th>Power State Aligned with FF recommendations</th>
<th>Host Managed Dynamic Drive Power States</th>
<th>Drive Managed Progressive Thermal Throttling (HCTM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host or device manufacture sets default power state. Unmodified over device life.</td>
<td>Host periodically polls CTEMP and changes drive power state. Polling frequency important.</td>
<td>Host configures TMT1/2</td>
</tr>
<tr>
<td>Greater throttling residency</td>
<td>Reduced throttling residency</td>
<td>Drive internally polls CTEMP and adjusts throttling progressively.</td>
</tr>
<tr>
<td>Throttling when not always thermally necessary.</td>
<td>Throttle when thermally necessary.</td>
<td>Reduced throttling residency.</td>
</tr>
<tr>
<td>Full Gen6x4 burst not available.</td>
<td>Latency profiles shift for each power state command issued.</td>
<td>Consistent latency profile to a steady state workload.</td>
</tr>
<tr>
<td></td>
<td>Enables Gen6x4 Burst Capability.</td>
<td>Reduced thermal stress on drive components.</td>
</tr>
</tbody>
</table>

Two options to enable dynamic bursting above FF limits when thermal margin exists.
Call to Action

- **Devices**
  - Respecting PCIe® Slot Power
  - Supporting NVMe™ PS0 above “thermal TDP”
  - Progressive Thermal Throttling vs emergency throttling

- **Hosts**
  - Host to determine best methods between BMC managed burst power states and/or drive managed HCTM
  - Participate in SNIA Storage Management Initiative, OCP HW management, and Linux Foundation’s OpenBMC

- **Future**
  - Standardized power efficiency metrics similar to Client’s battery life workload.
Please take a moment to rate this session.

Your feedback is important to us.