STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

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

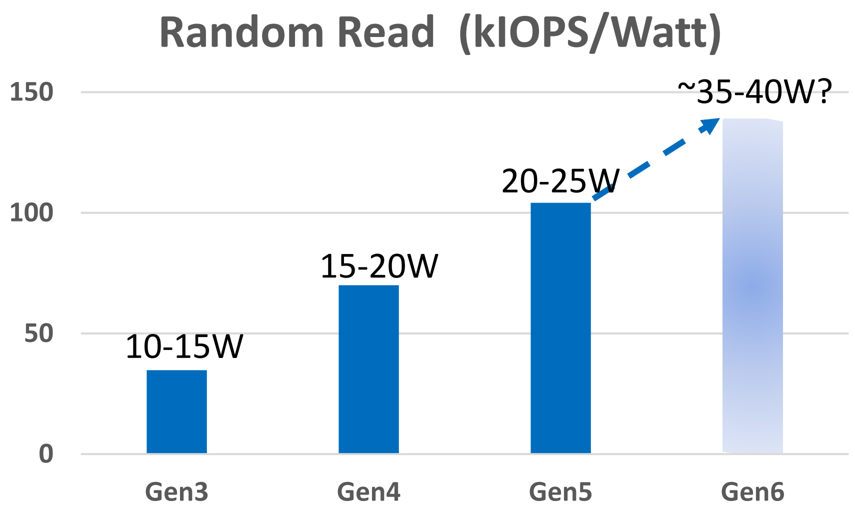
Steven Wells – Fellow

Suresh Rajgopal - Distinguished Engineer

Micron

Trends to Gen6 suggest > 25W

Power Efficiency Trends



- 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

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

PUBLISHED **SFF-TA-1008 Rev 2.0**

6. Informative: E3 Thermal Characteristics

Table 6-1 defines the recommended maximum sustained power allowed by each device variation.

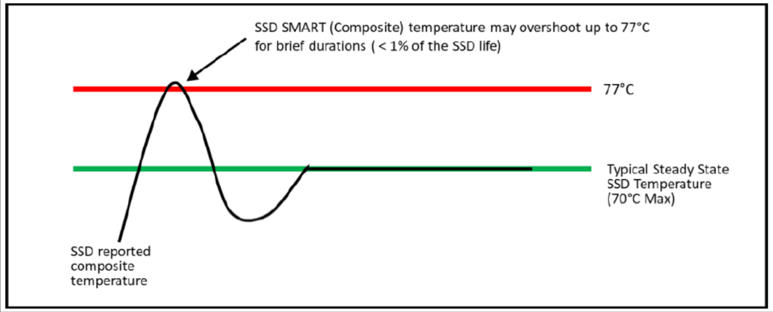
Table 6-1 Maximum Form Factor Power

Device	**E3S**	**E3L**	**E3S 2T**	**E3L 2T**
Max Power	25W	40W	40W	70W

For detailed device thermal requirements refer to SFF-TA-1023 Thermal Specification for EDSFF Devices.

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

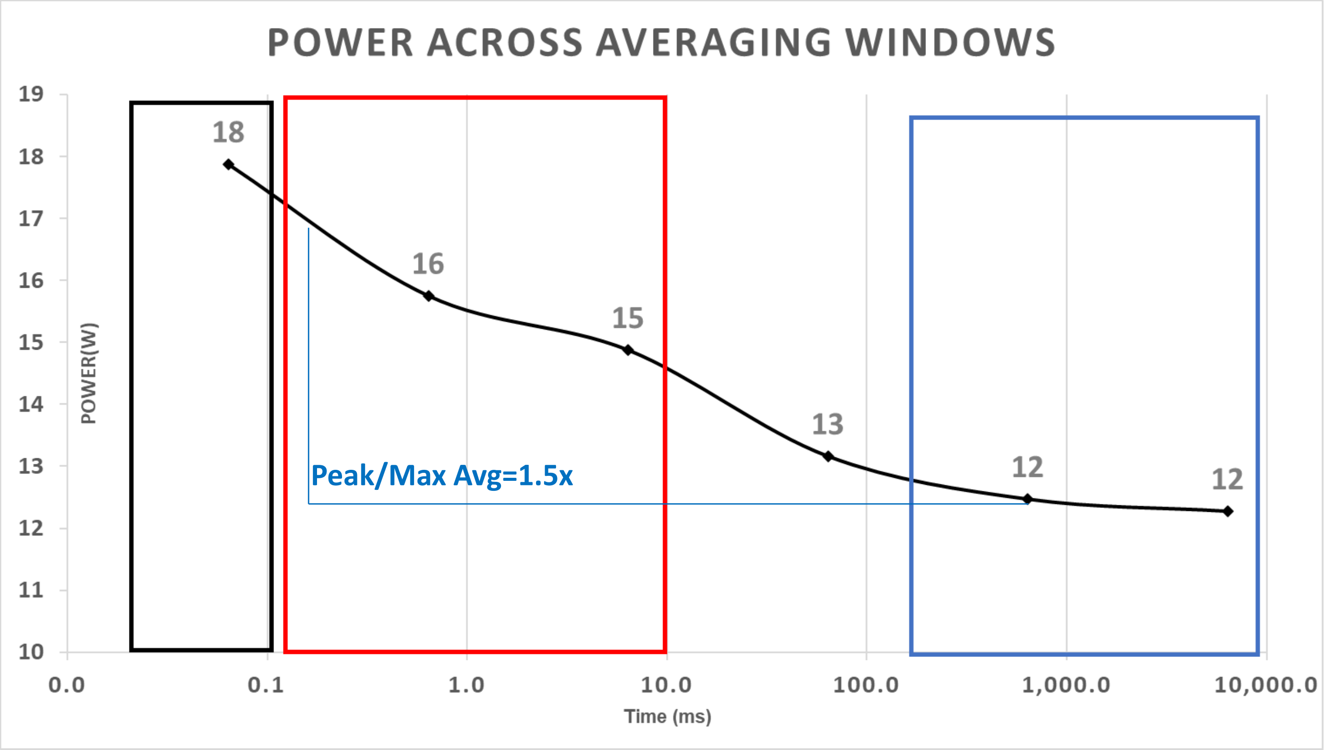
TTHROTTLE-3	Thermal throttling shall only engage under certain failure conditions such as excessive server ambient temperature or beyond the server's fan failure redundancy limit. The required behavior is illustrated below:



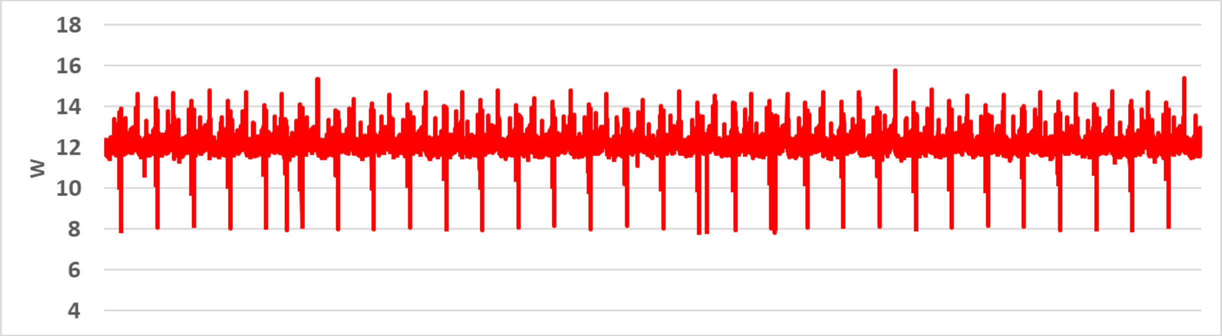
- OCP – Sets a paradigm that thermal throttling is only for failure conditions

SDC 23

OCP Power Measurement Guidance

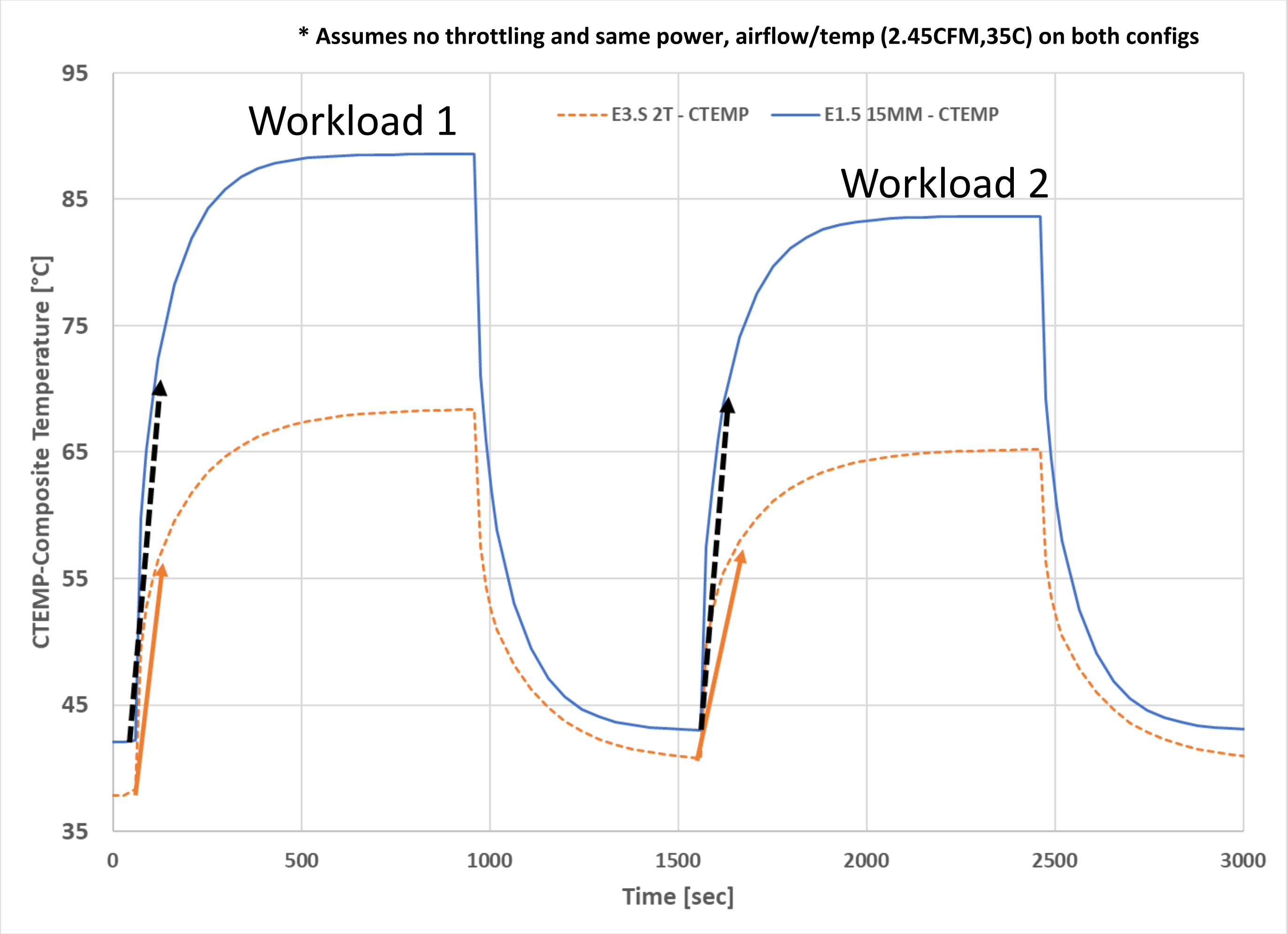


- Sub 100uS peaks are covered by filter capacitors
- Peak Power (100us window) is beyond what on- board capacitors can filter - required on platform regulators to track noise; IR drop and brown-out conditions
- Max Average is typically considered thermally relevant (1 second or greater)



How fast can temp change in E1.S and E3.S?

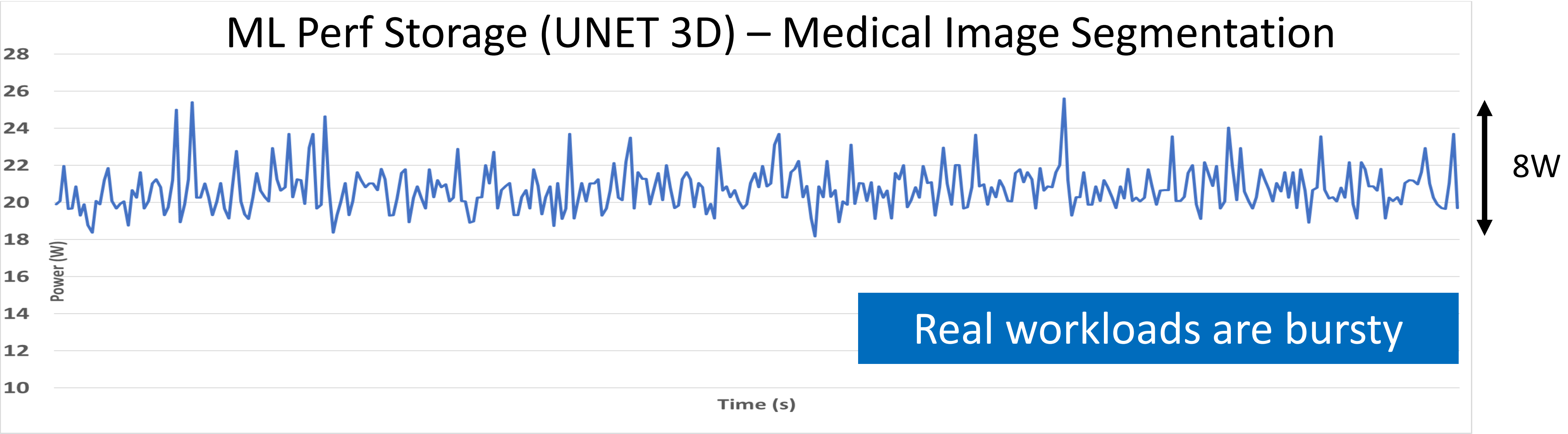
- Composite temperature change during operation*
- E1.S 15mm - smaller FF
  - **0.5-0.75 degrees per sec from Idle**
- E3.S 2T - larger FF
  - **Less than 0.25 degrees per sec from Idle**

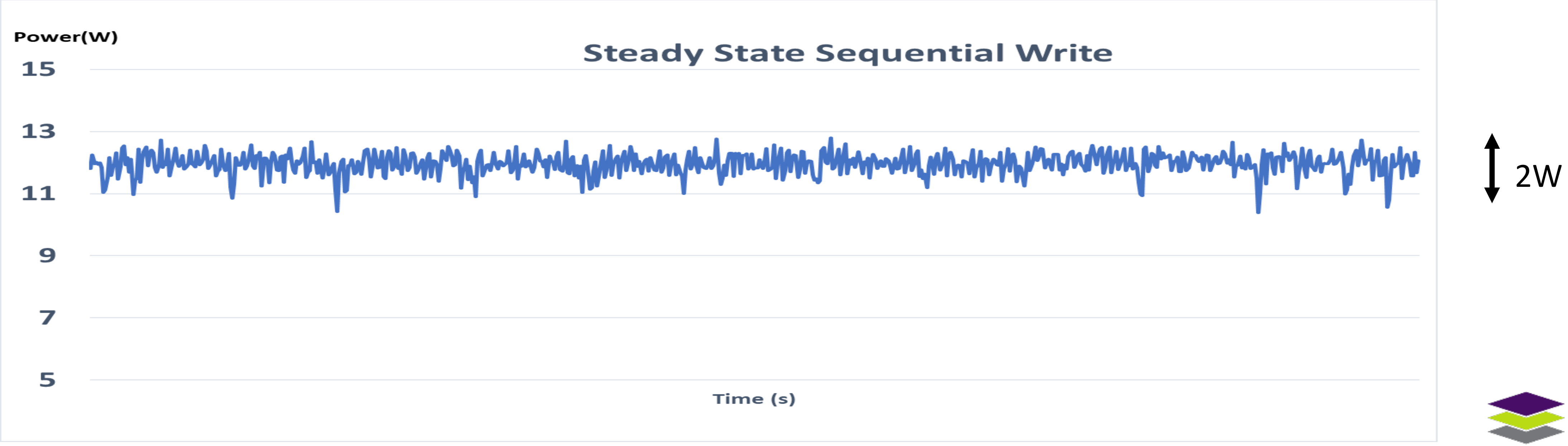


Depending on FF and system airflow capabilities, the temperature gradient is between 0.25-1C/s

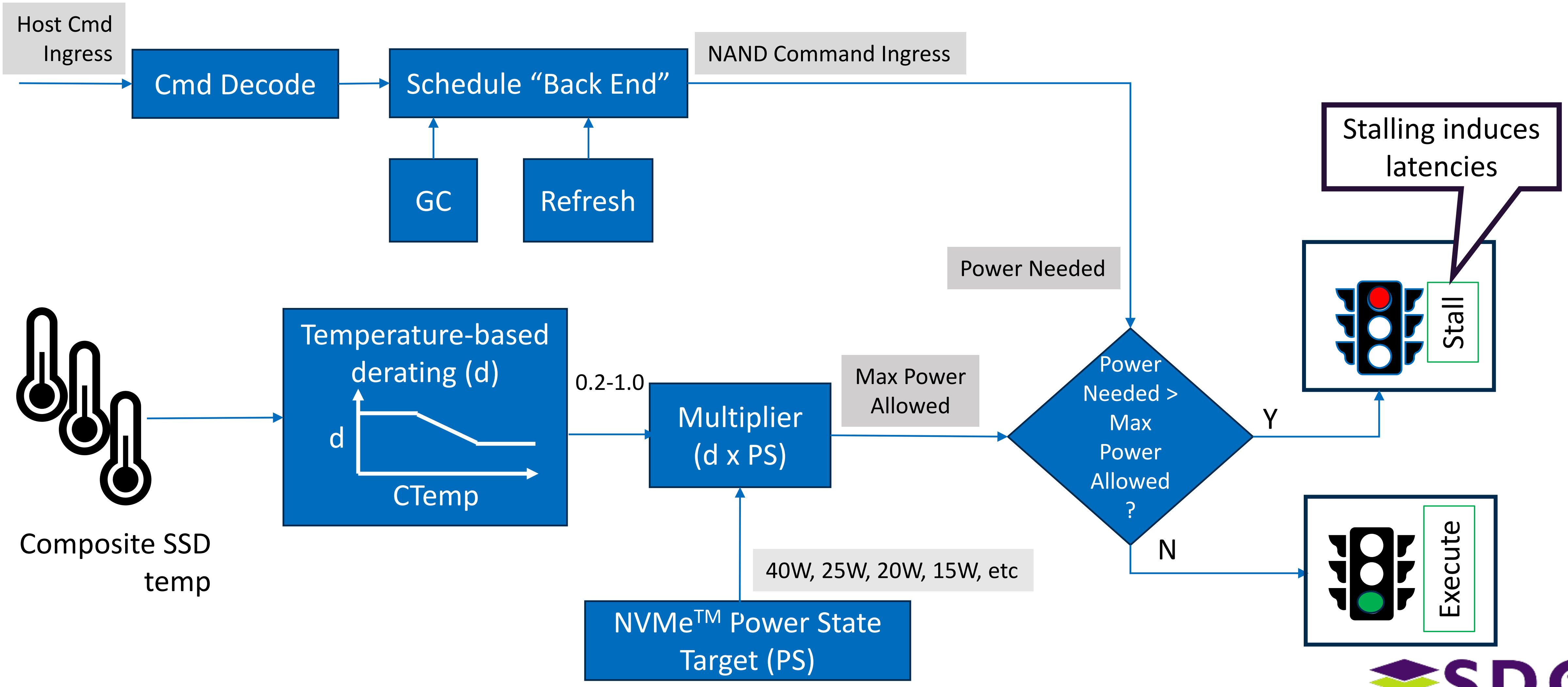
SDC 23

Max Average Power from a real workload



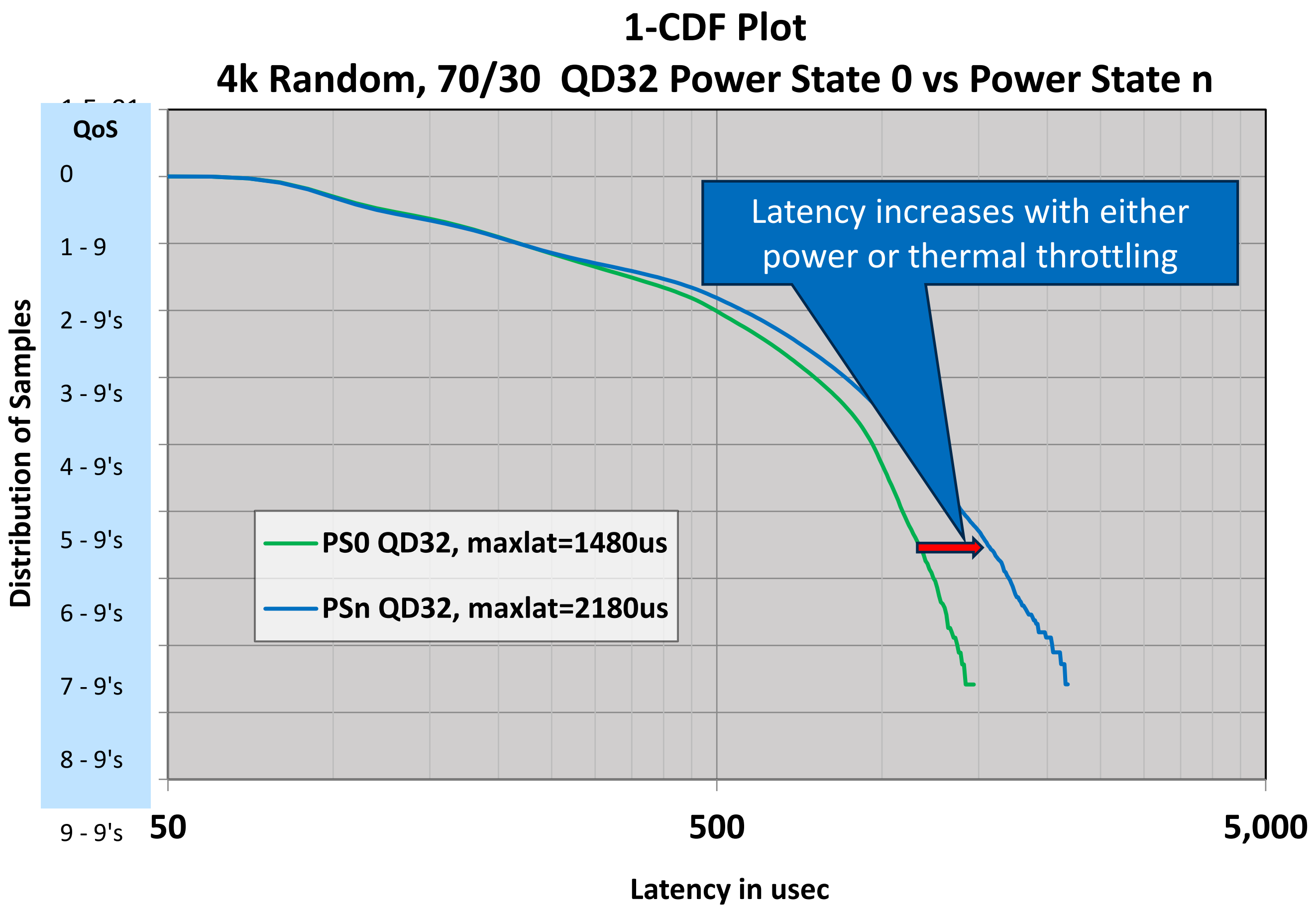


SSD Internal Power and Thermal Throttling – One Possible Conceptual Implementation



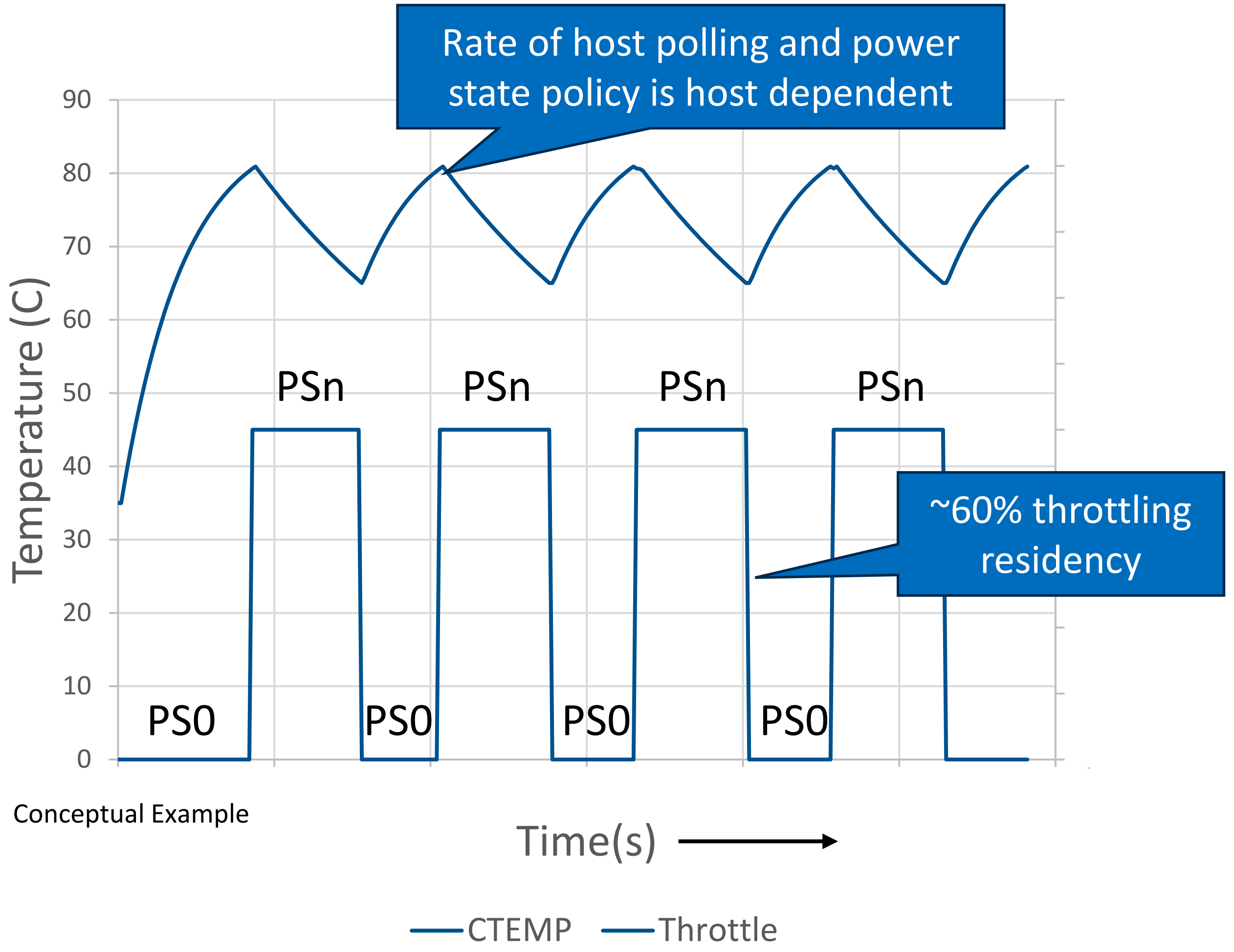
Latency Impacts to Throttling

Mixed workload read latencies during 2 NVMe™ power states



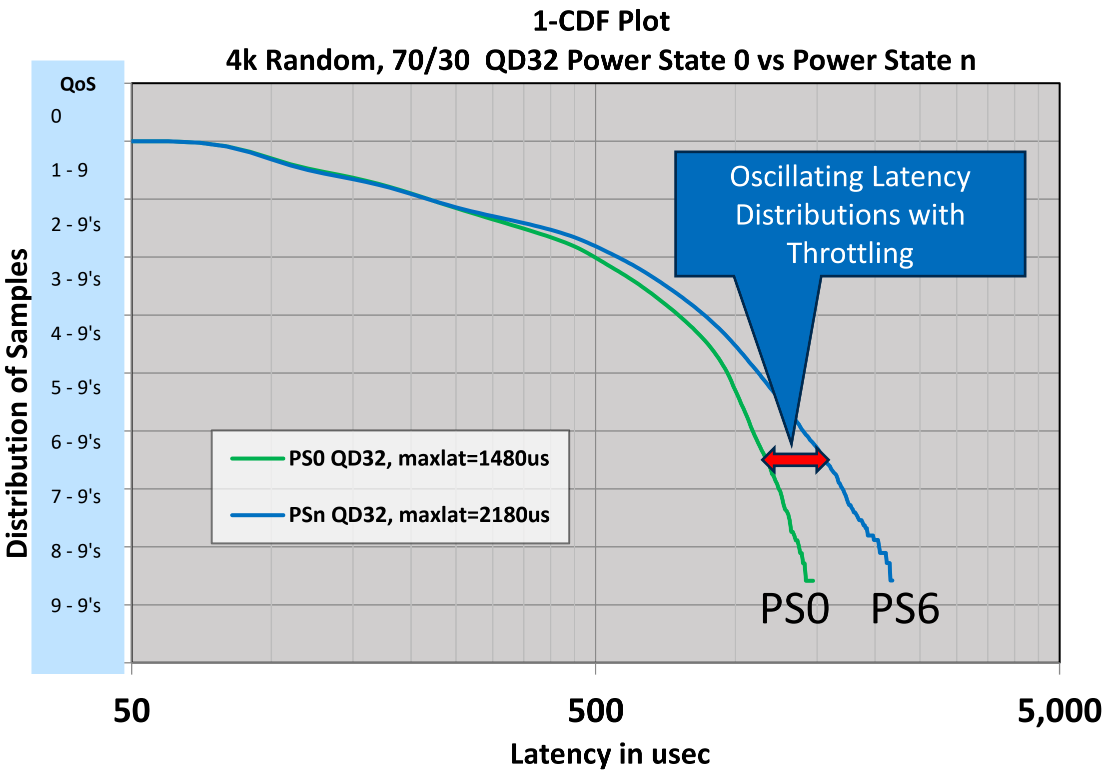
- Cases with throttling will have extended NVMe™ command completion latencies (avg and/or tails)
- The goal is to minimize residency in throttling

Workload Managed by Host Initiated Power States



Conceptual Example

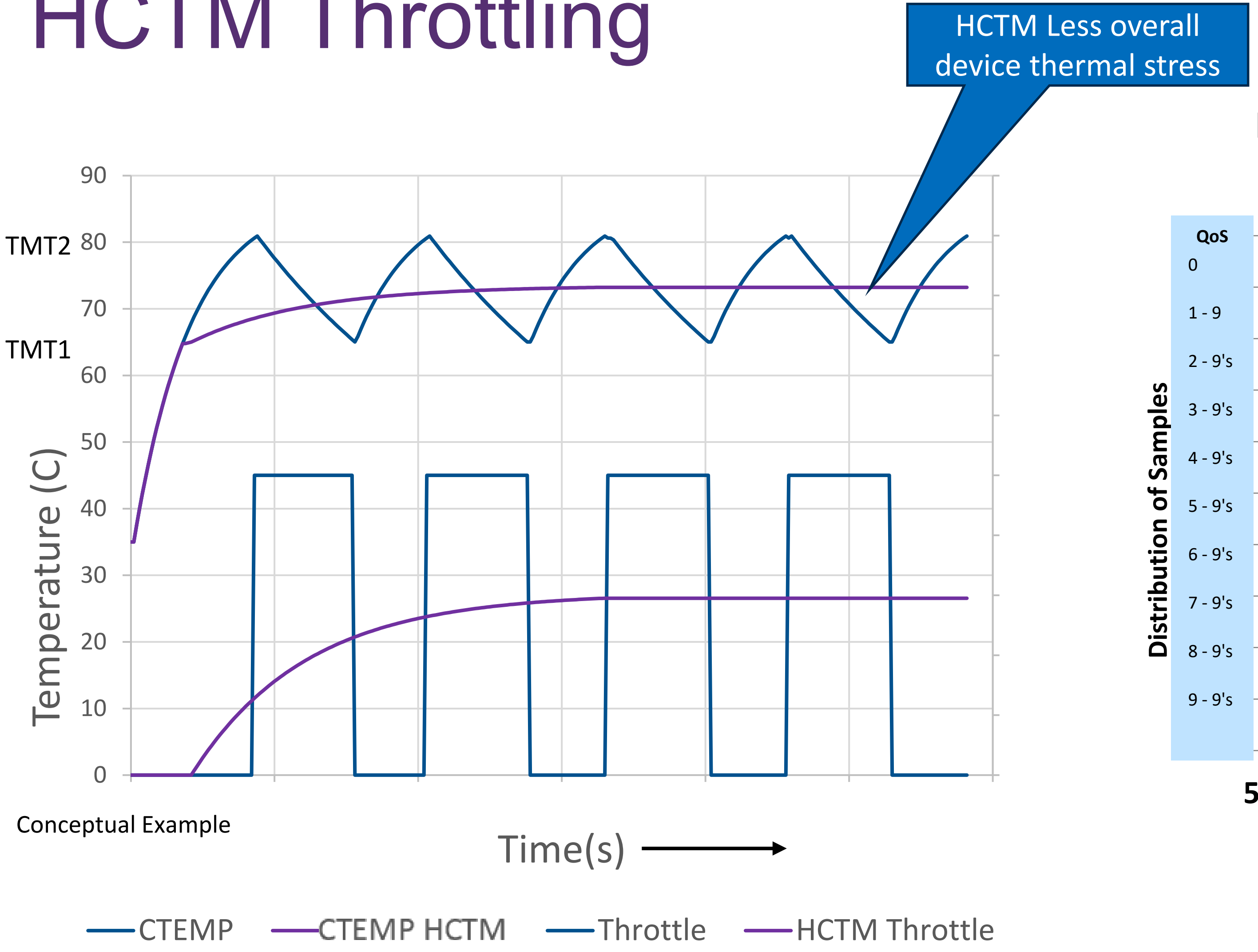
Example Latency Distributions across power states



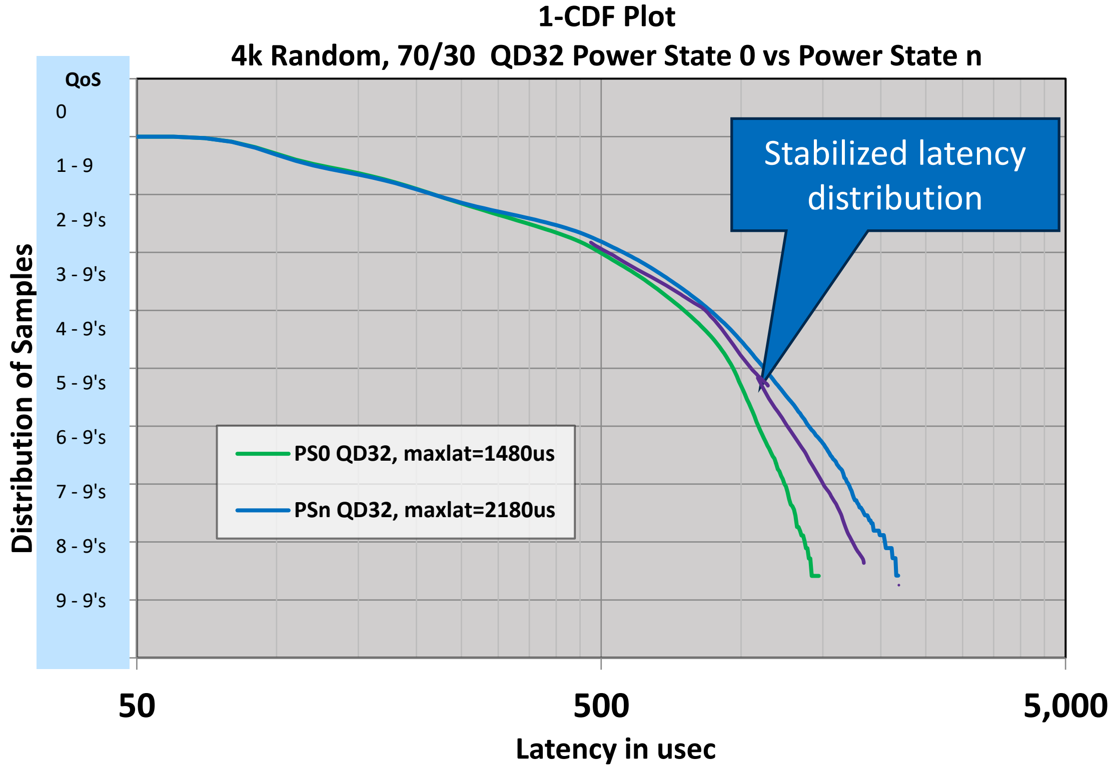
Oscillating Power States Creates Oscillating Latency Distributions

SDC 23

HCTM Throttling



Example Latency Distributions across power states



HCTM enables stable latency distribution on stable workloads

Three Options for Power Management

Power State Aligned with FF recommendations

Host or device manufacture sets default power state. Unmodified over device life.

Greater throttling residency

- Throttling when not always thermally necessary

Full Gen6x4 burst not available

Host Managed Dynamic Drive Power States

Host periodically polls CTEMP and changes drive power state

- Polling frequency important

Reduced throttling residency

- Throttle when thermally necessary

Latency profiles shift for each power state command issued

Enables Gen6x4 Burst Capability

Drive Managed Progressive Thermal Throttling (HCTM)

Host configures TMT1/2

- Drive internally polls CTEMP and adjusts throttling progressively

Reduced throttling residency

Consistent latency profile to a steady state workload

Reduced thermal stress on drive components

Enables Gen6x4 Burst Capability

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.

