NVMe HDD SpinUp and Power Management

Tim Walker

4 Feb 2021

SpinUp and Power Management Goals

- Keep NVMe HDD power properties compliant with the NVMe and PCIe specs.
- Provide means to control the device's spinup power usage by the host or chassis, as appropriate.
  - e.g. a fabric-attached chassis cannot delegate power mgmt. to a connected host.
    - Difficult to keep host aware of the chassis power state
    - Difficult to select a controlling host + provide failover
  - System should be configurable to manage power via
    - NVMe – for easy host management
    - PCIe – to manage power internal to chassis without NVMe root complex
    - Unmanaged – simplest implementation or for bench use.
- No NVMe-MI required

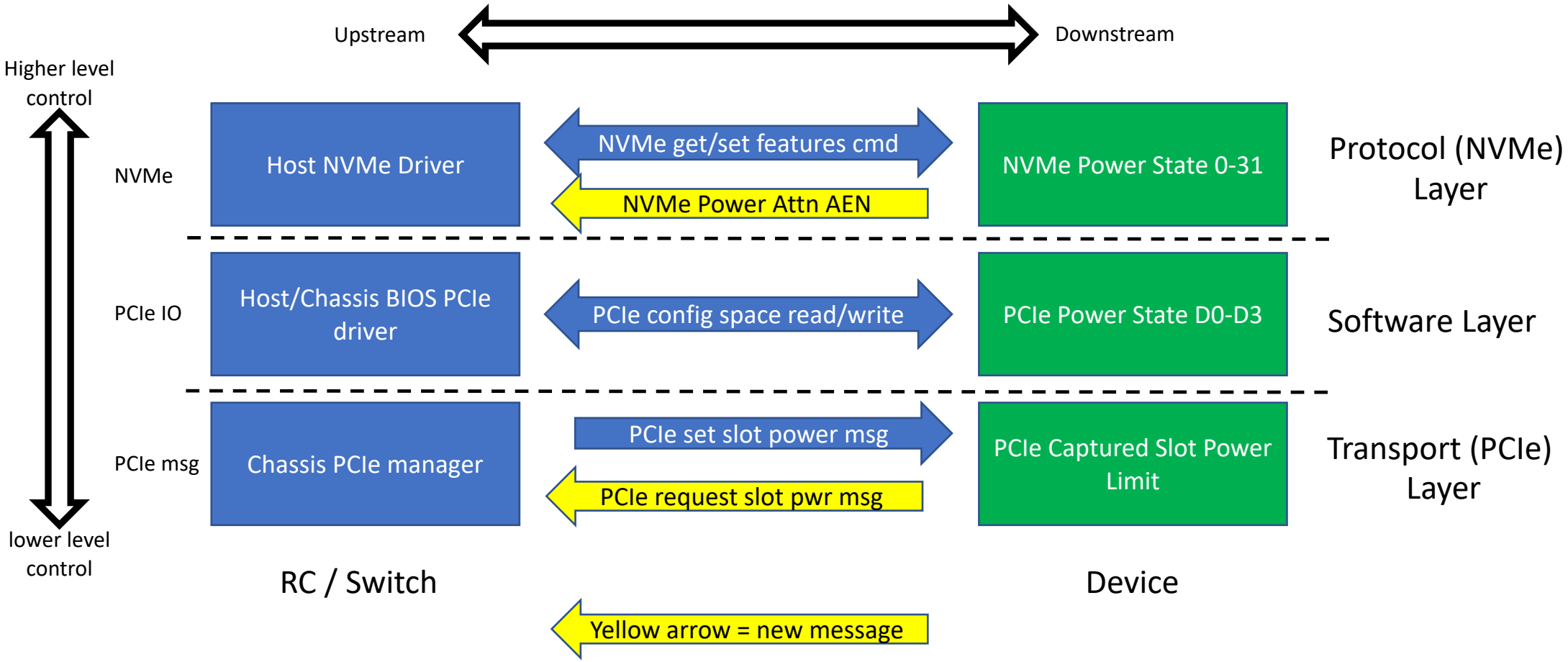
NVMe HDD Power Mgmt Overview

- There are (at least) four different controls for the power consumption of an NVMe device (excluding the PCIe link power state). From top of stack downwards:
  - NVMe power state
    - Up to 32 discrete power states
    - Each state publishes its
      - Active Power (ACTP) indicating the largest average power over a ten second window.
      - Max Power (MP) indicating the maximum power the device may draw. This should be sufficient for spin-up in spinning states.
    - Managed by NVMe commands
  - PCIe Power state
    - 4 states (D0-D3hot), only D0 is operational. (D1 & D2 not used by NVMe, D3 nonoperational)
    - Set by writing value to PCIe config register (Offset PMCAP + 4h: PMCS – PCI Power Management Control and Status)
  - PCIe Slot Power Limit
    - Hardware power limit contained in upstream port register, copied to downstream register at link init and if changed.
    - Communicated by PCIe TLPs (transport layer)
  - SFF Module Spec max power limit
    - A default maximum power prescribed by the SFF Module Spec
    - Applicable until overridden by PCIe Slot Power Limit
      - Assumed to be applicable at device boot.

Higher stack

Lower stack

Multiple Level Power Management



NVMe HDD Feature: Spinup Control

Only one mandatory change to NVMe to support HDD spinup: new NVMe feature: Spinup Control

Example Case: An NVMe HDD implements PS0-PS3

- 0=fully active
- 1=heads idle but not retracted
- 2=heads retracted
- 3=spun down

Only 0 is mandatory, others are implementation specific

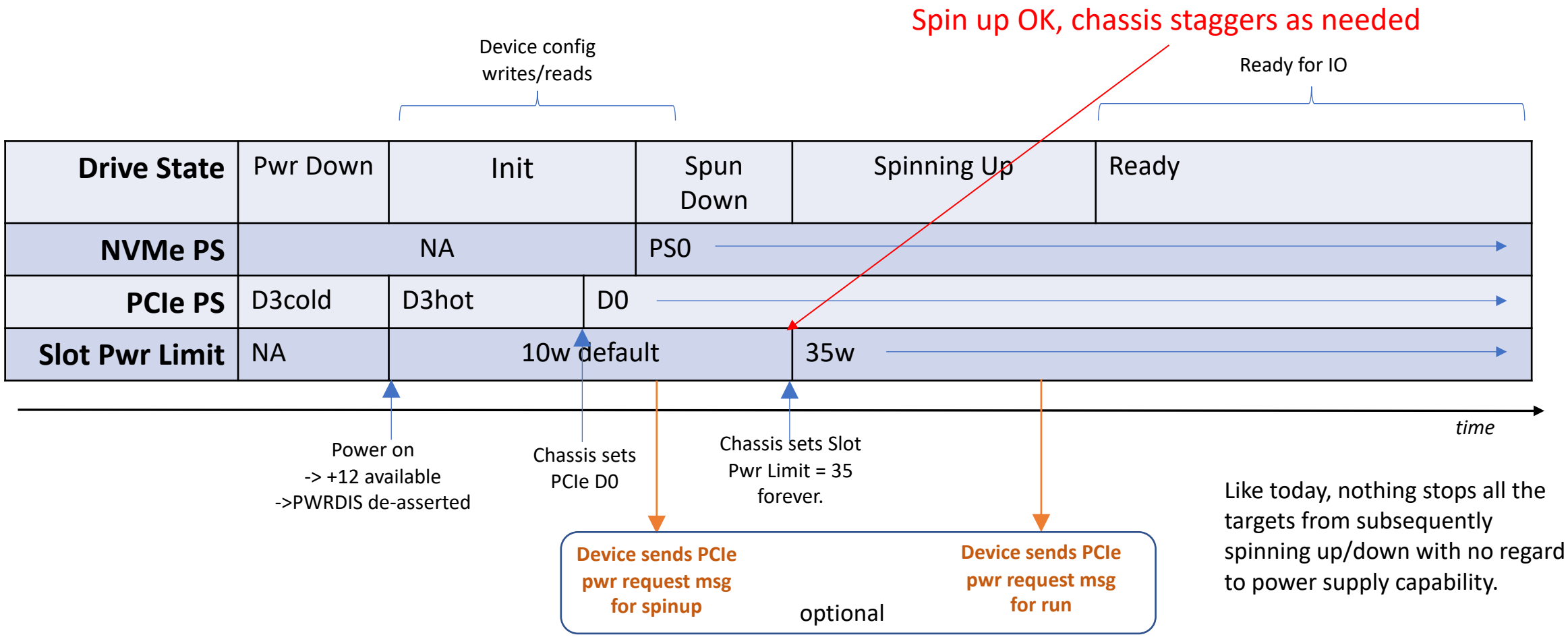
After a Power On Reset:

- if Spinup Control is ***enabled***, power up in NVMe PS3 – (a non-spinning power state)
  - defer initial spinup until a controller processes a Set Features command (Power Management) that requests transition to an operational power state. E.g. power state 2, 1, or 0
- if Spinup Control is ***disabled***, power up in NVMe PS0 (operational)
  - begin initial spinup routine at NVM Subsystem Reset.
- Use Set Features (Power Management) to transition between operational and non-operational power states.

Supporting “enabled” is optional.

NVMe-HDD Initial Power On – Dumb Chassis

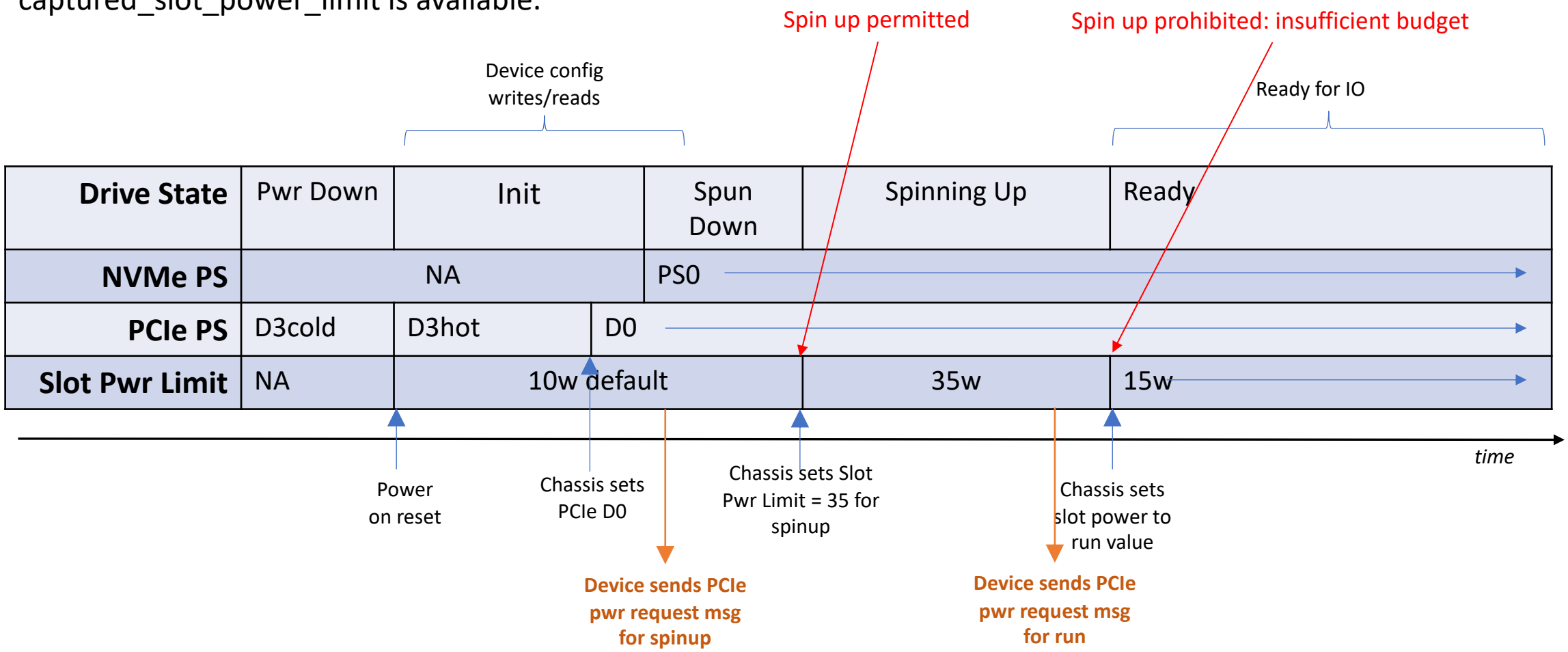
Chassis does not manage power after initial spinup
Spinup Enable = False; drive spins up after power reset when sufficient captured_slot_power_limit is available.



Like today, nothing stops all the targets from subsequently spinning up/down with no regard to power supply capability.

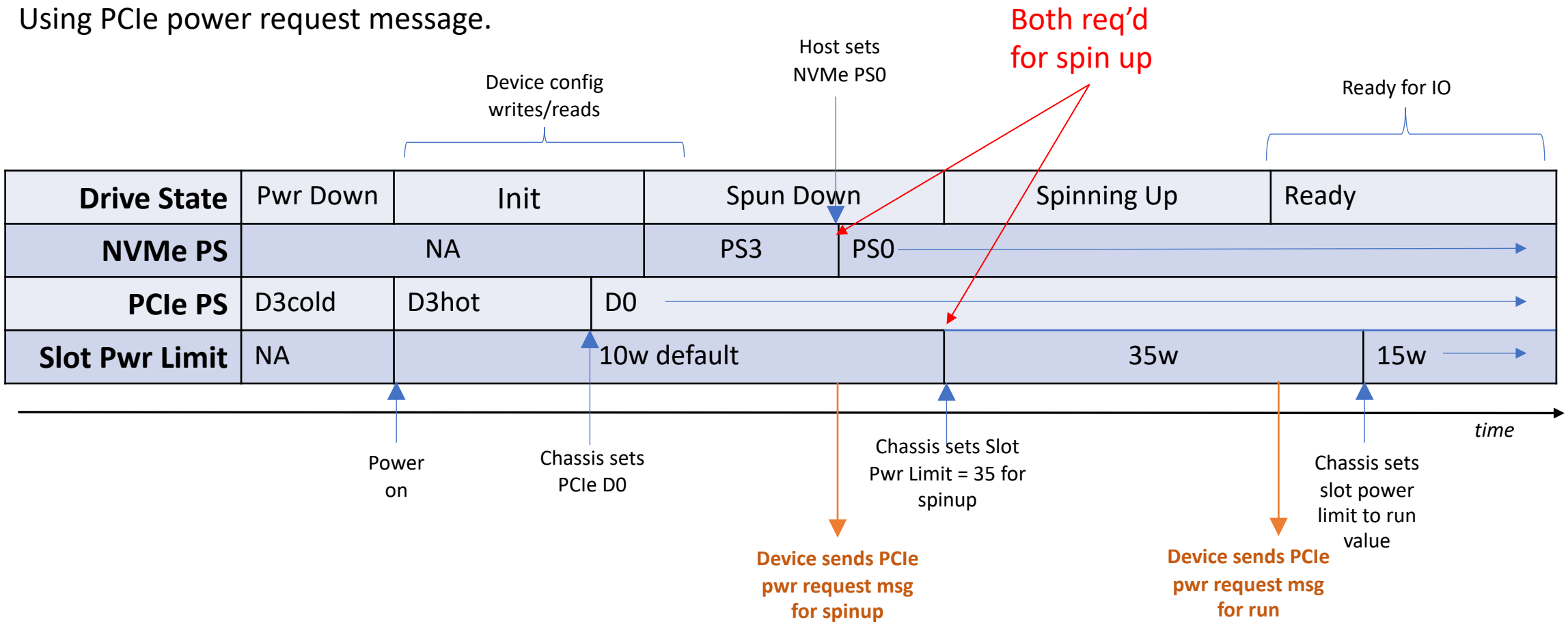
NVMe-HDD Initial Power On – Chassis Manage Pwr

Chassis actively manages power via Slot Pwr Limit
Spinup Enable = False; drive comes up in NVMe PS0 and spins up after power reset when sufficient captured_slot_power_limit is available.



NVMe-HDD Initial Power On – Host Commanded Spin-up

NVMe host controls spinup
Spinup Enabled = True; drive comes up in NVMe PS3; spins up when sufficient slot_power_limit is available and NVMe PS <= 3.
Using PCIe power request message.

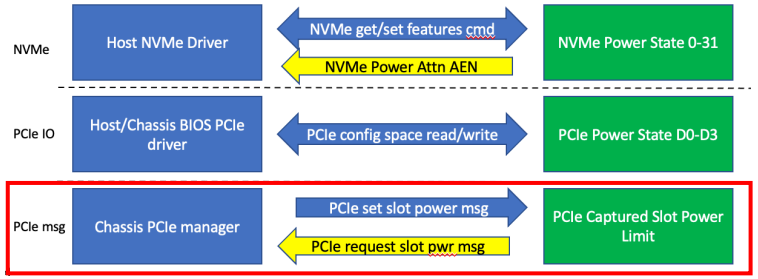


Backup

Spinup In A Nutshell

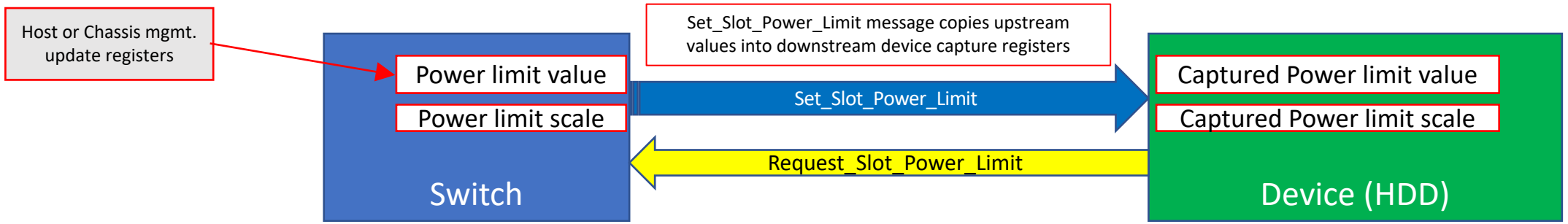
- Spinup Control Enable feature bit:
  - If Spinup Control enabled, the NVMe host will control HDD spin status via NVMe Set Features using existing NVMe power states, if sufficient power budget is available.
    - Power up in NVMe PS3.
  - If Spinup Control disabled, the drive will spin up automatically as soon as it sees sufficient power budget in the PCIe CSPL (captured slot power limit).
    - Power up in NVMe PS0.
- Drive will never intentionally draw more power than allowed by the PCIe Captured Slot Power Limit (CSPL) register pair.
  - Switch manager has to set power budget to allow spinup power or no spin.
  - Two choices
    - Set *Slot_Power_Limit* to permit spin up, and trust the system to be power-conscious. (how we do it today)
    - Dynamically configure drives' *Slot_Power_Limit* to support operations within enclosure power budgets.
- If drive operation requires more power than allowed by the PCIe CSPL
  - If host-initiated operation, fail the command with unique power error code.
  - Optionally: Drive request a higher CSPL by sending *Request_Slot_Power_Limit* PCIe msg.
  - Optionally: Drive Generate a power budget AEN to notify NVMe driver.
    - Do we need a unique error code AND an AEN?
    - Handled by different drivers?
- Drive will not autonomously transition to a lower power state unless the current CSPL will support transition back to the current higher power state.

PCIe Slot Power Limit



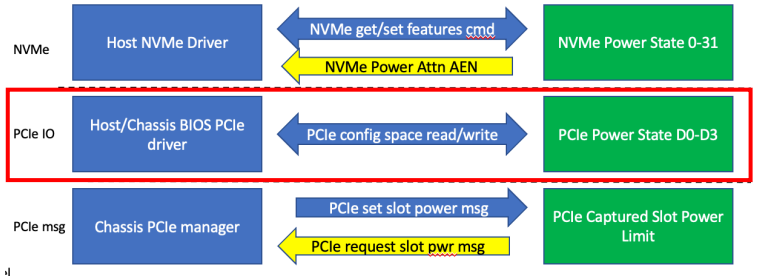
Ref: PCIe Base Specification 4.0 R1 Section 6.9 Slot Power Limit Control

- Slot Power Control is a software-managed method to limit a downstream device's maximum power.
  - *Slot Power Limit Value* and *Scale* fields of the Slot Capabilities register are implemented in the Downstream Ports of a Root Complex or a Switch
  - *Captured Slot Power Limit Value* and *Scale* fields of the Device Capabilities register are implemented in Endpoint, Switch, or PCI Express-PCI Bridge Functions present in an Upstream Port
  - ***Set_Slot_Power_Limit*** Message that conveys the content of the Slot Power Limit Value and Scale fields of the Slot Capabilities register of the Downstream Port (of a Root Complex or a Switch) to the corresponding Captured Slot Power Limit Value and Scale fields of the Device Capabilities register in the Upstream Port of the component connected to the same Link.
- The *Set_Slot_Power_Limit* message is sent at link initialization and any time the upstream registers values are updated.



- New ***Request_Slot_Power_Limit*** message
  - Same format as the Set_Slot_Power_Limit message with the power limit and scale values.
  - Device optionally sends this message if the Captured_Slot_Power_Limit value is insufficient for device operation.

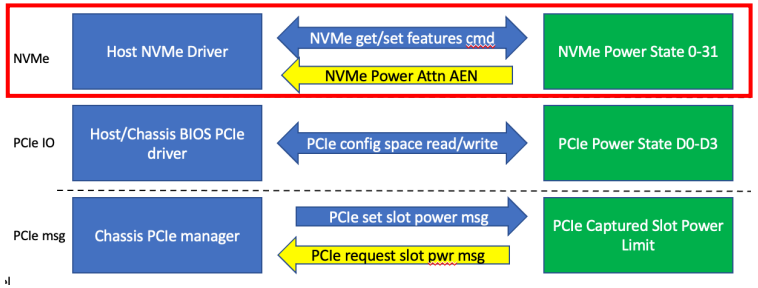
PCIe Power State



Ref: PCIe Base Specification 4.0 R1 Section 5.3 Device Power Management States (D-states) of a Function

- PCIe defines four device power states: D0-D3
  - The D0 power management state is the normal operation state of the Function. Other states are various levels of reduced power, where the Function is either not operating or supports a limited set of operations.
  - D1 and D2 are intermediate states that are not recommended for NVMe devices, per the NVMe spec
  - D3 comprises two substates, D3hot and D3cold, referring to whether main power is applied.
  - PCIe power state is configured by the PCIe driver. When powered up the device comes up in D3hot, then the PCIe power state register(Offset PMCAP + 4h: PMCS – PCI Power Management Control and Status) is written to bring the device (function) to the operational state D0.
  - Once the device is in PCIe D0 normal NVMe operation can commence.

NVMe Power State



Ref: PCIe Base Specification 4.0 R1 Section 5.3 Device Power Management States (D-states) of a Function

- NVMe defines 32 device power states: PS0-PS31. All but D0 are optional.
  - The D0 power state permits the maximum power usage. Subsequently numbered states must require less power than the next lower numbered state. PS0 > PS1 > ... > PS31
  - NVMe Set/Get Features commands set and report the current NVMe power state.
  - NVMe HDD POR is to implement PS0-PS3
    - 0=fully active
    - 1=heads idle but not retracted
    - 2=heads retracted
    - 3=spun down
- NVMe Power Attention AEN (New)
  - Device optionally sends when the device requires power management action
  - Host should validate power settings against the commanded or autonomous drive activity and increase/decrease power states as needed.

PCIe D0 is not the same as the *NVMe* power state PS0

- Both denote normal operating power states, but are controlled by different methods
- PCIe power state is configured by the PCIe driver. The device comes up in D3hot, then the PCIe power state register (Offset PMCAP + 4h: PMCS – PCI Power Management Control and Status) is written to bring the device (function) to an operational state. Once the device is in PCIe D0 normal NVMe operation can commence.
- NVMe power states are set by NVMe Set Features command, after the PCIe endpoint is up and configured with an admin queue established.
- Helpful to keep in mind PCIe is the *transport* while NVMe is the *protocol*. PCIe power states govern the low-level power enable. NVMe power states are device-dependent for power / performance optimization once fundamental PCIe and NVMe links are established.

Changes Required

- NVMe: NVMe Spinup Control – device configuration that determines if the device will spin-up up on power reset, or wait for a NVMe command.
  - Power reset comes from +12 available and PWRDIS de-asserted.
- NVMe: Define a power budget AEN – notify system when allocated power budget is inadequate for requested device operation.
- SFF Module Spec: Modify the 25W max power requirement from the 8639 module spec – allow chassis vendor to optimize the power supply limits.
- PCIe: Define a power request message companion to the Set Slot Power Limit -- Enable the device to request a specific slot power limit.

New Form Factor Definition

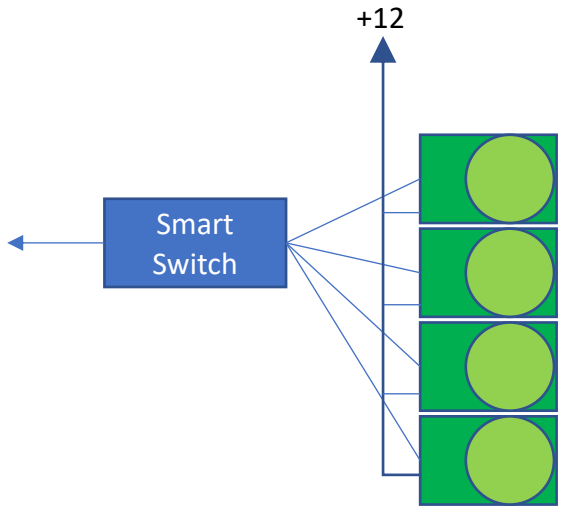
- New 3.5” form factor definition
- Establish a default slot power limit less than what most chassis vendors will budget per slot steady state.
  - 10-15W today
  - Enough for the drive to become able to be managed.
  - This is the power limit the HDD will observe until the upstream PCIe port has assigned a new one.
  - Drive is free to do any initialization, including spin, as long as power stays below form factor default.
  - Captured slot power limit will override the form factor default when the set_slot_power_limt message arrives from upstream port.

PCIe Power Request Message

- Define a new PCIe message to complement the set_slot_power message
- Downstream device sends the req_slot_power message to the upstream port to request a slot_power_limit.
- req_slot_pwer message is not acknowledged by the upstream switch, other than optionally setting a new slot_power_limit which will create a new set_slot_power message from the upstream port back down to the device.
- Use-case: NVMe HDD will temporarily request 35w for spin up. Will spin up when granted. Device will request it’s normal ACTP (active power) once spun up.

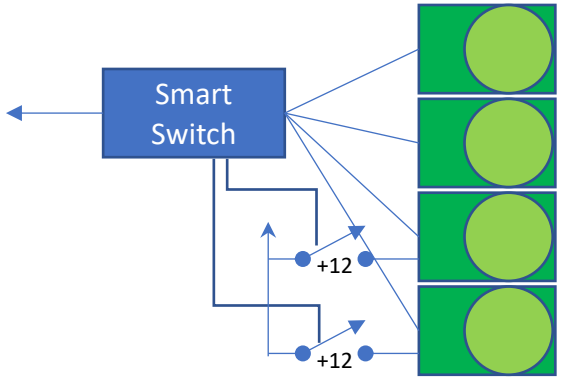
PCIe-Attached Chassis with No Slot Power Gating

- Chassis mgmt by smart switch CPU (implementation specific)
- HDD NVMe_Spinup_control -> disabled
- Chassis boots and sets Slot_Power_Limit to 15W.
- Chassis sequentially sets Slot_Power_limit to 35W using SAS-like timing to stagger spinup.
- Subsequent power excursions due to spin up/down unmanaged as they are today.



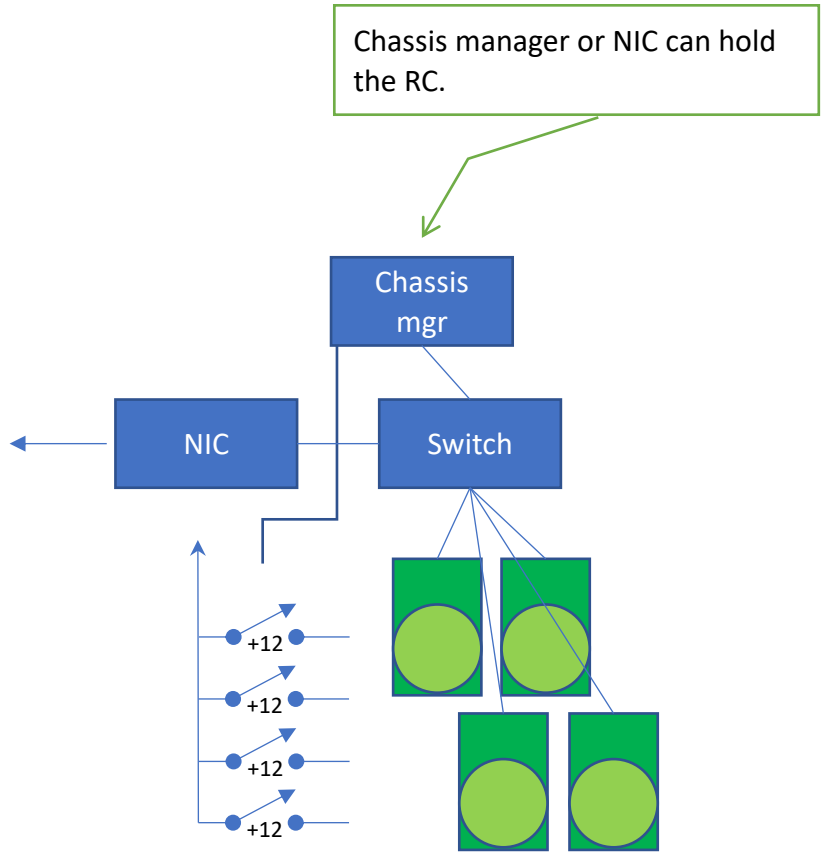
PCIe-Attached Chassis with Slot Power Gating

- HDD NVMe_Spinup_control -> disabled
- Smart switch boots and sets Slot_Power_Limit to 35W for spinup.
- Smart Switch sequentially enables +12 to slots to stagger spinup.
- Subsequent power excursions handled as today.



Fabric-Attached Chassis with Slot Power Gating

- PCIe enumeration and control is internal to the chassis.
- HDD NVMe_Spinup_control -> disabled
- Chassis boots and sets Slot_Power_Limit to form factor default: e.g. 10W.
- Chassis enables +12 to all slots.
- Chassis grants 35w to slots sequentially to stagger spinup
- As drives spin up chassis sets Slot_Power_Limit to run value (15w)
- Chassis sets next drive to 35w until it has gone through entire chassis.
- Drive generates AEN/PCIe message on power fault



NVMe HDD Spinup Control

Slot Power Limit control mechanism (ref: PCI Express Base Specification, Rev. 4.0 Version 1.0)

For Adapters:

- Until and unless a Set_Slot_Power_Limit Message is received indicating a Slot Power Limit value greater than the lowest value specified in the form factor specification for the adapter's form factor, the adapter must not consume more than the lowest value specified.
- Current 8639 specifies 25W max, which is also the “lowest value specified”. However, HDD chassis typically budget <15W steady state with ~40W spinup.
- An adapter must never consume more power than what was specified in the most recently received Set_Slot_Power_Limit Message or the minimum value specified in the corresponding form factor specification, whichever is higher.
- Components with Endpoint, Switch, or PCI Express-PCI Bridge Functions that are targeted for integration on an adapter where total consumed power is below the lowest limit defined for the targeted form factor are permitted to ignore Set_Slot_Power_Limit Messages, and to return a value of 0 in the Captured Slot Power Limit Value and Scale fields of the Device Capabilities register
- Such components still must be able to receive the Set_Slot_Power_Limit Message without error but simply discard the Message value

For Root Complex and Switches which source slots:

- Configuration software must not program a Set_Slot_Power_Limit value that indicates a limit that is lower than the lowest value specified in the form factor specification for the slot’s form factor.

Pertinent NVMe Power State Parameters

NVMe supports up to 32 power states. Each state includes a set of parameters describing its power profile.

Active Power Workload (APW): This field indicates the workload used to calculate maximum power for this power state. Refer to section 8.4.3 for more details on each of the defined workloads. This field shall not be “No Workload” unless ACTP is 0h.

Active Power (ACTP): This field indicates the largest average power consumed by the NVM subsystem over a 10 second period in this power state with the workload indicated in the Active Power Workload field. The power in Watts is equal to the value in this field multiplied by the scale indicated in the Active Power Scale field. A value of 0h indicates Active Power is not reported.

Idle Power (IDLP): This field indicates the typical power consumed by the NVM subsystem over 30 seconds in this power state when idle (i.e., there are no pending commands, register accesses, background processes, sanitize operation, nor device self-test operations). The measurement starts after the NVM subsystem has been idle for 10 seconds. The power in Watts is equal to the value in this field multiplied by the scale indicated in the Idle Power Scale field. A value of 0h indicates Idle Power is not reported. Refer to section 8.4.

Non-Operational State (NOPS): This bit indicates whether the controller processes I/O commands in this power state. If this bit is cleared to ‘0’, then the controller processes I/O commands in this power state. If this bit is set to ‘1’, then the controller does not process I/O commands in this power state. Refer to section 8.4.1.

Maximum Power (MP): This field indicates the sustained maximum power consumed by the NVM subsystem in this power state. The power in Watts is equal to the value in this field multiplied by the scale specified in the Max Power Scale bit. A value of 0h indicates Maximum Power is not reported. Refer to section 8.4.