Off Module Power Loss Protection

Cost Efficiencies for Enterprise Systems

Rob Sykes - Toshiba Memory America

Disclaimers

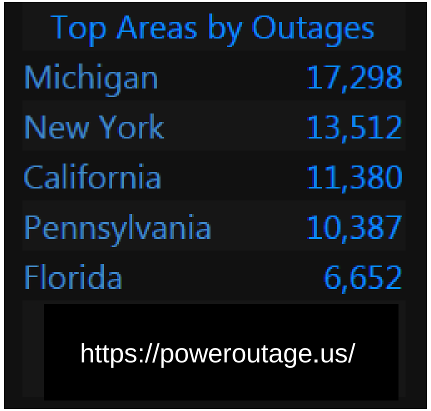
- The information in this presentation refers to specifications still in the development process. This presentation reflects the current thinking of various PCI-SIG® and NVMe™ workgroups. All material is subject to change before specifications are released!

Power Loss Types!

- Sudden Power Loss

- System Crash / Loss of Communication

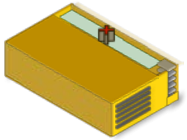
- Hot Removal



Top Areas by Outages	
Michigan	17,298
New York	13,512
California	11,380
Pennsylvania	10,387
Florida	6,652

https://poweroutage.us/

1000000000000000000000000000000000

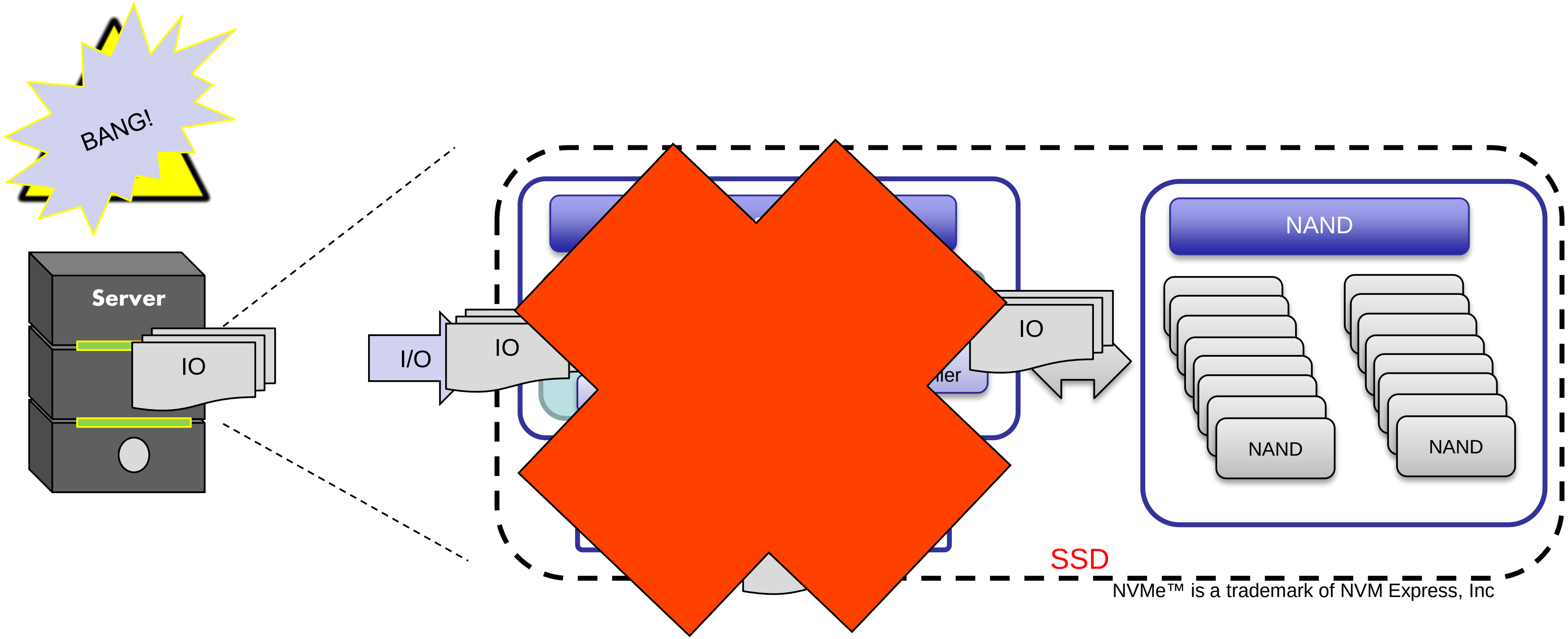


What is Power Loss Protection

- PLP is a hardware and firmware solution to ensure that the SSDs integrity is maintained should a power loss event occur !

  - Hardware – Capacitors / Battery / PMIC
  - FW – Clean shutdown / FTL flush

Data flow



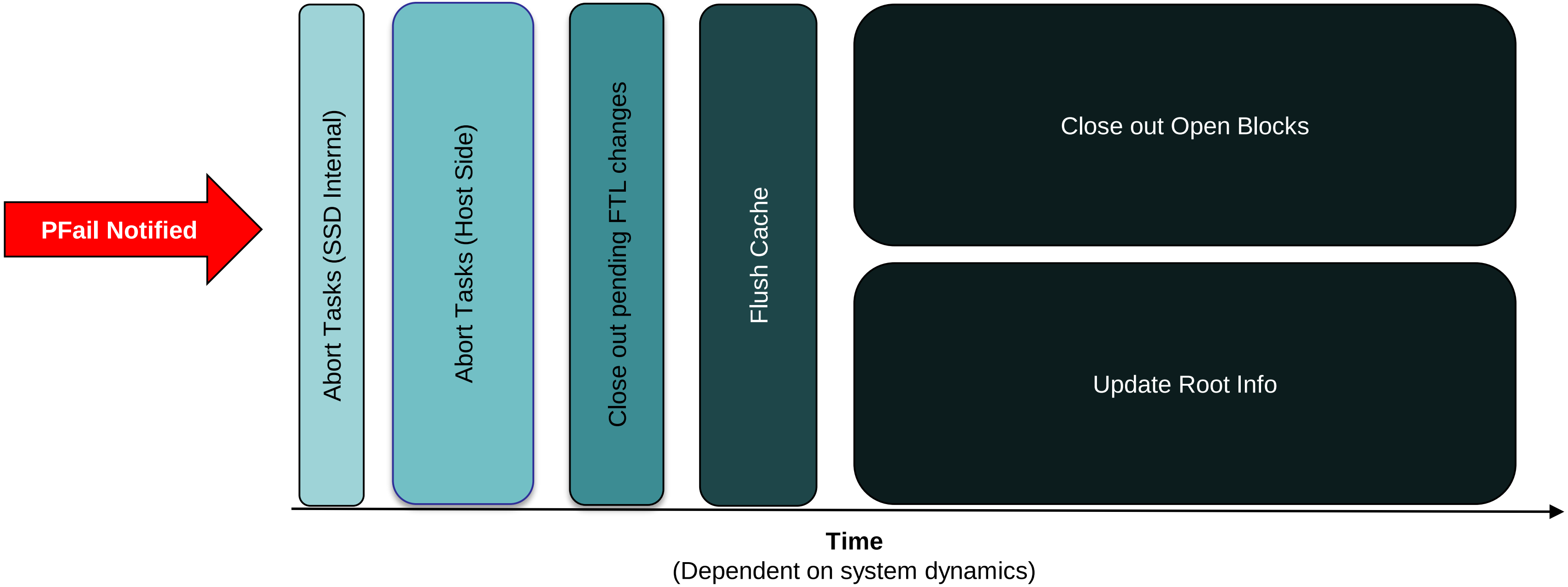
NVMe™ is a trademark of NVM Express, Inc

FTL Vulnerabilities

- We need to write the current metadata to the NAND that is temporarily stored in the DDR
- If we don’t, and power is lost, the metadata mapping table is lost and thus the data written on the NAND will not be accessible to the OS
- Note that its unlikely that the complete FTL is committed to the NAND on every transaction – this would impact performance so only the change in FTL is committed. This also reduces the time back up power is required to commit the data to the NAND
- Committing the data to the NAND (Flush) is only part of the transaction between the controller and the NAND
- NVMe™ provides another mechanism to keep the host and device in sync – if the device hasn’t sent a completion entry doorbell signal:
  - The host shouldn’t believe those commands were acted on
  - The device should roll back the changes to the FTL as much as possible, leaving any written blocks marked as dirty and unused

NVMe™ is a trademark of NVM Express, Inc

Time to ensure integrity



Enterprise Drives for Boot?

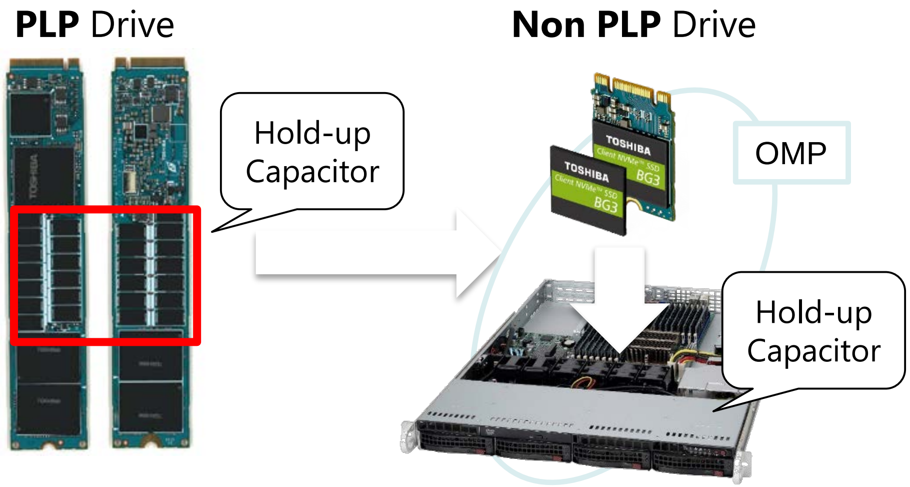
- Use of Enterprise class drives for boot is expensive – and you may not be benefiting from all the enterprise class features intended for Storage / Server Applications
    - May not require Dual Port and SR-IOV
    - May not need the higher OP to meet heavy workload DWPD found in enterprise drives
        - Boot drive IO is mostly reads!
    - May not require TCG enterprise requirements (Pyrite and Opal rather than Enterprise or Ruby)
    - May not require End to End data protection (PI)
    - May not require non native sector sizes
    - May not require SGL
    - May not require CMB
    - May not require support for many namespaces

OMP Introduction

- Typical boot drives for enterprise systems need to ensure data integrity in the event of system power losses and system crashes
- As such boot drives for enterprise systems are usually expensive enterprise grade drives with on board power loss protection to ensure integrity of data should a power loss event occur
- Off Module Power Loss Protection moves complexity in the SSD drive to the host system which may have integrated power loss protection
- The concept of Off Module Power Loss Protection is to move the detection and backup power mechanisms to the host system or daughter card
- Off Module Power Loss Protection is in the process of being standardized

OMP on the Host

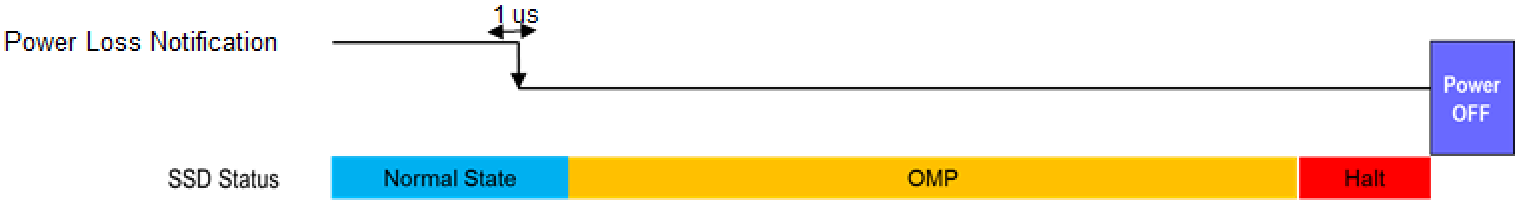
- Host needs to design the mother board (or carrier) to enable OMP
  - Power loss detection
  - Power Loss Notification signal
  - Optional Power Loss Acknowledge signal
  - Backup energy source



OMP Control Discussion

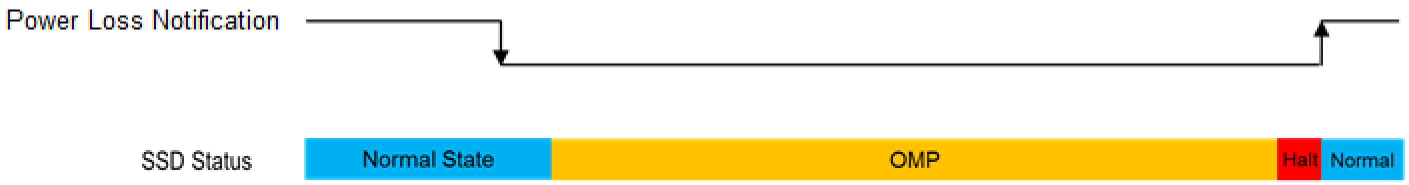
- “Off Module PLP” **controlled by two additional control signals**:
  - This approach is method to control “Off Module PLP” by two additional control signals. Additional signals are labelled as “Power Loss Notification” and “Power Loss Acknowledge”
- Power Loss Acknowledge is a useful adjunct
  - The Off board energy source and any host side timer needs to be based on worst case timing
  - Using Power Loss Acknowledge allows actual completion notice – could be 10x faster
    - Improves response time for user interactions
    - For battery powered energy sources, this extends batter life

Behavior of Off Module PLP controlled by Power Loss Notification – Notification Start



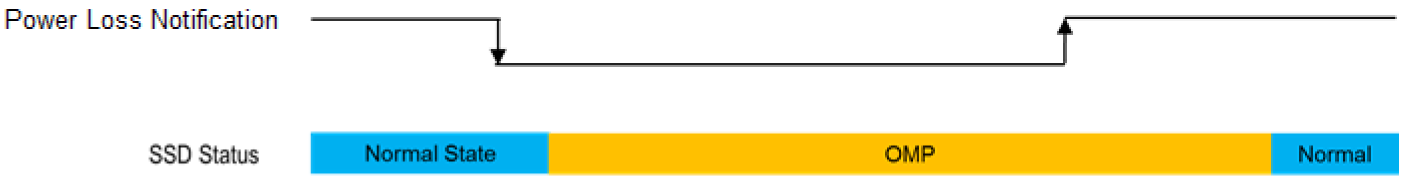
- Power Loss Notification is asserted, de-bounced and acted upon
- OMP Starts and SSD integrity is maintained

De-assert after complete – Discussion



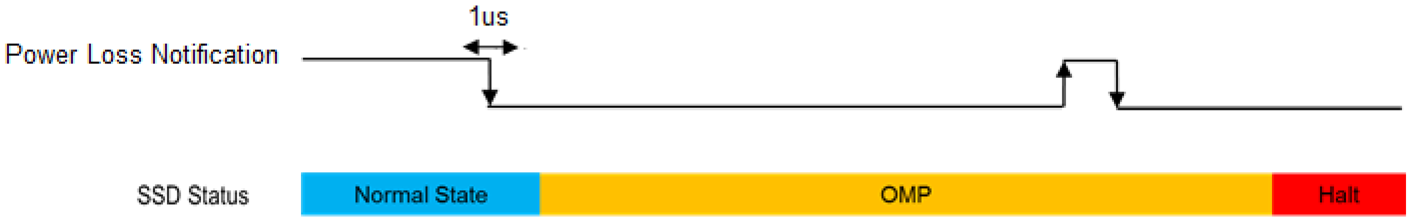
- If Power Loss Notification was de-asserted after completing Off Module PLP, SSD will return to normal operation state

De-assert during OMP - Discussion



- If Power Loss Notification was de-asserted during Off Module PLP operation, SSD will return to normal operation.

De-assert and assert during OMP - Discussion



- Power Loss Notification was de-asserted and asserted again during Off Module PLP (Flush)
  - SSD should complete the OMP task otherwise there is the potential of risking user data.

OMP Standardization Status

- OMP Standardization is occurring in PCI-SIG’s Mini Express committee and in NVMe™ (Focus on M.2)
    - It is proposed to be referred to as Power Loss Signals (Power Loss Notification and Power Loss Acknowledge)

- PCI-SIG status
    - This is an activity is in the Mini Express committee
        - Many Co-sponsor’s
        - ECR in discussion
- NVMe™ TPAR to be revisited as PCI-SIG reaches milestones
    - In progress

NVMe™ is a trademark of NVM Express, Inc

OMP – PCIe® SIG Status

- Current work is focused on the ECR
  - OMP control will be applied to M.2 WWAN cards
    - Pin selection is very difficult for Socket 2, Key B xxx-WWAN
  - There may be interest in providing Power Loss signaling outside of the scope of M.2, however pin availability is an issue
  - Such work will occur in the PCI-SIG domain
    - Join the PCI-SIG ElectroMechanical committee for more information

PCIe® is a registered trademark and/or service mark of PCI-SIG

OMP – NVMe™ Status

- TPAR 4029 (Off Module Power) held off to better develop how functionality is split between PCI-SIG and NVMe™
  - NVMe™ has formed a subgroup to work on this
    - Most are also supporting efforts in PCI-SIG
    - As PCI-SIG stabilizes, this activity will go back to NVMe™ for completion

NVMe™ is a trademark of NVM Express, Inc

A Possible Command Set

	Description
Set Features Command	**Off module PLP controlled by Notification signal Enable (OPIE):** This field specifies whether Off module PLP controlled by Notification signal is enabled. If this field is set to ‘1’, then OPIE is enabled. If this field is cleared to ‘0’, then OPIE is disabled. This field is cleared to ‘0’ by default.
Get Features Command	**Select (SEL):** This field specifies which value of the attributes to return in the provided data:
Identify - Controller Data Structure	**Off module PLP Capability (OPC):** This field indicates Off module PLP Capability that the controller supports.
	Off module PLP controlled by Notification signal Capability (OPIC): This field set to ‘1’ then the controller supports Off module PLP controlled by Notification signal. If cleared to ‘0’ then the controller does not support Off module PLP controlled by Notification signal.
	Typical Completion time of Off module PLP controlled by notification signal (TCOPI): This field indicates the typical maximum time when Off module PLP controlled by notification signal completes. If Off module PLP controlled by Notification signal is not supported, then this field is reserved. You can calculate typical energy budget that is identify power descriptor(PS0) x TCOPI
	Maximum Completion time of Off module PLP controlled by Notification signal (MCOPI): This field indicates the life end* of maximum time in microseconds when Off module PLP controlled by Notification signal completes. If Off module PLP controlled by Notification signal is not supported, then this field is reserved. You can calculate max energy budget that is identify power descriptor(PS0) x MCOPI
Get Log Page	**Accumulated Completion Count of Off module PLP controlled by Notification signal :** This field indicates accumulated count which Off module PLP controlled by Notification signal completes.
	Accumulated Incompletion Count of Off module PLP controlled by Notification signal This field indicates accumulated count which Off module PLP controlled by Notification signal doesn't complete.

Summary

- Power Loss Protection is a fact of life for SSDs
  - An upcoming development is Off Module Power
- Follow PCI-SIG (http://pcisig.com/specifications) for latest specifications and Engineering Changes (ECN's) and participate in PCI Express-Mini to provide feedback
- Join NVMe (https://nvmexpress.org/) to participate and provide feedback