NVM Express
The Interface Standard for PCI Express SSDs

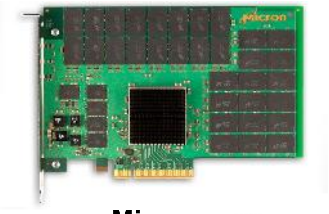
Kevin Marks
Principal Engineer / Technologist
Dell Inc.

Peter Onufryk
Director of Engineering
IDT

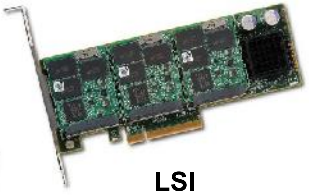
Benefits of PCIe as an SSD Interface

- **PCIe is High Performance**
  - Full duplex, multiple outstanding requests, and out of order processing
  - Scalable port width (x1 to x16)
  - Scalable link speed (2.5 GTps, 5 GTps, 8 GTps)
  - Low latency (no HBA overhead or protocol translation)
- **PCIe is Low Cost**
  - High volume commodity interconnect
  - Direct attach to CPU subsystem eliminates HBA cost
- **PCIe Provides Effective Power Management**
  - Direct attach to CPU subsystem eliminates HBA power
  - Link power management
  - Optimized Buffer Flush/Fill (OBFF)
  - Power Budgeting and Dynamic Power Allocation
  - Slot Power Limit

Fusion-io



Micron



LSI

OCZ

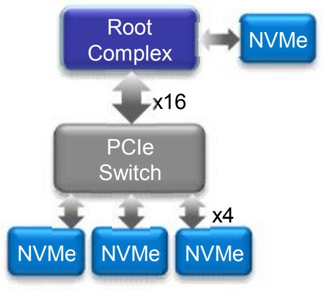
Virident

NVM Express (NVMe)

- NVMe defines an optimized queuing interface, command set, and feature set for PCIe SSDs
  - Architected to scale from client to enterprise
- Standardization accelerates industry adoption
  - Standard drivers
    - Eliminates need for OEMs to qualify a driver for each SSD
    - Enables broad adoption across a wide range of industry standard and proprietary OSes
  - Consistent feature set
    - All SSDs implement required features
    - Optional features are implemented in a consistent manner
  - Industry ecosystem
    - Development tools
    - Compliance and interoperability testing

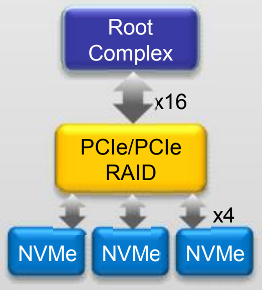
Example usage models for NVMe devices

Server Caching



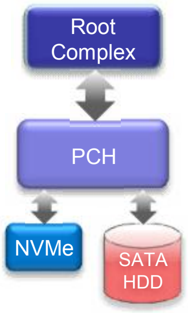
- Used for temporary data
- Non-redundant
- Used to reduce memory footprint

Server Storage



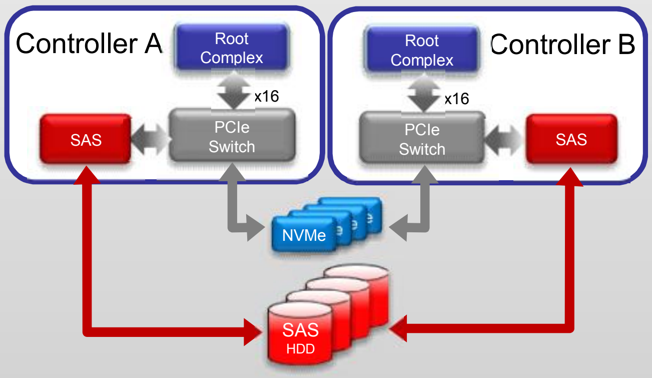
- Typically for persistent data
- Redundant (i.e., RAID'ed)
- Commonly used as Tier-0 storage

Client Storage



- Used for Boot/OS drive and/or HDD cache
- Non-redundant
- Power optimized

External Storage



- Used for Metadata or data
- Multi-ported device
- Redundancy based on usage

The Value of NVMe

NetApp

"The integration of solid state storage technology will have a profound impact on computer architectures over the next few years. The NVMe Work Group, driving a standard interface for the industry's most popular hardware building blocks, is ensuring that Solid State becomes a ubiquitous component of server and enterprise storage, enabling companies like NetApp to innovate at the system level and deliver unprecedented value to customers."

David Dale
Director of Industry Standards
NetApp

The Value of NVMe

NetApp

“The
prof
yea
the
ens
serv
to i
to c

EMC²

“EMC continues to be at the forefront of helping customers exploit the cost and performance benefits of SSDs, and is a strong supporter of standardization. Cloud and big data are prime examples of the drivers for insatiable performance demands customers are facing. Non-Volatile memory has the raw speed to meet these challenges, but requires an optimized interface across the industry to fully realize its potential. NVM Express can be that interface and EMC is pleased to help drive this important standard.”

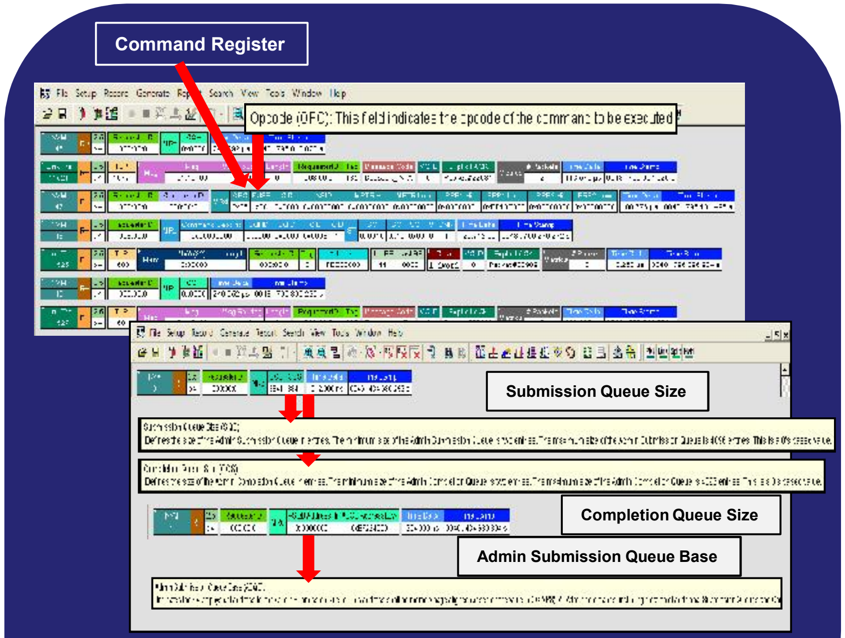
Bill DePatie
Vice President of Hardware Engineering
EMC

NVM Express Specification

- NVM Express 1.0 completed on March 1, 2011
  - Specification available at nvmexpress.org
- Specification cooperatively developed by more than **80 companies** within the NVMe Work Group
  - NVMe Work Group is directed by a multi-member Promoter Group of Companies consisting of Cisco, Dell, EMC, IDT, Intel, Micron, NetApp, Oracle, SandForce and STEC

NVMe Ecosystem

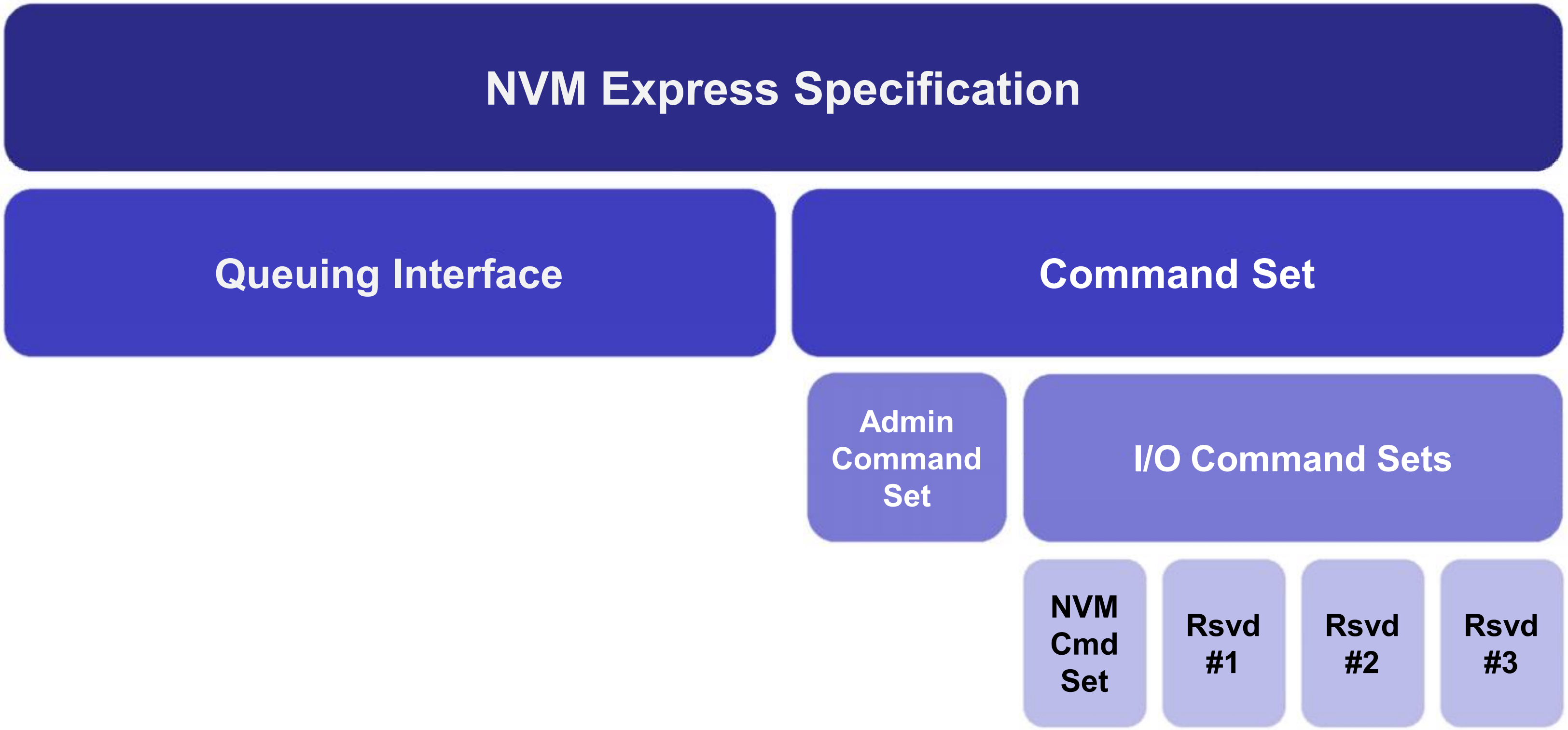
- **Products**
  - In development pipelines for 2012 releases
- **Drivers**
  - Linux driver available at nvmexpress.org
  - Windows driver in development (IDT, Intel, and SandForce effort)
- **Development Tools**
  - LeCroy PCIe analyzers available with NVMe decode
- **Compliance and Interoperability Testing**


Command Register
Opcode (OPC): This field indicates the opcode of the command to be executed
Submission Queue Size
Completion Queue Size
Admin Submission Queue Base

LeCroy PCIe Protocol Analyzer Trace from IDT 2011

NVMe Promoter Group has begun working with the University Of New Hampshire IOL to create an interoperability test suite and implementers list

NVMe Structure



NVM Express Specification

Queuing Interface

Command Set

Admin Command Set

I/O Command Sets

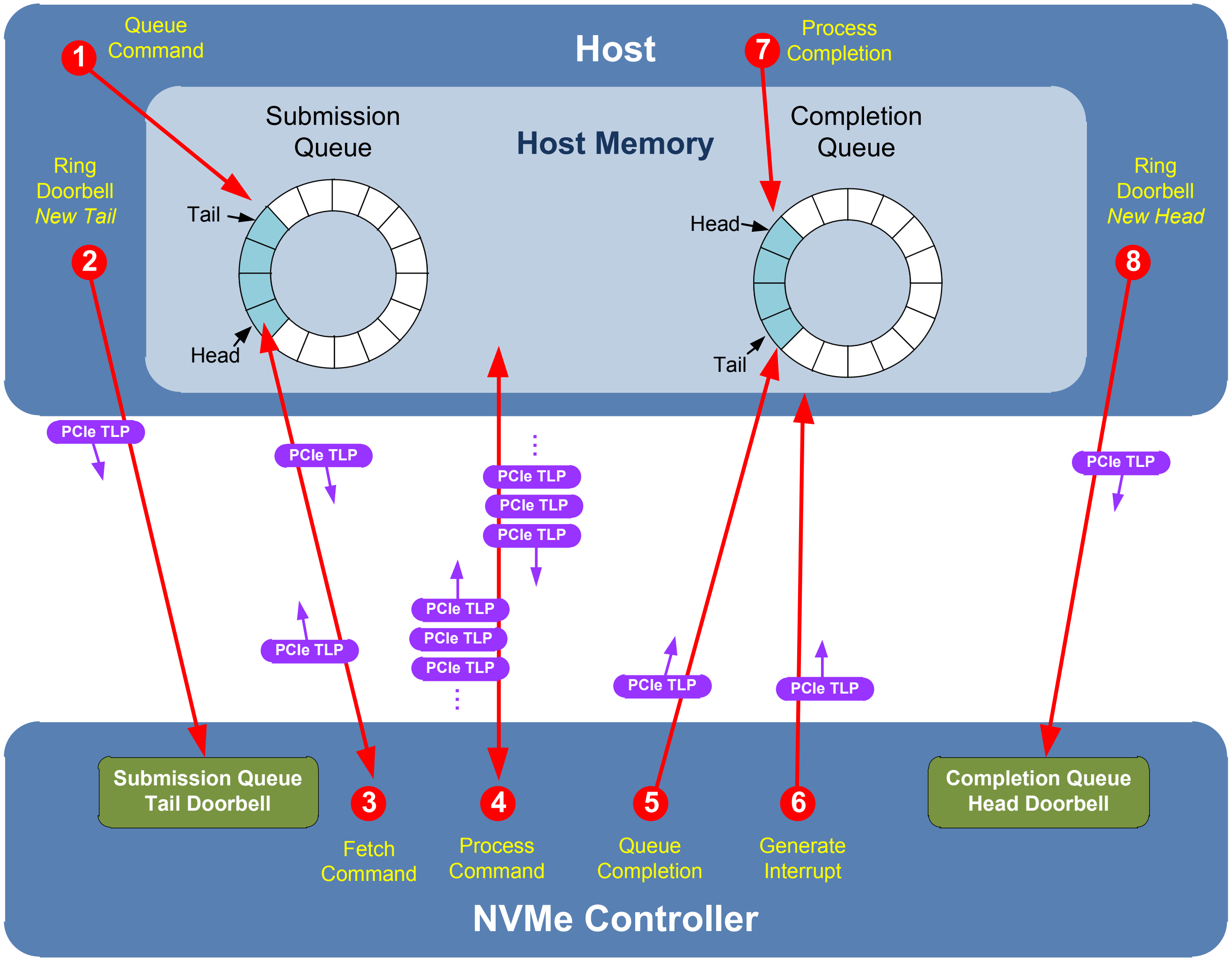
NVM Cmd Set

Rsvd #1

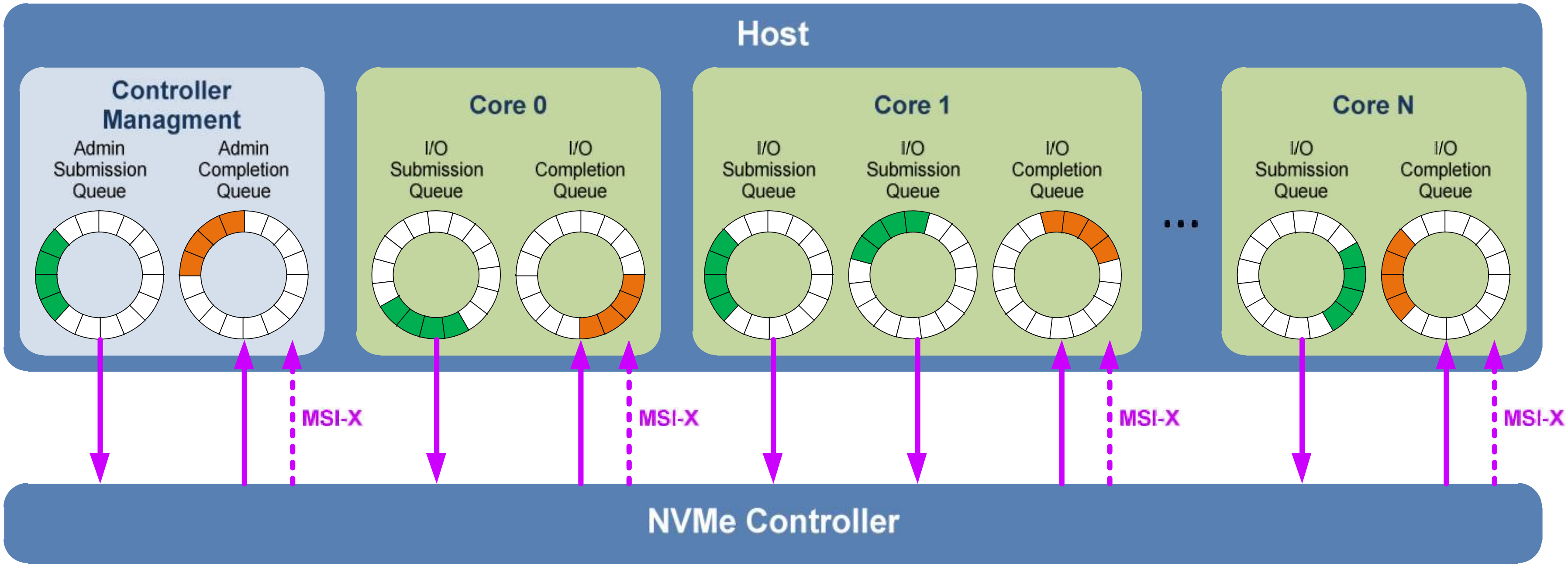
Rsvd #2

Rsvd #3

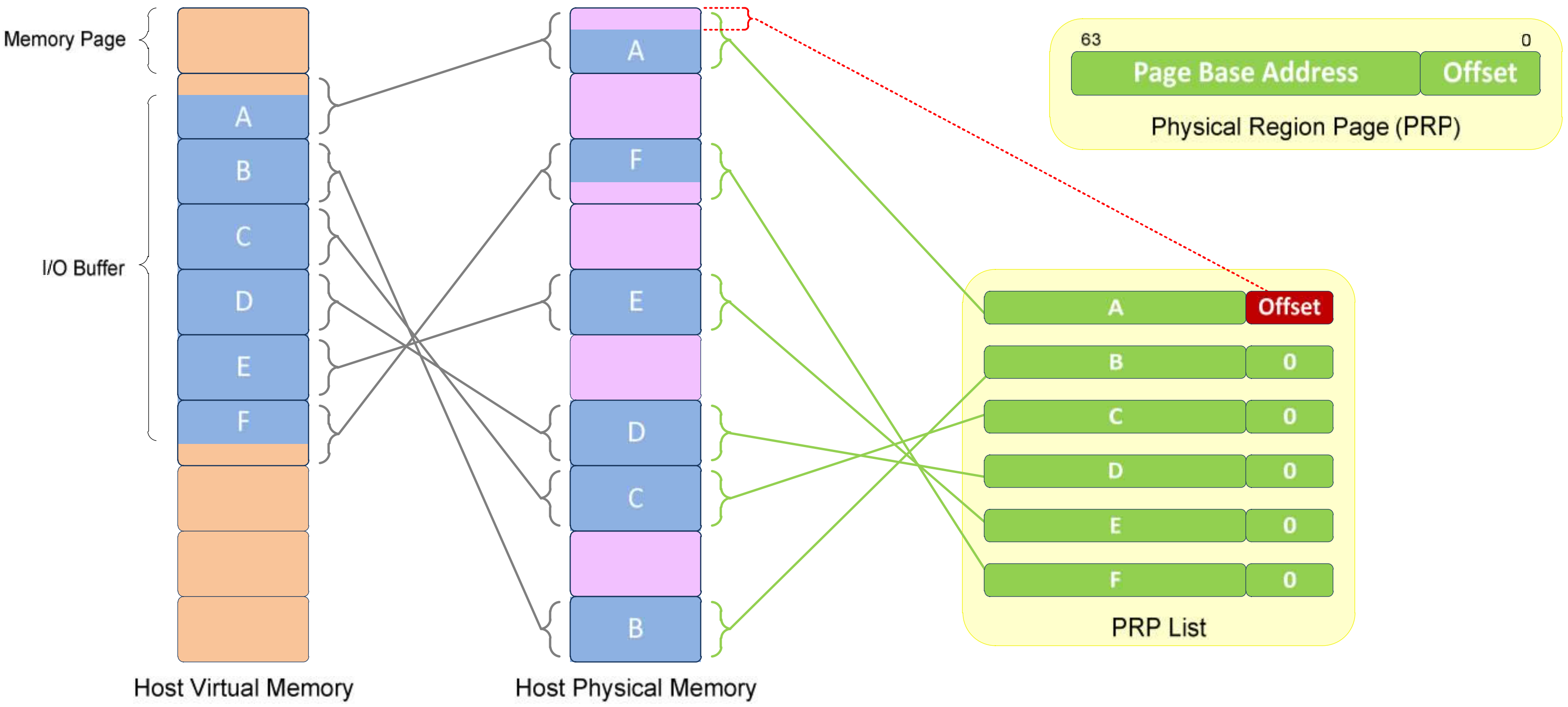
Queuing Interface



Admin and I/O Queues

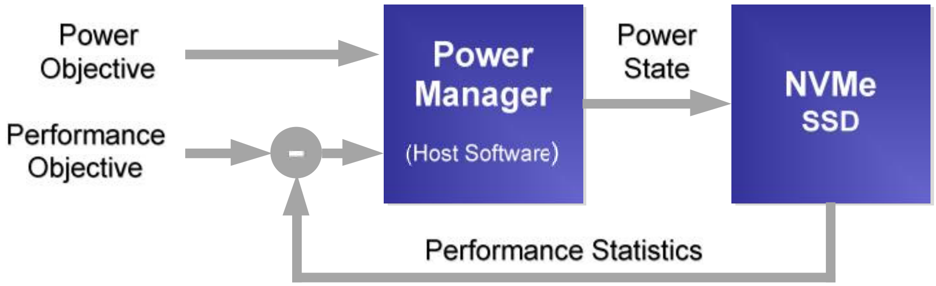


PRP Scatter/Gather Lists



Fixed Size PRP Lists Accelerate Out of Order Data Delivery

Power Management



Example Power State Descriptor Table
(Indentify Controller Data Structure)

Power State	Maximum Power	Entry Latency	Exit Latency	Relative Read Throughput	Relative Read Latency	Relative Write Throughput	Relative Write Latency
0	25 W	5 mS	5 mS	0	0	0	0
1	18 W	5 mS	7 mS	0	0	1	0
2	18 W	5 mS	8 mS	1	0	0	0
3	15 W	20 mS	15 mS	2	0	2	0
4	10 W	20 mS	30 mS	1	1	3	0
5	8 W	50 mS	50 mS	2	2	4	0
6	5 W	20 mS	5000 mS	5	3	5	1

Admin Command Set

Command	Required or Optional	Category
Create I/O Submission Queue	Required	Queue Management
Delete I/O Submission Queue	Required	
Create I/O Completion Queue	Required	
Delete I/O Completion Queue	Required	
Identify	Required	Configuration
Get Features	Required	
Set Features	Required	
Get Log Page	Required	Status Reporting
Asynchronous Event Reporting	Required	
Abort	Required	Abort Command
Firmware Image Download	Optional	Firmware Update / Management
Firmware Activate	Optional	
I/O Command Set Specific Commands	Optional	I/O Command Set Specific
Vendor Specific Commands	Optional	Vendor Specific

NVM Command Set

Command	Required or Optional	Category
Read	Required	Required Data Commands
Write	Required	
Flush	Required	
Write Uncorrectable	Optional	Optional Data Commands
Compare	Optional	
Dataset Management	Optional	Data Hints
Vendor Specific Commands	Optional	Vendor Specific

NVMe - Architected for Performance

- **No practical limit on number of outstanding requests**
  - Up to 64K I/O queues each with up to 64K entries
  - 32-bit controller unique command identifier (16-bit Queue ID + 16 Command ID) allows up to 2^{32} outstanding commands
- **Supports many-core processors without locking**
  - Each processor may be configured with its own submission/completion queues and MSI-X interrupt
- **At most one doorbell write to issue a command**
  - Multiple commands may be issued with a single doorbell write
- **Streamlined NVM command set avoids burdening controller with legacy command support requirements**
- **Fixed size 64B commands and 16B completions enable fast and efficient command decode and execution**
  - Commands contain 2 PRPs allowing 4KB or 8KB reads and writes be processed without fetching any additional information (e.g., scatter/gather list)
- **PRP based scatter/gather list allow efficient out-of-order data delivery**

NVMe Standardized Features

- **Logical block data and metadata**
- **End-to-end data protection (T10 DIF and DIX compatible)**
- **Security (Trusted Computing Group collaboration)**
- **Submission queue arbitration and QoS**
- **Firmware update and activation**
- **Dynamic power management**
- **Robust error reporting**
- **Interrupt coalescing configuration/control**
- **Capability discovery and configuration**

Summary

- **NVMe is a high performance queuing interface and command set optimized for PCIe SSDs**
- **NVMe is scalable from client to enterprise applications**
- **NVMe 1.0 specification is complete and available at** www.nvmexpress.org
- **Industry ecosystem is forming around NVMe**
  - Standard drivers and development tools