



NVMe™ Software Drivers: What's New and What's Supported?

Scott Lee – Windows; Sudhanshu Jain / Murali Rajgopal – vmware; Jim Harris – SPDK
Uma Parepalli, Session chair; Cameron Brett - Organizer

August 06, 2019

Sponsored by NVM Express™ organization, the owner of NVMe™, NVMe-oF™ and NVMe-MI™ standards

Speakers

Scott Lee



Sudhanshu (Suds)
Jain



Murali Rajagopal



Jim Harris

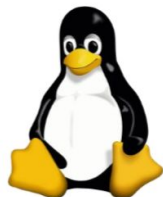


Uma Parepalli, Session Chair

Cameron Brett, Organizer

NVMe Driver Ecosystem

Robust drivers available on all major platforms



freeBSD®



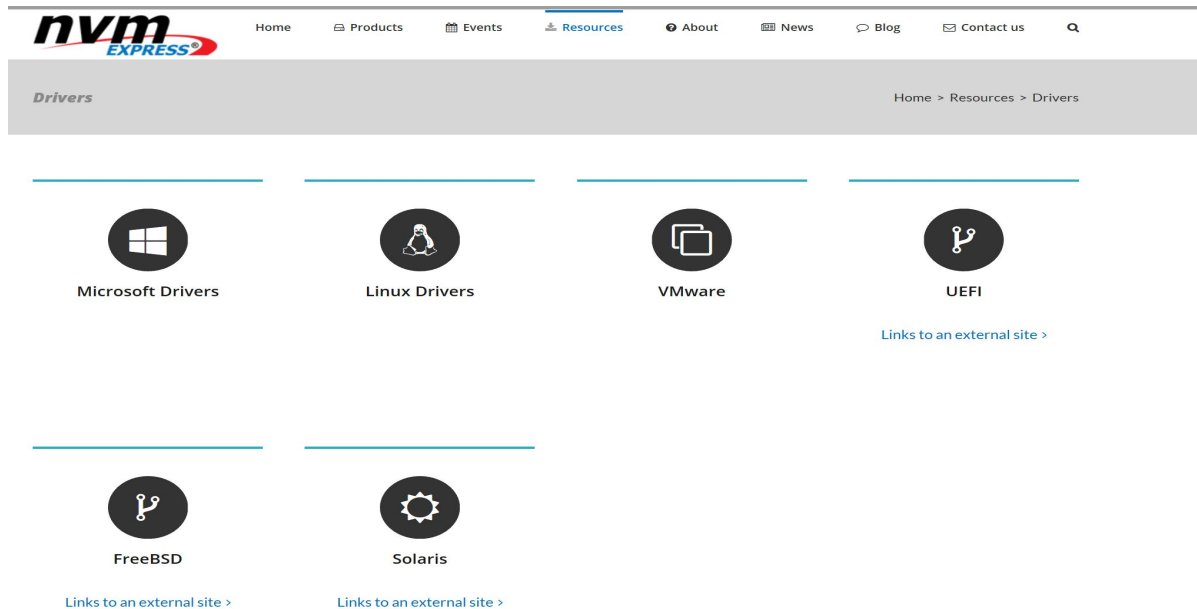
ORACLE®
SOLARIS



Flash Memory Summit

nvm
EXPRESS®

Visit NVM Express Website <http://nvmexpress.org> for Drivers related resources



The screenshot shows the NVM Express website's Drivers page. The header includes the NVM Express logo and a navigation menu with links for Home, Products, Events, Resources (highlighted), About, News, Blog, and Contact us. Below the header, a breadcrumb trail reads "Home > Resources > Drivers". The main content area features six driver categories, each with an icon and a label: Microsoft Drivers (Windows icon), Linux Drivers (Tux penguin icon), VMware (VMware logo icon), UEFI (UEFI icon), FreeBSD (FreeBSD icon), and Solaris (Solaris icon). Each category has a "Links to an external site >" link below it.

nvm
EXPRESS®

Home Products Events **Resources** About News Blog Contact us

Drivers Home > Resources > Drivers

Microsoft Drivers Linux Drivers VMware UEFI
Links to an external site >

FreeBSD Solaris
Links to an external site >

UEFI NVMe Drivers – Very stable in 2019

- Highly stable UEFI NVMe drivers available on Intel and ARM platforms
- NVMe support available from preboot UEFI to booting all major Operating Systems.



Flash Memory Summit

nvm
EXPRESS®

Windows Inbox NVMe™ Driver

Scott Lee, Principle Software Engineer Lead, Microsoft



Flash Memory Summit

nvm
EXPRESS®

6

Agenda

- New Additions for Windows 10 version 1903, May 2019 Update (19H1)
- Windows NVMe™ Diagnostic
- New Additions for Next Windows version
- Futures



Flash Memory Summit

nvm
EXPRESS®

Windows 10 version 1903, May 2019 Update

- TP4018/4018a: NVM Set & Endurance Group
- Improved diagnostics of NVMe hardware issues
 - Controller Fatal Status (CFS)
- Device Self-Test
- Runtime D3 for NVMe™
- Host Controlled Thermal Management Feature



Flash Memory Summit

nvm
EXPRESS®

NVMe™ Diagnostic – Controller Fatal Status

- Checked when Async Event Notification (AEN), controller reset (e.g. IO timeout), invalid command ID in completion entry or command failure
- Storport event 534 in Microsoft-Windows-Storage-Storport/Operational channel

Event 534, StorPort

General Details

Miniport logs an error for device (Port = 1, Path = 255, Target = 255, Lun = 255).
Id: 7
Error description: Controller Fatal Status is set
Corresponding Class Disk Device Guid: {00000000-0000-0000-0000-000000000000}
Adapter Guid: {5ef3fa6c-a98e-11e9-b93b-806e6f6e6963}
Miniport driver name: stornvme
VendorId:
ProductId:
SerialNumber:
DataLength: 4
Data: 0x0B000000

Log Name: Microsoft-Windows-Storage-Storport/Operational
Source: StorPort
Event ID: 534
Level: Error
User: N/A
OpCode: (109)
More Information: [Event Log Online Help](#)

Logged: 7/19/2019 5:24:55 PM
Task Category: Miniport logs an error.
Keywords: Event logged by Miniport
Computer: ██████████

NVMe™ Diagnostic – AEN

- Driver will send Asynchronous Event Request as part of controller initialization
- Event logged when AEN indicates a warning or error event
 - Error Event - Critical warning bit set
 - Warning Event - Available spare below 2
 - Warning Event – Percentage used above 95
- Storport event 539 for error events in Microsoft-Windows-Storage-Storport/Health channel
- Storport event 543 for warning events in Microsoft-Windows-Storage-Storport/Health channel



NVMe™ Diagnostic – AEN (cont)

- Example AEN Error Event - Critical Failure

Event 539, StorPort

General Details

The miniport logged a health event.

Log Name:	Microsoft-Windows-Storage-Storport/Health		
Source:	StorPort	Logged:	7/19/2019 4:49:26 PM
Event ID:	539	Task Category:	Port
Level:	Error	Keywords:	Asynchronous Event,Asynchronous Event
User:	N/A	Computer:	[REDACTED]
OpCode:	Info		
More Information:	Event Log Online Help		

Event 539, StorPort

General Details

Friendly View XML View

+ System

- EventData

- MiniportName** stornvme
- MiniportEventId** 14
- MiniportEventDescription** Health Status-Critical Warning
- PortNumber** 1
- AdapterGuid** {5ef5fa6c-a98e-11e9-b93b-806e6f6e6963}
- Parameter1Name** CriticalWarning
- Parameter1Value** 9
- Parameter2Name** Spare Below Threshold
- Parameter2Value** 1
- Parameter3Name** Temperature Threshold
- Parameter3Value** 0
- Parameter4Name** NVM Reliability Degraded
- Parameter4Value** 0
- Parameter5Name** Read Only Mode
- Parameter5Value** 1
- Parameter6Name** Volatile Backup Device Failure
- Parameter6Value** 0

NVMe™ Diagnostic – AEN (cont)

- Example AEN Warning Event – Percentage Used Above Threshold

Event 543, StorPort

General Details

The miniport logged a health event.

Log Name: Microsoft-Windows-Storage-Storport/Health
Source: StorPort Logged: 7/19/2019 4:49:55 PM
Event ID: 543 Task Category: Port
Level: Warning Keywords: Asynchronous Event, Asynchronous Event
User: N/A Computer: ██████████
OpCode: Info
More Information: [Event Log Online Help](#)

Event 543, StorPort

General Details

Friendly View XML View

+ System

-EventData

MiniportName stornvme
MiniportEventId 14
MiniportEventDescription Health Status-Endurance Warning
PortNumber 1
AdapterGuid {5ef5fa6c-a98e-11e9-b93b-806e6f6e6963}
Parameter1Name Percentage Used
Parameter1Value 99
Parameter2Name Endurance Threshold Limit
Parameter2Value 95

NVMe™ Diagnostic – IO Performance

- Classification of IO performance into pre-defined latency buckets
- Storport event 505 in Microsoft-Windows-Storage-Storport/Operational channel

Event 505, StorPort

General Details

Performance summary for Storport Device (Port = 2, Path = 0, Target = 0, Lun = 0) whose Corresponding Class Disk Device Guid is {5abac4d5-fbc6-04df-7de3-160367b04ce6}:
Total IO:492071
For latency buckets of 256us, 1ms, 4ms, 16ms, 64ms, 128ms, 256ms, 2000ms, 6000ms, 10000ms, 20000ms, 20000+ms,
The IO success counts are 463118, 22354, 5116, 1018, 462, 3, 0, 0, 0, 0, 0,
The IO failed counts are 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
The IO total latency (in 100ns) are 541484659, 102064523, 89287409, 93720526, 89488587, 3266952, 0, 0, 0, 0, 0,
Total Bytes Read:12292797952
Total Bytes Written:4095750144

Log Name: Microsoft-Windows-Storage-Storport/Operational
Source: StorPort Logged: 7/19/2019 5:36:02 PM
Event ID: 505 Task Category: Port
Level: Information Keywords: Write request,Read request
User: N/A Computer: [Redacted]
OpCode: Info
More Information: [Event Log Online Help](#)



NVMe™ Diagnostic – Command Tracing

- Support for tracing of NVMe command and response data
- Turn on by enabling Miniport and CommandTrace keywords for Microsoft-Windows-Storport ETW provider.
 - Method 1: Download Windows Performance Toolkit and run following commands in Command Prompt. Use Windows Performance Analyzer to view storport.etl.
 1. `xperf -start STORPORT -on Microsoft-Windows-Storport:0x0000200000000080:4 -BufferSize 1024 -MinBuffers 4096 -MaxBuffers 4096`
 2. `<run test>`
 3. `xperf -stop STORPORT -d storport.etl`
 - Method 2: Microsoft Message Analyzer.
 1. Add a new Live Trace and specify Microsoft-Windows-Storport as system provider. Configure the provider and select Miniport and CommandTrace keywords.
 2. Start the session and run your test. You should start to see some events.
 3. Stop the session to stop tracing.



NVMe™ Diagnostic – Command Tracing (cont)

- Example output from Windows Performance Analyzer

storport.etl - Windows Performance Analyzer

File Trace Profiles Window Help

1 Graph Explorer - storport.etl

System Activity

Generic Events Activity by Provider, T...

Getting Started Analysis

Generic Events Activity by Provider, Task, Opcode

Line #	Id	Srb (Field 1...	Para...	Para...	Parameter...	Parameter...	Pa...	Parameter4Na...	Para...	Parameter...	Para...	Parameter...	Para...	Parameter...	Para...	Parameter...	Count	Sum	Time (s)	
1	258																45			
2	00...	0xFFFFC481...	CID	134	OPC	9	FUSE	0	PSDT	0	NSID	0	MPTR	0	PRP1	0	PRP2	0	1	3.204390200
3	00...	0xFFFFC481...	CID	134	CDW10	2	CDW11	0	CDW12	0	CDW13	0	CDW14	0	CDW15	0	DPTR	0	1	3.204392500
4	00...	0xFFFFC481...	CID	134	Status.SC	0	Status.SCT	0	Complete Status	1	DW0	0	DW2	211	NULL	0	NULL	0	1	3.204613400
5	1...	0xFFFFC481...	CID	567	OPC	2	FUSE	0	PSDT	0	NSID	1	MPTR	0	PRP1	10282471424	PRP2	0	1	3.205655100
6	1...	0xFFFFC481...	CID	567	LBALOW	86492440	LBAHIGH	0	CDW12	7	CDW13	0	CDW14	0	CDW15	0	NULL	0	1	3.205657600
7	1...	0xFFFFC481...	CID	567	Status.SC	0	Status.SCT	0	Complete Status	0	DW0	0	DW2	131640	NULL	0	NULL	0	1	3.205770000
8	00...	0xFFFFC481...	CID	135	OPC	9	FUSE	0	PSDT	0	NSID	0	MPTR	0	PRP1	0	PRP2	0	1	3.413674900
9	00...	0xFFFFC481...	CID	135	CDW10	2	CDW11	3	CDW12	0	CDW13	0	CDW14	0	CDW15	0	DPTR	0	1	3.413678800
10	00...	0xFFFFC481...	CID	135	Status.SC	0	Status.SCT	0	Complete Status	1	DW0	0	DW2	212	NULL	0	NULL	0	1	3.413733000
11	00...	0xFFFFC481...	CID	136	OPC	9	FUSE	0	PSDT	0	NSID	0	MPTR	0	PRP1	0	PRP2	0	1	5.429367300
12	00...	0xFFFFC481...	CID	136	CDW10	2	CDW11	0	CDW12	0	CDW13	0	CDW14	0	CDW15	0	DPTR	0	1	5.429370500
13	00...	0xFFFFC481...	CID	136	Status.SC	0	Status.SCT	0	Complete Status	1	DW0	0	DW2	213	NULL	0	NULL	0	1	5.429434500
14	00...	0xFFFFC481...	CID	137	OPC	9	FUSE	0	PSDT	0	NSID	0	MPTR	0	PRP1	0	PRP2	0	1	5.430471000
15	00...	0xFFFFC481...	CID	137	CDW10	2	CDW11	4	CDW12	0	CDW13	0	CDW14	0	CDW15	0	DPTR	0	1	5.430473900
16	00...	0xFFFFC481...	CID	137	Status.SC	0	Status.SCT	0	Complete Status	1	DW0	0	DW2	214	NULL	0	NULL	0	1	5.430535400
17	00...	0xFFFFC481...	CID	138	OPC	9	FUSE	0	PSDT	0	NSID	0	MPTR	0	PRP1	0	PRP2	0	1	6.454252300
18	00...	0xFFFFC481...	CID	138	CDW10	2	CDW11	0	CDW12	0	CDW13	0	CDW14	0	CDW15	0	DPTR	0	1	6.454253800
19	00...	0xFFFFC481...	CID	138	Status.SC	0	Status.SCT	0	Complete Status	1	DW0	0	DW2	215	NULL	0	NULL	0	1	6.454316100
20	1...	0xFFFFC481...	CID	568	OPC	2	FUSE	0	PSDT	0	NSID	1	MPTR	0	PRP1	27188666368	PRP2	0	1	6.455463000
21	1...	0xFFFFC481...	CID	568	LBALOW	34712688	LBAHIGH	0	CDW12	0	CDW13	0	CDW14	0	CDW15	0	NULL	0	1	6.455466600
22	1...	0xFFFFC481...	CID	568	Status.SC	0	Status.SCT	0	Complete Status	0	DW0	0	DW2	131641	NULL	0	NULL	0	1	6.455567600



NVMe™ Diagnostic – Command Tracing (cont)

- Example output from Microsoft Message Analyzer

The screenshot displays the Microsoft Message Analyzer interface. The main window shows a list of messages with columns for MessageNumber, Timestamp, TimeElapsed, Source, Destination, Module, and Summary. The messages are filtered by the expression `tcp.port==80` and `*address==192.168.1.1`. The selected message (Message 1) is expanded to show its details in the 'Details 1' pane. The details pane shows the following fields:

Name	Value	Bit Offset	Bit Length	Type
MiniportName	stornvme	0	144	String
MiniportEventId	4 (0x00000004)	144	32	UInt32
MiniportEventDescrip...	Admin Command	176	224	String
PortNumber	2 (0x00000002)	400	32	UInt32
AdapterGuid	c5220643-2e8c-11e7-9b38-ecb1d7548344	432	128	Guid
PathID	255 (0xFF)	560	8	Byte
TargetID	255 (0xFF)	568	8	Byte
LUN	255 (0xFF)	576	8	Byte
ClassDeviceGuid	00000000-0000-0000-0000-000000000000	584	128	Guid
VendorId		712	8	String
ProductId		720	8	String
SerialNumber		728	8	String
Irp	0x0000000000000000	736	64	Etw.Etw...
Srb	0xFFFF48166103350	800	64	Etw.Etw...
Parameter1Name	CID	864	64	String
Parameter1Value	189 (0x00000000000000BD)	928	64	UInt64
Parameter2Name	QRC	992	64	String
Parameter2Value	9 (0x0000000000000009)	1056	64	UInt64
Parameter3Name	FUSE	1120	80	String
Parameter3Value	0 (0x0000000000000000)	1200	64	UInt64
Parameter4Name	PSDT	1264	80	String
Parameter4Value	0 (0x0000000000000000)	1344	64	UInt64
Parameter5Name	NSID	1408	80	String
Parameter5Value	0 (0x0000000000000000)	1488	64	UInt64

The 'Field Data' pane shows the value 189. The status bar at the bottom indicates 'Ready', 'Session Total: 207', 'Available: 207', 'Selected: 1', 'Viewpoint: Default', 'Truncated Session: False', 'Parsing Level: Full', and 'Build: 4.0.8112.0'.

Next Windows Version

- Development for next Windows version in progress
- Non-Operational Power State Config Feature
- LED for NVMe™ Devices
 - ACPI-based: PCIe® SSD Status LED Management _DSM
 - PCI-based: Native PCIe Enclosure Management (NPEM)



Flash Memory Summit

nvm
EXPRESS®

Futures*

- Native NVMe™ Storage Stack
- Zoned Namespace (ZNS)
- Device Firmware Hang Detection
- Runtime Hardware Reset of NVMe Devices

* Not plan of record



Flash Memory Summit

nvm
EXPRESS®



Flash Memory Summit



vSphere NVMe™ Driver Support

Sudhanshu (Suds) Jain and Murali Rajagopal, VMware

NVMe™ Focus @VMWare

vSphere 6.5

vSphere 6.7

Future Direction

Driver

- Boot (UEFI)
- Firmware Update
- End-to-end protection
- Deallocate/TRIM/Unmap
- 4K
- SMART, Planned hot-remove

- Performance enhancements
- Extended CLI
- Name space management
- Async event error handling
- Enhance diagnostic logs

- PCIe Native Hot-plug
- LED Management
- NVMe Over Fabric
- Multiple fabric option
- Sanitize

Core Stack

- Reduced serialization
- Locality improvements
- vNVMe Adaption layer
- Multiple completion worlds support in NVMe

- Optimized stack - Highly parallel execution for single path local NVMe devices
- Reach target of 90%+ performance of device spec

- Next Generation Storage Stack with ultra-high IOPS
- End-to-end NVMe Stack
- NVMe Multi-pathing, ANA

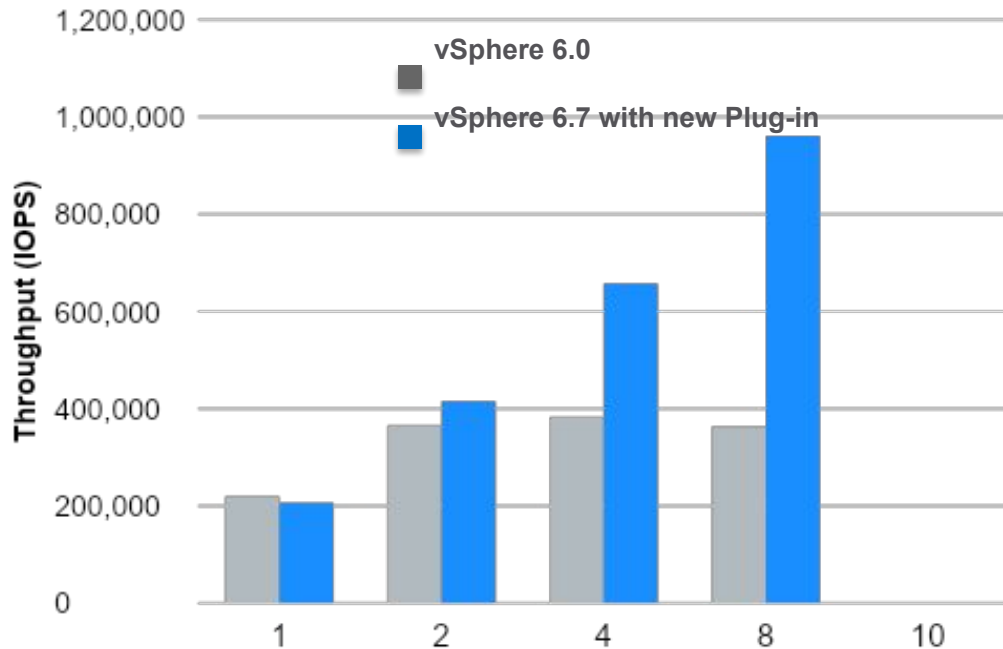
Virtual Devices

- NVMe™ 1.0e spec
- Hot-plug support
- VM orchestration

- Performance improvements
- Async mode support
- unmap support

- Rev the specification
- Parallel execution @backend
- 4K Support
- Scatter-gather support
- Interrupt coalescing

NVMe™ Performance Boost



Hardware:

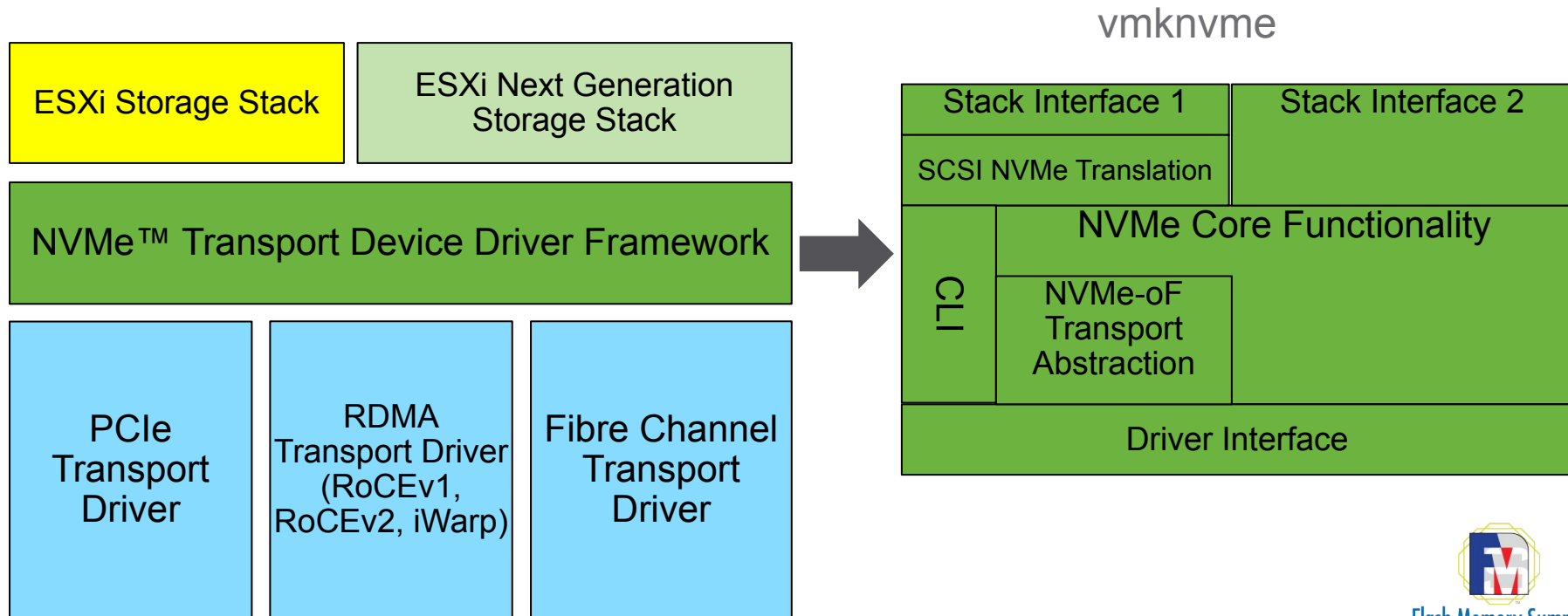
- Intel® Xeon® E5-2687W v3 @3.10GHz (10 cores + HT)
- 64 GB RAM
- NVM Express* 1M IOPS @ 4K Reads

Software:

- vSphere* 6.0U2 vs. Future prototype
- 1 VM, 8 VCPU, Windows* 2012, 4 VMDK eager-zeroed
- IOMeter:
 - 4K seq reads, 64 OIOs per worker, even distribution of workers to VMDK

The information in this presentation is intended to outline our general product direction and it should not be relied on in making a purchasing decision. It is for informational purposes only and may not be incorporated into any contract.

(Future) NVMe™ Driver Architecture



VMware's NVMe™ Driver Ecosystem

- Available as part of base ESXi image from vSphere 6.0 onwards
 - ❑ Faster innovation with async release of VMware NVMe™ driver
- VMware Opensource its NVMe Driver to encourage ecosystem to innovate
 - ❑ <https://github.com/vmware/nvme>
- Broad VMware NVMe Driver Ecosystem
<https://www.vmware.com/resources/compatibility/search.php?deviceCategory=io>
 - ❑ Close to 300 third party NVMe devices certified on VMware NVMe driver
- Beyond NVMe PCI Driver (Future)
 - ❑ Actively working with broad I/O controller and storage array partners to bring NVMe-oF solutions





Accelerating NVMe[™] with SPDK

Jim Harris, Principal Software Engineer, Intel

Notices and disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.
- Some results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance..
- Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.
- Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
- Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804.
- The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations.
- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.
- Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.
- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.
- The cost reduction scenarios described are intended to enable you to get a better understanding of how the purchase of a given Intel based product, combined with a number of situation-specific variables, might affect future costs and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs or cost reduction.
- No computer system can be absolutely secure.
- © 2019 Intel Corporation. Intel, the Intel logo, Xeon and Xeon logos are trademarks of Intel Corporation in the U.S. and/or other countries.
- *Other names and brands may be claimed as the property of others.



Flash Memory Summit

26 **nvm**
EXPRESS®

Storage Performance Development Kit



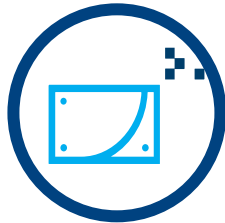
User Space Storage Software Stack

- Extreme performance (>10M IO/s on one thread)
- Block device abstraction and device drivers
- Network and virtualization protocols
- Resets, timeouts, I/O splitting, volume management



Widely Adopted

- Powering major storage systems in production today



C Libraries and Applications

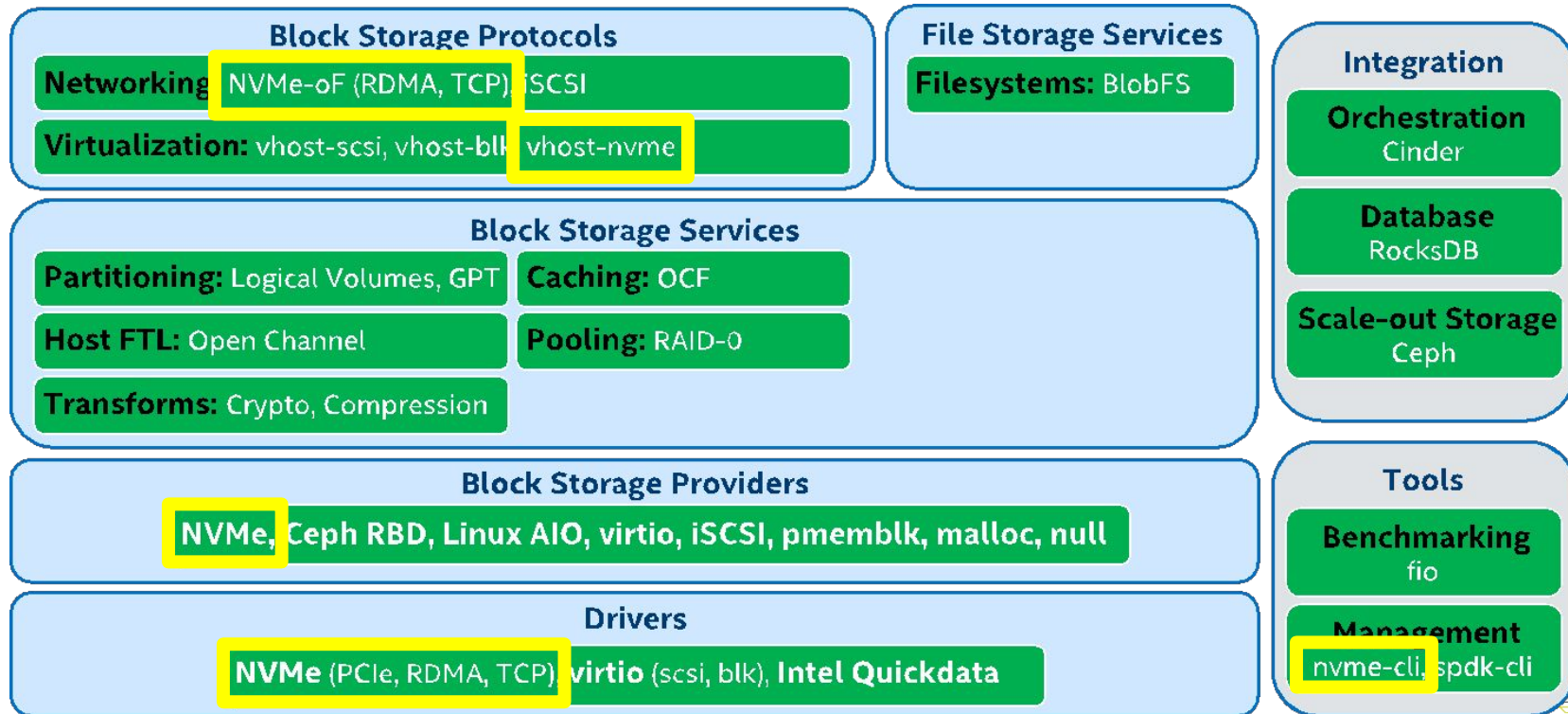
- Open Source (GitHub, BSD License)
- Active Community (~50 contributors each quarter)



Flash Memory Summit

nvm
EXPRESS®

SPDK Architecture



SPDK and Kernel

SPDK has better performance and efficiency compared to interrupt-driven kernel mode approaches

BUT...

SPDK is not a general-purpose solution

- covers some use cases very well – others not at all (or at least not well)

Polled mode design and userspace implementation drove much of the SPDK design



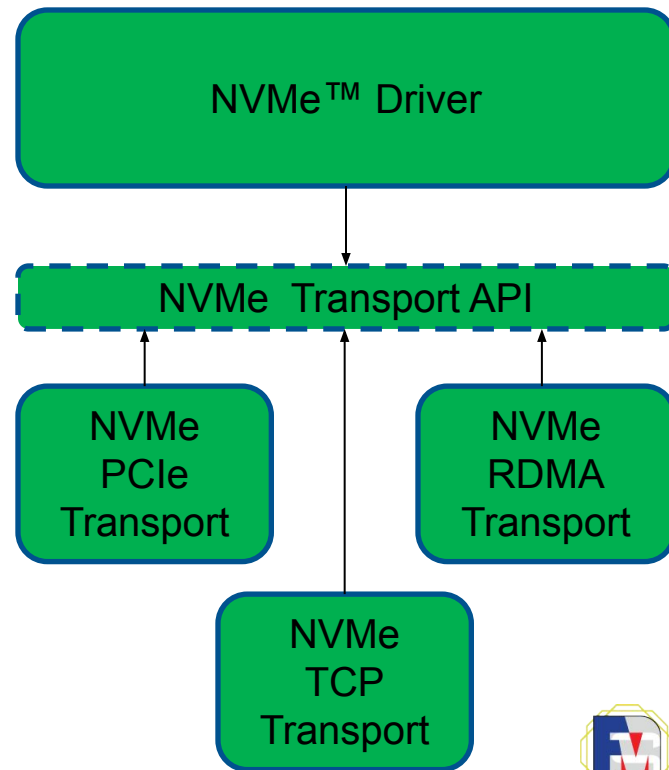
Flash Memory Summit

nvm
EXPRESS®

NVMe™ Transport Abstraction

Enables different implementations for different transports

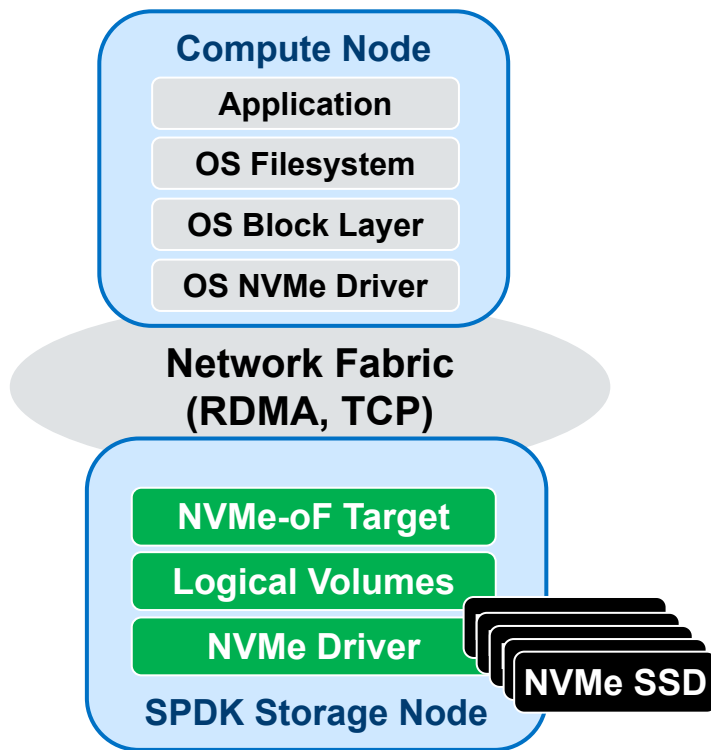
- construct/destroy controller
- set/get register value
- create/delete I/O queue pair
- submit request
- process completions



Flash Memory Summit

nvm
EXPRESS®

NVMe-oF™ Target



Spec-compliant, fully functional NVMe-oF™ target

- No modifications on client/compute node

Supports broad range of storage services – including:

- Sharing SSD across multiple clients (Logical Volumes)
- At-rest data encryption with crypto offload
- SSD pooling/stripping

Supported Features

Explicit Queue Pair Allocation

Metadata and Data Protection

Controller Memory Buffer

Timeout Handling

SGL

Asynchronous Attach

AER

NVMe-oF™ Persistent Reservations

NVMe™ /TCP

NVMe™ TP ratified November 2018

SPDK added TCP transport for

- NVMe driver
- NVMe-oF™ target

Supports alternative TCP stack implementations

Host Block FTL

Host FTL enabling smart data placement

- Based on OC2.0 specification

Block FTL support added to bdev nvme module

Long term goal: Zoned Namespace API

- With ZNS/OC adapters

NVMe™ Performance: Avoid MMIO

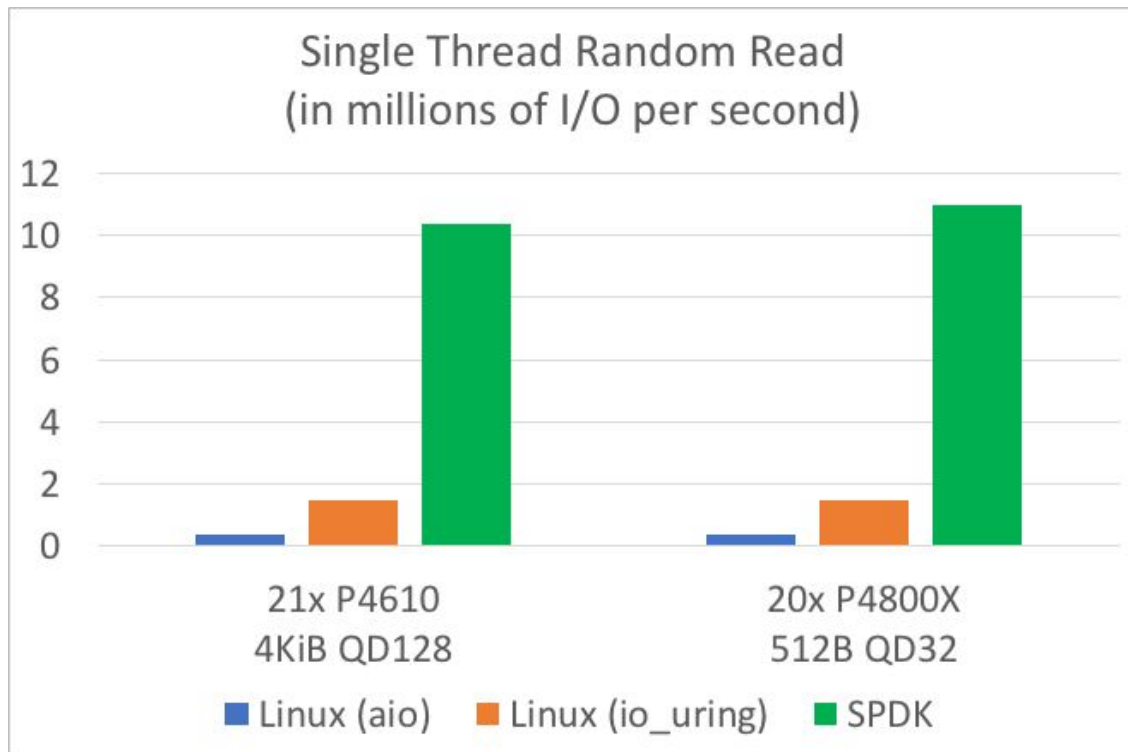
- Past: Simple completion queue doorbell batching



- Ring doorbell after processing first 3 completions
- Recent: Leverage polling
 - Delay ringing submission queue doorbell until end of poll call
- Future: Advanced completion queue batching
 - Track number of free cq slots
 - Only ring doorbell when slots are needed

SPDK NVMe™ Driver Performance

<https://spdk.io/news/2019/05/06/nvme/>



System Configuration: 2S Intel(R) Xeon(R) Platinum 8280L (use single thread for testing), 192GB DDR4 Memory, 6x Memory Channels per socket, Fedora 29, Linux kernel 5.0.0-rc+, BIOS: HT enabled, p-states enabled, turbo enabled, SPDK 19.04+, SPDK nvme-perf tool used for benchmarking, numjobs=1, direct=1, 21x Intel P4610 1.6T SSD or 20x Intel P4800X 375GB SSD.



Flash Memory Summit

nvm
EXPRESS®

Questions?



Flash Memory Summit

nvm
EXPRESS®

