Linux NVMe Driver

Keith Busch

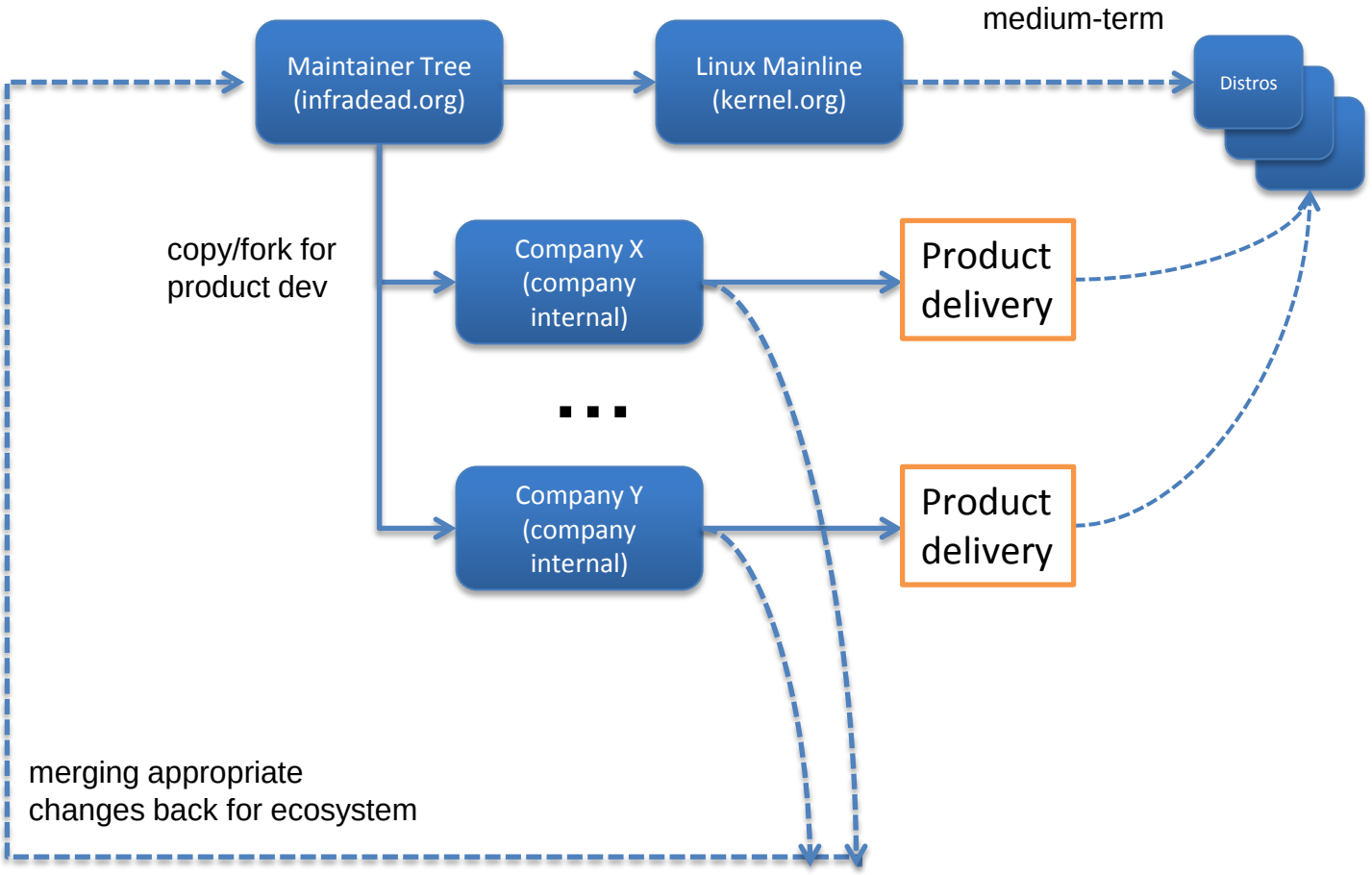
Software Engineer

Intel Corp

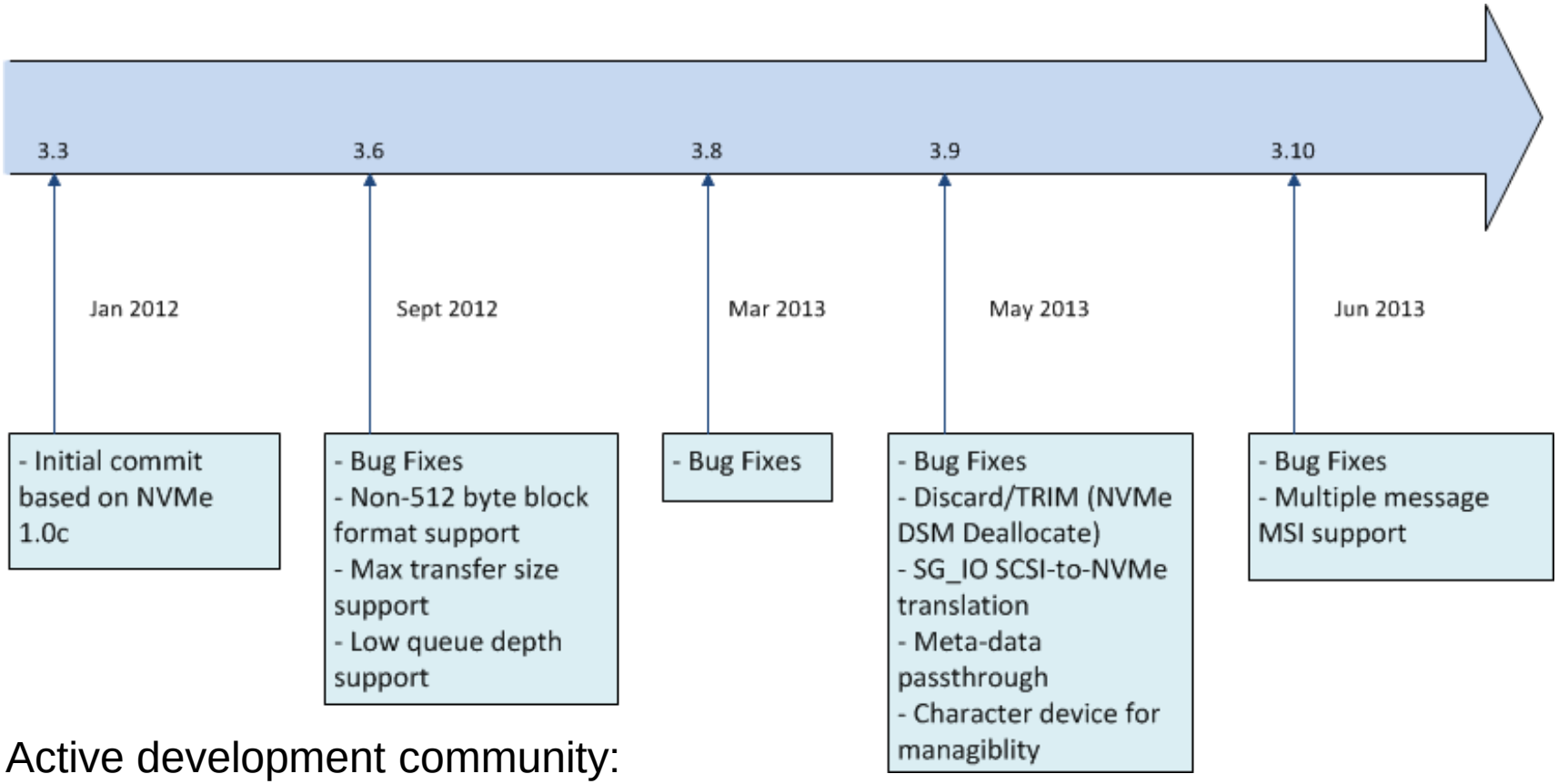
NVMe & Linux: Agenda

- Linux driver development process and history
- Implementation details
- NVMe inspired kernel optimizations
- How to get involved

NVMe: Linux community development process



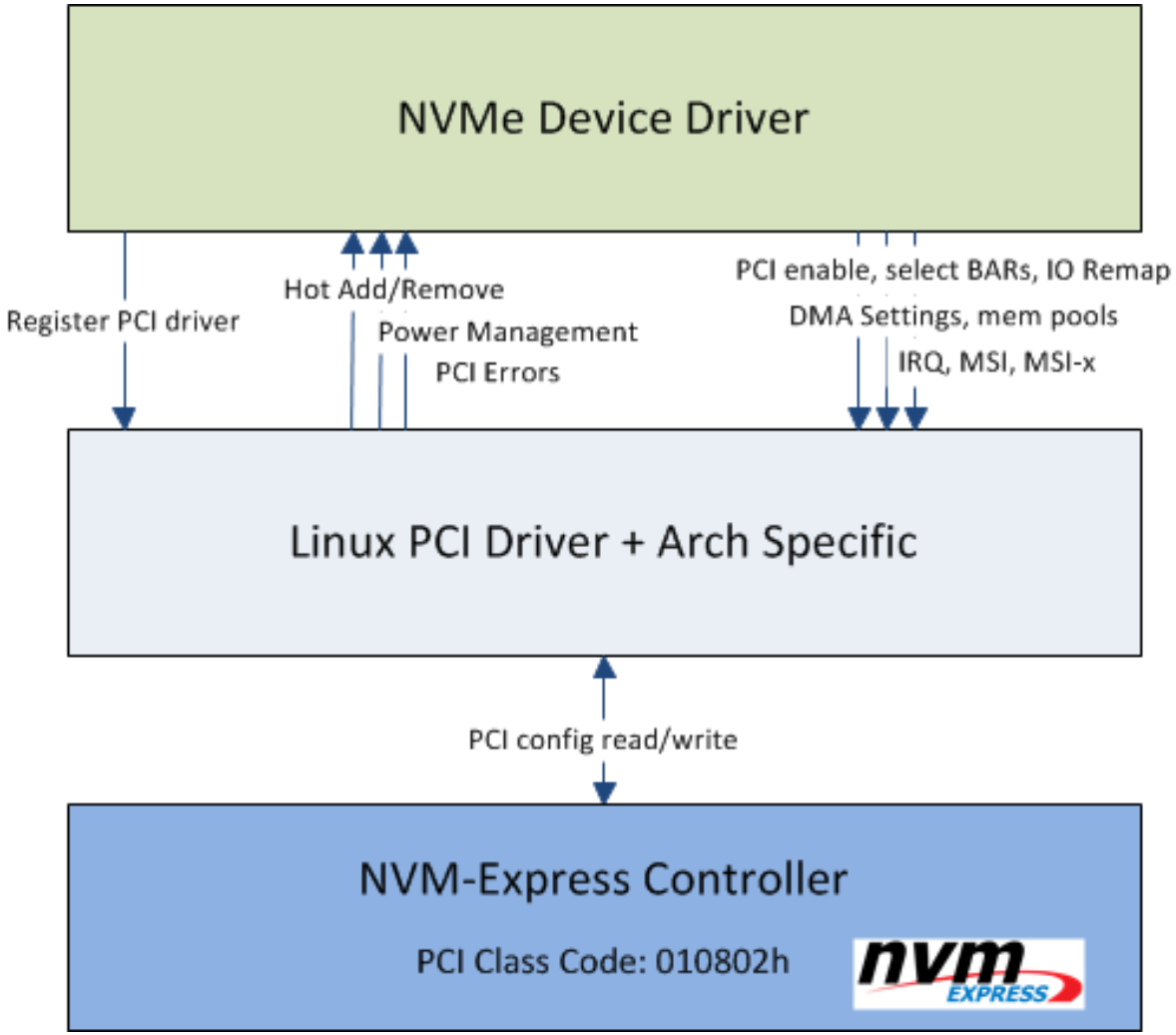
NVMe: Linux History



Active development community:

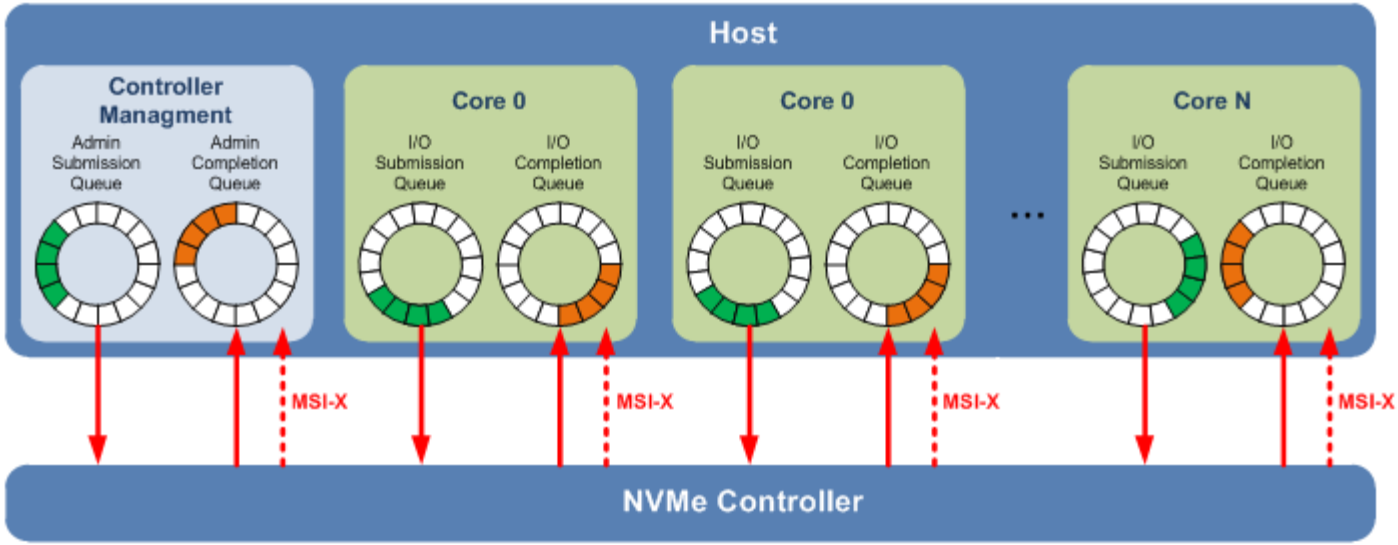
- 59 Change sets since initial commit
- 7 companies contributed patches
- Multiple Linux distributors ported driver to earlier kernel releases

NVMe: Linux PCI-e Driver

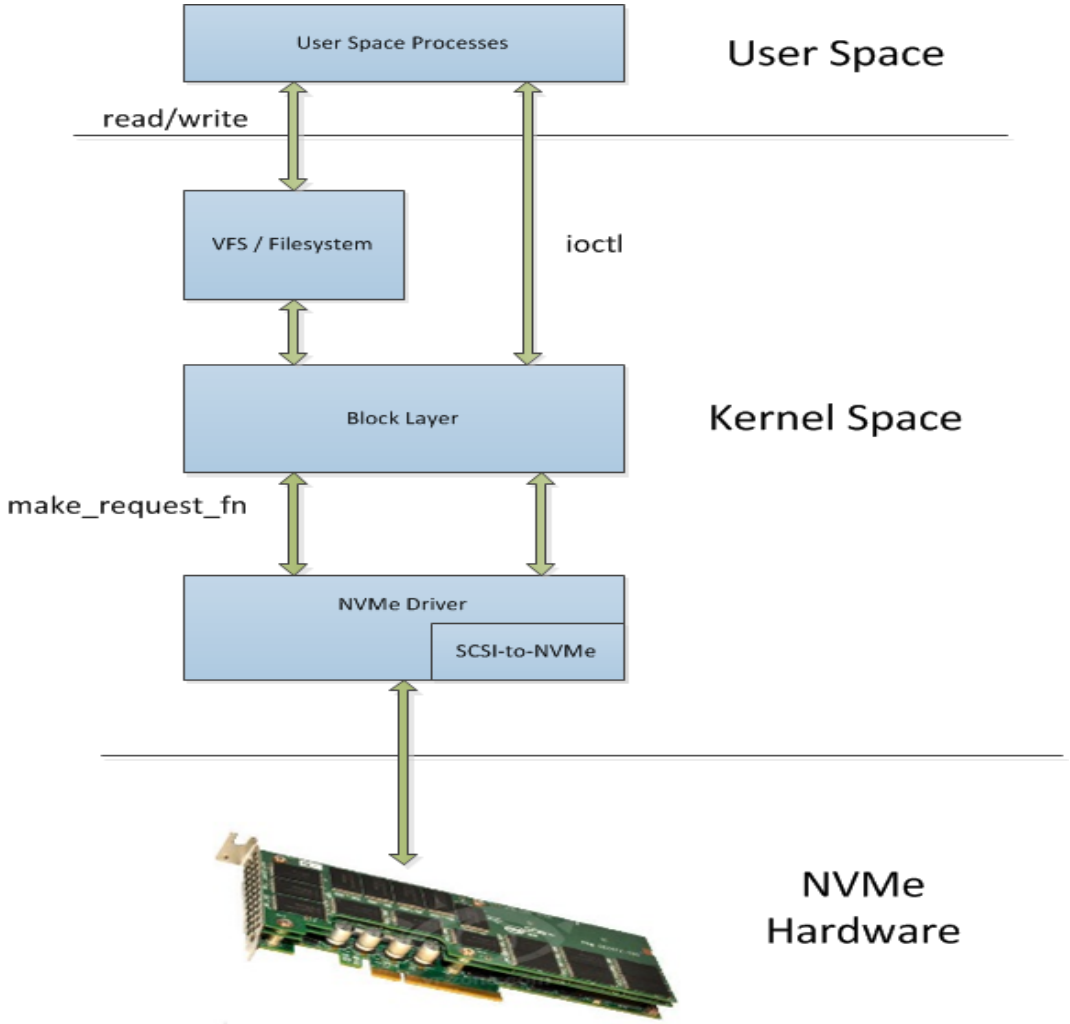


NVMe: Queue allocation details

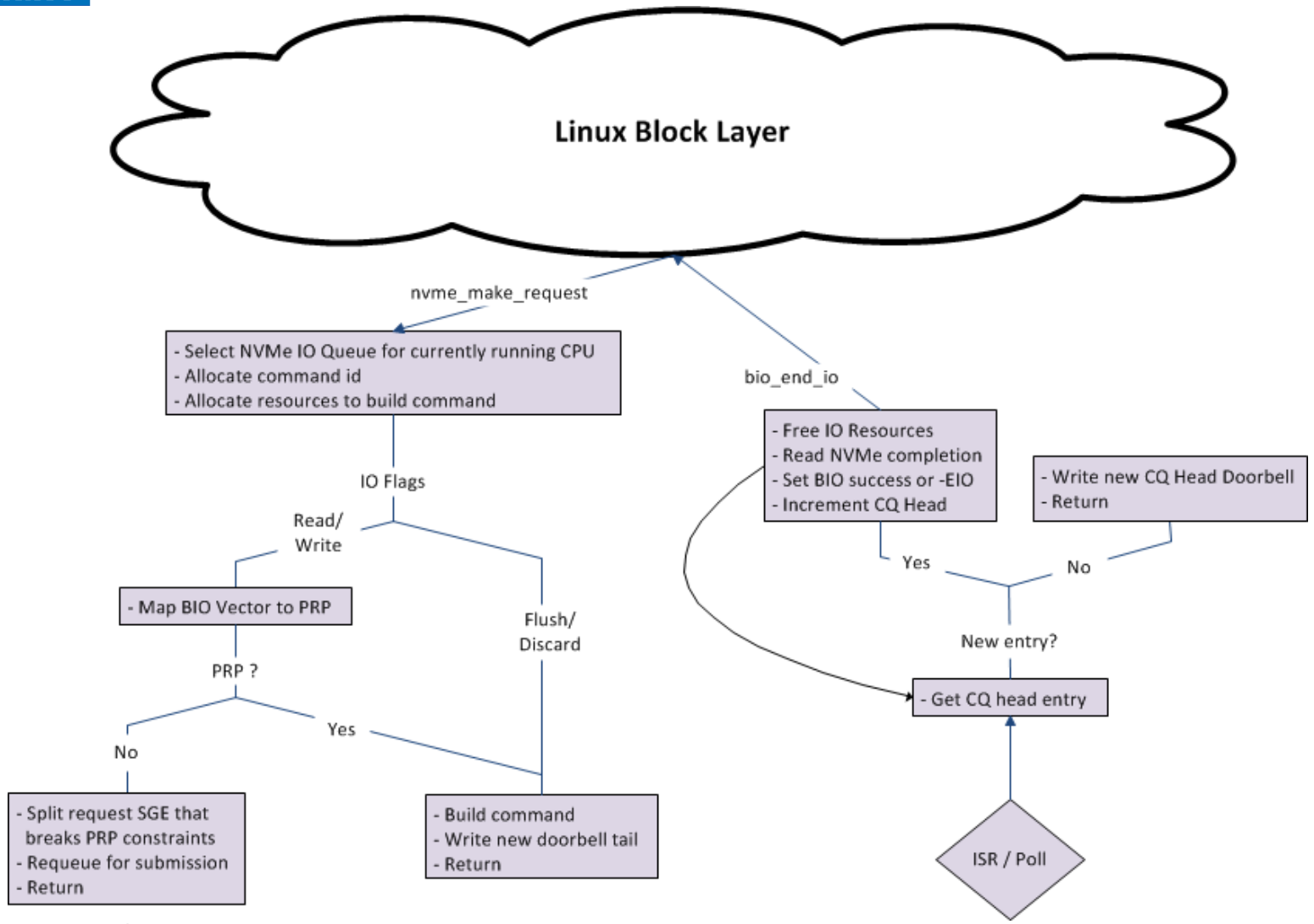
- Submission/Completion queue pairs
  - Round Robin Arbitration
  - One pair per CPU and assigned to that CPU
  - MSI-x interrupt affinity pinned to a CPU core per pair
    - Resort to MSI, then INTx, and finally polling if all else fails
  - Scalable: minimize lock contention, maximize cache hits



NVMe: Anatomy of Linux block software stack



NVMe: Detailed IO Process

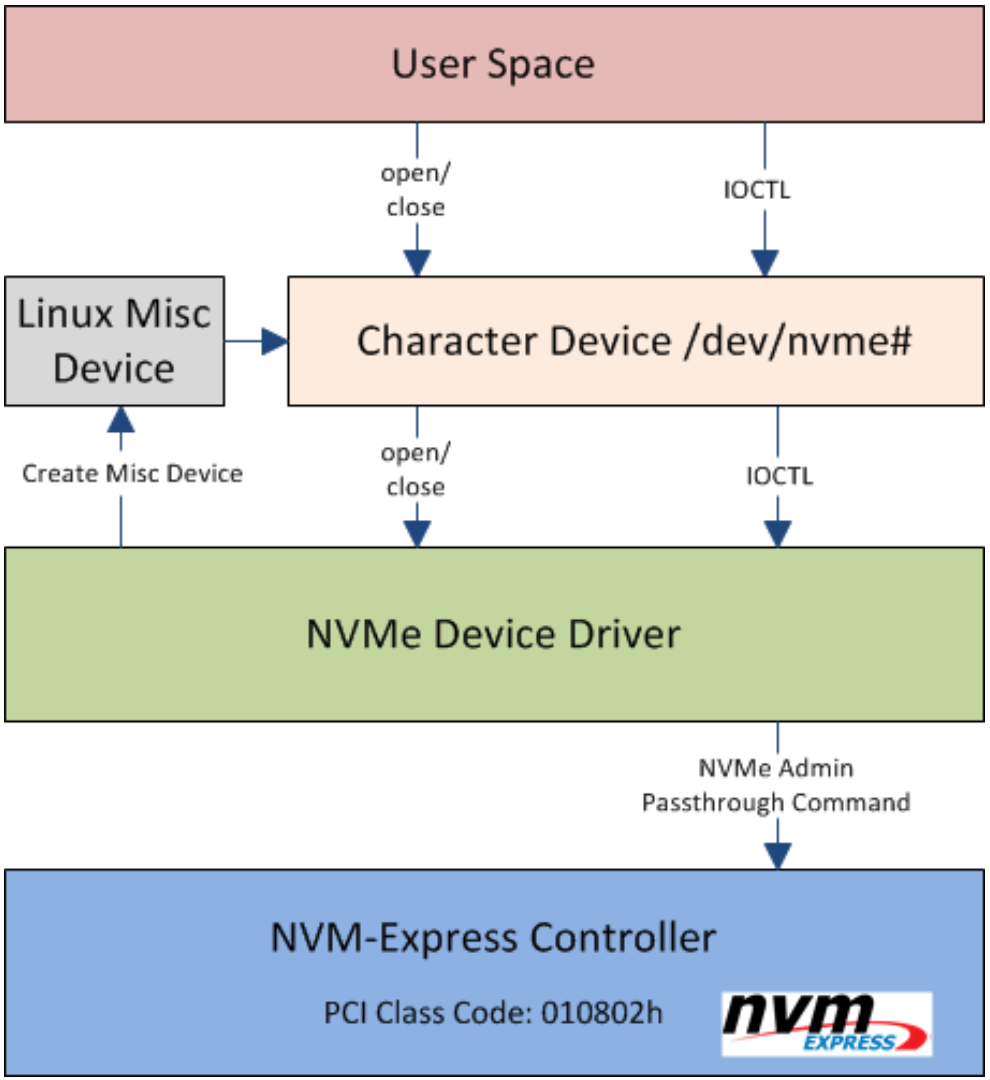


NVMe: Linux SCSI SG_IO IOCTL

For “Legacy” SCSI Management (not fast-path)

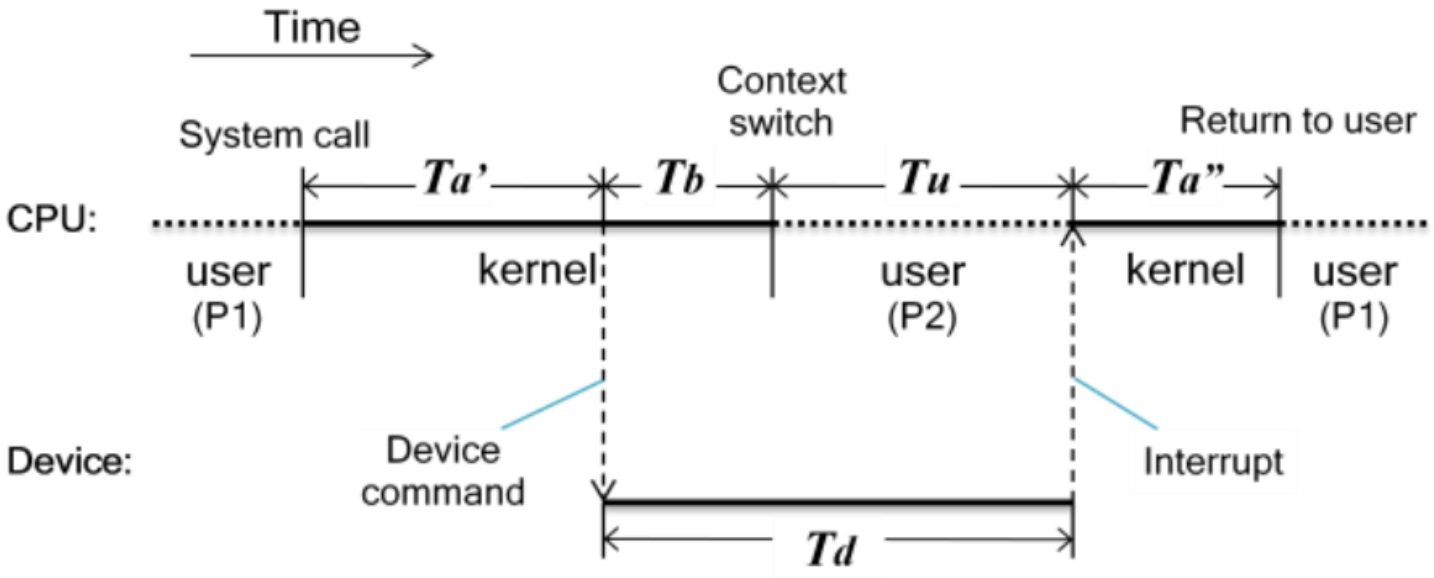
- Read/Write 6, 10, 12, 16
- Inquiry (Std, VPD 0, 80, 83, 86, B1)
- Mode Sense 10/16
- Mode Select 10/16
- Log Sense
- Read Capacity 10/16
- Report LUNS
- Request Sense
- Security Protocol In/Out
- Start Stop Unit
- Test Unit Ready
- Write Buffer
- Unmap

NVMe: Device management



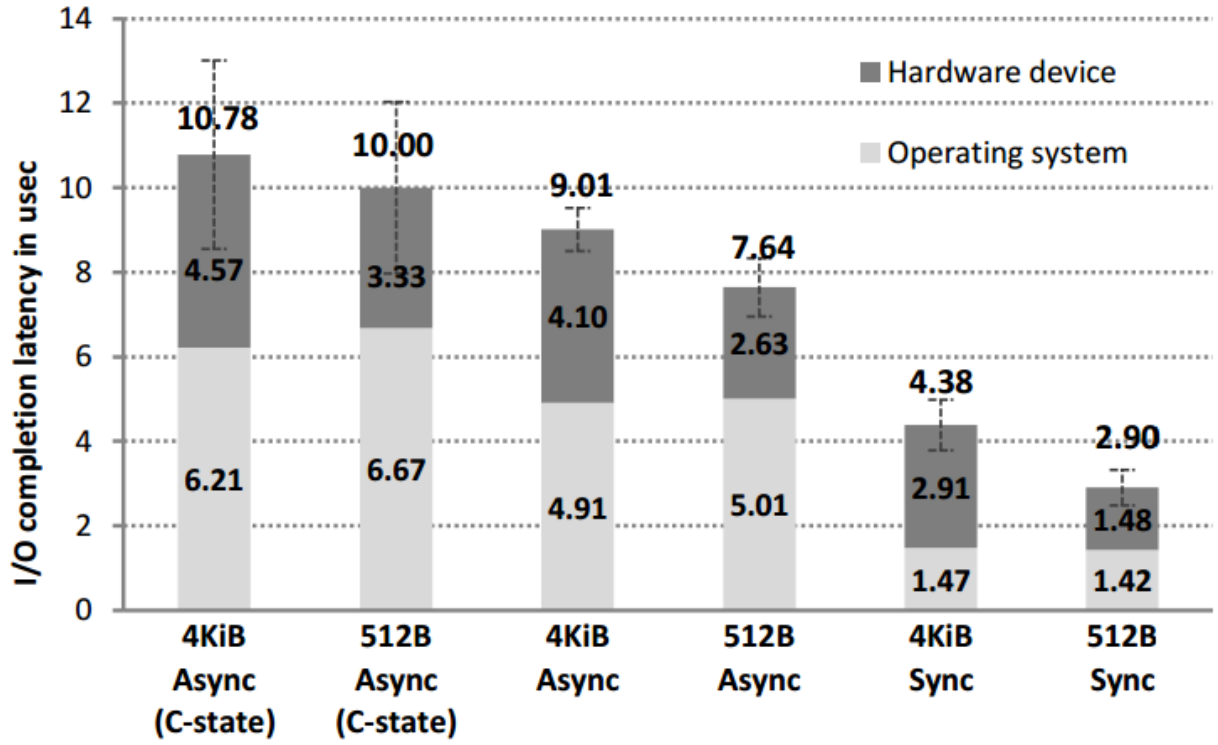
Linux block layer performance optimizations: beyond NAND

Asynchronous IO Latency sources:

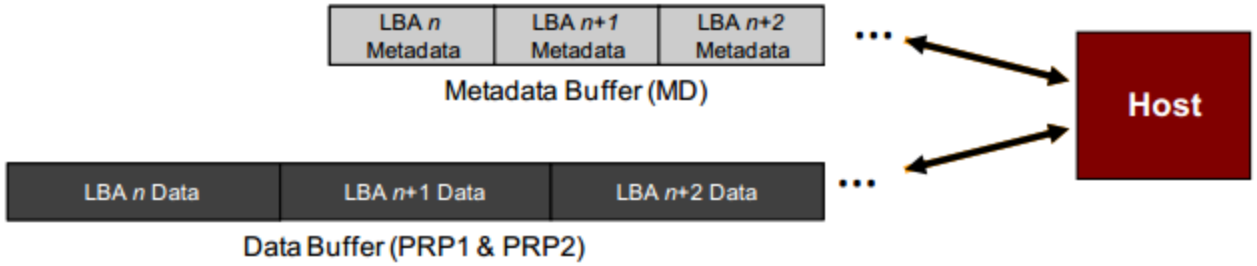


- For low latency devices, context switch and interrupt dominate user observed latency.

Linux block layer performance optimizations: beyond NAND

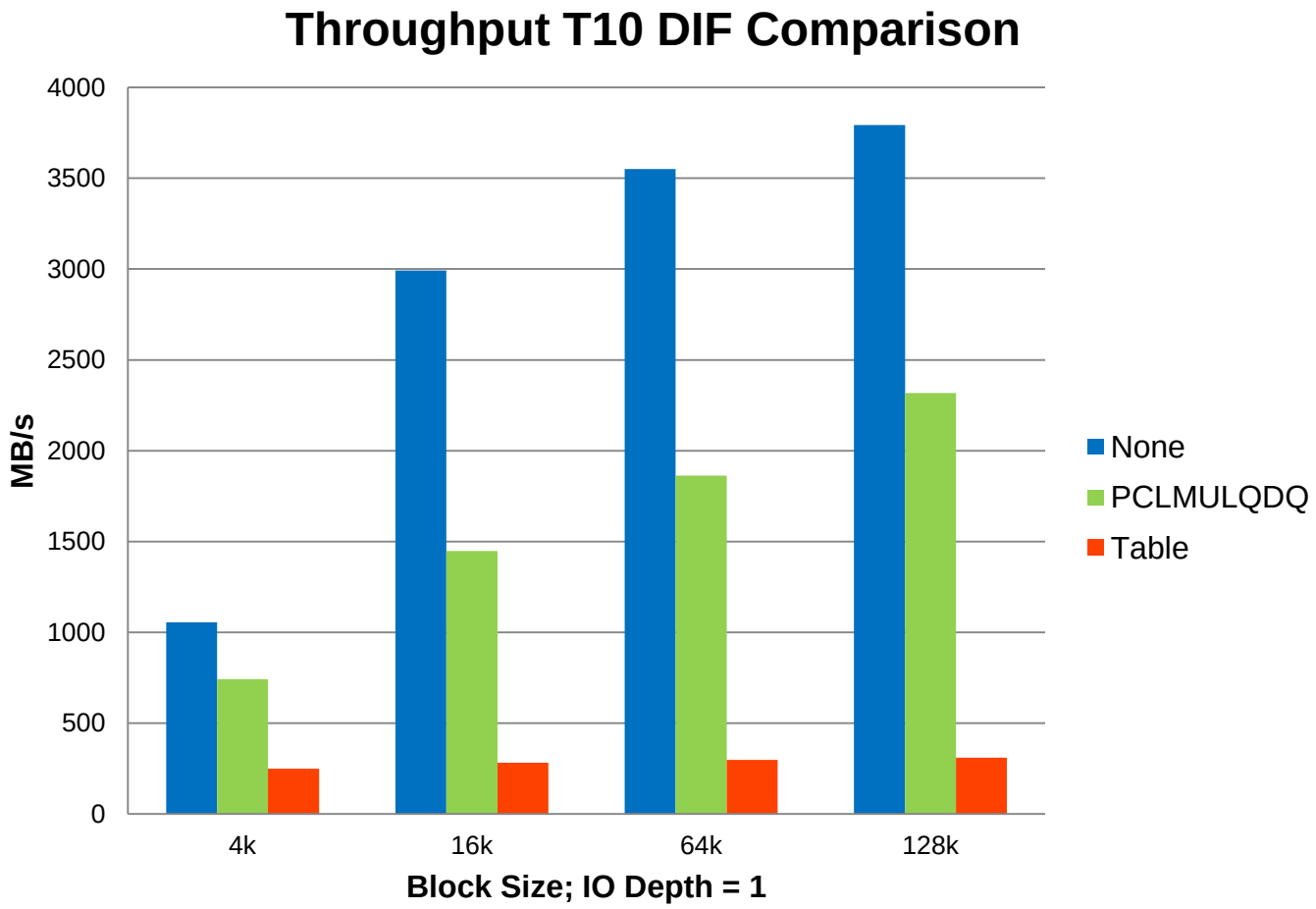


Linux performance optimizations: T10 DIF Protection Information



- Linux calculates CRC-16 Guard via table lookups and is **expensive**!
- x86-64 improvement: PCLMULQDQ; merged in linux crypto-dev tree

Linux performance optimizations: T10 DIF Protection Information

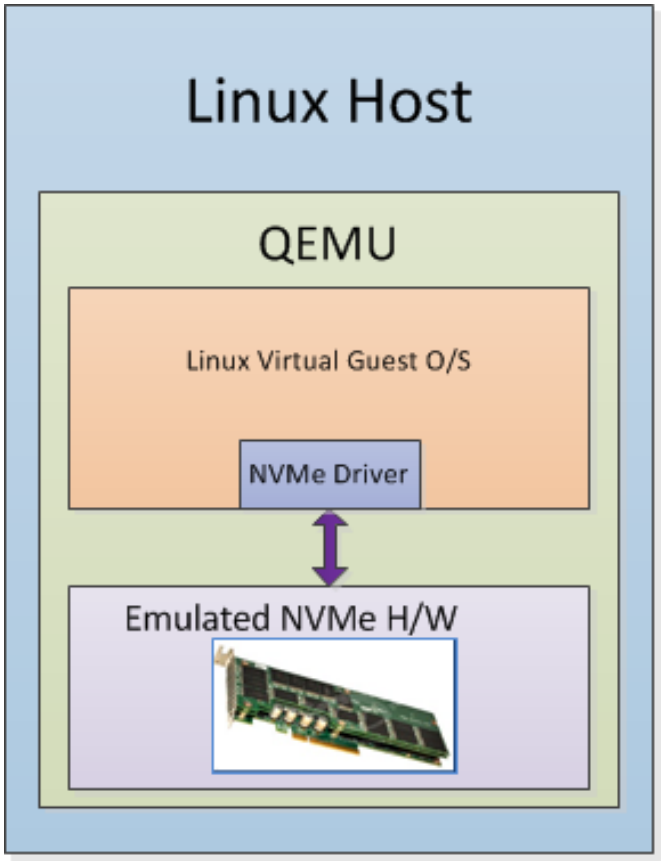


Linux NVMe: Get involved!

- Subscribe and contribute to mailing list: http://lists.infradead.org/mailman/listinfo/linux-nvme
- Clone, compile, and enhance driver: http://git.infradead.org/users/willy/linux-nvme.git
- Some TODO items:
  - Enhanced manageability via sysfs
  - Asynchronous events
  - Meta-data, T10 DIF/DIX
  - Power management
  - Performance enhancements/experiments
  - CPU hotplug
  - Advanced error handling
  - Enhanced PCI error handling
  - NVMe 1.1 spec updates
  - Device specific handling

Linux NVMe: Get involved!

- No hardware? No problem.
  - Machine emulator and virtualizer with NVMe support freely available from qemu.org
    - Good for testing features and basic functionality
    - Bad for analyzing performance and power characteristics



Questions:

keith.busch@intel.com

References

- NVM-Express
  http://nvmexpress.org/
- Linux NVMe Repository:
  http://git.infradead.org/users/willy/linux-nvme.git
- Linux NVMe Mailing list:
  http://merlin.infradead.org/pipermail/linux-nvme/
- When Polling is Better than Interrupt:
  https://www.usenix.org/system/files/conference/fast12/yang.pdf
- Block polling in Linux:
  http://lwn.net/SubscriberLink/556244/309ec42e8b9a4fcf/
- CRC-16 T10 DIF PCLMULQDQ:
  https://lkml.org/lkml/2013/5/1/449

NVMe OFA Open Source Windows Driver

Kwok Kong
Director of Software Engineering
PMC-Sierra

Agenda

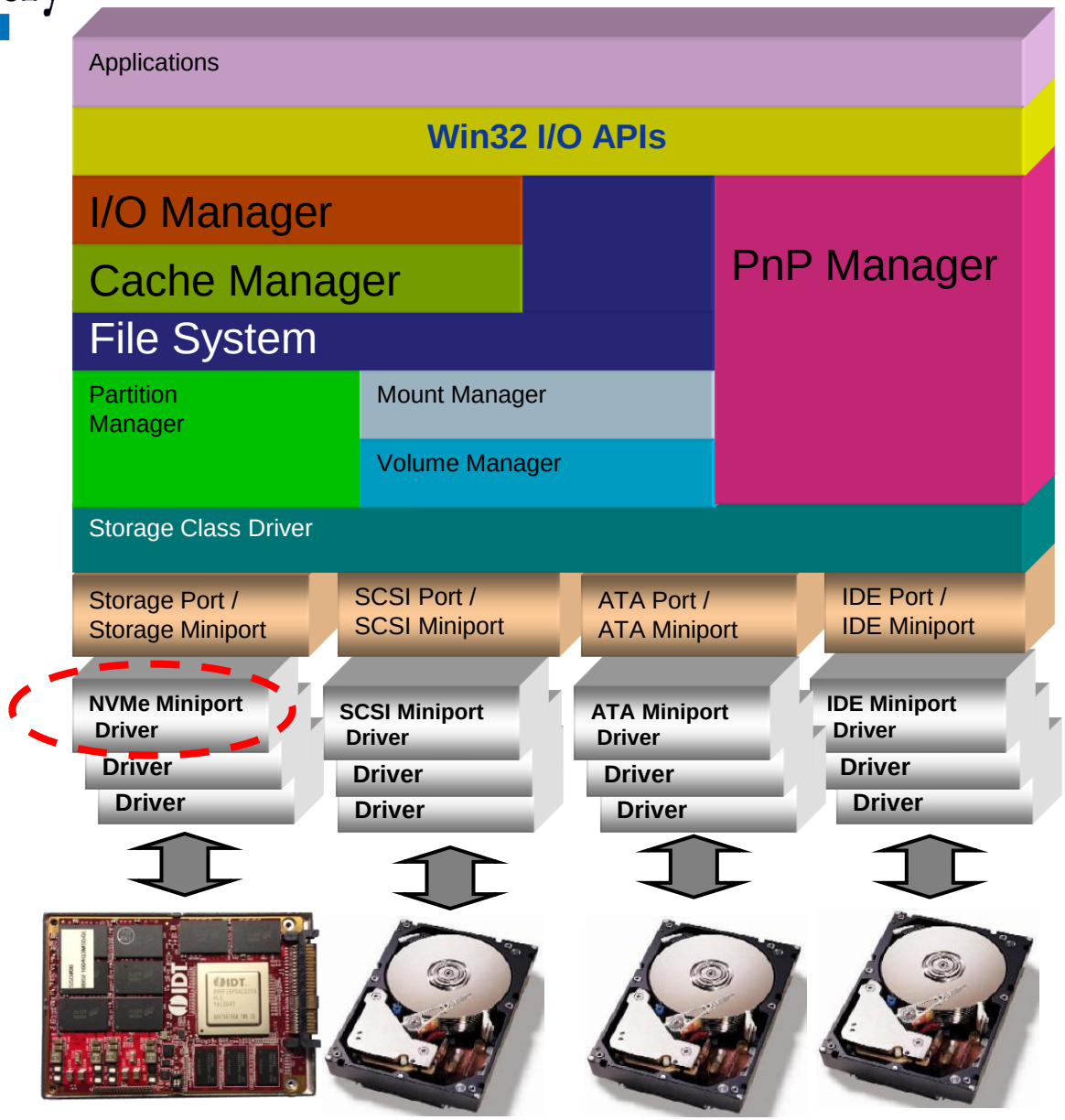
- Status Update
- Driver Architecture
- Driver Features
- Future Features

Status Update

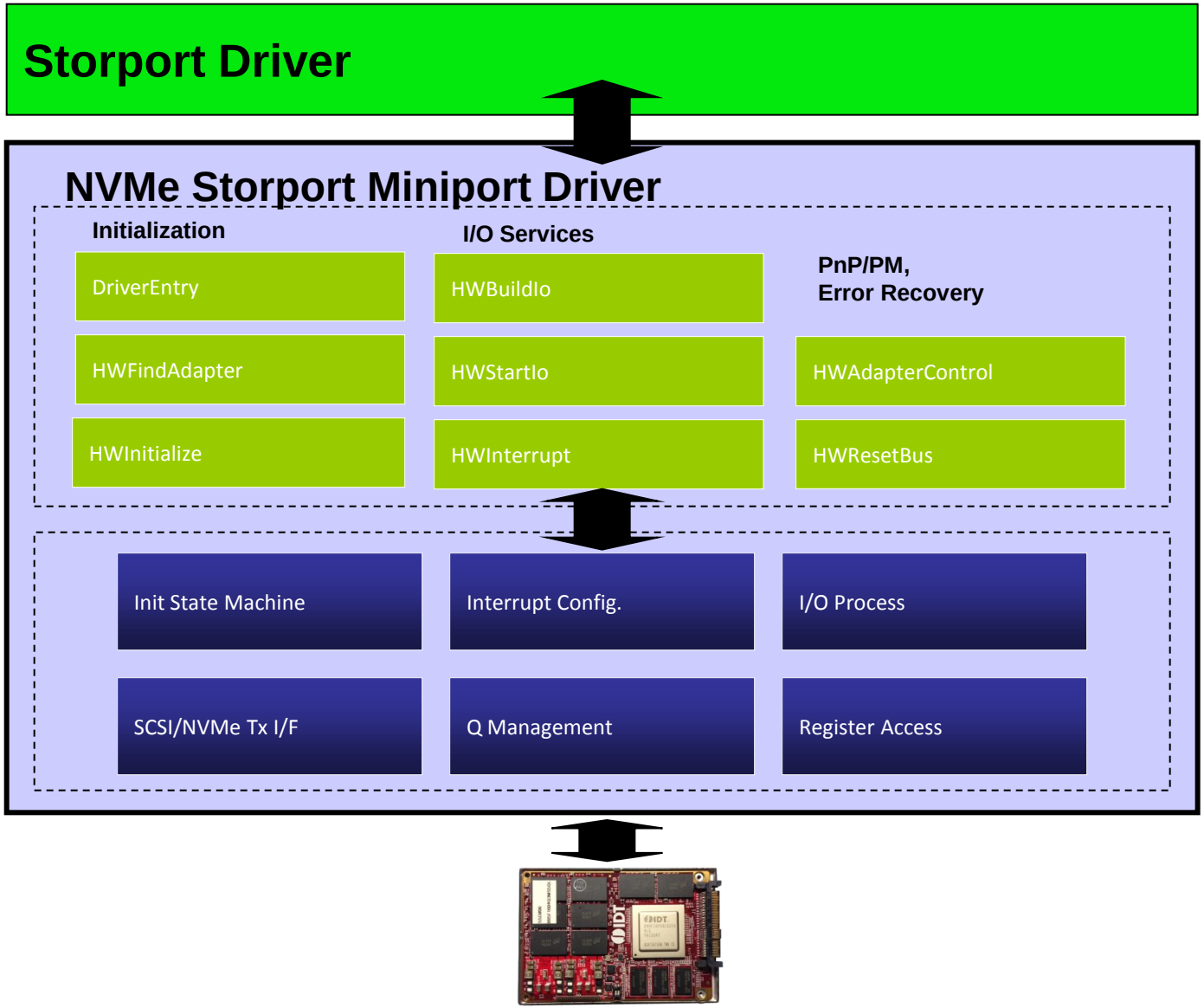
Release 1	• Q2 2012 (released) • 64-bit support on Windows* 7, Windows* Server 2008 R2 • Mandatory features
Release 1.1	• Q4 2012 (released) • Added 64-bit support Windows* 8 • Public IOCTLs and Windows* 8 Storport updates
Release 1.2	• Q2 2013 (released) • Added 64-bit support on Windows* Server 2012 • Signed executable drivers
Release 1.3	• Target: Q4 2013 • Added 32-bit support on all supported OS versions • End-to-end Data Protection

Three major releases of the Windows* OFA community driver since 2012.
Code contributions from Huawei, IDT, Intel, LSI, and SanDisk.

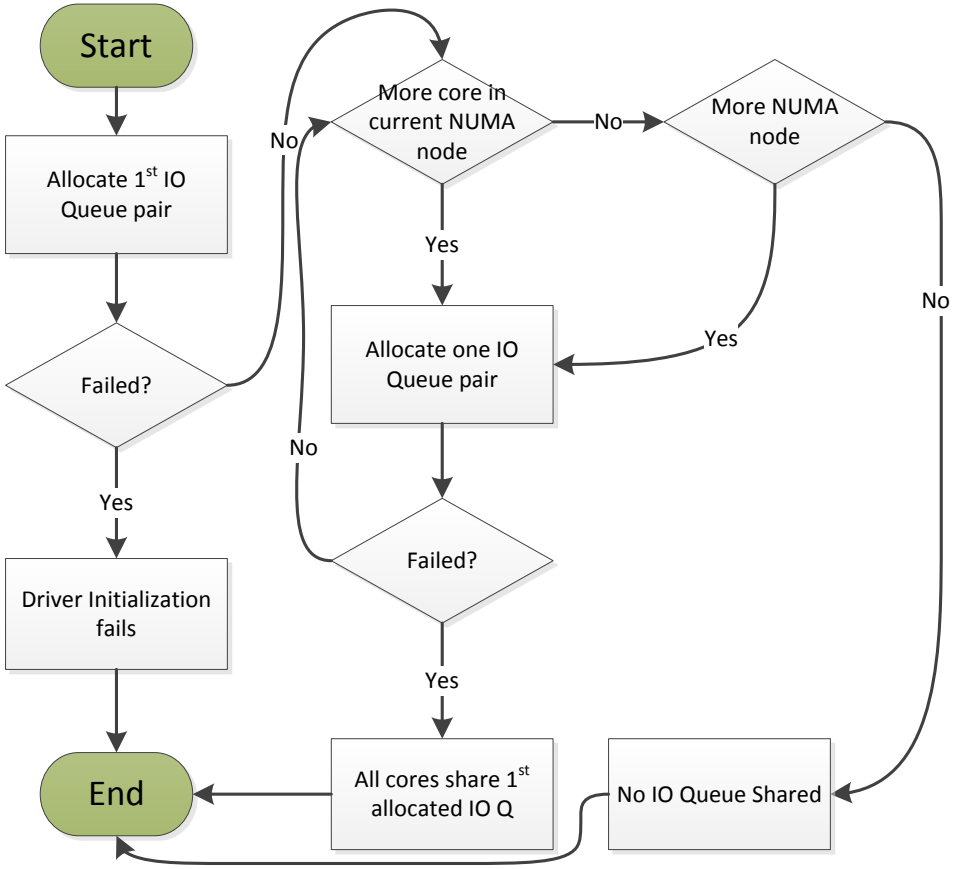
Windows Storage Architecture



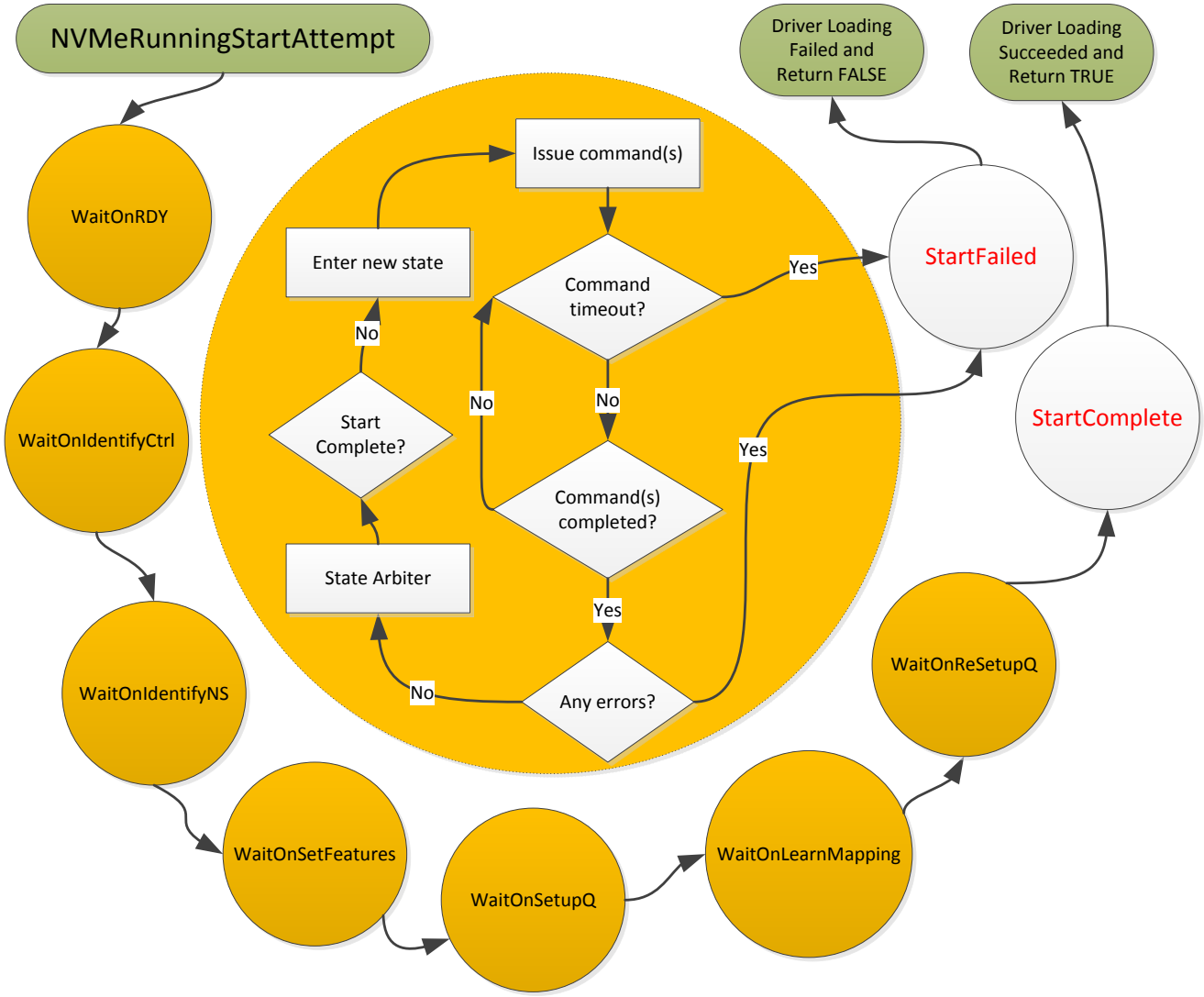
Driver Architecture



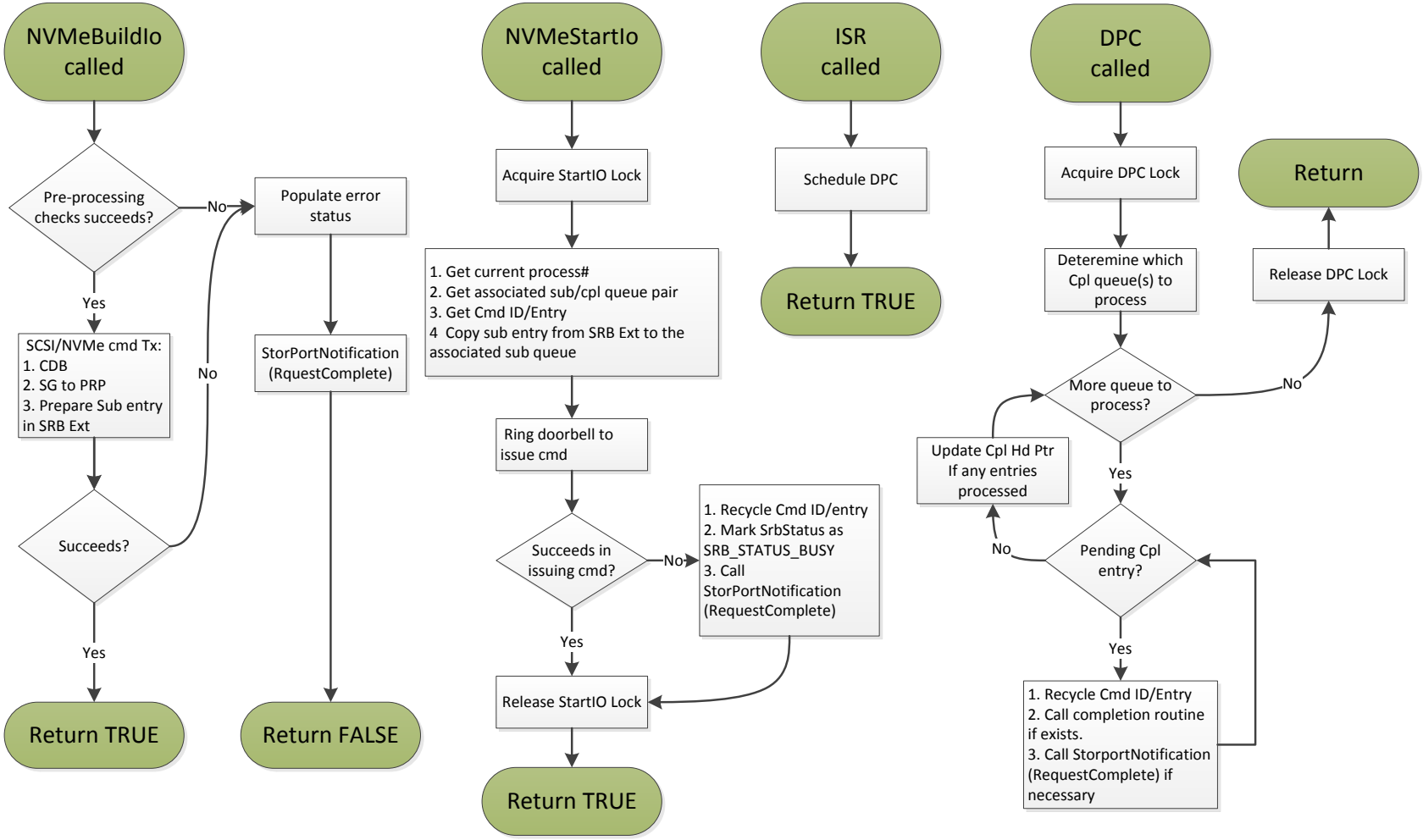
IO Queue Allocation Diagram



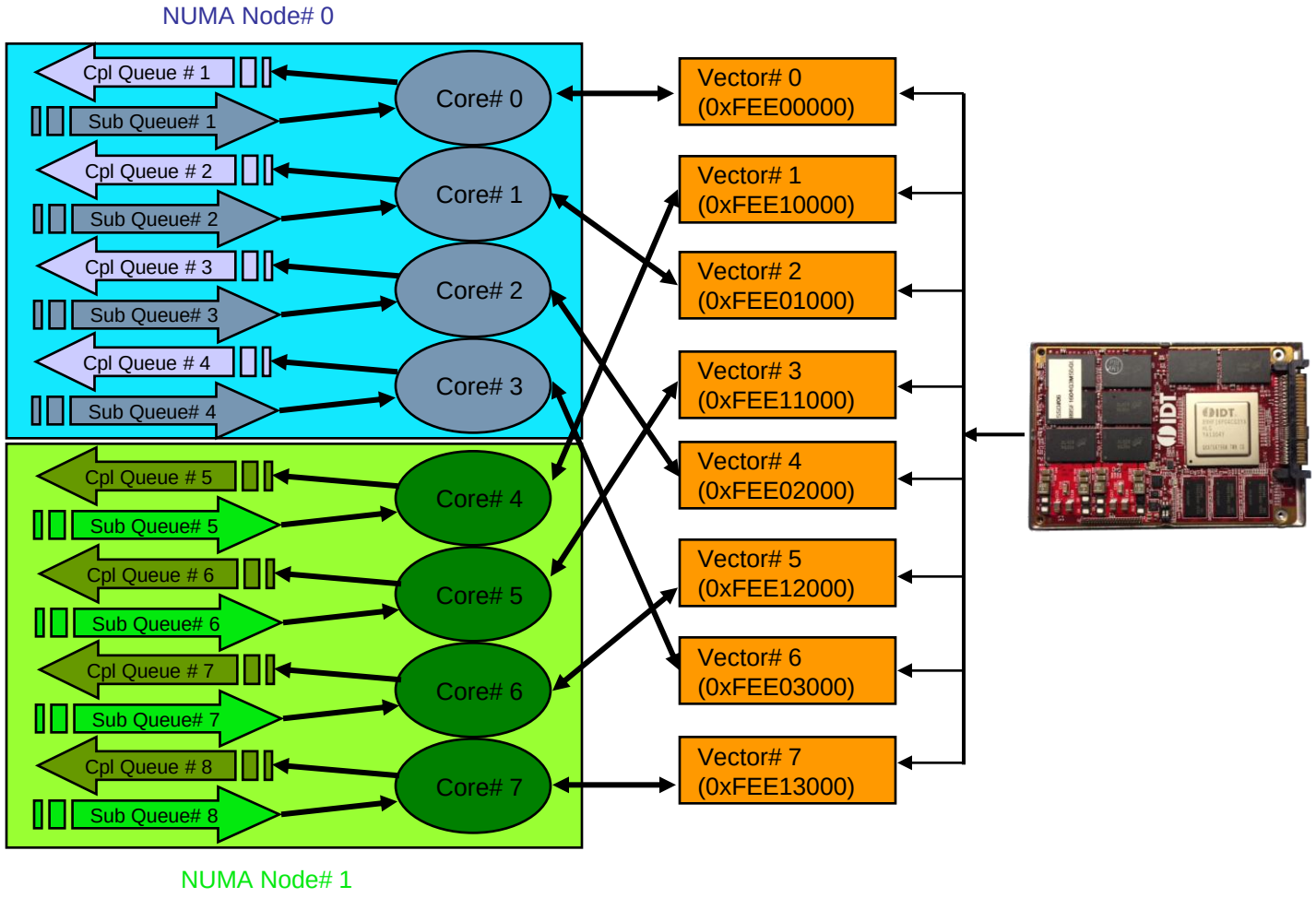
Driver Initialization State Machine



IO Process Diagram (Read)



Queue-Core-Vector Mappings



System Features Overview

Features	Supported
Windows Versions (64-bit only)	7, 8, Server 2008R2, Server 2012
NUMA Optimized Queues and Memory	Dedicated IO queues per CPU core Single Admin Queue for all CPU cores Queue memory allocated local to NUMA Node
Interrupt	MSI-X (Vectors mapped to NUMA optimized IO Queue Pairs) MSI INTX
Queue Arbitration / Priority	Round Robin only 1:1 mapping between Submission and Completion queues
Pass Through	Yes (with DeviceIoControl())
Multiple Namespaces	Up to 16, Mapped to Bus 0, Tgt 0, LUN 0-15
Registry Parameters	Name Space, Max Transfer Size, Admin Queue Size, IO Queue size, Interrupt Coalescing Time/Entries)
End to End Protection	No

Admin command Sets

Commands	Native Support	Pass Through
Delete I/O Submission Queue	Yes	No
Create I/O Submission Queue	Yes	No
Get Log Page	Yes	Yes
Delete I/O Completion Queue	Yes	No
Create I/O Completion Queue	Yes	No
Identify	Yes	Yes
Abort	No	No
Set Features	Yes	Yes
Get Features	Yes	Yes
Asynchronous Event Request	Yes	No
Firmware Activate	No	Yes
Firmware Image Download	No	Yes
Format NVM	No	Yes
Security Send	Yes	Yes
Security Receive	Yes	Yes
Vendor specific	No	Yes

NVM command Sets

Commands	Native Support	Pass Through
Flush	Yes	Yes
Write	Yes	No
Read	Yes	No
Write Uncorrectable	No	Yes
Compare	No	Yes
Dataset Management	Yes (Deallocate)	Yes
Vendor specific	No	Yes

Future Features

- 1.3 Release (end of 2013)
  - Windows 32-bit
  - End to End Protection
  - Hibernation Support on Boot Drive
  - NVM Format Enhancement
- 2014 and Beyond
  - NVMe 1.1 Features

Need you contribution to write the future

OFA NVMe Driver Working Group

- Founding Companies
  - PMC Sierra (IDT) – Chairperson
  - Intel – Code maintainer
  - LSI (SandForce)
- To Contribute
  - Join the mailing list
  - Email your patch to the WG mailing list
  - Code is checked in if approved by two out of three founding companies

Useful Information

- Join the Driver Mailing List to Contribute
  - http://lists.openfabrics.org/cgi-bin/mailman/listinfo/nvmewin
- Development Tools Description
  - https://www.openfabrics.org/developer-tools/nvme-windows-development.html
- Driver Source Code
  - http://www.openfabrics.org/svnrepo/nvmewin/