Linux NVM-Express

Keith Busch
Software Engineer
Intel

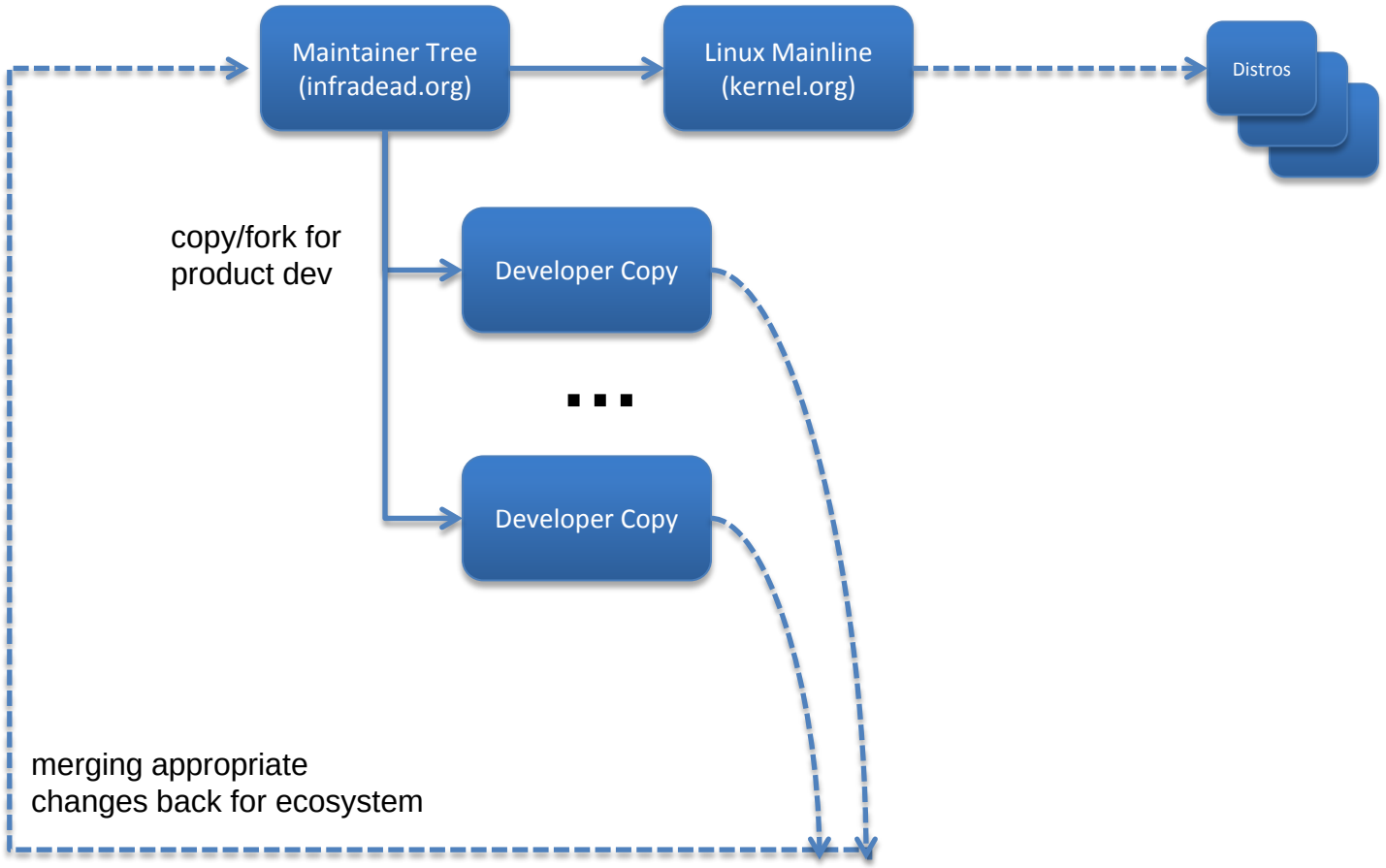
Agenda

- Driver Development History
- Current State
- Kernel Enhancements
- Future Work
- Getting Involved

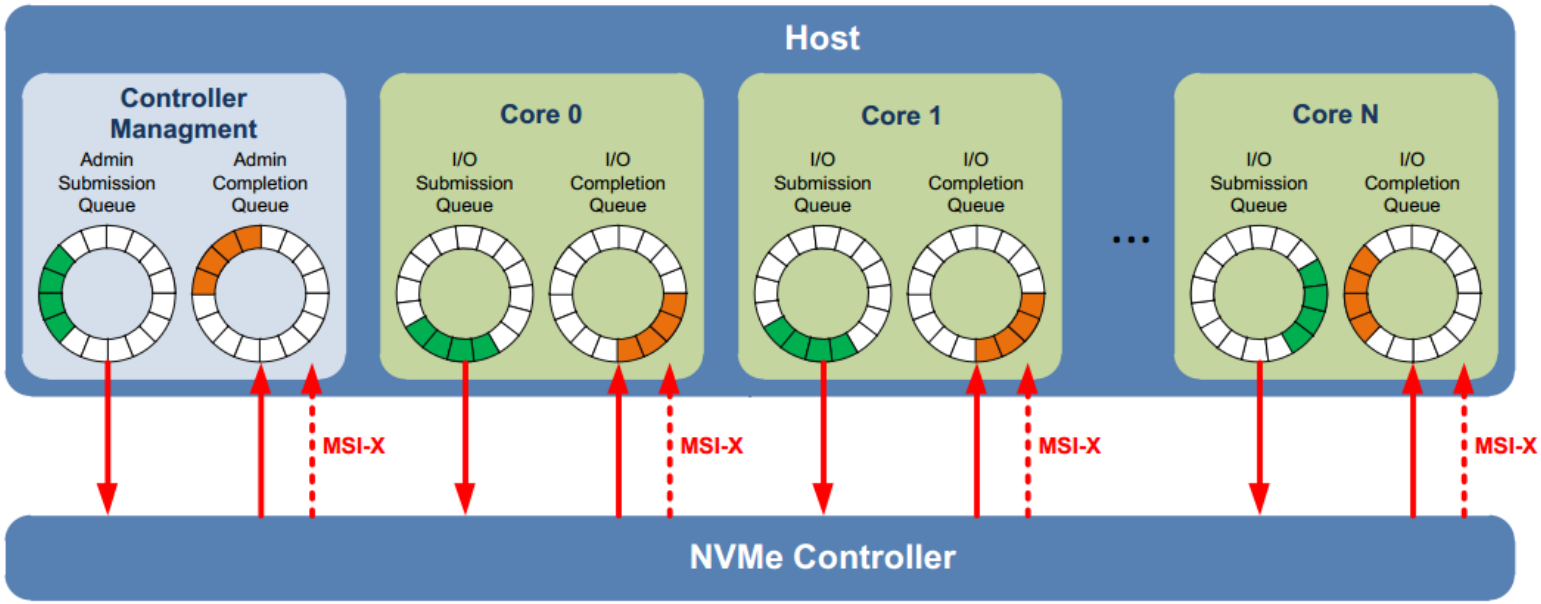
In the beginning …

- Originally revealed with 1.0 spec, March 2011, contributed by Matthew Wilcox
- Merge to mainline January, 2012 with 3.3.
- Since has seen over 150 commits from 25 individuals contributing bug fixes, features and enhancements.

Getting new updates Upstream

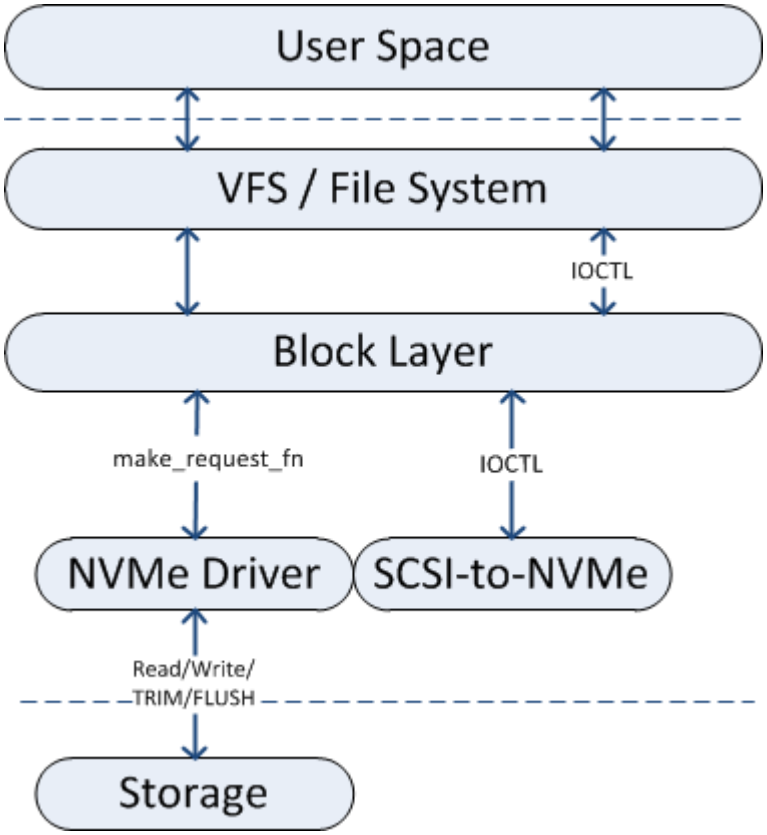


Driver Details: Queue Allocation



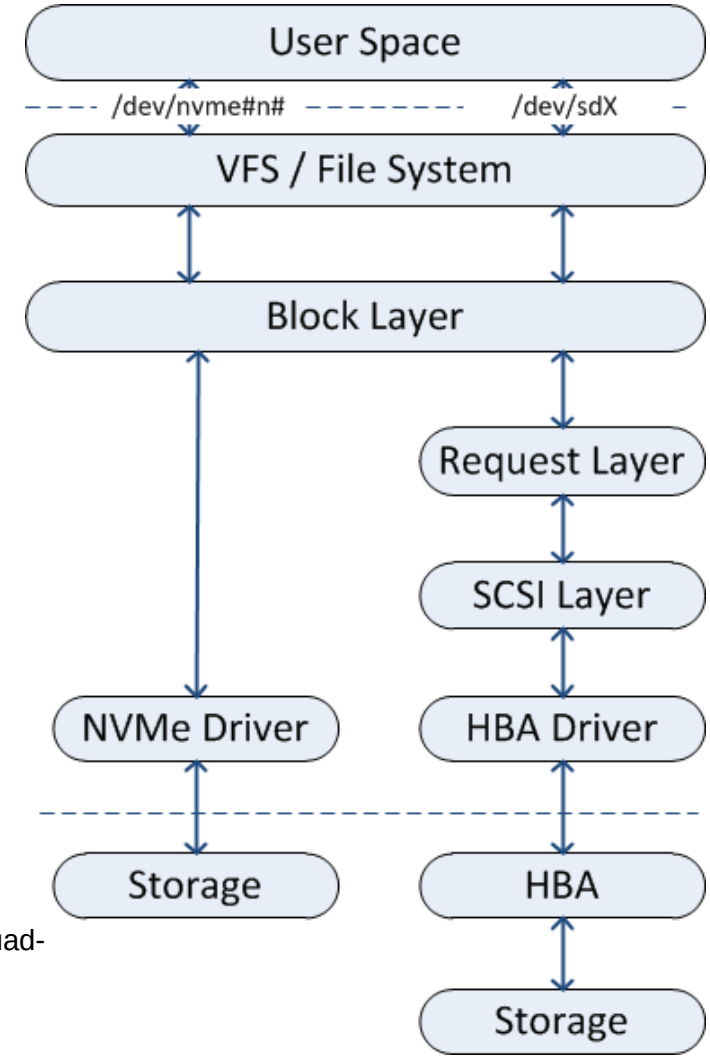
- Ideal case: one SQ/CQ pair per cpu core
- MSI-x IRQ affinity assigned to CPU associated Queue

Driver Details: Block Stack Anatomy



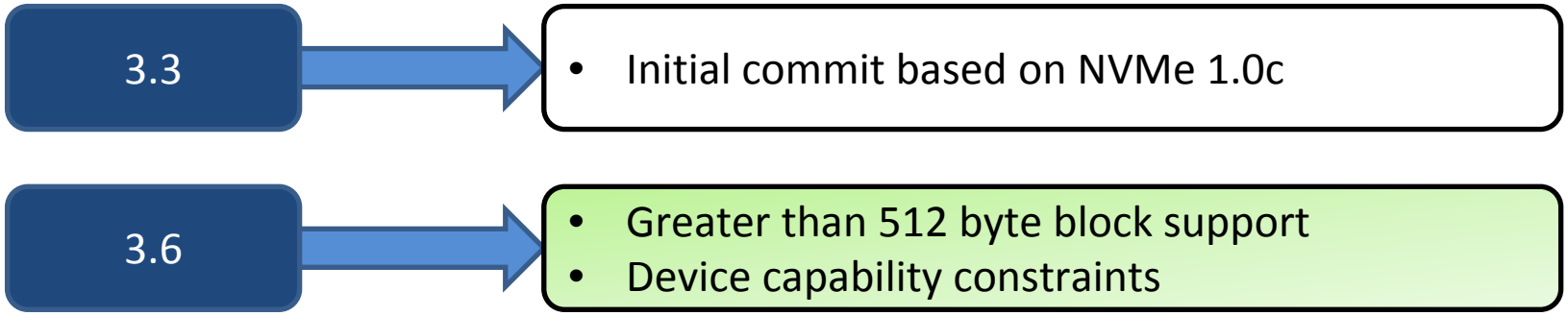
Storage Stack Comparison

- SAS vs. NVMe Device:
- Latency and CPU utilization reduced by 50+%*:
  - NVMe: 2.8us, 9,100 cycles
  - SAS: 6.0us, 19,500 cycles

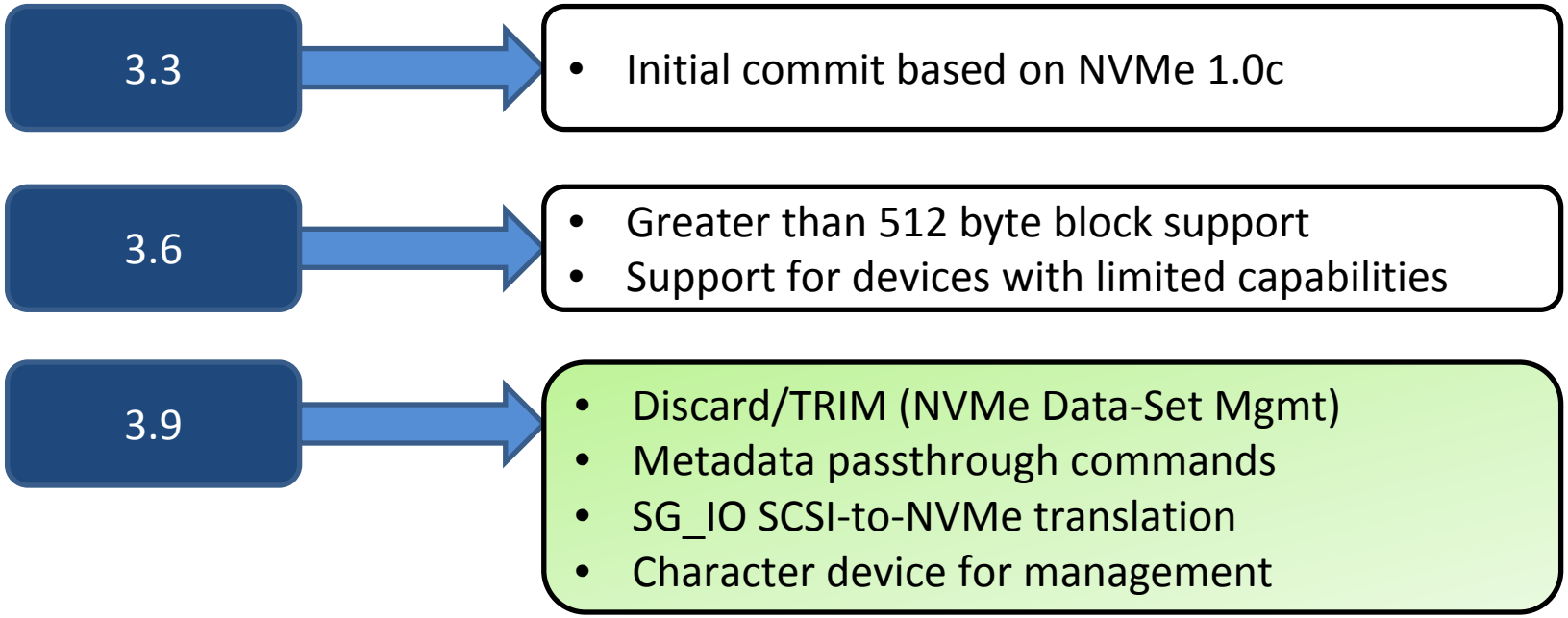


* Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux

Kernel 3.6 …


3.3
Initial commit based on NVMe 1.0c
3.6
Greater than 512 byte block support
Device capability constraints

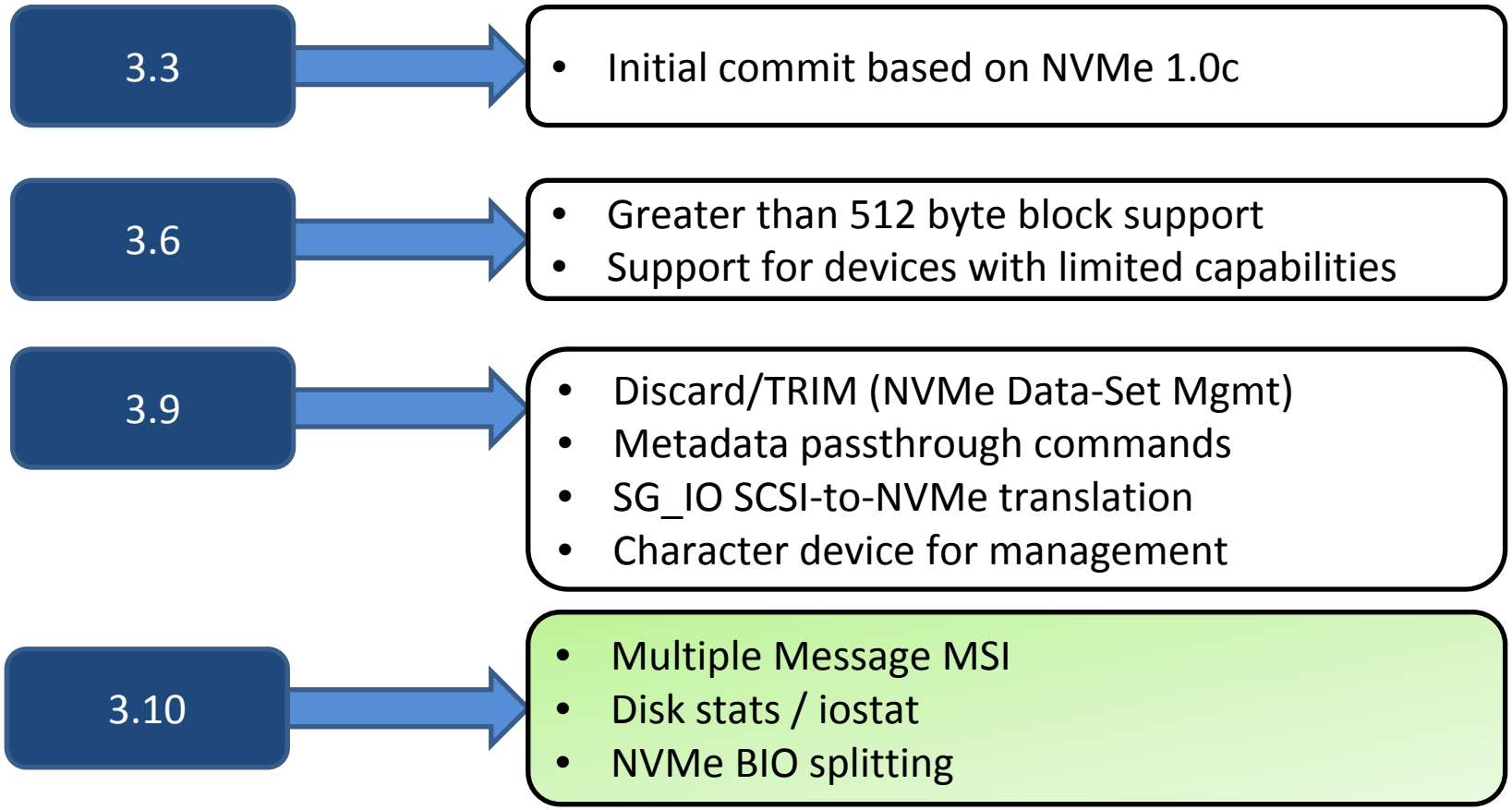
Kernel 3.9 …


3.3
Initial commit based on NVMe 1.0c
3.6
Greater than 512 byte block support
Support for devices with limited capabilities
3.9
Discard/TRIM (NVMe Data-Set Mgmt)
Metadata passthrough commands
SG_IO SCSI-to-NVMe translation
Character device for management

3.9: SCSI-NVMe SG_IO

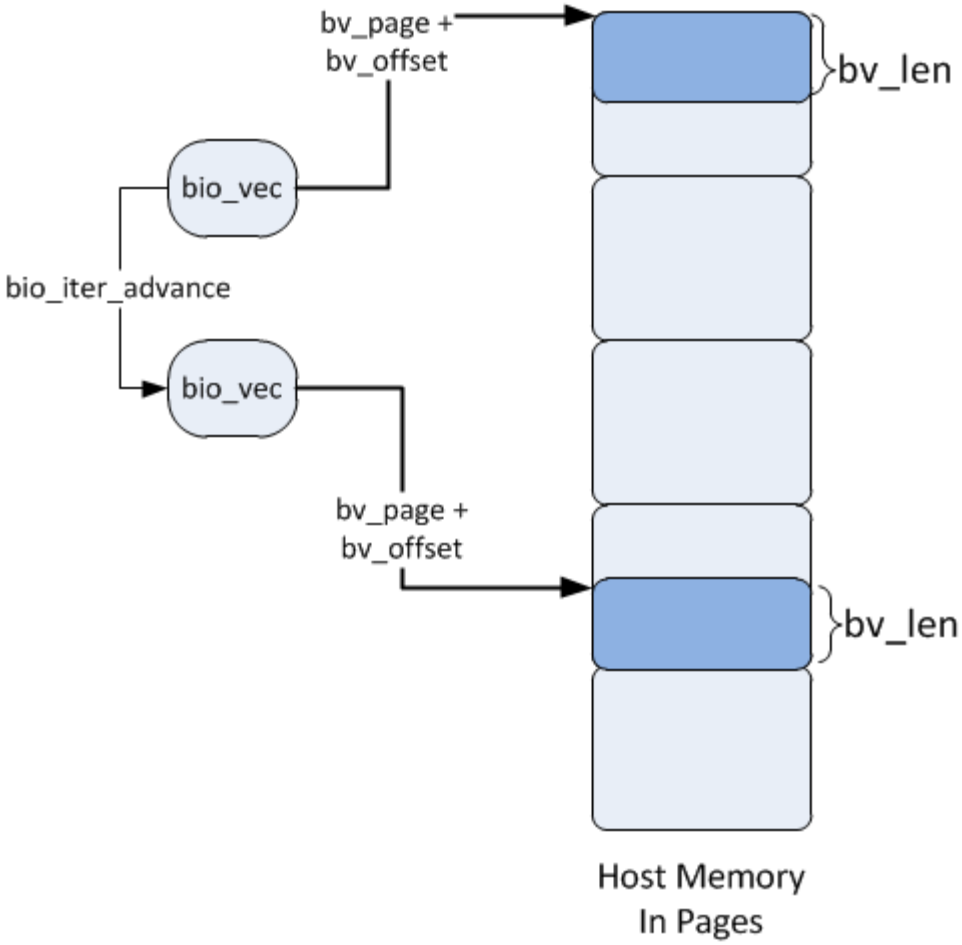
- Read/Write 6, 10, 12, 16
- Inquiry
- Mode Sense 10/16
- Mode Select 10/16
- Log Sense
- Read Capacity 10/16
- Report LUNS
- Request Sense
- Security Protocol In/Out
- Start Stop Unit
- Test Unit Ready
- Write Buffer
- Unmap

Kernel 3.10 …


3.3
Initial commit based on NVMe 1.0c
3.6
Greater than 512 byte block support
Support for devices with limited capabilities
3.9
Discard/TRIM (NVMe Data-Set Mgmt)
Metadata passthrough commands
SG_IO SCSI-to-NVMe translation
Character device for management
3.10
Multiple Message MSI
Disk stats / iostat
NVMe BIO splitting

3.10: BIO Splitting

- Not all I/O vectors can be mapped to an NVMe command's PRP list
- Requires virtually contiguous buffers

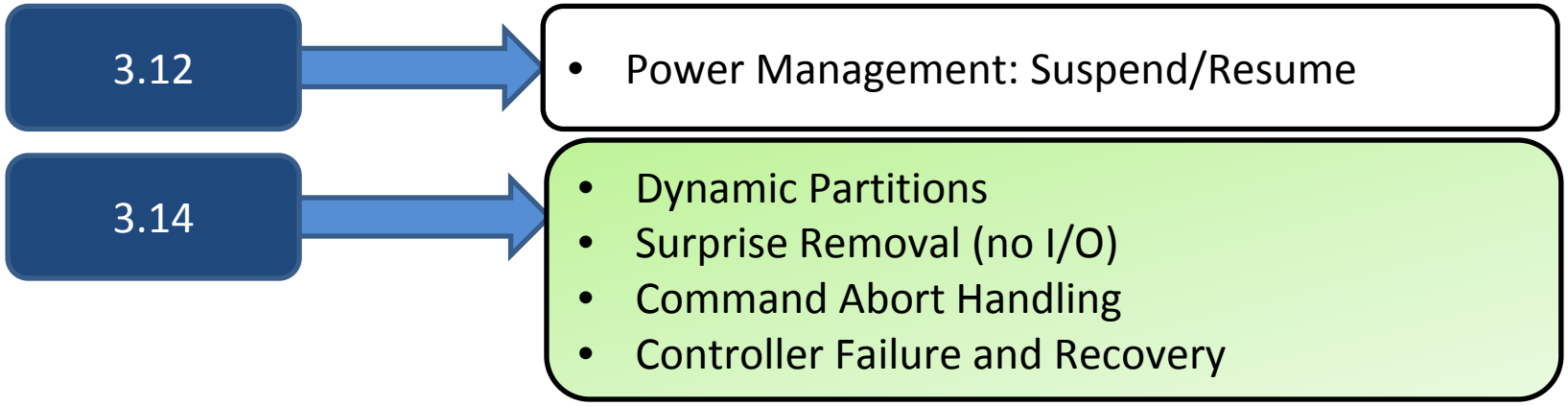


Kernel 3.12

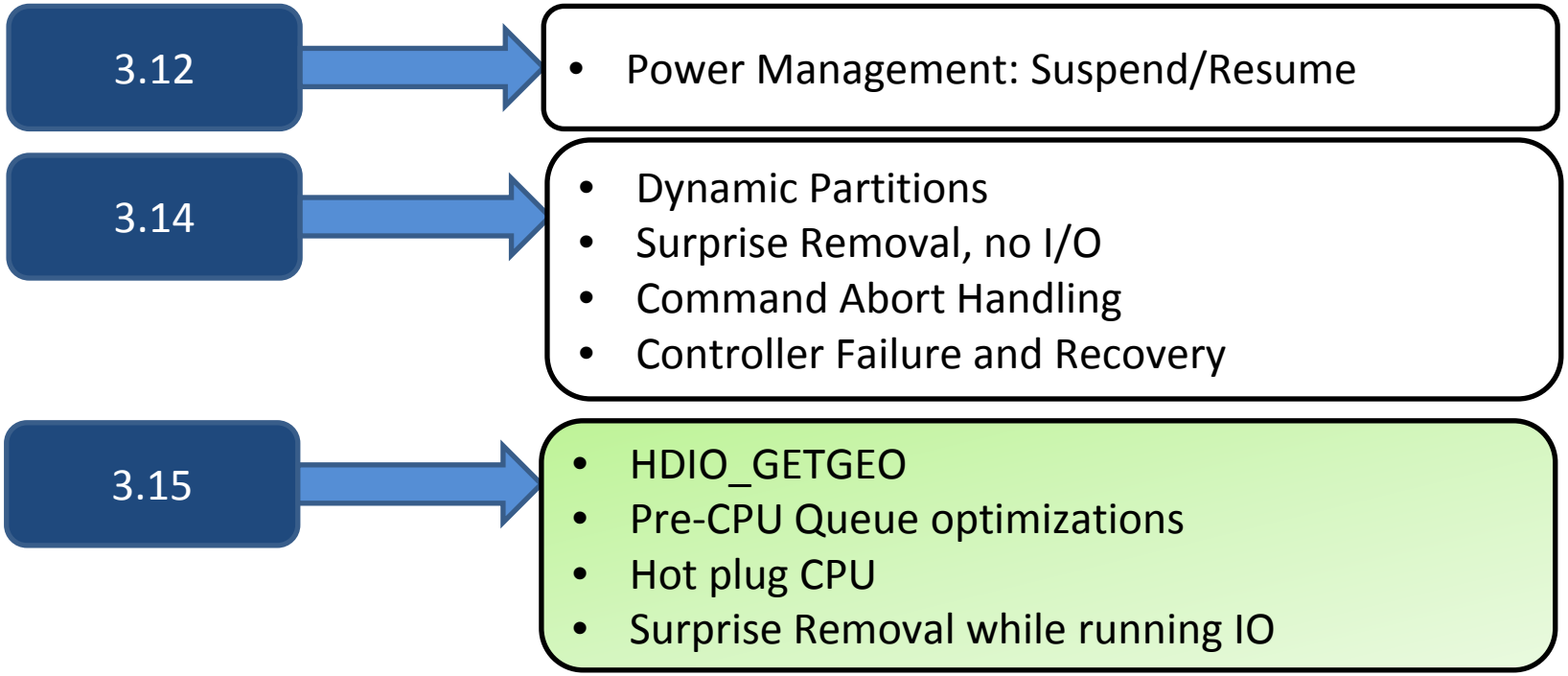
3.12

- Power Management: Suspend/Resume

Kernel 3.14


3.12
• Power Management: Suspend/Resume
3.14
• Dynamic Partitions
• Surprise Removal (no I/O)
• Command Abort Handling
• Controller Failure and Recovery

Kernel 3.15

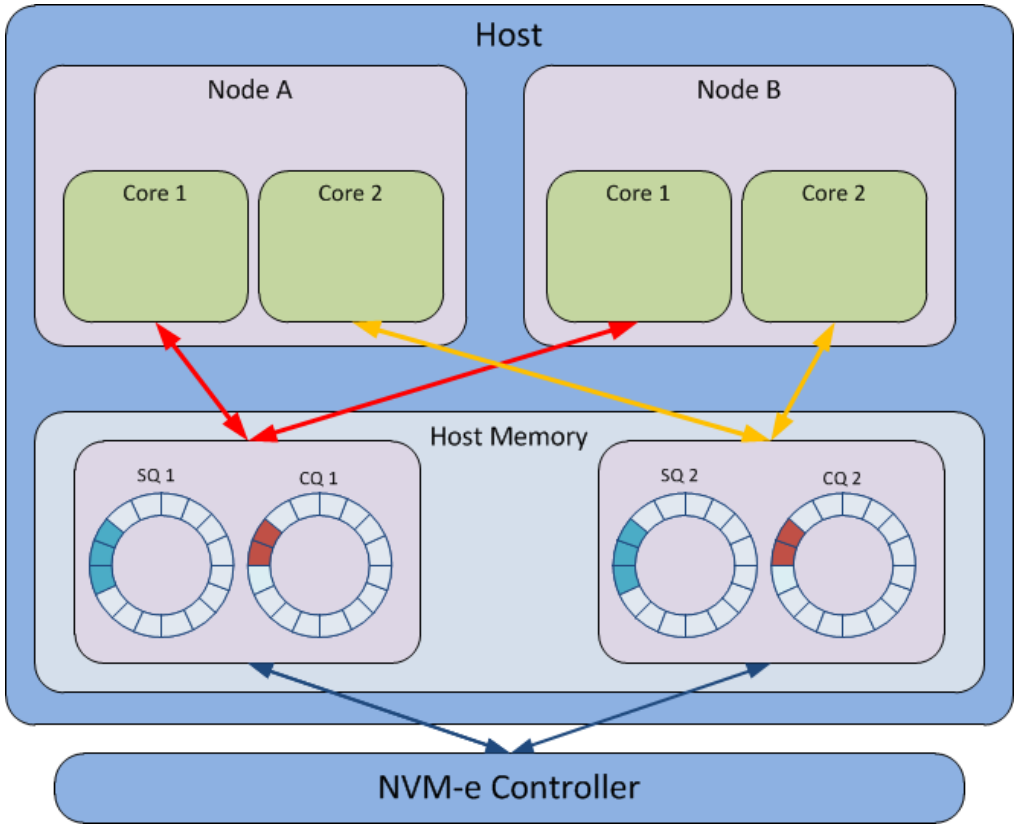

3.12
Power Management: Suspend/Resume
3.14
Dynamic Partitions
Surprise Removal, no I/O
Command Abort Handling
Controller Failure and Recovery
3.15
HDIO_GETGEO
Pre-CPU Queue optimizations
Hot plug CPU
Surprise Removal while running IO

3.15: Disk Geometry

- Prevent partitions that create this scenario:

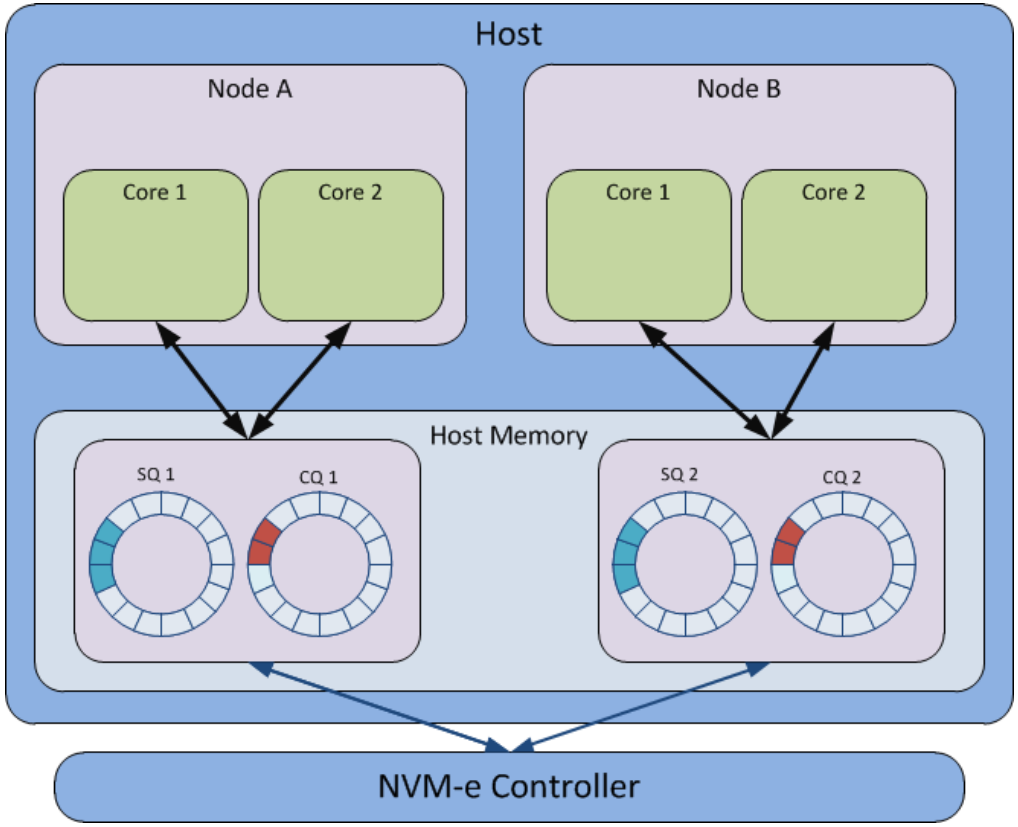
3.15: Per-CPU Optimization

- When more cores than queues, before:



3.15: Per-CPU Optimization

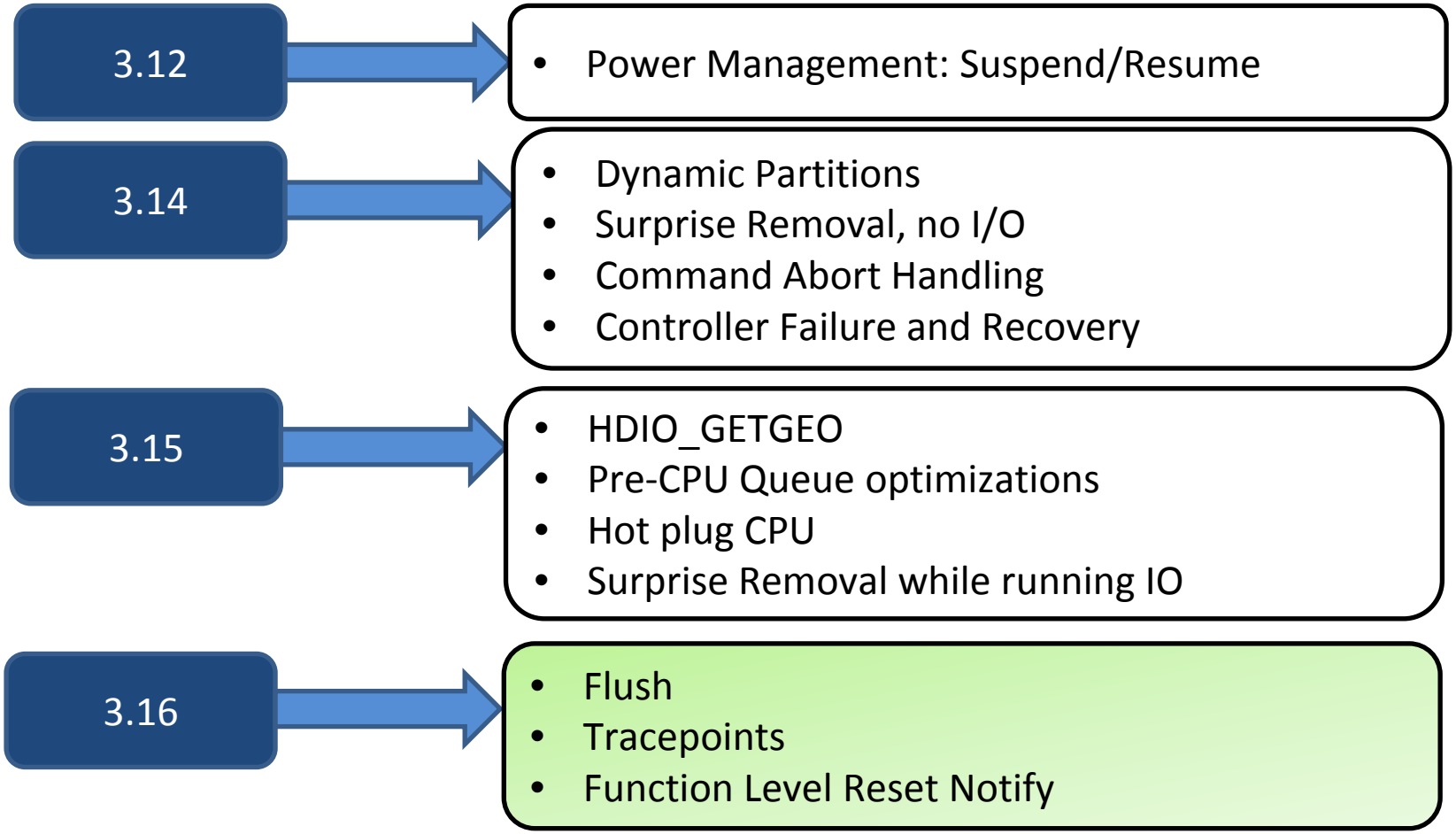
- When more cores than queues, after:



3.15: Surprise Removal

- Additional synchronization and reference counting software need for controller+storage removal safe without sacrificing performance

Kernel 3.16


3.12
Power Management: Suspend/Resume
3.14
Dynamic Partitions
Surprise Removal, no I/O
Command Abort Handling
Controller Failure and Recovery
3.15
HDIO_GETGEO
Pre-CPU Queue optimizations
Hot plug CPU
Surprise Removal while running IO
3.16
Flush
Tracepoints
Function Level Reset Notify

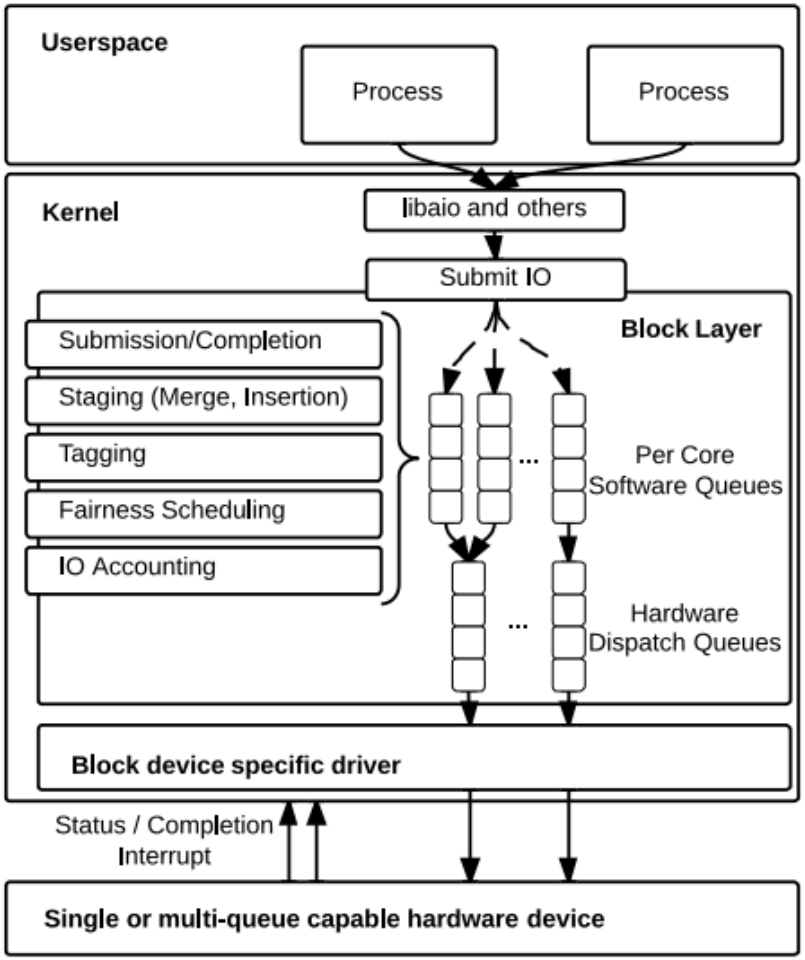
Kernel 3.17 (expected)

- Asynchronous Event Notification
- Mismatched Host-Device Pages
- PCI-e Advanced Error Reporting
- Write Zeroes command

Kernel Enhancements & Future Work

Block Multi-Queue

- Removes request based block driver software queue contention bottleneck
- Moves contention closer to the h/w, reducing latency on multi-threaded IO
- Merged into Linux 3.13; virtio-blk only mainline driver

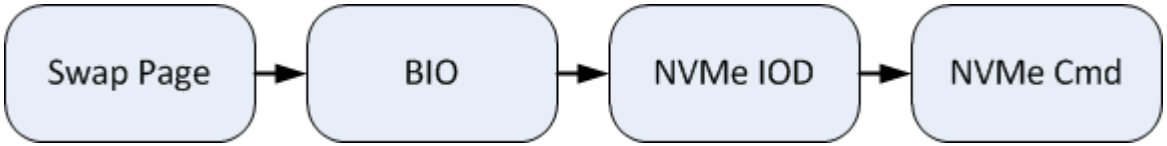


Block Multi-Queue

- IO Merging and Scheduling
- Removes duplicated code in bio-based drivers:
  - Timeouts, Tracing, Tagging, Diskstats, Runtime Power Management, Device Removal
- Latest nvme-mq tree based on 3.15, passes stability testing.
- Mainline merge date uncertain

Page IO

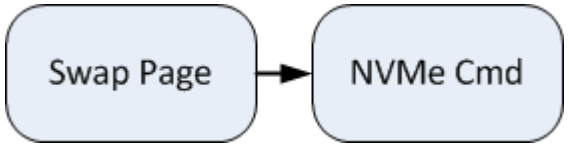
- Reading/writing a page (swap example):



- Multiple allocations, additional CPU and Memory resources required
- … but NVMe is optimized to handle pages

Page IO

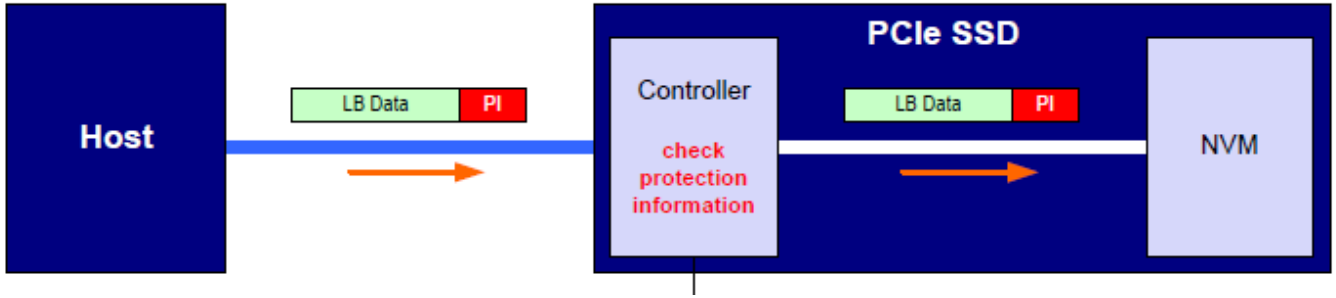
- New function to “bdev” ops with block layer read/write page functions:



- Benchmarks performance increase by 20%
  - No additional memory required!
- Applied in Linux 3.16
- TODO: Implement for NVMe

CRC T10 DIF/DIX

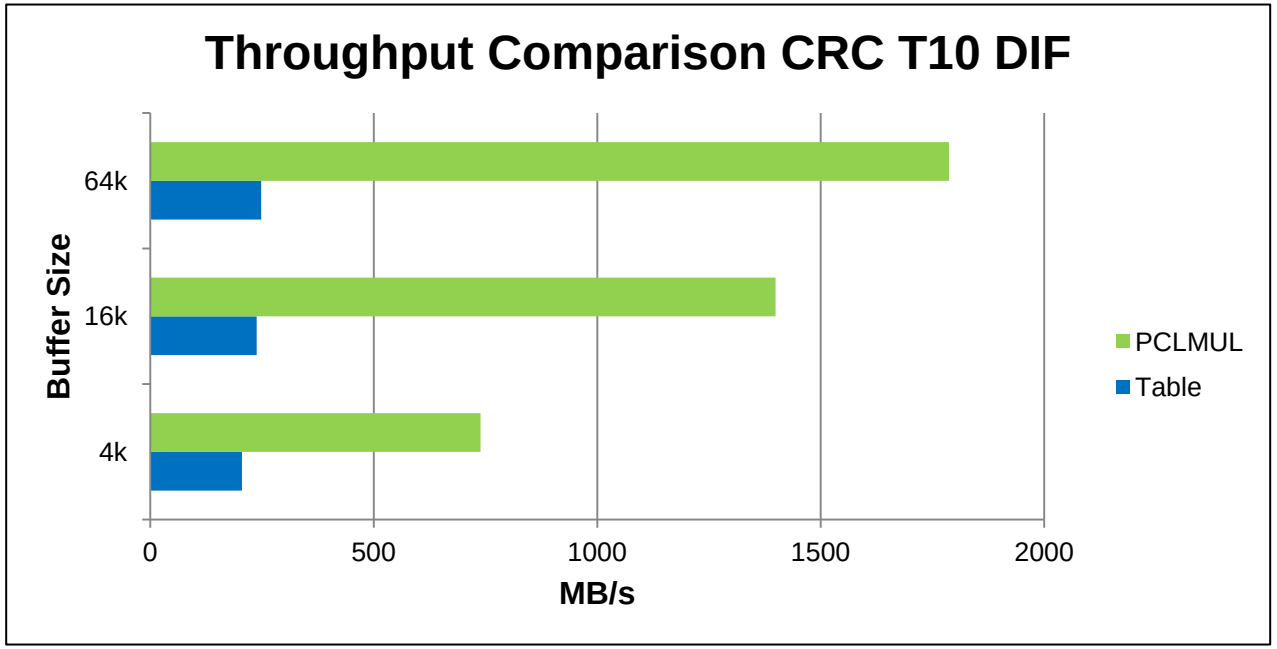
- Meta-data Formats used with protection information:



- Guard calculated as CRC-16, costly on CPU cycles.
- HW Acceleration?

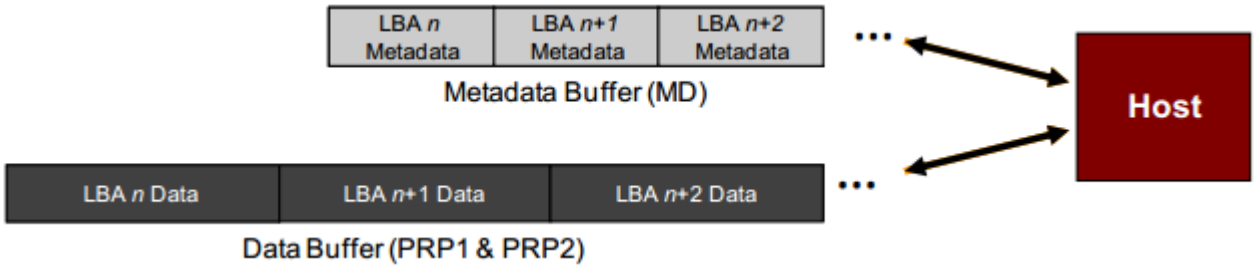
CRC T10 DIF/DIX

- X86_64 HW acceleration: PCLMULQDQ
  - Up to 8x calculation speed improvement
    - 3.5x on SSD I/O throughput
  - Added in 3.11 through linux-crypto



CRC T10 DIF/DIX

- Still not able to use blk-integrity extensions
  - Ref Tag consistent only with 512-byte block formats
    - NVM-e supports all powers of 2 >= 512
  - … and 8 bytes separate meta-data
    - NVM-e Metadata may be 8 to 65536 bytes



Other Future Work

- Block Polling
- NVMe 1.1 and 1.2 Support
  - SGL
  - Subsystem
  - Namespace List
- Performance Tuning
- Aggregate Doorbell writes
- Asynchronous Start/Stop
- Runtime Power Management
- SR-IOV
- Anything else ???

Getting Involved

- Subscribe to Linux-NVMe Mailing List

  http://lists.infradead.org/mailman/listinfo/linux-nvme
- Clone and enhance Linux-NVMe Repository:

  http://git.infradead.org/users/willy/linux-nvme.git
- Read up on NVM-Express:

  http://www.nvmexpress.org/

FIN

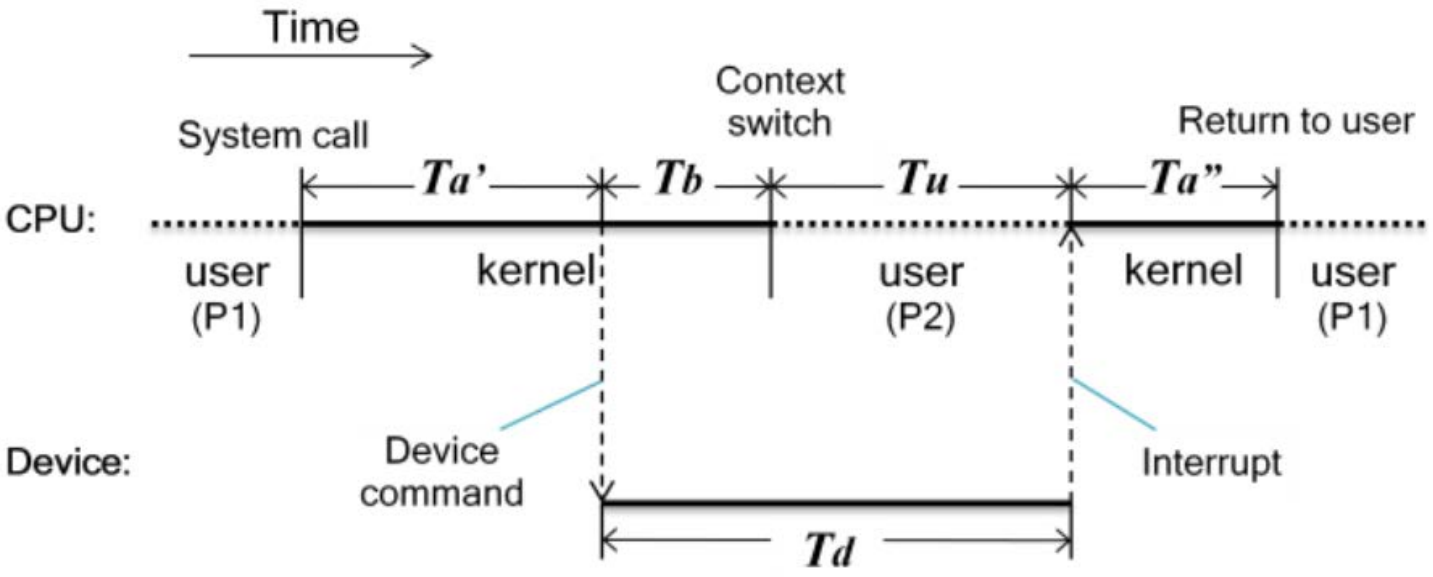
Questions?

keith.busch@intel.com

Backup

Block Polling

- IO Latency Sources:



- Beyond NAND: For low-latency device, context switch and interrupt dominate observed latency.


Flash Memory
SUMMIT

User-space Tooling

- Examples of user space tools for device management:

  http://git.infradead.org/users/kbusch/nvme-user.git

- Provides examples using the NVMe IOCTL interface to send commands to controller and parse output

Emulation

- Wanting to try nvme software, but lacking hardware:

  http://git.infradead.org/users/kbusch/qemu-nvme.git

- Useful if you just want to test driver and user space tools.
- Performs poorly on imitating nvme performance characteristics, among other things.

References

- NVM-Express:
  http://www.nvmexpress.org/
- Linux-NVMe Mailing List
  http://lists.infradead.org/mailman/listinfo/linux-nvme
- Linux-NVMe Repository:
  http://git.infradead.org/users/willy/linux-nvme.git
- CRC T10 DIF
  https://lkml.org/lkml/2013/5/1/449
- Page IO
  http://marc.info/?l=linux-kernel&m=139843448100786&w=2
- Block poll
  - http://lwn.net/SubscriberLink/556244/309ec42e8b9a4fcf/
- Block Multi-queue
  http://kernel.dk/systor13-final18.pdf
- SAS vs NVMe:
  http://www.nvmexpress.org/wp-content/uploads/2013/04/IDF-2012-NVM-Express-and-the-PCI-Express-SSD-Revolution.pdf