HISTORY OF NVMe IN LINUX

Understanding NVM Express primary features and timeline in the Linux upstream committed kernel through May 2015

Linux Development @ Intel

Matthew Wilcox, Keith Busch and Jon Derrick

Legal Disclaimer

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to http://www.intel.com/design/literature.htm.

This document contains information on products in the design phase of development.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

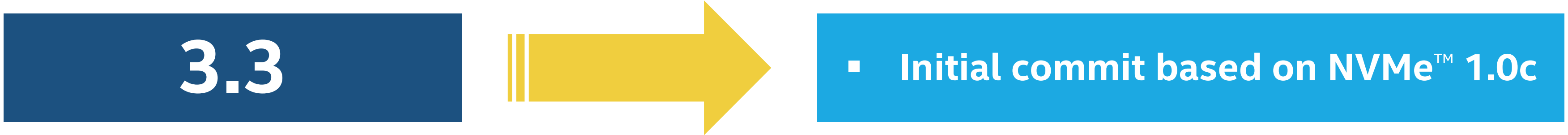
Results have been simulated and are provided for informational purposes only. Results were derived using simulations run on an architecture simulator or model. Any difference in system hardware or software design or configuration may affect actual performance.

Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others.

The Beginning

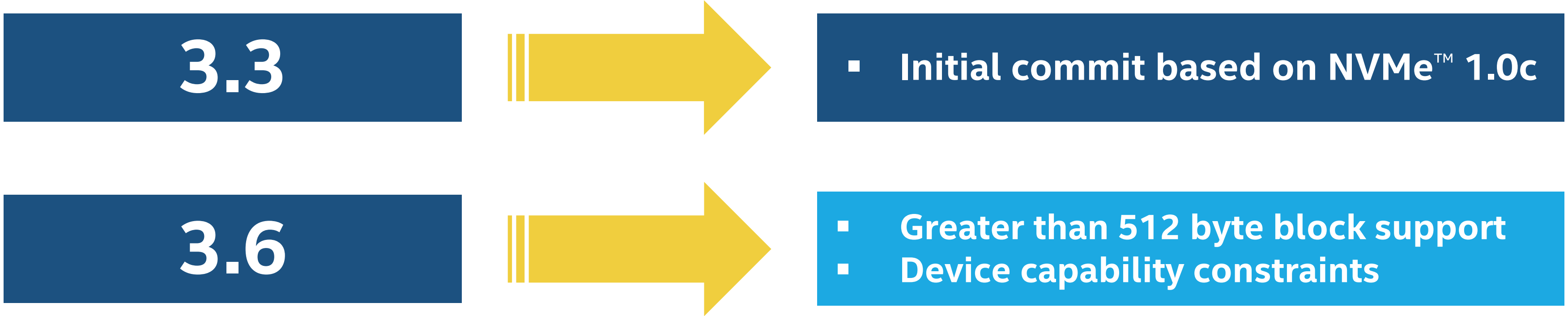


3.3

- Initial commit based on NVMe™ 1.0c

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

Kernel 3.6 …



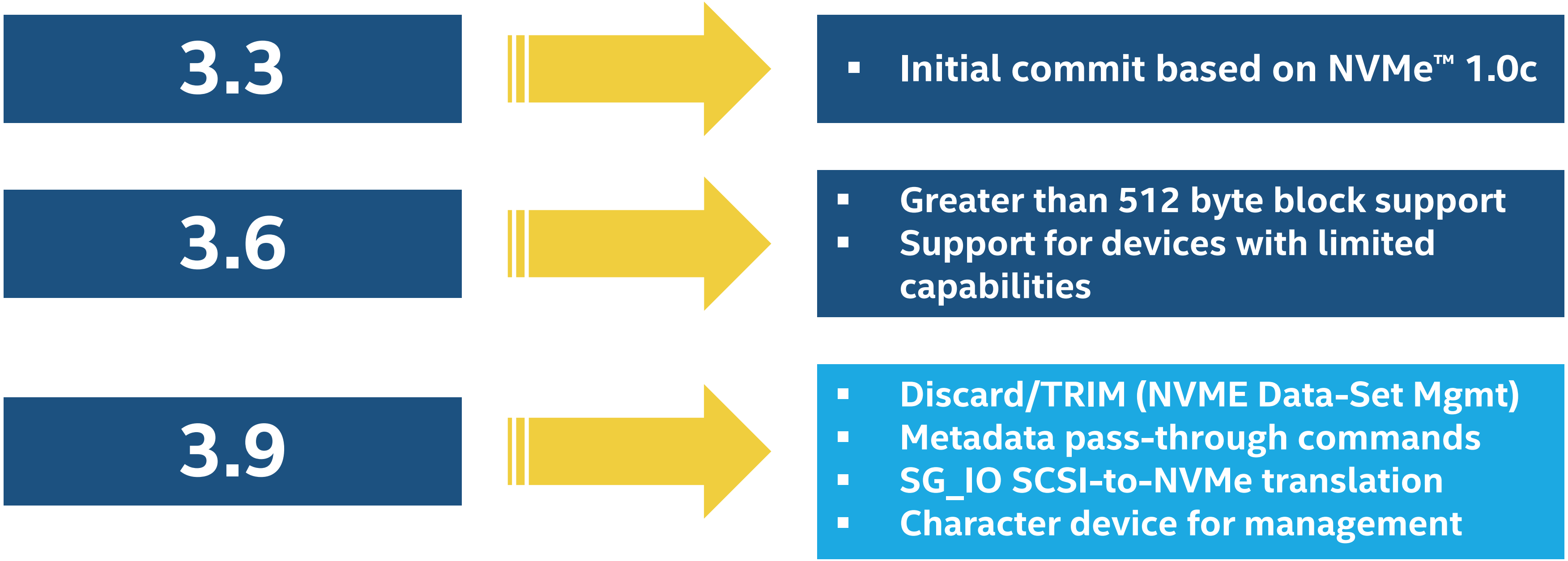
3.3

- 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

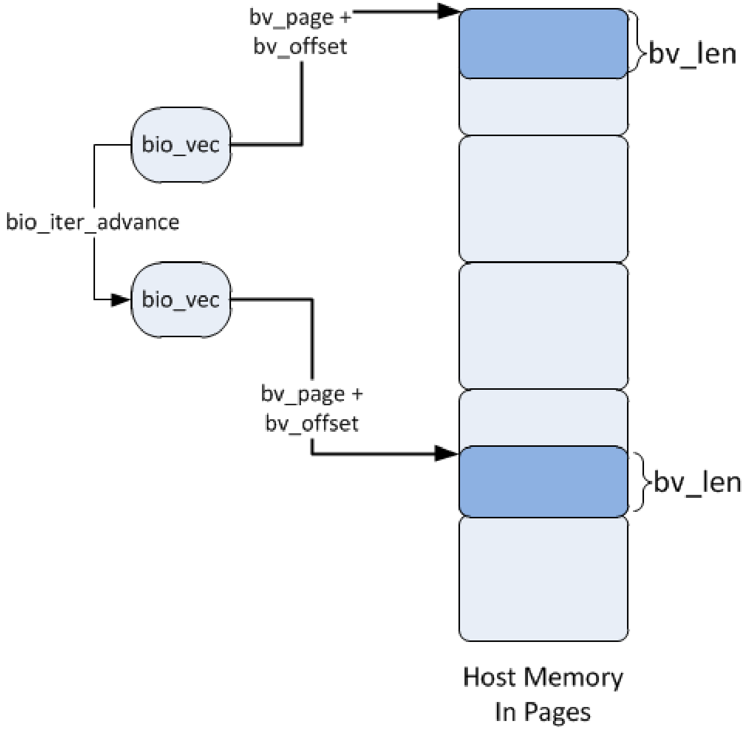
- Greater than 512 byte block support
- Support for devices with limited capabilities

3.9

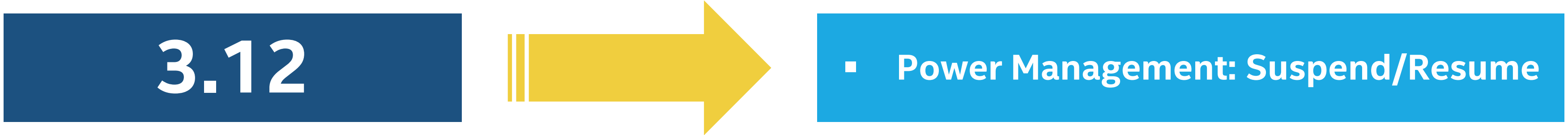
- Discard/TRIM (NVME Data-Set Mgmt)
- Metadata pass-through commands
- SG_IO SCSI-to-NVMe translation
- Character device for management

Kernel 3.10: Bio Splitting

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

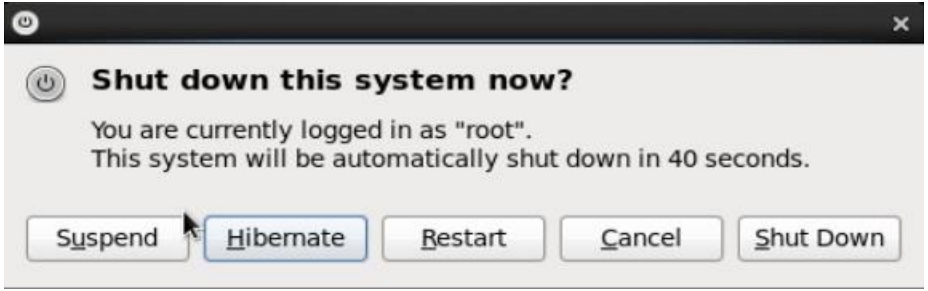


Kernel 3.12 …



3.12

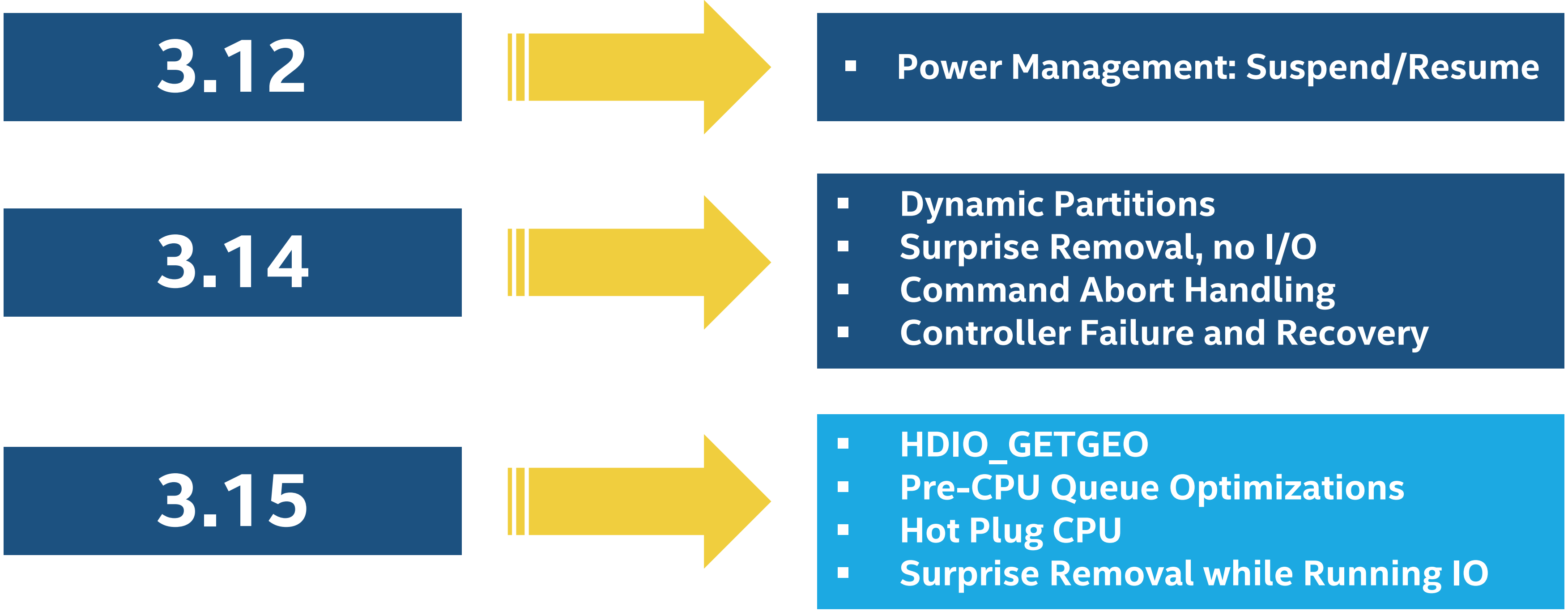
- Power Management: Suspend/Resume



Shut down this system now?

You are currently logged in as "root".
This system will be automatically shut down in 40 seconds.

Kernel 3.15 ...



3.12

- Power Management: Suspend/Resume

3.14

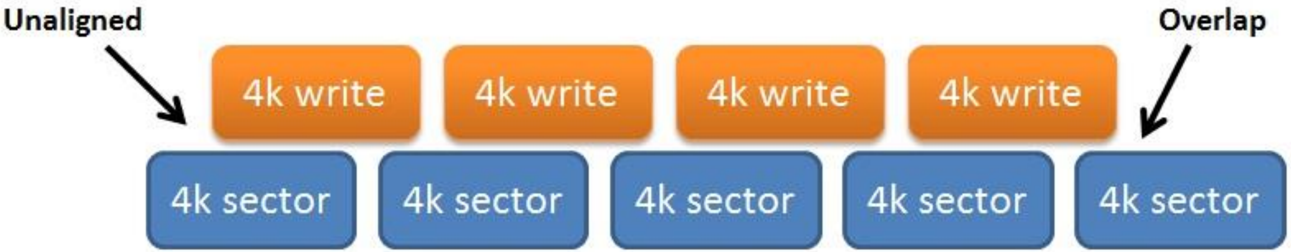
- Dynamic Partitions
- Surprise Removal, no I/O
- Command Abort Handling
- Controller Failure and Recovery

3.15

- HDIO_GETGEO
- Pre-CPU Queue Optimizations
- Hot Plug CPU
- Surprise Removal while Running IO

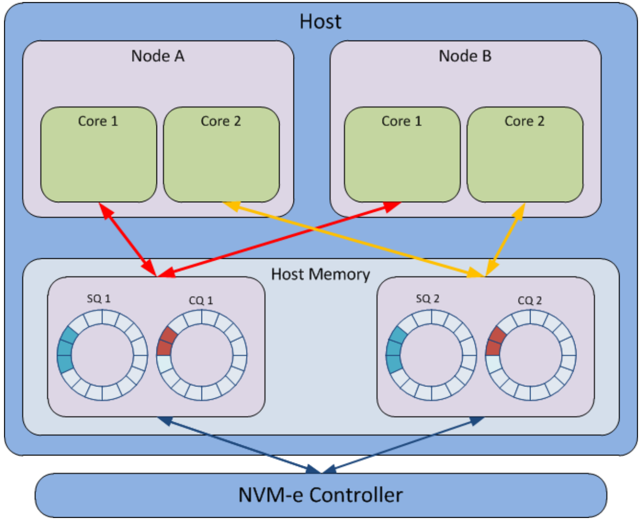
Kernel 3.15: Disk Geometry

- Prevent partitions that create this scenario:



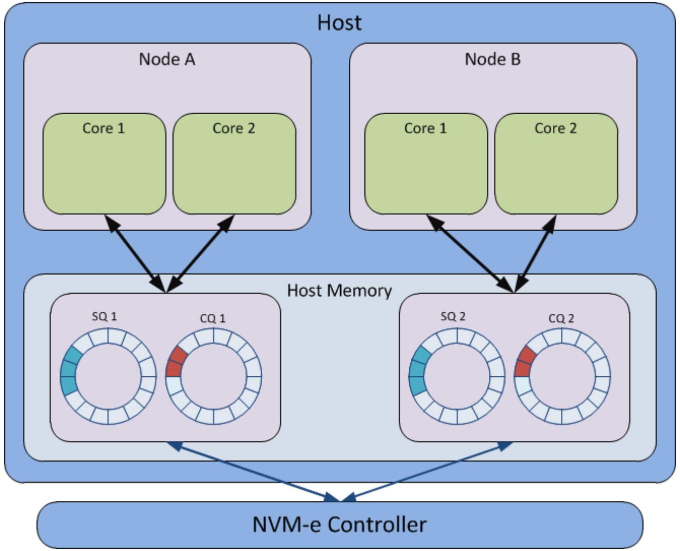
3.15: Per-CPU Optimization

- 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

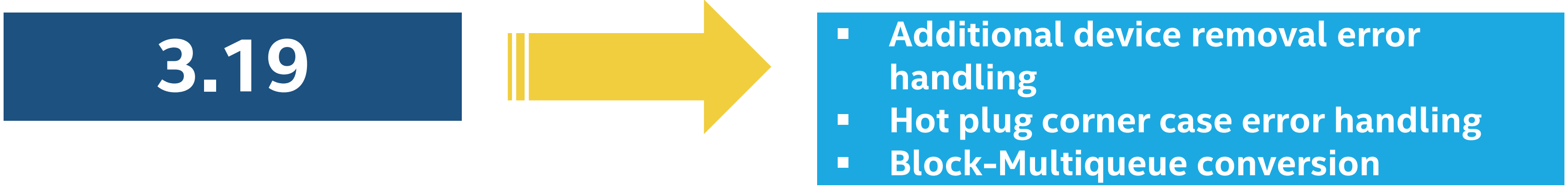
3.15

- HDIO_GETGEO
- Pre-CPU Queue Optimizations
- Hot Plug CPU
- Surprise Removal while Running IO

3.16

- Flush
- Tracepoints
- Function Level Reset Notify

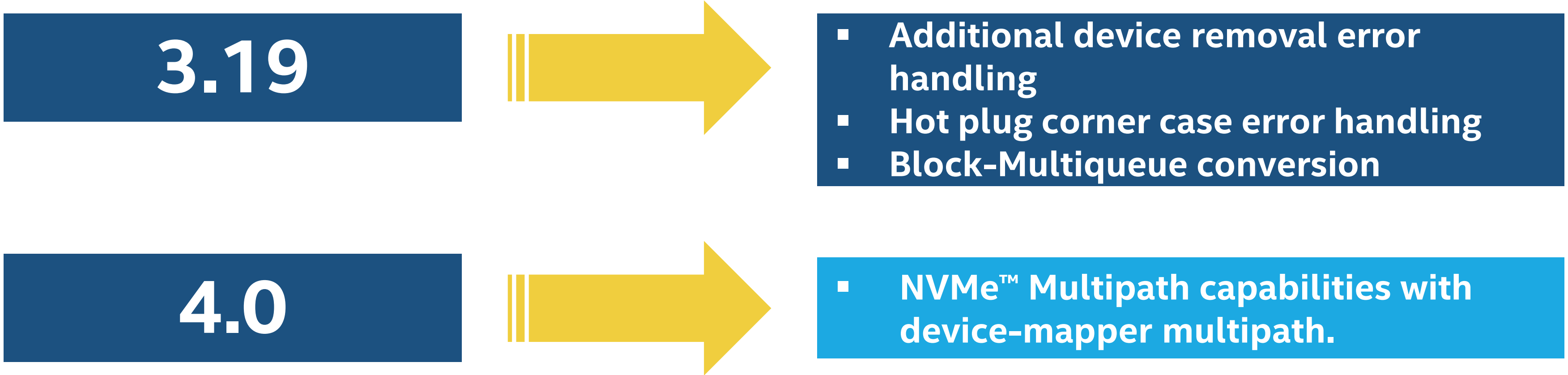
Kernel 3.19 …



3.19

- Additional device removal error handling
- Hot plug corner case error handling
- Block-Multiqueue conversion

Kernel 4.0 …



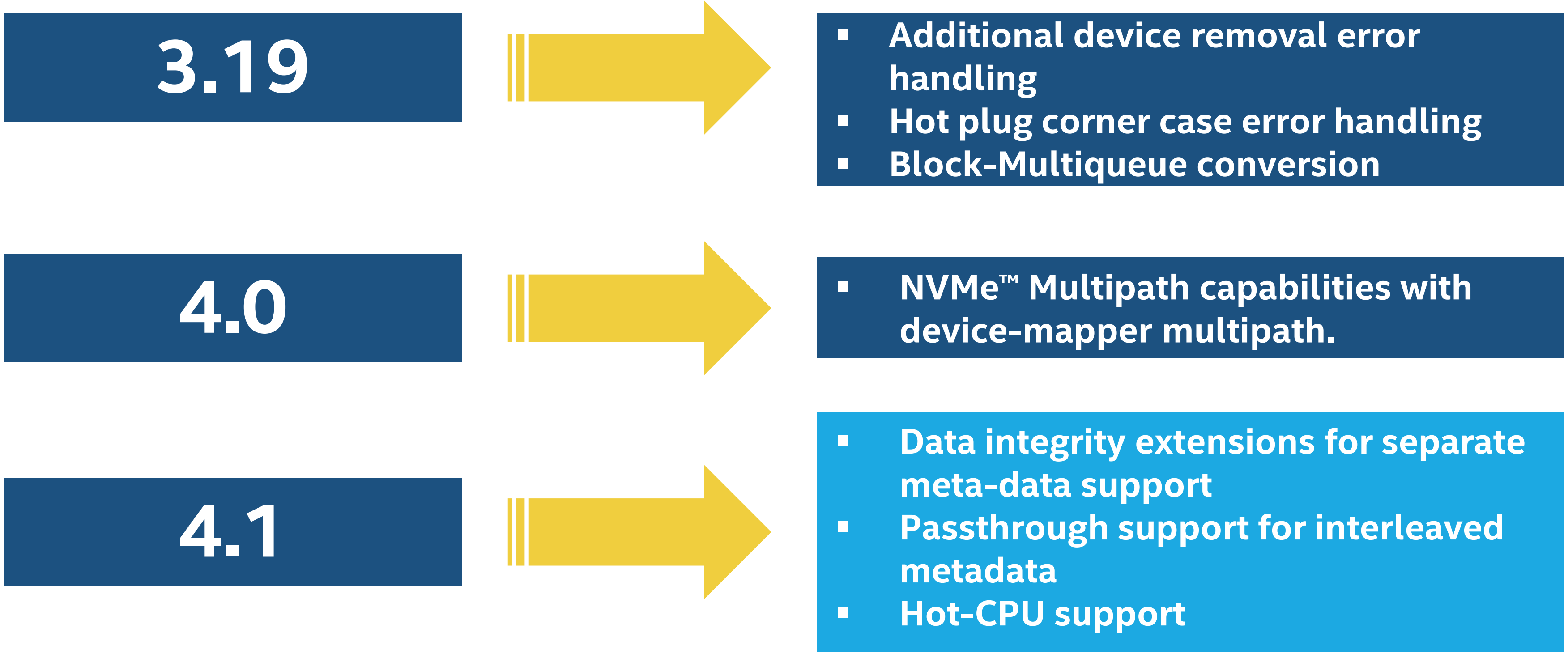
3.19

- Additional device removal error handling
- Hot plug corner case error handling
- Block-Multiqueue conversion

4.0

- NVMe™ Multipath capabilities with device-mapper multipath.

Kernel 4.1 …



3.19

- Additional device removal error handling
- Hot plug corner case error handling
- Block-Multiqueue conversion

4.0

- NVMe™ Multipath capabilities with device-mapper multipath.

4.1

- Data integrity extensions for separate meta-data support
- Passthrough support for interleaved metadata
- Hot-CPU support

experience
what's inside™