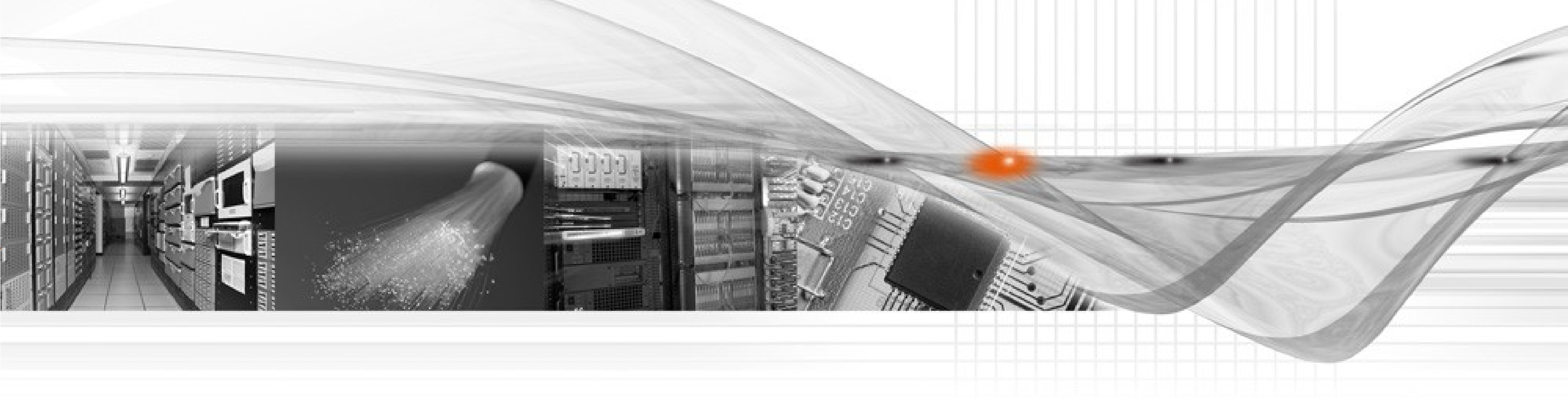


OPENFABRICS ALLIANCE

Introduction to the NVMe Working Group Initiative

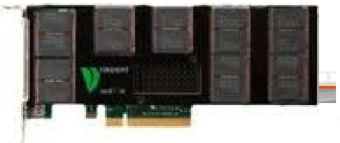
Author: Paul Luse
Date: 3/27/2012

Agenda

- NVM Express: Accelerating the PCI Express* SSD Transition
- Driver Ecosystem
- Looking to the Future

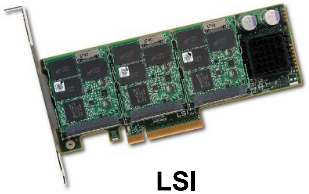
PCI Express* Ideal for SSDs

- PCI Express* is high performance
  - Full duplex, multiple outstanding requests, and out of order processing
  - Scalable port width (x1 to x16)
  - Scalable link speed (250/500/1000 MB/s)
  - Low latency
- PCIe is low cost
  - High volume commodity interconnect
  - Direct attach to CPU
- PCIe power management capabilities
  - Features include: Link power management, Optimized Buffer Flush/Fill (OBFF), Dynamic Power Allocation, etc
  - Optimized link idle power with L1.OFF


Virident

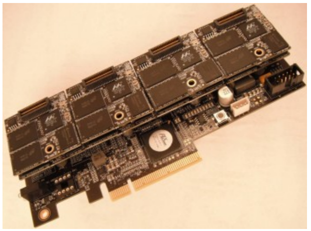
Fusion-io

Micron


LSI

OCZ

Intel


Marvell

PCI Express is a great interface for SSDs, and is making its presence known

Why NVM Express

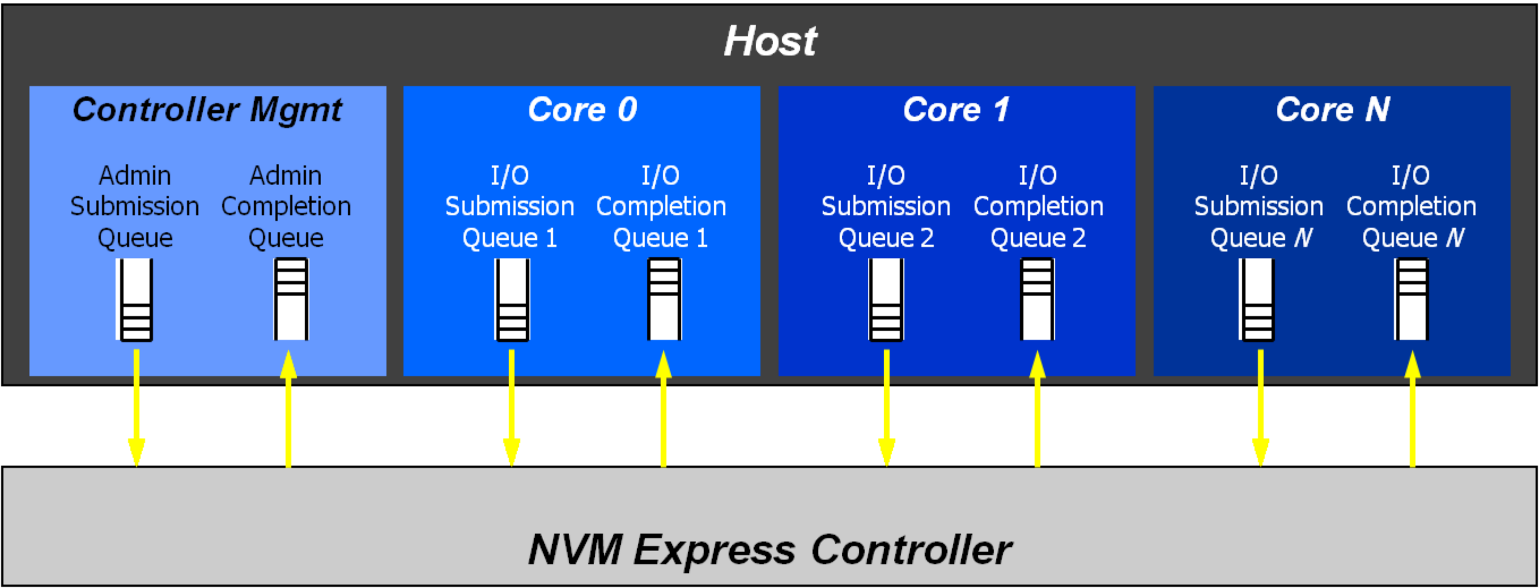
- Standards are needed for widespread industry adoption of PCIe SSDs
- NVM Express is a scalable host controller interface standard designed for Enterprise and Client systems that use PCI Express* SSDs
  - Includes optimized register interface and command set
- NVMe was developed by industry consortium of 80+ members and is directed by an 11 company Promoter Group
- NVMe 1.0 was published March 1, 2011, available at *nvmexpress.org*

NVMe enjoys wide industry support.

Product introductions starting later this year.

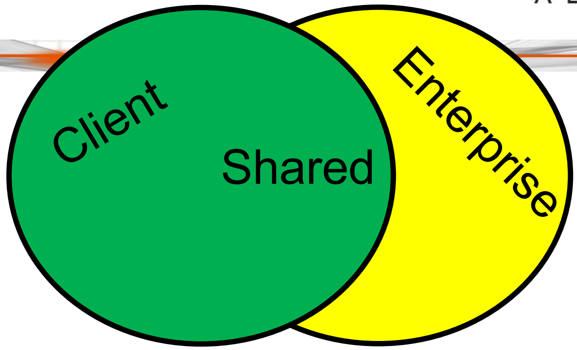
NVM Express Key Features

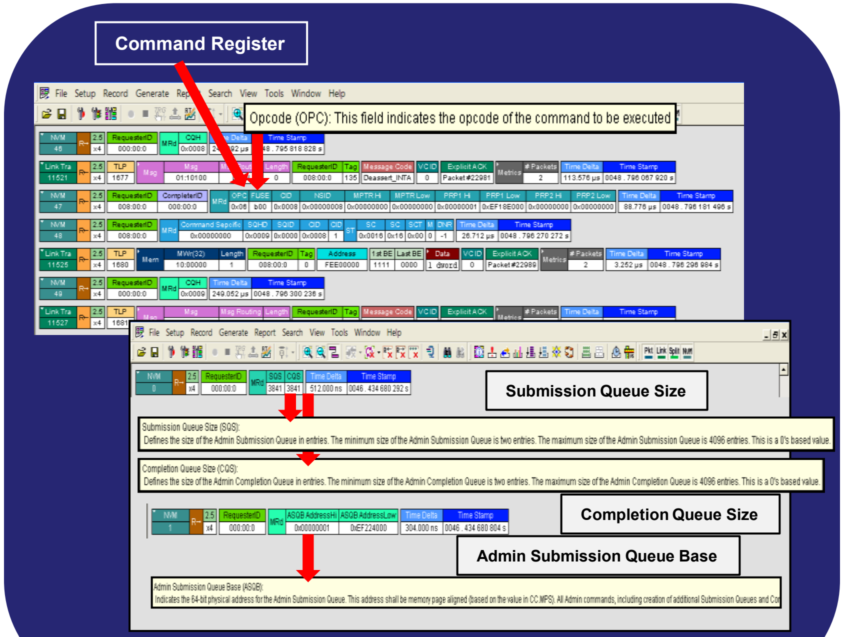
- The focus of the effort is efficiency, scalability, and performance
  - All parameters for 4KB command in single 64B DMA fetch
  - Supports deep queues (64K commands per queue, up to 64K queues)
  - Supports MSI-X and interrupt steering
  - Streamlined & simple command set (drops HDD legacy)
  - Enterprise: Support for end-to-end data protection (i.e., DIF/DIX)
  - NVM technology agnostic



NVMe Ecosystem Development

- By spanning client and Enterprise, infrastructure may be shared and benefit all segments
- Wide deployment of drivers are required for client, and may be leveraged in Enterprise
- UNH-IOL Interoperability program investment may be used for client & Enterprise
- Tools and IP may be leveraged
  - E.g., LeCroy* analyzer support
  - E.g., Synopsys* IP block





Opcode (OPC): This field indicates the opcode of the command to be executed

Submission Queue Size (SQS):
Defines the size of the Admin Submission Queue in entries. The minimum size of the Admin Submission Queue is two entries. The maximum size of the Admin Submission Queue is 4096 entries. This is a 0's based value.

Completion Queue Size (CQS):
Defines the size of the Admin Completion Queue in entries. The minimum size of the Admin Completion Queue is two entries. The maximum size of the Admin Completion Queue is 4096 entries. This is a 0's based value.

Admin Submission Queue Base (ASQB):
Indicates the 64-bit physical address for the Admin Submission Queue. This address shall be memory page aligned (based on the value in CC.MPS). All Admin commands, including creation of additional Submission Queues and Co

LeCroy PCIe Protocol Analyzer Trace

Agenda

- NVM Express: Accelerating the PCI Express* SSD Transition
- Driver Ecosystem
- Looking to the Future

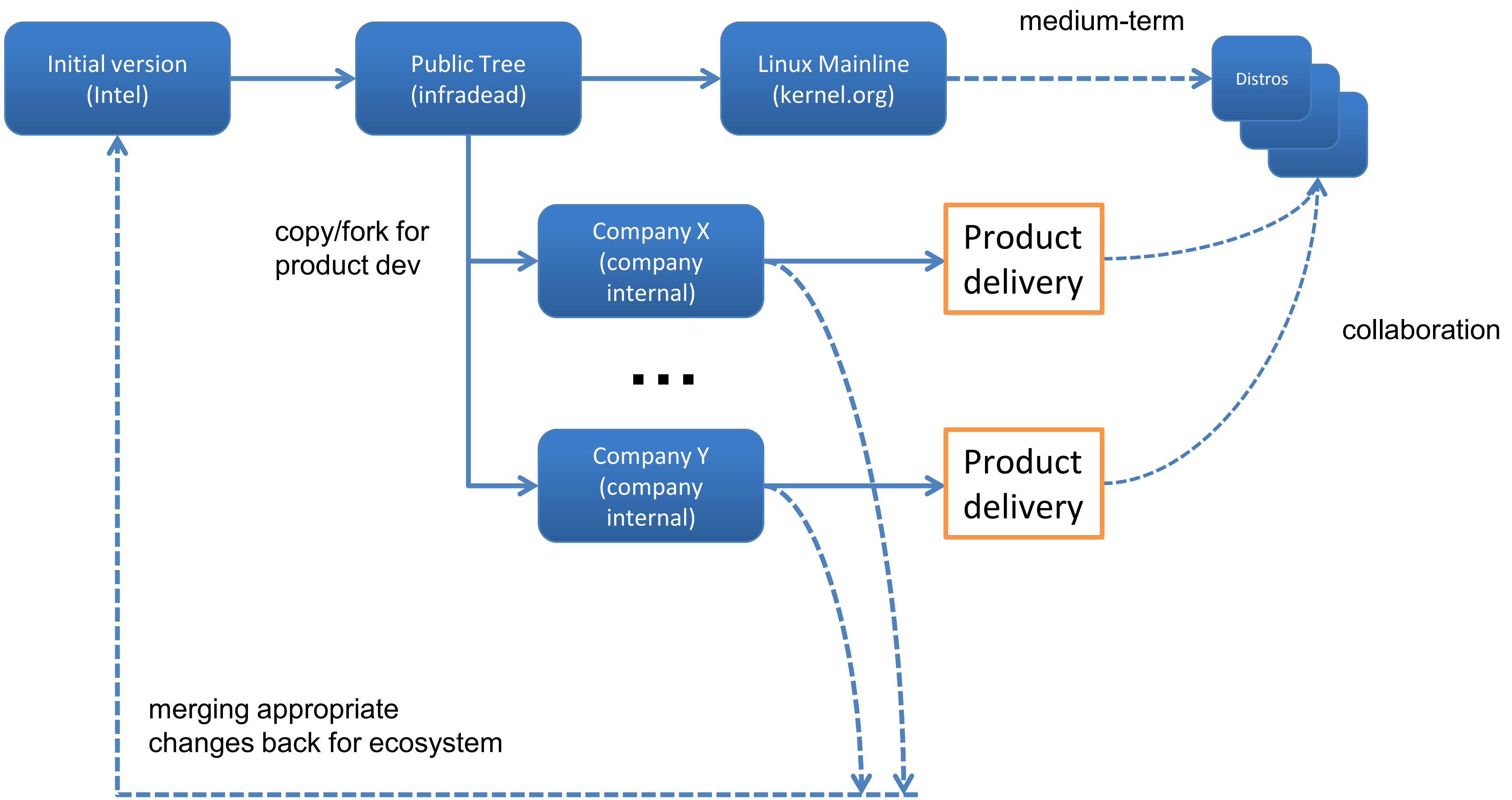
Reference Drivers for Key OSes

- Reference drivers complete for Linux*, Windows*, and Vmware*
- Linux
  - Already accepted into the mainline kernel on kernel.org
  - Open source with GPL license
  - Refer to http://git.infradead.org/users/willy/linux-nvme.git
- Windows
  - Baseline developed in collaboration by IDT, Intel, and LSI
  - Open source with BSD license
  - Maintenance is collaboration by NVMe WG and Open Fabrics Alliance
  - Refer to https://www.openfabrics.org/resources/developer-tools/nvme-windows-development.html
- VMware
  - Initial driver developed by Intel
  - Based on VMware advice, “vmk linux” driver based on Linux version
  - NVMe WG will collaborate with VMware on delivery/maintenance

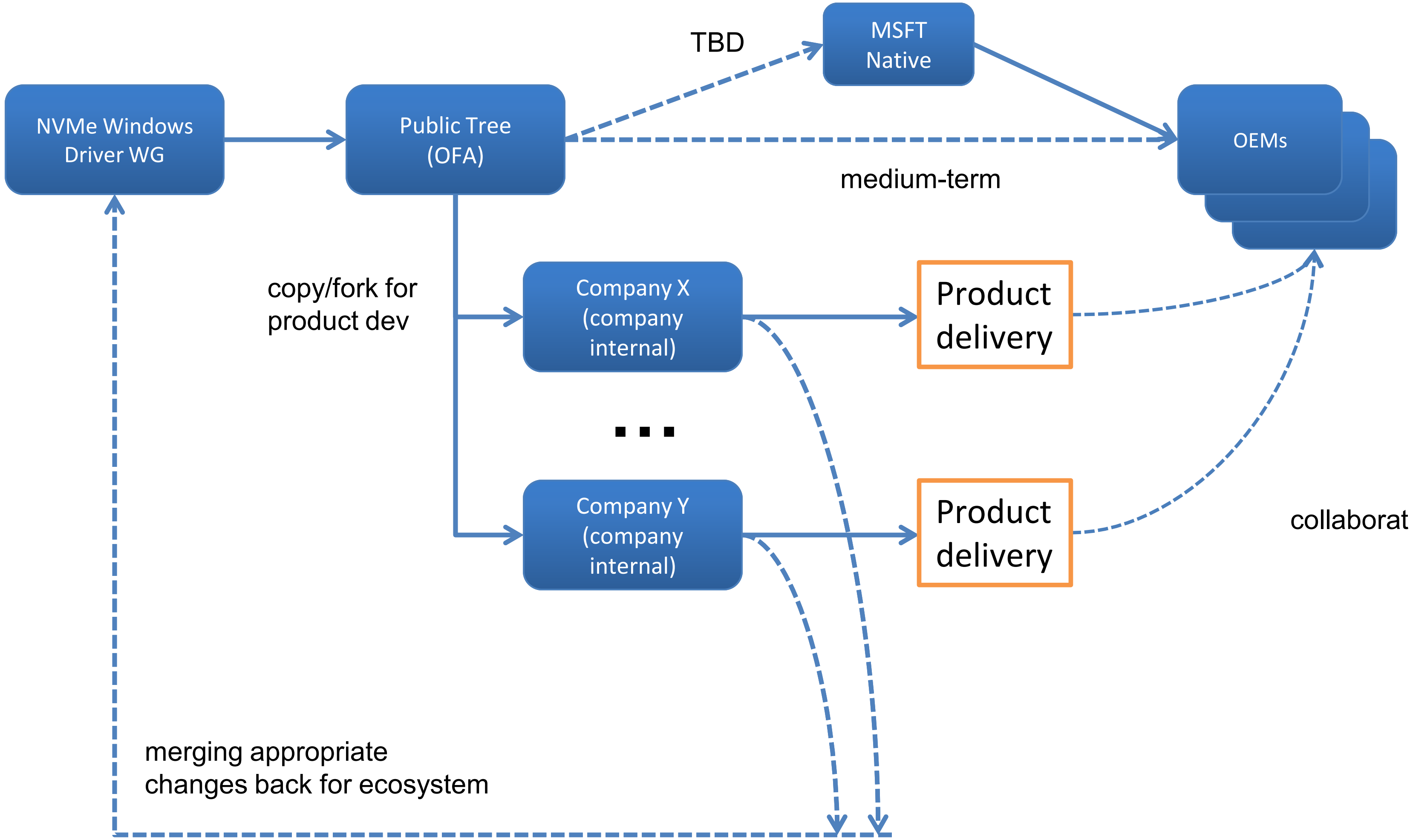
Driver Ecosystem Goals

- The long term goal is for each major OS to ship with a standard NVM Express driver
- The short term goal is to allow NVMe device manufacturers to provide the drivers they need with their products leveraging the reference drivers
- The reference drivers provide high performance, validated and fully compliant drivers to the ecosystem with reasonable licenses (e.g., GPL, BSD)
- "Fork and Merge" to achieve short term with reference drivers
  - Each NVMe device manufacturer "forks" the reference driver
  - Each NVMe device manufacturer adds in any product specific features
  - Each NVMe device manufacturer "merges" industry-wide applicable changes back to the reference driver

Linux* “Fork and Merge”



Windows* “Fork and Merge”



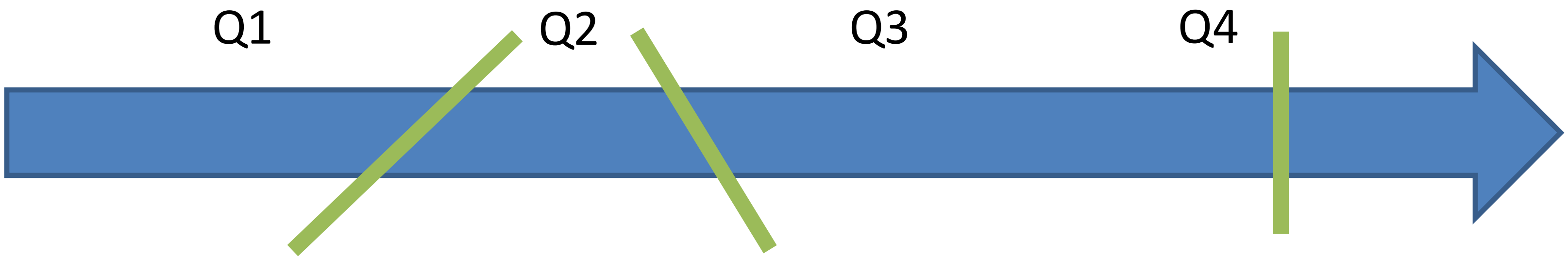
TBD
MSFT Native
NVMe Windows Driver WG
Public Tree (OFA)
OEMs
medium-term
copy/fork for product dev
Company X (company internal)
Product delivery
...
Company Y (company internal)
Product delivery
collaboration
merging appropriate changes back for ecosystem

Agenda

- NVM Express: Accelerating the PCI Express* SSD Transition
- Driver Ecosystem
- Looking to the Future

Reference Driver Releases in 2012

- Linux* and Windows* drivers are targeting two releases per year
- This cadence supports both NVMe and OS evolution



Linux Release 1

- Part of 3.3 kernel

Windows Release 1

- Initial release
- NVMe 1.0b support
- Win7,2008R2

Windows Release 1.1

- NVMe spec updates
- Public IOCTLs
- Win8 Storport specific
- Bug Fixes
- Win7,2008R2,Win8

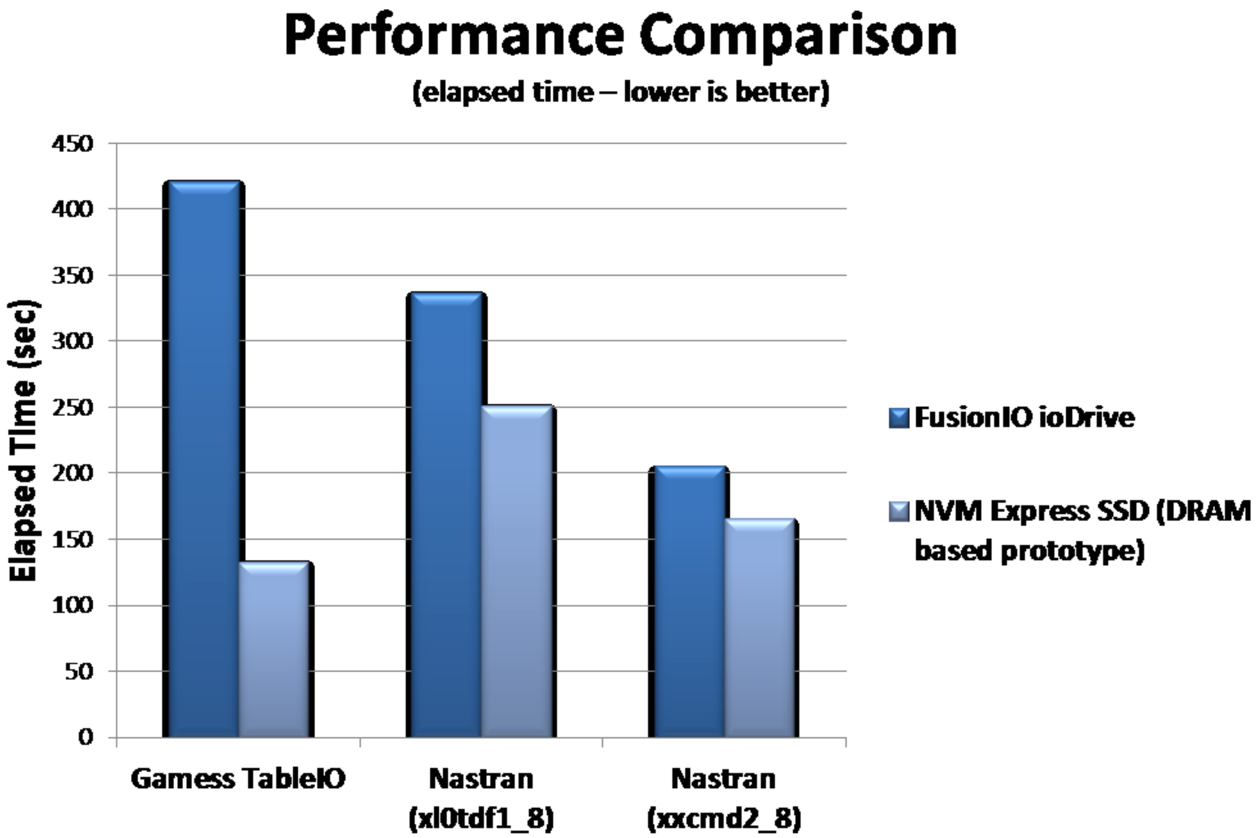
Backup

NVMe: Architected for Efficiency

- Efficiency = Lower Power & Higher Performance

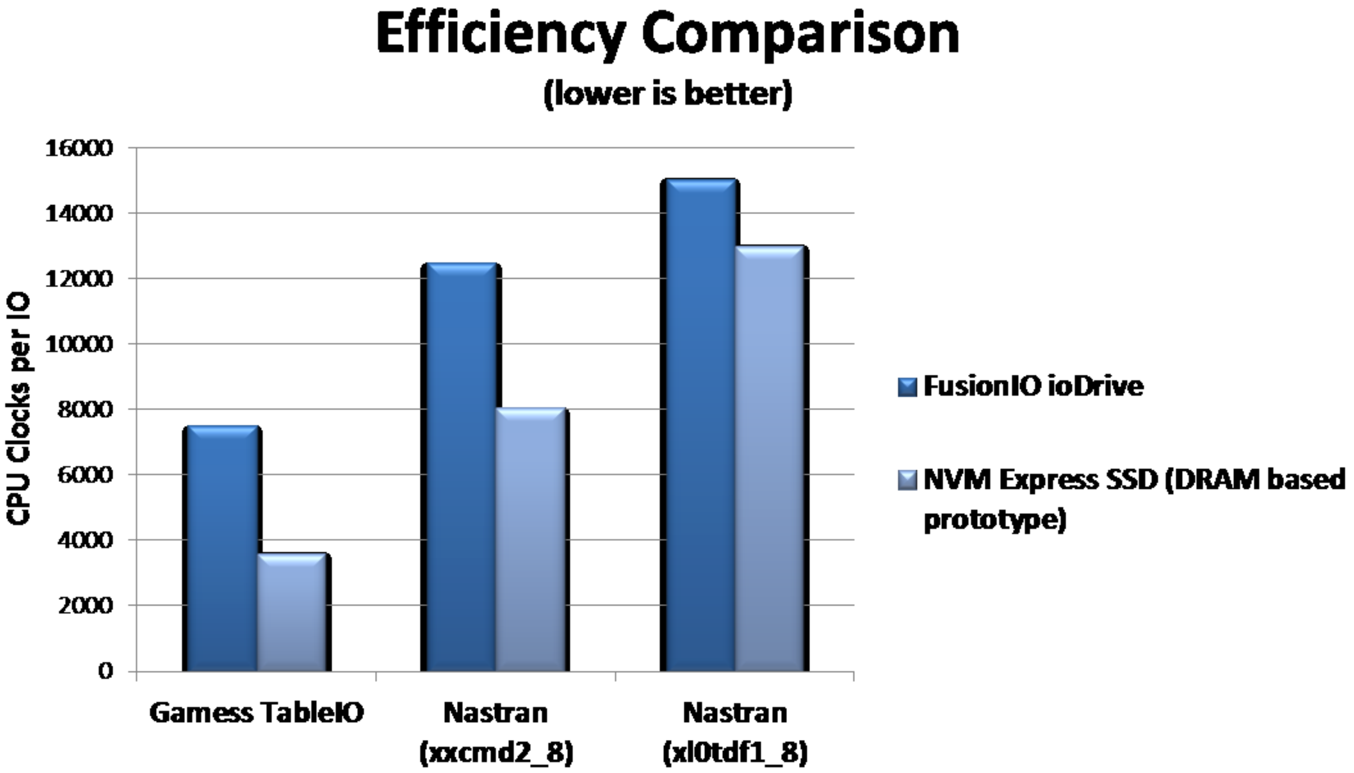
	AHCI	nvm EXPRESS
Uncacheable Register Reads Each consumes 2000 CPU cycles	4 per command 8000 cycles, ~ 2.5 µs	**0** per command
MSI-X and Interrupt Steering Ensures one core not IOPs bottleneck	No	Yes
Parallelism & Multiple Threads Ensures one core not IOPs bottleneck	Requires synchronization lock to issue command	No locking, doorbell register per Queue
Maximum Queue Depth Ensures one core not IOPs bottleneck	32	64K Queues 64K Commands per Q
Efficiency for 4KB Commands 4KB critical in Client and Enterprise	Command parameters require two serialized host DRAM fetches	Command parameters in one 64B fetch

NVM Express Has Headroom for Future Faster NVM Technologies



NVMe has scalability to support 10 year roadmap.

NVM Express: An Efficient Interface



NVMe prototype delivers is lower clocks per IO while delivering much higher performance.

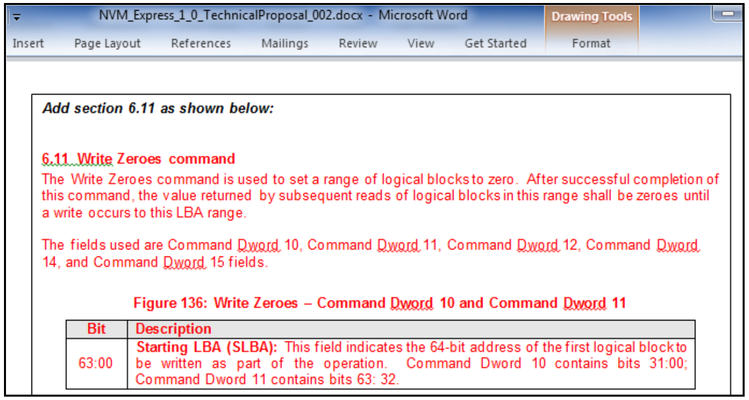
Continuing to Add Value

- NVMe 1.0 was published March 1, 2011
- In the fall of 2011, the NVMe Workgroup began looking at adding new features to provide more value
- NVMe 1.1 definition is underway; features include:
  - Write Zeroes command
  - Data Copy command
  - Generalized SGL support
  - Enabling efficient Multi-Path solutions
- NVMe 1.1 is targeted for an August release
  - Note: All features are optional and add more value. There is no NVMe 1.1 feature required for a first generation NVMe device.

Write Zeroes command

- Filesystems spend a lot of time zeroing blocks of data under certain workloads, can this be optimized?
- Zeroing blocks using the Write command is inefficient
  - The host transfers lots of zeros over the bus wasting power
- Zeroing blocks using Deallocate cannot be relied upon
  - Deallocate (i.e. Trim) does not guarantee the state of the data
  - The value read may be all zeros, all ones, or last data written
- Solution: Add Write Zeroes command
  - Filesystem is guaranteed final state of data is zeros
  - No data buffer is transferred
  - End-to-end data protection supported

Ratified NVMe Technical Proposal



Add section 6.11 as shown below:

6.11 Write Zeroes command

The Write Zeroes command is used to set a range of logical blocks to zero. After successful completion of this command, the value returned by subsequent reads of logical blocks in this range shall be zeroes until a write occurs to this LBA range.

The fields used are Command Dword 10, Command Dword 11, Command Dword 12, Command Dword 14, and Command Dword 15 fields.

Figure 136: Write Zeroes – Command Dword 10 and Command Dword 11

Bit	Description
63:00	**Starting LBA (SLBA):** This field indicates the 64-bit address of the first logical block to be written as part of the operation. Command Dword 10 contains bits 31:00; Command Dword 11 contains bits 63: 32.

Data Copy command

- In order to save power, it is important to minimize sending data unnecessarily over the bus
- Filesystems and applications frequently copy data and the data is going from one LBA to another on the same SSD…
- Add Data Copy command to optimize power & performance
  - Specifies source & destination LBA and length; no data buffer
  - The device may internally move the data
  - In some cases, the SSD may modify logical to physical translation tables without any data movement

Typical Read/Write Copy Steps	Power Savings
No bus transfer for read data	~ 0.8 W for ~ 30 ms
No bus transfer for write data	~ 0.8 W for ~ 30 ms
No NAND write*	~ 4.0 W for ~ 30 ms
Total Savings	***~ 1.8 W for ~ 100 ms***

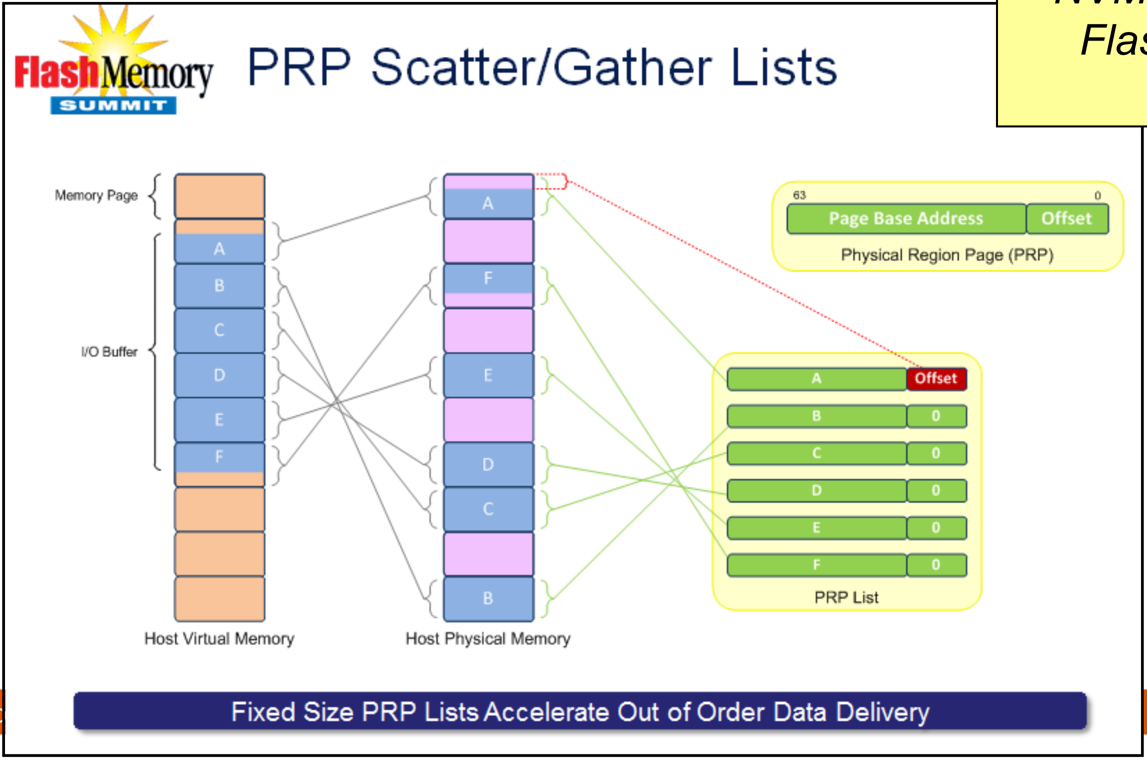
Assumptions: Modeled client x2 PCIe SSD. Data transfer and write bandwidth of 1 GB/s. 400 mW consumed per lane during data transfer for total of 1.6 W. Assumed NAND write (i.e. Program) power of 4 W.

* Assumes the SSD updates its logical to physical translation tables and does not do a data write.

Recall: Enabling Out of Order Data

- In previous talks, PRPs and their benefits have been described
  - A PRP has a fixed SGL entry size
  - Enables hardware to know where data starts in physical memory without walking the SGL table
- Benefit: Hardware may optimize data delivery
  - E.g. Transfer last half of data first

NVM Express talk at 8/11 Flash Memory Summit by Dell & IDT

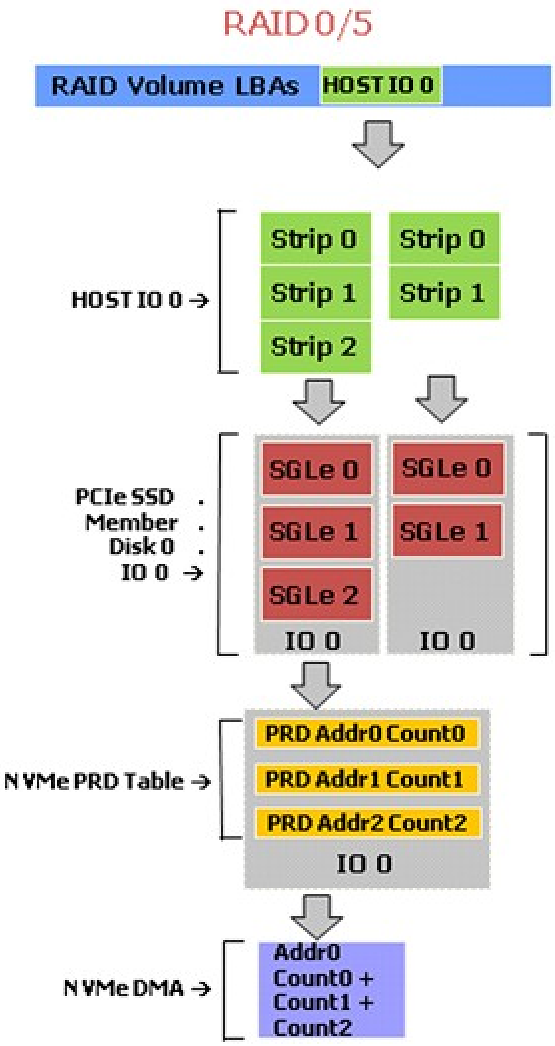


PRP Scatter/Gather Lists

Fixed Size PRP Lists Accelerate Out of Order Data Delivery

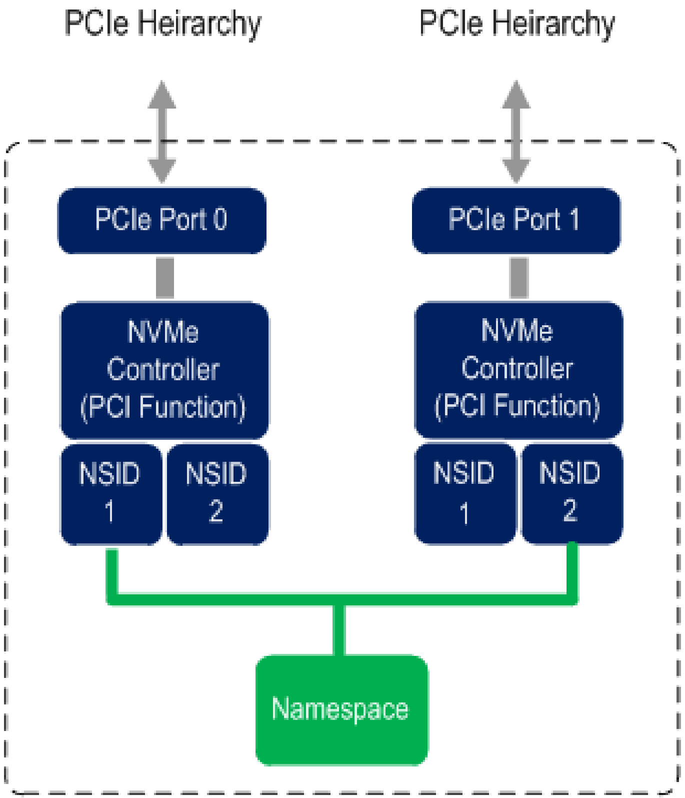
Generalized SGL Support

- There are a few unique cases where generalized SGL support is beneficial
- Example: Low level software RAID
  - If the first strip starts at an offset that is not page aligned then each new strip may have a non-zero offset
- NVMe 1.1 will add generalized SGL support as an option
  - Alternatively, command may be split into multiple commands
- Drivers should only use the generalized SGL when needed to avoid "losing out" on out of order data delivery efficiencies



Enabling Multi-Path

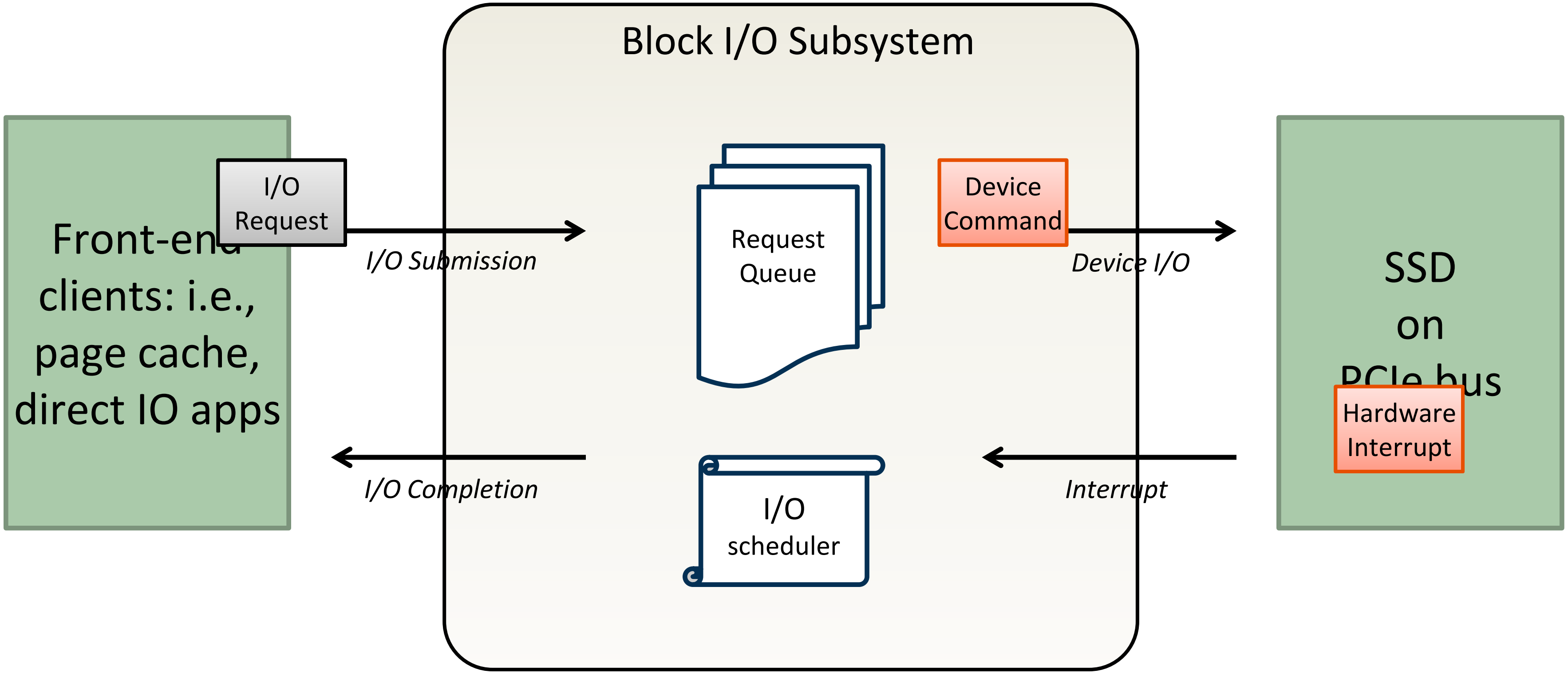
- An NVMe namespace may be accessed via multiple "paths"
  - SSD with multiple PCI Express* ports
  - SSD behind a PCIe switch to many hosts
- Two hosts accessing the same namespace must be coordinated
- NVMe is adding capabilities to enable effective host coordination
  - Unique ID for a namespace enables hosts to determine if accessing the same namespace (or not)
  - Reservation capability, allowing exclusive or shared access on a namespace basis



Take new optional NVMe 1.1 features into consideration in future designs.

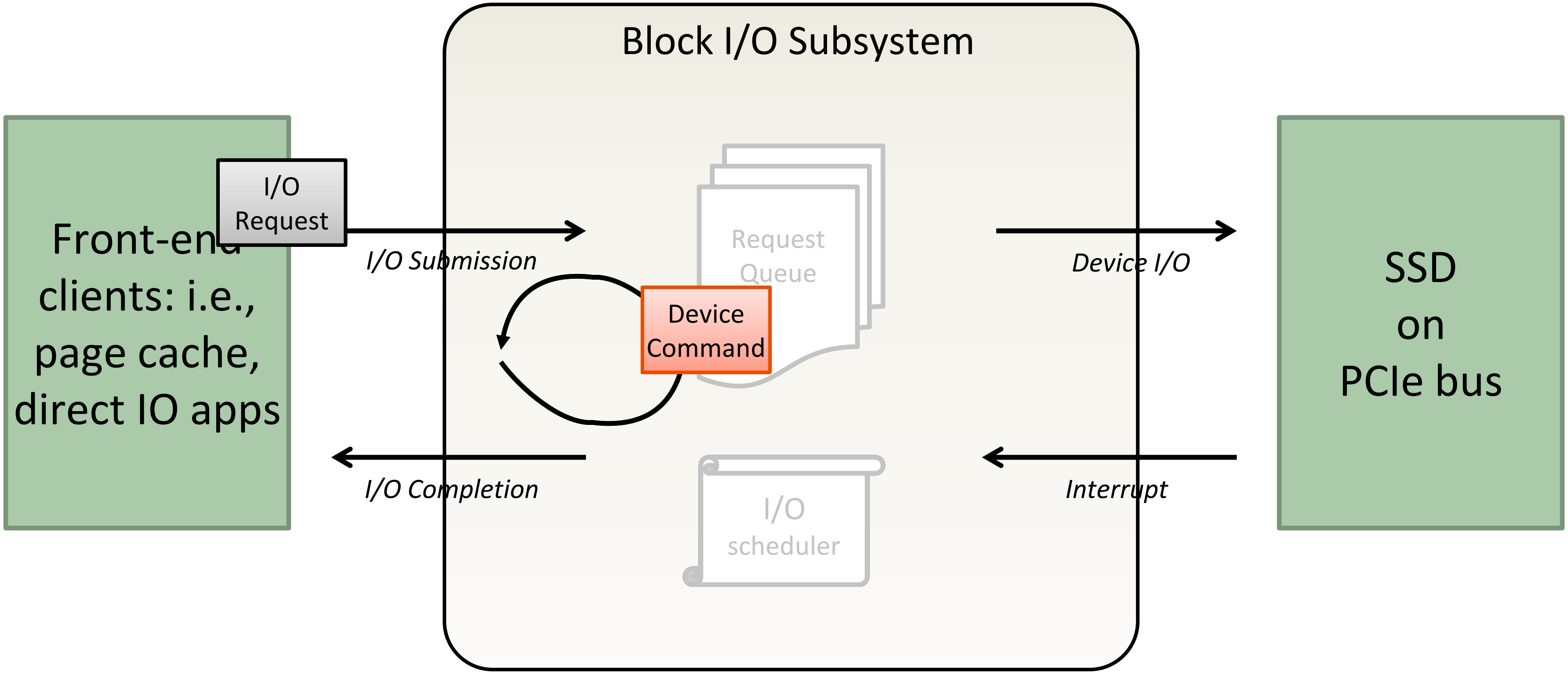
Asynchronous I/O Completion Flow

- Storage requests have traditionally had long latency
- This has led to an interrupt driven completion model



Synchronous I/O Completion Flow

- For devices that have lower latency, waiting for the completion may be shorter than the context switch time



Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

- A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.
- 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.
- Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to: http://www.intel.com/products/processor_number.
- 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
- Code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user
- Intel, Sponsors of Tomorrow and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

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

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the first quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “believes,” “seeks,” “estimates,” “may,” “will,” “should” and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel’s actual results, and variances from Intel’s current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the company’s expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions, including supply constraints and other disruptions affecting customers; customer acceptance of Intel’s and competitors’ products; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Uncertainty in global economic and financial conditions poses a risk that consumers and businesses may defer purchases in response to negative financial events, which could negatively affect product demand and other related matters. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel’s response to such actions; and Intel’s ability to respond quickly to technological developments and to incorporate new features into its products. Intel is in the process of transitioning to its next generation of products on 22nm process technology, and there could be execution and timing issues associated with these changes, including products defects and errata and lower than anticipated manufacturing yields. The gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; product mix and pricing; the timing and execution of the manufacturing ramp and associated costs; start-up costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; product manufacturing quality/yields; and impairments of long-lived assets, including manufacturing, assembly/test and intangible assets. The majority of Intel’s non-marketable equity investment portfolio balance is concentrated in companies in the flash memory market segment, and declines in this market segment or changes in management’s plans with respect to Intel’s investments in this market segment could result in significant impairment charges, impacting restructuring charges as well as gains/losses on equity investments and interest and other. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. Intel’s results could be affected by the timing of closing of acquisitions and divestitures. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting us from manufacturing or selling one or more products, precluding particular business practices, impacting Intel’s ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel’s results is included in Intel’s SEC filings, including the report on Form 10-Q for the quarter ended Oct. 1, 2011.

Rev. 1/19/12