NVMHCI: The Optimized Interface for Caches and SSDs

Amber Huffman
Intel

Agenda

- Hidden Costs of Emulating Hard Drives
- An Optimized Interface for Caches and SSDs

Hidden Costs of Serial ATA based SSDs

- Serial ATA SSDs take advantage of existing infrastructure to effectively bring SSDs to market
  - Slots are there, software is there
- However, Serial ATA SSDs are emulating hard drives which carries hidden costs
  - Complexity of the ATA command set
  - Tunneling over Serial ATA (an unnecessary intervening bus)
- There is also a power/latency penalty from tunneling over SATA
  - Roughly 200 mW during active transfers
  - Latency penalty of 10 µs to 10 milliseconds when coming out of a low power interface state to resume commands

What are the hidden costs in terms of performance and complexity of emulating HDDs?

ATA Has a Long History

- ATA (AT Attachment) defines the standard command set for hard drives
  - ATA was first developed by the X3T9.2 group starting in 1986
- ATA has adapted over the past 20 years to continue to serve the needs of the HDD industry
  - Features to serve industry needs (queuing, security, power management)
  - New commands or changes to commands to evolve with hard drives (e.g. CHS addressing to LBA addressing)

Portion of foreword from ATA-1 standard

> This standard defines the AT Attachment Interface. This standard defines an integrated bus interface between disk drives and host processors. It provides a common point of attachment for systems manufacturers, system integrators, and suppliers of intelligent peripherals.
>
> This standard was developed by Task Group X3T9.2 of Accredited Standards Committee X3 during 1986-90. The standards approval process started in 1991. This document includes annexes which are informative and are not considered par of the standard.

With the 20 year evolution of ATA, there is naturally a legacy burden

The ATA Command Set

- ATA-8 has well over 50 commands
  - Note that only a subset are mandatory (e.g. there are 4 mandatory read commands)
- Due to legacy infrastructure, devices are forced to continue to support unnecessary commands like PIO reads and writes
- This adds burden to SSD firmware designs: additional commands to support, additional code space consumed, additional latency to decode a more complex command set

ATA-7 Hard Drive Commands			
Read Commands	Write Commands	Power Management	Other
READ BUFFER	WRITE BUFFER	CHECK POWER MODE	CONFIGURE STREAM
READ DMA	WRITE DMA	IDLE	DEVICE CONFIGURATION
READ DMA EXT	WRITE DMA EXT	IDLE IMMEDIATE	DOWNLOAD MICROCODE
	WRITE DMA FUA EXT	SLEEP	EXECUTE DEVICE DIAGNOSTIC
READ DMA QUEUED	WRITE DMA QUEUED	STANDBY	FLUSH CACHE
READ DMA QUEUED EXT	WRITE DMA QUEUED EXT	STANDBY IMMEDIATE	FLUSH CACHE EXT
	WRITE DMA QUEUED FUA EXT		IDENTIFY DEVICE
READ FPDMA QUEUED	WRITE FPDMA QUEUED		NOP
READ MULTIPLE	WRITE MULTIPLE		READ LOG EXT
READ MULTIPLE EXT	WRITE MULTIPLE EXT		READ NATIVE MAX ADDRESS
	WRITE MULTIPLE FUA EXT		READ NATIVE MAX ADDRESS EXT
READ SECTOR(S)	WRITE SECTOR(S)		SERVICE
READ SECTOR(S) EXT	WRITE SECTOR(S) EXT		SET FEATURES
READ STREAM DMA EXT	WRITE STREAM DMA EXT		SET MAX ADDRESS
READ STREAM EXT	WRITE STREAM EXT		SET MAX ADDRESS EXT
READ VERIFY SECTOR(S)			SET MULTIPLE MODE
READ VERIFY SECTOR(S) EXT			SMART
			WRITE LOG EXT

Serial ATA Protocol Basics

- Serial ATA Native Command Queuing is the highest performance read protocol for SATA SSDs
  - Up to 32 read/write commands may be outstanding to the device, allowing for increased bus efficiency and re-ordering optimizations
- Each NCQ Read command includes:
  - H2D Register FIS to communicate Read FPDMA Queued command
  - D2H Register FIS to accept the command
  - DMA Setup FIS to setup DMA context for data transfer
  - Data FIS with up to 8KB of data per frame
  - Set Device Bits FIS to complete the command

Timestamp	Speed	Direction	FIS Type	Description	Tag
88.732 us	3 Gbps	H->D	FIS 27 - Cmd: 0x60=READ FPDMA QUEUED	LBA = 0x000000027B7C Sec Cnt = 0x0004	0x0D
89.672 us	3 Gbps	D->H	FIS 34 - Status: 0x40 - DRDY	LBA = 0x0027B7C Sec Cnt = 0x6C	
99.068 us	3 Gbps	D->H	FIS 41 - DMA Setup - A: 0 I: 0 D: 1	DMA Tx Count = 0x00000800	
100.224 us	3 Gbps	D->H	FIS 46 - Payload Data	Bytes Transferred: 2048	
107.928 us	3 Gbps	D->H	FIS A1 - Set Device Bit - I: 1 Err: 0x00	Status Hi: 0x04 Status Lo: 0x00 SActive: 0x00002000	

How efficient is the SATA protocol for SSDs?

Issue Command to Device

Timestamp	Speed	H->D Data	H->D Count	Host->Device Description	D->H Data	D->H Count	
88.396 us	3 Gbps	5757B57C		X_RDY			
88.408 us	3 Gbps	5757B57C		X_RDY			
88.552 us	3 Gbps				4A4A957C		R_RDY
88.564 us	3 Gbps				4A4A957C		R_RDY
88.732 us	3 Gbps	3737B57C		SOF			
88.744 us	3 Gbps	04608027	0	FIS 27 - Reg Host->Device			
				Features = 0x04=Obsolete			
				Command = 0x60=READ FPDMA QUEUED			
				C = 1 - Command Register Updated			
				PM Port = 0x0 - Default Port			
88.756 us	3 Gbps	40027B7C	1	Dev/Head = 0x40			
				Cyl High = 0x02			
				Cyl Low = 0x7B			
				Sec Num = 0x7C			
88.772 us	3 Gbps	00000000	2	Features(exp) = 0x00			
				Cyl High(exp) = 0x00			
				Cyl Low(exp) = 0x00			
				Sec Num(exp) = 0x00			
88.784 us	3 Gbps	00000068	3	Control = 0x00			
				Sec Cnt(exp) = 0x00			
				Sec Cnt = 0x68			
88.796 us	3 Gbps	00000000	4	Reserved			
88.812 us	3 Gbps	8373C6EF	5				
88.824 us	3 Gbps			CRC - Good			
		D5D5B57C		EOF			
88.836 us	3 Gbps	5858B57C		WTRM			
88.852 us	3 Gbps	5858B57C		WTRM			
88.872 us	3 Gbps				5555B57C		R_IP
88.884 us	3 Gbps				5555B57C		R_IP
88.992 us	3 Gbps				3535B57C		R_OK

Overhead

596 ns for H2D Register FIS

Command Accepted by Device

Timestamp	Speed	H->D Data	H->D Count	Host->Device Description	D->H Data	D->H Count	Device->Host Description
89.004 us	3 Gbps				3535B57C		R_OK
89.324 us	3 Gbps				5757B57C		X_RDY
89.340 us	3 Gbps				5757B57C		X_RDY
89.492 us	3 Gbps	4A4A957C		R_RDY			
89.504 us	3 Gbps	4A4A957C		R_RDY			
89.672 us	3 Gbps				3737B57C		SOF
89.684 us	3 Gbps				00400034	0	FIS 34 - Reg Device->Host
							DRDY
							BSY = 0, DRDY = 1, DF = 0, DSC = 0,
							DRQ = 0, CORR = 0, IDX = 0, ERR = 0
							Error = 0x00
							I = 0
							PM Port = 0x0 - Default Port
89.700 us	3 Gbps				40027B7C	1	Dev/Head = 0x40
							Cyl High = 0x02
							Cyl Low = 0x7B
							Sec Num = 0x7C
89.712 us	3 Gbps				00000000	2	Cyl High(exp) = 0x00
							Cyl Low(exp) = 0x00
							Sec Num(exp) = 0x00
89.720 us	3 Gbps	4A4A957C		R_RDY			
89.724 us	3 Gbps				0000006C	3	Sec Cnt(exp) = 0x00
							Sec Cnt = 0x6C
89.732 us	3 Gbps	4A4A957C		R_RDY			
89.736 us	3 Gbps				00000000	4	Reserved
89.752 us	3 Gbps				F2D77DD4	5	
89.764 us	3 Gbps						CRC - Good
					D5D5B57C		EOF
89.776 us	3 Gbps				5858B57C		WTRM
89.792 us	3 Gbps				5858B57C		WTRM
89.852 us	3 Gbps	5555B57C		R_IP			
89.864 us	3 Gbps	5555B57C		R_IP			
89.960 us	3 Gbps	3535B57C		R_OK			

Overhead

320 ns R_OK to X_RDY latency

636 ns for D2H Register FIS

Device Says *Let's Do the Data Tango*

Timestamp	Speed	H->D Data	H->D Count	Host->Device Description	D->H Data	D->H Count	Device->Host Description
98.720 us	3 Gbps				5757B57C		X_RDY
98.732 us	3 Gbps				5757B57C		X_RDY
98.880 us	3 Gbps	4A4A957C		R_RDY			
98.896 us	3 Gbps	4A4A957C		R_RDY			
99.068 us	3 Gbps				3737B57C		SOF
99.080 us	3 Gbps				00002041	0	FIS 41 - DMA Setup
							A = 0
							I = 0 D = 1
							PM Port = 0x0 - Default Port
99.092 us	3 Gbps				0000000D	1	Buf ID L = 0x0000000D
99.104 us	3 Gbps				00000000	2	Buf ID H = 0x00000000
99.120 us	3 Gbps				00000000	3	
99.132 us	3 Gbps				00000000	4	Buf Offset = 0x00000000
99.144 us	3 Gbps				00000800	5	Tx Count = 0x00000800
99.160 us	3 Gbps				00000000	6	Reserved
99.172 us	3 Gbps				48724D2D	7	
99.184 us	3 Gbps						CRC - Good
					D5D5B57C		EOF
99.200 us	3 Gbps				5858B57C		WTRM
99.212 us	3 Gbps				5858B57C		WTRM
99.240 us	3 Gbps	5555B57C		R_IP			
99.256 us	3 Gbps	5555B57C		R_IP			
99.388 us	3 Gbps	3535B57C		R_OK			

Overhead

668 ns for DMA Setup FIS

Device Completes Command

Timestamp	Speed	H->D Data	H->D Count	Host->Device Description	D->H Data	D->H Count	Device->Host Description
107.388 us	3 Gbps	3535B57C		R_OK			
107.568 us	3 Gbps				5757B57C		X_RDY
107.580 us	3 Gbps				5757B57C		X_RDY
107.736 us	3 Gbps	4A4A957C		R_RDY			
107.752 us	3 Gbps	4A4A957C		R_RDY			
107.928 us	3 Gbps				3737B57C		SOF
107.940 us	3 Gbps				004040A1	0	FIS A1 - Set Device Bit
							Error = 0x00
							Status Hi = 0x4
							Status Lo = 0x0
							I = 1
							PM Port = 0x0 - Default Port
107.956 us	3 Gbps				00002000	1	SActive
107.968 us	3 Gbps				43B8CA03	2	CRC
107.980 us	3 Gbps						CRC - Good
					D5D5B57C		EOF
107.996 us	3 Gbps				5858B57C		WTRM
108.008 us	3 Gbps				5858B57C		WTRM
108.112 us	3 Gbps	5555B57C		R_IP			
108.124 us	3 Gbps	5555B57C		R_IP			
108.192 us	3 Gbps	3535B57C		R_OK			

Overhead

180 ns R_OK to X_RDY latency

624 ns for Set Device Bits FIS

4KB Read SATA Bus Efficiency

- SSDs are an enormous jump in performance over hard drives
- As SSDs become mainstream, the protocol needs to be streamlined to account for this new level of performance
- The SATA bus overhead is 15% for a 4KB sequential read
  - Total time for 4KB read was 19.796 µs
  - Total SATA overhead was 3.024 µs
  - This does not account for any firmware processing overhead for SATA packets or ATA commands
- 3 µs is fantastic bus overhead for traditional HDDs, but the game changes with SSDs…

With SSDs, microseconds matter!

Agenda

- Hidden Costs of Emulating Hard Drives
- *An Optimized Interface for Caches and SSDs*

An Optimized Interface for NVM

- NVMHCI: Non-Volatile Memory Host Controller Interface
- NVMHCI is a clean and optimized interface for SSDs and caches
- NVM equivalent of the SATA AHCI controller interface

Quick Links | Home | Worldwide

Microsoft

Search Microsoft.com for: Go

PressPass - Information for Journalists

PressPass Home | PR Contacts | Fast Facts About Microsoft | Site Map | Advanced Search | RSS Feeds

Microsoft News
Product News
Consumer News
International Contacts
Legal News
Security & Privacy News
Events
News Archive

Corporate Information
Microsoft Executives
Fast Facts About Microsoft
Image Gallery
Broadcast Room

Related Sites
Analyst Relations
Community Affairs
Essays on Technology
Executive E-Mail
Global Citizenship
Investor Relations
Microsoft Research

The PressPass Broadcast Room
Broadcast-standard media for download

PressPass Subscriptions RSS

Dell, Intel and Microsoft Join Forces to Increase Adoption of NAND-Based Flash Memory in PC Platforms

Newly formed group to provide standard interface for nonvolatile memory subsystems.

REDMOND, Wash. — May 30, 2007 — Broad adoption of NAND flash memory technology in the PC platform received a boost with the formation of the Non-Volatile Memory Host Controller Interface (NVMHCI) Working Group. The NVMHCI Working Group is chaired by Intel Corporation with core contributors including Dell Inc. and Microsoft Corp.

Related Links
External Resources:
- Dell Web site
- Intel Web site

NVMHCI will provide a standard software programming interface for nonvolatile memory subsystems. The interface would be used by operating system drivers to access NAND flash memory storage in applications such as hard drive caching and solid-state drives.

"Several NAND solutions are coming on the scene to take advantage of the ReadyBoost™ and ReadyDrive™ features of the Windows Vista® operating system," said Bob Rinne, general manager of Windows Hardware Ecosystem at Microsoft. "Standardizing on a common controller interface will enable more integrated operating system support of these solutions moving forward."

Industry momentum for standardization in NAND storage solutions is building, especially as NAND moves into the PC platform. NVMHCI complements standardization work being done in the Open NAND Flash Interface (ONFI) Working Group.

"We've got a performance-enhancing NAND-based product in the market with our new Centrino mobile technology platform called Intel Turbo memory, and this newly formed working group will help make that and a number of other NAND-based solutions more prolific, faster," said Rick Coulson, senior fellow and director of I/O Architecture at Intel. "ONFI formed last year to standardize the interface between the Flash controller and the NAND itself, and standardizing the register level interface between the Flash controller and the operating system driver is the logical next step."

"Nonvolatile memory solutions enable better system performance and lower power consumption as well as facilitate additional benefits such as smaller form factors, quieter systems and improved robustness," said Liam Quinn, director of communications for technology strategy and architecture at Dell. "Dell looks forward to working with industry partners and extending the benefits NVMHCI will bring to our customers."

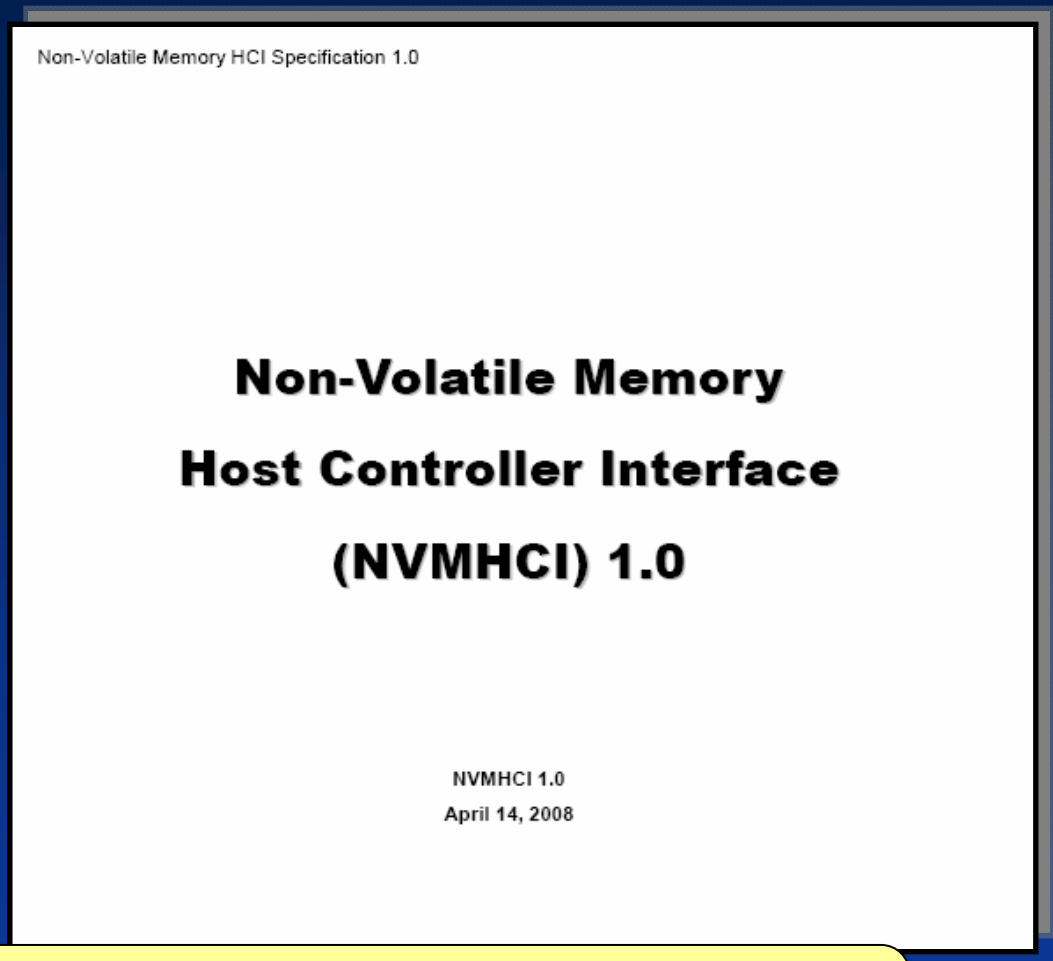
The group is actively expanding its membership to include other industry-leading companies

NVMHCI Membership

NVMHCI continues to grow with over 35 members.

NVMHCI 1.0 Specification Ratified

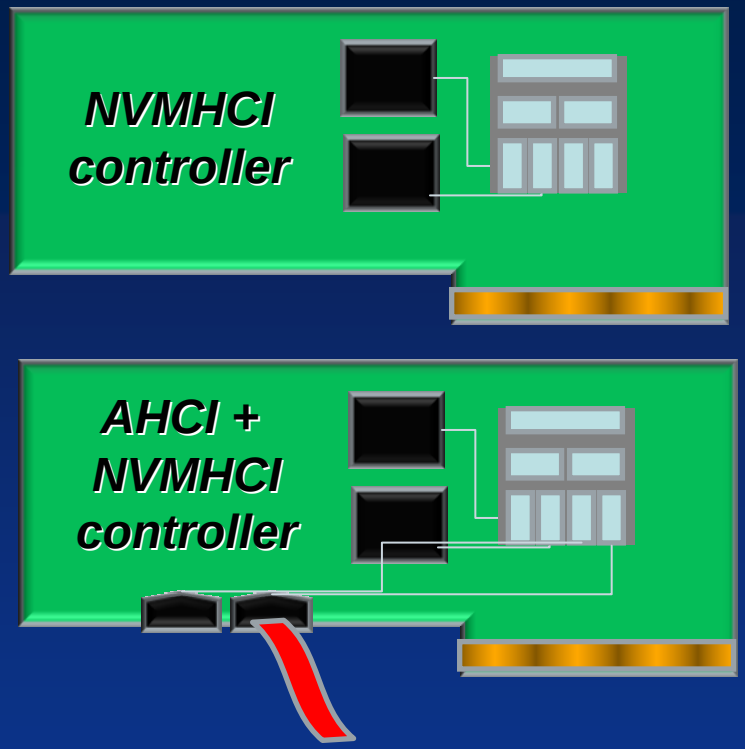
- NVMHCI 1.0 complete!
- Less than one year from team formation to ratification
- Includes register set, DMA engine, and command set definitions



Available for download at:
http://www.intel.com/standards/nvmhci

Technical Essence of NVMHCI

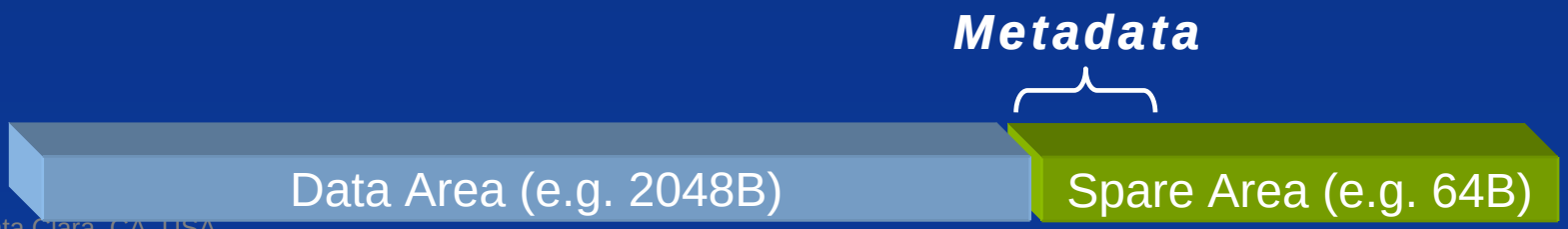
- NVMHCI defines a standard programming interface for non-volatile memory subsystems
- Leverage AHCI to provide best infrastructure for caching
  - One driver for HDDs and NAND
- Allows NVMHCI registers to appear as:
  - A separate PCI device
  - A port within an existing AHCI controller



- NVMHCI is a logical interface
  - All NAND management abstracted out: NAND technology changes too quickly
  - All caching algorithms are outside the spec: NVMHCI only defines how caching software gets access to the NAND
- Optimized interface for both cache and SSD usage models

Optimizing for Cache: Atomic Metadata

- NAND pages each have spare area
  - Used primarily for storing ECC syndromes
  - Also used for storing some NAND management or caching information
- NVMHCI exposes some spare area as metadata to the host
  - May be used in caching applications
    - e.g. What disk LBA is this data associated with?
  - Metadata is on an NVM page basis
  - Written atomically with the NVM page
  - The host may use metadata as it chooses
- Atomic metadata is not available in traditional HDD interfaces, making NVMHCI ideally suited for caches



Optimizing for NVM

- NVMHCI has eight commands total
  - One read, one write

Enumeration	Identify
Configuration	Get Features, Set Features
Health Monitoring	Get Status
IO	Read, Write, Flush
Management	Data Set Management

- NVMHCI provides priority per command and information on the subsequent workload for better NVM subsystem optimizations
- NVMHCI added a Physical Region Descriptor (PRD) Index Table so that unaligned writes could be optimized for cached/multi-plane programs
  - Out of order data delivery in SATA NCQ not used due to difficulty in host walking PRD table on each DMA Setup

PR: Priority	Indicates the priority of the request: The NVM subsystem may use this information to help determine the command to service
UR: Upcoming Reads	Indicates the number of queued read r are yet to be issued by host software. requests may be for any sector locatio
UW: Upcoming Writes	Indicates the number of queued write r are yet to be issued by host software. requests may be for any sector locatio
Avg WS: Average Write Size	This field indicates the average size in NVM pages for the queued write reque yet to be issued by host software.

	31	23	15	7	0
NVM Page n	Dword Offset		PRD Entry		
NVM Page n+1	Dword Offset		PRD Entry		
NVM Page ...	...				
NVM Page n + x	Dword Offset		PRD Entry		

Optimizing for NVM, continued

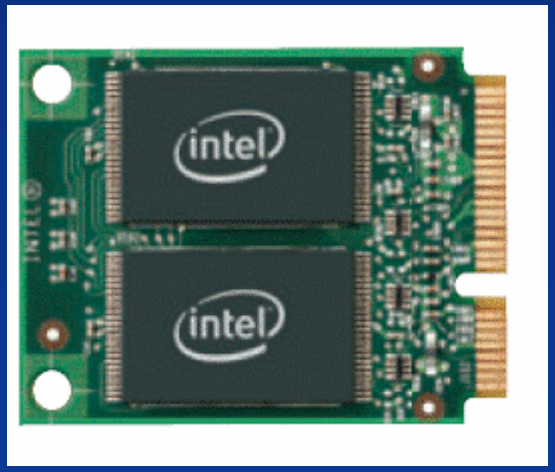
- NVMHCI allows commands to be executed out of order, and includes priority and timeout information
- NVMHCI allows interrupt combining on a per-command basis
  - In AHCI, every command interrupts on completion
- Dataset management enables performance, latency, and endurance to be optimized by the NVMHCI controller
  - Dataset management communicates read/write frequency, read/write latency, access size, deletes/trims for LBA ranges

WF: Write Frequency	05:04	

Value	Definition
00b	No write frequency information given.
01b	Long term storage. Written less than once on average per NVM device power cycle.
10b	User's current working set. Written once on average every NVM device power cycle.
11b	Dynamic data. Written more than once on average per NVM device power cycle.

NVMHCI in the Grand Scheme

- NVMHCI is a register interface and command set used by software drivers to communicate with NVM
- NVMHCI does not define the underlying NVM hardware architecture used
  - Could be discrete PCIe card
  - Could be a direct NAND interface (e.g. ONFI)



Hardware interface to NAND is NVMHCI implementation specific.

Summary

- Serial ATA SSDs deliver great performance beyond hard drives and take advantage of infrastructure in place today
- We can do even better with NVMHCI
  - Streamlined command set
  - Optimized features (metadata, PRD Index Table, dataset management, etc.)
  - Combined with AHCI for best caching interface (one driver controlling HDDs and cache)
- The NVMHCI 1.0 specification and information on joining the committee is at www.intel.com/standards/nvmhci

Get involved with NVMHCI today!

Risk Factors

This presentation contains forward-looking statements that involve a number of risks and uncertainties. These statements do not reflect the potential impact of any mergers, acquisitions, divestitures, investments or other similar transactions that may be completed in the future. The information presented is accurate only as of today's date and will not be updated. In addition to any factors discussed in the presentation, the important factors that could cause actual results to differ materially include the following: Demand could be different from Intel's expectations due to factors including changes in business and economic conditions, including conditions in the credit market that could affect consumer confidence; 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. Intel's results could be affected by the timing of closing of acquisitions and divestitures. 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 new 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; Intel's ability to respond quickly to technological developments and to incorporate new features into its products; and the availability of sufficient supply of components from suppliers to meet demand. The gross margin percentage could vary significantly from expectations based on changes in revenue levels; product mix and pricing; capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; excess or obsolete inventory; manufacturing yields; changes in unit costs; impairments of long-lived assets, including manufacturing, assembly/test and intangible assets; and the timing and execution of the manufacturing ramp and associated costs, including start-up costs. Expenses, particularly certain marketing and compensation expenses, vary depending on the level of demand for Intel's products, the level of revenue and profits, and impairments of long-lived assets. Intel is in the midst of a structure and efficiency program that is resulting in several actions that could have an impact on expected expense levels and gross margin. Intel's results could be impacted by adverse economic, social, political and physical/infrastructure conditions in the countries in which 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. 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. 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 June 28, 2008.

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. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.
- Intel may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards 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.
- 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
- Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.
- Intel, Intel Inside 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.
-