Standards Enabling Enterprise Class SSDs

Amber Huffman
Principal Engineer

MEMS001

Sponsors of Tomorrow. intel

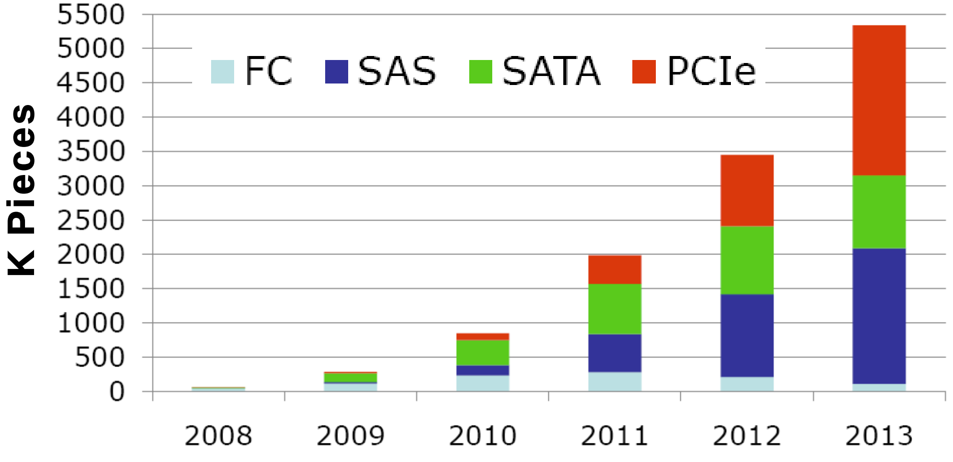
Agenda

- **Enterprise NVMHCI**
- ONFi 3.0
- Summary

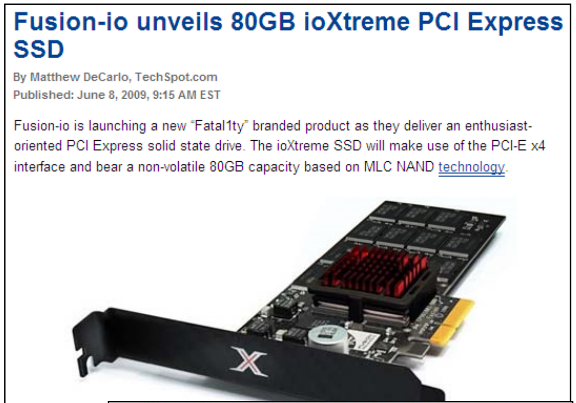
PCI Express* Technology SSD Trend

- PCI Express (PCIe) Technology SSDs are happening
  - Plentiful PCIe lanes
  - Enables > 600 MB/s in one device
  - Lower latency (µsec matter)
  - Opportunity to minimize infrastructure cost with direct attach

Enterprise SSD Outlook by Interface



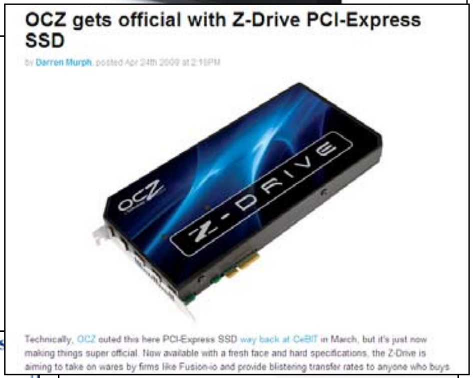
Source: Preliminary Gartner estimates, 12/09



Fusion-io unveils 80GB ioXtreme PCI Express SSD

By Matthew DeCarlo, TechSpot.com
Published: June 8, 2009, 9:15 AM EST

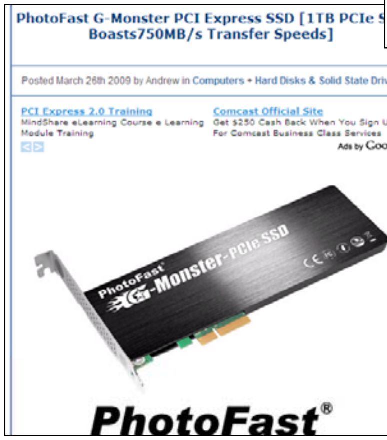
Fusion-io is launching a new "Fatal1ty" branded product as they deliver an enthusiast-oriented PCI Express solid state drive. The ioXtreme SSD will make use of the PCI-E x4 interface and bear a non-volatile 80GB capacity based on MLC NAND technology.



OCZ gets official with Z-Drive PCI-Express SSD

by Darren Murph, posted Apr 24th 2009 at 2:15PM

Technically, OCZ outed this here PCI-Express SSD way back at CeBIT in March, but it's just now making things super official. Now available with a fresh face and hard specifications, the Z-Drive is aiming to take on wares by firms like Fusion-io and provide blistering transfer rates to anyone who buys



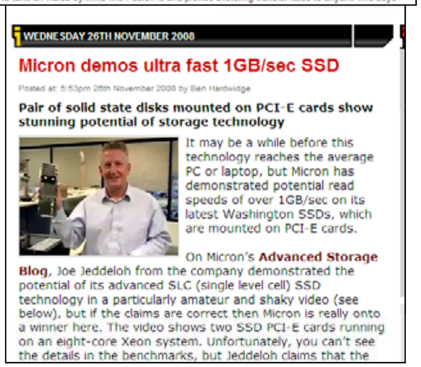
PhotoFast G-Monster PCI Express SSD [1TB PCIe S Boasts750MB/s Transfer Speeds]

Posted March 26th 2009 by Andrew in Computers • Hard Disks & Solid State Drives

PCI Express 2.0 Training
MindShare eLearning Course e Learning Module Training

Comcast Official Site
Get $250 Cash Back When You Sign Up For Comcast Business Class Services

Ads by Google



WEDNESDAY 26TH NOVEMBER 2008

Micron demos ultra fast 1GB/sec SSD

Posted at: 5:53pm 26th November 2008 by Ben Hardwidge

Pair of solid state disks mounted on PCI-E cards show stunning potential of storage technology

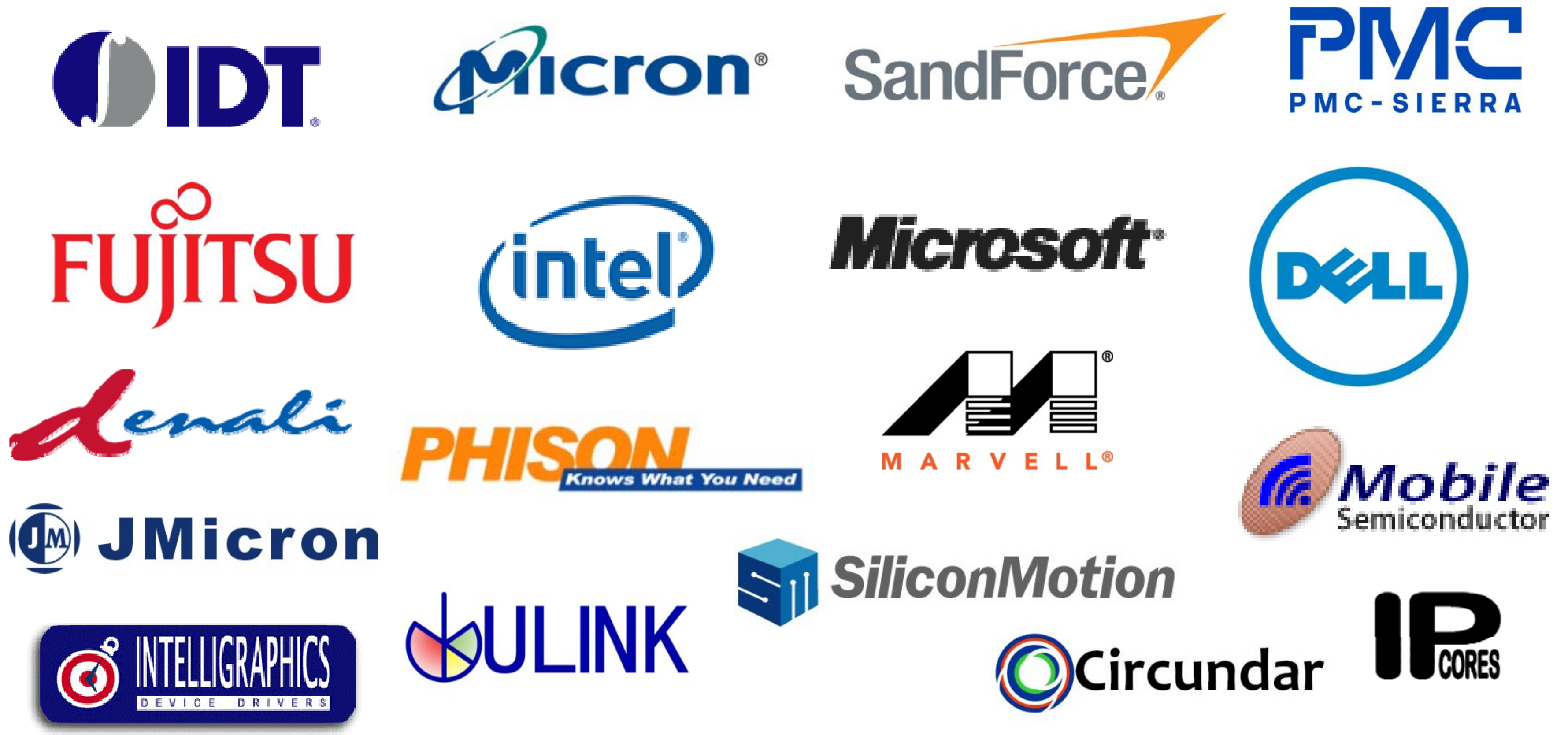
It may be a while before this technology reaches the average PC or laptop, but Micron has demonstrated potential read speeds of over 1GB/sec on its latest Washington SSDs, which are mounted on PCI-E cards.

On Micron's **Advanced Storage Blog**, Joe Jeddeloh from the company demonstrated the potential of its advanced SLC (single level cell) SSD technology in a particularly amateur and shaky video (see below), but if the claims are correct then Micron is really onto a winner here. The video shows two SSD PCI-E cards running on an eight-core Xeon system. Unfortunately, you can't see the details in the benchmarks, but Jeddeloh claims that the

Adoption Challenge

- The PCI Express* (PCIe) Technology SSDs that are emerging do not have a standard host controller interface (i.e., register i/f)
  - This requires each vendor to provide a driver with the SSD
- Lack of standard drivers is a challenge to PCIe SSD adoption
  - Requires each SSD vendor to provide a driver for each OS
  - Requires OEMs to validate each SSD with its own driver, increasing validation time and cost
  - Makes adoption of PCIe SSDs more difficult
- To resolve this, industry leaders are defining Enterprise NVMHCI
  - Standard host controller interface (register programming interface) for Enterprise class PCIe SSDs
  - Addresses Enterprise server scenarios with a streamlined and efficient interface enabling very high IOPs

Industry Leaders Driving Enterprise NVMHCI



The Workgroup includes 50+ member companies, continuing to grow.

The Value of Enterprise NVMHCI

Microsoft®

"A standardized interface functions as a foundation, enabling a volume market for technology innovation while avoiding the compatibility issues that arise from multiple, proprietary interfaces. Enterprise customers are requesting standard interfaces be used on non-volatile-memory products as an enabler to assist broad adoption."

John Loveall
Manager of Program Management, Storage and File Systems
Microsoft

The Value of Enterprise NVMHCI

Microsoft

"A standardized interface functions as a foundation, enabling a volume market f
compatibility is
interfaces. E
interfaces be
enabler to ass

Manager of F

FUJITSU

"The lack of a standard register level interface presents numerous problems when integrating PCIe SSDs into our products, including longer qualification times and functionality that is not uniformly implemented across vendors. Fujitsu Technology Solutions sees Enterprise NVMHCI as an important part of enabling broad adoption in PCIe SSDs emerging in the Enterprise space by resolving these concerns. Joining the working group was a natural choice to foster this industry leading standardization effort."

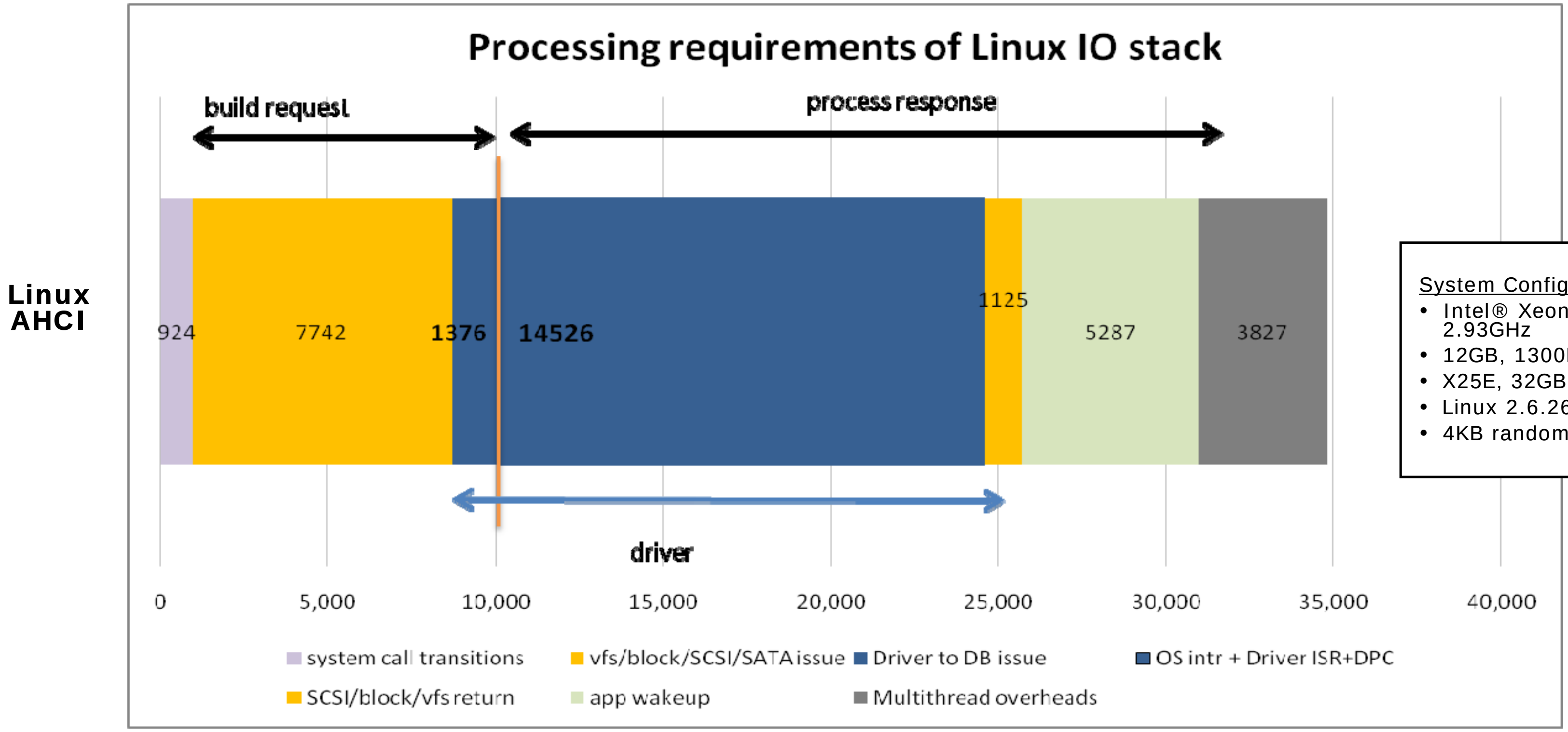
Jens-Peter Seick
Senior Vice President x86 Server Product Unit
Fujitsu Technology Solutions

Enterprise NVMHCI Goals

- Goals
  - Address Enterprise server scenarios
  - Enable an efficient interface that can achieve very high IOPs
  - Enable OS vendors to deliver a standard & high performance driver for all PCI Express* (PCIe) Technology SSDs
  - Enable OEMs to qualify / validate a single driver on each OS
  - Ensure features are implemented in a consistent fashion, and thus reduce time to market for PCIe SSDs
- To realize goals quickly
  - Utilize existing & efficient NVMHCI Workgroup team
  - Leverage from NVMHCI 1.0 where it makes sense
  - Take advantage of extending drivers already written or underway for NVMHCI 1.0

Optimization Points, example

- The Linux* stack using AHCI is ~ 35,000 clocks / IO
- A large impact is uncacheable reads, ~ 2000 clocks each
  - Minimum of 4 uncacheable reads required with AHCI
- Enterprise NVMHCI is eliminating uncacheable reads for command issue/completion



System Config

- Intel® Xeon® X5570 2.93GHz
- 12GB, 1300MHz DRAM
- X25E, 32GB
- Linux 2.6.26
- 4KB random RD IO

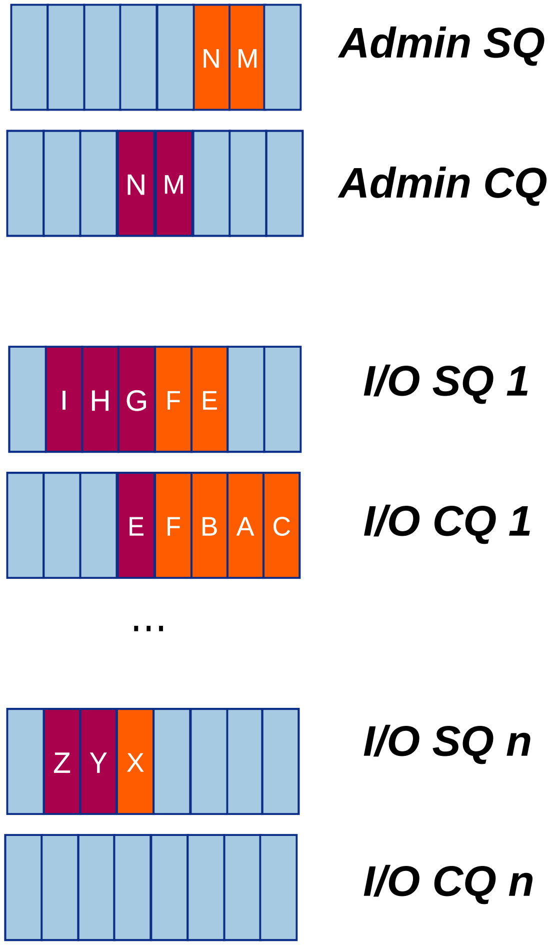
Source: Intel internal analysis

Interface Attributes

- Eliminate performance bottlenecks seen in other interfaces
  - Eliminate uncacheable reads from command issue/completion
  - Ensure maximum of 1 MMIO write in command issue path
  - Support deep command queues
  - Avoid “pointer chasing”, aggregate command information
- Support efficient and streamlined command set
  - ~10 to 15 optimized NVM commands
  - Do not carry forward HDD command set legacy
- Support for MSI-X and interrupt aggregation
- Support for IO virtualization (e.g. SR-IOV)
- Efficient error handling & *driver* translation into SCSI management architectures prevalent in Enterprise

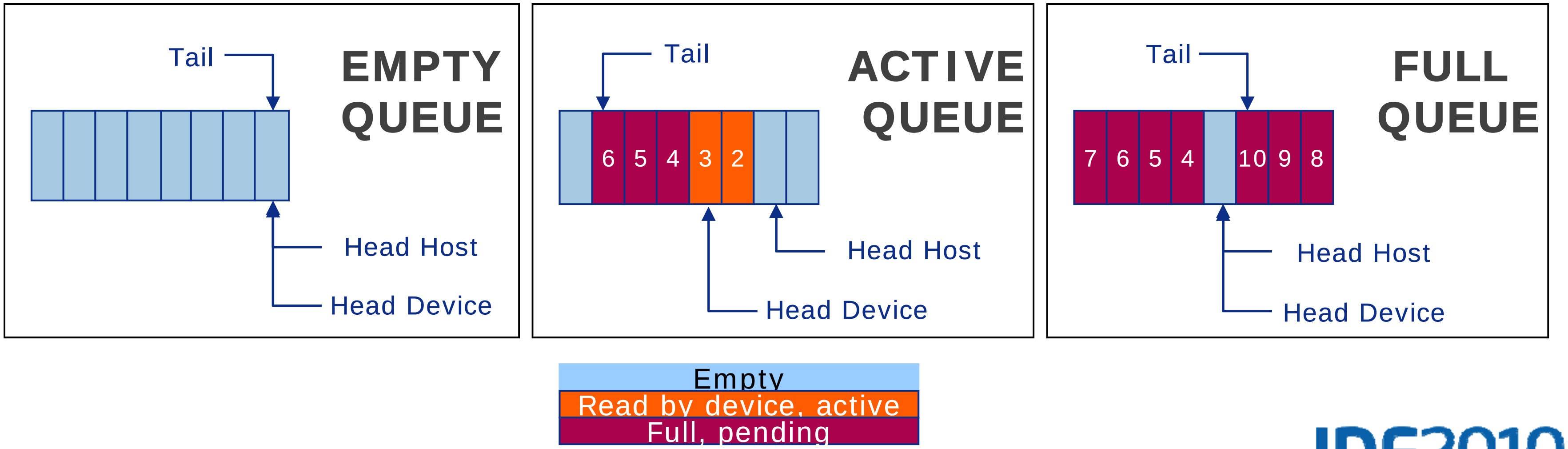
Paired Queue Mechanism

- Command submission & completion use a paired queue mechanism (submission and completion queues)
- The Admin queue carries out functions that impact the entire device
  - E.g. Queue creation and deletion, command abort
- Driver creates the number of queues that match system config & workload
  - E.g. On a 4 core system, devote a queue pair per core to avoid locking and ensure structures in right core's cache



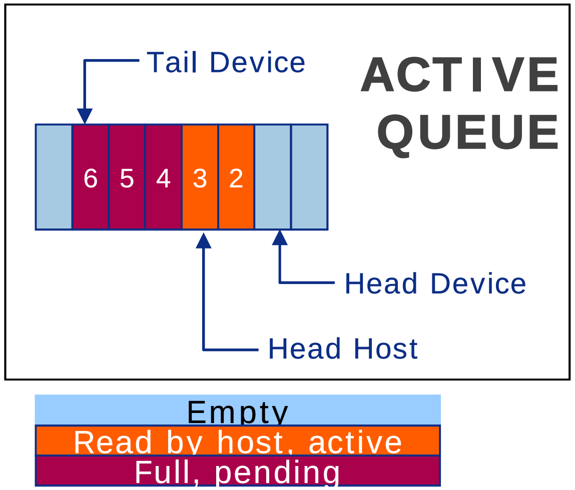
Submission Queue Details

- A submission queue (SQ) is a circular buffer with a fixed slot size that the host uses to submits commands for execution
- The host updates an SQ Tail doorbell register when there are 1 to n new commands to execute
  - The old SQ Tail value is simply overwritten in the device
- The device reads SQ entries in order and removes them from the SQ, then may execute those commands out of order

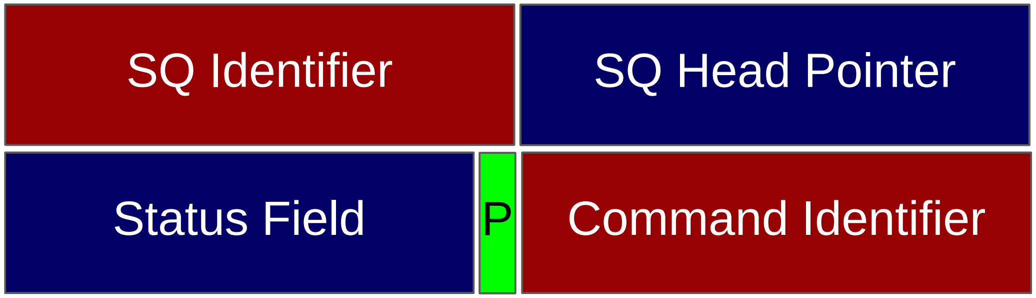


Completion Queue Details

- A completion queue (CQ) is a circular buffer with a fixed slot size that the device posts status to for completed commands
- The device identifies the command that completed by the SQ Identifier and the Command Identifier (assigned by software)
  - Multiple SQs can use the same completion queue
- The latest SQ Head pointer is returned in the status to avoid a register read for this information
- The Phase (P) bit indicates whether an entry is new, and inverts each pass through the circular buffer



Completion Queue Entry



+ 2 reserved Dwords

Register Set

- The device indicates capabilities and version

- Interrupts are indicated via register for device failures only
  - Normal command driven interrupts processed via CQ

- Admin Queue configuration is via registers
  - All other queues are created via Admin Queue commands

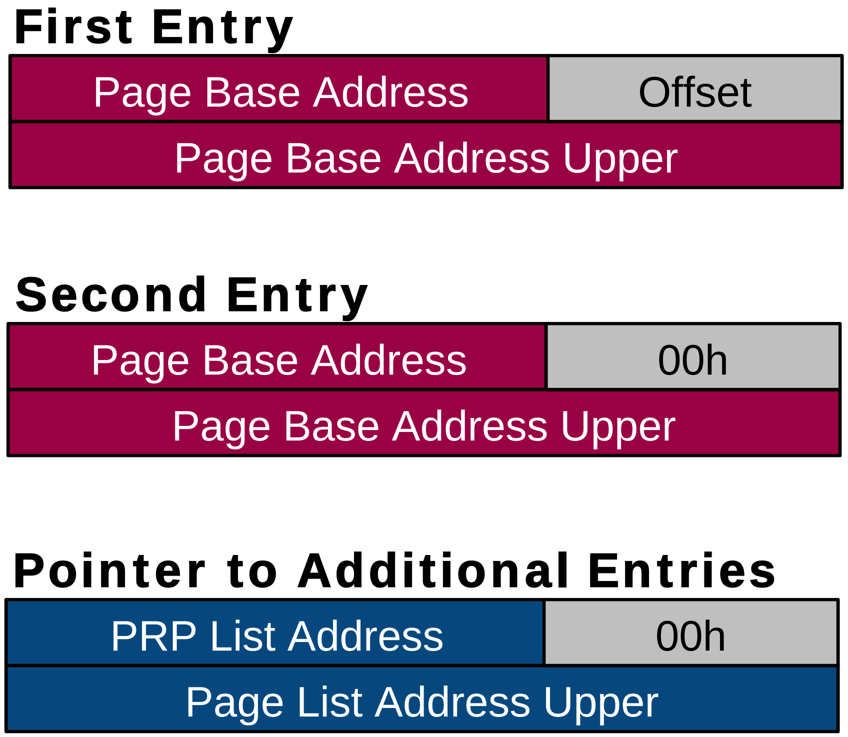
- Scalable to large number of SQ/CQ based on requirements

- Single “port” per virtual device, with multiple LBA namespace access

Symbol	Description
CAP	Controller Capabilities
VS	Version
IS	Interrupt Status
CMD	Command and Status
Reserved	Reserved
AQA	Admin Queue Attributes
ASQ	Admin Submission Queue Base Address
ACQ	Admin Completion Queue Base Address
SQ0TDBL	Submission Queue 0 Tail Doorbell (Admin)
CQ0HDBL	Completion Queue 0 Head Doorbell (Admin)
SQ1TDBL	Submission Queue 1 Tail Doorbell
CQ1HDBL	Completion Queue 1 Head Doorbell
SQ2TDBL	Submission Queue 2 Tail Doorbell
CQ2HDBL	Completion Queue 2 Head Doorbell
…	…
SQyTDBL	Submission Queue y Tail Doorbell
CQyHDBL	Completion Queue y Head Doorbell

Data Transfer Optimization

- Out of order data transfer is important for SSDs
- Walking a scatter/gather list (SGL) to determine where data starts for a transfer is inefficient
- A fixed size SGL entry enables efficient out of order data & simplifies hardware
- Better approach: Page lists
  - First entry contains an offset
  - “Middle” entries are full page in size
  - Last entry may be less than a page
- The command includes three entries to optimize for 4KB & 8KB I/O
  - For a larger transfer, third entry points to a list of entries


First Entry

Page Base Address | Offset

Page Base Address Upper

Second Entry

Page Base Address | 00h

Page Base Address Upper

Pointer to Additional Entries

PRP List Address | 00h

Page List Address Upper

Timeline & Opportunity

- Schedule

– Apr 2010:	0.5 revision
– Jul 2010:	0.7 revision
– Sep 2010:	0.9 revision (erratum only after this point)
– Oct 2010:	Release candidate (30-day review)
– Nov 2010:	1.0 release

- To get involved in the specification definition, join the NVMHCI Workgroup
  - Details at http://www.intel.com/standards/nvmhci

Schedule enables product intercept in 2012.

Agenda

- Enterprise NVMHCI
- **ONFi 3.0**
- Summary

ONFi Workgroup History & Results

- NAND was the only commodity memory with no standard i/f
- The Open NAND Flash Interface (ONFi) Workgroup was formed in May 2006 to drive standardization for the raw NAND Flash interface

	Q3 '06	Q4 '06	Q1 '07	Q2 '07	Q3 '07	Q4 '07	Q1 '08	Q2 '08	Q3 '08	Q4 '08	Q1 '09	Q2 '09	Q3 '09	Q4 '09	Q1 '10
Specification	ONFI 1.0			ONFI 2.0					ONFI 2.1			ONFI 2.2			
Features	Standard electrical & protocol interface, including base command set			Defined a high speed DDR interface, tripling the traditional NAND bus speed in common use					Included additional features and support for bus speeds up to 200 MB/s			Added features including LUN reset, enhanced program page register clear, and ICC measurement specs			
Maximum Speed	50 MB/s			133 MB/s					200 MB/s			200 MB/s			
Other Activities		Block Abstracted NAND 1.0						ONFI – JEDEC Collaboration							

ONFi
OPEN NAND FLASH INTERFACE

hynix SONY numonyx Micron
intel SanDisk PHISON Knows What You Need SPANSION

ONFi **Members**

A-Data
AboUnion Technology
Afa Technologies
Alcor Micro
Aleph One
Anobit Tech.
Apacer
Arasan Chip Systems
ASMedia Technology
ATI
Avid Electronics
BitMicro
Biwin Technology
Chipsbank
Cypress
DataFab Systems
Data I/O
Datalight
Denali Software
Densbits Technologies
ENE Technology
Entorian
Eonsil LLC
Evatronix
FCI
FormFactor
Foxconn
Fresco Logic
Fusion Media Tech
Genesys Logic
Hagiwara Sys-Com
HiperSem
Hitachi GST
Hyperstone
InCOMM
Indilinx
Inphi
Intelliprop
ITE Tech
Jinvani Systech
Kingston Technology
Lauron Technologies
Lotes
LSI
Macronix
Marvell
MemoCom Corp
Mentor Graphics
Metaram
Moai Electronics
Mobile Semiconductor
Molex
Nvidia
Orient Semiconductor
P.A. Semi
PMC Sierra
Power Quotient Int.
Powerchip Semi.
Prolific Technology
Qimonda
SandForce
Seagate
Shenzhen Netcom
Sigmatel
Silicon Integrated Sys.
Silicon Motion
Silicon Storage Tech.
STEC
Skymedi
Smart Modular Tech.
Solid State System
Super Talent Elec.
Synopsys
Tandon
Tanisys
Telechips
Teradyne
Testmetrix
Transcend Information
Tyco
UCA Technology
University of York
Viking InterWorks
Vident Systems
Western Digital
WinBond

ONFI enjoys support of 90+ members.

ONFi Workgroup Development

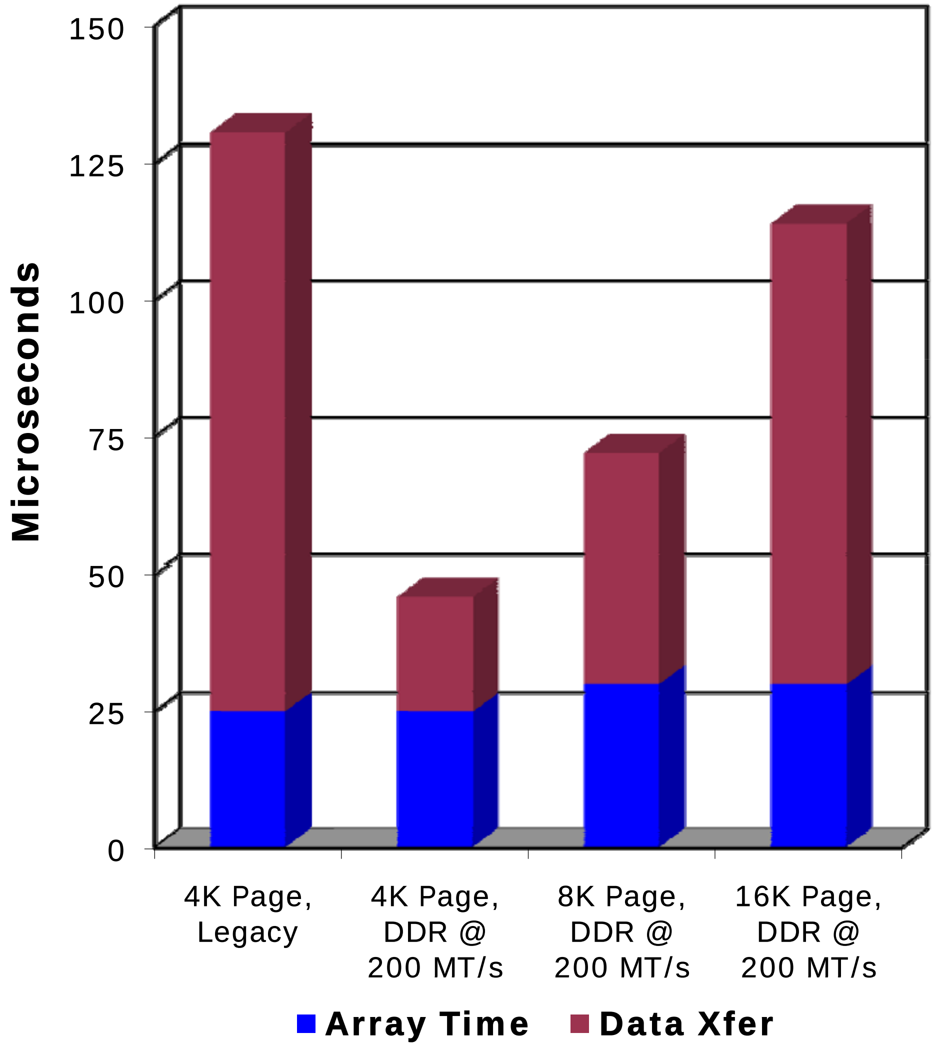
- The ONFi Workgroup has two active development efforts

- ONFi 3.0
  - Scales the NAND interface from 200 MT/s to 400 MT/s
  - Backwards compatible with ONFi 1.x and 2.x

- EZ NAND (Error Correction Code Zero NAND)
  - Enables NAND lithography dependent functions to be separated from the host
  - ECC (and other functions) can be offloaded into an ASIC stacked in the NAND package, while still using normal NAND protocol (Program, Erase, etc)
  - Alleviates the burden from the host of keeping pace with rapidly changing NAND if desired

Continuing Scaling Requirement

- As NAND lithographies shrink, page sizes increase
- The page size increases for performance scaling
  - Array times slow down a bit, so send more data in each page to compensate
- The NAND bus needs to keep pace with page scaling, or the bus will be the limiting factor
  - E.g. 16KB page data transfer time is **2.5X** the array time

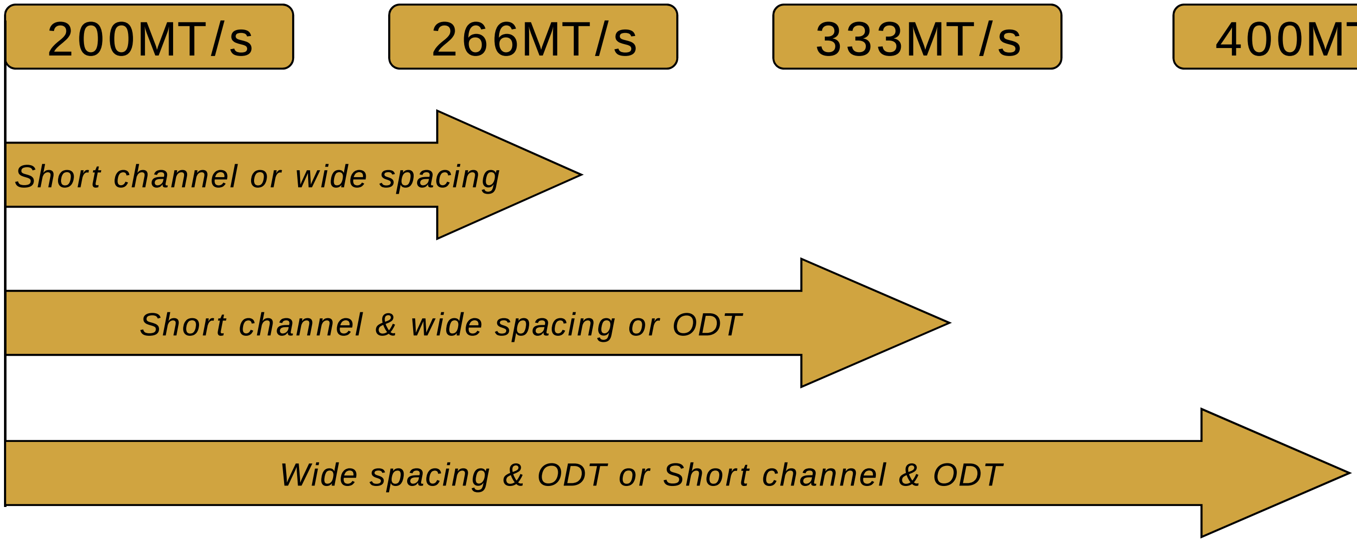


*Array time based on SLC NAND vendor data sheets.

The Path to 400 MT/s

- Key enablers for ONFi 3.0 include:
  - A shorter channel (controller distance to the NAND)
  - Wider spacing between signals
  - On-die termination
  - Complementary clock and DQS signals

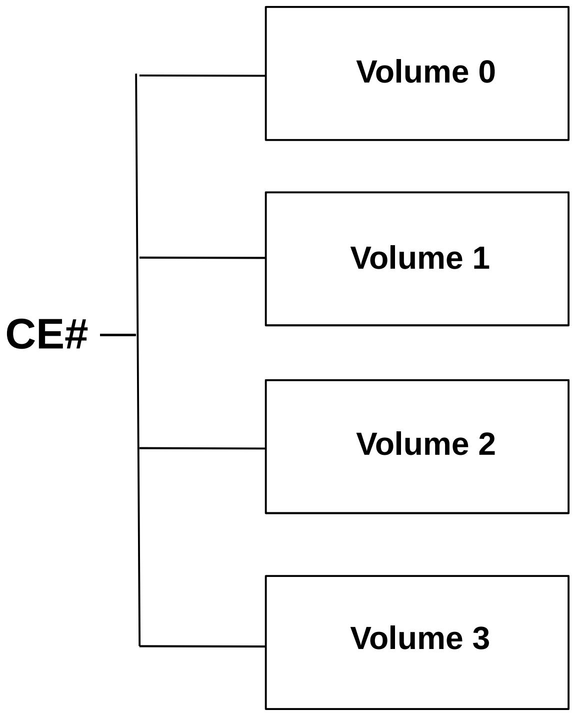
- Reaching speed grades requires a combination of enablers


200MT/s
266MT/s
333MT/s
400MT/s
Short channel or wide spacing
Short channel & wide spacing or ODT
Wide spacing & ODT or Short channel & ODT

IDF2010
INTEL DEVELOPER FORUM

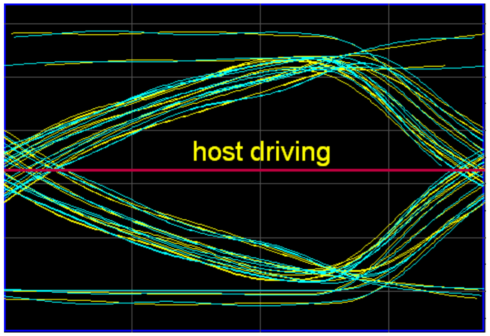
CE# Pin Reduction

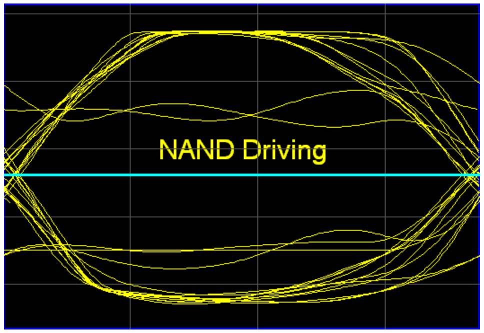
- For a reasonable capacity SSD, there are often 32 or more CE# pins
  - Each NAND package has 2 to 4 CE#s
- CE# is a slow signal, and all of these pins add cost to the SSD controller
- ONFI 3.0 includes a mechanism to share CE#s across die & packages
- Each target is assigned a “volume” at power-on
- Each volume is addressed by the host and stays active until the next volume is addressed



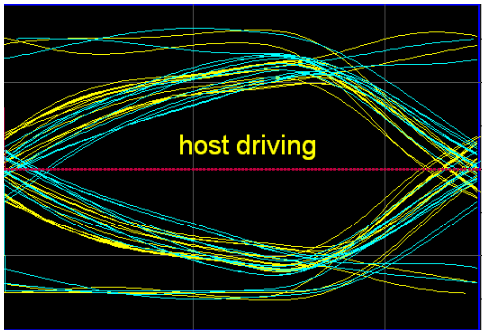
Benefit of On-Die Termination

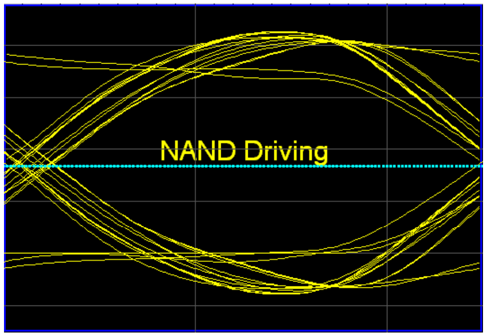
- Data eye at 266 MT/s without ODT
  - Passable, but getting close to the edge





- Data eye at 400 MT/s with On-Die Termination
  - ODT is the key enabler for 400 MT/s operation





On-die Termination Control

- Each LUN (die) may be the terminator for any volume
  - Terminator for its volume: Target termination
  - Terminator for another volume: Non-target termination

- At initialization, the LUN is configured with the volumes it will terminate for
  - Note: Many LUNs will not terminate for any volume

- If a command is issued to a volume the LUN is the terminator for, it snoops and terminates when data is transferred

Matrix of volumes that LUN may terminate for

Volume	7	6	5	4	3	2	1	0
Volume Byte 0	VOL 7	VOL 6	VOL 5	VOL 4	VOL 3	VOL 2	VOL 1	VOL 0
Volume Byte 1	VOL 15	VOL 14	VOL 13	VOL 12	VOL 11	VOL 10	VOL 9	VOL 8

ONFi 3.0 targeted for spec completion in Q3.

Agenda

- Enterprise NVMHCI
- ONFi 3.0
- Summary

Summary

- Enterprise NVMHCI enables broader adoption of PCI Express* (PCIe) Technology SSDs
  - Enables standard OS driver support, simplifies OEM qualification, and decreases time to market
  - Specification will be completed in November to enable PCIe SSD product intercept in 2012

- ONFI 3.0 delivers 400 MB/s per NAND channel, enabling next generation SSDs
  - Specification will be completed in 2H of the year

Enterprise NVMHCI and ONFI 3.0 standards enable next generation high performance SSDs.

Call to Action

- Get involved in the Enterprise NVMHCI specification development
  - Information on joining the Workgroup at http://www.intel.com/standards/nvmhci
- Get involved in the ONFi 3.0 specification development
  - Information on joining the ONFi Workgroup at http://onfi.org/membership/join
- Use these standards to deliver your next generation SSD solution

Additional sources of information on this topic:

- Other Sessions
  - MEMS002: Designing Solid-State Drives (SSDs) into Data Center Solutions
  - MEMS003: Understanding the Performance of Solid-State Drives (SSDs) in the Enterprise
  - MEMS004: Enterprise Data Integrity and Increasing the Endurance of Your Solid-State Drive (SSD)

- Learn more about Enterprise NVMHCI at:
  http://www.intel.com/standards/nvmhci

- Learn more about ONFI 3.0 at:
  http://www.onfi.org

Session Presentations - PDFs

The PDF for this Session presentation is available from our IDF Content Catalog at the end of the day at:

intel.com/go/idfsessionsBJ

URL is on top of Session Agenda Pages in Pocket Guide

Please Fill out the Session Evaluation Form

Give the completed form to the room monitors as you exit!

Thank You for your input, we use it to improve future Intel Developer Forum events

Q&A

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.
- 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 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.
- Copyright © 2010 Intel Corporation.

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. 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 corporation's expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions; 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 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. Additionally, Intel is in the process of transitioning to its next generation of products on 32nm process technology, and there could be execution issues associated with these changes, including product defects and errata along with lower than anticipated manufacturing yields. 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; defects or disruptions in the supply of materials or resources; and Intel's ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on changes in revenue levels; product mix and pricing; start-up costs, including costs associated with the new 32nm process technology; 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; the timing and execution of the manufacturing ramp and associated costs; and capacity utilization;. 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. The majority of our 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 our 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 impacted 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. 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 our ability to design our products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other risk factors that could affect Intel's results is included in Intel's SEC filings, including the report on Form 10-Q.

Rev. 1/14/10