PCI Express Storage in Client Systems

John Carroll
Storage Architect
Intel

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.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

All products, computer systems, 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.

Intel, Core i7, Core i5, Core i3, Ultrabook^{TM}, and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

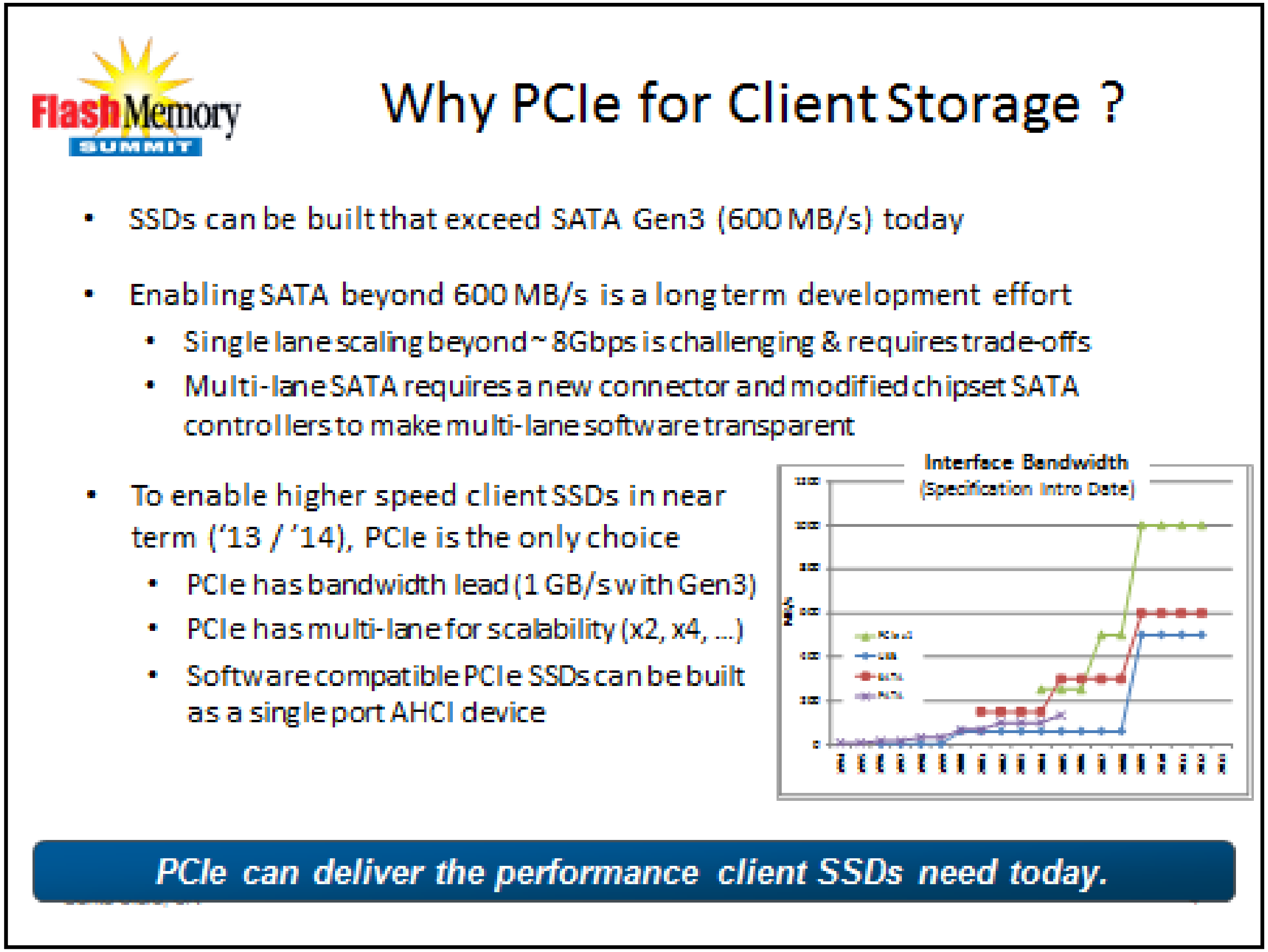
Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

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

Intel® Smart Response Technology requires a select Intel® processor, Intel® chipset, Intel® software and BIOS update. Depending on system configuration, your results may vary. Contact your system manufacturer for more information.

Intel® Smart Connect Technology requires a select Intel® processor, Intel® wireless product, Intel® software and BIOS update. Depending on system configuration, your results may vary. Contact your system manufacturer for more information.

2012: PCIe* Storage Coming to Client



Why PCIe for Client Storage ?

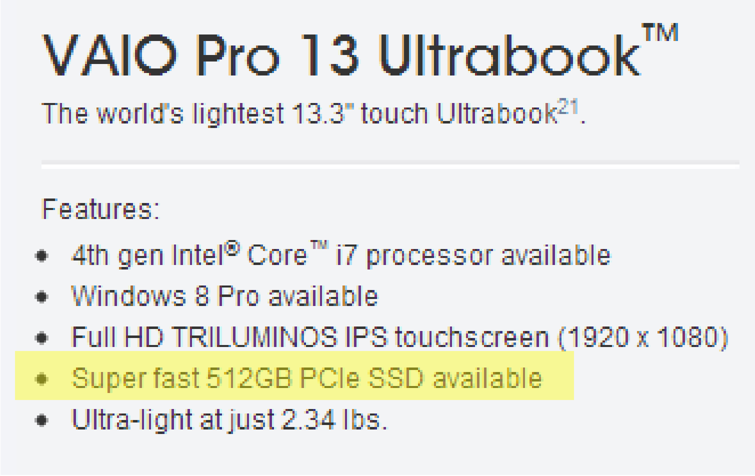
- SSDs can be built that exceed SATA Gen3 (600 MB/s) today
- Enabling SATA beyond 600 MB/s is a long term development effort
  - Single lane scaling beyond ~ 8Gbps is challenging & requires trade-offs
  - Multi-lane SATA requires a new connector and modified chipset SATA controllers to make multi-lane software transparent
- To enable higher speed client SSDs in near term ('13 / '14), PCIe is the only choice
  - PCIe has bandwidth lead (1 GB/s with Gen3)
  - PCIe has multi-lane for scalability (x2, x4, …)
  - Software compatible PCIe SSDs can be built as a single port AHCI device

PCIe can deliver the performance client SSDs need today.

Source: FMS '12 Amber Huffman

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

2013: Client PCIe* SSDs Have Arrived

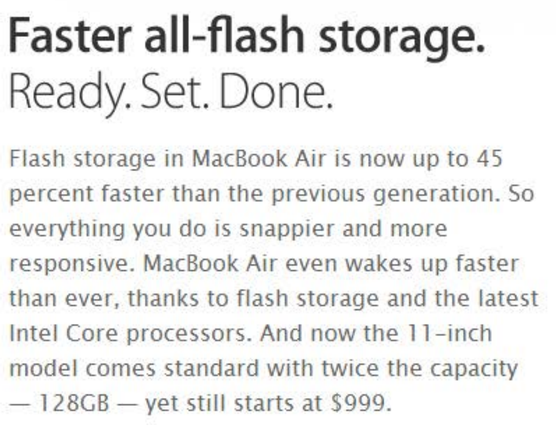


VAIO Pro 13 Ultrabook™

The world's lightest 13.3" touch Ultrabook[21].

Features:

- 4th gen Intel® Core™ i7 processor available
- Windows 8 Pro available
- Full HD TRILUMINOS IPS touchscreen (1920 x 1080)
- Super fast 512GB PCIe SSD available
- Ultra-light at just 2.34 lbs.



Faster all-flash storage.

Ready. Set. Done.

Flash storage in MacBook Air is now up to 45 percent faster than the previous generation. So everything you do is snappier and more responsive. MacBook Air even wakes up faster than ever, thanks to flash storage and the latest Intel Core processors. And now the 11-inch model comes standard with twice the capacity — 128GB — yet still starts at $999.

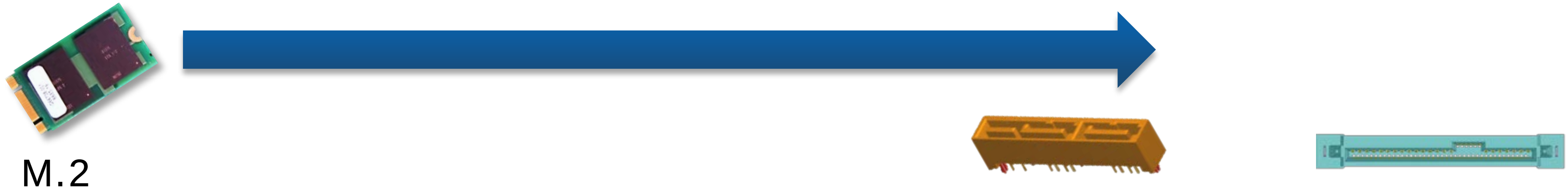
* Other names and brands may be claimed as the property of others.

Agenda

- Client PCIe Form Factors
- SATA to PCIe Transition
  - Power Optimizations for PCIe SSDs
- Controller Interfaces

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

Form Factor & Connector Landscape



- M.2* was designed for the unique needs of Ultrabook™
  - However, M.2 is being used in a wide variety of devices
  - Cannot support HDDs or SSHDs

- 2.5” SATA Express* and SFF-8639 connectors provide flexibility to support HDDs, SSHDs, & SSDs on the same connector

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

EMI Challenges for PCIe* Cabling

- A reference clock in the cable for 2.5” PCIe drives causes EMI issues for client PCs

- PCI-SIG solved this challenge with a new clocking mechanism (SRIS)

Subpart A—General

§ 15.1 Scope of this part.

(a) This part sets out the regulations under which an intentional, unintentional, or incidental radiator may be operated without an individual license. It also contains the technical specifications, administrative requirements and other conditions relating to the marketing of part 15 devices.

(b) The operation of an intentional or unintentional radiator that is not in accordance with the regulations in this part must be licensed pursuant to the provisions of section 301 of the Communications Act of 1934, as amended, unless otherwise exempted from the licensing requirements elsewhere in this chapter.

(c) Unless specifically exempted, the operation or marketing of an intentional or unintentional radiator that is not in compliance with the administrative and technical provisions in this part, including prior Commission authorization or verification, as appropriate, is prohibited under section 302 of the Communications Act of 1934, as amended, and subpart I of part 2 of this chapter. The equipment authorization and verification procedures are detailed in subpart J of part 2 of this chapter.

SRIS clocking enables cabling for PCIe* SSDs in client systems

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

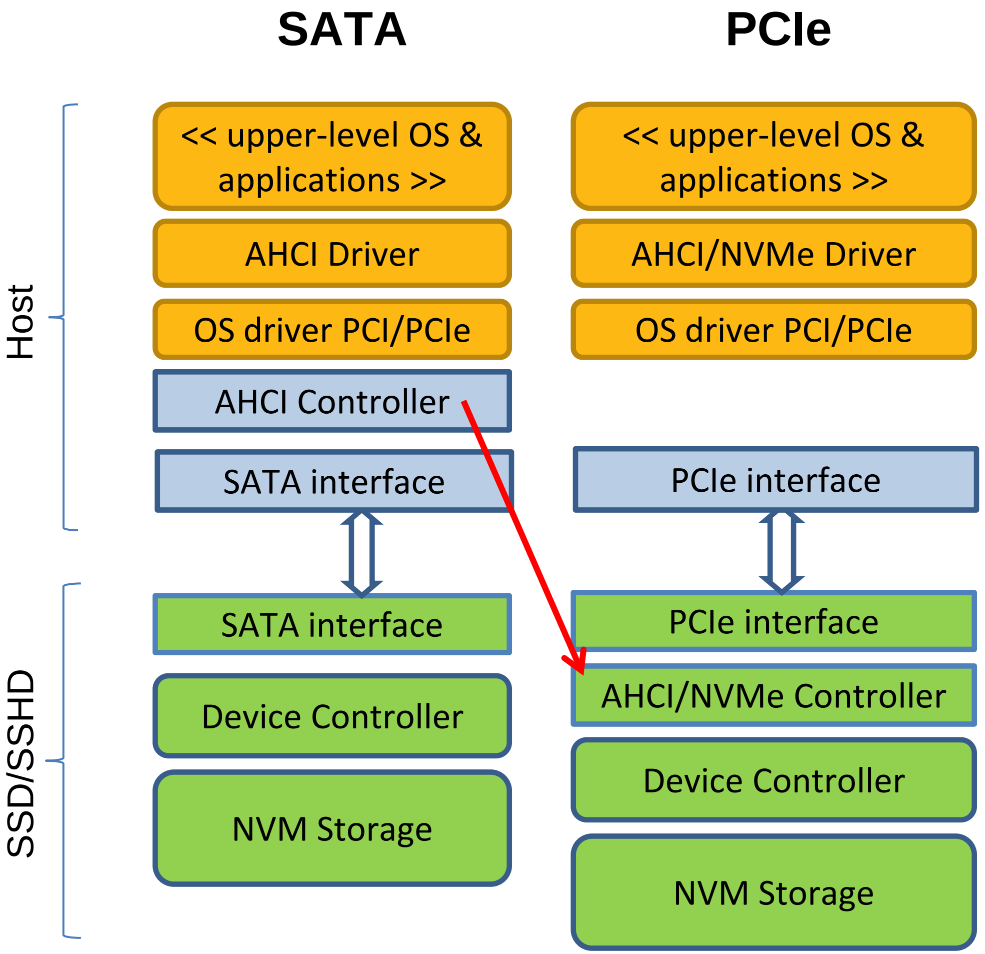
Agenda

- Client PCIe Form Factors
- SATA to PCIe Transition
    - Power Optimizations for PCIe SSDs
- Controller Interfaces

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

Transitioning from SATA* to PCIe*

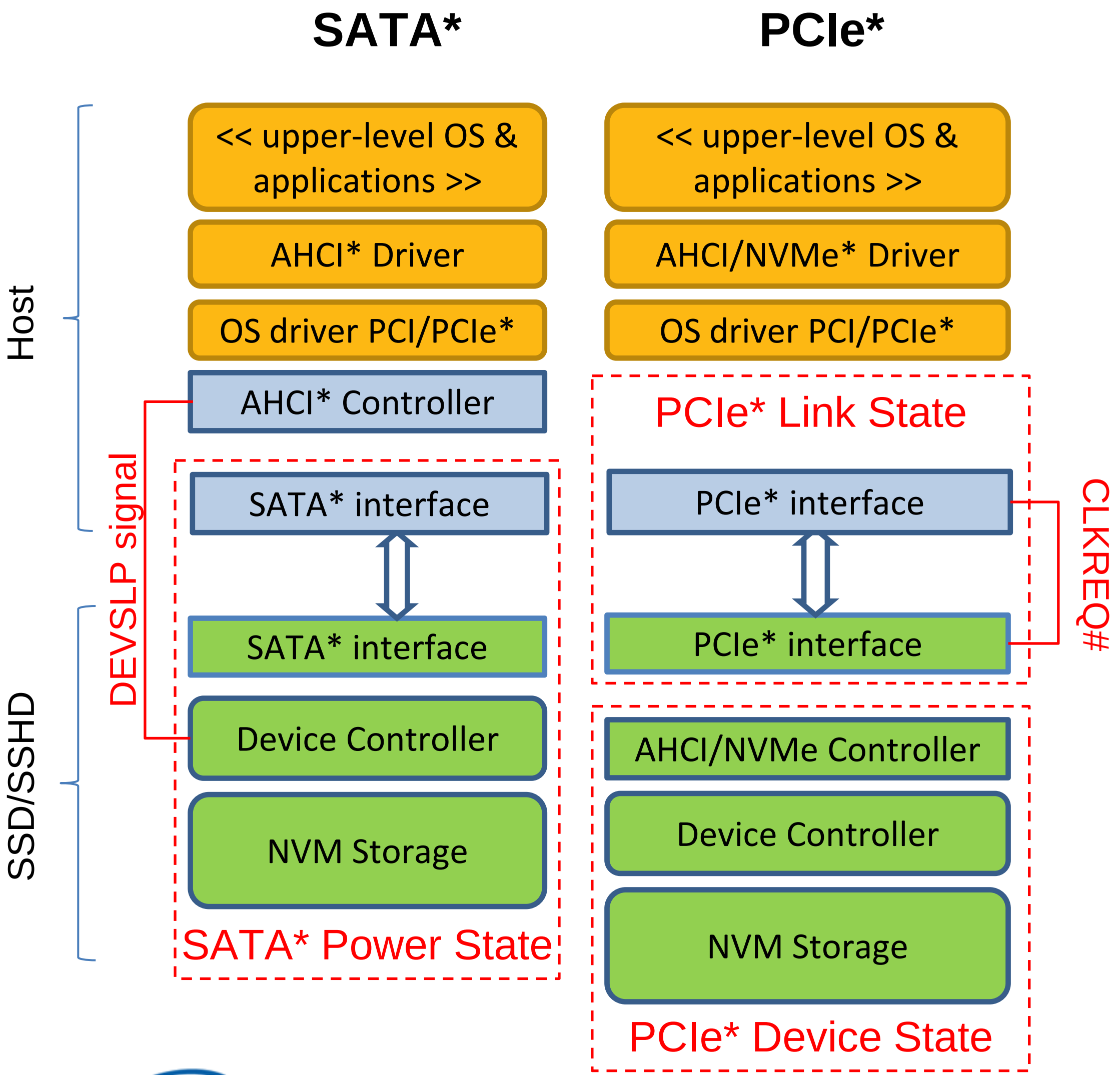
- PCIe* storage devices incorporate controller functionality into SSD/SSHD



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

Independent Power States

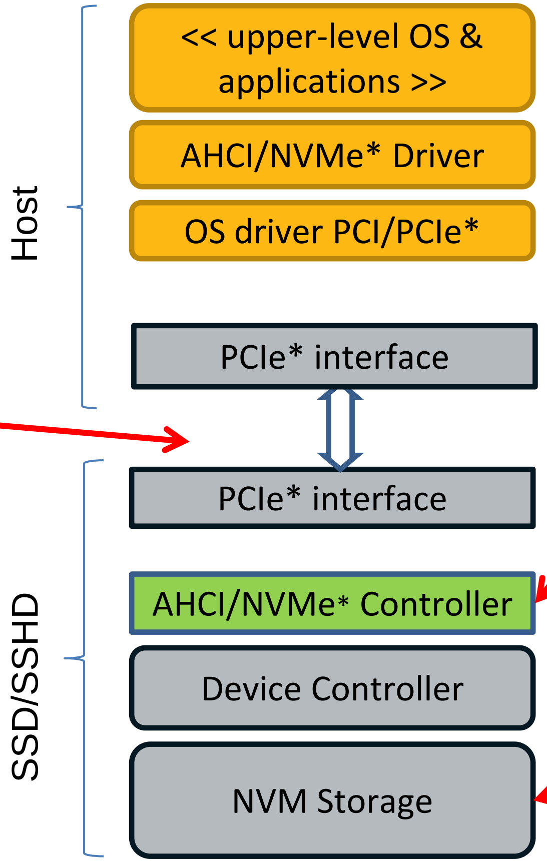
- PCIe* separates link and device state into two independent states
- Link states defined by PCIe* spec
- Device states defined by controller interface



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

PCIe* DevSleep Equivalent

The lowest non-zero power state for a PCIe* SSD



Host

<< upper-level OS & applications >>

AHCI/NVMe* Driver

OS driver PCI/PCIe*

PCIe* interface

SSD/SSHD

PCIe* interface

AHCI/NVMe* Controller

Device Controller

NVM Storage

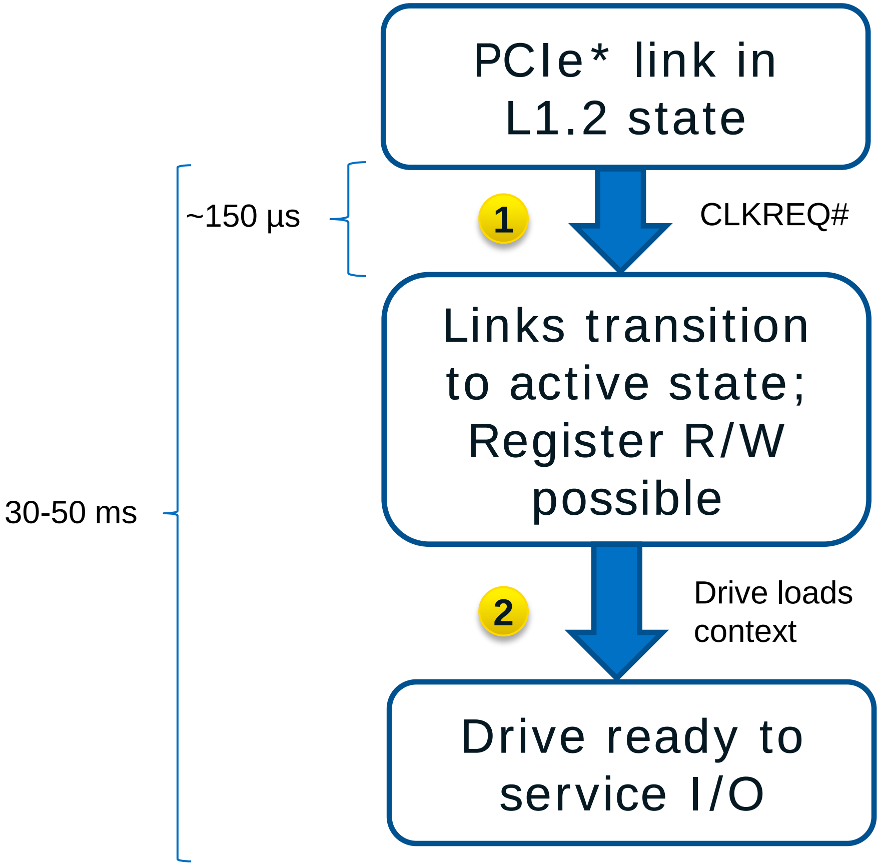
Link in low power L1.2 state

Controller registers active for faster recovery

Rest of drive powered down

Resuming from DevSleep Equivalent

- For SATA*, AHCI controller in the host allows slower SSD recovery from DevSleep
- With PCIe*, register reads can stall the CPU consuming watts while waiting for controller to respond
- 2 step resume process for responsiveness and save power



PCIe* link in L1.2 state

~150 µs

1 CLKREQ#

Links transition to active state; Register R/W possible

30-50 ms

2 Drive loads context

Drive ready to service I/O

Link transitions first; drive catches up later

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

Agenda

- Client PCIe Form Factors
- SATA to PCIe Transition
  - Power Optimizations for PCIe SSDs
- Controller Interfaces

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

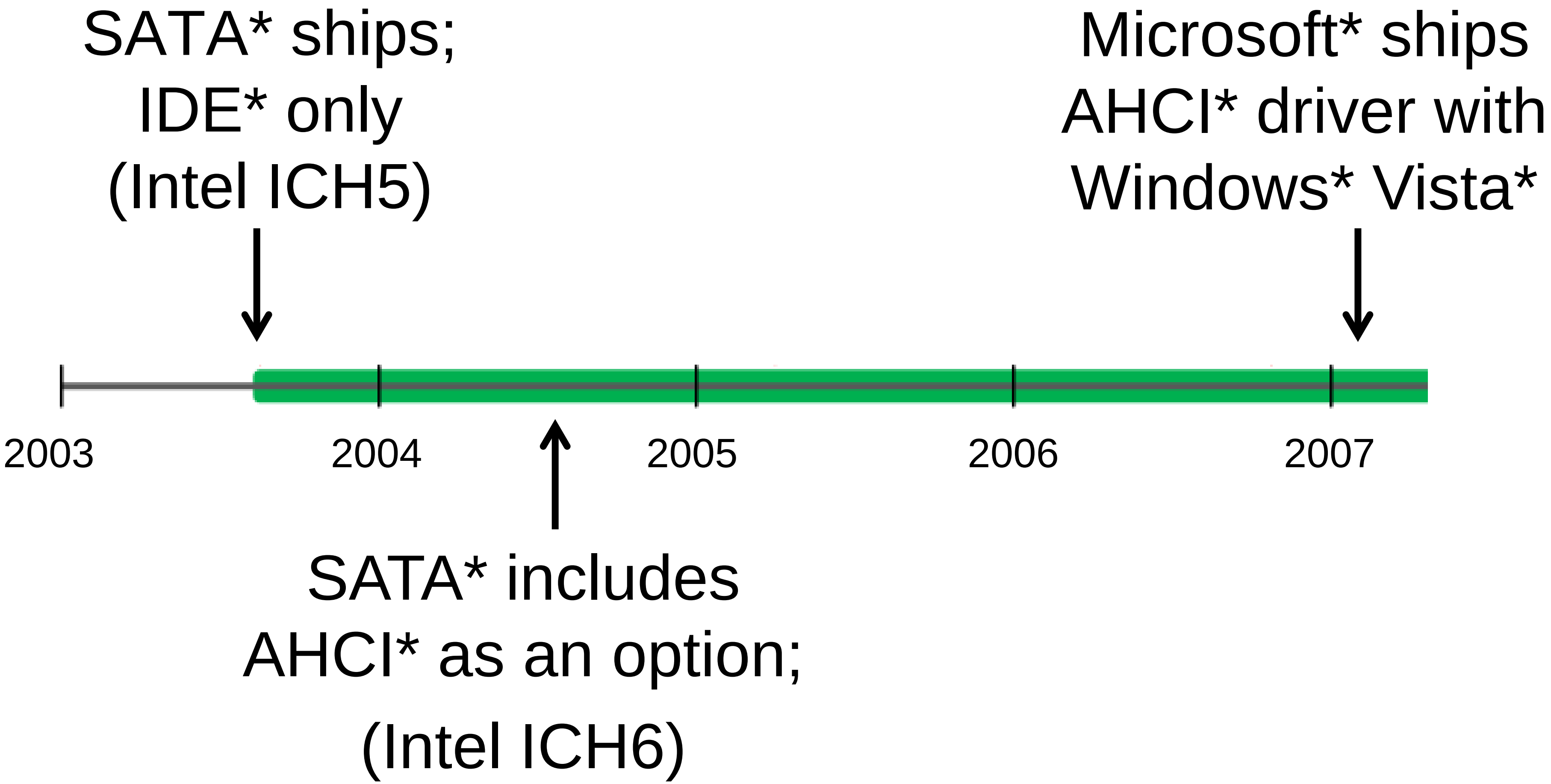
AHCI Lacks Scalability for the Future

A huge leap ahead of IDE, but still designed for hard drives…

	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* is architected from the ground up for NAND today and next gen NVM of tomorrow.

Another Controller Interface Transition



SATA* ships; IDE* only (Intel ICH5)

Microsoft* ships AHCI* driver with Windows* Vista*

SATA* includes AHCI* as an option; (Intel ICH6)

Summary

- Client PCIe* storage shipping now – primarily M.2
- Transitioning to PCIe* brings new tools to reduce storage power
- PCIe* AHCI* devices leverage existing AHCI* SW
- NVMe* the long term PCIe* solution; ecosystem establishing quickly

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