PCI Express SSD Technology In Practice

Hojun Shim
Samsung Electronics
Controller Development Team
hojun.shim@samsung.com

Abstract:

본 세미나에서는 PCIe SSD가 탄생하게 된 배경을 살펴보고, PCIe SSD 관련 표준화 및 시장 트렌드를 소개한다. 그리고 향후 Computer System Architecture에서 PCIe SSD 응용처에 대해서 논의한다.

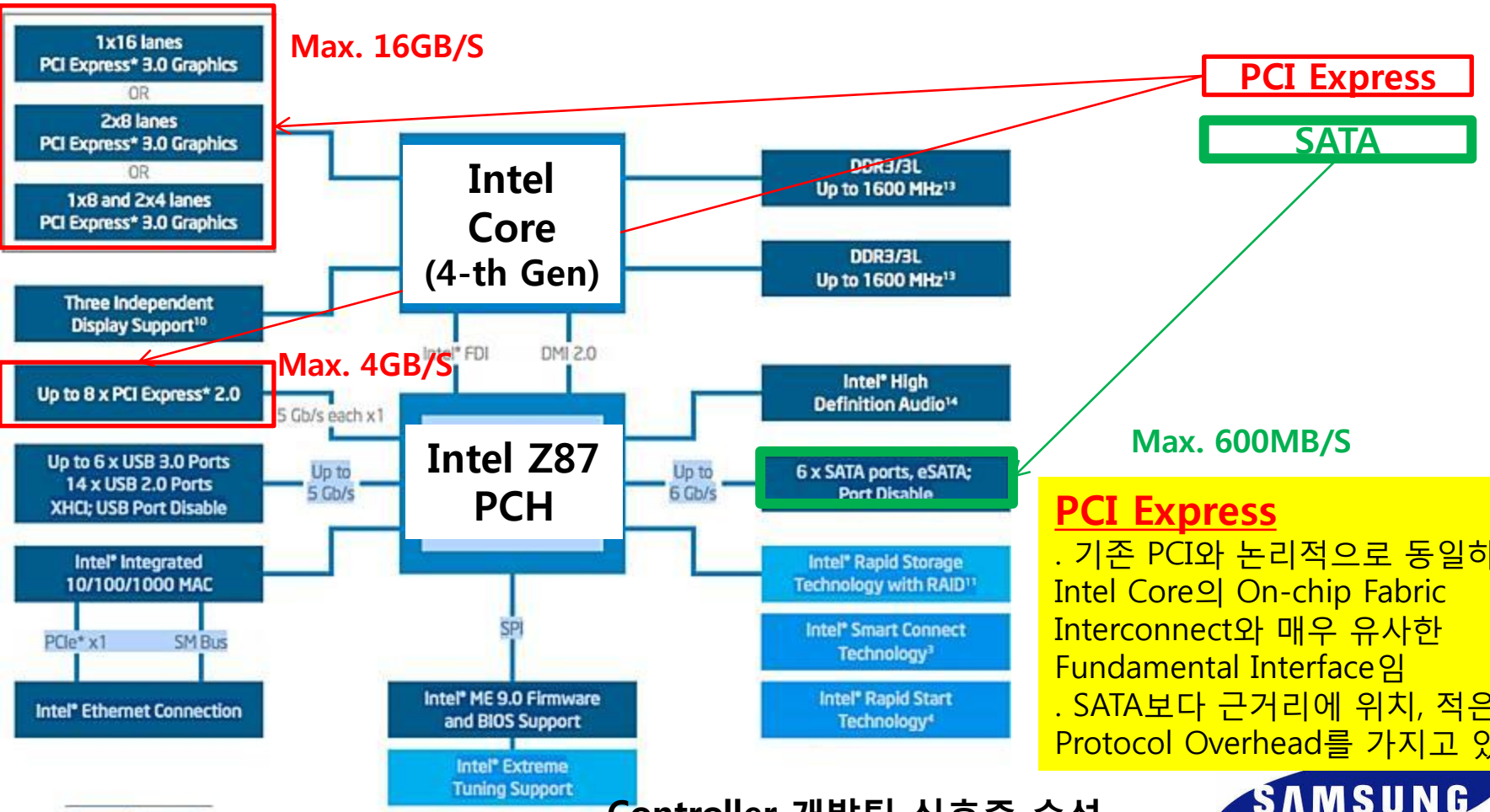
SAMSUNG

Contents

- Why PCI Express?
- Memory Hierarchy
  - SAS/SATA SSD
  - Caching Hot Data into PCIe SSD
  - PCIe SSD
- Product Examples
  - Non-native RAID (OCZ)
  - PCIe AHCI (Marvell)
  - Native PCIe NVMe (Micron)
  - Drive Backplane (HP)
- New Physical FF Specifications
  - SFF-8639 for Enterprise
  - SATA Express & M.2 (NGFF) for Clients
- New Logical Protocol Specifications
  - AHCI
  - NVMe (NVM Express)
  - SCSIe (SCSI Express)
- Samsung PCIe SSD
- Future Works
  - Memory Hierarchy Vertical Optimization: Virtual Memory Expansion in MemCached Servers
  - SSD Interface Optimization: PCIe → DRAM Interface

Why PCI Express?

- Intel Haswell (Lynx Point) Platform



PCI Express

. 기존 PCI와 논리적으로 동일하며, Intel Core의 On-chip Fabric Interconnect와 매우 유사한 Fundamental Interface임

. SATA보다 근거리에 위치, 적은 Protocol Overhead를 가지고 있음

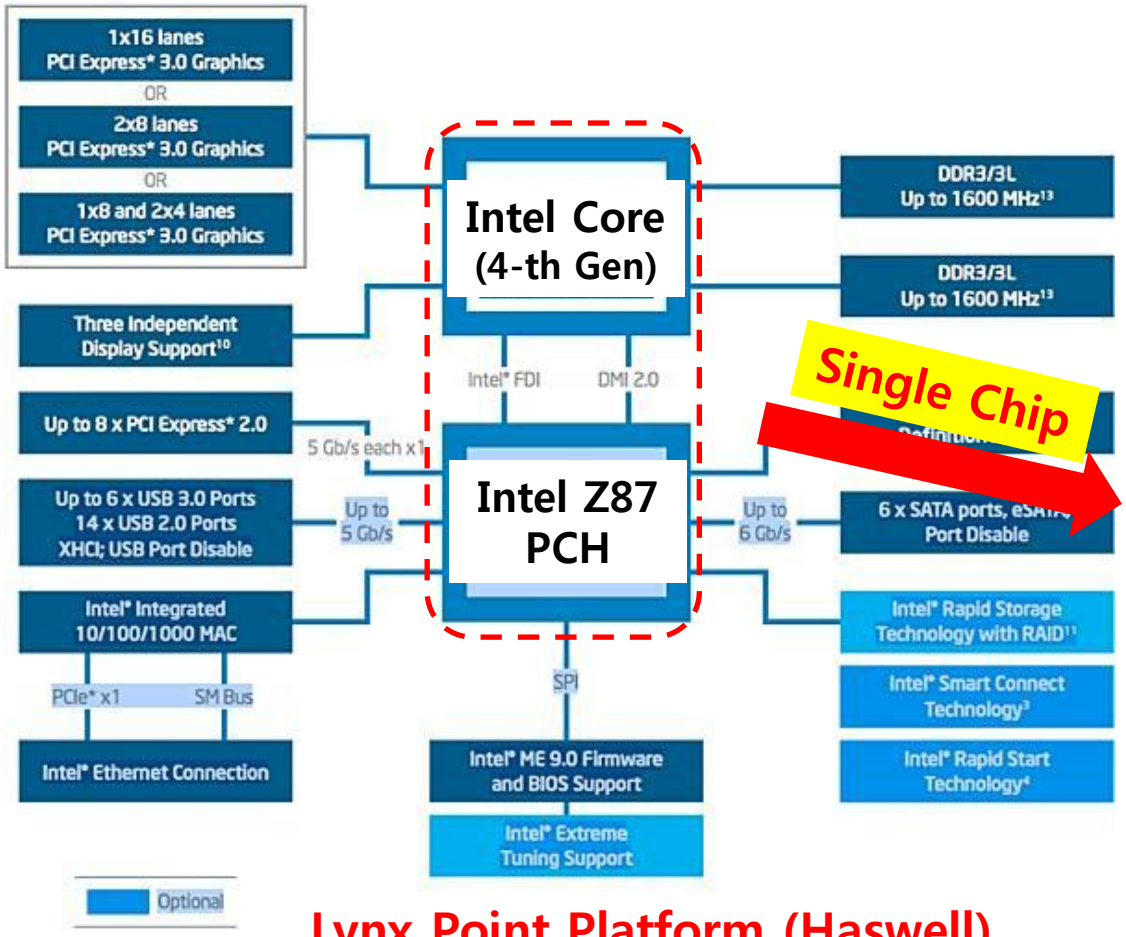
Controller 개발팀 심호준 수석

SAMSUNG

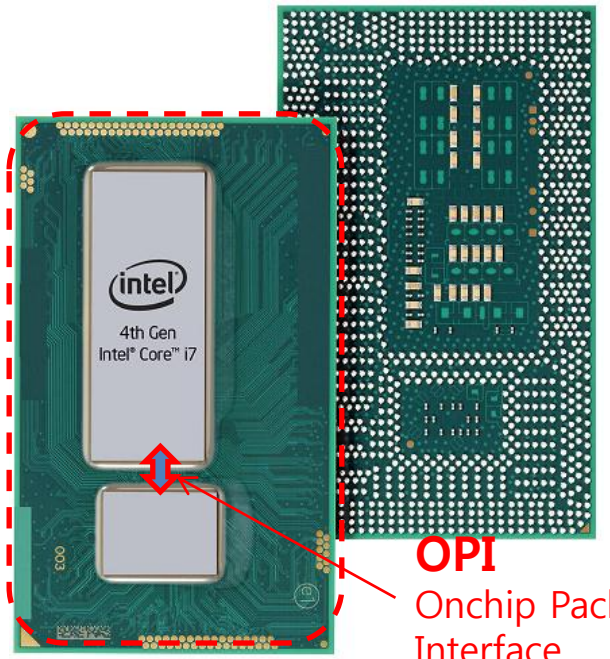
Why PCI Express?

- Haswell vs. Haswell-ULT

Haswell-ULT
. 전체적인 플랫폼 크기를 줄이기 위해서 2개의 물리적으로 다른 칩을 OPI (Onchip Package Interface)로 하나의 칩으로 묶고 기능을 축소함



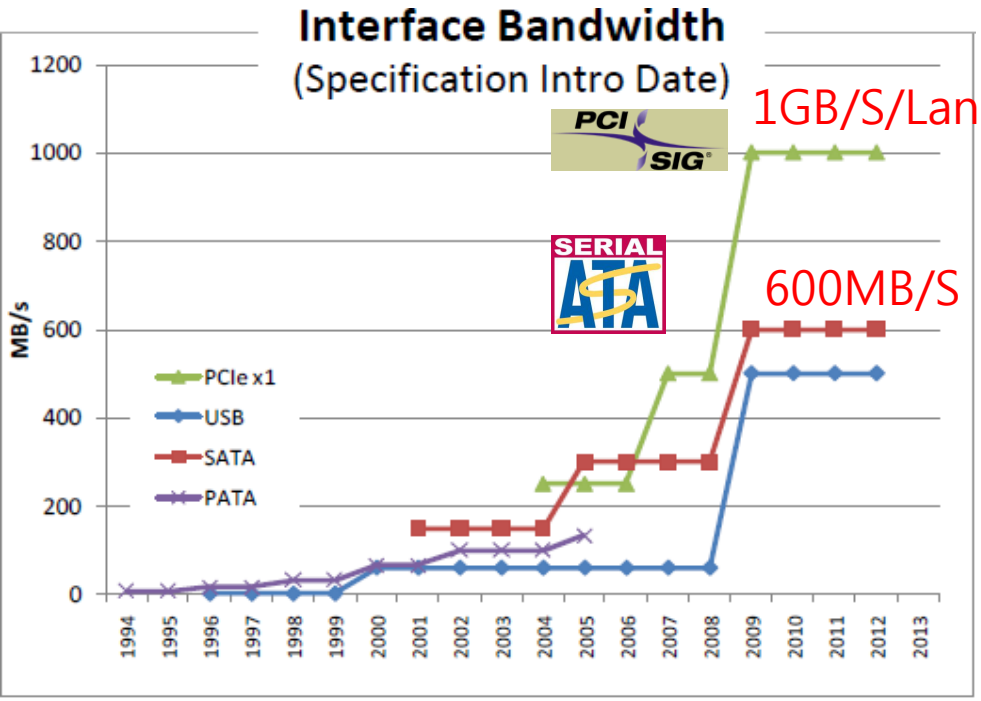
Lynx Point Platform (Haswell)



Lynx Point-LP

Why PCI Express?

- PCIe is the Only Interface which Enable SSD to achieve Higher Performance SSD than current SATA/SAS Interface
  - Higher Bandwidth (1GB/S/Lane @Gen3)
  - Scalable Bandwidth with Multi Lanes
  - Software Compatibility (@Single Port AHCI Device)



SATA 6G
PCIe Gen2 x4
PCIe Gen3 x4

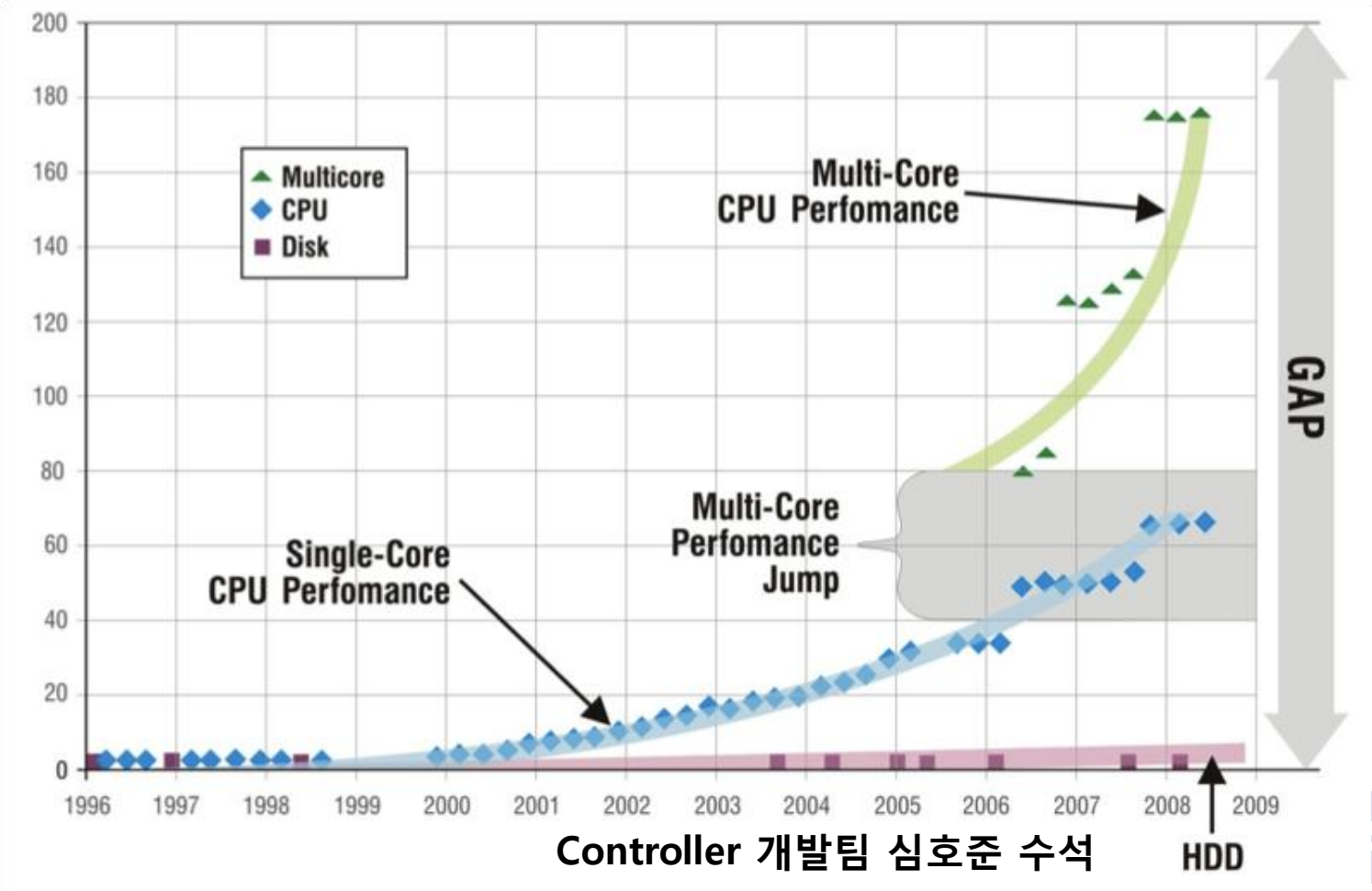
600MB/S

PCIe Bandwidth Comparison (Each Direction)			
	PCIe 1.x	PCIe 2.x	PCIe 3.0
x1	250MB/sec	500MB/sec	1GB/sec
x2	500MB/sec	1GB/sec	2GB/sec
x4	1GB/sec	2GB/S	4GB/S
x8	2GB/sec	4GB/sec	8GB/sec
x16	4GB/sec	8GB/sec	16GB/sec

SAMSUNG

Memory Hierarchy

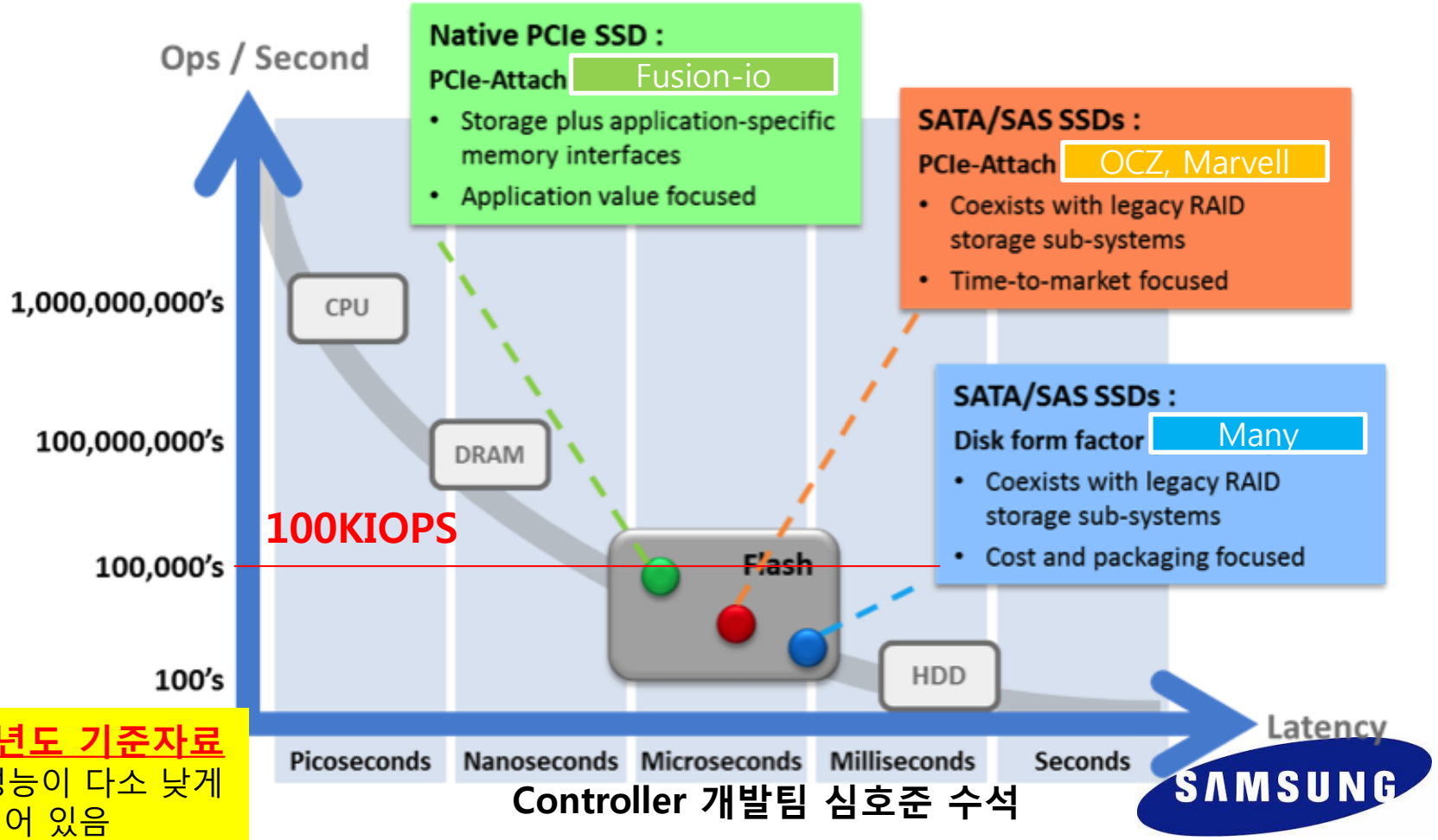
- Each Component's Relative Performance



Controller 개발팀 심호준 수석

Memory Hierarchy

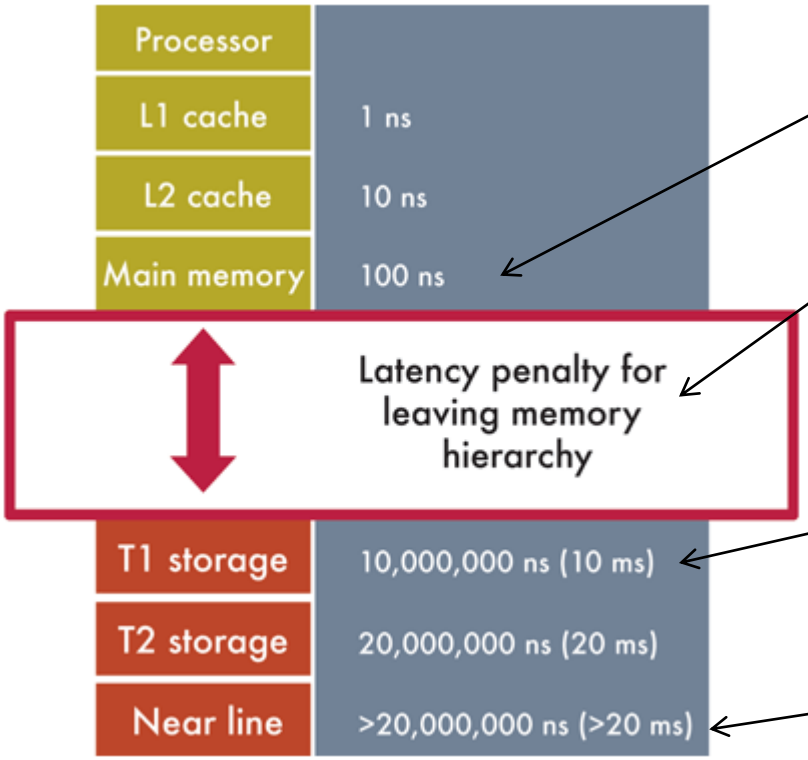
- Each Component's Absolute Performance



2009년도 기준자료
SSD 성능이 다소 낮게 표기되어 있음

Memory Hierarchy

- HDD Storage와 Memory 사이의 커다란 Latency 차이 존재
- NAND Flash Memory로 이 커다란 Latency 차이로 인한 Application Performance 및 User Experience 저하를 어떻게 극복할 것인가?



DRAM

Caching Hot Data to System Main Memory (DRAM, **<10us**)
Caching Hot Data to **PCIe SSD** with NAND Flash
PCIe SSD with NAND Flash (**<100us**)
SATA/SAS SSD with **NAND Flash** (**100us**)

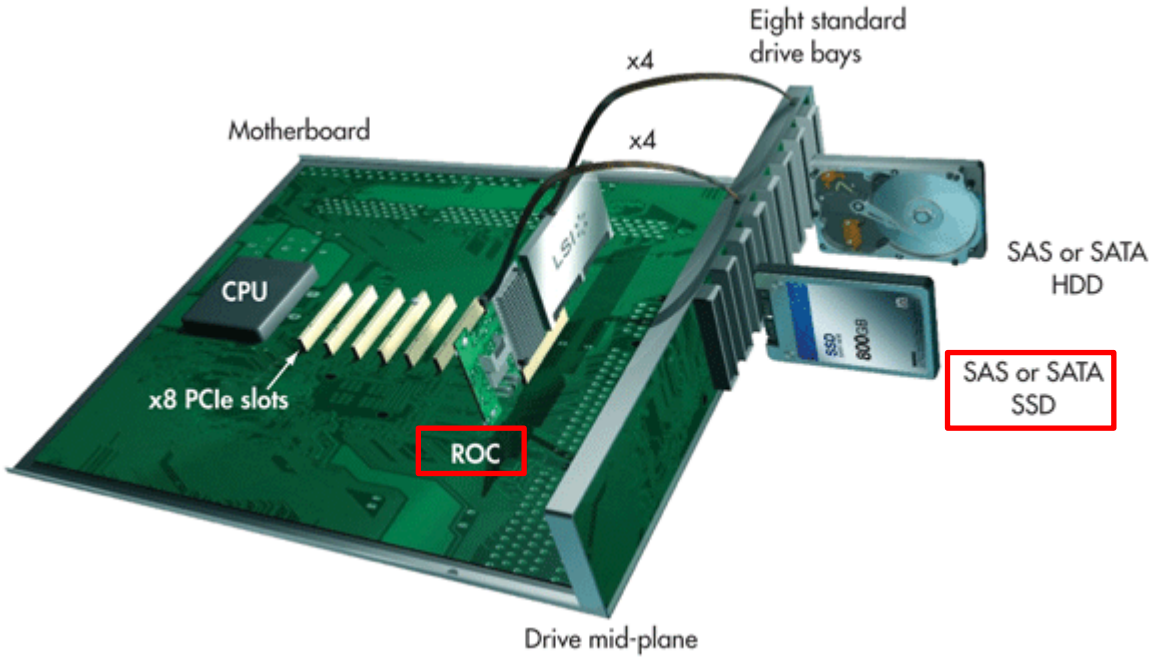
SATA/SAS HDD with Rotating Media

NAS (Network Attached Storage, Ethernet)
or SAN (Storage Area Network, Fiber Channel)

http://electronicdesign.com/memory/evolution-solid-state-storage-enterprise-servers

Memory Hierarchy

- SATA/SAS SSD with NAND Flash



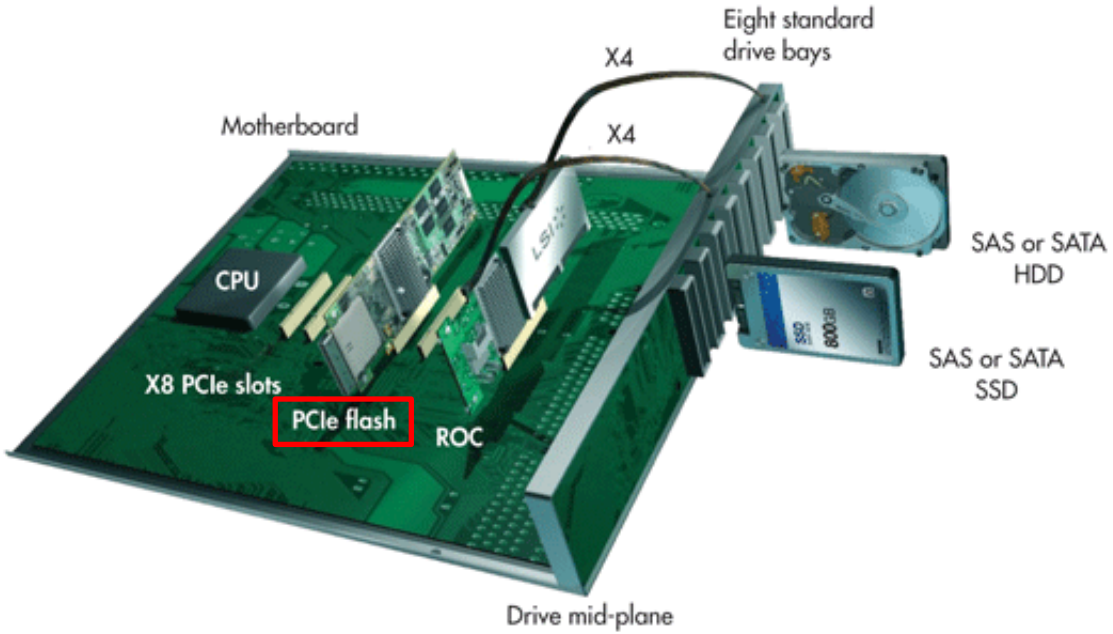
SAS and SATA SSDs are supported today in standard storage bays with a RAID-on-chip (ROC) controller on the server's PCIe bus.

Controller 개발팀 심호준 수석

SAMSUNG

Memory Hierarchy

- Caching Hot Data to PCIe SSD with NAND Flash

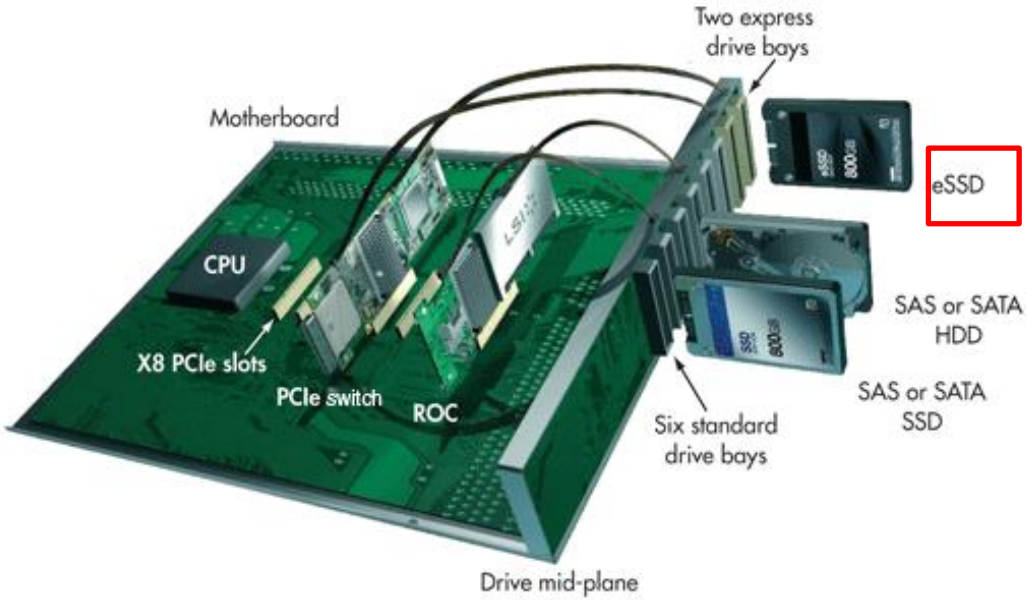


PCIe flash adapters overcome the limitations imposed by legacy storage protocols, but they must be plugged directly into the server's PCIe bus.

SAMSUNG

Memory Hierarchy

- PCIe SSD with NAND Flash



Express Bay fully supports the low latency of flash memory with the high performance of PCIe, while maintaining backwards compatibility with existing SAS and SATA HDDs and SSDs.

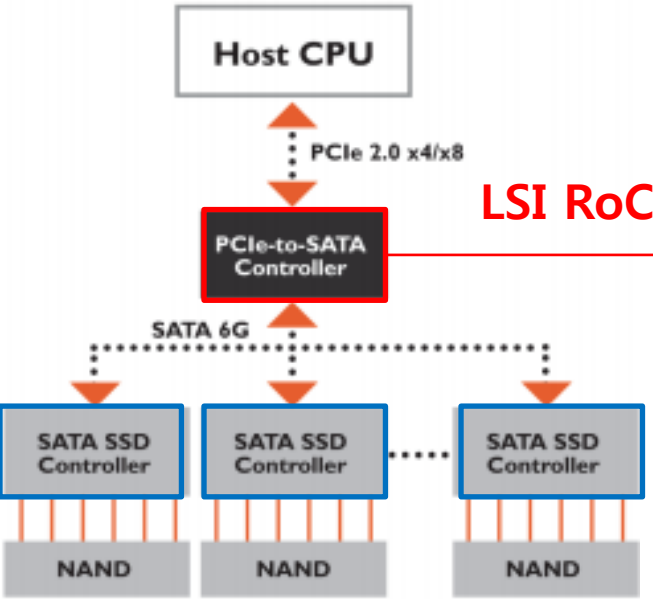
Controller 개발팀 심호준 수석

SAMSUNG

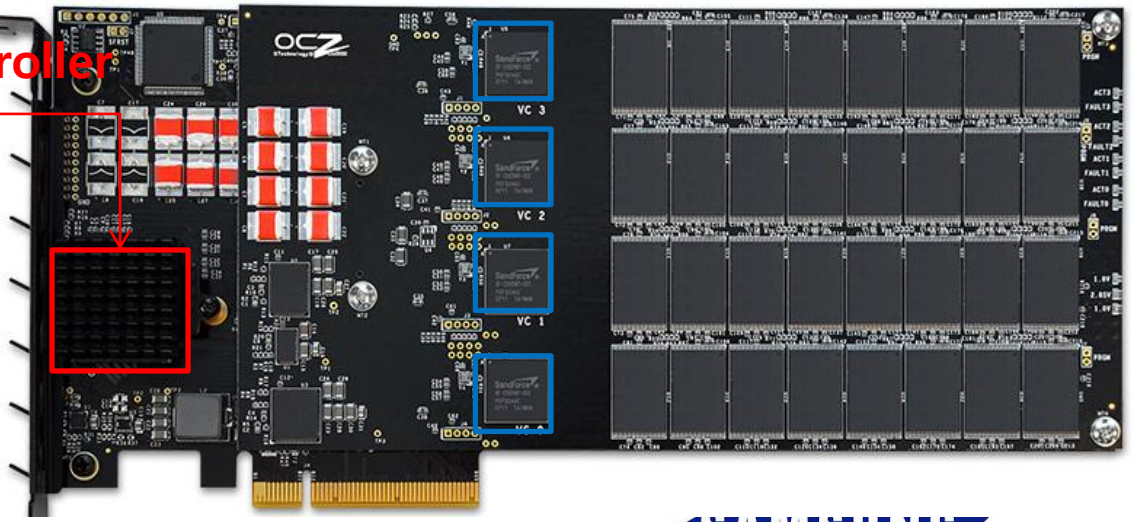
Product Examples

- PCIe CEM Add-in Card SSD (Non-native PCIe SATA or SAS RAID)
  - OCZ Z-Drive R4 (PCIe Gen2 x8)
  - LSI RoC (RAID On Chip) Controller + 8 x Sandforce 2218 Controllers (SATA Gen3)

Sandforce is acquired by LSI. (2011, October)



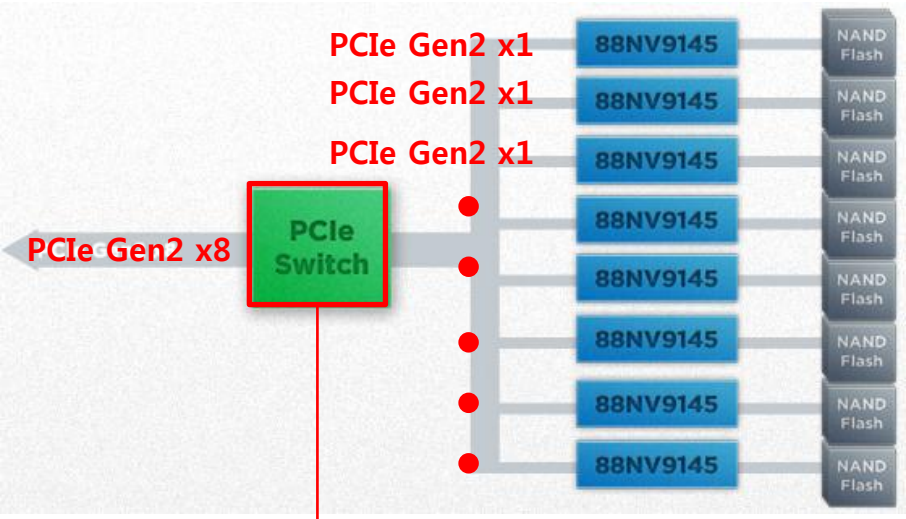
LSI RoC Controller

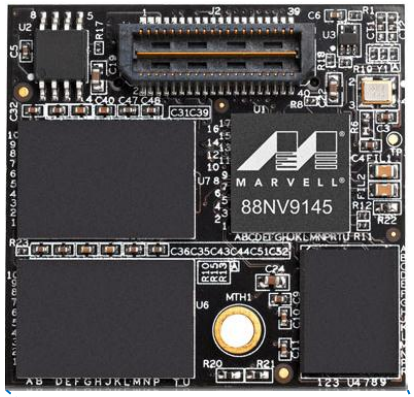


Controller 개발팀 심호준 수석

Product Examples

- PCIe CEM Add-in Card SSD (PCIe AHCI)
  - Marvell 88NV9145 (PCIe Gen2 x1)
    - 4-ch 4-way NAND flash system
  - Scalable PCIe SSD through PCIe Switch for High-end Servers (PCIe Gen2 x8)





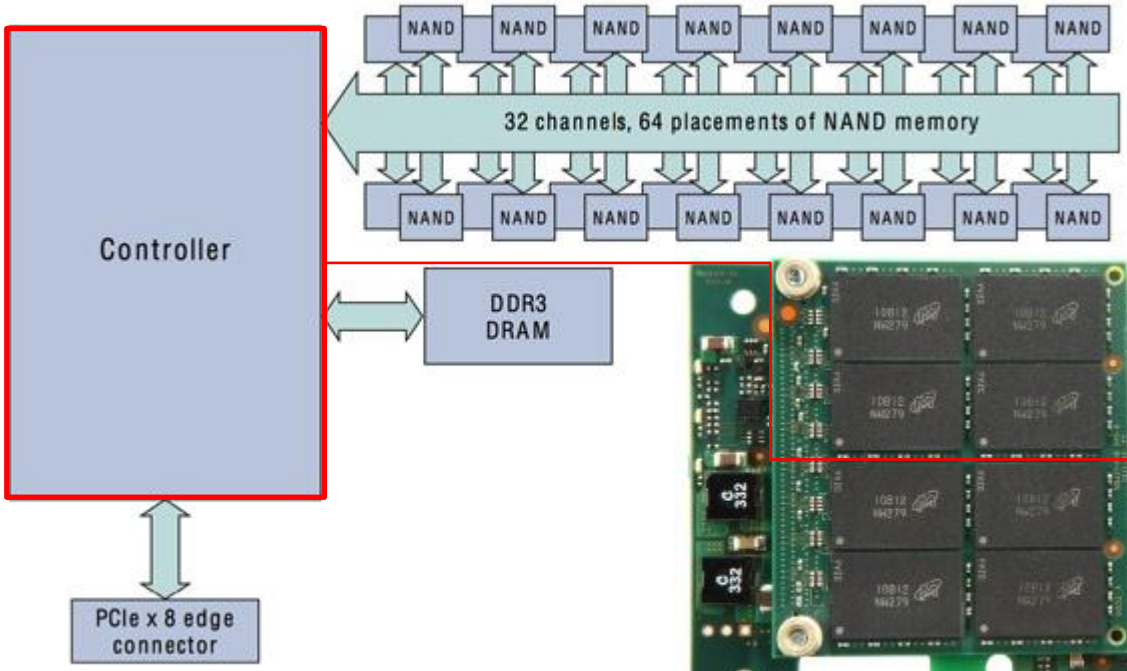
Software RAID is Necessary! ➔ BAD!

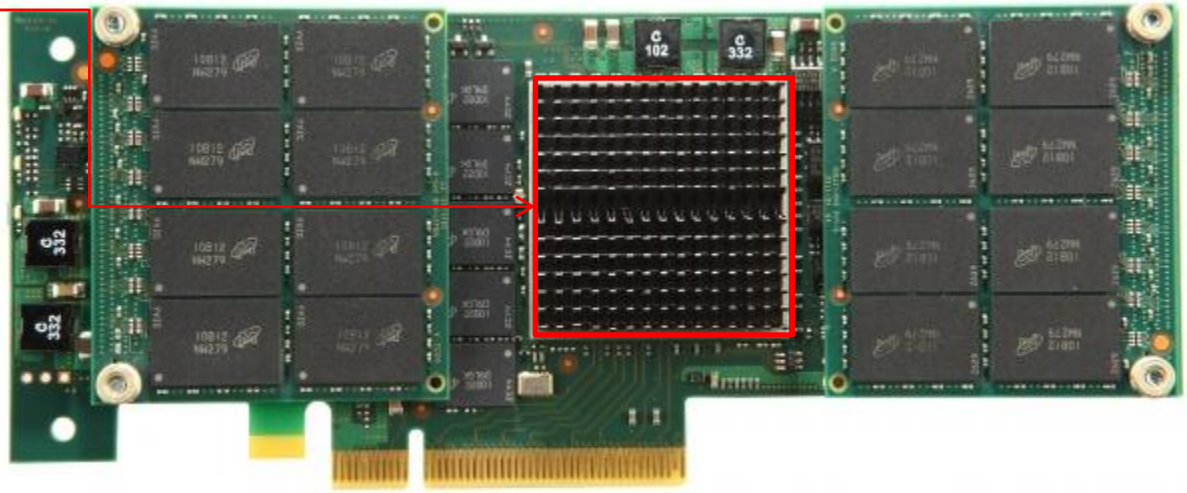
PLX PCIe Switch

Product Examples

- PCIe CEM Add-in Card SSD (Native PCIe NVMe)
  - Micron P320h
  - PCIe SSD for High-end Servers (PCIe Gen2 x8)
  - IDT Controllers (NVM Express)
    - 1517 pins, 32 NAND flash channels

IDT is acquired by PMC Sierra. (2013, May)





Controller 개발팀 심호준 수석

Product Examples

- 2.5-inch Express SSD Drives (Native PCIe NVMe)
  - Micron P320h (with Dell)
  - Hot swappable 2.5-inch FF (PCIe Gen2 x4)
    - Improved Serviceability and Scalability
    - Lowered TCO (Total Cost of Ownership)

Three 2.5-inch Express Drive Bays

Dell's 12-th Generation Server

Product Examples

- Drive Back-Plane

http://electronicdesign.com/memory/evolution-solid-state-storage-enterprise-servers

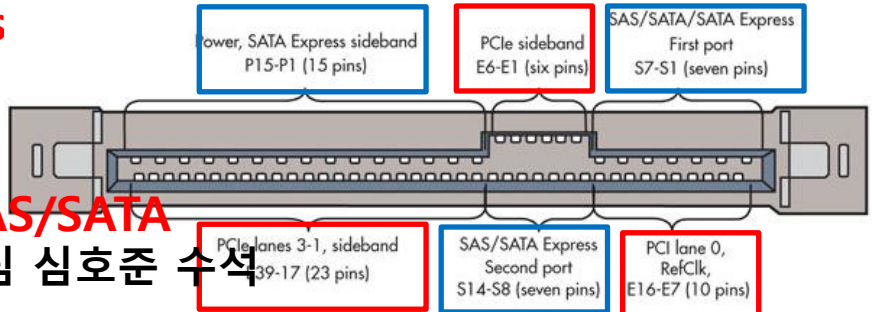
For Six 2.5-inch SATA/SAS Drive Bays SFF-8680 Spec.

For Two 2.5-inch Express Drive Bays

SFF-8639 Spec.

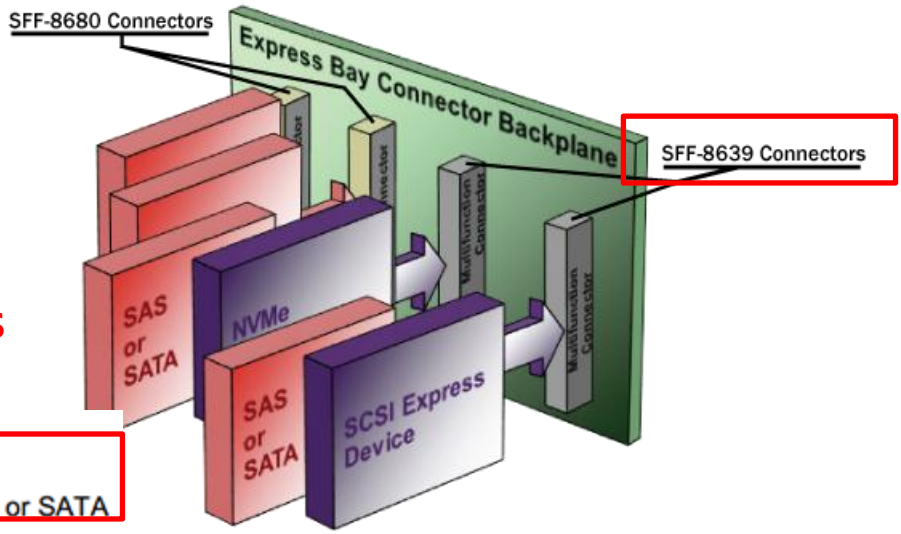
4-lane PCIe, 2-port SAS/SATA

Controller 개발팀 심호준 수석

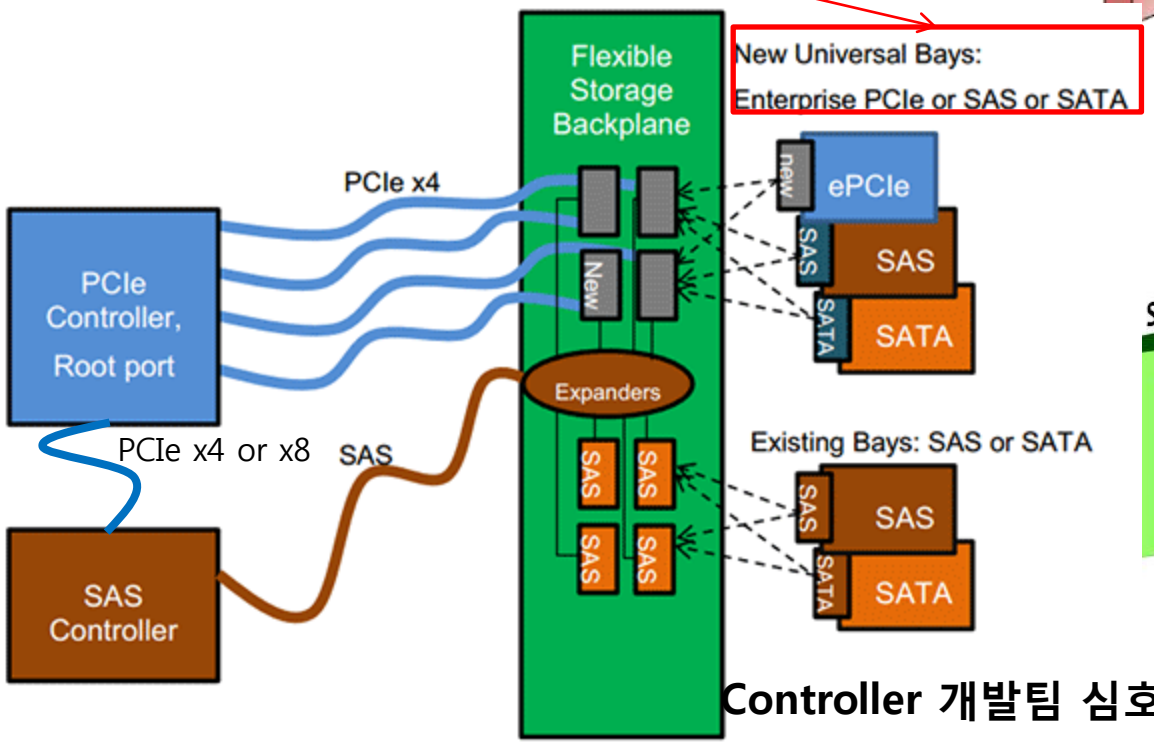


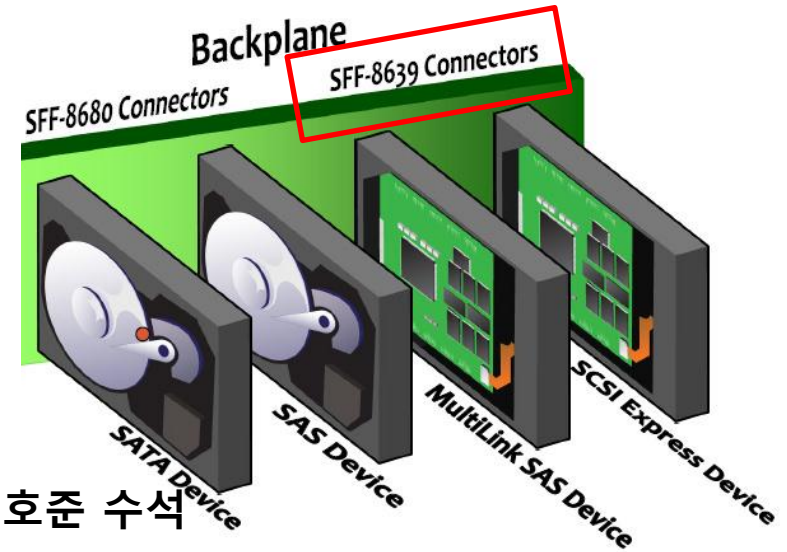
New Physical FF Specifications

- For Servers **SFF Committee**
  - SFF-8680 → SFF-8639
  - Flexible & Universal Back-Plane Implementation



For Two 2.5-inch Express Drive Bays

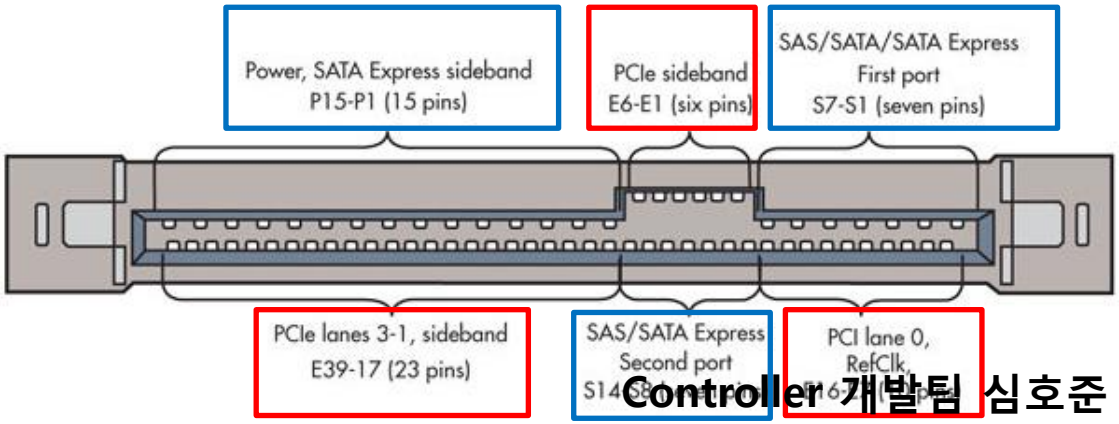




Controller 개발팀 심호준 수석

New Physical FF Specifications

- SFF-8639 SFF Committee
  - Enterprise Backplane Connector for 2.5-inch SSD covering PCIe, SATA, and SAS.
  - 4 Lanes (red below) are Connected to CPU or Chipset for PCIe Support.
  - 2 lanes (blue below) are Connected to an HBA/RAID Controller or Chipset for SAS & SATA Support.
  - Allows Client PCIe SSDs to be Used in Enterprise Backplanes.
- Why 2.5-inch FF is Important?
  - 2.5-inch is Critical Enterprise FF to Support Hot Swap Backplanes.



SFF-8639 Spec.

Enterprise PCIe

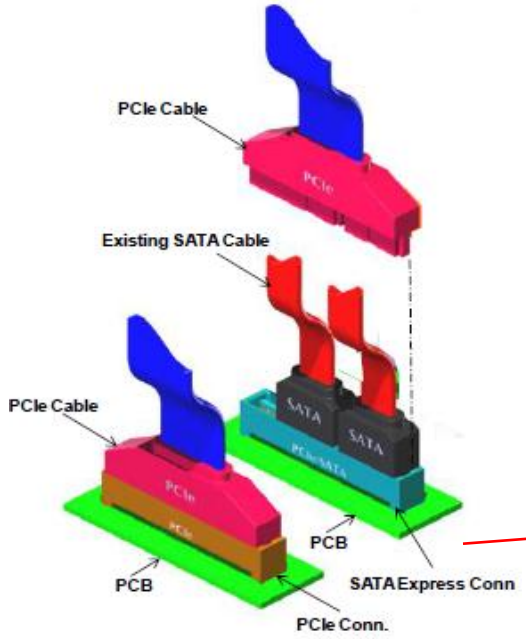
SATA/SAS

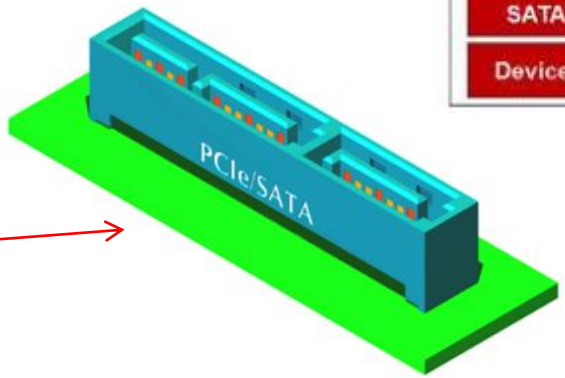
SAMSUNG

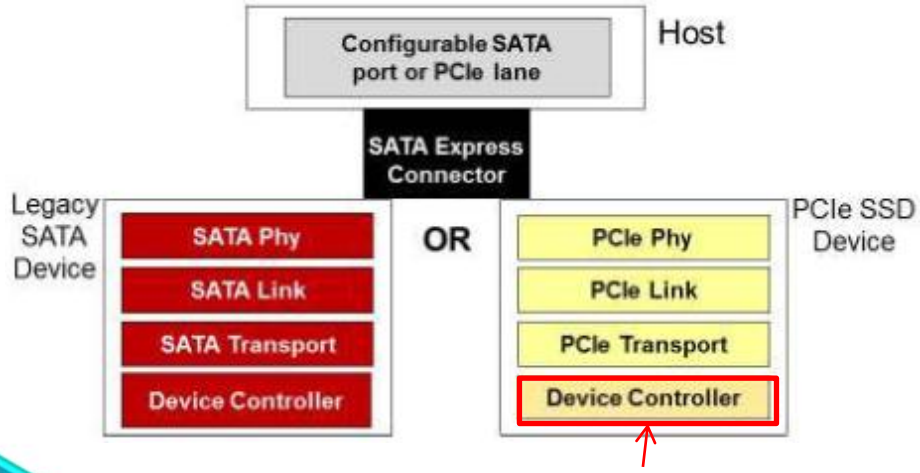
New Physical FF Specifications

- For Clients (**SATA Express**)
  - SATA → SATA Express
  - Flexible Motherboard Implementation

SATA Express
. SATA-IO의 Marketing Name임
. **FF, Connector만 정의**
. Software I/F는 AHCI/NVMe/SCSIe 상관없이 쓸 수 있음







AHCI / NVMe / SCSIe

PCIe 2 lanes or 2-port SATA

The new SATA Express connector contains both Serial ATA and PCI-Express.

New Physical FF Specifications

- For Clients (**M.2**)
  - mSATA → M.2 (NGFF)
  - Faster & Smaller & Thinner SSD Implementation

M.2

1 port SATA or PCIe 2 or 4 lanes

mSATA

1 port SATA

M.2

PCIe 4 lanes

NGFF modules can come in different lengths, including 42, 60 and 80 mm. The **longer** modules can **fit more flash chips**. The combination of the PCI-Express interface and multiple flashchips that are parallel-controlled makes it so Ultrabook SSDs can also be much **faster** than they are now. Another advantage is that NGFF modules are **even thinner** than mSATA, so laptops can also be made thinner.

Controller 개발팀 심호준 수석

New Physical FF Specifications

51mm x 30mm
z: 4.85mm

mSATA

42mm x 22mm
z: 2.75mm single side[1]
z: 3.85mm double side[1]

NGFF Card **M.2**

Proposed Draft

Serial ATA International Organization

Version 04
August 14, 2012

Serial ATA Technical Proposal #TPR_C112
Title: NGFF Card Format for SSDs

This is an internal working document of the Serial ATA International Organization. As such, this is not a completed standard and has not been approved. The Serial ATA International Organization may modify the contents at any time. This document is made available for review and comment only.

Specification optimized for caching devices or SSDs, includes a series of module lengths and connector keys enabling SATA*, 2x or 4x PCI Express*

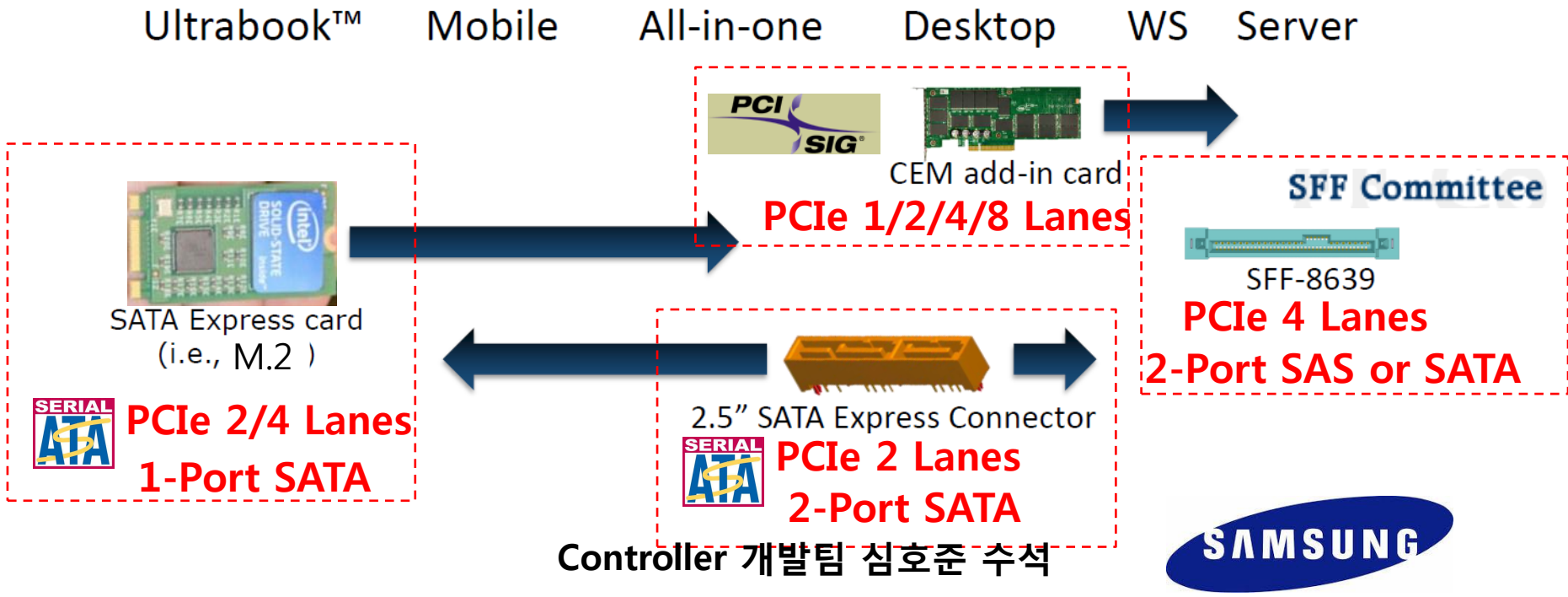
Smaller, thinner, SSD optimized form factor

Controller 개발팀 심호준 수석

SAMSUNG

New Physical FF Specifications

- Form Factor & Connector Landscape
  - CEM Add-in Card for Workstations and Servers
  - SFF-8639 for Enterprise
  - SATA Express (M.2 (NGFF)) for Ultrabook
  - SATA Express (2.5-inch) for Desktop



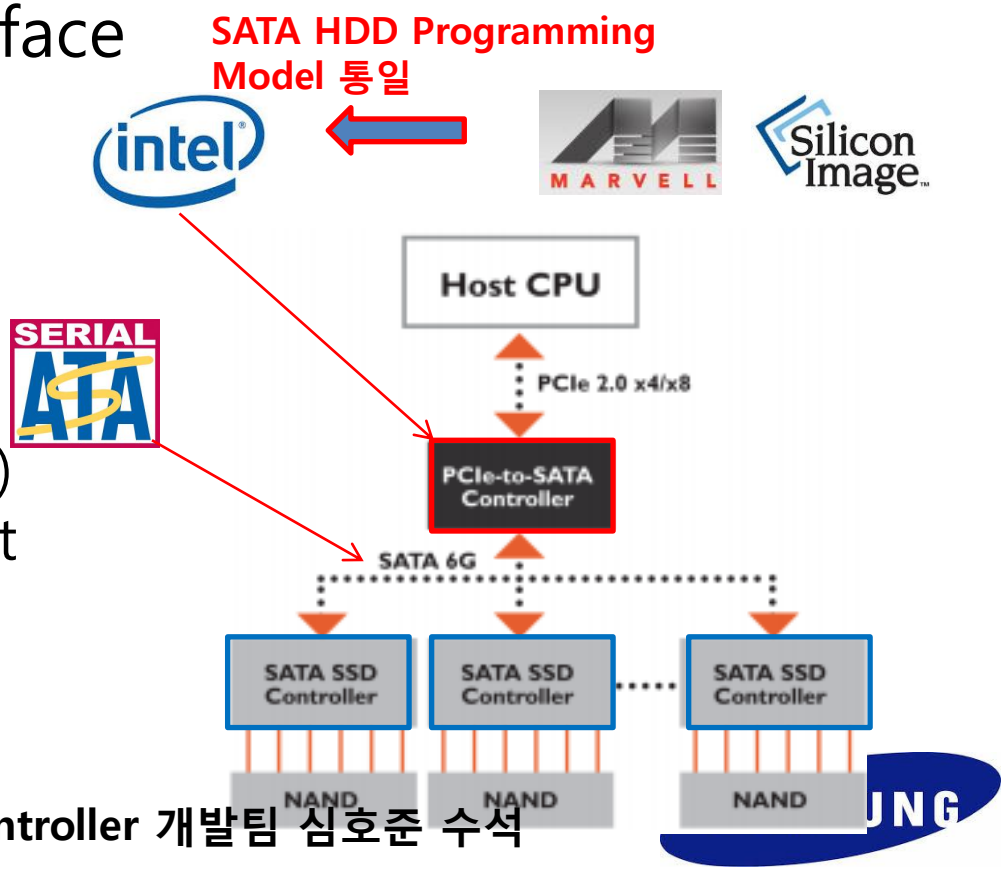
Controller 개발팀 심호준 수석

New Logical Protocol Specifications

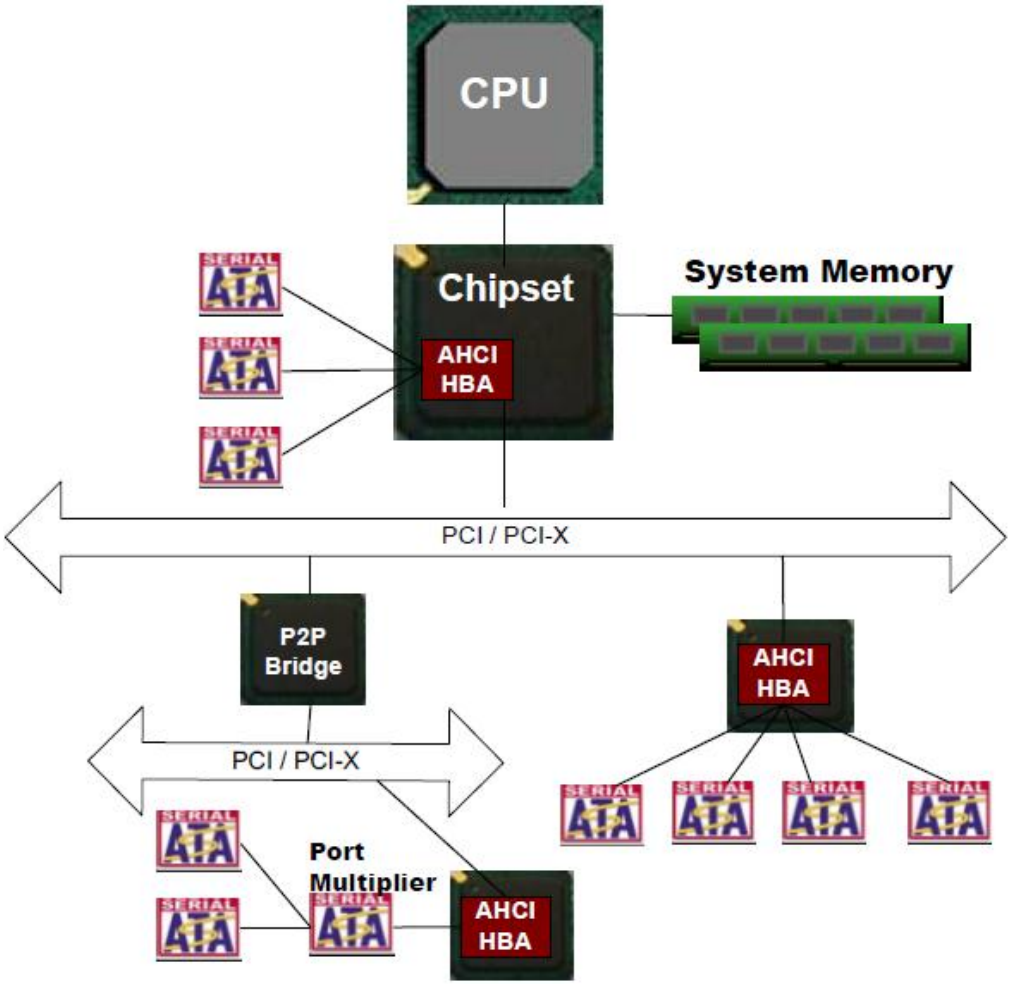
- AHCI
  - Introduced as the Serial ATA Programming Interface in 2004 by Intel
    - James Boyd
  - Designed for HDD
  - Key Features:
    - Native Command Queuing (**32 Cmds**)
    - Power Management Features (Slumber, Partial, etc)

AHCI

. Intel이 다수의 Vendor에 의한 Proprietary Protocol SATA HBA Controller들로 인해서 SATA HDD가 효율적으로 Intel Platform에 통합되지 못하자, AHCI Spec을 제정하고 AHCI Controller를 Intel Chipset에 내장함.



New Logical Protocol Specifications



AHCI

. AHCI는 10년 이상된 Spec으로 SATA HDD를 위해서 만들어졌으며, 당시는 SSD가 탄생하기 전에 만들어진 Spec임

AHCI & SATA Deployment in Intel Architecture

Controller 개발팀 심호준 수석

SAMSUNG

New Logical Protocol Specifications

- NVM Express
  - Introduced as the PCIe SSD Programming Interface in 2011 by NVM Express Working Group
    - Lead by Intel (Amber Huffman)
  - Architected from the ground up for performance
  - Designed for SSDs, with Scalability for Future NVM Technologies (**64K cmds, 64K Qs**)
  - Key features:
    - Optimized Interrupt Architecture for Scalable IOPs
    - Large Scale Parallelism with Many & Deep Command Queues

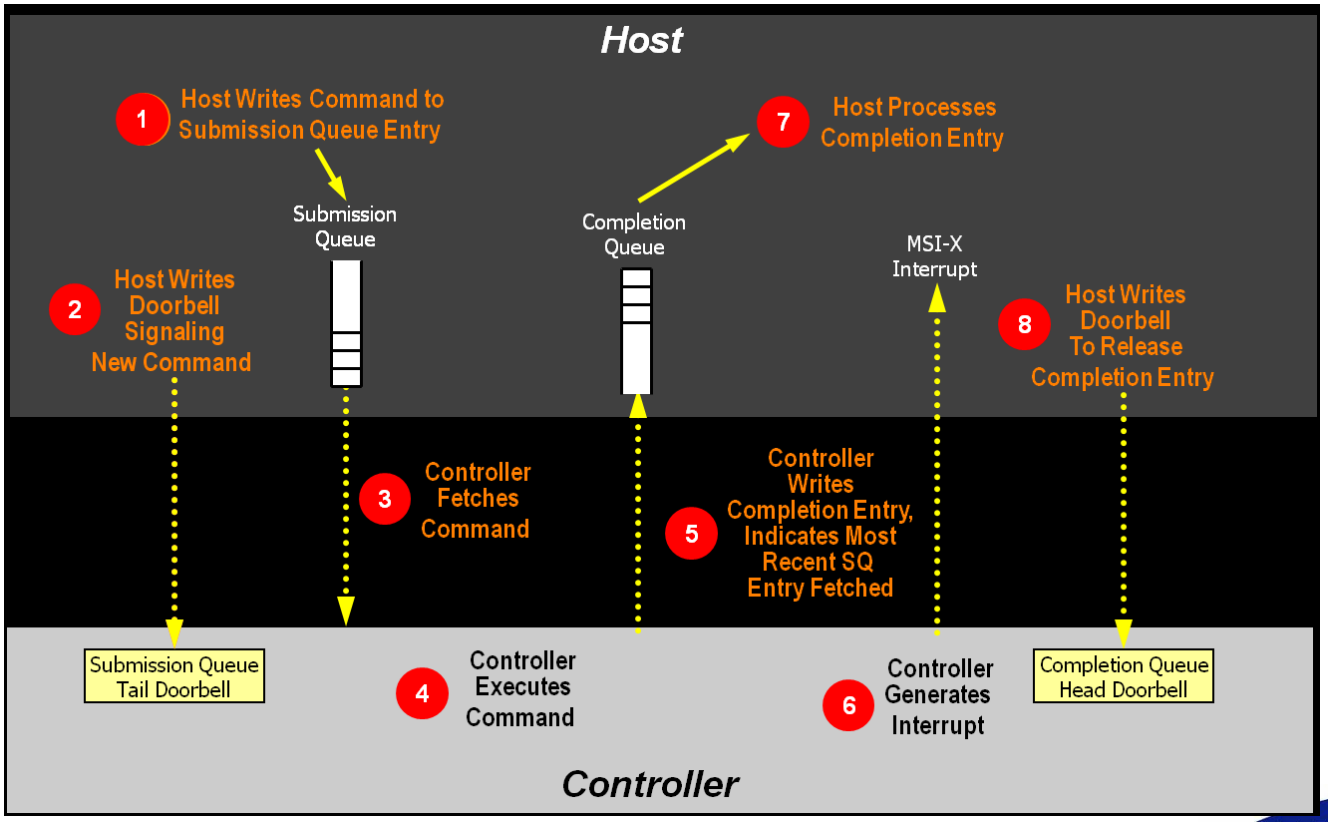
Key NVMe Players

http://uk.hardware
nvm-express-wher

Controller 개발팀 심호준 수석

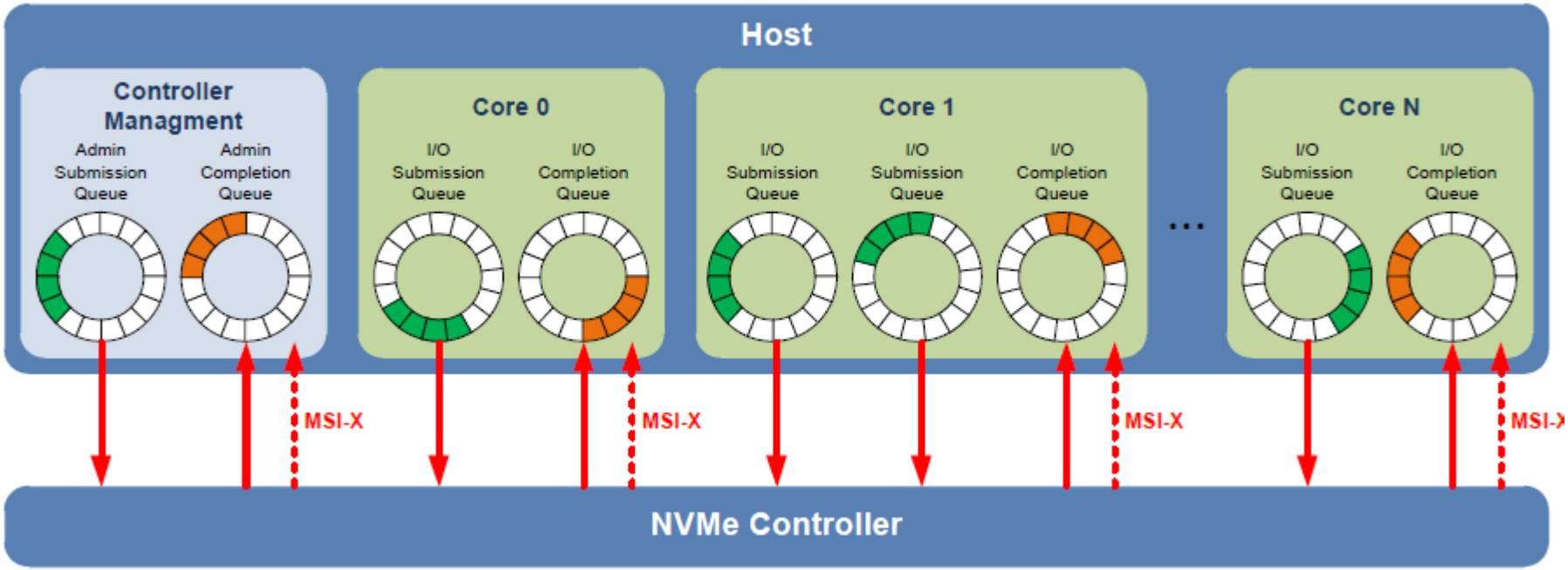
New Logical Protocol Specifications

- NVM Express
  - Command Processing Flow

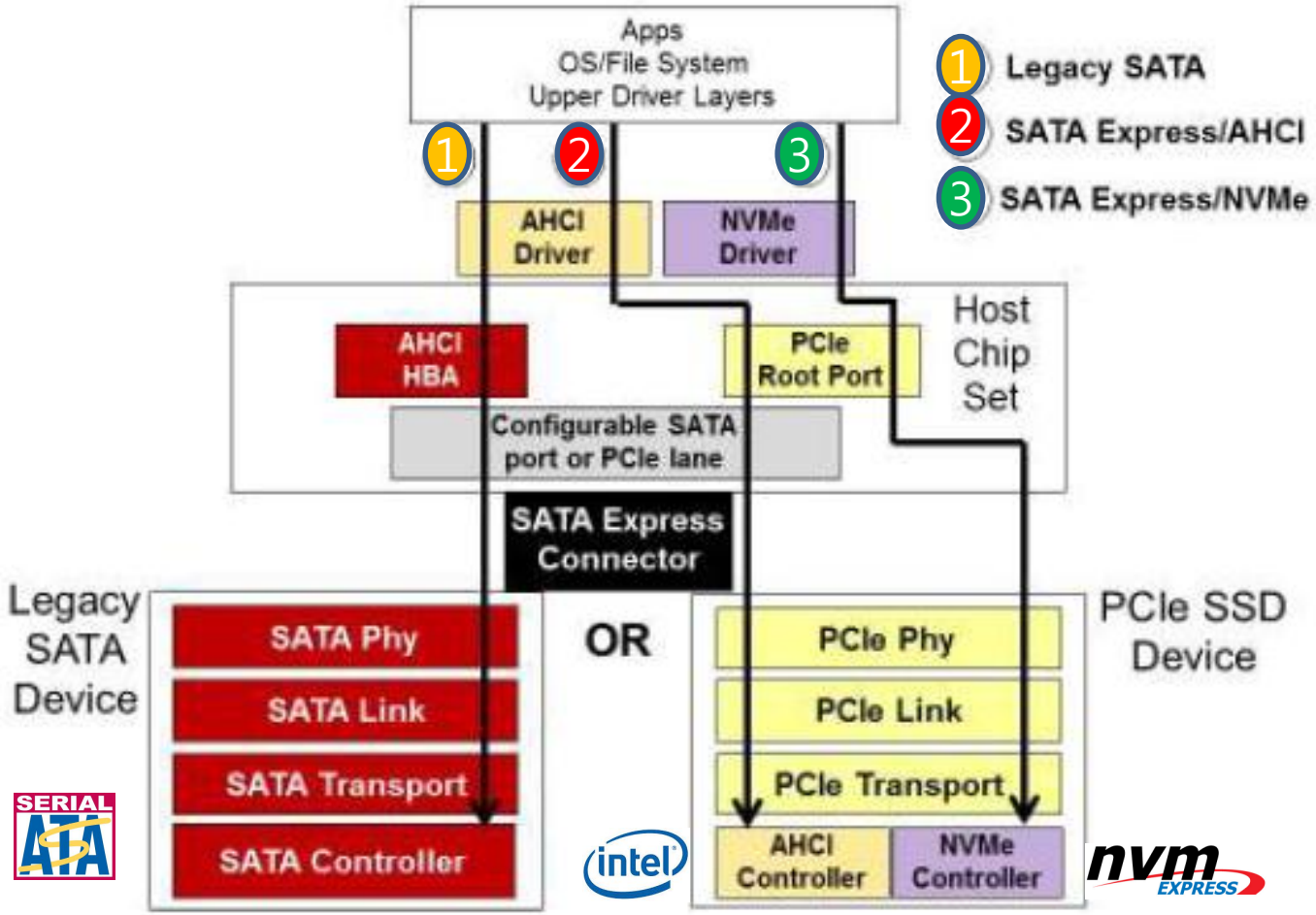


New Logical Protocol Specifications

- NVM Express
  - Large Scale Parallelism: Many & Deep Qs
  - Optimized Interrupt Architecture: MSI-X Vectored Interrupt



New Logical Protocol Specifications



SATA Express SSD System Example (SATA / PCIe (AHCI or NVMe))

SAMSUNG

New Logical Protocol Specifications

- AHCI vs. NVM Express

	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	1 Queue 32 Commands per Q	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

How?
. NVM 특성을 최대한 활용하기 위해서 새로운 Programming Model 정립

How?
. MSI-X Vector Interrupt 사용

How?
. NVMe에서 서로 다른 Submission Q를 사용해서 Name Space를 공유함으로써 가능

NVM Express is optimized for SSDs, replacing the decade old AHCI standard designed for the hard drive era

http://www.nvmexpress.org/wp-content/uploads/2013/04/IDF-2012-NVM-Express-and-the-PCI-Express-SSD-Revolution.pdf

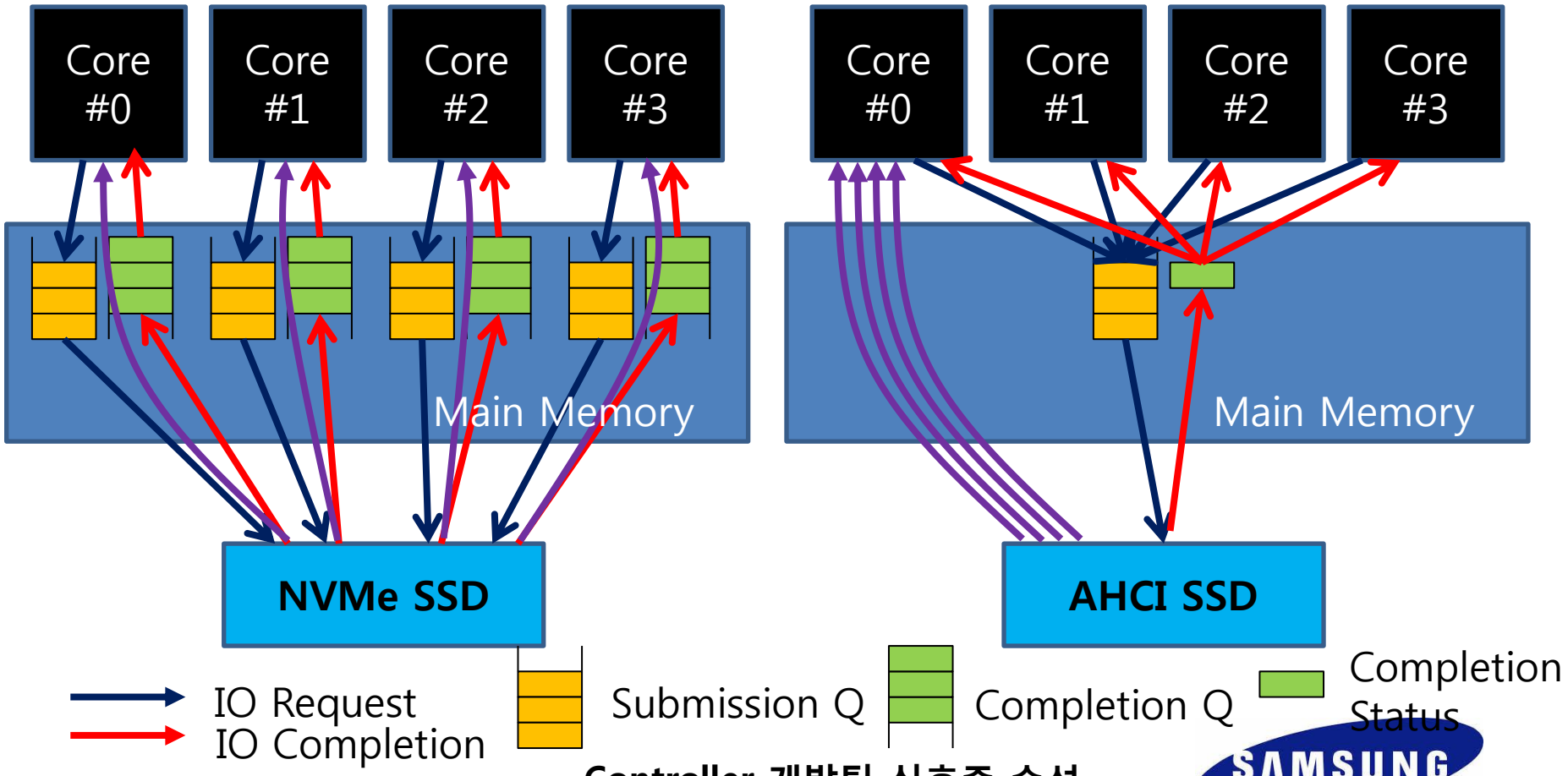
IDF2012 INTEL DEVELOPER FORUM

SAMSUNG

New Logical Protocol Specifications

NVMe is Superior to AHCI
. Lock-less Protocol
. Scalable

- NVM Express
- AHCI

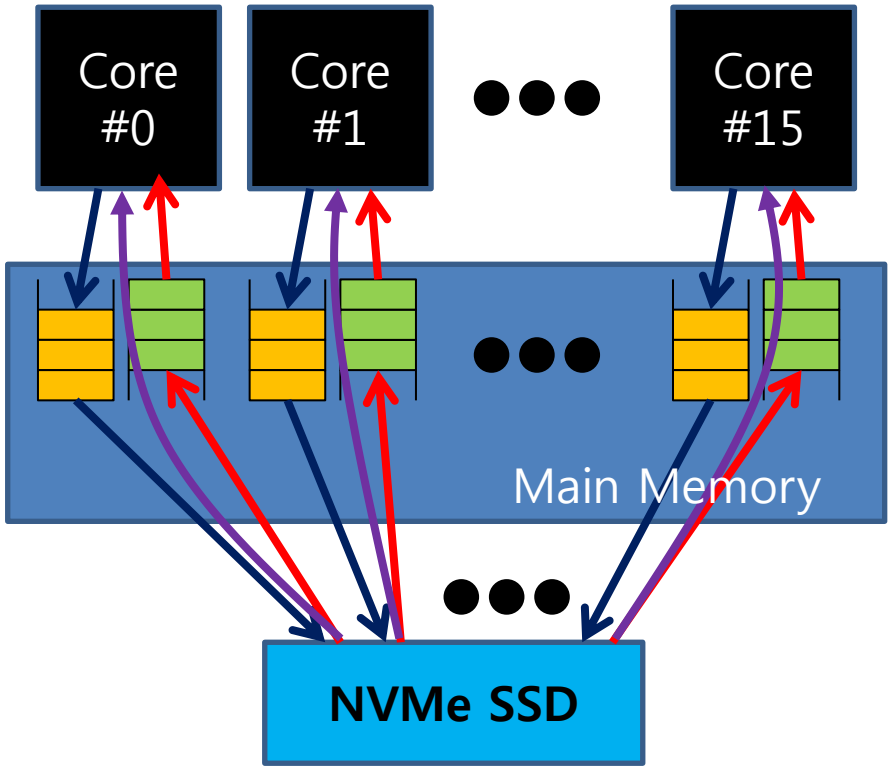


Controller 개발팀 심호준 수석

SAMSUNG

New Logical Protocol Specifications

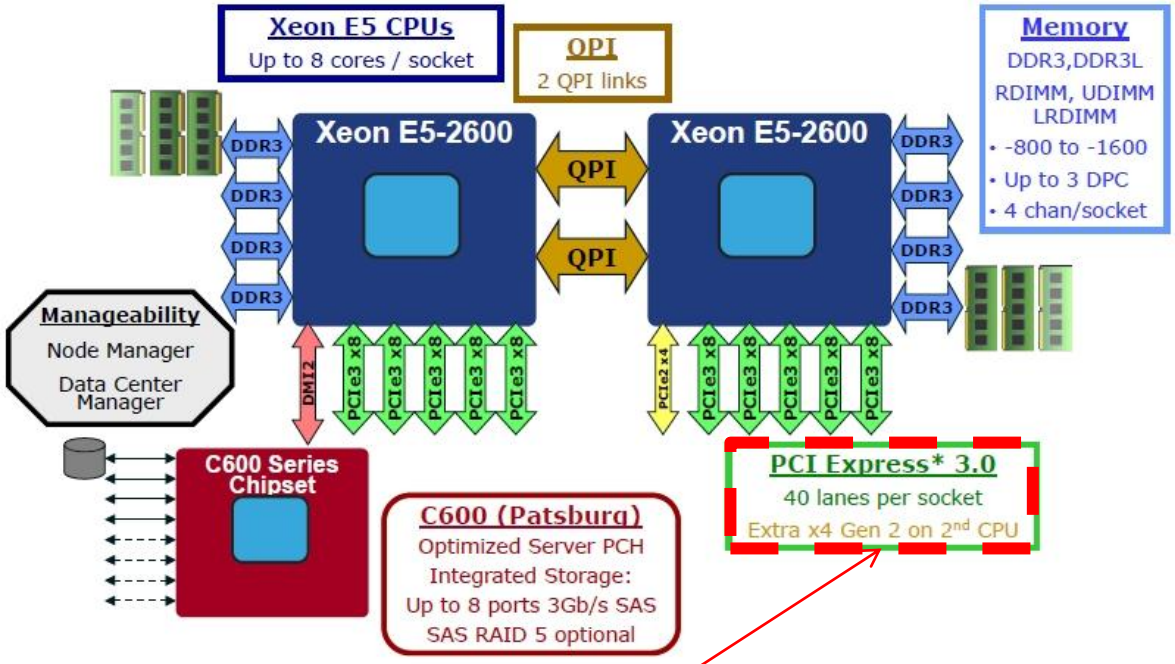
- NVM Express is Capable of Achieving Higher than 1 Million IOPS in Servers or Workstations.
  - 100KIOPS / Core
  - In Total, **1.6MIOPS**



SAMSUNG

New Logical Protocol Specifications

- 16-Core Server Example

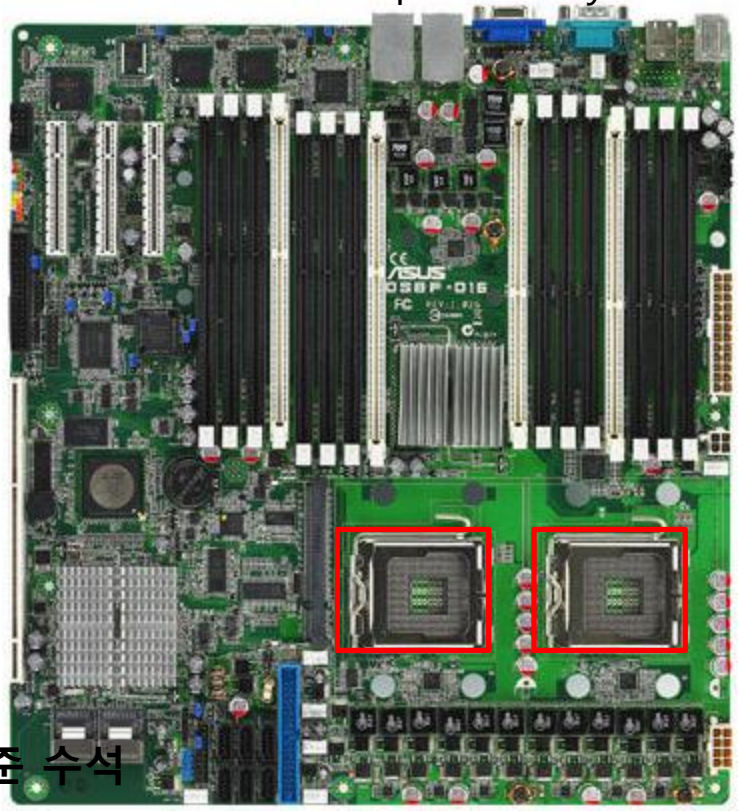


NVM Express SSDs are Attached Here.

DSBF-D16/SAS

. Dual Processor Server Board

. 16-DIMM Expandability



Controller 개발팀 심호준 수석

New Logical Protocol Specifications

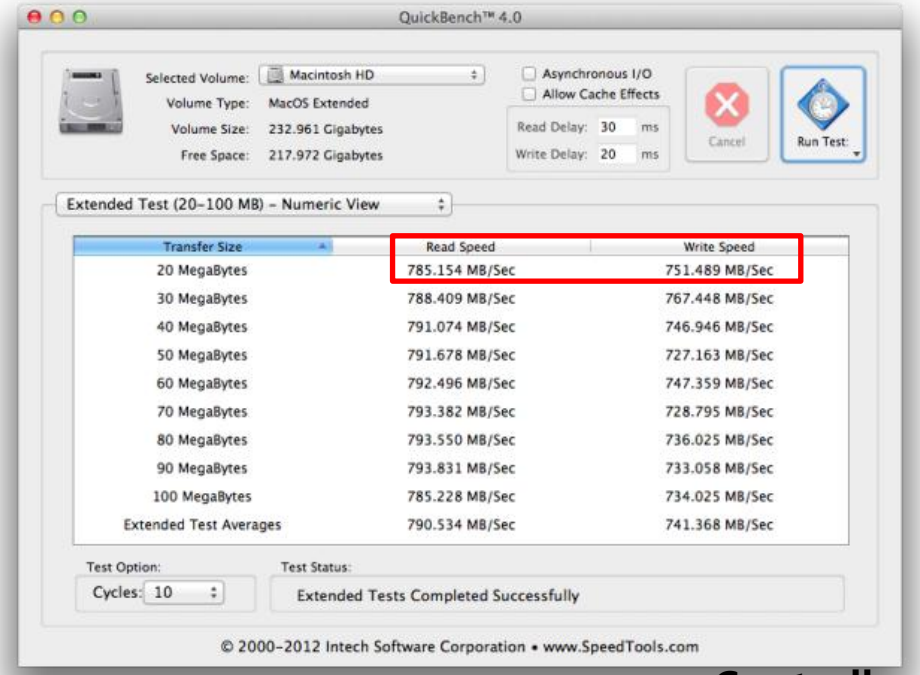
- SCSI Express
  - SOP (SCSI Over PCIe) + PQI (PCIe Queuing Interface)
  - Still in the Draft Stage in T10 (INCITS) Discussion

Samsung PCIe SSD Controllers

- XP941 SSD (UAX)
  - Performance
    - Sequential R/W @Gen2 2-lane = 790/750 MB/S

http://www.anandtech.com/show/7058/2013-macbook-air-pcie-ssd-and-haswell-ult-inside

New MacBook Air (WWDC 2013)



SAMSUNG

SM0256F

Controller 개발팀 심호준 수석

Future Works

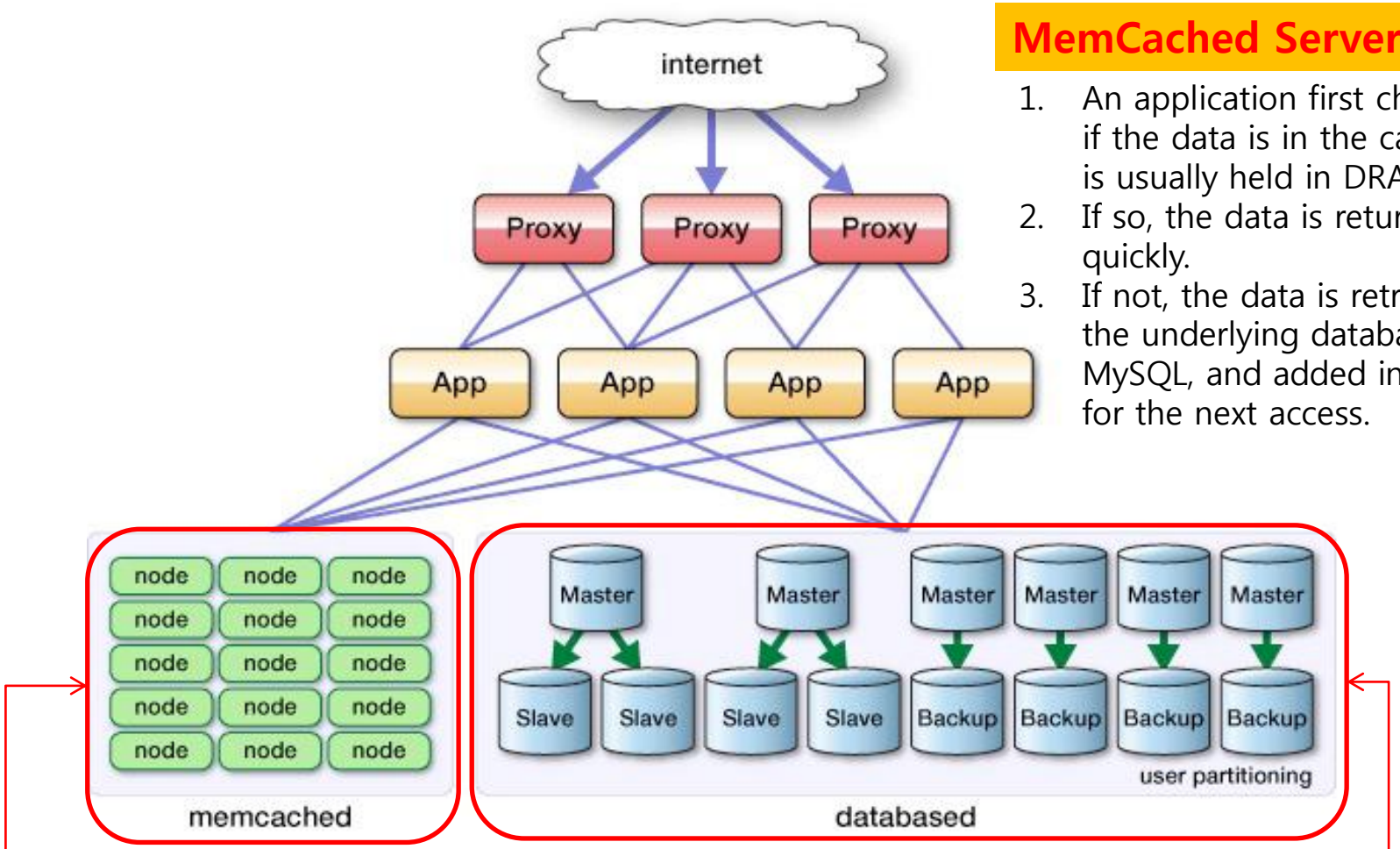
- Memory Hierarchy Vertical Optimization Using a PCIe I/F SSD
  - Low Protocol Overhead With PCIe I/F
  - Example: Virtual Memory Expansion in **MemCached Servers**

Future Works

MemCached Server Example

- Virtual Memory Expansion Example: MemCached Server Virtual Memory
  - Large Percentage of Data Remains Relatively Constant
    - Wikipedia Page Contents
    - YouTube Video Links
    - NetFlix Video Links
  - Poorly Designed Solutions Regenerate Data on Each Request
    - No Regenerate → Regurgitate
    - Caching != Buffering

Future Works



MemCached Server Example

1. An application first checks to see if the data is in the cache, which is usually held in DRAM.
2. If so, the data is returned very quickly.
3. If not, the data is retrieved from the underlying database, such as MySQL, and added into the cache for the next access.

Conventional DRAM Main Memory → PCIe SSDs through Virtual Memory Expansion

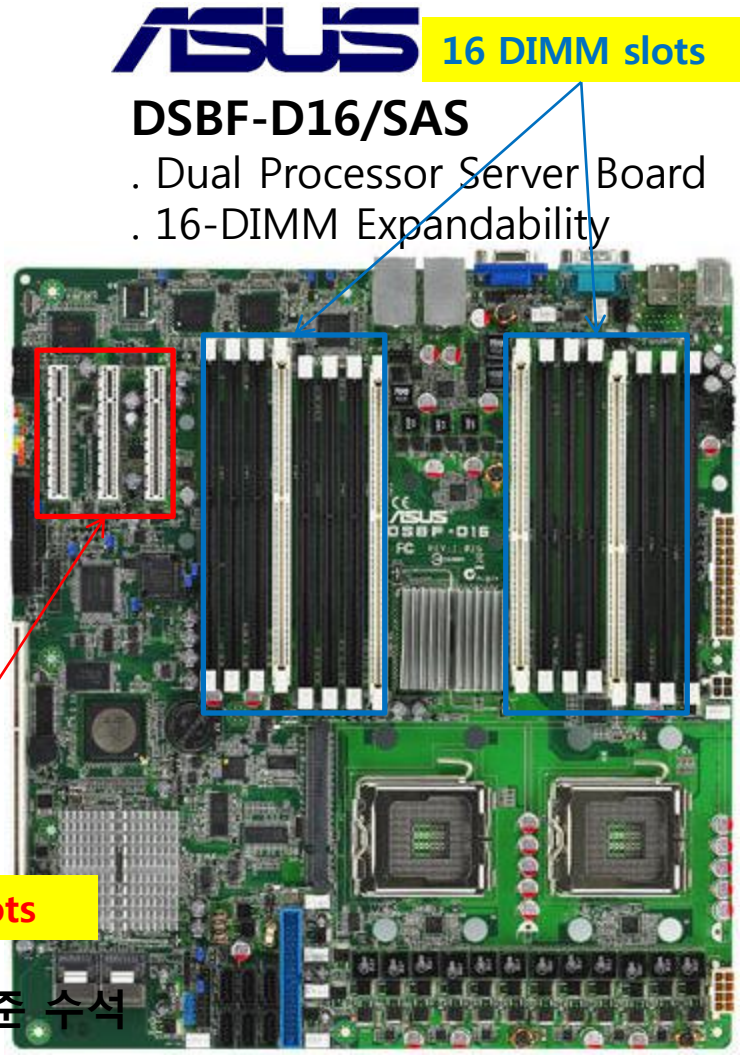
Conventional HDD → SATA/SAS SSDs

SAMSUNG

Future Works

- Virtual Memory Expansion Example: MemCached Server Virtual Memory
  - # of Nodes Decreases by 1/16
    - 16 Nodes, 128GB / Node in DRAM
    - 1 Node, 2TB / Node in PCIe SSD
  - Value Shifts From Processor & DRAM To Flash

MemCached Server Example



Future Works

Value Positioning

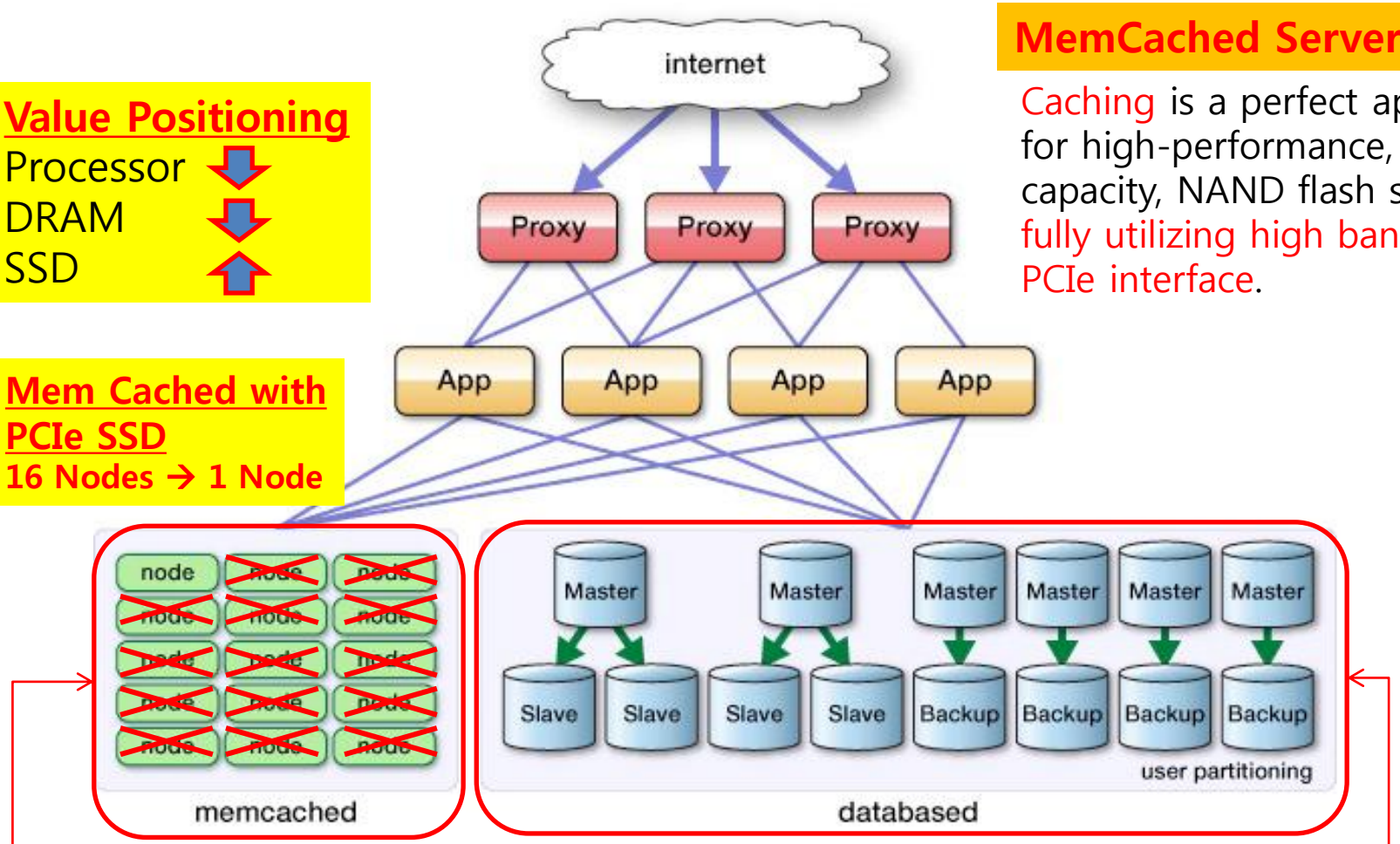
Processor ↓

DRAM ↓

SSD ↑

Mem Cached with PCIe SSD

16 Nodes → 1 Node



MemCached Server Example

Caching is a perfect application for high-performance, high-capacity, NAND flash storage by fully utilizing high bandwidth of PCIe interface.

Conventional DRAM Main Memory → PCIe SSDs through Virtual Memory Expansion

Conventional HDD → SATA/SAS SSDs

Controller 개발팀 심호준 수석

SAMSUNG

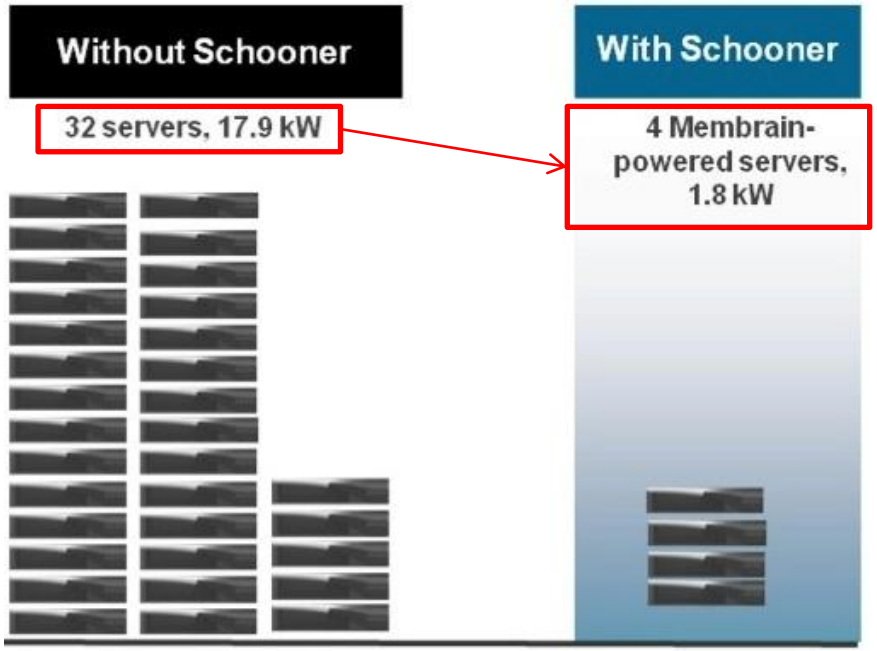
Future Works

MemCached Server Example

- Real World Example

FUSION-IO

Mem Cached with PCIe SSD
. TCO saving & Energy saving
. 32 Nodes → 4 Nodes (12.5%)
. 18kW → 2kW (11.1%)



Traditional memcached in DRAM versus
Schooner Membrain and Fusion-io drives

http://www.sandisk.com/products/enterprise-software/membrain/

Future Works

DRAM Interface SSD Example

- Interface Optimization
- Example: PCIe → DRAM Interface

Publication number	US20130086311 A1
Publication type	Application
Application number	13/629,642
Publication date	Apr 4, 2013
Filing date	Sep 28, 2012
Priority date	Dec 10, 2007
Inventors	Ming Huang, Zhiqing Zhuang
Original Assignee	Ming Huang, Zhiqing Zhuang

DIMM-based SSD technology should be looked at as a serious alternative to expensive high capacity DRAM. Since a single SSD DIMM provides far inure capacity than DRAM DIMM can, the system can then use this storage as a cache or paging area for DRAM operations.

http://www.storage-switzerland.com/Articles/Entries/2011/8/26_SSD_DIMM_-_An_Alternative_to_PCIe_SSD.html

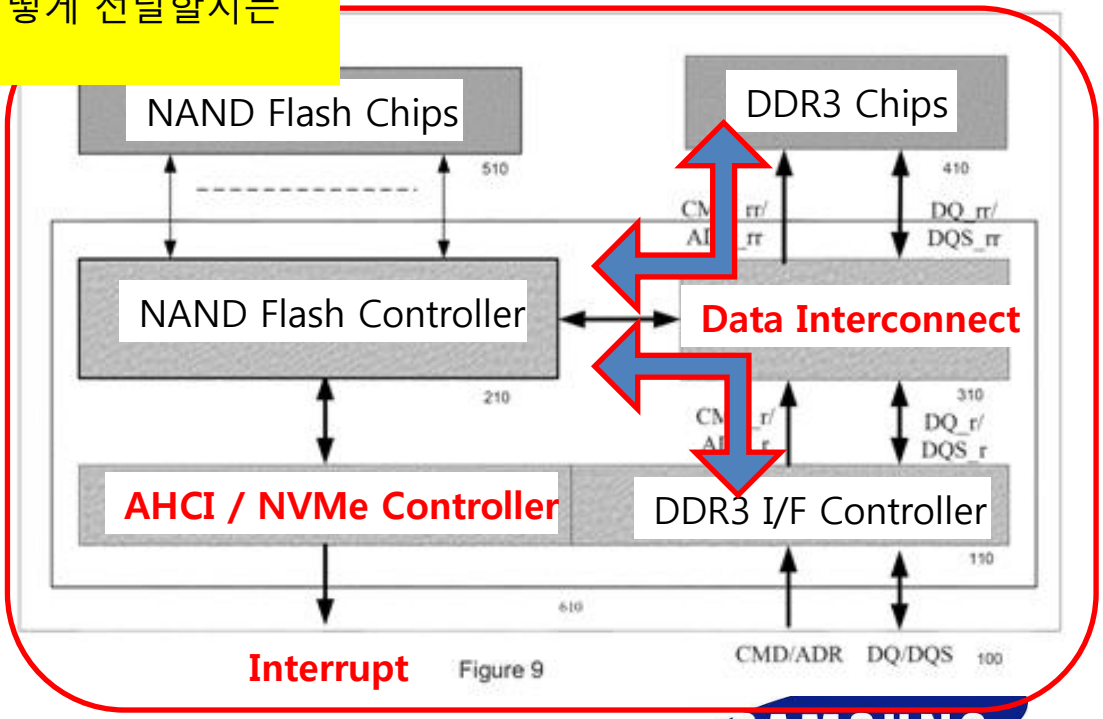
http://www.google.com/patents/US20130086311

Future Works

DRAM Interface SSD
- 싸고 **고용량**의 Main Memory 제공 가능한 솔루션
- NAND Flash Subsystem의 성능이 PCIe 대역폭을 넘는 **초고성능 SSD** 개발 가능
- PCIe I/F 대비 System 전체적인 PCIe Traffic의 양이 획기적으로 줄어들기 **때문에 System 전체적인 Thermal/Power 개선 효과**
- 현재는 특허만 나온 **개념 수준 (시제품 존재여부?)**
- Interrupt 신호를 DIMM 소켓을 통해서 어떻게 전달할지는 현재 불분명

DRAM Interface SSD Example

DRAM Interface SSD Architecture



Value Positioning

DRAM ⇓

SSD ⇑

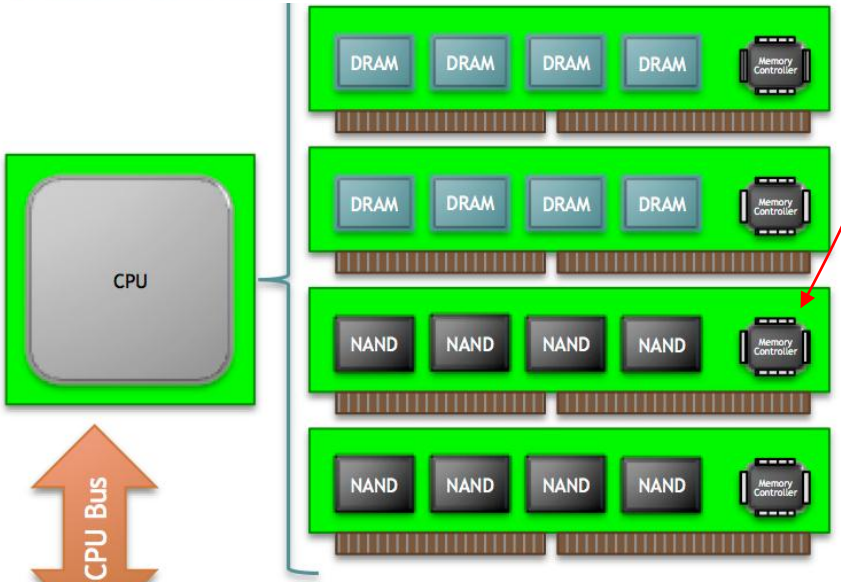
Controller 개발팀 심호준 수석

SAMSUNG

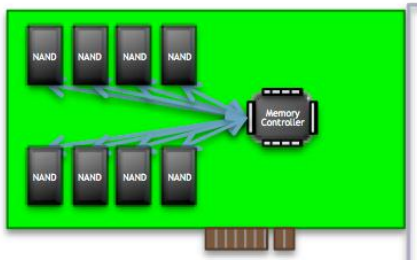
Diablo Technologies:
- RR/RW = 150K / 60KIOPS
- SR/SW = 1GB/S / 760MB/S
- 200GB or 400GB capacity

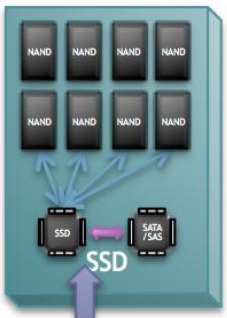
Future Works

DRAM Interface SSD Example



PCIe AHCI / PCIe NVMe /
PCIe SCSIe / …

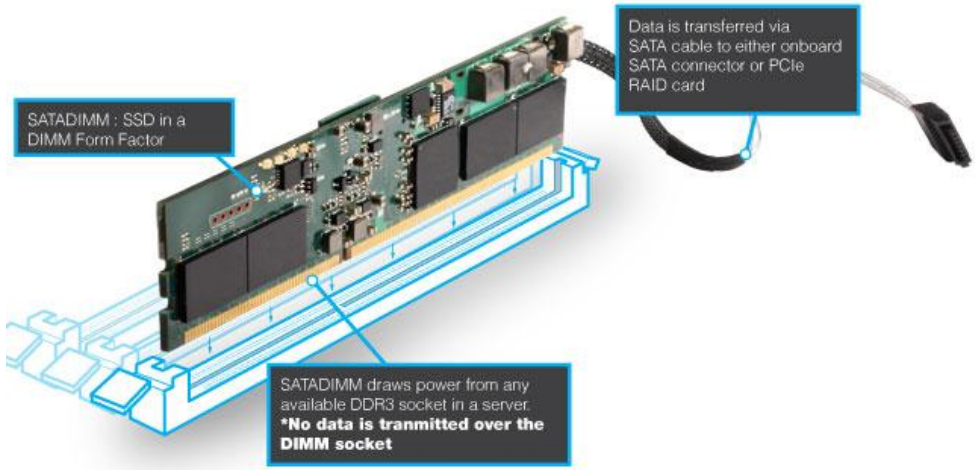




SATA /
SAS / …

Controller 개발팀 심호준 수석

Future Works



DRAM Interface (?) SSD Example

- Cf. NVDIMM (Non-Volatile DIMM)
  - Battery-backed DIMM (DRAM Only, No Flash)

SATADIMM http://www.vikingtechnology.com

NVDIMM (SSD 아님)

viking TECHNOLOGY

SAMSUNG