PCIe SSD Devices - A Year Later
Standards and Drivers

Robert Randall
Windows Device Driver Architect
Micron Technology, Inc.
Learning Objectives

- Review from last year – where did we start?
- Understand the new developments in the PCIe SSD technology space; standards and open source.
- Understand what PCIe SSDs are available today and their characteristics.
- Understand NAND factors; SLC vs. MLC, risks, limitations, fit.
- Understand deployment models; caching, hot-spots, etc.
Review – why PCIe SSDs?

- No 2.5” form factor restrictions
  - More physical space
    - Higher capacity
    - Room for active cooling – NAND gets hot during writes

- PCIe Performance
  - GEN2 500 MB/sec per lane per direction
  - GEN3 1 GB/sec per lane per direction
    - Example: GEN3 x16
    - Nominal performance of 16 GB/sec
Review – Protocol status last year

- T10 SCSI Over PCIe® Architecture (SOP)
  - Transport layer spec for SCSI of PCIe
  - Last year, in early stage, work in progress
  - Strong support, SCSI compliant
- T10 PCIe® architecture Queuing Interface (PQI)
  - Companion document to SOP
  - Host to controller specification.
    - Well, that is new.
    - May use NVM-Express as the PQI implementation
Review – Protocol status last year

- NVM-Express
  - Design goals
    - Scalability, performance, simplicity
    - Ease of implementation
  - Last year, industry support (80 companies) and some adoption
    - Intel demos at Intel Developers Conference.
  - [Open source driver for Linux](#) (GPL)
Protocol status this year

- T10 SOP and PQI
  - SOP almost a standard (letter ballot) and PQI not far behind. Amazing progress in one year.
- NVM-Express (NVMe)
  - Open source driver for Microsoft Windows (BSD) hosted at Open Fabrics Alliance.
  - NVMe Compliance Suite testing framework
    - Intel in partnership with UNH-IOL
    - Wiki
- Open source projects are active.
Protocols – What’s the difference?

- **T10 SCSI Over PCIe (SOP) & PCIe Queuing Interface (PQI)**
  - SOP - Defines a standard way of transporting SCSI commands over PCIe.
  - PQI – Defines a standard way for a host and a controller to queue SCSI commands.
    - You may use NVMe as the PQI for SOP.

- **NVM-Express**
  - Simple protocol (not SCSI).
  - Defines host to controller interface.
Protocols – What’s the difference

- Complimentary, not competitive
  - T10 SOP & PQI
    - Preserves investment in SCSI protocol
      - Existing testing tools can be re-used.
      - Storage may be shared by several initiators
    - May use NVM-Express as PQI implementation
  - NVM-Express
    - No SCSI target implementation needed
    - Promotes a vision of tiered memory similar to tiered storage.
    - Post NAND like Phase Change Memory (PCM).
PCIe SSDs available now - solutions

- Proprietary host to controller interface
  - No NVM-Express or SOP & PQI solutions in the market but expect some soon.
- A wide variety of proprietary options from many vendors
  - DELL, Fusion IO, Intel, LSI, Micron, OCZ, Virident, and others
  - Range from consumer grade to enterprise grade
  - Price and performance vary
Intel has demonstrated NVMe PCIe SSD at IDCs.

IDT NVMe controllers available
- PCIe GEN3 x 8 and 2 x 4 in single package.
- See IDT website for parts.

PCle GEN3 improvements
- Encoding overhead drops from ~20% to ~1.5%
- Over twice as fast as PCIe GEN2
NAND factors SLC vs. MLC

- **Single Level Cell**
  - One bit per cell
  - More reliable
  - Longer wear life
    - High TBW
  - Lower power consumption
  - More expensive / MB
    - Much more…
    - Usually reserved for enterprise class devices

- **Multi Level Cell**
  - Two bits per cell
    - Twice the capacity per cell
  - Less expensive / MB
  - Shorter wear life
    - Lower TBW
  - Slower write speed
    - 1/3 that of SLC
  - Less reliable
    - Higher bit error ratio
NAND factors SLC vs. MLC

- MLC storage devices typically use firmware and/or drivers to improve performance, extend TBW and improve reliability
  - Caching algorithms
    - Improve write performance (write-back)
    - Improve read performance
  - More robust error correction
    - Intelligent wear leveling
    - Extend the life of each cell
Deployment models

- Where do enterprise PCIe SSDs make sense?
  - If time is money, PCIe SSDs offer the best performance in the smallest package.
- Choice between
  - SATA or SAS attached SSDs
    - Traditional form factors
    - Hot swap is traditional and easy to adopt
  - PCIe SSDs
    - Non-traditional storage form factor – PCIe adapter cards
    - One traditional form factor – DELL 2.5” PCIe hot swap
Deployment models

- Example, requirements call for 500K IOPs
  - 1 15K SAS drive delivers ~175 - ~200
  - 1 SATA SSD drive delivers ~400 - ~20k
  - 1 PCIe SSD SLC drive delivers ~ 50K – ~1.6M IOPs
  - Power, cooling, space trade-offs.
  - FC disk array with ~2500 SAS drives or ~3 PCIe SSDs
    - Power consumption ~18525W vs. ~90W
    - Space - 3 PCIe Slots vs. rack of disk array.
    - Cooling – BTUs…
Deployment models

- Heavy write environments need to be careful
  - TBW – how long will the device last before it must be replaced?
  - Power consumption will be higher, more heat.
- Heavy read environments
  - TBW will be longer
  - Less power consumption, less heat.
- Best fit decisions are ALWAYS workload specific
  - Do your homework before making a decision
Deployment trends

- Cloud computing expected to drive adoption of SSDs in general
  - Less power per GB
  - More IOPs and higher bandwidth per device
  - Certain applications will experience big ROI
    - Think Netflix hosted on Amazon cloud
      - All content access is reads
      - Density, power, cooling, performance all favor PCIe SSDs
Wrap up

- Q & A
- Share opinions
- Discuss trends
- How will this impact management models?
References

- NVM-Express  http://www.nvmexpress.org
- Open Fabrics  http://www.openfabrics.org
- T10  http://www.t10.org/drafts.htm
- PCI  http://www.pci.sig