PCI Express®: What’s Next for Storage

Dr. Debendra Das Sharma
Member, PCI-SIG® Board of Directors
Intel Fellow and Director of I/O Technology and Standards
Intel Corporation
Agenda

- Evolution of PCI Express Technology
- Power-efficient Performance
- RAS Enhancements
- I/O Virtualization
- Form Factors
- Compliance
- Conclusions
PCI-SIG® Snapshot

Organization that defines the PCI Express® (PCIe®) I/O bus specifications and related form factors.
- 750+ member companies located worldwide

PCI-SIG continues its solid reputation of delivering low cost, high-performance, low-power specifications to support compliance and interoperability across multiple applications and markets.

- Australia
- Austria
- Belgium
- Brazil
- Bulgaria
- Canada
- China
- Czech Republic
- Denmark
- Finland
- France
- Germany
- Hong Kong
- Hungary
- India
- Ireland
- Israel
- Italy
- Japan
- Malaysia
- Norway
- Russia
- Singapore
- Slovak Republic
- South Korea
- Sri Lanka
- Sweden
- Switzerland
- Taiwan
- The Netherlands
- Turkey
- United Kingdom
- United States
Evolution of PCI/PCIe Technology

PCI started in 1992 as bus based to PCIe Full duplex differential signaling
Five generations of PCI Express architecture with backwards compatibility!
Maintained status as the ubiquitous I/O interconnect through three decades of (r)evolution in compute
Growth of PCIe Technology in Storage

- Data explosion is driving SSD adoption
  - SSD market CAGR of 14.8% during 2016-2021 Source: IDC
  - PCIe SSD market to surpass a CAGR of 33% during 2016-2020 Source: Technavio
- PCIe technology is outpacing other interconnect technologies in both units and bandwidth/capacity

Source: SSD Insights Q1/18, Forward Insights
PCIe in Storage

Performance and user benefits for current and future storage applications

Faster data transfer:

- **PCI Express 3.0 Specification (8GT/s) published in 2010**
  - Low power with high performance
  - Wide breadth of solutions available from numerous vendors
  - Provides the cost effective performance required for storage today

- **PCI Express 4.0 Specification (16GT/s) finalized and published in October 2017**
  - Numerous vendors confirmed with 16GT/s PHYs in silicon
  - Major IP vendors offering 16GT/s controllers

- **PCI Express 5.0 (32GT/s) Specification targeted for release in Q1 2019**
  - Protocol already supports higher speed via extended tags and credits and additional changes target speed transition

Better user experience:

- Client and enterprise storage applications using PCIe technology helps keep data closer to CPU
Evolution of PCIe in Platforms

Continuous Improvement: Data Rate, Protocol enhancements, Power enhancements, Form-factor, and Usage Models

Doubling Bandwidth & Improving Capabilities Every 3-4 Years

Relevant through evolution of platforms across multiple market segments
PCIe Architecture Layering for Modularity and Reuse

**Software**

**Transaction**
- Split-transaction, packet-based protocol
- Credit-based flow control, virtual channels

**Data Link**
- Logical connection between devices
- Reliable data transport services (CRC, Retry, Ack/Nak)

**Logical PHY**
- Physical information exchange
- Interface initialization and maintenance

**Electrical**
- Market segment specific form factors
- Evolutionary and revolutionary

**Mechanical**
PCI-SIG and SDA Liaison

- PCI-SIG and the SD Association have formed a liaison to advance SD Express as a component in the PCIe ecosystem
  - Collaborate on a technical interchange related to SD Express and SD Express Test Guidelines, as well as information related to PCIe electrical certification of SD Express products
  - Form the PCI-SDA Advisory Team and the SD-PCIe Technical Group whose members are from companies that belong to both the SDA and PCI-SIG

- New SD Express Card leverages PCIe 3.0 interface to deliver up to 985 MB/s transfer rate
  - Maintains backward compatibility with existing SD hosts in the market
  - Meets changing performance levels of mobile and client computing, imaging, gaming, IoT and automotive applications
Power Efficient Performance

- **Scalable Performance**
  - Width scaling: x1, x2, x4, x8, x12, x16,
  - Frequency scaling: Five generations
    - 2.5 and 5 GT/s with 8b/10b encoding
    - 8 and 32 GT/s with 128b/130b encoding

- **Low Power (Active/Idle)**
  - Rich set of Link and Device States
  - L0s, L1, L1-substates, L2/L3
  - D0, D1, D2, D3_hot/cold
  - Platform-level power optimization hooks: Dynamic Power Allocation, Optimized Buffer, Flush Fill, Latency Tolerance Reporting
  - Active power –5pJ/b, Standby power: 10uW/Lane

- **Vibrant ecosystem with IP Providers**
RAS Features

- PCIe architecture supports a very high-level set of Reliability, Availability, Serviceability (RAS) features
  - All transactions protected by CRC-32 and Link level Retry, covering even dropped packets
  - Transaction level time-out support (hierarchical)
  - Well defined algorithm for different error scenarios
  - Advanced Error Reporting mechanism
  - Support for degraded link width / lower speed
  - Support for hot-plug
DPC/ eDPC Motivation and Mechanism

• Recently added (enhanced) Downstream Port Containment (DPC and eDPC) for emerging usages
• Emerging PCIe usage models are creating a need for improved error containment/recovery and support for asynchronous removal (a.k.a. hot-swap)
• Defines an error containment mechanism, automatically disabling a Link when an uncorrectable error is detected, preventing potential spread of corrupted data
• Reporting mechanism with Software capability to bring up the link after clean up
• Transaction details on a timeout recorded (side-effect of asynchronous removal)
• eDPC: Root-port specific programmable response to gracefully handle DPC downstream
I/O Virtualization

- Reduces System Cost and power
- Single Root I/O Virtualization Specification
  - Released September 2007
  - Allows for multiple Virtual Machines (VM) in a single Root Complex to share a PCI Express* (PCIe*) adapter
- An SR-IOV endpoint presents multiple Virtual Functions (VF) to a Virtual Machine Monitor (VMM)
  - VF allocated to VM => direct assignment
- Address Translation Services (ATS) supports:
  - Performance optimization for direct assignment of a Function to a Guest OS running on a Virtual Intermediary (Hypervisor)
- Page Request Interface (PRI) supports:
  - Functions that can raise a Page Fault
- Process Address Space ID enhancement to support Direct assignment of I/O to user space
Range of SFF Form Factors

Current PCIe Form Factors

- Low Power NVMe
  - M.2 80mm and 110mm
  - U.2 2.5in x 7mm

- Server Performance NVMe
  - Low profile HHHL x4 AIC
  - U.2 2.5in x 15mm

- Server Performance NVMe
  - Low profile HHHL x8 AIC

EDSFF family

- EDSFF 1U Short
- EDSFF 1U Long x4, x8 (ruler)
- EDSFF 1U Long 18mm heatsink
PCIe Compliance Process – Enabling a Robust Ecosystem

PCI-SIG® Specs

Describes
Device requirements
- 3.0 Base and CEM specs

C&I Test Specs
Define
Test criteria based on spec requirements
- Test Definitions
- Pass/Fail Criteria

Test H/W & S/W
Validates
Test criteria
- Compliance
- Interoperability

Clear Test Output Maps
• Directly to Test Spec

Test Tools And Procedures
PASS
FAIL

Predictable path to design compliance
Conclusions

- Single standard covering systems from handheld to data center
- Predominant direct I/O interconnect from CPU with high bandwidth
- Low-power
- High-performance
- Predictive performance growth spanning five generations
- A robust and mature compliance and interoperability program