

# PCIe SSD Devices

**Robert Randall**  
**Micron Technology, Inc.**

- ❑ PCIe and its impact on SSDs
  - ❑ **WARNING:** a bit of a hardware chat...
  - ❑ Explore the history and evolution of drive interconnects – how did we get to PCIe?
  - ❑ Introduce the industry standards available for PCI SSD implementation – how do you do it?
  - ❑ Speculate on futures that will extend the reach of PCIe SSD applications – what might be next?

# Block Storage Interconnects

- ❑ Parallel IDE/ATA/ATAPI and parallel SCSI
- ❑ Fibre Channel SCSI drives and SANs
- ❑ Serial ATA → SATA
- ❑ Serial attached SCSI → SAS
- ❑ iSCSI → SCSI encapsulated in TCP, iSER, iWARP
- ❑ Infiniband fabrics → SCSI RDMA Protocol (SRP)
- ❑ FCoE → FC encapsulated in TCP / GFPT / MPLS / raw Ethernet
- ❑ ATAoE → ATA over raw Ethernet

# Why PCIe for SSDs?

- ❑ Be honest, lower cost of packaging
  - ❑ adapter card not disk form factor
- ❑ Performance
  - ❑ Serialized transfers and multiple lanes
    - ❑ Simultaneous DMA transfers from multiple adapters
    - ❑ Nominal 250 MB/sec per lane per direction
  - ❑ No bus arbitration, elimination of mastering
  - ❑ Message Signaled Interrupts (MSI)
    - ❑ Not limited by a single line based interrupt

# PCIe GEN 2 is Enabler

- ❑ Message Signaled Interrupt – Extended (MSI-X)
  - ❑ Up to 2048 interrupts (minimum of 64)
    - ❑ Interrupts target a processor for better scaling
- ❑ 64bit addressing and interrupt masking
- ❑ Nominal 500 MB/sec per lane per direction
  - ❑ 4 lanes → 2 GB/sec (20 GT/sec)
  - ❑ 8 lanes → 4 GB/sec (40 GT/sec)
  - ❑ 16 lanes → 8 GB/sec (80 GT/sec)
- ❑ **POWER!** (from Jezzers on TopGear)

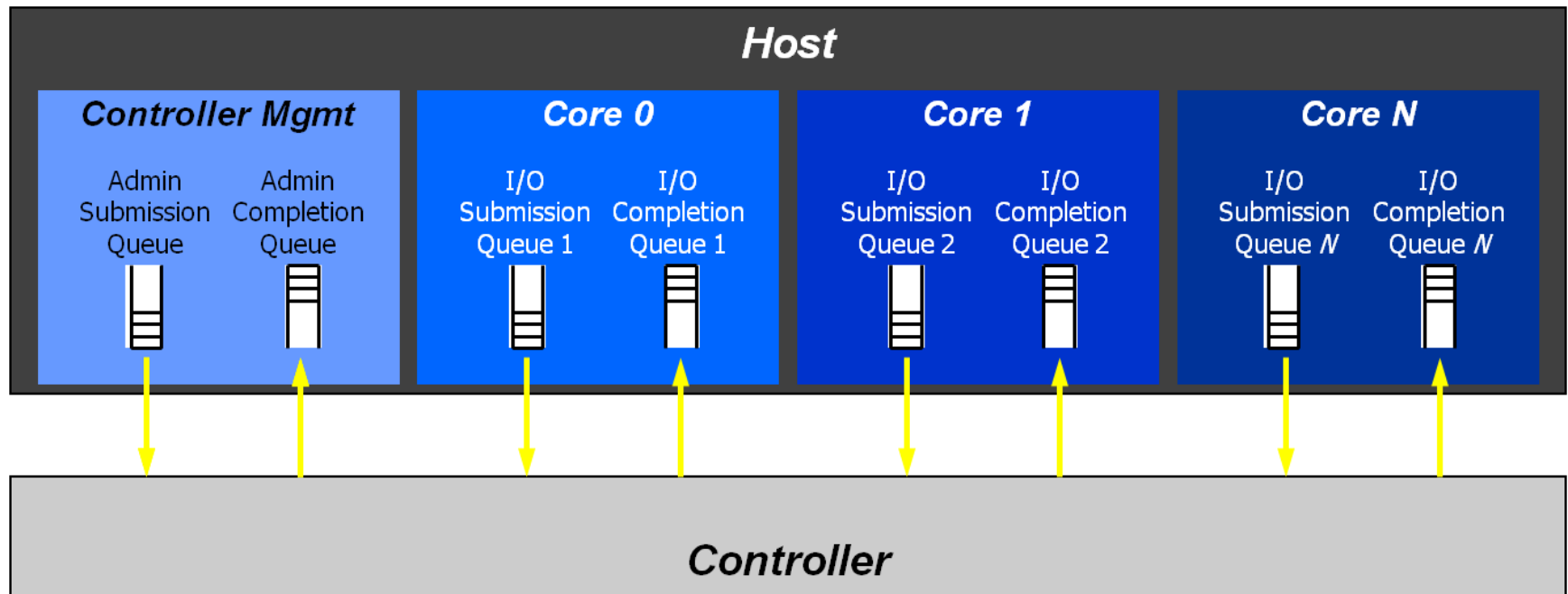
- ❑ PCIe and NVM storage protocols
  - ❑ NVM-Express ([www.nvmexpress.org](http://www.nvmexpress.org))
    - ❑ Started as derivation of AHCI called NVMHCI
    - ❑ In development for many years
    - ❑ Specification 1.0b available for download
    - ❑ Open source Linux driver contributed by Intel.
  - ❑ T10 Standards in development ([www.t10.org](http://www.t10.org))
    - ❑ SCSI Over PCIe® Architecture (SOP)
      - ❑ The transport layer
    - ❑ PCIe® architecture Queuing Interface (PQI)
      - ❑ The PCIe  $\leftrightarrow$  SCSI device layer

- NVM Express is a scalable host controller interface designed to address the needs of Enterprise and Client systems that utilize PCI Express based solid state drives. The interface provides an optimized command issue and completion path. It includes support for parallel operation by supporting up to 64K I/O Queues with up to 64K commands per I/O Queue. Additionally, support has been added for many Enterprise capabilities like end-to-end data protection (compatible with T10 DIF and SNIA DIX standards), enhanced error reporting, and virtualization.

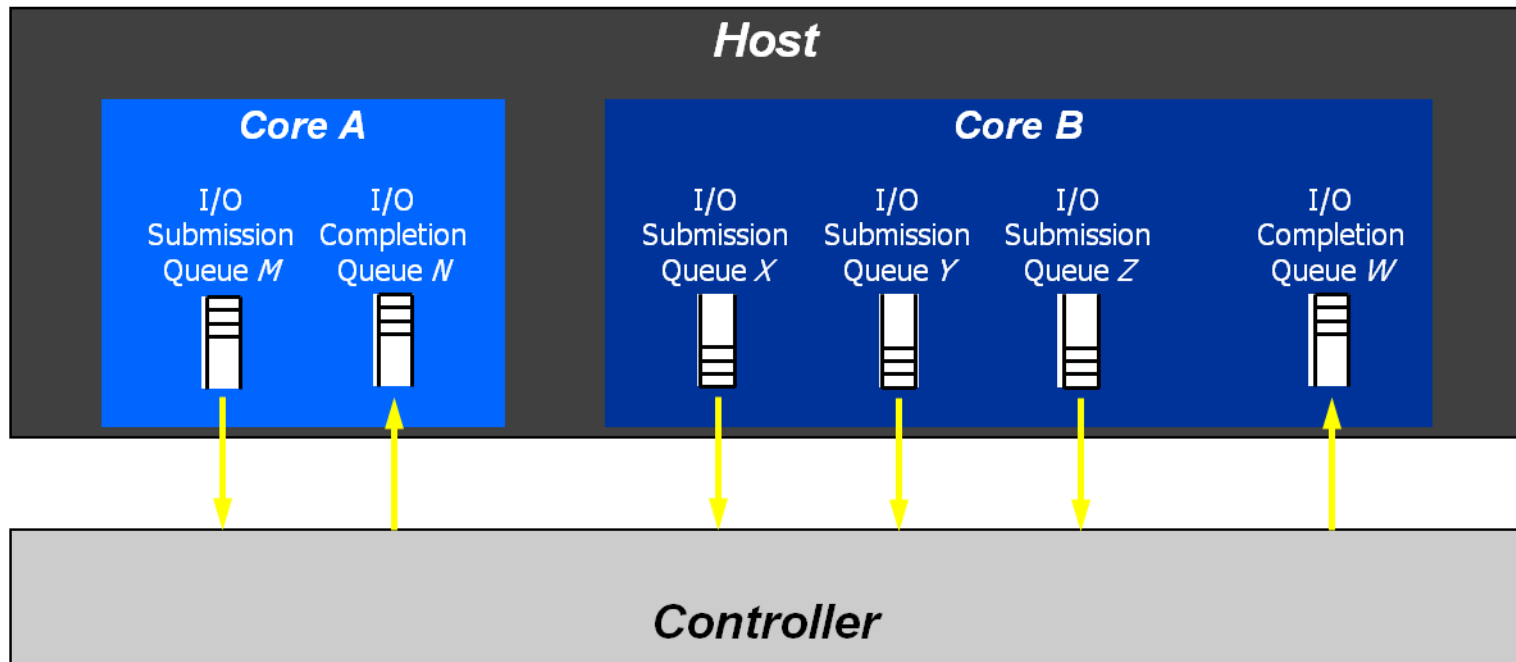
# Important Features

- ❑ Does not require un-cacheable / MMIO register reads in the command issue or completion path.
- ❑ A maximum of one MMIO register write is necessary in the command issue path.
- ❑ Support for up to 64K I/O queues, with each I/O queue supporting up to 64K commands.
- ❑ Priority associated with each I/O queue with well-defined arbitration mechanism.
- ❑ All information to complete a 4KB read request is included in the 64B command itself, ensuring efficient small I/O operation.
- ❑ Efficient and streamlined command set.
- ❑ Support for MSI/MSI-X and interrupt aggregation.
- ❑ Efficient support for I/O virtualization architectures like SR-IOV.
- ❑ Robust error reporting and management capabilities.

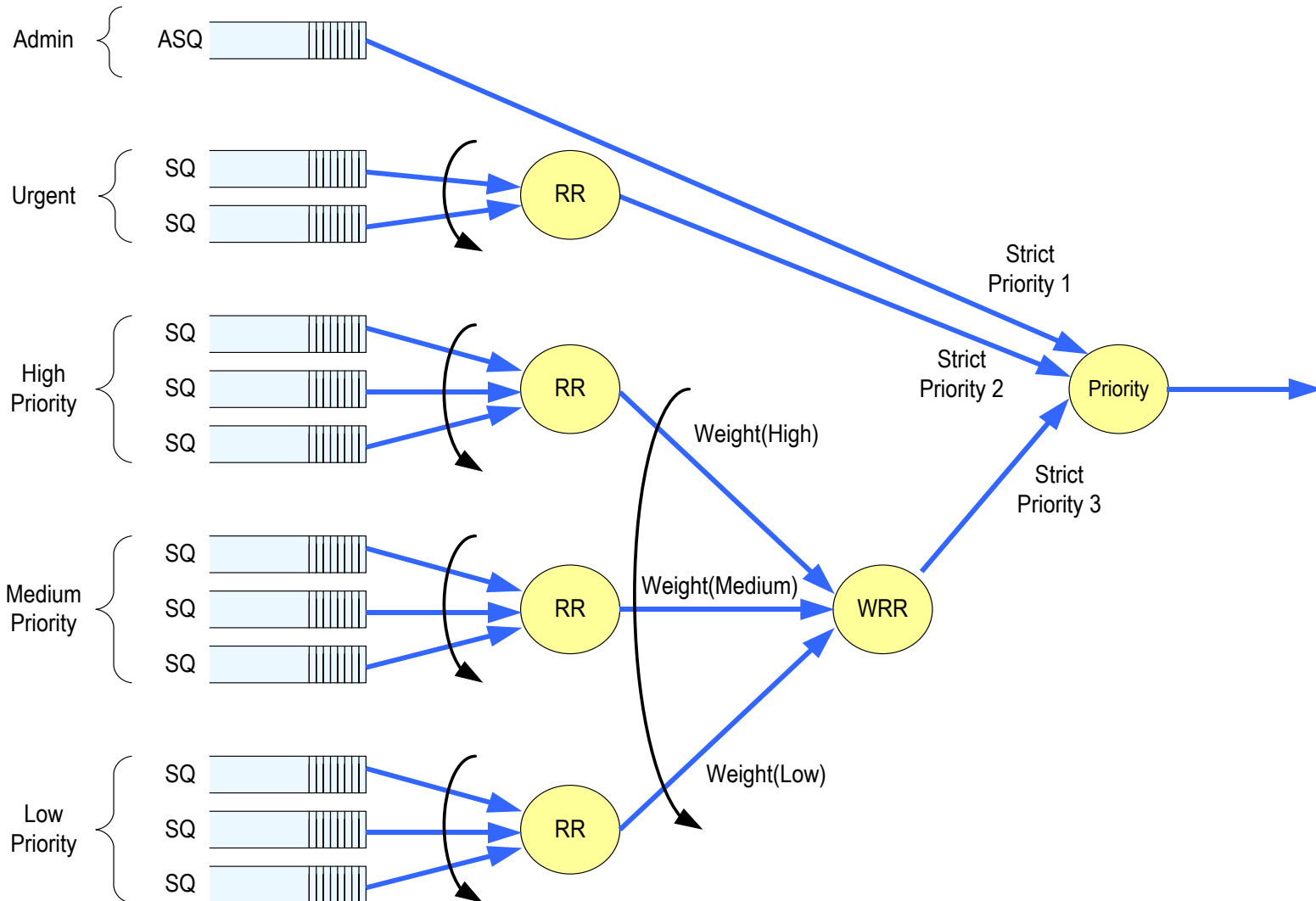
# Basic Architecture



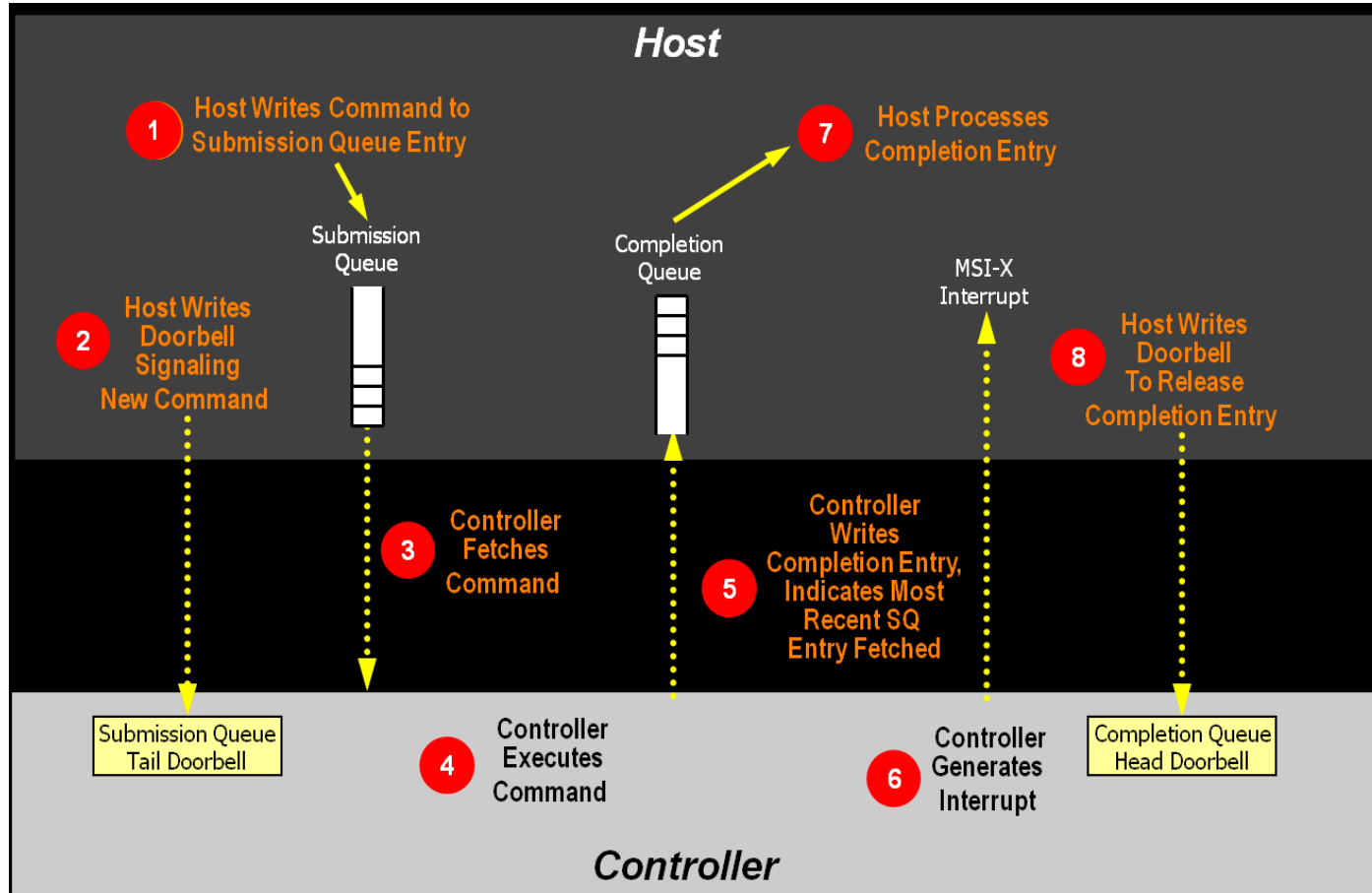
# N Submission → I Completion



# Priority Based Command Queuing



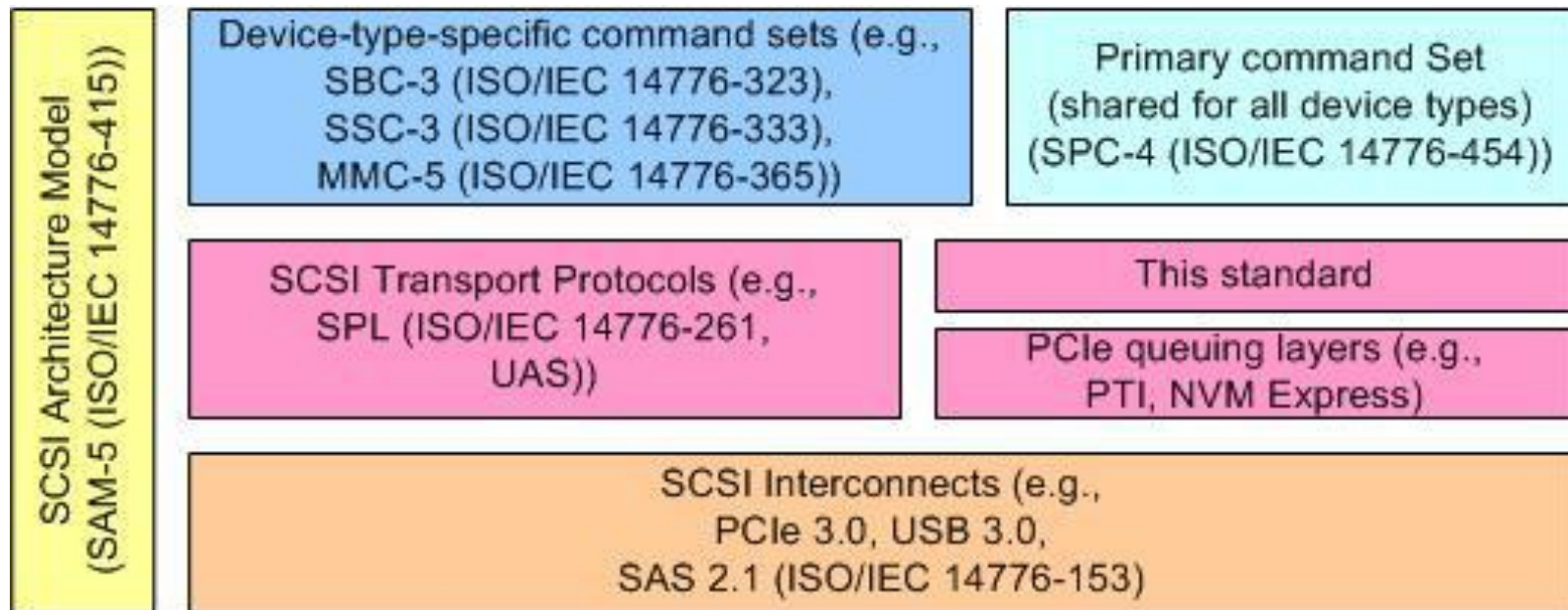
# Command Processing



- ❑ Designed to take full advantage of PCIe GEN 2 features.
- ❑ Designed to minimize host CPU load
- ❑ Designed for high performance.
- ❑ Design is light weight, low overhead

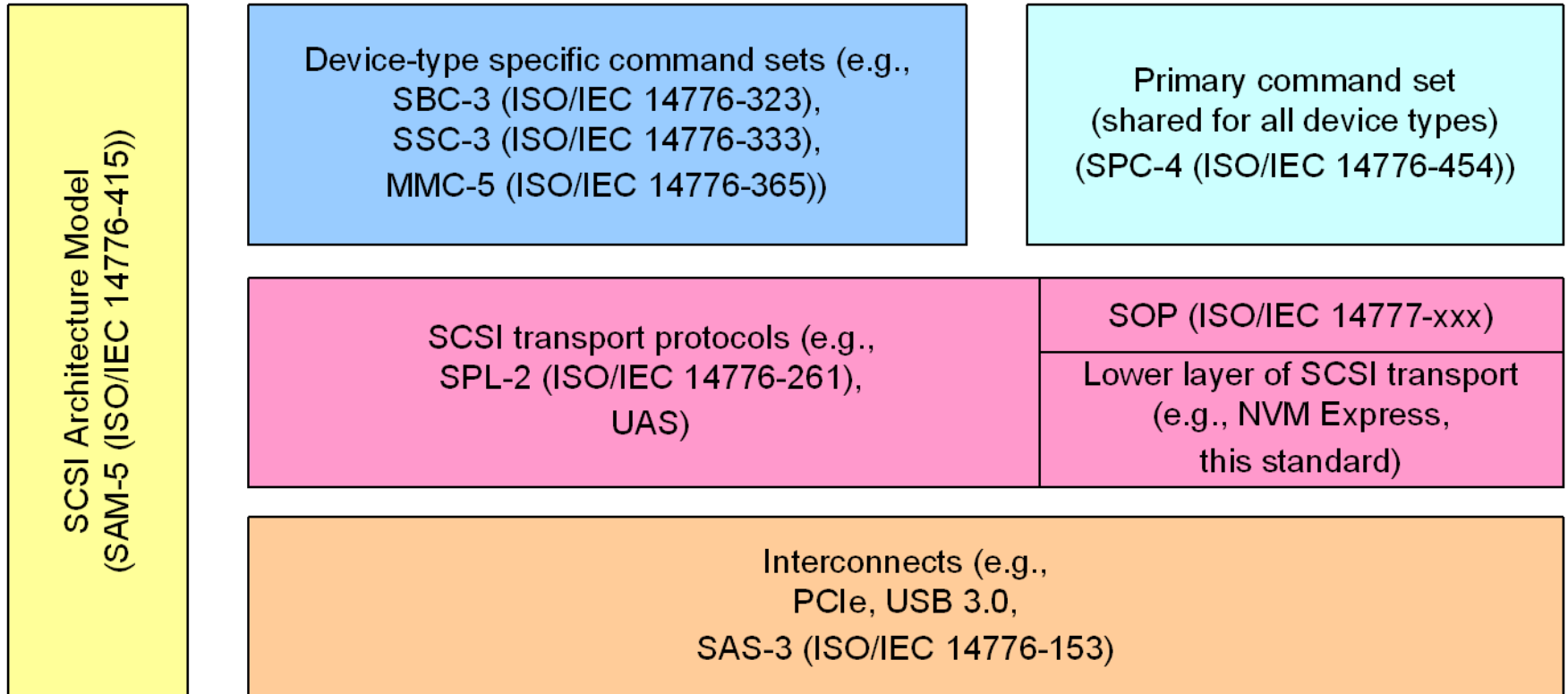
# T10 SCSI over PCIe® (SOP)

- Use PCIe® as a transport for SCSI commands
  - Complimentary to NVM-Express



- ❑ In the early stages (draft 0 in the works)
- ❑ Appears to target PCIe GEN 3
- ❑ SCSI protocol is THE standard in block devices.
- ❑ Allows for re-use of existing storage industry testing technologies for testing PCIe SSD compliance, integrity, etc.
- ❑ Very active working group based on teleconference traffic and meeting minutes.

# T10 PCIe® Queuing Interface (PQI)



- ❑ Ambitious
- ❑ Target and Initiator support
- ❑ Alternative to NVM-Express
- ❑ Very active working group based on teleconference traffic and meeting minutes.

- ❑ PCIe GEN 3
  - ❑ Twice as fast as GEN 2
- ❑ PCIe over cable with external switches?
  - ❑ Thunderbolt (Intel)
- ❑ Multiport PCIe SSD devices?
- ❑ PCIe SSD on main board?
  - ❑ Blades, high density, low power / cooling
- ❑ NVM-Express vs. T10 SOP & PQI

# Wrap up

## □ Q & A

# References

- ❑ NVM-Express <http://www.nvmexpress.org>
- ❑ T10 [http://www.t10.org/drafts.htm#SCSI3\\_PCI](http://www.t10.org/drafts.htm#SCSI3_PCI)
- ❑ PCI <http://www.pci.sig>