Towards Million IOPS

Wei Zhou

Marvell Semiconductor Corp

SSD is all about IOPS

- NAND flash / SSD offers dramatic small block random read/write performance enhancement compared with traditional hard drive
- A typical state of art SATA SSD can reach 50 – 70K IOPS for 4K random read, 20 – 40K IOPS for 4K random write
- PCIE based SSD showing another level of leap forward in performance
- With the right design & architecture choice, PCIE SSD could reach Million IOPS, both 4K random read and 4K random write

The Hybrid Approach

+

Modularize and Everything on a single card

HBA
Host bus adaptor

SATA SDD
Drive

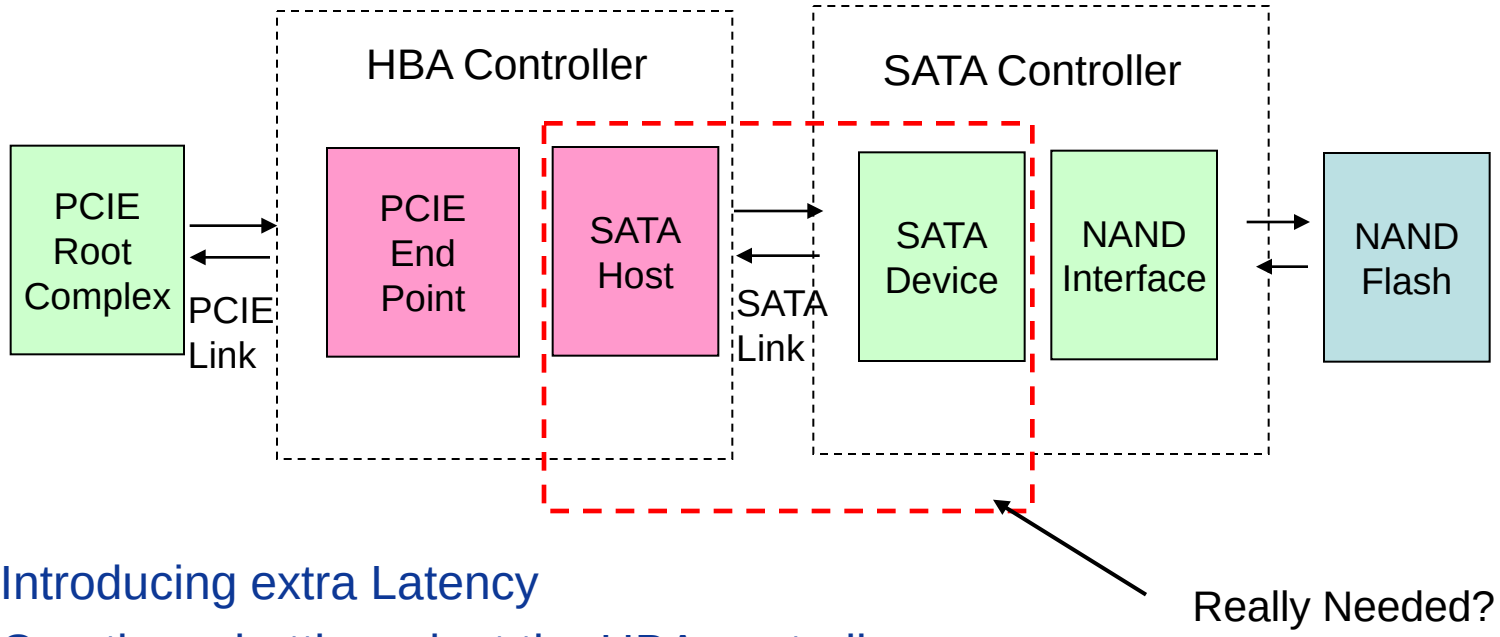
PCIE SDD
Drive

Hybrid Approach - Pros

- HBA has been around for decades with mature software stack for RAID, storage management etc
- Off shelve, readily available chipset and components
- SATA SSD firmware development independent of HBA and driver
- Plug & Play, Software compatibility

Hybrid Approach - Cons

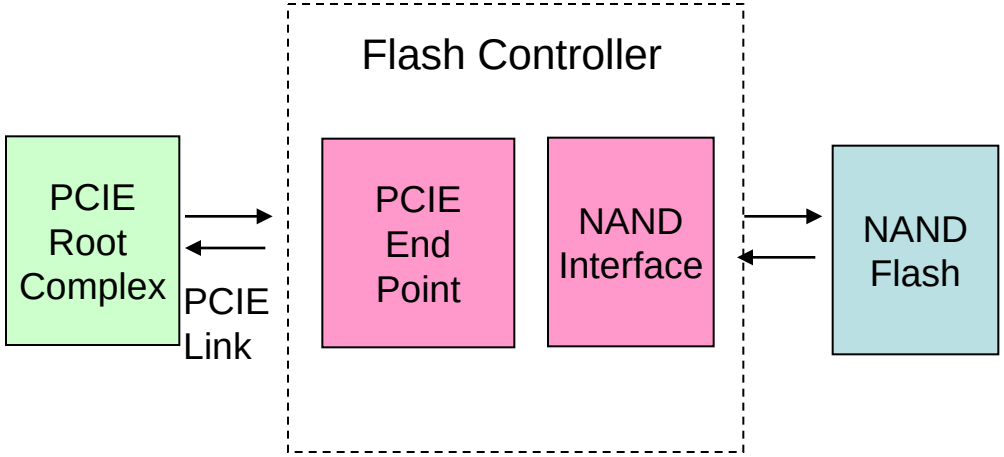
Unnecessary Protocol Conversions



- Introducing extra Latency
- Creating a bottleneck at the HBA controller
- More silicon and board real estate
- More power consumption

Direct Approach

Much simplified data path



- The key is the design of a high performance PCIE based flash controller
- Different ways of doing it: FPGA vs ASIC

FPGA doesn't scale

- Take PCIE x8 as example
  - At Gen1, PCIE offers 1.6GB/s bandwidth
  - As rule of thumb, the controller internal switching bandwidth should double that ~ 3.2GB/s
  - Considering 256 bit internal data path, internal clock needs to run at 100 Mhz
  - Already stressing the state of art FPGA
- At Gen2, needs to go up to 200 Mhz
- At Gen3 -- **400Mhz,** FPGA just won't make it

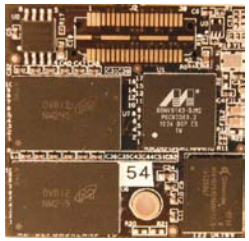
FPGA is expensive and has hidden cost

- It is very challenging to put a CPU inside FPGA to run faster than 200Mhz, while 600Mhz – 1Ghz ARM CPU is common practice, which implies:
  - In an FPGA controller system, host CPU needs to step in to take a majority of the work, put extra burden on the host CPU
  - Complicating the design for backup and recovery from power loss
  - Data recovering is very time consuming

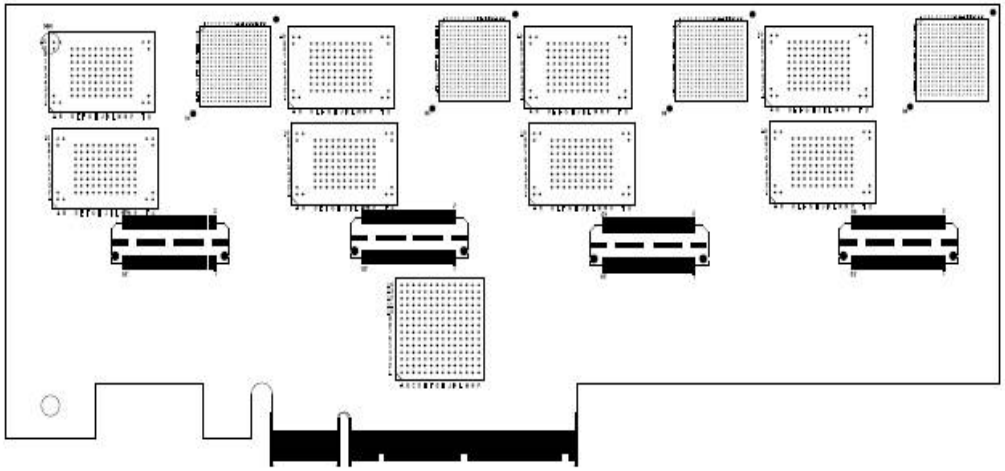
ASIC & Modularized

- A monolithic huge ASIC controller may not be the best choice for system integration
  - The routing and interconnecting difficulty
  - 32 NAND FLASH channel ~ 540 signals need to be routed
  - Cost could be high
  - Reliability/Thermal could be a concern
- Modularized design offers better scalability, flexible system interconnect and desired scalable performance for different applications.

Modularized design



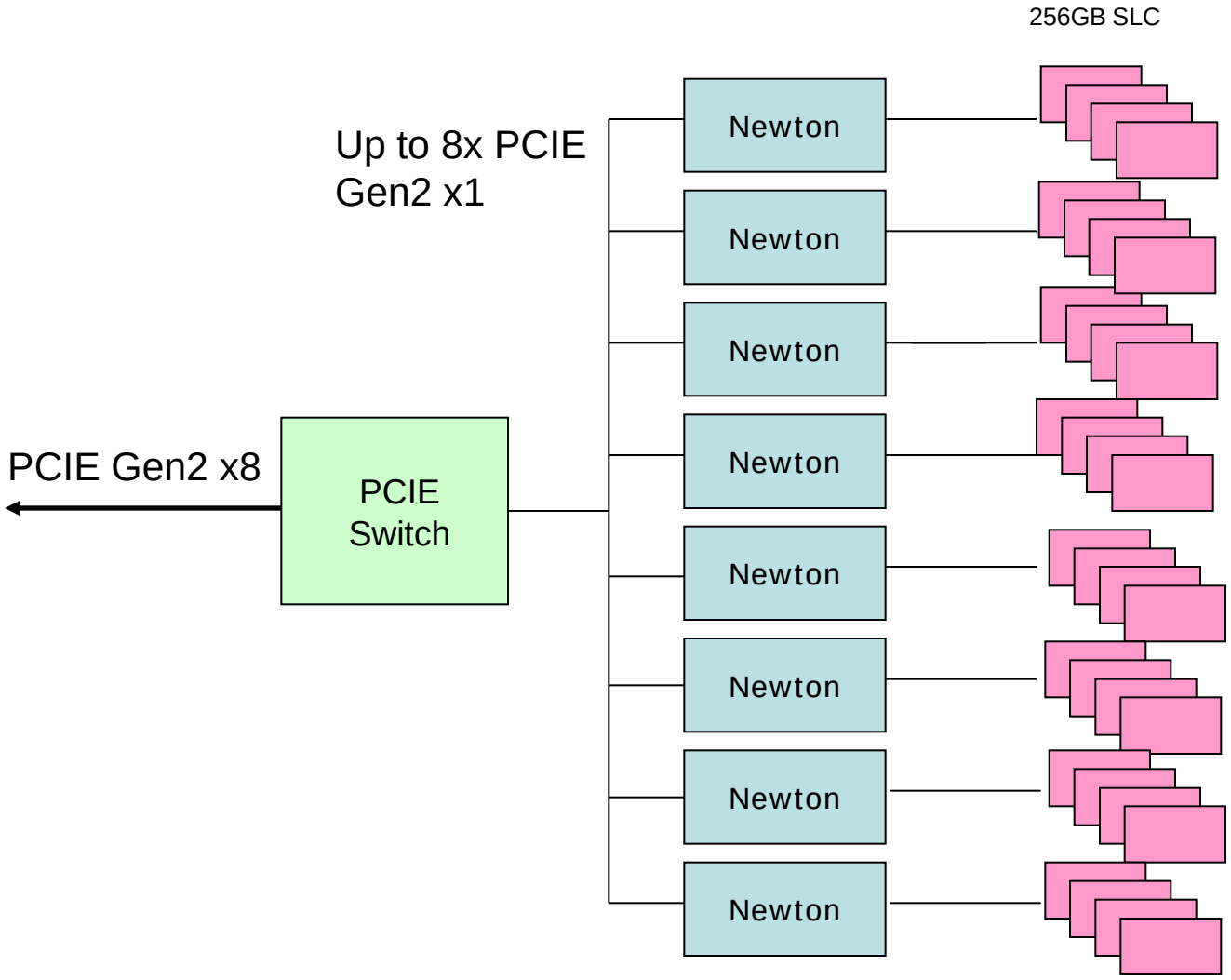
Integrate x8



Module

Low Profile, half length x8 PCIE SSD Card

System block diagram



Performance & Scalability

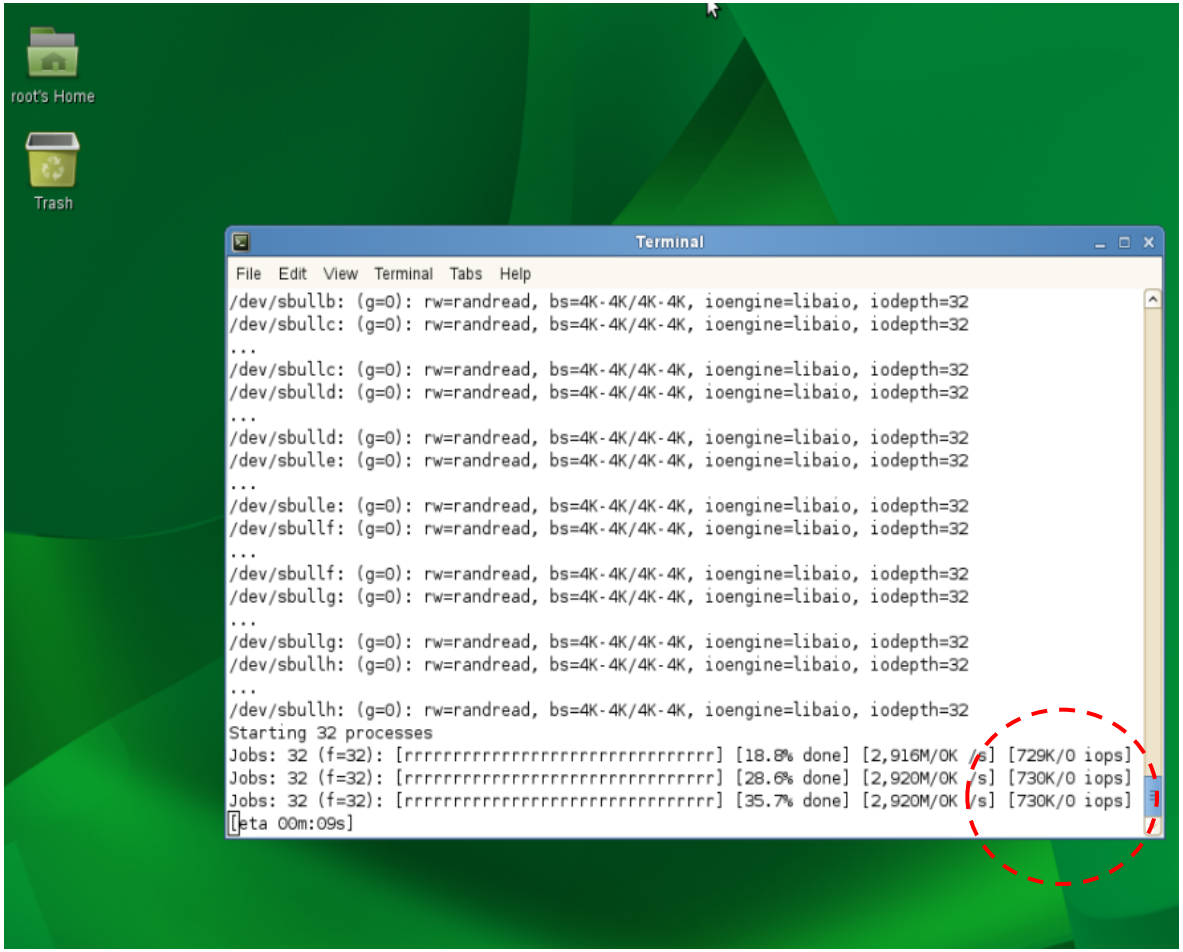
4K Random Read	
1 Module	93 K
2 Modules	186 K
4 Modules	371 K
8 Modules	730 K
16 Modules	**1.4 Million**

4K Random Write (clean drive)	
1 Module	70 K
2 Modules	140 K
4 Modules	277 K
8 Modules	530 K
16 Modules	**1.04 Million**

Inside a Shuttle PC

The low profile card could be readily inserted into a small Shuttle PC

13” x 8.5” x 8”



Screen shot of running fio under Linux on the shuttle PC
730K IOPS

Implications

- For the past few decades, IO performance increase has always been lagging behind CPU
- With the right storage architecture and storage device, any PC today could be turned into a powerful IO machine
- There are plenty of IO bandwidth and IOPS to explore
- It is up to the OS and application to fully leverage the power of the new generation of PCIE flash SSD

Q & A