NVMe Host Acceleration

Hiren Patel

IntelliProp Inc.

Abstract

- Benefits of using hardware automation.
  - Compare Linux driver vs. acceleration IP in hardware.
  - Acceleration will maximize performance of SSD.
  - Reduce CPU overload.
- Based on a white paper - available for download from IntelliProp.

Agenda

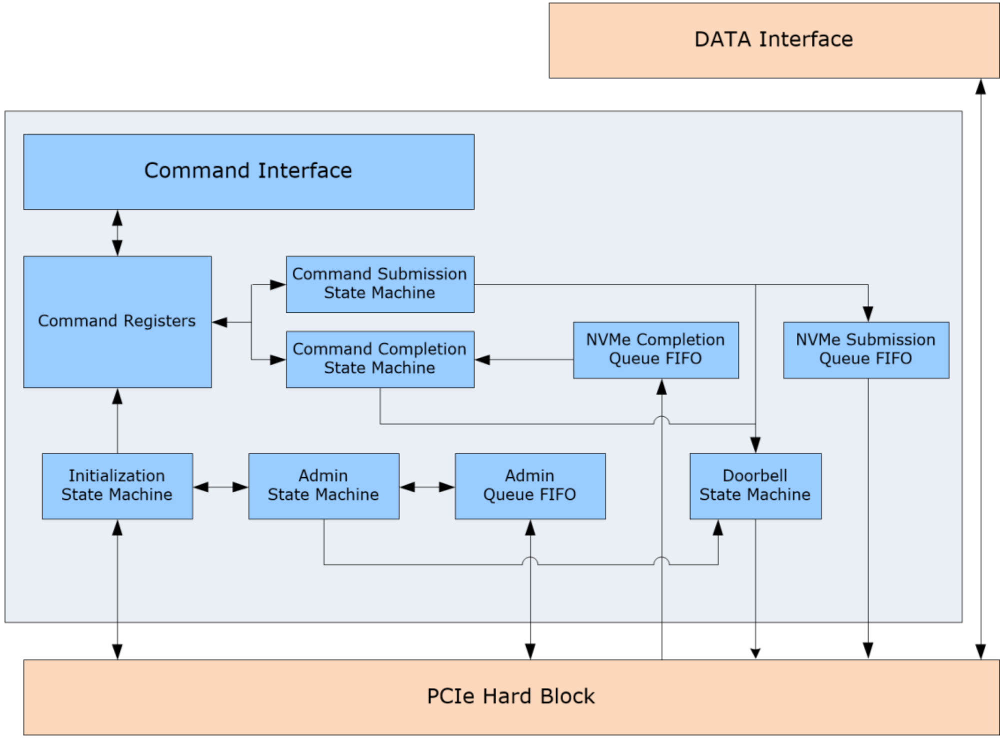
- NVMe Overview
- NVMe Accelerator IP
- Testing Strategy
- Measurements
- Analysis
- Summary/Conclusion
- Next Steps
- References

NVMe Overview

- Storage specification designed to access NVM over PCIe.
- Commands in the Submission queue.
- Completions in the Completion queue.
- Flow control from endpoint.
- Linux NVMe drivers available (off-the-shelf implementation of NVMe).

NVMe Accelerator IP

- Alternative or Complement to Linux NVMe driver.
- Hardware Logic, low latency and high bandwidth.
- Supports PCIe Gen1, 2, 3 and 4.
- Offload NVMe data transfers and burden from processor (ARM).



Testing Strategy

- Test Platform
  - Xilinx ZC706 FPGA development board (Cortex A9).
  - Opsero FPGA Drive board (FMC to M.2).
  - Uses AXI MM to the PCIe Hard block (Gen2).
  - Samsung 950 Pro M.2 NVMe SSD.
  - ARM processor (on Zynq) @ 667MHz clock and @ 800MHz clock.
  - Xilinx Petalinux Tools + NVMe Linux driver.

Testing Strategy - cont.

- Test Methodology (Driver)
  - NVMe SSD enumeration + EXT4 partition (/mnt/nvme0)
  - 8 parallel dd commands
  - 512 bytes to 128kB range of data transfer size (power of 2 increments)
  - 2^{20} I/O commands or 1M commands per data transfer size
  - Sequential access for read/write testing
  - 97% CPU Utilization during the above test and 0.0% with no commands
- Test Methodology (NVMe Host Accelerator)
  - Single submission and completion queue
  - 256 queue depth
  - Raw storage device
  - 512 Bytes to 128kB range of data transfer size (power of 2 increments)
  - 2^{20} I/O commands or 1M commands per data transfer size
  - Sequential access for read/write testing

Test Strategy - cont.

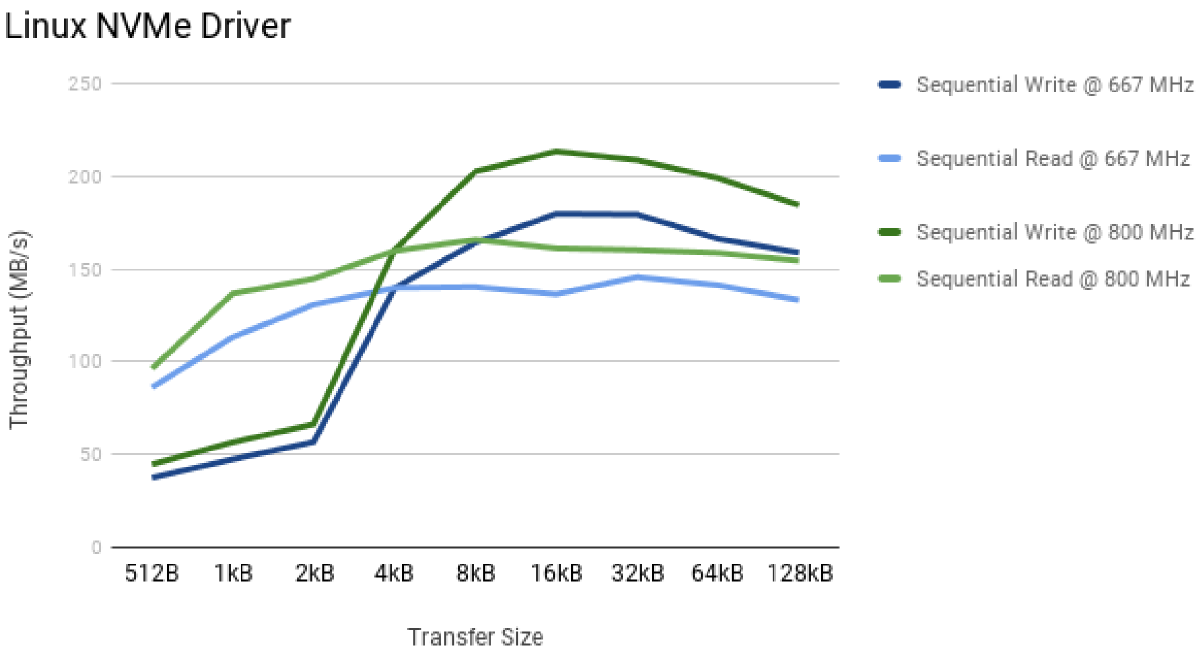
- Performance Calculations
  - Linux Driver

$$Throughput\ (MB/s) = \frac{(Block\ Size * Count)}{Time} * \frac{1}{1e^6}\frac{MB}{s}$$

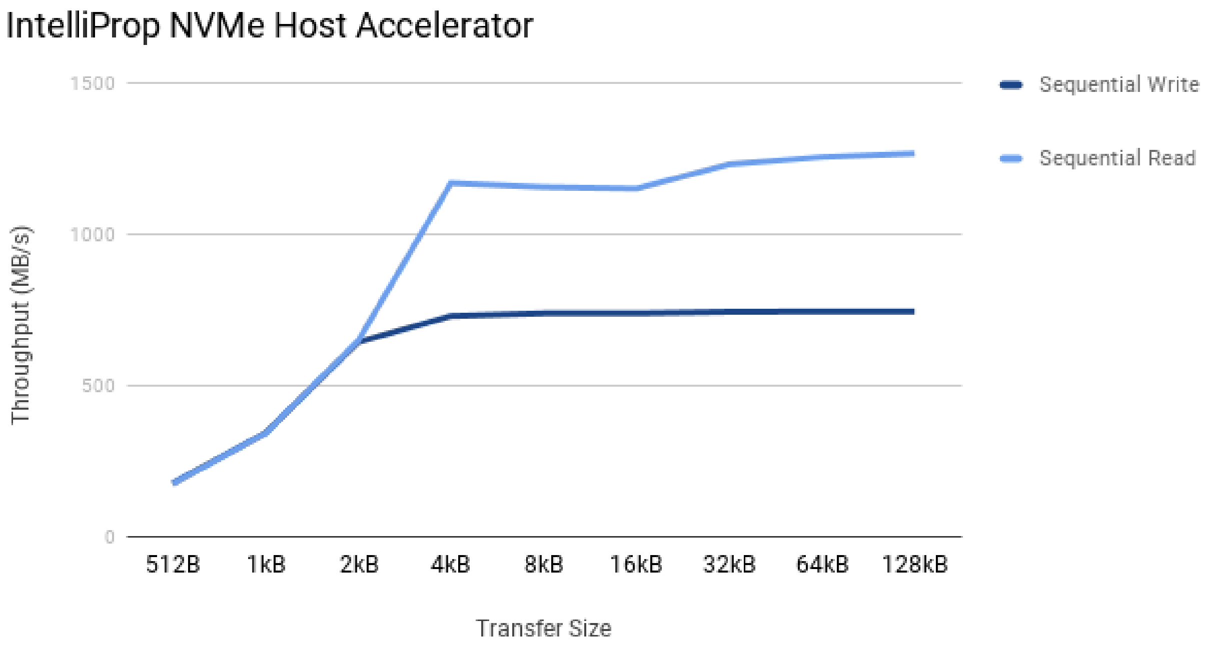
  - NVMe Host Accelerator @125MHz

$$Throughput\ (MB/s) = \frac{(Block\ Size * Count)}{Clocks} * \frac{1}{8ns}\left(\frac{1e^9\,ns}{1\ s}\right) * \frac{1}{1e^6}\frac{MB}{s}$$

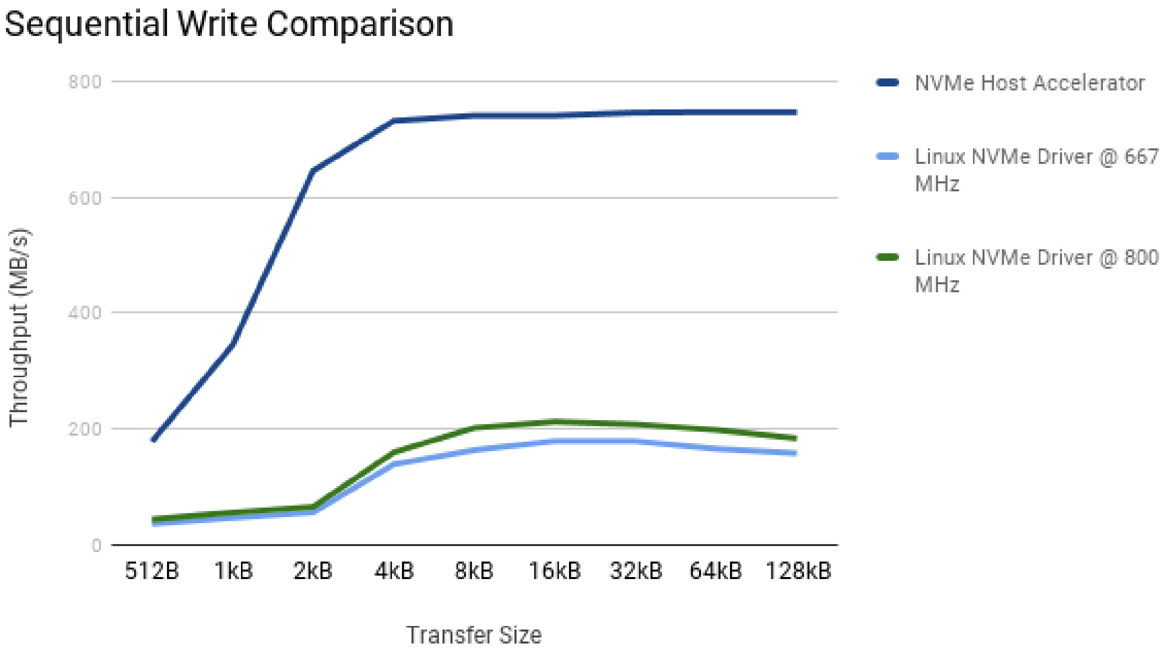
Measurements/Graphs



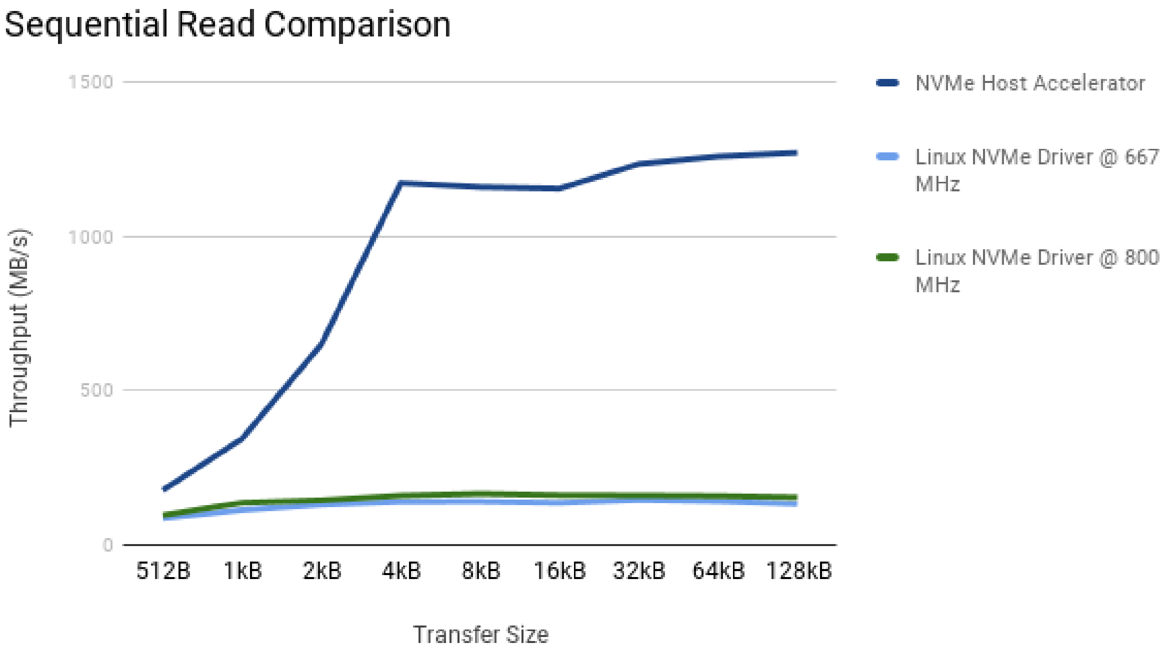
Measurements/Graphs



Measurements/Graphs



Measurements/Graphs



Analysis

Performance in MB/s

Size	IP Write	IP Read	Driver Write	Driver Read	Analysis
512B	179	177	45	96	The accelerator is already faster.
4kB	732	1172	140	160	Performance is 5x better on writes and 7x on reads. Driver is at 160 MB/s.
16kB	741	1154	213	161	Drive limitations show up (flat on writes and reads) on the Host Acceleration IP. The IP is still faster than the driver.
64kB	747	1258	199	159	Both Driver and Host Acceleration IP remain flat.

- ✓ Update: 960 Samsung drives (Gen3) are showing closer to 2000MB/s writes with some newer OEM drives reaching 2400MB/s. A new updated white paper will be available soon.

Summary/Conclusion

- Acceleration or offloading the CPU helps with performance. The graphs show a limitation unrelated to the drive for the Linux driver. The NVMe Host Accelerator IP shows some read/write (max performance) limitations on the drive.

- Why is the driver slow? Some thoughts are: the ARM and NVMe driver have to fetch instructions from DDR, build the command in memory and ring the doorbell. The IntelliProp NVMe Host Accelerator IP performs all of these functions in hardware - offloading the work from the driver and ARM.

- For the ARM and driver to achieve the same performance as the NVMe Host Accelerator IP, it would need to run at

6.4 GHz!

Next Steps

- Repeat the whitepaper testing on the Zynq Ultrascale+ board with PCIe Gen3 and ARM Cortex A53.
- Test with the newer M.2 drives (960 Pro) and new U.2 drives.
- Test with the newly developed Linux driver from IntelliProp to use on the Host Accelerator IP.

References/More Information

- References
  - IntelliProp NVMe Host Accelerator IP White Paper by Collin Butler @ IntelliProp
  - Xilinx Zynq documentation, pictures, product briefs and user guides (UG954, UG477, dh0016)
  - Opsero FPGA Drive FMC to M.2 board: http://fpgadrive.com/
- More Information:
  - IntelliProp Inc.
  - info@intelliprop.com
  - www.intelliprop.com