

NVMe Over Fabrics

Performance and emerging NVMs

Zvonimir Bandic
Storage Architecture, HGST, a Western Digital Co.

August 11, 2015



Credits and acknowledgements

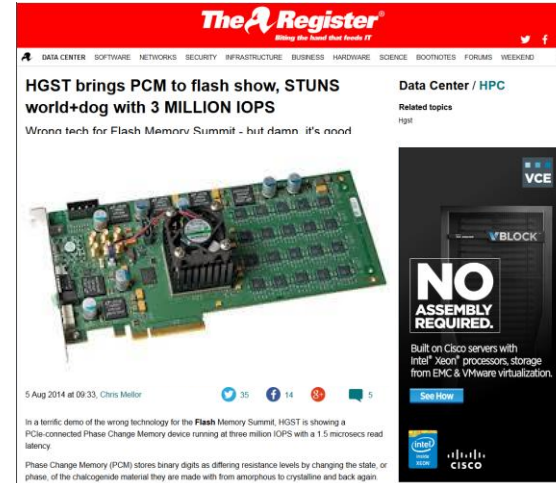
- Christoph Hellwig
- Qingbo Wang
- Paul Suhler
- Zvonimir Bandic

Outline

- NVMe over fabrics Architecture
 - Host driver architecture
 - Controller driver architecture
- NVMe over fabrics performance Outlook:
 - FIO measurements on sub-10 us latency
- Emerging NVMs Application Synergies
 - As emerging NVM SSDs break into sub-10us local latency category, there is tremendous incentive to ensure that network latency is comparable/smaller

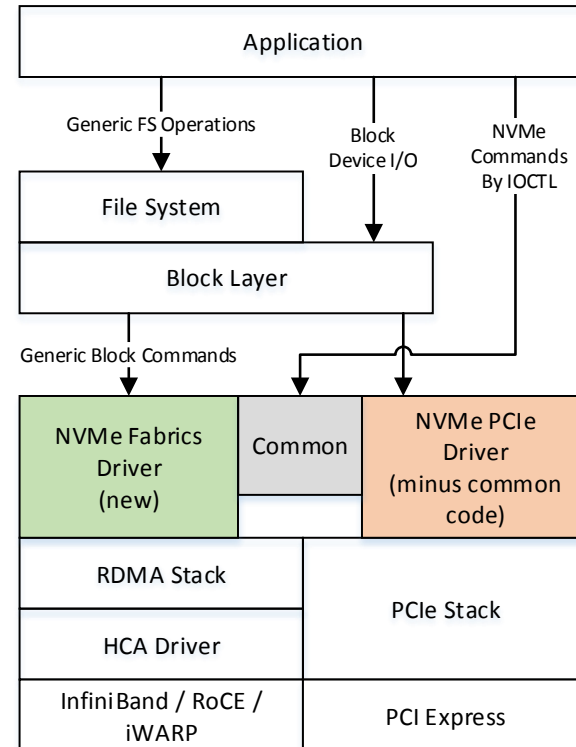
Impact of emerging NVM based SSDs

- NVMe devices have low latency using NAND flash devices
- Emerging NVM based SSD devices will have even lower latency:
 - Enabled by improved read latency when compared to NAND
- Low SSD latency puts pressure on delivering low-latency networking architecture:
 - NVMe over fabrics!



Host Software Architecture

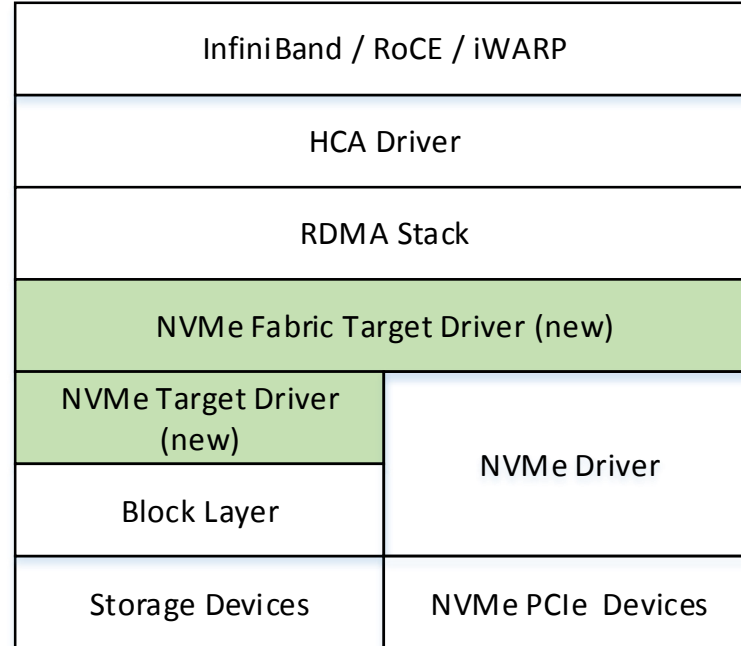
- Common code was extracted from NVMe PCIe driver.
- NVMe Fabrics driver is new.
- Other driver, stack, and FS modules are unmodified.



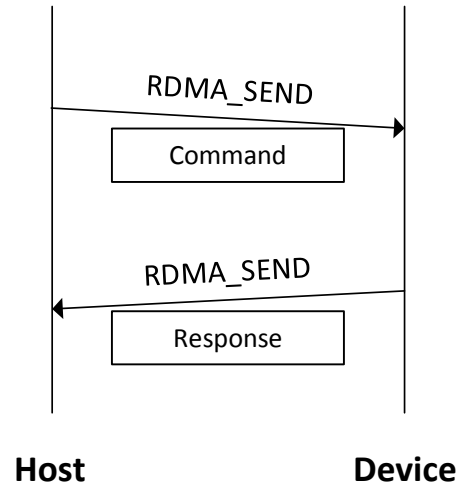
NVMe over Fabrics Controller Architecture

- Target devices include:
 - RAM disk
 - NVMe device
 - Other NVM SATA/SAS devices

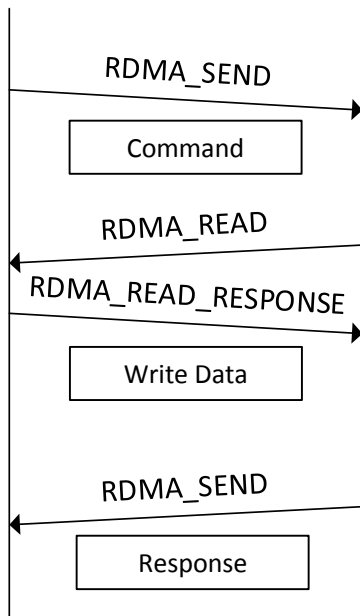
Initial transport is InfiniBand



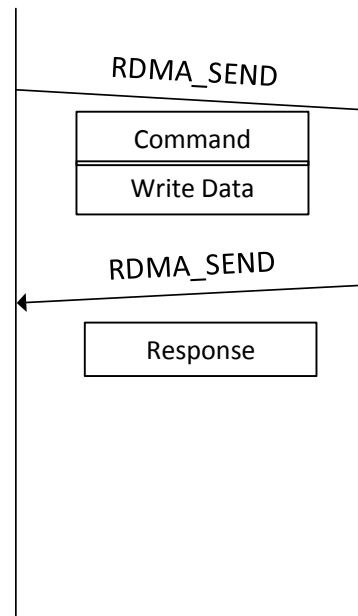
Non-Data Commands



Data-Out Commands



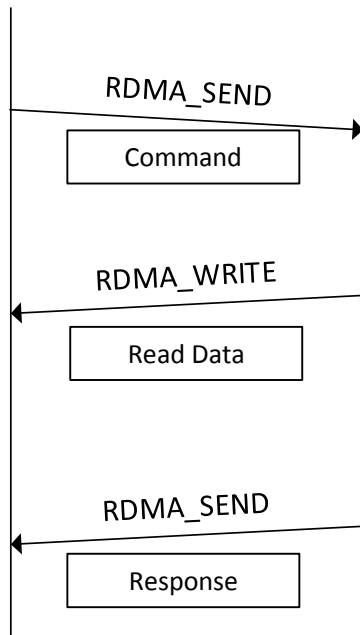
Separate data transfer phase



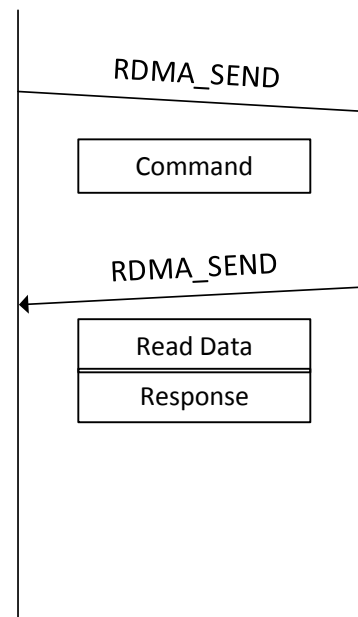
In-capsule data

In-capsule data saves one round trip when data size is small

Data-In Commands



Separate data transfer phase



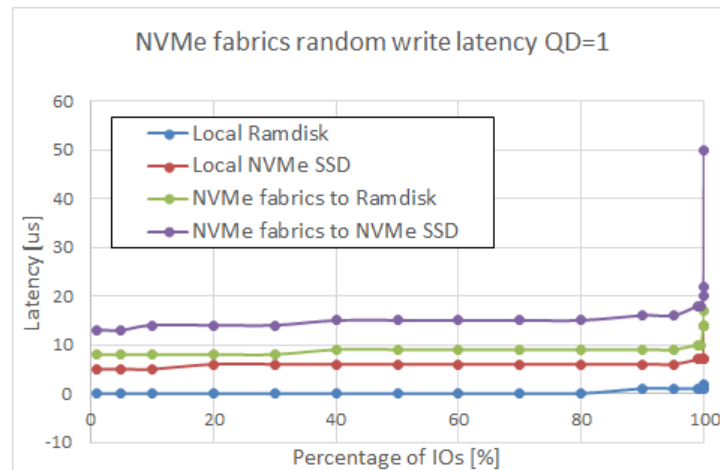
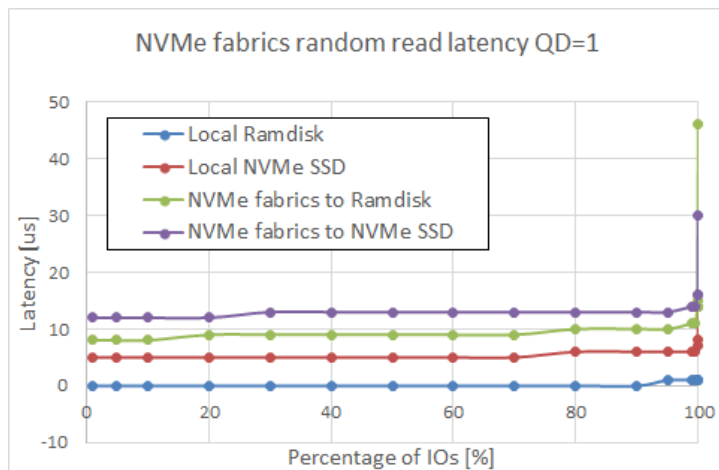
In-capsule data

In-capsule data saves one round trip when data size is small

Test Configuration

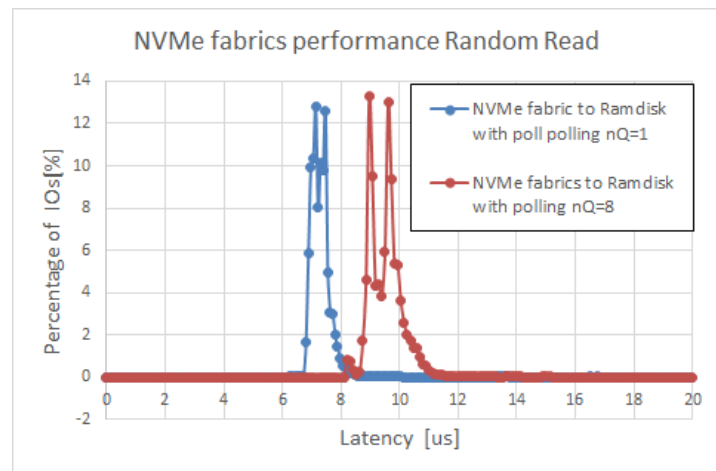
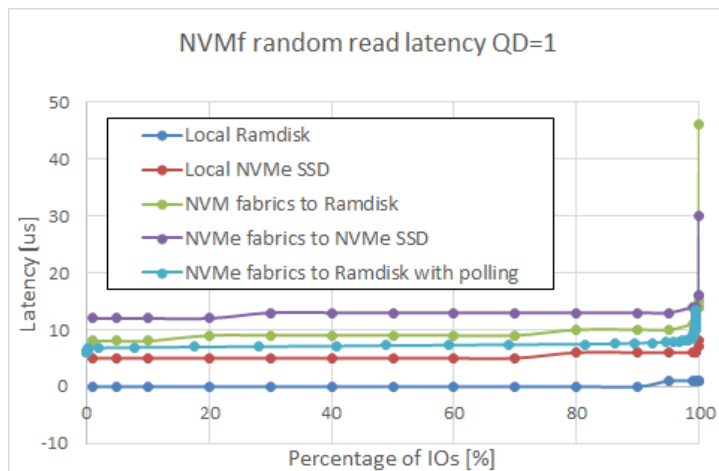
- Subsystem:
 - ASUS Z97-Deluxe, i7-4790, 3.6 GHz
 - Ubuntu 14.04.02, 4.1.0 kernel
 - Mellanox ConnectX-3 (InfiniBand)
 - HGST Research prototype NVMe card
- Host:
 - ASUS X99-Rampage V Extreme, i7-5930K, 3.5 GHz
 - DDR4-3200
 - Mellanox ConnectX-3

Performance measurements (InfiniBand)



- 13 us latency at QD=1 for random reads
 - Sub-10 us network contribution
- Further improvements:
 - Polling library should remove 3 us from the local device
 - 2-3 us additional improvement in network contribution should be possible

Performance measurements (with polling)



- Added polling on the initiator side:
 - Linux side Ramdisk target driver synchronous
- Latency (end-to-end) is 8 us:
 - Network latency contribution is <7 us

Conclusions

- Emerging NVM based SSD are likely going to have 10x improvement in latency compared to NAND based SSDs
- For cluster-based compute and storage, this brings new requirements on network latency:
 - Network becomes new bottleneck
- Networking protocols have to be redesigned to allow for significant latency improvement:
 - We have currently achieved sub-7 us network latency contribution
- Work identified within NVMe over fabrics that minimizes network contribution to latency



Architected for Performance