NVMe™ Performance
Local vs. Remote

Oscar Pinto, Ming Lin, Gunna Marripudi

Samsung Semiconductor Inc.

Legal Disclaimer

This presentation is intended to provide information concerning SSD and memory industry. We do our best to make sure that information presented is accurate and fully up-to-date. However, the presentation may be subject to technical inaccuracies, information that is not up-to-date or typographical errors. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of information provided on this presentation.

The information in this presentation or accompanying oral statements may include forward-looking statements. These forward-looking statements include all matters that are not historical facts, statements regarding the Samsung Electronics' intentions, beliefs or current expectations concerning, among other things, market prospects, growth, strategies, and the industry in which Samsung operates. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements contained in this presentation or in the accompanying oral statements. In addition, even if the information contained herein or the oral statements are shown to be accurate, those developments may not be indicative developments in future periods.
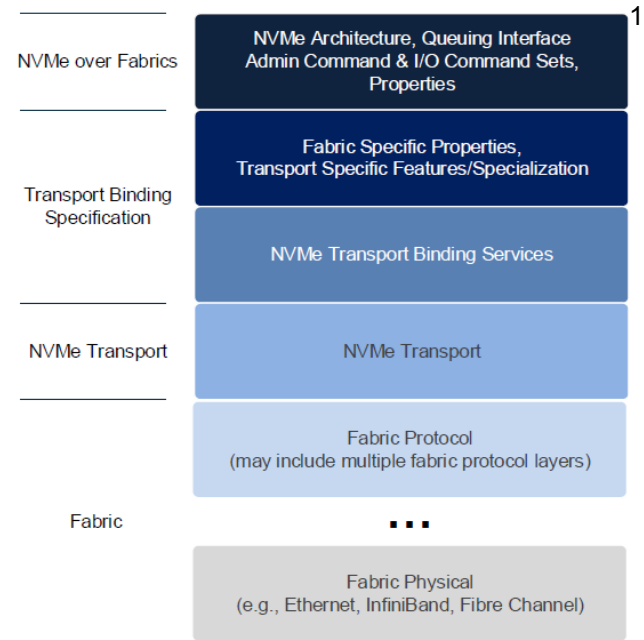
Agenda

- NVM Express® over Fabrics Overview
- Test Configuration
- Performance Comparison

NVM Express® over Fabrics Overview

- Maintains consistency with base NVMe™ definition but for fabrics support[*]
- Support for multiple transport types
- Exposes NVMe™ parallelism to host
- Performance closer to local NVMe drives

NVMe over Fabrics	NVMe Architecture, Queuing Interface Admin Command & I/O Command Sets, Properties
Transport Binding Specification	Fabric Specific Properties, Transport Specific Features/Specialization
	NVMe Transport Binding Services
NVMe Transport	NVMe Transport
	Fabric Protocol (may include multiple fabric protocol layers)
Fabric	…
	Fabric Physical (e.g., Ethernet, InfiniBand, Fibre Channel)

[*] http://www.nvmexpress.org/wp-content/uploads/NVMe_over_Fabrics_1_0_Gold_20160605.pdf
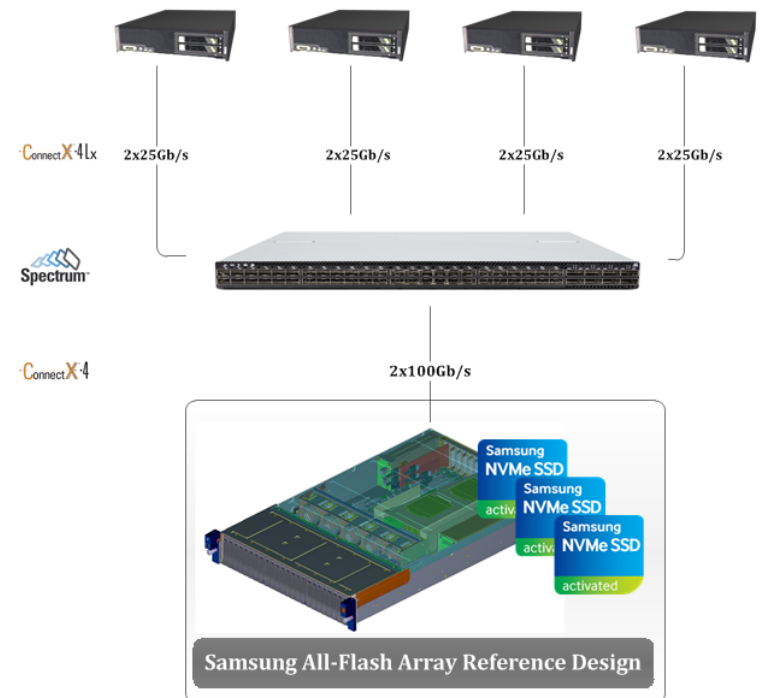
NVM Express® over Fabrics Linux Stack

- Kernel stack
- Leverages OFED RDMA stack
- NVMe™ command processing in
  - Initiator: *nvme_core*
  - Target: *nvmet_core*
- NVMe™ transport binding in
 - Initiator: *nvme_rdma*
  - Target: *nvmet_rdma*

Linux Open Source Stack
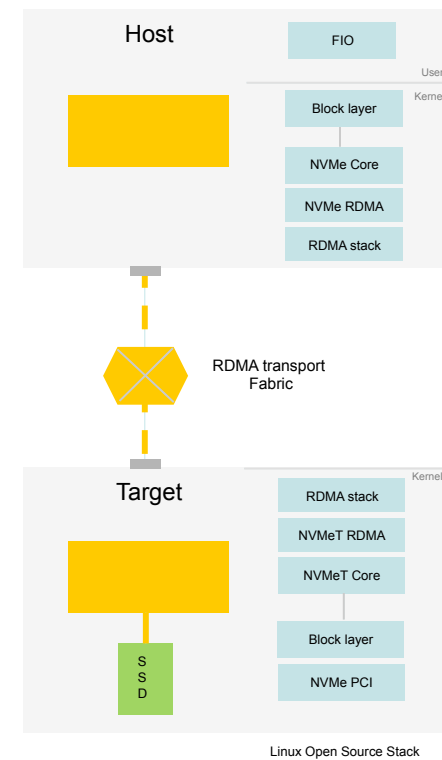
Test Configuration

- ■ Configuration
  - • 1x NVMe-oF™ target
    - – 24x Samsung NVMe™ Datacenter 2.5" SSDs
    - – 2x 100Gb/s ConnectX®-4 EN
  - • 4x initiator hosts
    - – 2x25Gb/s each

  - • Ubuntu 14.04.4 LTS Linux 4.7.0-rc2 kernel
  - • Open Source NVMe-oF™ kernel drivers

Samsung All-Flash Array Reference Design

Latency Comparison

Latency (micro seconds)
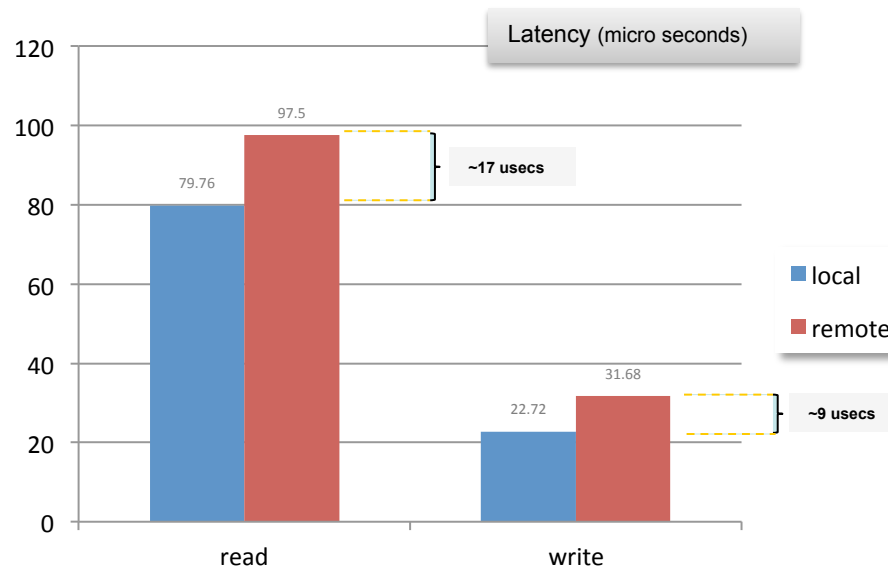
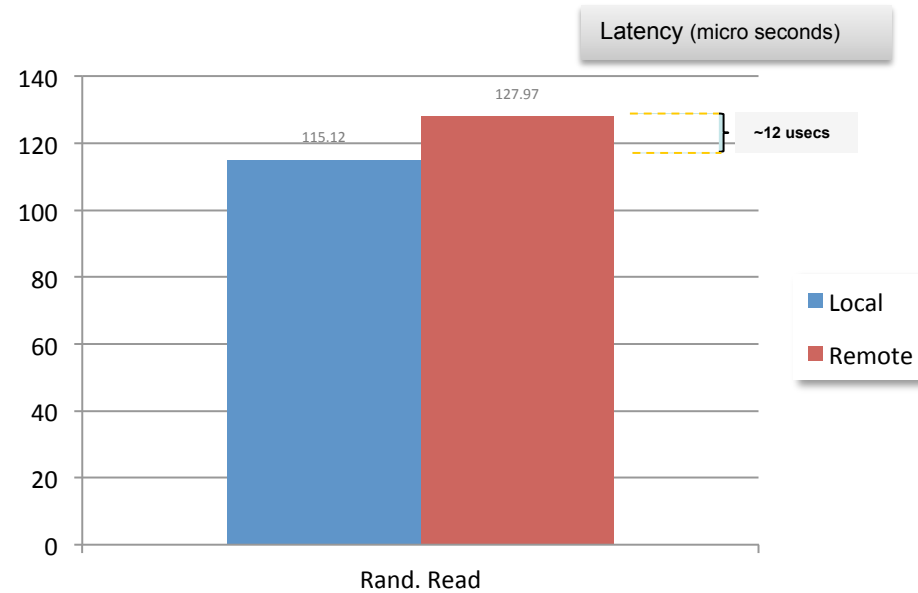**Random IO at QD1, 1 job**

- Round-trip delta: Reads ~17usecs; Writes ~9usecs

Latency Comparison - Loaded

Random IO at QD16, 2 jobs

- Performance delta: ~12 usecs

Latency Comparison - Scale
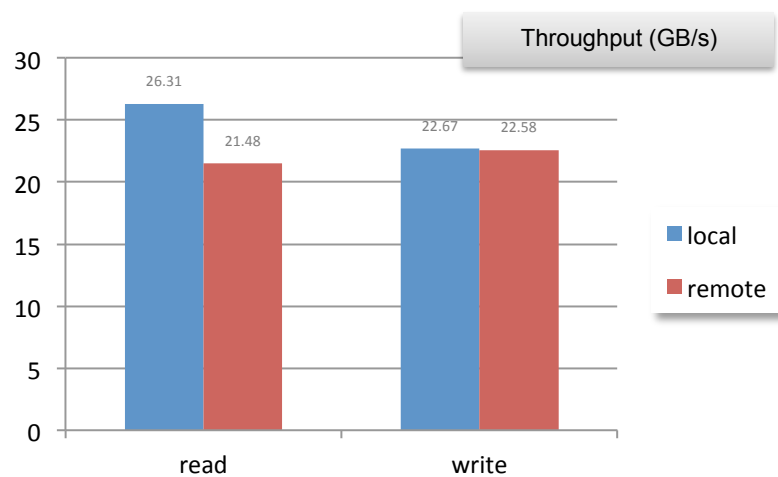

Latency (micro seconds)

	1	4	8	16	20	24
Rand. Read	127.97	127.97	131.88	146.62	156.37	179.94
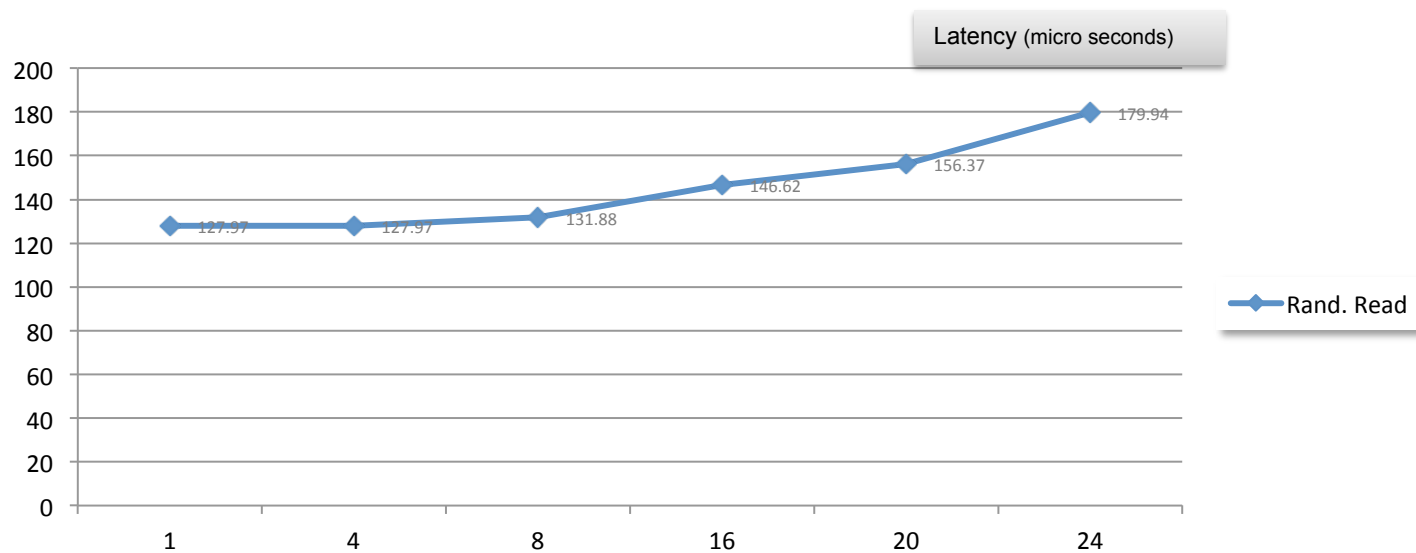
Random IO QD16

- NVMe-oF™ latency scales linearly from 1 to 24 drives

Performance (24 SSDs)

IOPS

	read	write
local	4,884	886
remote	4,304	871

Throughput (GB/s)

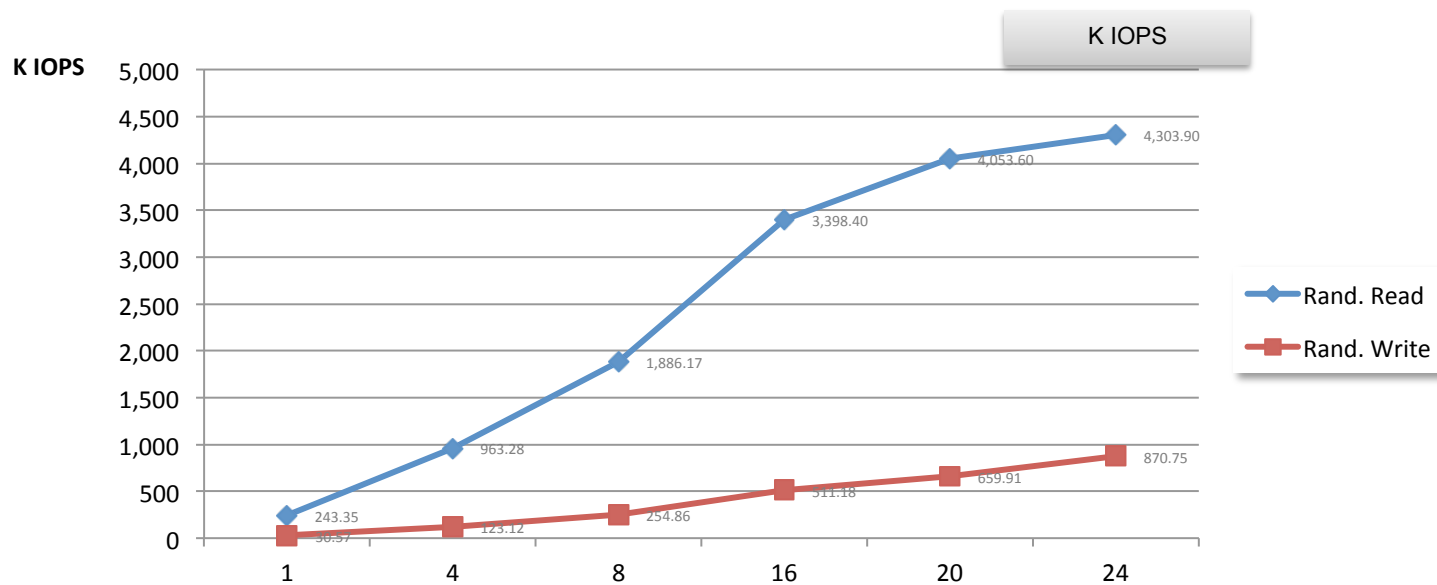|	read	write
local	26.31	22.67
remote	21.48	22.58

- High aggregate NVMe-oF™ performance: 4.3M IOPS & 21.5GB/s throughput
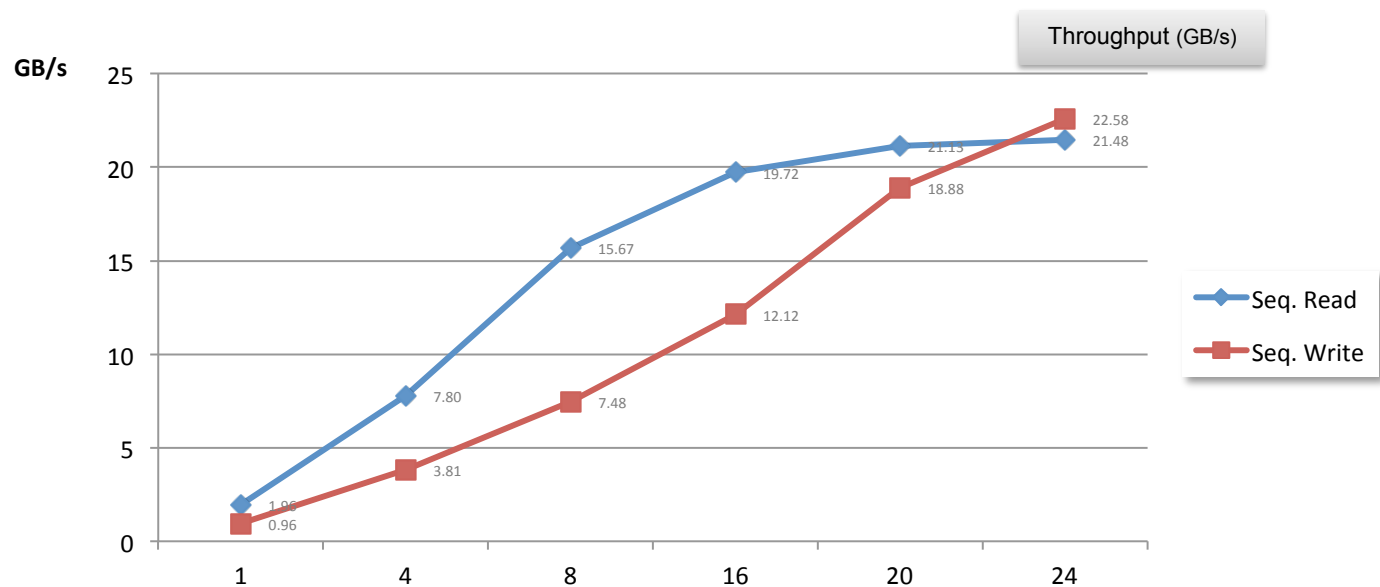- Further optimizations needed for performance to scale

IOPS Scaling

Random IO at QD16, 2 jobs per drive

- NVMe-oF™ stack scales linearly for Random IO

Throughput Scaling

Throughput (GB/s)

GB/s

Legend: Seq. Read, Seq. Write

Data points (Seq. Read, blue): 1.96, 7.80, 15.67, 19.72, 21.13, 21.48
Data points (Seq. Write, red): 0.96, 3.81, 7.48, 12.12, 18.88, 22.58

X-axis: 1, 4, 8, 16, 20, 24

Sequential IO at QD32, 1 job per drive

- NVMe-oF™ stack scales linearly for Sequential IO

Summary: NVMe™ Local vs. Remote

Performance Delta		1-drive	24-drive
Latency (QD1)	**Read**	~17usecs	-
	Write	~9usecs	-
IOPS	**Read**	10%	12%
	Write	On par	2%
Throughput	**Read**	On par	18%
	Write	On par	On par

Call for Action

- Further performance analysis & tuning of Linux Open Source stack

- Transport binding optimizations

- Feature enhancements
  - Faster failure recover scenarios
  - Reservations support

Thanks

Test Methodology

- Host – Target Setup
  - Ubuntu 14.04.4 LTS Linux 4.7.0-rc2 kernel
  - Each host mapped with 2 subsystems each with 3 SSDs
  - Each subsystem mapped to 1x 25Gb/s NIC
- Benchmark tool
  - fio 2.6-20-g2caf
  - ioengine=libaio
  - Random IOPS: 4k, iodepth=16, numjobs=2
  - Sequential IO: 128k, iodepth=32, numjobs=1
  - Latency: random IO 4k, iodepth=1, numjobs=1