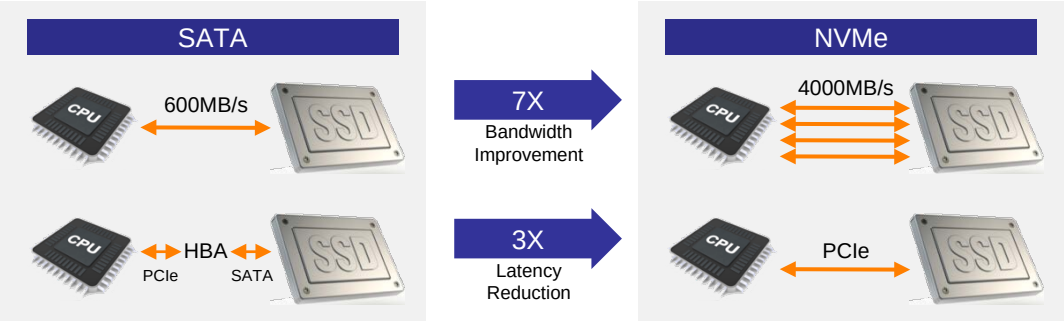
How NVMe can support Remote Access via RDMA

NVMe and RDMA – A Perfect Match?

Yaniv Romem, CTO, Excelero

NVMe Drives and Interface

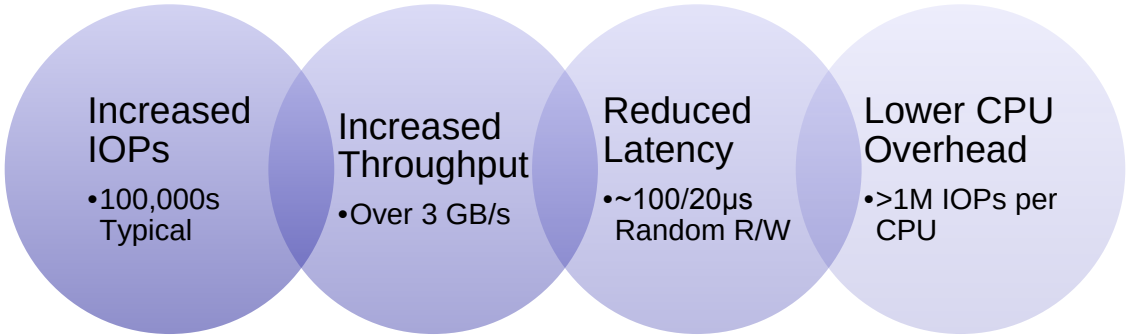
- The NVMe interface at the hardware level yields an immediate 7X BW improvement and a greater than 3X reduction in latency vs. SATA (AHCI)



…and *10X better random performance!!*

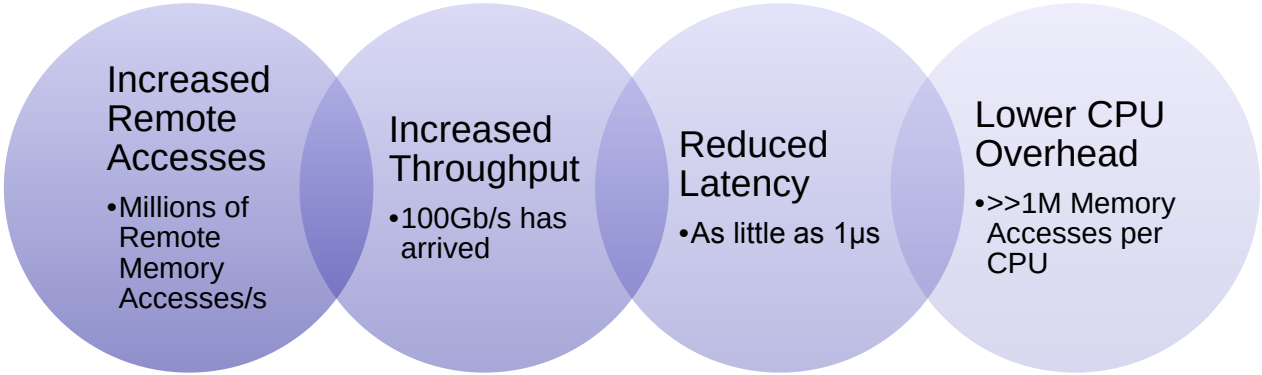
Why NVMe?

- NVMe (NVM express)
  - A streamlined protocol for accessing NonVolatile memories via PCIe
  - Justification = Take advantage of:



RDMA

- Remote Direct Memory Access
  - Enables a network adapter to access remote memory directly
  - Justification = Utilize new memories remotely:



Similarities, Coincidence?

NVMe

RDMA

Both Work via Completion Queues

- Increased IOPs
  - 100,000s Typical
- Increased Throughput
  - Over 3 GB/s
- Reduced Latency
  - ~100/20µs Random R/W
- Lower CPU Overhead
  - >1M IOPs per CPU

- Increased Remote Accesses
  - Millions of Remote Memory Accesses/s
- Increased Throughput
  - 100Gb/s has arrived
- Reduced Latency
  - As little as 1µs
- Lower CPU Overhead
  - >>1M Memory Accesses per CPU

Fabric NVMe & RDMA

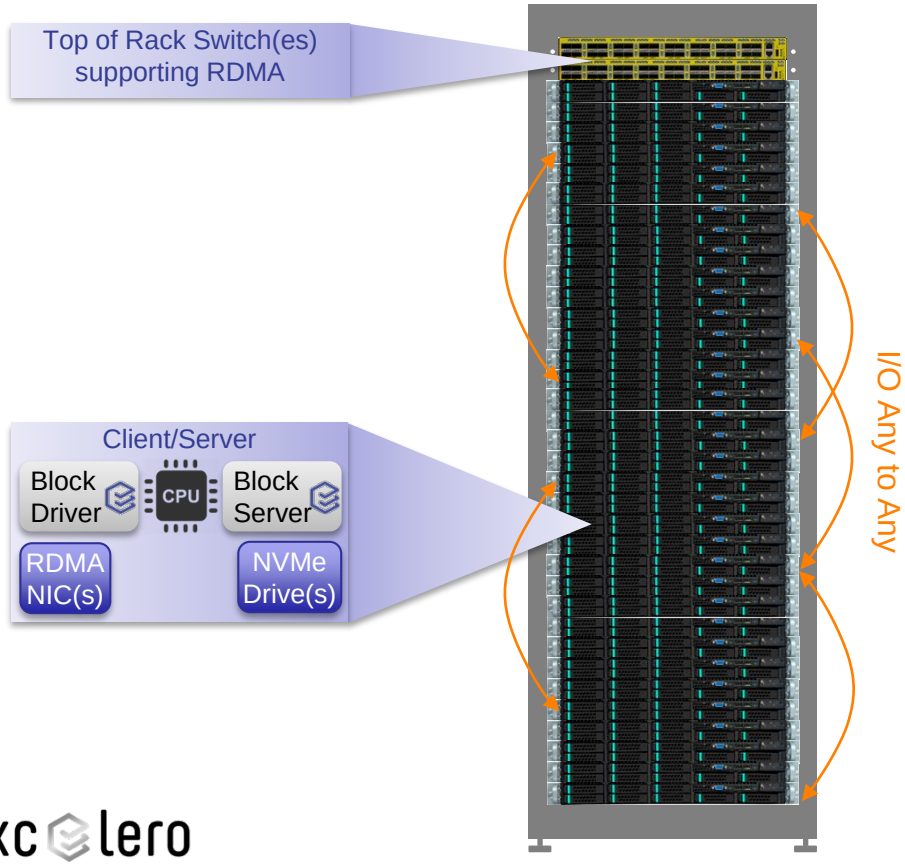
- NVMe-over-Fabrics
  - As a protocol, fabric agnostic, but:
    - Why would you want to add 200µs-1000µs latency to a 20us device???
- NVMe interface is layered above the networking
    - NVMe layer = control layer
    - RDMA = data layer
- Established RDMA Fabrics allow for choices in storage topology

Fabric NVMe Quirks/Challenges

- NVMe over Fabric dictates
  - 1-to-1 NVMe namespace over a link
  - Within the target, no implementation definitions
- Challenges
  - Volume definition & management
  - Device Aggregation, extension and redundancy
  - Centralized management
  - Storage service from a single target
    - CPU can become a bottleneck if it is doing control translation and / or data manipulation
  - Storage service from a cooperative group of targets
    - CPU can still be a bottleneck
    - Inter-communication between targets puts additional stress on network

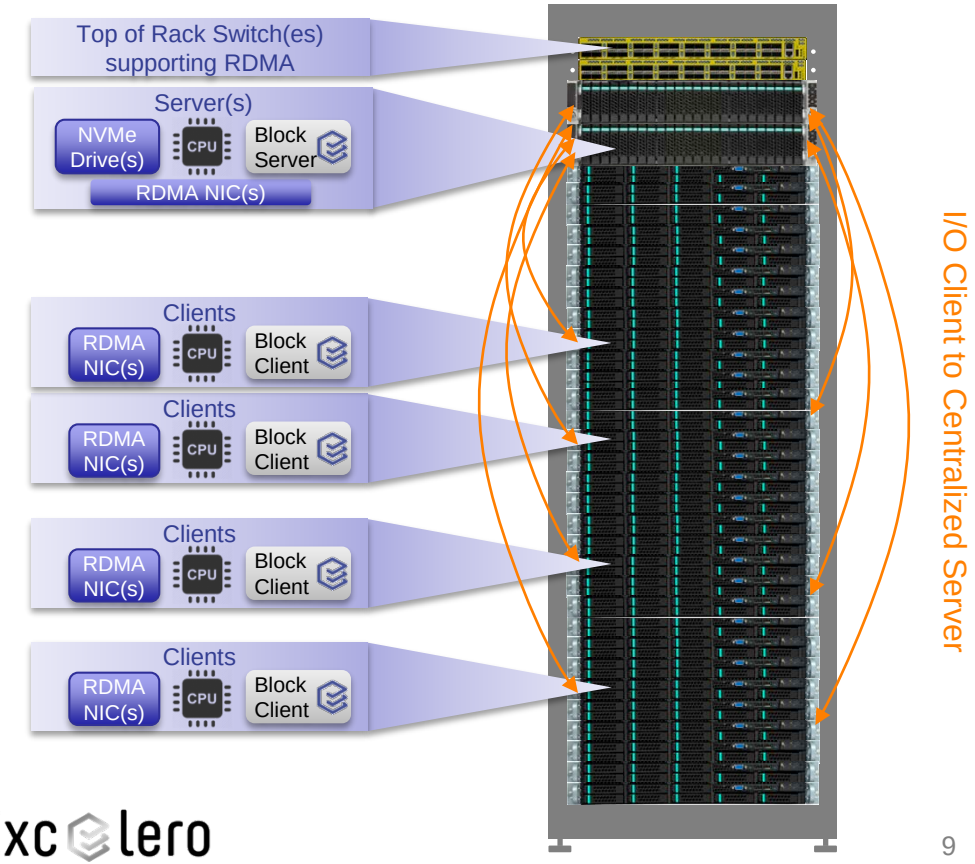
Fabric NVMe – RDMA Topology 1

- Converged Infrastructure
  - Each Server has 1 or more NVMe SSDs
  - All SSDs are treated as one storage pool
  - RDMA peer-to-peer makes all storage high performance



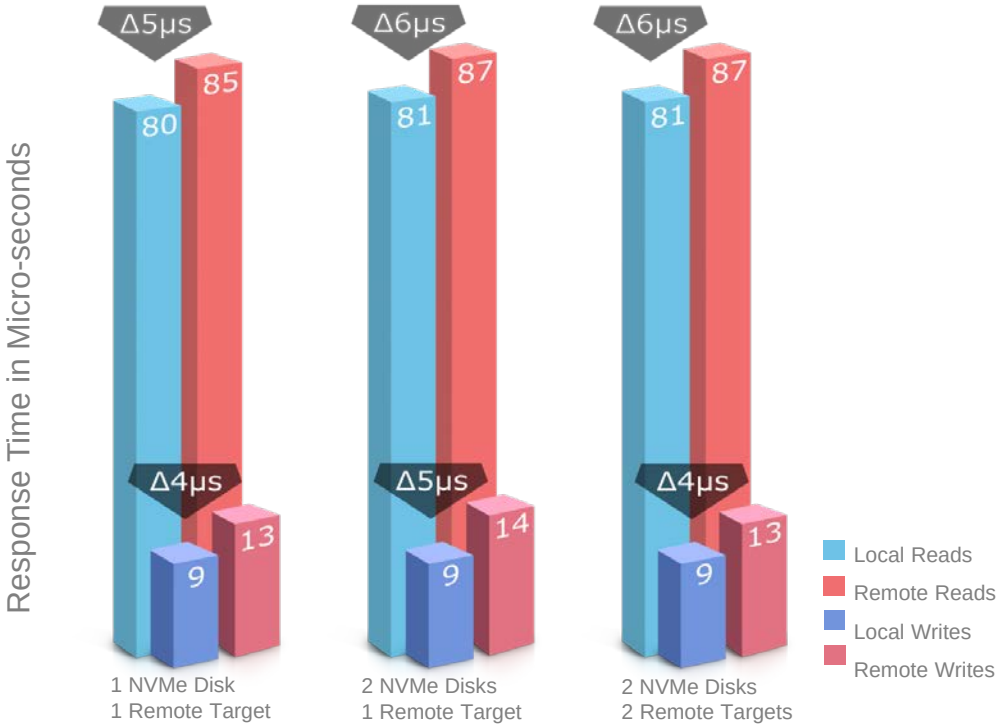
Fabric NVMe – RDMA Topology 2

- Centralized Model
  - Optional Redundancy
  - Serviceability
  - RDMA can make centralized, remote storage as fast as local

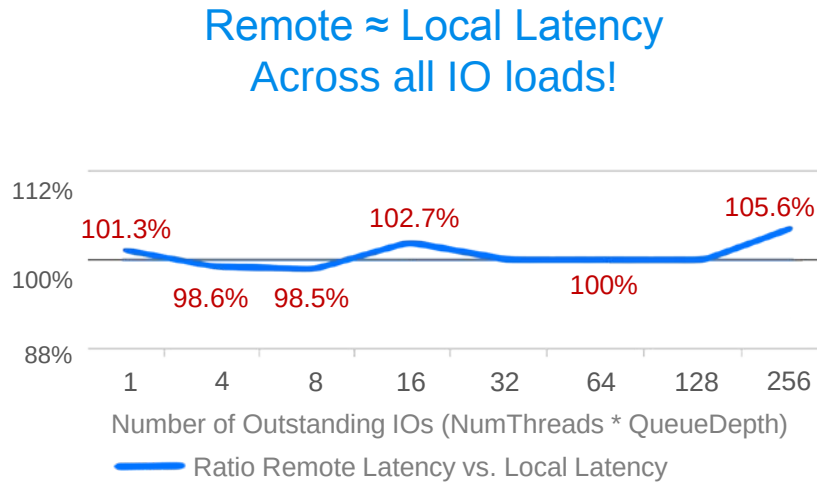


Putting It All Together…

Local vs. Remote Flash, 4K Reads and Writes, Single Queue Depth



Ratio of Latency, Remote vs. Local, 4K Reads in the 99.9th Percentile



Excelero

Thank You

Yaniv Romem, CTO, Excelero

info@excelero.com

http://www.excelero.com

“Free Your Flash!”