NVMe-oF: What Performance Can You Expect for Real Applications?

Andy Walls

IBM Fellow, CTO and Chief Architect Flash Storage - IBM

Flash Adoption Progression

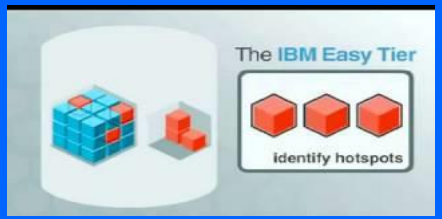
Flash Memory Summit

Began as Flash attached using Interfaces designed for HDDs.

(and Formfactors)

Serial Attached SCSI

Usually as tiers



External Storage Increased in Speed and Functionality

And All Flash Arrays became ubiquitous

NVMe: An interface and protocol designed for Solid State Media

nvm EXPRESS

Flash Memory Summit

Foretelling the NVMe Advantage

- **Flash System 840/900 designed with flash in mind**
- **Achieved the goal of NVMe with hardware data path**
- **No Firmware in data path**
- **Done with our own proprietary Interface**
- **Showed what is possible with flash optimized protocols**

Achieving the NVMe Advantage

Merged with 2 Controllers into a POWERFUL Full Function AFA

Uses IBM FlashCore Modules

2U24 NVMe usable capacity in a **2U** enclosure

Up to 378TB usable capacity in a **2U** enclosure

Up to 865TB effective capacity with in line hardware compression.

Up to 2PB effective capacity using Data Reduction Pools in the software.

Fully Redundant full function all flash array with world class performance

3X faster access to valuable analytics driving better business results with NVMe technology

A Great Example of End to End NVMe – The FS 9100

- Backend redesigned for NVMe
- Spectrum Virtualize Software Stack
- Astounding Capacity – 19.2TB FCMs today
- At Speed Hardware Compression
- Very Low latency, High IOPs and 37GB/sec Throughput
- Support for NVMeFC already.

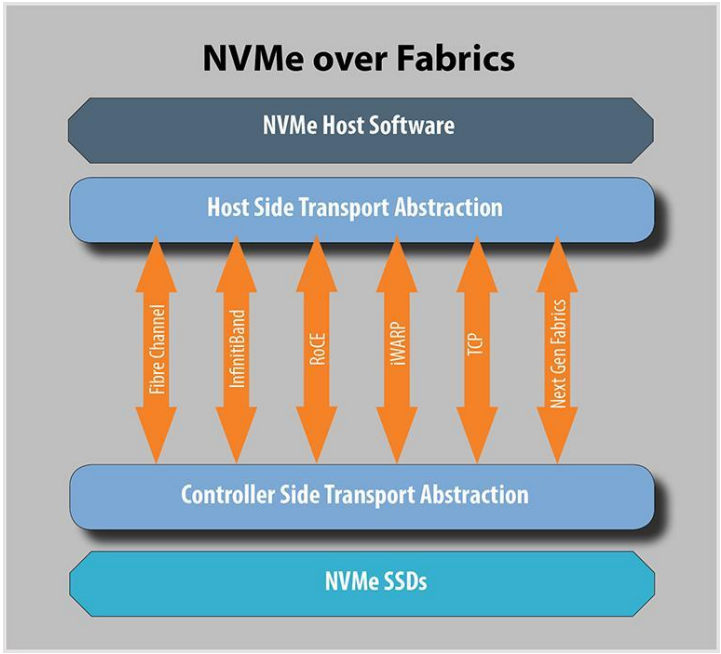
Redundant Controllers
Dual Socket Skylake
24 DIMMs. 3 HBAs.
Battery for persistent write Cache.

Up to 24 IBM Flash Core Modules

- **NVMe**
- **Hardware Compression**
- **Dual ported**
- **IBM Designed flash controllers**
- **4.8, 9.6 and 19.2TB capacity**

Flash Memory Summit

NVMe inside the storage or server is great, but what about the connection to external Storage?

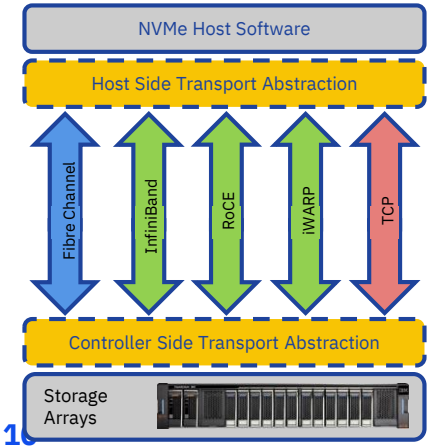


NVMeoF

- Disaggregated storage has advantages
- Reduces overhead in drivers and OS for network attached storage
- Full support just coming available – Multi Pathing drivers

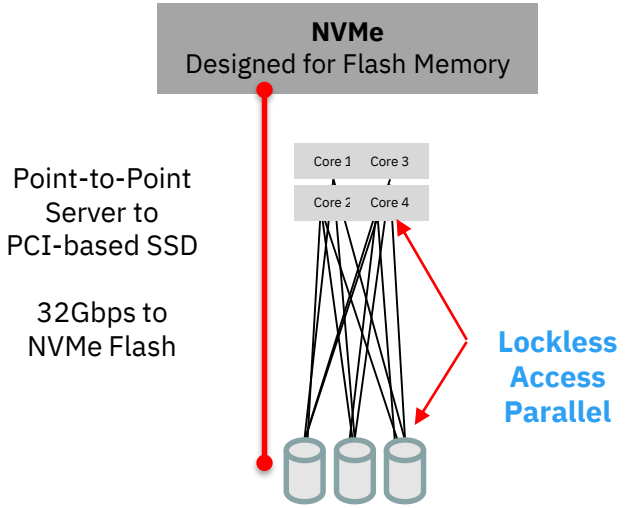
NVMeOF Attachment

- **NVMe is inside the server or storage array whereas NVMe over Fabrics is across the network**
  - Direct Attached SSD (PCIe based) doesn't scale
  - Networked storage is a must for large customers
  - Only 13% of storage capacity shipped is DAS (inside the server), 87% of the total storage capacity shipped is external storage
- **Three types of fabric transports for NVMe currently part of the standard**
  - *NVMe over Fabrics using the Fibre Channel Protocol (FCP)*
    - NVMe/FC or FC-NVMe
  - *NVMe over Fabrics using Remote Direct Memory Access (RDMA)*
    - InfiniBand, RoCE / iWARP
  - *NVMe over Fabrics using TCP*
    - NVMe/TCP
- **Goal of NVMe over Fabrics is to provide distance connectivity to NVMe devices with no more than 10 microseconds (µs) of additional latency over a native NVMe device inside**



Performance advantages

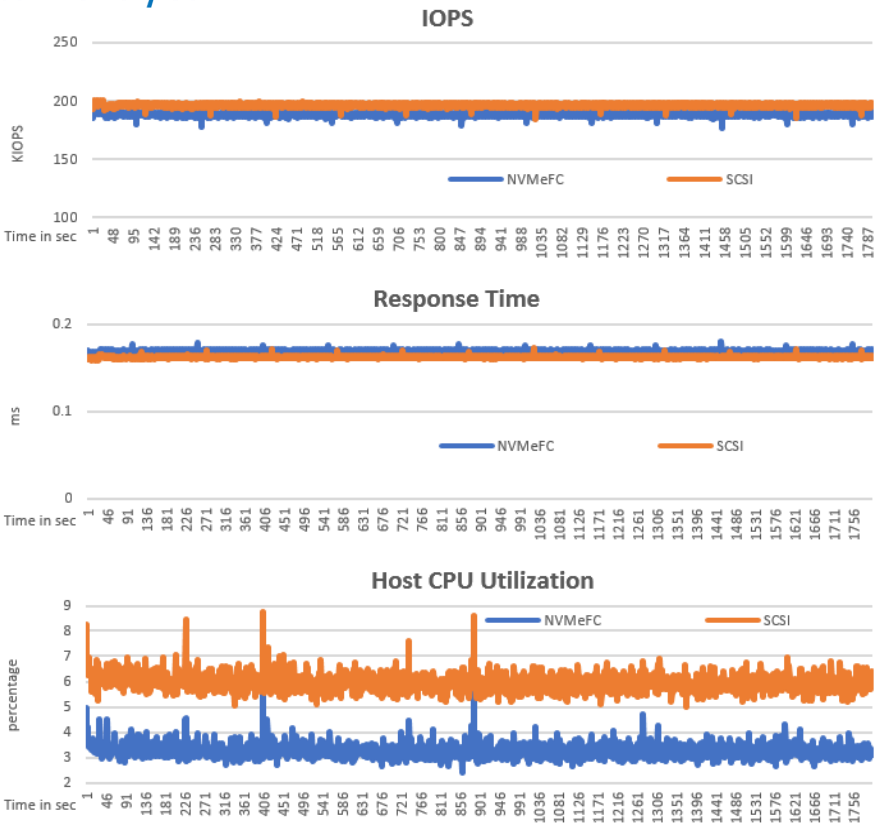
- Reduced Stack Overhead
- More Parallelism
- Lower CPU Utilization
- Higher IOPs driven by less CPU utilization
- CPUs can spend more resources on application



Non-Volatile Memory (NVMe) Protocol

Up to 64,000 Queues
64,000 Commands per Queue
Each Core has dedicated queues per SSD

FS9100 SCSI vs NVMe/FC: Typical 70% READ workload



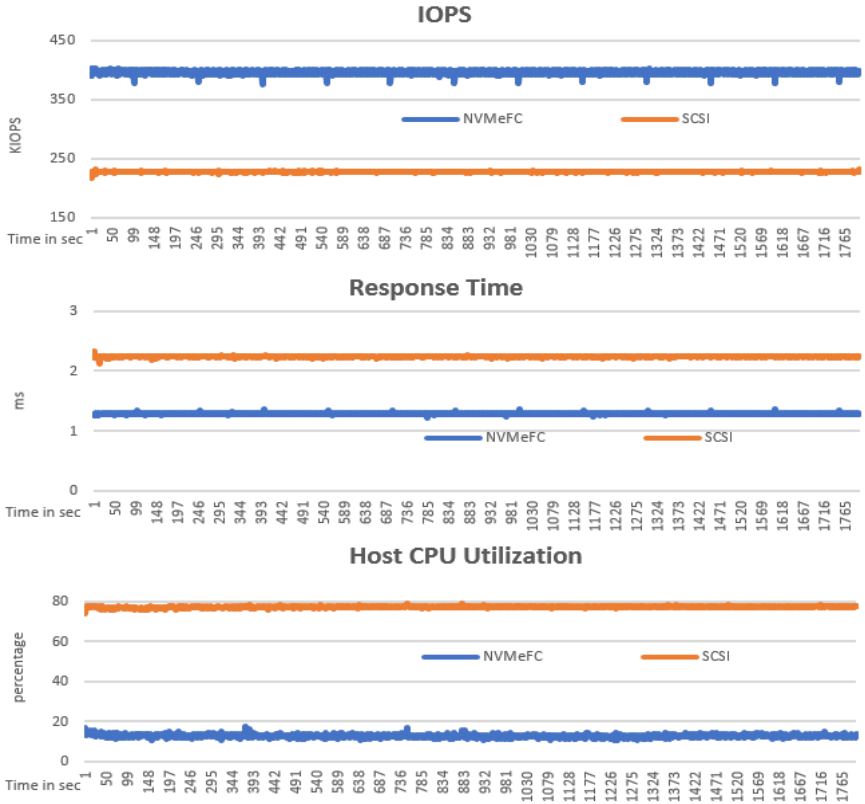
Workload

- IO size 4KB @70% Read / 30% Write cache hit
- Identical workload of 200K IOPs on SCSI and NVMe/FC
- 32 SCSI devices with total QD=32
- 32 NVMe/FC devices with 4 associations and total 64 queues

Performance

- SCSI and NVMe/FC IOs show identical response time
- SCSI performance max out at QD32
- ***NVMe/FC delivers same IOPs at half CPU consumption***
- Code efficiency in NVMe/FC host stack
- Potential queuing in NVMe/FC gives same response time with lesser CPU cost than SCSI

FS9100 SCSI Vs NVMe/FC: On IO intensive 70% READ workload



Workload

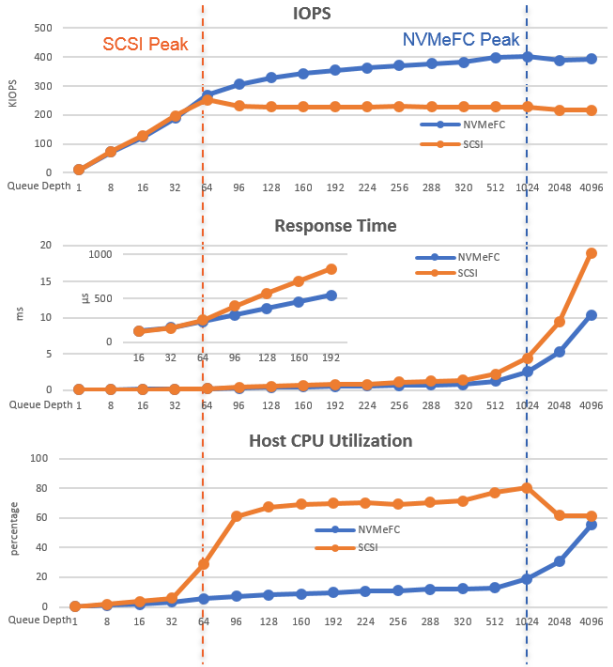
- IO size 4KB @70% Read / 30% Write cache hit
- Maximum workload with QD 512
- 32 SCSI devices
- 32 NVMe/FC devices with 4 associations and total 64 queues

Performance

- ***NVMe/FC IOPs scale to 400K IOPs***
- ***NVMe/FC show 50% latency drop over SCSI***
- NVMe/FC IOPs limited by Storage target port capability
- SCSI IOPs limited to 220K IOPs
- SCSI performance limited by host stack bottleneck
- SCSI drives CPU usage almost to 70%

FS9100 70% READ: SCSI Vs NVMe/FC

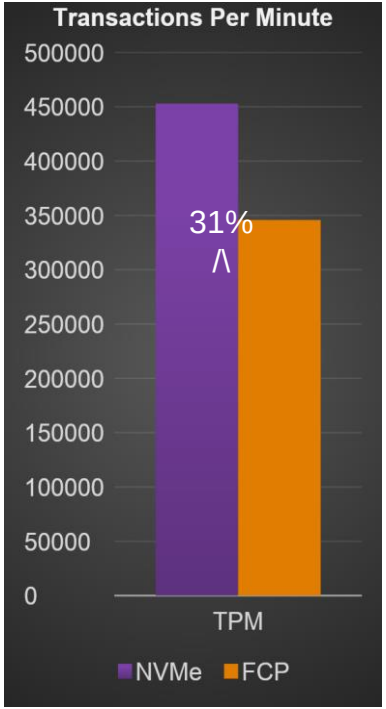
Workload: IO size 4KB @70% Read, 30% Write 32 devices, total 64 NVMe queues, varying QD

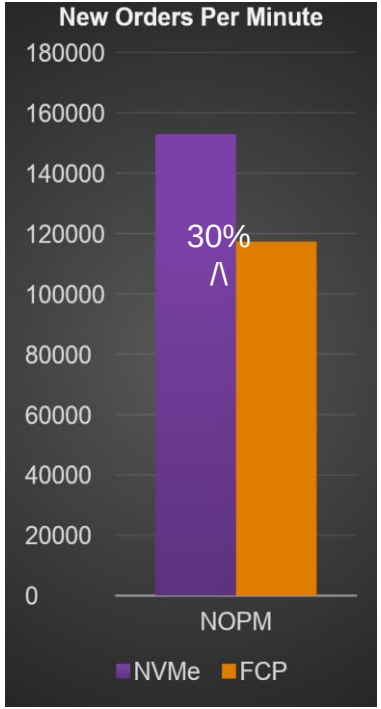


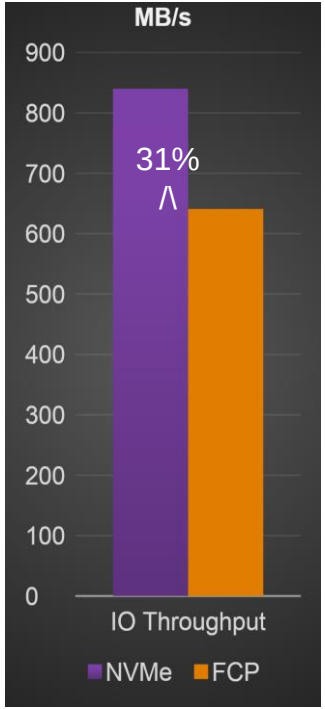
SCSI Vs NVMe/FC

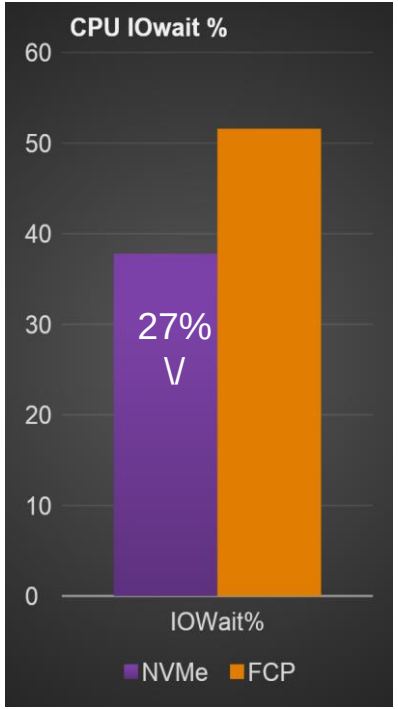
- ***Exhibit identical IOPs, Response time till 270K IOPs.***
  SCSI hits hockey stick curve @270K IOPs, NVMe/FC perf peaks at 400K IOPs
- Sub-millisec Response time up to 512 QD for NVMe/FC
- Exhibit similar Response time up to 512 QD
- Less room for application after SCSI jumps CPU usage with higher QD
- ***Fairly low CPU usage with high NVMe/FC IOPs***

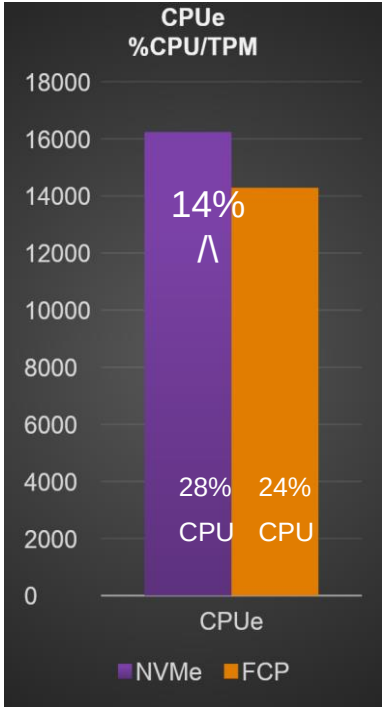
Oracle 12C TPC-C FCP vs NVMeFC on FS9150











*Bypass file system cache, NOOP FCP scheduler, Nomerges, array cache on (default)

Going Beyond Low Latency

- NVMe provides opportunity for Accelerators
- Modern Storage stacks are very complex – Data Reduction and Log Structured Array, Encryption, Erasure Codes (RAID), Replication, etc
- Doing everything in firmware affects the performance significantly
  - Higher response times
  - Longer tail latencies – Inconsistent performance
  - Lower IOPs

Examples of Accelerators

- High Speed Compression
- Other data reduction assists
- Searching and Sorting