Western Digital.

TECH BRIEF

Next Evolution of Disaggregated Storage Services

With the ever-increasing compute capabilities of modern CPUs and GPUs, the affliction of being "data-starved" has become a major pain point. This term perfectly illustrates that companies and research institutions require more performance and lower latency from their local and remote storage. Enter NVMe™ SSDs. These devices can produce significantly higher throughput and lower latencies because they are attached directly to the PCIe® bus and have parallel connections and multiple queues. However, to begin fully incorporating these devices, companies are redesigning their entire data center infrastructure (servers, drives, network, etc.) and possibly accelerating server refresh schedules, costing significantly more money. Enter NVMe over Fabrics (NVMe-oF™).

Disaggregating NVMe Storage

Utilizing NVMe-oF standards offers the benefits of NVMe SSDs, low latency, and high-performance parallel data paths, while adding the benefits of disaggregation, customizable capacity sizing, and allocation. Utilizing Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) enhances these amazing capabilities by adding RDMA functionality to eliminate many wasted CPU cycles, TCP overhead, and SCSI protocol translation. Western Digital's OpenFlex™ and RapidFlex™ products use RoCE v2, which provides low latency and high throughput to disaggregated NVMe storage.

RoCE v2 Advantages

- Low CPU utilization by applying RDMA to NVMe-oF namespaces
- Low latency by minimizing memory copies on send and receive
- Increased throughput efficiency by using streamlined network protocols
- Reduced investment by utilizing Ethernet-based network infrastructure
- Extended access with RoCE v2's introduction of routable packets

Realizing Storage Performance

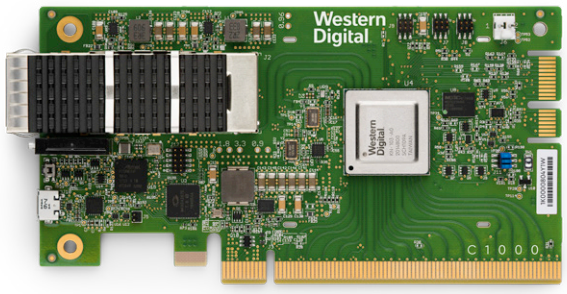
NVMe provides improved storage performance; however, in order to achieve this blazing speed, it uses a very limited internal resource: PCIe lanes. For a standard server utilizing only the CPUs for compute this might only be a minor concern, but GPU clusters place a much higher importance on available PCIe lanes. This is where utilizing fabric-attached storage comes in, and with RoCE v2 your GPUs can now directly access your storage without wasting time waiting on CPU interrupts.

In the data center, using RDMA offloads data movement, allowing higher availability of CPU resources to the application. Adopters of RoCE v2 gain the benefit from RDMA's capabilities, possibly without changing their network infrastructure. By reducing Ethernet network latency and offloading CPU overhead, RoCE v2 can increase performance in search, storage, database, financial and high transaction rate applications. By increasing CPU efficiency and improving application performance, RoCE v2 can reduce the number of servers needed, which reduces costs by lowering CPU-based software licensing fees and increasing the available network ports.

How Does NVMe-oF on RoCE v2 Stack Up?

Like every other protocol, there are multiple ways to implement an NVMe-oF storage network. Alternatives such as NVMe-oF support in Fibre Channel (FC), Transmission Control Protocol (TCP), internet Wide Area RDMA Protocol (iWARP), and RoCE v2 each has its advantages. Some institutions will choose a solution based on cost, others on familiarity, some on performance, and lastly some by future outlook. Western Digital chose RoCE v2 because we believe it satisfies each objective.

With NVMe-oF on FC, companies may be able to maintain their existing familiar storage infrastructure, which helps lower upfront costs. This assumes their existing storage platforms, switches, and HBAs can be firmware-upgraded to support such a new protocol. However, in order to take full advantage of the available throughput, companies may to need to purchase newer FC switches and HBAs in addition to replacing fiber cables and transceivers. While FC is the most frequently deployed storage fabric, its implementation with NVMe-oF is still tiny due mostly to its newness.



RapidFlex NVMe-oF Controller

With NVMe-oF using iWARP, companies again may be able to reuse their existing familiar infrastructure of TCP-based network and add the advantage of RDMA. However, low latency could be lost with a significantly larger network stack which includes Direct Data Placement, marker PDU aligned framing, and separate RDMA protocol all on top of TCP. Not only does this build up more latency, it could also limit the effective throughput gains of newer high-bandwidth networks. While latency and throughput of iWARP is limited, it does still offer some of the same benefits of RDMA, though sometimes with reduced performance. Of the various NVMe-oF protocols, iWARP is the least supported in network switches and controllers.

With NVMe-oF on TCP, companies may be able to maintain their existing familiar TCP-based infrastructure. Unlike iWARP, this solution allows for direct TCP transport binding from NVMe, instead of overlaying other protocols. This means your network infrastructure should not need upgrading except to achieve higher throughput. However, just like any TCP-based communication, low latency may not be realized, host processor utilization is typically higher, and throughput efficiency can be constrained. In order to achieve the desired performance of NVMe, companies may have to purchase extra equipment to isolate NVMe-oF storage traffic from standard TCP/IP traffic. While this does counter the cost benefits, this solution allows for both shared and dedicated access to NVMe-oF storage.

NVMe-oF on RoCE v2 provides the best of all worlds. It is Ethernet-based and thus uses familiar network components requiring some easy configuration changes. Being User Datagram Protocol (UDP)-based allows for efficient communications and throughput while minimizing latency. With a properly configured Priority Flow Control (PFC) and Explicit Congestion Notification (ECN), concerns about congestion or packet loss are mitigated. Like TCP, the cost per port is known and controllable on both switch and network interface card (NIC) while using the same cabling and transceivers. With RoCE v2's support for Layer 3 routing, distance limitations are less of a concern. RoCE v2 is an ideal choice for your disaggregated NVMe storage with its known and controllable costs, use of familiar technology, optimal performance, and use of a network protocol designed specifically for the current and future of clustering and storage network.

Looking to the Future of NVMe-oF

As we continue moving deeper into this high-performance data-centric world, Western Digital is continuously developing new and exciting storage platforms within the OpenFlex and RapidFlex product lines, enabling customers to move into this performance-based NVMe-oF ecosystem. We believe the future of NVMe-oF will be the enhancement of storage flexibility while maintaining mission-critical reliability and performance. To answer this need, Western Digital recently released the OpenFlex Data24, featuring the Western Digital RapidFlex NVMe-oF controllers, which offers flexible deployment options by allowing for multiple servers to directly attach to this platform. Removing the requirement of a RoCE v2-supported switch, the Western Digital OpenFlex Data24 reduces deployment complexity, lowers NVMe-oF entry cost, increases performance, and lowers latency while maintaining all of the previously mentioned RoCE v2 benefits.

OpenFlex Data24 NVMe-oF Storage Platform

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com