SC22 Network Research Exhibition: Demonstration Abstract

NRE-002: High-Performance Data Transfer Nodes for Petascale Science with NVMe-over-Fabrics as Microservice

Se Young Yu, Jim Chen, Fei Yeh, Joe Mambretti
International Center for Advanced Internet Research - Northwestern University,
young.yu, jim-chen, fyeh, j-mambretti@northwestern.edu

Abstract

The PetaTrans with NVMe-over-Fabrics as microservice is a research project aimed at improving large-scale WAN microservices for streaming and transferring large data among high-performance Data Transfer Nodes (DTNs). Building on earlier initiatives, for SC22, we are designing, implementing, and experimenting with NVMe-over-Fabrics on 400 Gbps Data Transfer Nodes (DTNs) over large-scale, long-distance networks with direct NVMe-to-NVMe service over RDMA over Converged Ethernet (RoCE) and TCP fabrics using SmartNICs. NVMe-over-Fabrics microservice connects remote NVMe devices without userspace applications, reducing overhead in high-performance transfer and offloading NVMe-over-Fabrics initiators software stack in SmartNICs. The primary advantage of the NVMe-over-Fabrics microservice is that it can be deployed in multiple DTNs as a container with lower overhead.

We optimized NVMe/RDMA performance to achieve 133 Gbps streaming throughput over a 22ms path between StarLight and McLean using StarLight 200 Gbps WAN testbed.

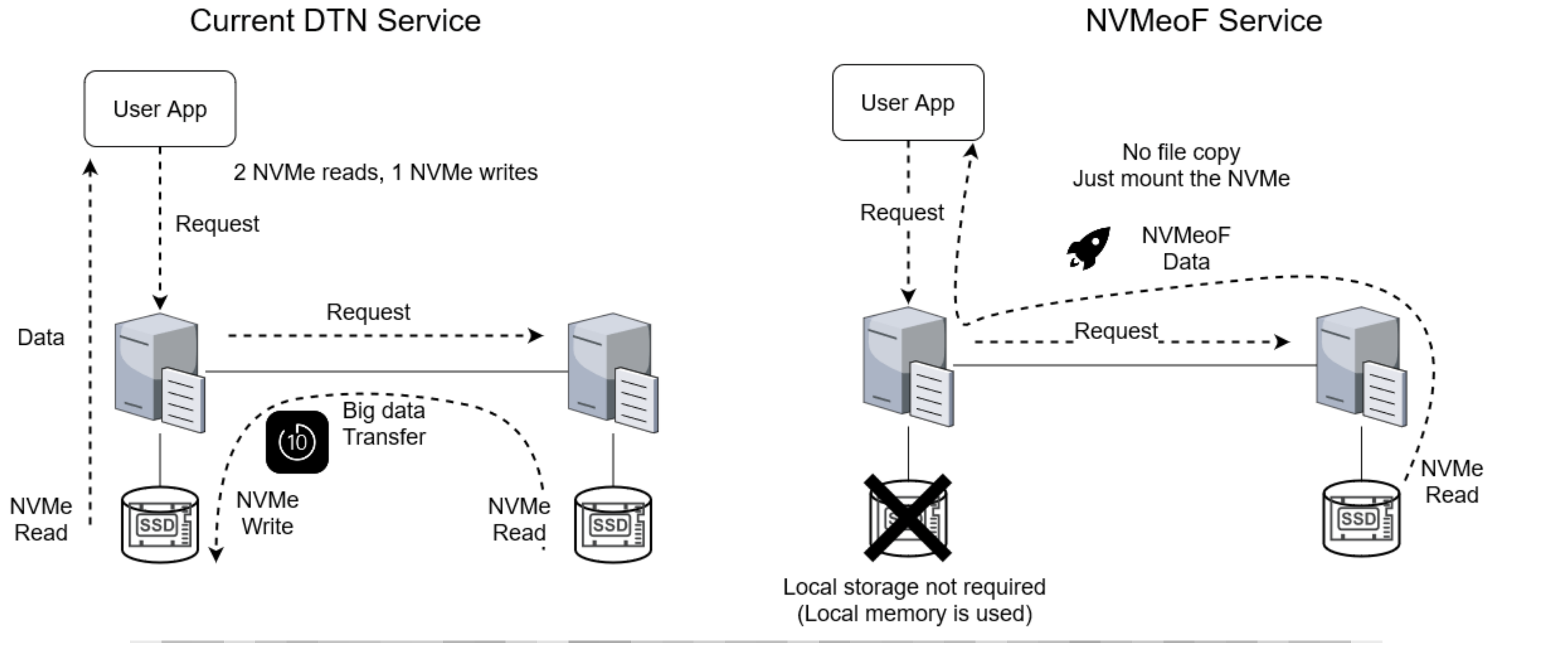
Goals

1. The PetaTrans -400 Gbps Data Transfer Node (DTN) with NVMe-over-Fabrics research project is aimed to improve large-scale WAN microservices for high-performance data streaming and transfer using a novel NVMe-over-Fabrics technique.

2. We are designing, developing, and experimenting with 400 Gbps Data Transfer Nodes (DTNs) with NVMe-over-Fabrics as microservice over multi-200 and 400 Gbps Wide Area Networks (WANs) to demonstrate the capabilities of NVMe-to-NVMe direct connections in high-performance long-distance networks.

3. NVMe-over-Fabrics microservice is being designed to optimize capabilities for supporting E2E (e.g., edge servers with 200 Gbps NICs and multiple NVMes) large scale, high capacity, high capacity, high performance, reliable, high quality, sustained individual data streams for science research.

4. Related research includes DTN components, such as NUMA and NVMe, multiple flows of transport protocols, system and network monitoring, orchestration of various systems, pipelined workflows, NVMe-over-Fabrics as microservice, and other considerations.



5. Utilize SmartNICs for offloading NVMeoF functions to minimize overhead for accessing remote NVMe devices in the host system. Performance optimization for SmartNIC-NVMe as well as host-SmartNIC communications.

6. Enhancements include additional capabilities for controlling NVMe-over-Fabrics using microservice architecture to establish a direct mapping between remote NVMe devices and transparent data streaming and transfer with low overhead over high-performance long-distance networks.

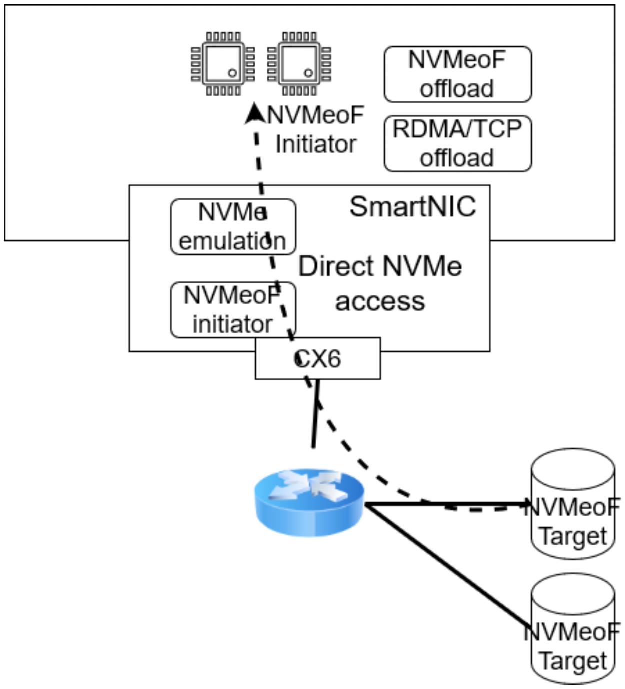


Figure 2. SmartNIC NVMeoF offloading

Resources

Required resources from SCinet WAN are 1 Tbps E2E WAN services from the StarLight International/National Communications Exchange Facility in Chicago to the SC22 venue, between StarLight and the JBDT Facility in McLean, between the JBDT Facility and the SC22 venue and among all sites. In addition, another site utilized will be a 400 Gbps ESnet testbed at Berkeley connected to the StarLight Facility.

Involved Parties

- Se-Young Yu, iCAIR, young.yu@northwestern.edu
- Jim Chen, iCAIR, jim-chen@northwestern.edu
- Fei Yeh, iCAIR, fyeh@northwestern.edu
- Joe Mambretti, iCAIR, j-mambretti@northwestern.edu