

NVMe RDMA 그리고 그 외 각광 받는 기술들

글. 유재근 쿼텀코리아 실장

서론

지난 10년 동안 엔터프라이즈급 비휘발성 '플래시' 스토리지가 주를 이루었다. 엔터프라이즈 플래시의 가격은 '올 플래시 데이터 센터'에 더 이상 예전만큼 눈길을 주지 않을 정도로 하락했다. 스피닝 디스크의 용량은 여전히 비교적 저렴하지만 HDD 밀도와 성능 향상에 대한 기술적 장벽은 높아지고 있다. 대부분의 데이터 센터에는 기업의 요구 사항에 따라 두 가지 기술을 모두 수용하도록 플래시와 스피닝 디스크(HDD)를 혼합하여 사용하고 있다.

이제 모든 개인용 컴퓨터 사용자들도 HDD 스토리지를 SSD로 전환하는 것만으로도 얻을 수 있는 성능 향상에 익숙해져 있다. 많은 사람들이 알아차리지 못하고 있는 사실은 이러한 업그레이드에 따른 모든 이점을 제대로 누리지 못한다는 것이다. 플래시 스토리지는 매우 빠르기 때문에 스토리지 시스템의 병목 현상은 더 이상 스토리지 디바이스 자체가 아니다. 이제는 스토리지 인터페이스로 인해 문제가 발생하기도 한다.

이 모순에 대한 해결방안은 NVMe(비휘발성 메모리 익스프레스)이다. NVMe는 플래시 SSD에서 최신 영구적인 메모리 기술에 이르기까지 모든 유형의 비휘발성 메모리의 성능을 발휘하도록 설계되었다. 이 기고에서는 NVMe의 정의, 독립 실행형 및 네트워크 구현 방식 모두에서 보다 나은 성능을 구현하는 방법과 속도의 이점을 누릴 수 있는 몇 가지 주요 구축 사례를 소개한다.

엔터프라이즈 비휘발성 스토리지 환경은 광범위하고 복잡하다. 스타트업과 기존 스토리지 제공업체들 모두는 이 모든 문제를 해결하겠다고 약속한다. 새로운 기술을 적용하는 것은 비용이 많이 든다. DRAM 및 3D-Xpoint와 같은 최첨단 영구 메모리 장치가 가장 비용이 많이 들며, 엔터프라이즈 NVMe SSD와 SAS SSD는 중간이다. TCO(총소유비용)가 중요하다. 성능 밀도, 인프라 비용과 시간 같은 요소들은 모두 NVMe 스토리지가 특정 애플리케이션에 적합한지 여부에 영향을 미친다.

가장 포괄적인 의미에서 디지털 제품을 빠른 속도로 생산함으로써, 더 높은 매출 또는 경쟁 우위로 이어지는 거의 모든 분야 고객들에게 NVMe 스토리지 도입이 적합한 후보라고 할 수 있다. 왜 NVMe가 이렇게 중요한지 이해하기 위해서는 과거를 되돌아보는 것이 도움이 될 것이다.

데이터 스토리지의 간략한 발전사

꽤 오랜 기간 하드디스크는 HDD를 의미했다. 실시간 운영을 위한 최초의 저렴한 대용량 저장 매체로써, 저장 용량이 커진 HDD를 탑재한 PC의 혁명은 유비쿼터스를 만들었고 그에 따라 비용도 지속적으로 감소했다. 폼 팩터가 감소하였고, 플래터당 밀도와 플래터 수는 증가했다. 하지만 장치의 본질적인 물리(형태)적 연결방식은 변하지 않았다. 결과적으로, HDD와 통신하는 데 사용되는 통신 프로토콜인 ATA와 SAS는 하나의 대기열만 가지고 있다.

초기 또는 적어도 꽤 오래전에는 비휘발성 스토리지도 있었다. 현대의 NAND 플래시는 1980년대에 발명되었다. 이후 소비자 전자 기기에서 널리 사용되면서 개발이 촉진되었고 궁극적으로는 가격 하락을 주도했다. 플래시의 장점은 기계식이 아니며 명령을 직렬로 처리하는 것에 국한되지 않는다는 것이다. 또한, 짧은 대기시간과 높은 병렬 액세스는 플래시의 가장 큰 장점 중 하나이다. HDD는 성능이 우수하고, 용량이 비교적 크고, TB당 비용이 저렴하여 오랫동안 지배적이었다. 하지만 시간이 지남에 따라 플래시 가격은 떨어지고 있으며, 가장 큰 플래시 SSD의 용량은 이제 가장 큰 HDD보다 훨씬 더 크다. 플래시는 높은 가격 및 고성능 애플리케이션에서 가장 많이 사용되지만, 최근 대용량과 높은 성능 밀도로 HDD보다 TCO가 낮아 전력, 냉각 및 물리적 공간 소비가 상대적으로 매우 적다.



그림 1. HDD와 NAND 플래시 기술의 시작

NVMe 입문

플래시 기반 스토리지는 간단하게 HDD를 SSD로 교체하여 기존 디스크 어레이의 속도를 높일 수 있다. 그러나 플래시의 짧은 대기시간과 높은 병렬 처리를 최대한 활용하려면 단순히 스토리지를 바꾸는 것 이상의 작업을 수행해야 한다. 즉, 새로운 연결 기술이 필요하다. 병렬적인 측면에서 SAS의 단일 대기열 아키텍처에는 병렬 아키텍처가 없으므로 병렬적인 대기열을 처리할 수 있는 플래시와 일치하지 않는다. HDD의 경우 메카닉으로 인해 많은 대기시간이 발생한다. 만약 소프트웨어 업스트림이 하드웨어보다 빠르면 문제없다. 그러나 이제 플래시를 사용하면 스토리지 디바이스 자체의 대기시간이 너무 낮아 소프트웨어와 프로토콜이 가장 느린 링크가 된다. 마지막으로, PCIe에 도달하기 위해 하드웨어 스토리지 컨트롤러를 경유하여 I/O를 처리하는 것은 HDD와 SSD 모두에 더 많은 대기시간과 또 다른 병목 현상을 가중시킨다.

이러한 문제에 대한 해답은 NVM Express(nvmeexpress.org)를 통해서 확인할 수 있는 NVMe라는 인터페이스 사양이다. NVMe는 비휘발성 메모리 디바이스의 속도, 짧은 지연시간 및 기기 내 병렬 처리를 활용하도록 특별히 설계되었다.

	NVMe	SAS	SATA
시작 연도	2009	2005	2000
처리량	~8Gbps/lane	12Gbps	6Gbps
지연 세트	20 μ s 미만	500 μ s 미만	500 μ s 미만
IOP 읽기	최대 100만	최대 20만	최대 50만
커맨드 세트	NVMe	SCSI	ATA
커맨드 수	13	최대 200	
큐(Queue) 수	65,535	1	1
커맨드 큐 깊이	65,536	254	32
폼 팩터	2.5인치, U.2, M.2, Rules	2.5인치, 3.5인치	2.5인치, 3.5인치

표 1. NVMe, SAS, SATA 비교

[표 1]은 NVMe와 기존 인터페이스의 여러 가지 중요한 차이점을 요약하여 보여주고 있다. 이 중에 강조되어야 할 몇 가지 사항이 있다. 하나는 NVMe의 대규모 단일 디바이스 병렬 처리 능력이다. SAS에는 최대 254개의 항목이 포함된 단 하나의 큐가 있는 반면, NVMe 드라이브에는 각각 65,536개의 항목이 포함된 65,535개의 Queue를 처리할 수 있다. 몇 가지를 구현할지는 장치의 제조사에 달려있지만, 제품의 사양은 많은 헤드룸으로 구축된다.

실제로, 대기열이 많다는 것은 [그림 2]의 오른쪽에 표시된 것처럼 여러 CPU 코어가 각 코어/드라이브 쌍에 대한 큐와 함께 많은 NVMe 드라이브에 동시에 액세스할 수 있다는 것을 의미한다.

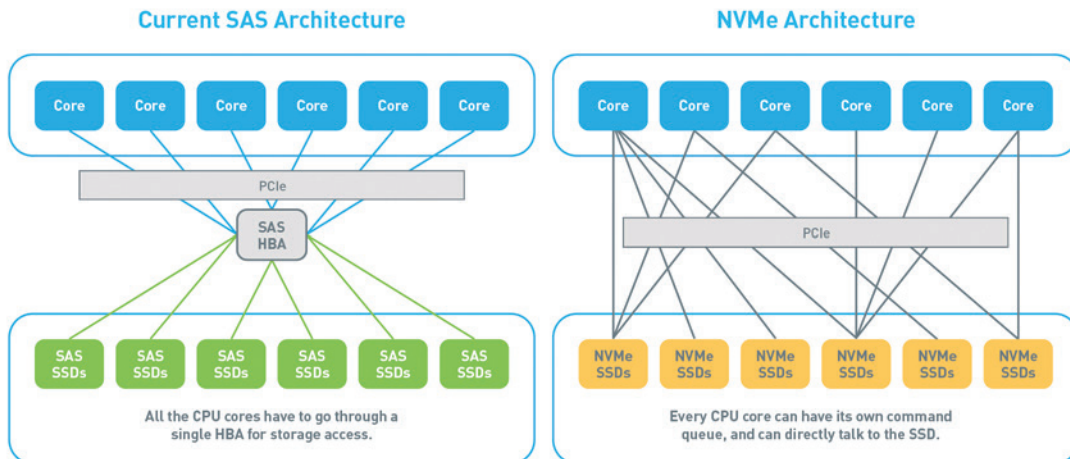


그림 2. NVMe 전과 후

[그림 2]에서 설명하는 또 다른 장점은 NVMe 디바이스가 SAS 컨트롤러 또는 HBA와 같은 중간 매체 대신 PCIe 버스에 직접 연결되어 있다는 것이다. 공유 메모리는 CPU를 우회하여(거치지 않고) 애플리케이션에서 스토리지로 데이터를 바로 전송하는 데 사용된다. 중요한 사실은, NVMe 인터페이스는 지연시간 최소화, 향상된 처리량과 IOP 기능으로 모든 면에서 훨씬 더 빠르다는 것이다.

공유 스토리지 및 NVMe 네트워킹

지금까지 설명한 모든 사항은 단일 서버 내에 연결된 저장 장치에 적용된다. 퀀텀은 8K 비디오 재생 및 자율 주행 차량 연구와 같은 까다로운 워크플로우를 위한 고성능 공유 스토리지 전문기업이다. 공유를 위해서는 네트워크가 필요하다.

NVMe의 장점을 공유하기 위해서 공유 메모리를 사용하는 대신, 네트워크를 통해 호스트와 스토리지 간에 메시지가 전송된다는 것이다. 네트워크는 파이버 채널(FC), 인피니밴드(IB) 또는 이더넷이 있지만, 요즘 대부분의 사용자들은 이더넷을 선호한다. 일단, 이더넷 어댑터와 스위치는 저렴하다. 그러나 장비 비용보다 더 중요한 것은 편의성이다. 신규 도입을 위해 FC 또는 IB 네트워킹의 복잡한 기술을 배울 필요가 없으며, 대부분의 조직은 이미 이더넷 전문 지식을 보유하고 있다. 오늘날 FC 또는 IB 스토리지 네트워크가 구축되어 있더라도 단일 네트워크 인프라로 통합할 수 있는 이점이 있다. FC와 IB의 연결이 중요하지 않다는 것은 아니다. 이러한 기술에 기반한 전용 스토리지 네트워크를 당분간 계속 사용할 수 있을 것이다. 이에 반대하는 경쟁사가 있다면 그것은 아마도 이더넷만 지원하고 있을 가능성이 크다.

FC, IB 및 이더넷의 차이점은 FC와 IB가 처음부터 스토리지 트래픽을 효율적으로 처리하도록 설계되었다는 점이다. 이들은 데이터를 ‘안정적’이고 (누락되지 않고) 순서대로 전달한다. 이더넷은 범용 네트워크로 설계되었으며 설계상으로는 “안정적이지 않다”고 한다. 데이터는 전달되지만, 일부는 다른 경로를 통해 순서에 맞지 않게 전송되거나 전송 중에 누락될 수도 있다. 발신자와 수신자의 상위 계층 프로토콜은 이더넷의 기본 신뢰성, 데이터 버퍼링, 데이터 재정렬, 손실된 패킷

감지 및 재전송과 같은 근본적인 불안정성에 대해 고려해봐야 한다. 이 모든 것은 시간이 걸린다. 다시 말해, 스토리지 네트워크의 대표적인 불편사항인 대기시간이 발생하게 된다.

이더넷을 통해 대기시간이 짧고 빠른 스토리지 네트워크를 구축하려면 대부분의 조직에서 이미 보유하고 있는 기술 외에 추가 기술을 구축해야 한다. '무손실 이더넷' 구현을 위해서는 PFC(Priority Flow Control) 또는 DifServ(Differentiated Services)와 같은 DCB(Data Center Bridging) 기능이 필요할 수 있으며, RDMA(Remote Direct Memory Access)는 절대적인 요구 사항이다. DCB에 대한 철저한 처리는 이 기고의 범위를 벗어나지만, RDMA는 이더넷 네트워크에서 NVMe 장치를 성공적으로 사용하는 비결이기 때문에 자세히 살펴보고자 한다.

비결은 RDMA

이더넷 네트워크 NVMe의 지연시간 절감과 성능에 가장 큰 기여를 하는 요소는 RDMA이다. RDMA (iSER) 및 NVMeoF (NVMe over Fabrics)용 iSCSI 확장은 RDMA를 기반으로 구축되었으며 IB에서 영구적으로 사용되어 왔다. RDMA 컨소시엄(rdmaconsortium.org)은 2002년 TCP/IP 네트워크를 통해 RDMA 기술을 개발하고 홍보하기 위해 설립되었다. 이러한 오래된 개념이 어떻게 새로운 해결책이 되었는지 이해하기 위해서는 일반 TCP/IP 이더넷 네트워크가 어떻게 구축되는지 검토해 볼 필요가 있다.

TCP/IP와 같은 기존 네트워크 프로토콜 스택은 운영 체제 커널에서 여러 계층으로 구성되어 있다. 단계마다 버퍼가 있다. CPU는 해당 버퍼와 데이터를 주고받는 애플리케이션 간에 데이터를 복사하는 작업에 관여한다. 계산 비용이 많이 드는 컨텍스트 스위치가 필요하기 때문에 대기 시간과 CPU의 사용률이 높아지게 된다. 스토리지 프로토콜 더미가 위에 배치되면 상황은 더욱 악화된다.

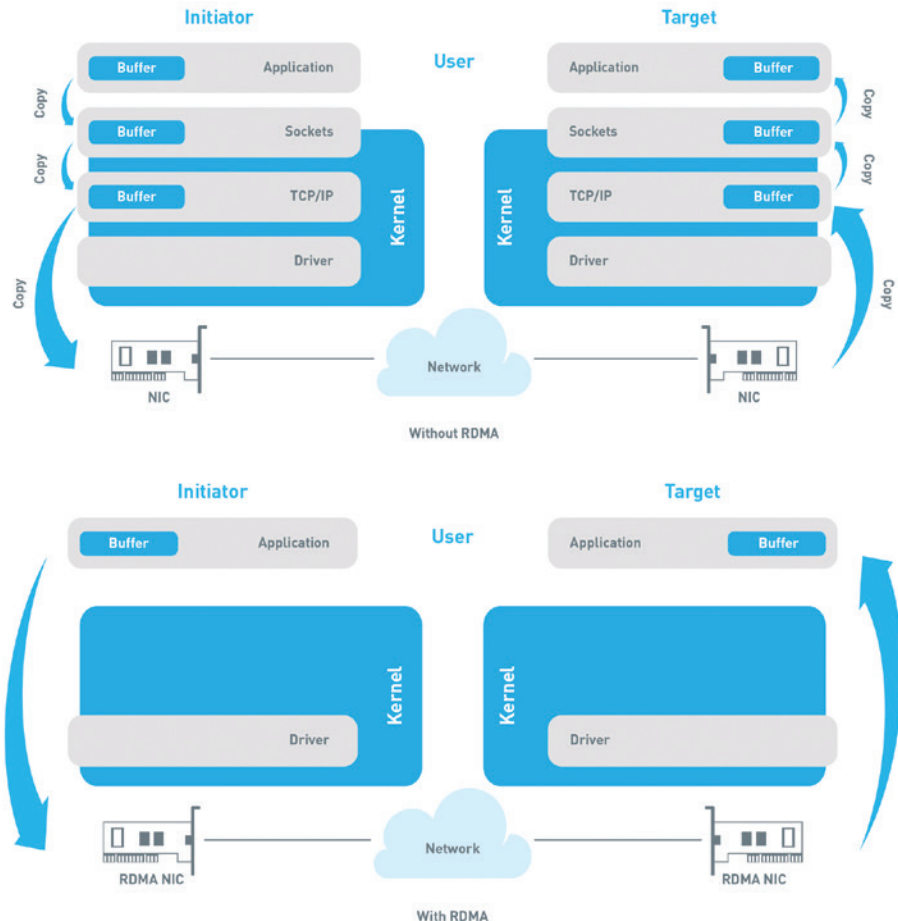


그림 3. RDMA 전과 후

대조적으로, RDMA는 데이터 전송을 위한 확실한 지름길이다. 시내 중심가에서 빨간 불을 기다리는 대신 고속도로를 타고 도시를 가로지르는 것과 같다. RDMA를 사용하면 데이터가 사용자 공간 애플리케이션 간에 직접 이동된다. 거의 모든 세부 사항이 네트워크 인터페이스 하드웨어로 오프로드 된다.

CPU는 연결 설정 및 감독에 관여하지만 데이터 이동은 NIC로 완전히 오프로드 된다. 이렇게 하면 다른 작업에 필요한 CPU 주기를 확보할 수 있다. 모든 애플리케이션 I/O는 운영 체제 커널을 우회하여 사용자 공간에서 발생하므로 버퍼링과 컨텍스트 전환이 필요하지 않다. 기존에 커널에서 처리된 네트워크 프로토콜은 NIC 하드웨어로 오프로드 된다. 이러한 오프로드 및 스트리밍 효과는 네트워크에 의해 부과되는 추가 지연시간을 최대 10 μ s 이하로 감소시키는 것이다. 이것이 RDMA가 이더넷 네트워크 NVMe의 비결인 이유이다.

이더넷 RDMA 옵션

RDMA가 중요한 역할을 하는 만큼, RDMA 지원 네트워크를 구현하는 것이 편리하다면 매우 좋을 것이다. 하지만, 이를 구현하는 데에는 여러 방법들이 있다. 다음 내용을 통해 RDMA 지원 이더넷의 기존 옵션에 대한 간략한 개요를 확인할 수 있다. 항상 그렇듯이, 특정한 구성에 가장 적합한 선택은 개별 상황과 요구 사항에 따라 다르다.

새로운 기술을 적용할 경우 초기 제품들은 흔히 독점적인 구현 방법을 사용한다. 큰 틀이 잡히기 시작하면 표준이 더욱 명확해진다. 표준 기반 솔루션이 증가함에 따라 독점 옵션이 사라지거나 자신만의 고유한 이점이 있고 틈새에 맞는 구축 사례가 된다. 이러한 방법론은 이더넷 네트워크 NVMe 스토리지에 적용된다. 독점적 또는 일부 독점적인 솔루션이 존재하지만, 표준 기반 제품들에 의해 빠르게 복잡성이 사라지고 있다. NVMe 네트워크 RDMA 지원 이더넷을 구축하기 위한 두 가지 관련 표준은 RDMA(iSER) 및 NVMeoF(NVMe over Fabrics)용 iSCSI 확장이다.

iSER는 RDMA를 사용하기 위해 iSCSI를 확장하는 것으로 고안되어, 2007년에 RFC 5046과 함께 IETF 표준으로 전환되었다. iSER는 세 가지 네트워크 전송 기술 중 하나인, RDMA 서비스(iWARP)가 포함된 TCP, RoCE(RDMA over converged Ethernet) 또는 IB를 사용한다. iSER를 과거의 iSCSI와 혼동하지 않아야 한다. 이 둘은 호환되지 않으며, iSER는 RDMA를 통해 훨씬 더 높은 성능을 제공할 수 있다. 하지만 기본적으로는 여전히 SCSI 프로토콜이 포함되어 있다. 이론적으로 이것은 NVMeoF와 같은 순수한 NVMe 구현보다 대기시간이 더 길고 잠재적으로 성능이 낮다는 것을 의미하지만, 실제로 이것이 중요한지의 여부는 스토리지 시스템의 설계와 사용 방법에 달려 있다.

NVMeoF는 더욱 새로워졌다. 이 사양의 버전 1.0은 2016년 발표되었으며 그 이후 계속 발전해 오고 있다. NVMeoF는 오늘날 이더넷뿐만 아니라 IB 및 FC를 포함한 모든 네트워크 패브릭에서 NVMe의 효율성을 확장하도록 설계되었다. RoCE와 iWARP는 지원되는 RDMA 이더넷 전송이며, NVMe/TCP라고 알려진 기존 TCP 소켓을 기반으로 하는 새로운 non-RDMA 이더넷 구현이 가능하다. FC와 IB는 non-Ethernet의 대안이다.

요약하자면 iSER 또는 NVMeoF를 통해 RDMA over Ethernet으로 NVMe를 실행하려면 RoCE 또는 iWARP를 지원하는 이더넷 네트워크(NIC 및 스위치 모두)가 필요하다. 이 둘은 호환되지 않으므로, RoCE 클라이언트는 iWARP를 통해 제공되는 스토리지 리소스에 액세스할 수 없으며 그 반대의 경우도 마찬가지이다.

NVMe 스토리지의 장점

앞서 설명한 것처럼 NVMe 장착 스토리지의 가장 큰 장점 중 하나는 짧은 지연시간이다. 따라서 누락된 프레임이 있어서는 안 되는 고성능 고해상도 스트리밍 미디어에 특히 적합하다. 4K 및 8K, UHD 및 HDR 비디오 워크로드가 대표적이지만, 편집 및 그래픽 작업, 많은 수의 저해상도 스트림을 수집하거나 전달할 경우에도 적합하다.

AI/ML 애플리케이션, IOT 및 고성능 분석, 특히 실시간 요구 사항이 있는 경우 NVMe 스토리지의 높은 IOPS와 성능 장점은 다양한 유형의 워크로드(작업 요구사항)에 많은 도움이 될 수 있다.

NVMe 스토리지의 가격이 비싸다고 생각될 수 있지만, 초기 도입 비용은 종종 환경 요인에 의해 상쇄될 수 있다. 고성능 HDD 기반 시스템을 구축하려면 많은 스피ンを 통합해서 사용해야 한다. 이는 용량과 성능이 크게 일치하지 않아 성능 요건을 충족하기에 충분한 스피너들을 획득하기 위해서는 필요한 용량 이상으로 훨씬 많은 용량을 구매해야 한다. 랙 공간 외에도 작동하는 모든 하드 드라이브에는 전원과 냉각시스템이 갖춰져야 한다. NVMe 기반 스토리지 어플라이언스는 단 몇 개의 랙 유닛으로 전체 많은 랙을 차지하는 HDD 기반 시스템과 동일한 성능을 제공할 수 있다. 이것을 IOP 수, 8K 비디오 스트림 또는 랙 단위당 수행할 수 있는 기타 작업 단위의 성능 밀도라고 한다. NVMe 스토리지는 HDD 기반 솔루션, 심지어 SAS로 연결된 SSD 기반 솔루션(일반적인 SSD 스토리지)보다 성능 밀도가 매우 높다.

일부 데이터에만 초고속 액세스가 필요한 애플리케이션의 경우에는 NVMe 스토리지를 하이브리드 스토리지 어플라이언스에서 '티어 제로'로 사용하여 특징점을 확장하고 비용을 절감할 수 있다. 퀀텀의 Primary Storage Tiering 기술을 활용하면 자동화된 데이터의 계층 간 이동 기능을 적용하여 성능과 용량 요구 사항을 모두 만족하도록 사용할 수 있다.

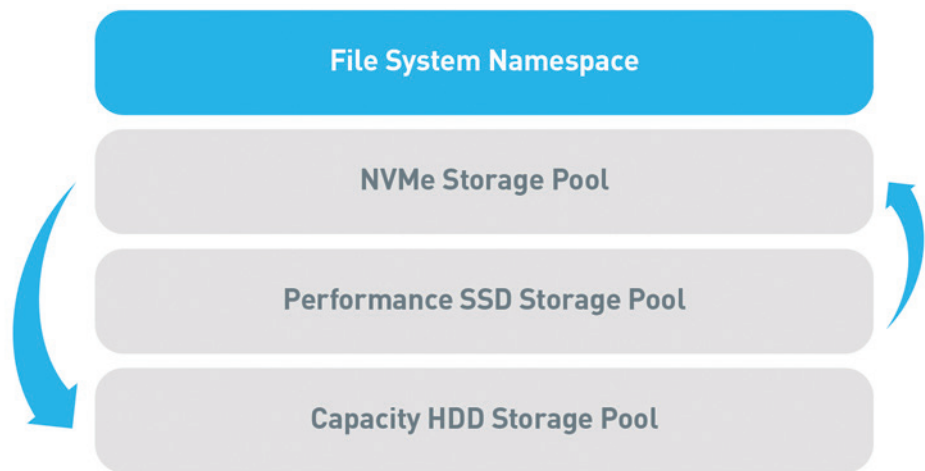


그림 4. NVMe 통합 계층 스토리지 시스템

결론

고성능 스토리지의 현재와 미래는 NVMe 장착 비휘발성 메모리 기술이다. RDMA를 활용하는 NVMe는 짧은 지연시간과 고대역폭 공유를 통해 기존 네트워크 스토리지의 병목 현상을 최소화하거나 제거할 수 있다. 계층화를 통해 지능적으로 구축할 경우 NVMe 스토리지는 비용 효율적이며 비교적 적은 용량으로 상시 액세스가 가능하다. 즉, NVMe 드라이브를 하이브리드 시스템에서 다른 유형의 스토리지와 결합하면, 비용을 제어하면서 성능을 향상시킬 수 있다. 비디오 및 기타 고대역폭 애플리케이션을 위한 공유 스토리지 시스템 설계에 대한 퀀텀의 전문 지식은 이제 미디어 및 엔터테인먼트, 비디오 감시, 자율 주행 차량 연구 및 유사한 시장을 대상으로 하는 NVMe 기반 제품까지 확대되었다. 자세한 내용은 www.quantum.com/kr 에서 확인할 수 있다. 📄