(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년08월05일
(11) 등록번호 10-2430187
(24) 등록일자 2022년08월03일

(51) 국제특허분류(Int. Cl.)
G06F 13/16 (2006.01) *G06F 13/28* (2006.01)
G06F 3/06 (2006.01) *G06F 9/44* (2018.01)
(52) CPC특허분류
G06F 13/16 (2013.01)
G06F 13/28 (2013.01)
(21) 출원번호 10-2016-0058833
(22) 출원일자 2016년05월13일
심사청구일자 2020년11월17일
(65) 공개번호 10-2017-0007103
(43) 공개일자 2017년01월18일
(30) 우선권주장
3494/CHE/2015 (Provisional Application) 2015년07월08일 인도(IN)
3494/CHE/2015 (Non-provisional Application) 2015년11월25일 인도(IN)
(56) 선행기술조사문헌
US08463881 B1
US08554968 B1

(73) 특허권자
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
삼맛세띠 산디프 아난드쿠마르
인도, 방갈로르-37, 에이이씨에스 레이아웃, 씨 블록, 1번 메인, 수카사 애비뉴, #747, 지2
(74) 대리인
특허법인가산

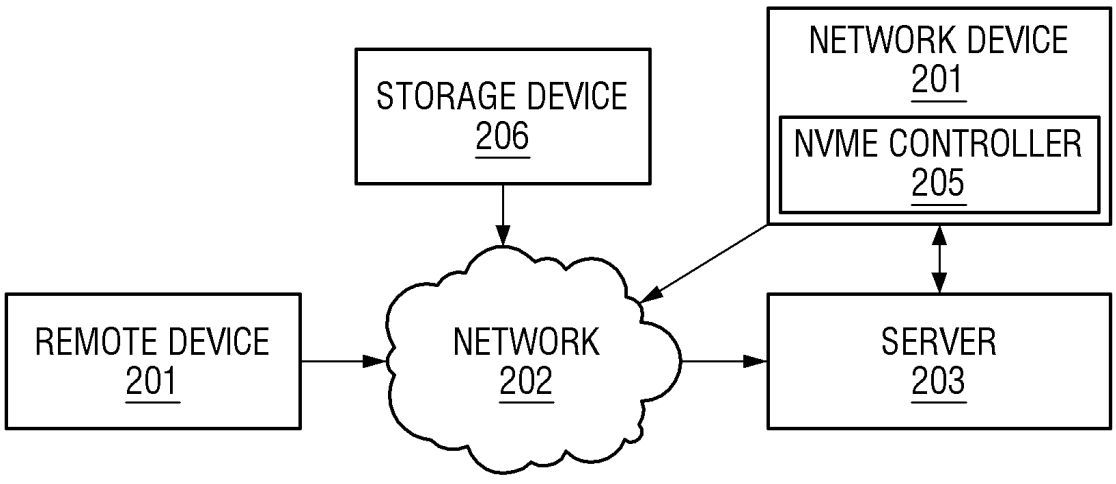
전체 청구항 수 : 총 10 항 심사관 : 지정훈

(54) 발명의 명칭 **RDMA NVMe 디바이스의 구현 방법**

(57) 요 약

네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법이 제공된다. 상기 방법은, 서버에 연결된 네트워크 디바이스로, NVMe(Non-Volatile Memory Express) 컨트롤러를 초기화하고, 상기 네트워크 디바이스로, 원격 디바이스 검색 프로세스를 처리하기 위한 NVMe 큐 페어(NVMe queue pair)를 설정하고, 상기 네트워크 디바이스 측에서, 상기 서버 측에서 유지되는 상기 NVMe 컨트롤러에 의해 제어되는 상기 스토리지 디바이스를 접근하기 위한 요청을 상기 원격 디바이스로부터 수신하고, 상기 네트워크 디바이스로, 상기 원격 디바이스를 찾기 위한 상기 검색 프로세스를 초기화하고, 상기 원격 디바이스가 검색되면, 상기 네트워크 디바이스로, 상기 NVMe 큐 페어를 RDMA 큐 페어(Remote Direct Memory Access queue pair)에 매핑함으로써 상기 원격 디바이스와의 접속을 수립하는 것을 포함한다.

대 표 도 - 도2



(52) CPC특허분류
G06F 3/0655 (2013.01)
G06F 3/067 (2013.01)
G06F 3/0679 (2013.01)
G06F 9/4403 (2013.01)
G06F 2213/0026 (2013.01)

명 세 서

청구범위

청구항 1

서버에 연결된 네트워크 디바이스, NVMe(Non-Volatile Memory Express) 컨트롤러를 포함하는 상기 네트워크 디바이스 및 상기 네트워크 디바이스에 연결된 적어도 하나의 스토리지 디바이스를 제어하는 상기 NVMe 컨트롤러에 의해 상기 NVMe 컨트롤러를 초기화하고,상기 네트워크 디바이스를 사용하여, 원격 디바이스 검색 프로세스를 처리하기 위한 NVMe 큐 페어(NVMe queue pair)를 설정하고,

상기 네트워크 디바이스를 사용하여, 상기 원격 디바이스로부터 상기 스토리지 디바이스에 접근하라는 요청을 수신하고,상기 네트워크 디바이스를 사용하여, 상기 원격 디바이스를 찾기 위한 상기 검색 프로세스를 초기화하고,

상기 네트워크 디바이스를 사용하여, 상기 원격 디바이스가 검색되면, 상기 NVMe 큐 페어를 RDMA 큐 페어(Remote Direct Memory Access queue pair)에 매핑함으로써 상기 원격 디바이스와의 접속을 수립하는 것을 포함하는,

네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

청구항 2

제1항에 있어서,

상기 원격 디바이스와의 접속을 수립하는 것은,

상기 네트워크 디바이스와 연관된 상기 NVMe 큐 페어를 상기 원격 디바이스와 연관된 상기 RDMA 큐 페어에 매핑하고,

관리자 전송 큐 기반 어드레스(admin submission queue base address) 및 완료 큐 기반 어드레스(completion queue base address)를 RDMA 관리 큐 페어(RDMA admin queue pair)에 매핑하는 것을 포함하는 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

청구항 3

제1항에 있어서,

상기 원격 디바이스를 사용하여, 상기 네트워크 디바이스에 NVMe 명령을 전송하고,

상기 NVMe 컨트롤러에 의해, 상기 NVMe 명령을 인출하고, 상기 NVMe 명령은 큐 페어 수신 완료 이벤트(queue pair receive completion event)에 대한 참조(reference)를 포함하고,

상기 NVMe 컨트롤러를 사용하여 상기 NVMe 명령을 디코딩하고,

상기 NVMe 컨트롤러를 사용하여 상기 큐 페어 수신 완료 이벤트를 트리거(trigger)하는 것을 더 포함하는 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

청구항 4

제3항에 있어서,

상기 NVMe 명령은 플러시(flush), 라이트(write), 리드(read), 수정 불가능한 라이트(write uncorrectable) 및 비교 인스트럭션(instruction) 중 적어도 하나를 포함하는 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

청구항 5

제3항에 있어서,

상기 NVMe 명령은 PRP 필드(Physical Region Page field)를 더 포함하고,

상기 PRP 필드는 상기 네트워크 디바이스에 송신될 데이터 신원(data identity)을 포함하는 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

청구항 6

제1항에 있어서,

상기 NVMe 컨트롤러를 초기화하는 것은,

적어도 하나 이상의 프로세서를 부트 업(boot up)하고,

펌웨어(firmware)를 호출하고,

상기 펌웨어로 SFR(Special Function Register)을 포함하는 하나 이상의 시스템 모듈, 메모리 유닛을 초기화하고,

RNIC(RDMA Network Interface Card)를 초기화하는 것을 포함하는 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

청구항 7

삭제

청구항 8

삭제

청구항 9

제1항에 있어서,

상기 컨트롤러를 초기화하는 것은 PCIe(Peripheral Component Interconnect Express) 기반의 RNIC(RDMA Network Interface Card)를 초기화하는 것을 포함하고,

상기 네트워크 디바이스와 상기 원격 디바이스는, 상기 네트워크 디바이스 및 상기 원격 디바이스의 큐 페어들을 매핑하고 상기 네트워크 디바이스 및 상기 원격 디바이스 모두에 대해 동일한 메모리를 등록(register)함으로써 서로 연결되는 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

청구항 10

제1항에 있어서,

상기 원격 디바이스로 전송되는 완료 지시 필드(completion indication field)의 NVMe 명령의 상태를 지시(indicate)하는 것을 더 포함하고,

상기 완료 지시는, 상기 수신된 상기 스토리지 디바이스에 접근 요청이 상기 NVMe 컨트롤러에 의해 실행됨을 지시하는 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

청구항 11

삭제

청구항 12

서버에 연결된 네트워크 디바이스, 상기 네트워크 디바이스는 NVMe(Non-Volatile Memory Express) 컨트롤러를 포함하고,

상기 네트워크 디바이스는,

원격 디바이스 검색 프로세스를 처리하기 위한 NVMe 큐 페어(NVMe queue pair)를 설정하고,상기 NVMe 컨트롤러

에 의해 제어되는 스토리지 디바이스를 접근하기 위한 요청을 상기 원격 디바이스로부터 수신하고,

상기 원격 디바이스를 찾기 위한 상기 검색 프로세스를 초기화하고,

상기 원격 디바이스가 검색되면, 상기 NVMe 큐 페어를 RDMA 큐 페어(Remote Direct Memory Access queue pair)에 매핑함으로써 상기 원격 디바이스와의 접속을 수립하는 것을 포함하는 네트워크를 통해 원격으로 스토리지 디바이스에 대한 접근을 제공하는 스토리지 디바이스 접근 시스템.

청구항 13

삭제

청구항 14

네트워크 디바이스와 연관된 NVMe(Non-Volatile Memory Express) 컨트롤러, RDMA(Remote Direct Memory Access) 네트워크 인터페이스 카드(RNIC)와 연관된 적어도 하나 이상의 프로세서를 사용하여 초기화하고, 상기 초기화는 상기 RNIC를 서버로 설정하는 것을 포함하고,

적어도 하나의 프로세서를 사용하여, 네트워크 검색 프로세서를 처리하기 위한 마스터 큐 페어(master queue pair)를 생성하고,

적어도 하나의 프로세서를 사용하여, 네트워크를 통해 적어도 하나 이상의 원격 디바이스로부터 들어오는 연결 요청을 수신하고, 상기 연결 요청은 상기 NVMe 컨트롤러와 연관된 스토리지 디바이스에 대한 연결하기 위한 요청이고,

적어도 하나의 프로세서를 사용하여, 바인딩(binding)을 통해 상기 네트워크 디바이스와 상기 원격 디바이스를 연결하고,

적어도 하나의 프로세서를 사용하여, 상기 원격 디바이스로부터 상기 스토리지 디바이스로 들어오는 데이터 작업을 수신하고,

적어도 하나의 프로세서를 사용하여, 상기 스토리지 디바이스에서 상기 수신된 데이터 작업을 처리하는 것을 포함하는 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 일반적으로 네트워크 상의 데이터 스토리지에 관한 것이고, 더욱 구체적으로는 RDMA(Remote Direct Memory Access)로 인에이블(enable)되는 RNIC(RDMA-enabled Network Interface Controller) 및 컨트롤러/디바이스 측의 NVMe(Non-Volatile Memory Express)를 통합하기 위한 방법에 관한 것이다.

배 경 기 술

[0002] 종래의 컴퓨팅 장치의 구성에서, 클라이언트 및 서버는 각각, RDMA 프로토콜을 이용하여 서로 통신할 수 있는 NIC(network interface controller)들 또는 NW(network) I/O(input/output) 디바이스들을 포함한다. 서버는, 서버의 운영 체제 및 관련 드라이버를 실행시키는 호스트 프로세서를 포함한다. 서버는 또한 서버에 의해 또는 서버 측에서 유지되는 스토리지에 대한 접근을 관리하는 스토리지 컨트롤러를 포함한다. 클라이언트의 NW I/O 디바이스는, 서버에 의해 유지되는 스토리지에 데이터를 라이트(write)하거나, 상기 스토리지로부터 데이터를 리드(read)하기 위해, 서버의 NW I/O 디바이스에 요청을 발행한다. 서버의 운영 체제, 관련 드라이버 및 호스트 프로세서는 서버의 NW I/O 디바이스에 의해 수신한 요청을 처리하고, 스토리지 컨트롤러에 이에 대응 요청을 발행한다. 스토리지 컨트롤러는 이들 대응 요청을 수신 및 실행한다. 대응 요청을 실행한 후, 스토리지 컨트롤러는 요청 완료 정보(request completion information)(및 데이터가 스토리지로부터 리드된 경우 관련 데이터)를 서버의 운영 체제 및 관련 드라이버에 발행한다. 이로부터, 서버의 운영 체제, 관련 드라이버 및 호스트 프로세서는 대응 요청 완료 정보(corresponding request completion information) 및 관련 데이터를 생성하고, 대응 요청 완료 정보 및 관련 데이터를 서버의 NW I/O 디바이스에 발행한다. 이후 서버의 NW I/O 디바이스는 대응 요청 완료 정보 및 관련 데이터를 클라이언트의 NW I/O 디바이스에 발행한다.

[0003] NVMe 포럼(NVMe forum)은 현재 NVMe 프로토콜에 대해 서로 다른 전송을 대상으로 하는 패브릭 상의 NVMe에 관한

명세(NVMe over fabrics specification)의 표준화를 추진 중이다. 포럼은 RDMA 및 NVMe를 사용하는 네트워크 상에서 레이턴시(latency)가 적은 데이터 전송을 목표로 하고 있다. 네트워크 상에서 보다 빠른 데이터 스토리지에 대한 요구가 증가함에 따라, RDMA를 이용하는 저(low) 레이턴시 NVMe 기반의 SSD(solid-state drive)의 사용에 대한 기대가 높다.

[0004] 현재 기술에 따르면, 명령 및 데이터가 RNIC로부터 NVMe 메모리로 복사되거나 NVMe 메모리로부터 RNIC로 복사되는 경우 메모리는 별개로 관리된다. 각각의 모듈 레지스터는 자신의 인터럽트를 소유하고, 호스트 자원을 중복적으로 사용한다. 현재 방식에 따른 문제들은 호환 불가능한 인터페이스, 중복성의 증가, 변환 모듈 비용의 증가, 비효율적인 메모리 사용 및 다중 인터럽트 핸들러를 포함한다.

[0005] 이와 같은 관점에서, 초기화 디바이스(initiator device) 및 컨트롤러/디바이스 사이에서 원활한(seamless) 데이터 전송을 위해 네트워크 상에서 데이터를 처리하는 방법 및 시스템이 요구된다.

[0006] 앞서 언급한 결점, 단점 및 문제점은 본 명세서에서 다루어지고, 이어지는 명세서를 통해 이해될 수 있을 것이다.

발명의 내용

해결하려는 과제

[0007] 본 발명이 해결하고자 하는 기술적 과제는 RDMA NVMe 디바이스를 구현하기 위한 방법을 제공하는 것이다. 본 발명의 다양한 실시예에 따르면, NVMe SSD는, 네트워크 상의 블록 디바이스 데이터 스트로지를 지원하기 위한 RDMA를 채용하여 구현된다. 본 발명의 다양한 실시예에 따르면, 사용자로 하여금 네트워크 상에서 원격으로 디바이스를 접근하고, 로컬 머신에 접속된 것처럼 모든 디스크 오퍼레이션을 수행하는 것을 가능하게 한다.

[0008] 본 발명의 기술적 과제들은 이상에서 언급한 기술적 과제로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0009] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시예에 따른 네트워크를 통해 원격으로 스토리지 디바이스에 접근을 가능하게 하는 스토리지 디바이스 접근 방법은, 서버에 연결된 네트워크 디바이스로, NVMe(Non-Volatile Memory Express) 컨트롤러를 초기화하고, 네트워크 디바이스로, 원격 디바이스 검색 프로세스를 처리하기 위한 NVMe 큐 페어(NVMe queue pair)를 설정하고, 네트워크 디바이스 측에서, 서버 측에서 유지되는 NVMe 컨트롤러에 의해 제어되는 스토리지 디바이스를 접근하기 위한 요청을 원격 디바이스로부터 수신하고, 네트워크 디바이스로, 원격 디바이스를 찾기 위한 검색 프로세스를 초기화하고, 원격 디바이스가 검색되면, 네트워크 디바이스로, NVMe 큐 페어를 RDMA 큐 페어(Remote Direct Memory Access queue pair)에 매핑함으로써 원격 디바이스와의 접속을 수립하는 것을 포함한다.

[0010] 본 발명의 몇몇의 실시예에서, 상기 원격 디바이스와의 접속을 수립하는 것은, 상기 네트워크 디바이스와 연관된 상기 NVMe 큐 페어를 상기 원격 디바이스와 연관된 상기 RDMA 큐 페어에 매핑하고, 관리자 전송 큐 기반 어드레스(admin submission queue base address) 및 완료 큐 기반 어드레스(completion queue base address)를 RDMA 관리 큐 페어(RDMA admin queue pair)에 매핑하는 것을 포함할 수 있다.

[0011] 본 발명의 몇몇의 실시예에서, 상기 방법은, 상기 원격 디바이스로, 상기 네트워크 디바이스에 NVMe 명령을 전송하고, 상기 NVMe 컨트롤러로, 상기 NVMe 명령을 인출하고, 상기 NVMe 명령은 큐 페어 수신 완료 이벤트(queue pair receive completion event)에 대한 참조(reference)를 포함하고, 상기 NVMe 컨트롤러로 상기 NVMe 명령을 디코딩하고, 상기 큐 페어 수신 완료 이벤트를 트리거(trigger)하는 것을 더 포함할 수 있다.

[0012] 본 발명의 몇몇의 실시예에서, 상기 NVMe 명령은 플러시(flush), 라이트(write), 리드(read), 수정 불가능한 라이트(write uncorrectable) 및 비교 인스트럭션(instruction) 중 하나를 포함할 수 있다.

[0013] 본 발명의 몇몇의 실시예에서, 상기 NVMe 명령은 PRP 필드(Physical Region Page field)를 더 포함하고, 상기 PRP 필드는 상기 네트워크 디바이스에 송신될 데이터 신원(data identity)을 포함할 수 있다.

[0014] 본 발명의 몇몇의 실시예에서, 상기 NVMe 컨트롤러를 초기화하는 것은, 프로세서를 부트 업(boot up)하고, 펌웨어(firmware)를 호출하고, 상기 펌웨어로 SFR(Special Function Register)을 포함하는 하나 이상의 시스템 모듈, 메모리 유닛을 초기화하고, RNIC(RDMA Network Interface Card)를 초기화하는 것을 포함할 수 있다.

[0015] 본 발명의 몇몇의 실시예에서, 상기 네트워크 디바이스에 접속된 상기 스토리지 디바이스는 RDMA(Remote Direct Memory Access)를 이용하는 NVMe 스토리지 프로토콜(NVMe storage protocol)로 구현될 수 있다.

[0016] 본 발명의 몇몇의 실시예에서, 상기 NVMe 스토리지 프로토콜은 iWARP(Internet Wide Area RDMA protocol), 인피니밴드(Infiniband) 및 RoCE(RDMA over Converged Ethernet) 중 하나를 포함할 수 있다.

[0017] 본 발명의 몇몇의 실시예에서, PCIe(Peripheral Component Interconnect Express) 기반의 RNIC(RDMA Network Interface Card) 및 상기 네트워크 디바이스는, 상기 네트워크 디바이스 및 상기 원격 디바이스의 큐 페어들을 매핑하고 상기 네트워크 디바이스 및 상기 원격 디바이스 모두에 대해 동일한 메모리를 등록(register)함으로써 서로 연결될 수 있다.

[0018] 본 발명의 몇몇의 실시예에서, 상기 방법은, 상기 원격 디바이스에 완료 지시 필드(completion indication field)의 상기 NVMe 명령의 상태를 지시(indicate)하는 것을 더 포함하고, 상기 완료 지시 필드는, 수신된 상기 스토리지 디바이스를 접근하기 위한 상기 명령이 상기 NVMe 컨트롤러에 의해 실행됨을 지시할 수 있다.

[0019] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시예에 따른 네트워크를 통해 원격으로 스토리지 디바이스에 대한 접근을 제공하는 스토리지 디바이스 접근 시스템은, 원격 디바이스; NVMe(Non-Volatile Memory Express) 컨트롤러; 및 서버에 연결된 네트워크 디바이스를 포함하고, 네트워크 디바이스는, 원격 디바이스 검색 프로세스를 처리하기 위한 NVMe 큐 페어(NVMe queue pair)를 설정하고, 서버 측에서 유지되는 NVMe 컨트롤러에 의해 제어되는 스토리지 디바이스를 접근하기 위한 요청을 원격 디바이스로부터 수신하고, 원격 디바이스를 찾기 위한 검색 프로세스를 초기화하고, 원격 디바이스가 검색되면, NVMe 큐 페어를 RDMA 큐 페어(Remote Direct Memory Access queue pair)에 매핑함으로써 원격 디바이스와의 접속을 수립한다.

[0020] 본 발명의 몇몇의 실시예에서, 상기 NVMe 컨트롤러를 초기화하는 것은, 프로세서를 부트 업(boot up)하고, 펌웨어(firmware)를 호출하고, 상기 펌웨어로 SFR(Special Function Register)을 포함하는 하나 이상의 시스템 모듈, 메모리 유닛을 초기화하고, RNIC(RDMA Network Interface Card)를 초기화하는 것을 포함할 수 있다.

[0021] 기타 실시예들의 구체적인 사항들은 상세한 설명 및 도면들에 포함되어 있다.

도면의 간단한 설명

[0022] 도 1은 종래 기술에 따른 인터페이스 메커니즘을 설명하기 위한 개략도이다.

도 2는 본 발명의 일 실시예에 따른 네트워크 상의 원격 디바이스 데이터 스토리지에 대한 접근을 제공하는 시스템을 설명하기 위한 블록도이다.

도 3은 본 발명의 일 실시예에 따른 RDMA NVMe 서브 시스템을 설명하기 위한 블록도이다.

도 4는 본 발명의 일 실시예에 따른 RDMA NVMe 디바이스의 인터페이싱을 설명하기 위한 블록도이다.

도 5는 본 발명의 다른 실시예에 따른 RDMA NVMe 디바이스의 인터페이싱을 설명하기 위한 블록도이다.

도 6은 본 발명의 일 실시예에 따른 RDMA NVMe 디바이스를 이용하여 네트워크를 통해 원격으로 스토리지 디바이스에 대한 접근을 가능하게 하는 방법을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0023] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 것이며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 도면에서 층 및 영역들의 상대적인 크기는 설명의 명료성을 위해 과장된 것일 수 있다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[0024] 하나의 소자(elements)가 다른 소자와 "접속된(connected to)" 또는 "커플링된(coupled to)" 이라고 지칭되는 것은, 다른 소자와 직접 연결 또는 커플링된 경우 또는 중간에 다른 소자를 개재한 경우를 모두 포함한다. 반면, 하나의 소자가 다른 소자와 "직접 접속된(directly connected to)" 또는 "직접 커플링된(directly coupled to)"으로 지칭되는 것은 중간에 다른 소자를 개재하지 않은 것을 나타낸다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다. "및/또는"은 언급된 아이템들의 각각 및 하나 이상의 모든 조합을 포

함한다.

[0025] 소자(elements) 또는 층이 다른 소자 또는 층의 "위(on)" 또는 "상(on)"으로 지칭되는 것은 다른 소자 또는 층의 바로 위뿐만 아니라 중간에 다른 층 또는 다른 소자를 개재한 경우를 모두 포함한다. 반면, 소자가 "직접 위(directly on)" 또는 "바로 위"로 지칭되는 것은 중간에 다른 소자 또는 층을 개재하지 않은 것을 나타낸다.

[0026] 비록 제1, 제2 등이 다양한 소자, 구성요소 및/또는 섹션들을 서술하기 위해서 사용되나, 이들 소자, 구성요소 및/또는 섹션들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 소자, 구성요소 또는 섹션들을 다른 소자, 구성요소 또는 섹션들과 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 소자, 제1 구성요소 또는 제1 섹션은 본 발명의 기술적 사상 내에서 제2 소자, 제2 구성요소 또는 제2 섹션일 수도 있음은 물론이다.

[0027] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성요소, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0028] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0029] 도 1은 종래 기술에 따른 인터페이스 메커니즘을 설명하기 위한 개략도이다.

[0030] 도 2는 본 발명의 일 실시예에 따른 네트워크 상의 원격 디바이스 데이터 스토리지를 지원하기 위해 RDMA를 채용한 NVMe SSD를 설명하기 위한 블록도이다.

[0031] 본 시스템은 원격 디바이스(201), 네트워크(202), 서버(203), 서버(203)에 연결된 네트워크 디바이스(204) 및 NVMe 컨트롤러(205) 및 스토리지 디바이스(206)를 포함한다.

[0032] 네트워크(204)는 원격 디바이스(201), 서버(203), 네트워크 디바이스(204) 및 스토리지 디바이스(206) 사이의 임의의 수단에 의해 데이터를 전송하도록 하는 임의의 네트워크일 수 있다. 일반적으로, 데이터는 이미지, 비디오, 음악 및 문서의 형태일 수 있으나, 본 발명의 범위가 이에 제한되는 것은 아니다. 네트워크(204)는 개인용 또는 공용이거나, 유선 또는 무선이거나, 전체 또는 부분 네트워크일 수 있다.

[0033] 서버(203)는 네트워크(204)를 통해 다른 컴퓨팅 시스템에 서비스를 제공하는 임의의 컴퓨터 시스템 또는 그 균등물일 수 있다. 일반적으로, 서버(203)는 정보를 관리하고, 및/또는 정보의 검증을 위한 메커니즘을 구비한다. 본 시스템에서, 서버(203)는 정보를 관리하고, 및/또는 정보의 검증을 위한 메커니즘을 구비한다. 본 발명의 일 실시예에서, 서버(203)는 네트워크(202)에 접속된 디바이스의 신원을 제공한다.

[0034] 본 발명의 다른 실시예에서, 서버(203)는 디바이스에 의해 네트워크(202)로 전송된 자격 증명(credential)을, 네크워크(202)에 연결된 다른 디바이스에 의해 전송된 자격 증명으로 유효화한다. 본 발명의 또 다른 실시예에서, 서버(203)는 데이터를 저장하고, 네트워크(130)를 통한 저장된 데이터에 대한 접근을 관리한다.

[0035] 원격 디바이스(201)는 개인용 컴퓨터, 디지털 정지 영상 카메라, 디지털 비디오 카메라, 텔레비전 또는 디지털 읽기/표시 디바이스와 같은 네트워크(202)에 통신 가능하도록 연결된 디지털 라이프스타일 디바이스일 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다. 일반적으로, 원격 디바이스(201)는 원격 접근 정보(remote access information)를 생성하여 네트워크(202)를 통해 네트워크 디바이스(204)에 전송한다.

[0036] 스토리지 디바이스(206)는 HDD(hard disk drive) 및 SSD(solid state drive) 중 하나를 포함하고, SSD는 3차원 크로스 포인트 메모리(3-dimensional cross-point memory), 플래시 메모리(flash memory), 강유전체 메모리(ferroelectric memory), SONOS 메모리(silicon-oxide-nitride-oxide-silicon memory), 폴리머 메모리(polymer memory), 나노 와이어(nanowire), 강유전체 트랜지스터 RAM(ferroelectric transistor random access memory, FeTRAM 또는 FeRAM), 나노 와이어 및 EEPROM(electrically erasable programmable read-only memory) 중 적어도 하나를 포함하는 비휘발성 메모리를 포함한다. 네트워크 디바이스(204)에 접속된 스토리지 디바이스(206)는 RDMA를 이용하는 NVMe 스토리지 프로토콜로 구현된다. 여기서 NVMe 스토리지 프로토콜은 iWARP(Internet Wide Area RDMA protocol), 인피니밴드(Infiniband) 및 RoCE(RDMA over Converged Ethernet)

중 하나를 포함한다.

[0037] 도 3은 본 발명의 일 실시예에 따른 RDMA NVMe 서브 시스템을 설명하기 위한 블록도이다.

[0038] RDMA 및 NVMe 컨트롤러를 효율적으로 통합하기 위한 빌딩 블록을 포함하는 인터페이스는 여기서 설명된다. RDMA 큐 페어(RDMA queue pair)는 NVMe 큐로서 등록될 것이다. NVMe는 RDMA로 인에이블(enable)되는 RNIC(RDMA-enabled Network Interface Controller)에 의해 사용되는 동일한 DMA(Direct Memory Access)를 사용하고, 데이터를 직접 배치시킨다. 내부 RAM은 NSMR(Non-Shared Memory Region)로서 RDMA로 등록되고, NVMe 명령을 포함하는 RDMA_SEND 요청을 처리하고, RDMA_SEND를 이용한 응답을 포스팅(posting)하기 위해 사용된다. 또한, 내부 RAM은 RDMA_WRITE 및 RDMA_READ_RES에 사용되는 직접 데이터 배치(direct data placement)를 위해 RDMA로 등록된다. 초기화부(initiator)로부터의 데이터는 펌웨어(firmware, F/W)의 RAM으로 직접 배치되어 각각의 명령에 따라 처리된다.

[0039] 도 4는 본 발명의 일 실시예에 따른 RDMA NVMe 디바이스의 인터페이싱을 설명하기 위한 블록도이다.

[0040] 하나의 완전한 리드(read) 및 라이트(write) 트랜잭션을 처리하기 위한 원격 디바이스와 네트워크 디바이스 사이의 접속을 수립하는 것으로부터 여러 단계를 포함하는 메커니즘에 속한 과정은 다음과 같다. 네트워크 디바이스의 전원이 켜지면, 프로세서를 부트 업(boot up)하고 펌웨어를 호출하고, 펌웨어로 SFR(Special Function Register)을 포함하는 모든 모듈, 메모리 유닛 등을 초기화하는 것과, 최근의 RNIC 초기화를 포함하는 NVMe 컨트롤러 초기화가 수행된다. 초기화 동안, RNIC는 자신을 서버로 설정한다. 또한, 네크워크 디바이스는 디폴트로 검색 프로세스를 처리하기 위한 마스터 큐 페어(master queue pair)를 설정한다. 마스터 큐 페어를 설정하는 동안, 네트워크 디바이스는 임의의 원격 디바이스로부터 인입되는 접속 요청들을 리스닝(listen)한다.

[0041] 또한, 네트워크 디바이스는 원격 디바이스를 찾기 위한 검색 프로세스를 초기화한다. 원격 디바이스가 검색되면, 네트워크 디바이스는 상기 원격 디바이스와 자신을 바인딩(bind)하여, 접속이 수립된다. 접속 수립은 양 단에서의 마스터 큐 페어가 설정되었음을 의미한다. 마스터 큐 페어는 NVMe에 대한 관리자 큐(admin queue)와 매핑되고, 네트워크 디바이스에서의 마스터 큐 페어는, 관리자(Admin) 명령이 상기 큐 페어를 이용하여 네트워크 디바이스에 의해 요청되고 디폴트 NSMR이 상기 큐 페어에 대해 배타적으로 등록되는 속성을 갖는다.

[0042] 또한, 네트워크 디바이스는 NVMe 식별 명령을 생성하고, RDMA_SEND를 포스팅한다. NVMe 명령은 디코딩될 것이며, 디코딩은 NVMe 컨트롤러에 의해 명령을 처리하는 것에 불과하다. NVMe 명령은 큐 페어 수신 완료 이벤트(queue pair receive completion event)에 대한 참조(reference)로서 인출된다. NVMe 명령을 인출하는 동안, 큐 테일 포인터(queue tail pointer)가 항상 큐 페어 수신 완료 이벤트와 동기화됨에 따라, 명령은 NVMe 컨트롤러에 의해 직접 리드된다. 완료 이벤트는 펌웨어가 링 도어벨 동작(ring doorbell operation)과 유사한 방식으로 동작하도록 트리거(trigger)한다. 데이터 이동에 대해, 데이터는 특정 동작을 위해 등록된 RDMA SGL(RDMA Scatter/Gather Lists)로 직접적으로 DMA된다.

[0043] 본 발명에 따르면, 명령 캡슐(command capsule)은 또한 호스트에 등록된 공용 메모리 영역(host registered shared memory region)인 PRP 필드(Physical Region Page field)를 포함한다. 다음으로 식별 데이터는 펌웨어로부터 직접 리드되어 RDMA SGL로 DMA되어, 네트워크 디바이스에 전송될 것이다. 상기 단계들은 원격 디바이스로부터 네트워크 디바이스로 데이터 전송을 수행하는 임의의 명령에 대해서 동일하다. 원격 디바이스로부터 스토리지 디바이스로 데이터를 전송하는 경우, 각각의 라이트 명령은 NVMe 컨트롤러에 의해 디코딩된다. 펌웨어는 RDMA로 데이터를 저장하기를 원하는 어드레스를 등록할 것이다. 또한, SGL은 각각의 명령에 대해 등록된 어드레스로 형성되고 네트워크 디바이스로 전송될 것이다. 수신하는 동안, 네트워크 디바이스는 그것을 요청된 SGL에 어드레싱하고 데이터 패킷을 형성하고 그것을 포스팅할 것이다. 각각의 RDMA 패킷이 수신되면, RNIC는 데이터를 펌웨어의 버퍼에 직접 배치할 것이다. 수신 완료 이벤트가 RNIC에 의해 트리거되면 펌웨어는 통지를 받을 것이다.

[0044] 도 5는 본 발명의 다른 실시예에 따른 RDMA NVMe 디바이스의 인터페이싱을 설명하기 위한 블록도이다.

[0045] 스토리지 디바이스는 네트워크를 통해 접속되고 RDMA를 사용하는 NVMe 스토리지 프로토콜로 구현된다. 네트워크 프로토콜은 iWARP, 인피니밴드, RoCE 등을 포함한다. 또한, NVMe 큐는 RDMA 큐 페어와 매핑되고, 관리자 전송 큐 기반 어드레스(admin submission queue base address) 및 완료 큐 기반 어드레스(completion queue base address)는 RDMA 관리 큐 페어(RDMA admin queue pair)에 매핑된다. 다중 IO 큐의 경우도 또한 이와 유사하다. 임의의 큐 페어에 대한 완료 이벤트는 NVMe 컨트롤러가 요청된 동작을 처리하도록 구현된다. 테일 도어벨 업데이트(Tail doorbell update)가 필요하지 않다. 컨트롤러로부터 호스트로의 데이터 이동은 다른 메모리 영역으로

서 RDMA로 등록된 메모리를 통해 일어날 것이다.

[0046] 본 발명의 다른 실시예에 따르면, PCIe(Peripheral Component Interconnect Express) 기반의 NVMe 컨트롤러에 의해 제공되는 표준 인터페이스는 RNIC와 효율적으로 인터페이싱하기 위해 변형될 필요가 있다.

[0047] 본 발명의 일 실시예에 따르면, 기존의 PCIe 기반의 RNIC 및 NVMe 디바이스는 함께 연결되어, 큐 페어를 매핑하고, 양 드라이버를 위한 동일한 메모리를 등록함으로써 더 적은 메모리를 사용한다. NVMe 컨트롤러 동작을 호출하기 위해 RNIC 인터럽트를 사용함으로써 양 디바이스에 등록된 인터럽트는 감소된다.

[0048] 도 6은 본 발명의 일 실시예에 따른 RDMA NVMe 디바이스를 이용하여 네트워크를 통해 원격으로 스토리지 디바이스에 대한 접근을 가능하게 하는 방법을 설명하기 위한 순서도이다.

[0049] 단계(602)에서, 서버에 연결된 네트워크 디바이스로 NVMe 컨트롤러를 초기화한다. 단계(604)에서, 네트워크 디바이스는 원격 디바이스 검색 프로세스를 처리하기 위한 NVMe 큐 페어(NVMe queue pair)를 설정한다. 단계(606)에서, 네트워크 디바이스는 서버 측에서 유지되는 NVMe 컨트롤러에 의해 제어되는 스토리지 디바이스를 접근하기 위한 요청을 원격 디바이스로부터 수신한다. 단계(608)에서, 네트워크 디바이스는 원격 디바이스를 찾기 위한 검색 프로세스를 초기화한다. 단계(610)에서, 원격 디바이스가 검색되면, 네트워크 디바이스는 NVMe 큐 페어를 RDMA 큐 페어(Remote Direct Memory Access queue pair)에 매핑함으로써 상기 원격 디바이스와의 접속을 수립한다.

[0050] 본 발명의 다양한 실시예에 따르면, NVMe의 낮은 레이턴시와 RDMA의 속도가 결합되어 빠른 네트워크 스토리지 솔루션, 네트워크 기능이 구비된 스탠드얼론 SSD, 및 스토리지의 휴대성 및 확장성을 제공한다.

[0051] 본 발명은 특정 실시예를 참조하여 설명되었으나, 본 발명의 다양한 실시예의 더 넓은 사상 및 범위를 벗어나지 않고 이들 실시예들에 다양한 수정 및 변형이 이루어질 수 있음은 해당 기술 분야의 통상의 기술자에게 있어서 명백하다. 또한, 본 명세서에서 설명된 다양한 디바이스, 모듈 등은 하드웨어 회로, 예컨대, CMOS(complementary metal oxide semiconductor) 기반의 로직 회로, 펌웨어, 소프트웨어 및/또는 하드웨어, 펌웨어 및/또는 기계로 판독 가능한 매체에 포함된 소프트웨어의 임의의 조합을 이용하여 구현되고 동작할 수 있다. 예를 들어, 다양한 전기적 구조 및 방법은 트랜지스터, 로직 게이트, 및 ASIC(application specific integrated circuit)과 같은 전기 회로를 이용하여 구현될 수 있다.

[0052] 이상 첨부된 도면을 참조하여 본 발명의 실시예들을 설명하였으나, 본 발명은 상기 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 제조될 수 있으며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

부호의 설명

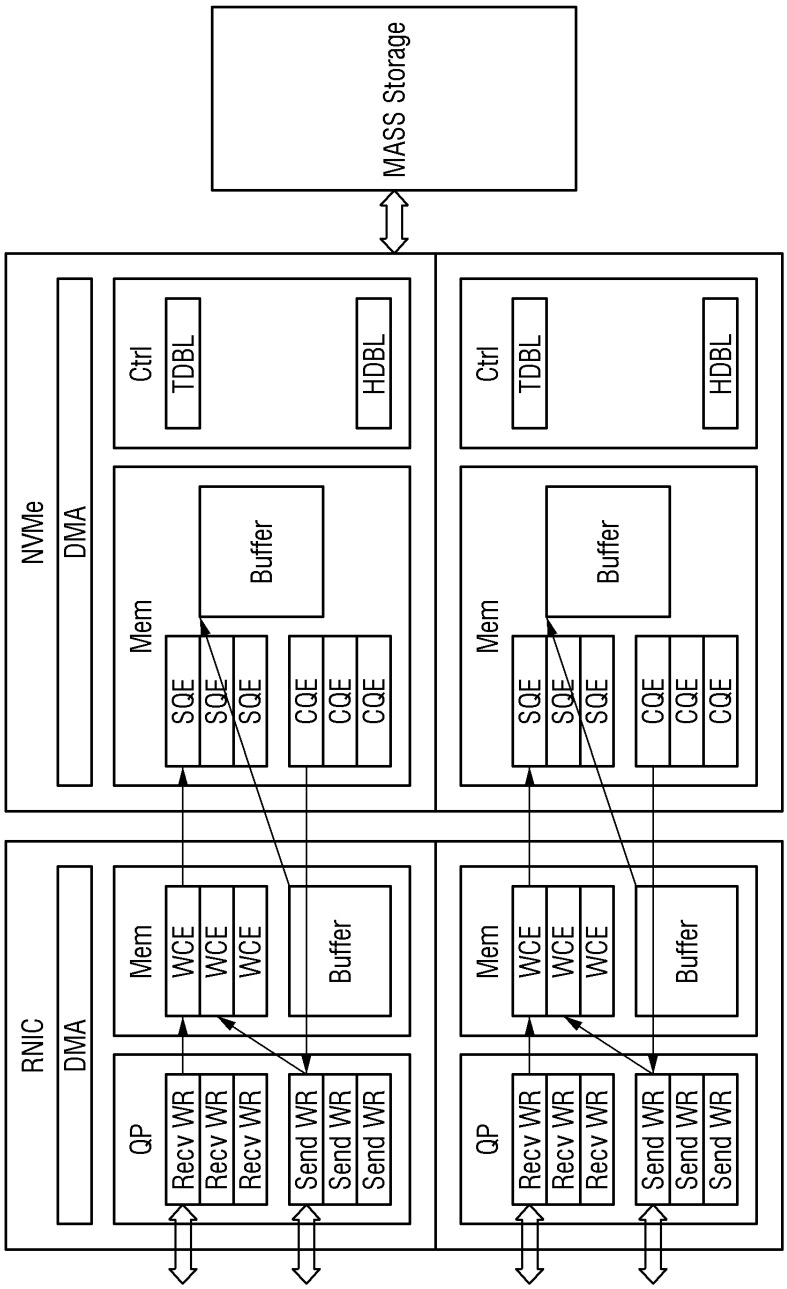
[0053] 201: 원격 디바이스 202: 네트워크

203: 서버 204: 네트워크 디바이스

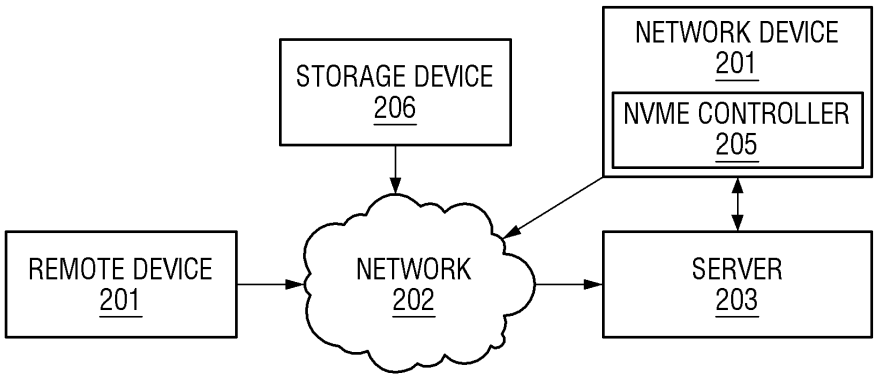
205: NVMe 컨트롤러 206: 스토리지 디바이스

도면

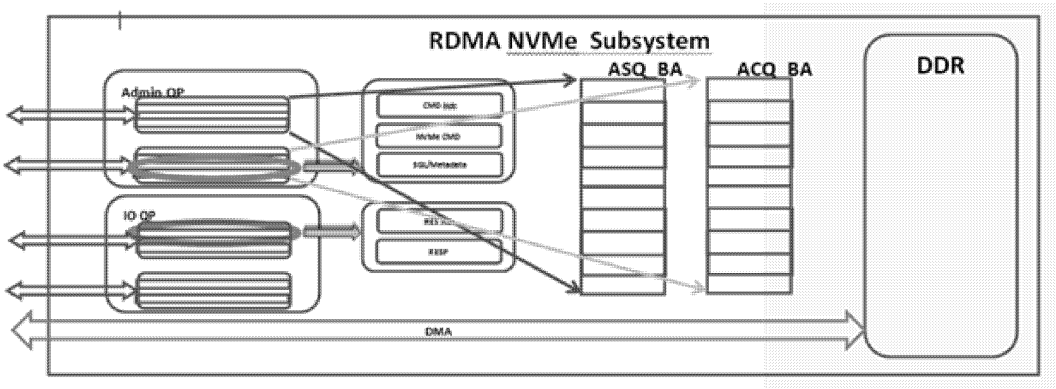
도면1


MASS Storage
NVMe
DMA
Ctrl
TDBL
HDBL
Mem
Buffer
SQE
SQE
SQE
CQE
CQE
CQE
RNIC
DMA
Mem
WCE
WCE
WCE
Buffer
QP
Recv WR
Recv WR
Recv WR
Send WR
Send WR
Send WR

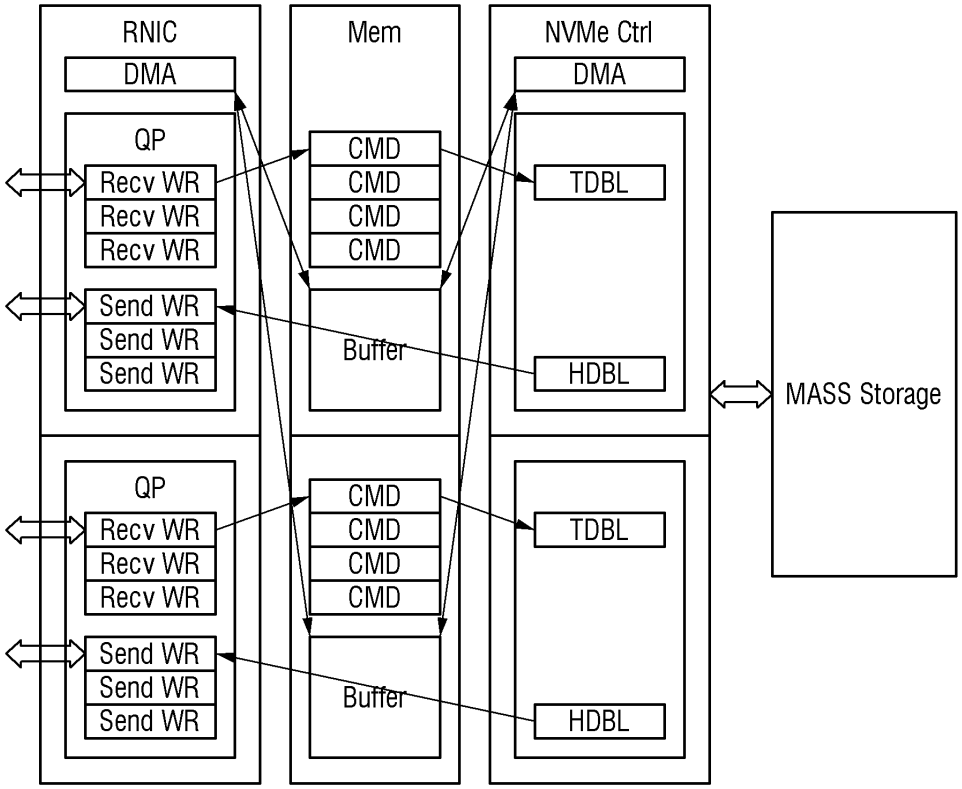
도면2


STORAGE DEVICE
206
NETWORK DEVICE
201
NVME CONTROLLER
205
REMOTE DEVICE
201
NETWORK
202
SERVER
203

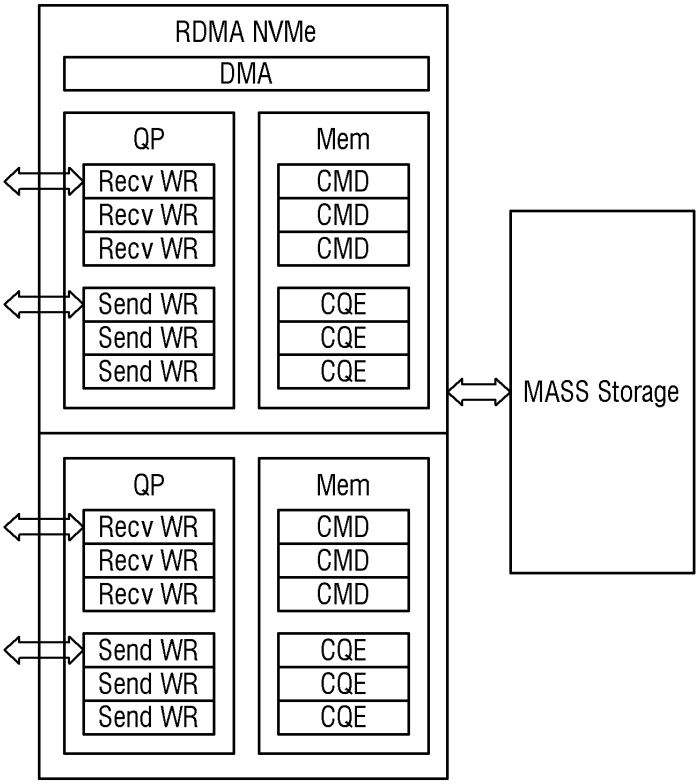
도면3


RDMA NVMe Subsystem
ASQ BA
ACQ BA
DDR
Admin QP
IO QP
DMA

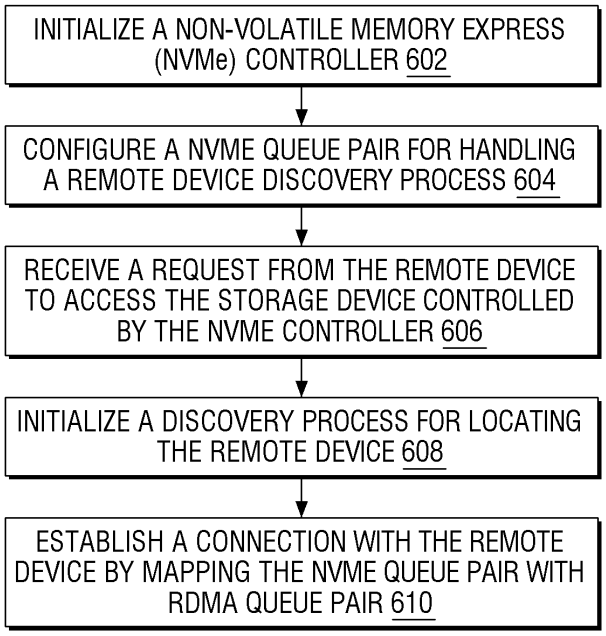
도면4


RNIC
DMA
QP
Recv WR
Recv WR
Recv WR
Send WR
Send WR
Send WR
QP
Recv WR
Recv WR
Recv WR
Send WR
Send WR
Send WR
Mem
CMD
CMD
CMD
CMD
Buffer
CMD
CMD
CMD
CMD
Buffer
NVMe Ctrl
DMA
TDBL
HDBL
TDBL
HDBL
MASS Storage

도면5


RDMA NVMe
DMA
QP
Recv WR
Recv WR
Recv WR
Send WR
Send WR
Send WR
Mem
CMD
CMD
CMD
CQE
CQE
CQE
QP
Recv WR
Recv WR
Recv WR
Send WR
Send WR
Send WR
Mem
CMD
CMD
CMD
CQE
CQE
CQE
MASS Storage

도면6


INITIALIZE A NON-VOLATILE MEMORY EXPRESS (NVMe) CONTROLLER 602
CONFIGURE A NVME QUEUE PAIR FOR HANDLING A REMOTE DEVICE DISCOVERY PROCESS 604
RECEIVE A REQUEST FROM THE REMOTE DEVICE TO ACCESS THE STORAGE DEVICE CONTROLLED BY THE NVME CONTROLLER 606
INITIALIZE A DISCOVERY PROCESS FOR LOCATING THE REMOTE DEVICE 608
ESTABLISH A CONNECTION WITH THE REMOTE DEVICE BY MAPPING THE NVME QUEUE PAIR WITH RDMA QUEUE PAIR 610