(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년11월28일
(11) 등록번호 10-2471713
(24) 등록일자 2022년11월23일

(51) 국제특허분류(Int. Cl.)
G06F 3/06 (2006.01) *H05K 7/18* (2006.01)
(52) CPC특허분류
G06F 3/0688 (2013.01)
G06F 3/0658 (2013.01)
(21) 출원번호 10-2018-0141180
(22) 출원일자 2018년11월15일
심사청구일자 2021년11월15일
(65) 공개번호 10-2019-0105490
(43) 공개일자 2019년09월17일
(30) 우선권주장
62/638,722 2018년03월05일 미국(US)
15/981,801 2018년05월16일 미국(US)
(56) 선행기술조사문헌
한국공개특허 제10-2018-0012190호(2018.02.05.) 1부.*
한국공개특허 제10-2018-0012181호(2018.02.05.) 1부.*
한국공개특허 제10-2018-0012201호(2018.02.05.) 1부.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
올라이그, 솜퐁 폴
미국 캘리포니아주 94566 플레젠튼 파세오 그라나다 3050
울리, 프레드
미국 캘리포니아주 95129 산호세 그린 드라이브 1471
(74) 대리인
특허법인 고려

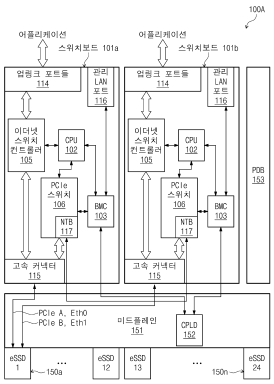
전체 청구항 수 : 총 15 항 심사관 : 김종기

(54) 발명의 명칭 **다수의 솔리드-스테이트 드라이브들을 지원하기 위한 모듈러 시스템 아키텍처**

(57) 요 약

랙-장착형 데이터 스토리지 시스템은: 하나 이상의 스위치보드들을 포함하는 섀시; 상기 하나 이상의 스위치보드들과 인터페이스하는 미드플레인; 및 커넥터를 이용하여 상기 미드플레인과 탈착 가능하게 연결된 하나 이상의 데이터 스토리지 장치들을 포함한다. 하나 이상의 데이터 스토리지 장치들 중 적어도 하나의 데이터 스토리지 장치는 미드플레인과 인터페이스하는 로직 장치를 포함한다. 로직 장치는 미드플레인과 해당 데이터 스토리지 장치의 장치-특정 인터페이스를 제공한다. 적어도 하나의 데이터 스토리지 장치는 커넥터의 핀 상의 신호에 기초하여 제 1 프로토콜에 따라 로직 장치를 이용하여 구성되고, 그리고 적어도 하나의 데이터 스토리지 장치는 로직 장치를 이용하여 커넥터의 핀 상의 신호의 변화에 기초하여 제 2 프로토콜에 따라 재구성 가능하다.

대 표 도 - 도1a



(52) CPC특허분류
G06F 3/067 (2013.01)
H05K 7/18 (2013.01)

명 세 서

청구범위

청구항 1

하나 이상의 스위치보드들을 포함하는 섀시(chassis);

상기 하나 이상의 스위치보드들과 인터페이스(interface)하는 미드플레인(midplane); 및

커넥터를 이용하여 상기 미드플레인과 연결된 적어도 하나의 데이터 스토리지 장치를 포함하되,

상기 적어도 하나의 데이터 스토리지 장치는 상기 미드플레인과 인터페이스하는 장치를 포함하고,

상기 장치는 상기 미드플레인과 상기 적어도 하나의 데이터 스토리지 장치의 인터페이스를 제공하고,

상기 적어도 하나의 데이터 스토리지 장치는 적어도 하나의 NVMe(non-volatile memory express) SSD(solid-state drive) 및 NVMe-oF(NVMe over fabrics) SSD를 포함하고,

상기 미드플레인은 HA(High Availability) 모드 또는 non-HA 모드에서 동작하는 상기 데이터 스토리지 장치를 구성하고,

상기 미드플레인은 프로그래머블 로직 장치를 포함하고,

상기 프로그래머블 로직 장치는:

상기 적어도 하나의 데이터 스토리지 장치의 존재를 감지하고; 그리고

상기 적어도 하나의 데이터 스토리지 장치의 동작을 관리하기 위해 상기 하나 이상의 스위치보드들의 하나 이상의 BMC(baseboard management board)들과 인터페이스를 제공하고,

상기 프로그래머블 로직 장치와 상기 하나 이상의 BMC들과의 인터페이스에 기초하여, 상기 적어도 하나의 데이터 스토리지 장치는 상기 커넥터의 제1 신호에 기초하여 제 1 프로토콜에 따라 구성되고,

상기 프로그래머블 로직 장치와 상기 하나 이상의 BMC들과의 인터페이스에 기초하여, 상기 적어도 하나의 데이터 스토리지 장치는 상기 커넥터의 제2 신호에 기초하여 제 2 프로토콜에 따라 재구성 가능한 스토리지 시스템.

청구항 2

제 1 항에 있어서,

상기 적어도 하나의 데이터 스토리지 장치는 프로세서, FPGA(field-programmable gate array) 또는 ASIC(application-specific integrated circuit)를 포함하는 스토리지 시스템.

청구항 3

제 1 항에 있어서,

상기 커넥터는 PCIe(peripheral component interconnect express) 프로토콜을 지원할 수 있는 스토리지 시스템.

청구항 4

제 1 항에 있어서,

상기 커넥터는 적어도 하나의 U.2 커넥터 또는 M.2 커넥터인 스토리지 시스템.

청구항 5

제 1 항에 있어서,

상기 적어도 하나의 데이터 스토리지 장치는 NF1(new form factor 1) 표준과 호환 가능한 스토리지 시스템.

청구항 6

삭제

청구항 7

제 1 항에 있어서,

상기 커넥터는 핀을 포함하고, 상기 핀은 공급자에 의해 정의된 예약된 핀인 스토리지 시스템.

청구항 8

제 1 항에 있어서,

상기 섀시는 2U 섀시이고, 그리고

상기 스토리지 시스템은 상기 2U 섀시에서 서로의 위에 배치되는 제 1 스위치보드 및 제 2 스위치보드를 포함하는 스토리지 시스템.

청구항 9

제 8 항에 있어서,

상기 미드플레인은 미리 결정된 수의 데이터 저장 장치들을 삽입하는 구동 장치 공간을 갖는 스토리지 시스템.

청구항 10

제 1 항에 있어서,

상기 스토리지 시스템은 상기 섀시 내의 스위치보드를 포함하고, 상기 섀시는 1U 섀시를 포함하는 스토리지 시스템.

청구항 11

제 1 항에 있어서,

상기 스토리지 시스템은 상기 섀시에서 나란히 배치되는 제 1 스위치보드 및 제 2 스위치보드를 포함하고, 상기 섀시는 2U 섀시를 포함하는 스토리지 시스템.

청구항 12

삭제

청구항 13

삭제

청구항 14

레지스터;

하나 이상의 BMC(baseboard management board) 멀티플렉서들; 및

슬롯 멀티플렉서를 포함하되,

프로그래머블(programmable) 로직 장치는 데이터 스토리지 시스템의 미드플레인에 집적되고, 그리고 상기 미드플레인은 적어도 하나의 데이터 스토리지 장치를 삽입하기 위한 커넥터를 포함하는 구동 장치 공간을 포함하고,

상기 프로그래머블 로직 장치는 상기 하나 이상의 BMC 멀티플렉서들의 각각의 BMC 멀티플렉서들을 이용하여 하나 이상의 스위치보드들의 하나 이상의 BMC들에게 인터페이스를 제공하여 상기 적어도 하나의 데이터 스토리지 장치의 동작을 관리하고,

상기 하나 이상의 BMC 멀티플렉서들 각각은 상기 슬롯 멀티플렉서와 연결되어 상기 하나 이상의 스위치보드들의 BMC들이 상기 적어도 하나의 데이터 스토리지 장치와 동시에 통신 가능하게 하고,

상기 프로그래머블 로직 장치는 상기 하나 이상의 BMC들 각각이 상기 레지스터에 접근하도록 함으로써 상기 커넥터 상의 제1 신호에 기초하여 제 1 프로토콜에 따라 상기 적어도 하나의 데이터 스토리지 장치로 인터페이스를 제공하고, 그리고

상기 프로그래머블 로직 장치는 상기 커넥터 상의 제2 신호에 기초하여 제 2 프로토콜에 따라 상기 적어도 하나의 데이터 스토리지 장치를 재구성 가능한 프로그래머블 로직 장치.

청구항 15

제 14 항에 있어서,

상기 프로그래머블 로직 장치는 상기 데이터 스토리지 시스템의 하나 이상의 스위치보드들과 인터페이스하는 프로그래머블 로직 장치.

청구항 16

제 14 항에 있어서,

상기 프로그래머블 로직 장치는 FPGA(field-programmable gate array)인 프로그래머블 로직 장치.

청구항 17

제 14 항에 있어서,

상기 레지스터는 구성 레지스터, 활성 BMC ID, 슬롯 ID, 및 존재 레지스터를 포함하는 프로그래머블 로직 장치.

청구항 18

제 14 항에 있어서,

상기 커넥터는 핀을 포함하고, 상기 핀은 공급자에 의해 정의된 예약된 핀인 프로그래머블 로직 장치.

청구항 19

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 일반적으로 데이터 스토리지 장치들, 좀 더 구체적으로, 다수의 SSD들(solid-state drives)을 지원하기 위한 모듈러 시스템 아키텍처에 관한 것이다.

배 경 기 술

[0002] 종단(edge) 장치는 엔터프라이즈(enterprise) 또는 서비스 제공자 코어 네트워크들(service provider core networks)에 다양한 모바일 장치들에 의해 생성되는 네트워크의 트래픽의 엔트리 포인트들(entry points)을 제공한다. 종단 장치의 예시들은 라우터(router), 라우팅 스위치(routing switch), 스위치, 통합 접근 장치(integrated access device), 멀티플렉서(multiplexer), 및 다양한 MAN(metropolitan area network) 및 WAN(wide area network) 접근 포인트들을 포함한다.

[0003] 종단 장치는 하나의 유형의 네트워크 프로토콜 및 다른 프로토콜 사이를 변환할 수 있다. 일반적으로, 종단 장치들은 보다 빠르고, 보다 효율적인 중추(backbone) 및 핵심(core) 네트워크들로의 접근을 제공한다. 종단 장치들은 VPN(Virtual Private Network) 지원, VoIP(Voice over IP), 및 QoS(Quality of Service)와 같은, 향상된 서비스들을 또한 제공할 수 있다. 종단 장치는 종단 장치의 기능 및 수신(incoming) 및 송신(outgoing) 트래픽(traffic)의 통신 프로토콜(들)에 따라, 수신된 트래픽을 수정할 수 있거나 또는 수정하지 않을 수 있다. 예를 들어, 간단한 스위치는 수신 패킷(packet)들을 수정하지 않고 수신 트래픽을 라우팅하지만, 반면에 SBC(session border controller)는 수정된 패킷들을 보내기 전에 수신 패킷들에 대한 약간의 데이터 변환들을 할 수 있다.

[0004] 모바일 장치들로부터 데이터가 점점 더 생성될수록, 이들 모바일 장치들로부터 데이터센터로 전송되는 데이터의 양은 매년 급격하게 증가한다. 종단 장치는 모바일 장치들에 의해 생성되는 로컬(local) 데이터를 전처리(pre-process)하거나 원격 클라우드(remote cloud)로부터 기지국(base station)으로 워크로드(workload)들을 오프로딩(offloading) 하기 위한 능력(capability), 즉 종단 컴퓨팅을 가질 수 있다. 종단 장치들의 종단 컴퓨팅 능력은 모바일 장치들과 서비스 네트워크의 엔터프라이즈 사이의 데이터 전송을 효율적이고 그리고 경제적이게 할 수 있다.

[0005] 하나의 섀시(chassis) 내의 데이터 스토리지 장치들은 상이한 공급 업체들에 의해 제조될 수 있고 그리고 그들의 의도된 기능들 및 실행하기 위한 타깃 어플리케이션들(applications)에 따라 상이하게 구성되어야 한다. 상이한 공급 업체들로부터의 상이한 유형들의 데이터 스토리지 장치들을 지원할 수 있는 공통 시스템 플랫폼(platform)이 매우 바람직하다. 또한, NF1(New Form Factor 1) 기반 SSD들과 같이 최근에 떠오르는 데이터 스토리지 장치들을 지원할 수 있는 공통 시스템 플랫폼을 갖는 것도 또한 바람직하다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 상술한 기술적 과제를 해결하기 위한 것으로, 본 발명은 다수의 솔리드-스테이트 드라이브들을 지원하기 위한 모듈러 시스템 아키텍처를 제공할 수 있다.

과제의 해결 수단

[0007] 일 실시 예에 있어서, 랙-장착형 데이터 스토리지 시스템은 포함한다: 하나 이상의 스위치보드들을 포함하는 섀시; 하나 이상의 스위치보드들과 인터페이스하는 미드플레인; 그리고 커넥터를 이용하여 미드플레인과 탈착 가능하게 연결된 하나 이상의 데이터 스토리지 장치들. 하나 이상의 데이터 스토리지 장치들 중 적어도 하나의 데이터 스토리지 장치는 미드플레인과 인터페이스하는 로직 장치를 포함한다. 로직 장치는 미드플레인과 해당 데이터 스토리지 장치의 장치-특정 인터페이스를 제공한다. 적어도 하나의 데이터 스토리지 장치는 커넥터의 핀 상의 신호에 기초하여 제 1 프로토콜에 따라 상기 로직 장치를 이용하여 구성되고, 그리고 적어도 하나의 데이터 스토리지 장치는 로직 장치를 이용하여 커넥터의 핀 상의 신호의 변화에 기초하여 제 2 프로토콜에 따라 재구성 가능하다.

[0008] 다른 실시 예에 있어서, 프로그래머블 로직 장치는 포함한다: 프로그래머블 로직 장치는 포함한다: 레지스터; 하나 이상의 BMC(baseboard management board) 멀티플렉서들; 및 슬롯 멀티플렉서. 프로그래머블 로직 장치는 데이터 스토리지 시스템의 미드플레인에 집적, 구현, 또는 통합되고, 그리고 미드플레인은 복수의 커넥터들을 포함하여 하나 이상의 데이터 스토리지 장치들을 삽입하는 구동 장치 공간을 포함한다. 프로그래머블 로직 장치는 하나 이상의 BMC 멀티플렉서들의 각각의 BMC 멀티플렉서들을 이용하여 하나 이상의 스위치보드들의 하나 이상의 BMC들과의 인터페이스를 제공하여 하나 이상의 데이터 스토리지 장치들의 동작을 관리한다. 하나 이상의 BMC 멀티플렉서들 각각은 슬롯 멀티플렉서와 연결되어 하나 이상의 스위치보드들의 BMC들이 하나 이상의 데이터 스토리지 장치들과 동시에 통신 가능하게 한다. 프로그래머블 로직 장치는 하나 이상의 BMC들 각각이 레지스터에 접근하도록 함으로써 제 1 프로토콜에 따라 하나 이상의 데이터 스토리지 장치들로 장치-특정 인터페이스를 제공한다. 프로그래머블 로직 장치는 적어도 하나의 데이터 스토리지 장치에 대응하는 복수의 커넥터들의 커넥터의 핀의 변화에 기초하여 제 2 프로토콜에 따라 하나 이상의 데이터 스토리지 장치들 중 적어도 하나의 데이터 스토리지 장치를 재구성 가능하다.

[0009] 이벤트들의 구현 및 조합의 다양하고 신규한 세부 사항을 포함하는, 상술한 특징들과 다른 바람직한 특징들은 첨부 도면들을 참조하여 좀 더 구체적으로 설명될 것이고 청구 범위에서 언급될 것이다. 본문에서 설명된 특정한 시스템들 및 방법들은 한정들로서가 아닌 단지 예시로 도시됨이 이해될 것이다. 당업자가 이해할 수 있는 바와 같이, 본문에서 설명된 원리들 및 특징들은 본 개시의 범위를 벗어나지 않고 다양하고 수많은 실시 예들에서 이용될 수 있다.

발명의 효과

[0010] 본 발명의 실시 예에 따른, 공통 시스템 플랫폼은 M.2 커넥터(connector)를 이용하는 NF1(New Form Factor 1) 기반 SSD들과 같은 최근에 만들어진 장치들뿐만 아니라 표준 U.2 커넥터(예를 들면, PM1725a/1735 SSD들)를 갖는 NVMe-oF 장치를 지원할 수 있다.

도면의 간단한 설명

[0011] 본 명세서의 일부로서 포함되는 첨부 도면들은, 현재의 바람직한 실시 예를 도시하고 그리고 상술한 일반적인 설명 및 이하에 주어진 바람직한 실시 예의 상세한 설명과 함께 본문에서 설명된 원리들을 설명하고 교시하기 위해 제공된다.

도 1a는 일 실시 예에 따라, 2U 섀시에서 구현된 예시적인 데이터 스토리지 시스템의 블록도를 도시한다;

도 1b는 다른 실시 예에 따라, 2U 섀시에서 구현된 예시적인 데이터 스토리지 시스템의 블록도를 도시한다;

도 2는 일 실시 예에 따라, 데이터 스토리지 시스템의 미드플레인 내에 포함된 예시적인 CPLD의 블록도를 도시한다;

도 3은 일 실시 예에 따라, 2개 소형 폼팩터 SSD들을 포함하는 예시적인 데이터 스토리지 장치를 도시한다;

도 4는 일 실시 예에 따라, 1U 섀시에서 구현되는 예시적인 데이터 스토리지 시스템의 블록도를 도시한다; 그리고

도 5는 일 실시 예에 따라, 도 4의 데이터 스토리지 시스템 내에서 이용되는 예시적인 데이터 스토리지 장치를 도시한다.

도면들은 반드시 일정한 비율로 도시되진 않고 그리고 유사한 구조들 또는 기능들의 요소들은 도면들 전체에 걸쳐 예시적인 목적들로 유사한 참조 번호들에 의해 일반적으로 표시된다. 도면들은 단지 본문에서 설명된 다양한 실시 예들의 설명을 쉽게 하기 위해 의도된다. 도면들은 본문에서 개시된 교시들의 모든 양상을 설명하지 않고 그리고 청구 범위를 한정하지 않는다.

발명을 실시하기 위한 구체적인 내용

[0012] 본문에서 개시된 특징들 및 교시들 각각은 다수의 솔리드-스테이트 드라이브들(solid-state drives)을 지원하기 위한 모듈러(modular) 시스템 아키텍처를 제공하기 위한 시스템 및 방법을 제공하기 위해 다른 특징들 및 교시들과 분리되어 또는 함께 활용될 수 있다. 많은 추가적인 특징들 및 교시들을, 분리하여 그리고 함께, 활용하는 대표적인 예시들은 첨부된 도면들을 참조하여 더 상세하게 설명된다. 상세한 설명은 단지 본 교시들의 실시하기 위한 양상들을 위한 추가 세부 사항을 당업자에게 교시하기 위해 의도된 것이고 그리고 청구 범위를 한정하기 위해 의도되지 않는다. 그러므로, 상세한 설명에서 상술한 특징들의 조합들은 가장 넓은 의미에서 교시들을 실시하는데 필수적이지 않을 수 있고, 그리고 대신에 본 교시들의 대표적인 예시들을 특히 설명하기 위해 교시된다.

[0013] 이하의 설명에서, 단지 설명의 목적들로, 특정 명칭이 본 개시의 완전한 이해를 제공하기 위해 제시된다. 그러나, 당업자에게는 특정 세부 사항이 본 개시의 교시들을 실시하는데 요구되지 않는다는 것이 명백할 것이다.

[0014] 본문의 상세한 설명들의 일부 부분들은 컴퓨터 메모리 내의 데이터 비트들에 대한 동작들의 알고리즘들 및 기호적 표현들 측면에서 제시된다. 알고리즘 설명들 및 표현들은 데이터 처리 분야의 당업자가 연구 내용을 다른 당업자에게 효과적으로 전달하는데 이용된다. 알고리즘은 이곳에 있고, 그리고 일반적으로, 원하는 결과를 유도하는 단계들의 일관성 있는 순서로 생각된다. 단계들은 물리적 양들의 물리적 조작들을 요구하는 단계들이다. 일반적으로, 반드시 그런 것은 아니지만, 이러한 양들은 저장, 전송, 결합, 비교, 및 기타 조작이 가능한 전기적 또는 자기적 신호들의 형태를 취한다. 주로 공통 활용의 이유들로, 비트들, 값들, 요소들, 기호들, 문자들, 용어들, 숫자들 등으로서 이들 신호들이 참조되는 것이 때로는 편리한 것으로 판명되었다.

[0015] 그러나, 이들 모든 용어들 및 유사한 용어들은 적절한 물리적 양들과 관련되고 단지 이러한 양들에 적용되는 편리한 라벨들인 것을 유념해야 한다. 이하의 설명에서 명백하게 다르게 특별히 기술하지 않는 한, 설명 전체에 걸쳐, “처리(processing)”, “컴퓨팅”, “계산”, “판별”, “표시(displaying)” 등과 같은 용어들을 활용하는 논의들은, 컴퓨터 시스템의 레지스터들 및 메모리들 내의 물리적 (전자) 양들로서 표현된 데이터를 컴퓨터 시스템 메모리들 또는 레지스터들 또는 다른 정보 스토리지, 전송 또는 디스플레이 장치들 내의 물리적 양들로서 유사하게 표현되는 다른 데이터로 조작하고 변형하는 컴퓨터 시스템, 또는 유사한 전자 컴퓨팅 장치의 행동 및 프로세스들을 참조하는 것이 명백하다.

[0016] 또한, 대표적인 예시들 및 종속항들의 다양한 특징들은 본 교시들의 추가적인 유용한 실시 예들을 제공하기 위해 구체적으로 그리고 명시적으로 열거되지 않은 방식들로 결합될 수 있다. 모든 값 범위들 또는 엔티티

(entity)들의 그룹들의 표시들은 청구된 주제를 제한할 목적뿐만 아니라, 본래의 개시를 목적으로 모든 가능한 중간 값 또는 중간 엔티티를 개시하는 것을 또한 명시한다. 도면들에 도시된 구성 요소들의 치수들 및 형상들은 본 교시들이 어떻게 실시되는지 이해를 돕기 위해 설계된 것이나, 예시들에서 도시된 치수들 및 형상들로 제한하려는 의도가 아님을 또한 명시한다.

[0017] 본 개시는 상이한 공급 업체들에 의해 만들어진 다른 NVMe-oF(non-volatile memory express(NVMe) over fabrics) 장치들을 지원할 수 있는 공통 시스템 플랫폼, 그리고 NVMe 장치 또는 NVMe-oF 장치로 구성될 수 있는 멀티-모드(multi-mode) 스토리지 장치를 설명한다. 일 실시 예에 따라, 본 공통 시스템 플랫폼은 복수의 NVMe 또는 NVMe-oF 장치들을 각각 수용할 수 있는 미드플레인(midplane) 및 하나 이상의 마더보드들(motherboards; NVMe 장치들의 경우) 또는 하나 이상의 스위치보드들(switchboards; NVMe-oF 장치들의 경우)을 포함하는 랙-장착형 섀시(rack-mountable chassis; 또는 인클로저(enclosure))를 지칭할 수 있다. 본 공통 시스템 플랫폼은 M.2 커넥터(connector)를 이용하는 NF1(New Form Factor 1) 기반 SSD들과 같은 최근에 만들어진 장치들뿐만 아니라 표준 U.2 커넥터(예를 들면, PM1725a/1735 SSD들)를 갖는 NVMe-oF 장치를 지원할 수 있다.

[0018] 멀티-모드 NVMe-oF(non-volatile memory express(NVMe) over fabrics) 장치는 알려진 위치로부터 또는 섀시 유형 핀, 예를 들어, 핀(E6) 또는 멀티-모드 NVMe-oF 장치가 삽입된 섀시 유형에 따른 U.2 커넥터의 공급자에 의해 정의된 예약된 핀으로부터 정보를 감지함으로써 NVMe 또는 NVMe-oF 프로토콜 중 어느 하나를 지원할 수 있다. 멀티-모드 NVMe-oF 장치가 NVMe 섀시의 구동 장치 공간(drive bay)에 삽입되면, U.2 커넥터의 4개 PCIe(peripheral component interconnect express) 레인(lane)들 모두는 내장된(embedded) PCIe 엔진에 의해 구동된다. 이 경우, NVMe-oF 장치는 내장된 이더넷(Ethernet) 엔진을 비활성화하고(disable), 그리고 모든 NVMe 명령들 및 기능들이 지원되거나 활성화된다(enable). 다른 한편으로는, 멀티-모드 NVMe-oF 장치가 NVMe-oF 섀시의 구동 장치 공간에 삽입되면, NVMe-oF 장치의 이더넷 포트들이 활성화되고 데이터-플레인(data-plane)으로서 사용한다. 이 모드에서, by-4 (X4) PCIe 레인들은 2개 by-2 (X2) PCIe 레인들로서의 2개 제어 플레인(control plane)들로서 동작된다.

[0019] 일 실시 예에 따라, 본 공통 시스템 플랫폼의 미드플레인은 NVMe 장치 또는 NVMe-oF 장치 중 어느 하나로 구성될 수 있는 본 멀티-모드 스토리지 장치뿐만 아니라 NVMe 및 NVMe-oF 장치들 모두를 지원할 수 있다. NVMe 모드에서 구성되면, 본 멀티-모드 스토리지 장치는 NVMe 장치로 행동하고(또는 동작하고), 그리고 NVMe-oF 모드에서 구성되면, 본 멀티-모드 스토리지 장치는 하나 이상의 이더넷 포트들에 대한 U.2 커넥터 상의 SAS 핀들을 이용하여 NVMe-oF 장치로 기능한다(또는 동작한다).

[0020] 일 실시 예에 따라, 본 공통 시스템 플랫폼의 미드플레인은 HA(high availability) (듀얼-포트(dual-port)) 모드 및 non-HA(비HA) 모드 (싱글-포트(single-port)) 모두를 지원할 수 있다. 미드플레인은 2개 스위치보드들이 2개 PCIe 스위치들의 NTB(Non-Transparent Bridge)를 통해 서로 통신할 수 있는 HA 모드를 지원하기 위해 고속의, 클럭(clock), 및 제어 신호들로 미리 라우팅(routing)될 수 있다. 일반적으로, HA 모드를 지원할 수 있는 NVMe-oF 장치는 더 비쌀 수 있다. 본 공통 시스템 플랫폼은 HA 및 non-HA 모드들 모두를 지원할 수 있는 미드플레인을 이용하는 시스템 구성 및 어플리케이션에 기초하여 HA 및 non-HA 모드들을 모두 지원함으로써 규모의 경제를 제공할 수 있다. non-HA 모드에서 구성되면, 본 공통 시스템 플랫폼은 제어 플레인으로서 모든 표준 특징들을 위해 하나의 by-2 (X2) PCIe 레인들을 이용한다. HA 모드에서 구성되면, 본 공통 시스템 플랫폼은 4개 PCIe 레인들을 포트(A) 및 포트(B)를 위한 2개 by-2 (X2) PCI 레인들로 각각 나눈다. 이더넷 신호들은 주 이더넷 포트(예를 들어, Eth0)를 위한 일부 SAS 핀들(예를 들어, S2, S3, S5, 및 S6) 그리고 부 이더넷 포트(예를 들어, Eth1)를 위한 다른 SAS 핀들(예를 들어, S9, S10, S12, 및 S13)을 이용한다. 핀(E25, DualPort_EN#)은 듀얼 포트 구성을 활성화하는데 이용된다. 예를 들어, 핀(E25)이 로우(low)로 되면, NVMe-oF 장치는 듀얼 포트 모드에서 동작하거나, 그렇지 않으면, NVMe-oF 장치는 싱글 포트 모드에서 동작한다. 2개 PCIe 스위치들(106)은 각각의 NTB 포트들을 통해 서로 연결된다.

[0021] NVMe-oF의 경우, 본 공통 시스템 플랫폼의 스위치보드는 제어 플레인으로서 2개 by-2 (X2) PCIe 레인들을 이용하여 추가적인 비용 없이 섀시에 부착된 NVMe-oF 장치들 각각과 통신한다. 동일한 미드플레인이 NVMe 기반의 섀시 또는 NVMe-oF 기반의 섀시 모두를 위해 이용될 수 있다. 그러므로, 본 공통 시스템 플랫폼은 더 빠른 시장 출시 기간 및 더 낮은 개발 위험을 제공할 수 있다. 본 공통 시스템 플랫폼은 데이터 스토리지 시스템의 성능을 선형적으로 조정할 수 있고 슬롯(slot) 당 소형 폼팩터(Small Form Factor) 당 더 많은 SSD들을 제공할 수 있다.

[0022] 일 실시 예에 있어서, 기존 PCIe 드라이버(driver)는 본 공통 시스템 플랫폼 및 본 멀티-모드 스토리지 장치를

지원하기 위한 수정이 없이 이용될 수 있다. 또한, 본 공통 시스템 플랫폼은 1U 및 2U 섀시들과 같은 다양한 폼 팩터들(form factors)의 마더보드들 또는 스위치보드들을 재사용할 수 있다.

[0023] 본 공통 시스템 플랫폼은 동일한 데이터 스토리지 장치가 NVMe 장치 또는 NVMe-oF 장치로 이용될 수 있으므로 데이터 스토리지 장치들의 유닛(또는 단위) 당 종단(edge) 장치 또는 데이터 스토리지 서버의 비용을 낮출 수 있다. 또한, 본 멀티-모드 데이터 스토리지 장치는 다양한 제품들에 사용될 수 있거나 그리고/또는 데이터센터 내의 랙 장착형 섀시에 부착될 수 있다.

[0024] 일 실시 예에 따라, 본 공통 시스템 플랫폼의 미드플레인은 미드플레인 동작들을 관리하는 CPLD(complex programmable logic device; 복합 프로그래머블 로직 장치)를 포함한다. CPLD는 I2C 및/또는 SMBus를 통해 부착된 마더보드들 또는 스위치보드들의 BMC들과 통신한다. 각 마더보드 또는 스위치보드의 BMC는 그것에 부착된 상이한 유형들의 데이터 스토리지 장치들의 행동들(또는 동작들)을 감지하고 조정하는데 도움이 될 수 있다.

[0025] 일 실시 예에 따라, 미드플레인은 컴퓨터를 이용한 지원을 제공하지 않을 수 있다. 대신에, 미드플레인과 SSD 사이에 배치되는 SSD 인터포저 카드(SSD interposer card, 미도시)는 FPGA(Field-Programmable Gate Array) 또는 ASIC(Application-Specific Integrated Circuit)을 포함하여 미드플레인과 인터페이스(interface; 접속 또는 연결)하여 자신의 인터페이싱(interfacing)의 최적화를 관리한다. 예를 들어, SSD 인터포저 카드는 BMC와 조율하여 미드플레인과 함께 부착된 SSD를 관리할 수 있다.

[0026] 본 공통 시스템 플랫폼의 미드플레인은 복수의 SSD들이 연결될 수 있는 하나 이상의 종단 SSD 컨트롤러를 포함할 수 있다. 미드플레인은 시스템의 구성, 예를 들어, 1U 및 2U 섀시에 따라 하나 이상의 종단 SSD 컨트롤러들을 마더보드 또는 스위치보드에 연결하도록 구성될 수 있다. 다수의 종단 SSD 컨트롤러들 각각은 복수의 SSD들, 예를 들어, 4개 이더넷 SSD들(Ethernet SSDs; eSSDs)과 직접 인터페이스할 수 있다. eSSD들 각각은 자신의 더 가벼운 ASIC을 가질 수 있고 예를 들어, NGSFF(Next Generation Small Form Factor) 또는 NF1(Next Generation Small Form Factor) 표준과 호환 가능한, 소형 폼팩터를 가질 수 있다.

[0027] 도 1a는 일 실시 예에 따라, 2U 섀시에서 구현된 예시적인 데이터 스토리지 시스템의 블록도를 도시한다. 데이터 스토리지 시스템(100A)은 2U 랙 장착형 섀시에서 나란히 배치될 수 있는 2개 스위치보드들(101a, 101b)을 포함한다. 2개 스위치보드들(101a, 101b)은 섀시 내 자신들의 배치를 제외하고는 동일할 수 있다. 시스템 구성에 따라, 데이터 스토리지 시스템(100A)은 단지 하나의 스위치보드를 포함할 수 있다. 이하에서, 2개 스위치보드들(101a, 101b)은 집합적으로 또는 독립적으로 스위치보드(101)로서 지칭될 수 있다.

[0028] 스위치보드들(101a, 101b) 각각은 CPU(central processing unit; 102), BMC(baseboard management controller; 103), 이더넷 스위치 컨트롤러(105), PCIe 스위치(106), 이더넷 포트들 및 PCIe 포트들을 포함하는 복수의 업링크(uplink) 포트들(114), 공통 미드플레인(151)을 통해 복수의 데이터 스토리지 장치들(150, 예를 들어, NVMe SSD들 또는 eSSD들)과 인터페이스하는 고속 커넥터(115), 그리고 관리 LAN(local area network) 포트(116)를 포함한다. 데이터 스토리지 시스템(150)의 예시는 삼성전자에 의해 설계되고 제조된 PM1725a NVMe SSD이다. 이하에서, 용어들, 데이터 스토리지 장치 및 eSSD는 일부 실시 예들에서 설명의 편의를 위해 교환해서 이용될 수 있다; 그러나, 데이터 스토리지 장치(150)는 임의의 유형의 데이터 스토리지 장치들, 예를 들어, NVMe SSD, 이더넷 SSD, 및 NVMe SSD 또는 NVMe-oF SSD로 구성될 수 있는 멀티-모드 SSD일 수 있다.

[0029] BMC(103)는 다양한 센서들, 예를 들어, 파워 상태 센서(미도시) 및 온도 센서를 이용하여 해당 스위치보드(101)의 물리적 상태를 모니터링(monitoring)할 수 있는 로컬 서비스 프로세서이다. BMC(103)는 디스플레이를 갖는 휴대용 장치를 이용하여 관리 LAN 포트(116) 또는 SMBus(system management bus, 시스템 관리 버스; 미도시)와 같은 독립적인 통신 경로를 통해 서비스 관리자와 통신할 수 있다.

[0030] 업링크 포트들(114)은 어플리케이션을 실행하는 호스트 컴퓨터에 연결될 수 있고, 그리고 호스트 컴퓨터에서 실행되는 어플리케이션은 데이터 스토리지 장치(150)에 접근하여 업링크 포트들(114)을 통해 데이터를 저장할 수 있고 저장된 데이터에 접근할 수 있다. 데이터 스토리지 시스템(100A)이 NVMe-oF 시스템이면, 호스트 컴퓨터는 이더넷, 파이버 채널(Fibre Channel), 인피니밴드(InfiniBand)와 같은, 패브릭(fabric) 네트워크를 통해 데이터 스토리지 장치(150)에 접근할 수 있다.

[0031] 예를 들어, 업링크 포트들 각각은 100 기가비트 이더넷(Gigabit Ethernet; Gbe) 포트이다. NVMe-oF의 경우, 호스트 컴퓨터는 이더넷 패킷들을 데이터 스토리지 장치들(150) 상의 데이터를 읽고, 수정하고, 그리고 쓰는 명령들을 포함하는 스위치보드(101)로 전송할 수 있다. NVMe의 경우, 데이터 스토리지 장치(150)는 종래의 X86 기반의 마더보드(미도시)에 부착될 수 있다.

[0032] 관리 LAN 포트(116)는 외부 관리 스위치(미도시)에 연결될 수 있다. 시스템 관리자는 IPMI(intelligent platform management interface) 프로토콜을 통해 관리 LAN 포트(116)를 통하여 다수의 스위치보드들의 상태를 직접 관리(또는 모니터링)할 수 있다. IPMI 프로토콜은 IPMI 메시지들을 이용하여 관리 LAN 포트(116)를 통해 시스템 관리자와 BMC(103) 사이의 통신을 허용한다. 스위치보드(101)는 다른 구성 요소들, 회로들, 및/또는 서브시스템들, 예를 들어, 하나 이상의 DDR4(dual data rate 4) DIMMs(dual in-line memory modules)을 포함하여 데이터 스토리지 장치들(150)로의 데이터 전송 및 데이터 스토리지 장치들(150)로부터의 데이터 전송을 쉽게 할 수 있고 그리고 데이터 스토리지 장치들(150)을 제어하고 효과적으로 관리할 수 있다.

[0033] 일 실시 예에 따라, 최대 24개 데이터 스토리지 장치들(150)이 데이터 스토리지 시스템(100A)의 스위치보드들(101a, 101b) 각각에 연결될 수 있다. 그러므로, 총 24개 데이터 스토리지 장치들(150)이 데이터 스토리지 시스템(100A)에 부착될 수 있다. 예를 들어, 스위치보드(101a)는 이더넷 포트(0)를 통해 eSSD1부터 eSSD24까지 연결되고, 그리고 스위치보드(101b)는 이더넷 포트(1)를 통해 eSSD1부터 eSSD24까지 연결된다. 각 eSSD는 최대 700k IOPs(input/output operations per second)까지 지원할 수 있다. 데이터 스토리지 시스템(100A)의 평가된 성능은 랜덤 읽기 입출력(I/O)에 대해 약 1680만 IOPs(또는 약 16,800,000 IOPs; 각 eSSD 당 24 곱하기 700k IOPS)이다.

[0034] eSSD1 내지 eSSD24 각각은 NVMe 모드 또는 NVMe-oF 모드에서 동작하도록 구성될 수 있다. 예를 들어, 데이터 스토리지 장치(150)는 NVMe-oF 모드에서 동작하도록 구성되는 NVMe-oF 장치(또는 eSSD)이다. 이 경우, 4개 PCIe 레인들 중 2개(2X PCIe)는 제 1 스위치보드(101a)의 고속 커넥터(115)에 연결되도록 구성되고, 그리고 나머지 2개 PCIe 레인들은 제 2 스위치보드(101b)의 고속 커넥터(115)에 연결되도록 구성된다. 유사하게, 제 1 이더넷 포트(Eth0)는 제 1 스위치보드(101a)의 고속 커넥터(115)에 연결되도록 구성되고, 그리고 제 2 이더넷 포트(Eth1)는 제 2 스위치보드(101b)의 고속 커넥터(115)에 연결되도록 구성된다. 데이터 스토리지 장치들(150) 각각으로 전송되고 데이터 스토리지 장치들(150) 각각으로부터 수신되는 이더넷 및 PCIe 트래픽(traffic)은 이더넷 스위치 컨트롤러(105) 및 PCIe 스위치(106)를 통해 각각 라우팅(routing)된다.

[0035] 일 실시 예에 따라, 미드플레인(151)은 미드플레인 동작들을 관리하는 CPLD(complex programmable logic device, 152)를 포함한다. CPLD(152)는 부착된 스위치보드들(101)의 BMC들(103)과 통신한다. CPLD(152) 및 각 스위치보드(101)의 BMC(103)는 섀시에 부착된 상이한 유형들의 데이터 스토리지 장치들(150)의 행동들(또는 동작들)을 감지하고 조정하는데 도움이 될 수 있다. 예를 들어, CPLD(152)는 고속 커넥터(115) 상의 데이터 스토리지 장치(150)에 대응하는 특정한 핀 상의 전압을 측정함으로써 각 슬롯 내 데이터 스토리지 장치(150)의 존재/부재를 감지한다.

[0036] CPLD(152)는 BMC(103)에 대한 다양한 지원들을 제공하여 데이터 스토리지 장치들(150)을 관리할 수 있다. 예를 들어, CPLD(152)는 전용의 SMBus 및/또는 I2C 포트를 통해 BMC(103)를 부착된 데이터 스토리지 장치들(150, 데이터 스토리지 시스템(100A) 내 최대 24개까지)로 연결한다. CPLD(152)는 각 데이터 스토리지 장치(150)가 동일한 센서 어드레스를 갖는 제약으로 인해 BMC(103)가 한번에 하나씩 각 데이터 스토리지 장치(150)와 통신하게 할 수 있다. 데이터 스토리지 장치(150)는 2.5” 소형 폼팩터를 갖는 드라이브들을 지원하는 섀시의 각 드라이브 슬롯에 부착될 수 있다. CPLD(152)는 BMC(103)에 연결된 I2C 포트를 또한 지원(또는 제공)하여 데이터 스토리지 장치들(150)의 각 그룹의 파워 온/오프를 가능하게 하는 보호 메커니즘(mechanism)과의 신뢰할 수 있는 I2C 통신을 지원(또는 제공)할 수 있다. CPLD(152)는 (가능한 경우) PM Bus를 통해 PSU(power supply unit; 파워 공급부)의 정보에 접근함으로써 최대 24개 데이터 스토리지 장치들(150)의 미드플레인(151) 상의 표시등(예를 들면, 구동 고장 LED)을 더 켤 수 있고 그리고 로직 제어 및 파워-온/다운(power-on/down) 타이밍 제어를 리셋(reset)할 수 있다.

[0037] 일 실시 예에 따라, 데이터 스토리지 시스템(100A)은 PDB(power distribution board; 분전반, 153)를 포함한다. 분전반(153)은 스위치보드들(101a, 101b) 각각의 핫 스왑 컨트롤러(hot swap controller, 미도시)로 파워(예를 들어, 12V)를 공급하는 여분의 PSU(power supply unit; 파워 공급부)를 포함할 수 있다. 핫 스왑 컨트롤러는 스위치보드들(101)을 파워 오프하지 않고 데이터 스토리지 장치(150)가 섀시에 부착되거나 또는 섀시로부터 탈착되게 할 수 있다(즉, 데이터 스토리지 장치(150)는 이동식 데이터 스토리지 장치일 수 있음).

[0038] 도 1b는 다른 실시 예에 따라, 2U 섀시에서 구현된 예시적인 데이터 스토리지 시스템의 블록도를 도시한다. 데이터 스토리지 시스템(100B)은 1개 또는 2개 스위치보드들(111a, 111b), 분전반(153), 및 미드플레인(161)을 포함한다. 데이터 스토리지 시스템(100B)이 2개 스위치보드들(111a, 111b)을 포함하면, 2개 스위치보드들(111a, 111b)은 섀시 내 자신들의 배치를 제외하고 동일할 수 있다. 시스템 구성에 따라, 데이터 스토리지 시스템

(100B)은 단지 1개 스위치보드를 포함할 수 있다. 이하에서, 2개 스위치보드들(111a, 111b)은 집합적으로 또는 독립적으로 스위치보드(111)로서 지칭될 수 있다.

[0039] 도 1a의 데이터 스토리지 시스템(100A)과 유사하게, 데이터 스토리지 시스템(100B)은 2U 섀시에 둘러싸이고 그리고 2개 스위치보드들(111a, 111b)을 포함한다. 데이터 스토리지 시스템(100B)은 상이한 유형의 이더넷 스위치 컨트롤러(107) 그리고 데이터 스토리지 시스템(100A)의 데이터 스토리지 장치(150)의 폼팩터와 상이한 폼팩터를 갖는 데이터 스토리지 장치들(250)을 수용하기 위한 미드플레인(161)의 구동 장치 공간을 포함할 수 있다. 데이터 스토리지 장치들(150, 250)은 동일한 M.2 커넥터를 이용할 수 있다.

[0040] 구조적인 차이로 인하여, 데이터 스토리지 장치들(250)과 인터페이스하기 위한 미드플레인(161)에 포함된 SSD 컨트롤러는 데이터 스토리지 시스템(100A)의 미드플레인(151)에 포함된 SSD 컨트롤러와 비교하여 상이한 구성 및 설계를 가질 수 있다. 예를 들어, 데이터 스토리지 시스템(100B)의 스위치보드(111)에 포함된 이더넷 스위치 컨트롤러(107)는 브로드컴(Broadcom)에 의해 설계되고 제작된 트라이던트 시리즈(Trident series) 스위치보드이지만 반면에 도 1a에서 도시된 데이터 스토리지 시스템(100A)의 스위치보드(101)에 포함된 이더넷 스위치 컨트롤러(105)는 브로드컴에 의해 설계되고 제작된 토마호크 시리즈(Tomahawk series) 스위치보드이다. 데이터 스토리지 시스템(100B)의 스위치보드(111)에 포함된 이더넷 스위치 컨트롤러(107)가 데이터 스토리지 시스템(100A)의 스위치보드(101)에 포함된 이더넷 스위치 컨트롤러(105)보다 싸기 때문에 데이터 스토리지 시스템(100B)의 비용은 데이터 스토리지 시스템(100A)의 비용과 비슷하거나 적다; 그러나, 이것은 네트워킹 어플리케이션보다 스토리지 어플리케이션에 더 적합한 더 큰 사이즈의 버퍼들을 가질 수 있다. 또한, 데이터 스토리지 시스템(100B)의 데이터 스토리지 장치(250)의 개별 IOPS 성능(예를 들면, 550k IOPS)이 데이터 스토리지 시스템(100A)의 데이터 스토리지 장치(150)의 IOPS 성능(예를 들면, 700k IOPS)보다 낮을 수 있음에도 불구하고 동시에 그리고 독립적으로 접근될 수 있는 데이터 스토리지 장치들(150)의 증가된 개수(예를 들어, 데이터 스토리지 시스템(100B)의 48개 데이터 스토리지 장치들(250) vs. 데이터 스토리지 시스템(100A)의 24개 데이터 스토리지 장치들(150))로 인하여 데이터 스토리지 시스템(100B)의 IOPS 성능은 데이터 스토리지 시스템(100A)의 IOPS 성능과 동일하거나 더 나을 수 있다.

[0041] 일 실시 예에 따라, 데이터 스토리지 시스템(100B)에 포함된 데이터 스토리지 장치(250)는 하나 이상의 NGSFF 폼팩터 SSD들(본문에서 NF1 SSD들로 또한 지칭됨)을 가질 수 있다. 일부 실시 예들에 있어서, NGSFF 또는 NF1 폼팩터는 M.3 폼팩터로 또한 지칭될 수 있다. NF1 SSD는 M.2 커넥터를 이용하나 폭넓은 PCB(printed circuit board)를 수용하는 폼팩터를 가져서 추가적인 회로들(예를 들어, FPGA 또는 ASIC) 및/또는 NAND 패키지들을 위한 더 많은 공간을 제공한다. 예를 들어, NF1 SSD는 30.5 ㎜의 폭과 110 ㎜의 길이를 갖는다. 반면에, 표준 M.2 모듈들은 다양한 폭들(예를 들면, 12, 16 22, 및, 30 ㎜)과 길이들(예를 들면, 16, 26, 30, 38, 42, 60, 80, 110 ㎜)이 가능하다. 일 실시 예에 있어서, 각 데이터 스토리지 장치(250)는 데이터 스토리지 시스템(100B)에 부착될 수 있고 동시에 그리고 독립적으로 접근될 수 있는 SSD들의 최대 개수가 48이도록 2개 SSD들을 수용할 수 있다. 반면에, 도 1a에서 도시된 데이터 스토리지 시스템(100A)의 데이터 스토리지 장치(150)는 단지 1개 SSD를 가질 수 있고, 그리고 데이터 스토리지 시스템(100A)에 부착될 수 있는 SSD들의 최대 개수는 24이다.

[0042] 일 실시 예에 따라, 스위치보드들(111a, 111b)의 이더넷 스위치 컨트롤러(107)는 이더넷 리피터(repeater) 또는 리타이머(re-timer)로 교체될 수 있다. 이더넷 스위치 컨트롤러(107)에 포함된 리피터는 고속 커넥터(115)를 통해 업링크 포트들(114)과 다운링크 포트들 사이의 이더넷 신호들의 액티브 패스-쓰루(active pass-through)를 데이터 스토리지 장치들(250)로 제공할 수 있다. 예를 들어, 제 1 스위치보드(111a)의 리피터는 제 1 스위치보드(111a)에 부착된 eSSD들(1-24)에 대한 포트(0's)뿐만 아니라 미드플레인(161)을 통해 제 2 스위치보드(111b)에 부착된 eSSD들(1-24)의 이더넷 포트(1's)의 이더넷 신호들을 액티브 패스 쓰루(즉, 장거리의 신호 전송을 위해 신호들을 증폭)할 수 있다. 유사하게, 제 2 스위치보드(111b)의 리피터는 제 2 스위치보드(111b)에 부착된 eSSD들(1-24)의 포트(1's)뿐만 아니라 미드플레인(161)을 통해 제 1 스위치보드(111a)에 부착된 eSSD들(1-24)의 이더넷 포트(0's)의 이더넷 신호들을 액티브 패스 쓰루할 수 있다.

[0043] HA 모드에서, 스위치보드(111a)의 PCIe 스위치(106)는 부착된 eSSD들의 U.2 커넥터의 2개 PCIe 레인들(0, 1)을 제 1 이더넷 포트(이더넷 포트 0)를 위한 제어 플레인으로서 이용할 수 있다. SAS 포트(0)의 제 1 쌍은 제 1 이더넷을 위해 이용될 수 있다. 스위치보드(111a 또는 111b)의 PCIe 스위치(106)는 모든 eSSD들(1-24)의 PCIe 포트(A)와 통신할 수 있다. 유사하게, 스위치보드(111b)의 PCIe 스위치(106)는 부착된 eSSD들의 U.2 커넥터의 2개 PCIe 레인들(2, 3)을 제 2 이더넷 포트를 위한 제어 플레인으로서 이용할 수 있다. SAS 포트(1)의 제 2 쌍은 제 2 이더넷 포트(예를 들어, 이더넷 포트(1))를 위해 이용될 수 있다. 스위치보드(111b)의 PCIe 스위치(106)는 모

든 eSSD들(1-24)의 PCIe 포트(1)와 통신할 수 있다.

[0044] 부착된 eSSD들 각각은 PCIe 스위치(106) 및 eSSD 사이에서 수립된 제어 플레인을 통해 PCIe 스위치(106)를 통해 BMC(103)로 일부 장치-특정 정보를 제공할 수 있다. 제어 플레인을 통해 전달될 수 있는 이러한 장치-특정 정보의 예시들은 eSSD의 발견(discovery) 정보 및 FRU 정보를 포함하나, 이에 한정되지는 않는다.

[0045] 미드플레인(161)은 도 1a의 데이터 스토리지 시스템(100A)에서 사용되는 동일한 미드플레인일 수 있다. 그러나, 데이터 스토리지 시스템(100B)은 단지 하나의 SSD를 포함(예를 들어, 데이터 스토리지 장치(150)는 1개 PM1725a SSD를 포함)하는 데이터 스토리지 시스템(100A)에 비해 동시에 그리고 독립적으로 접근될 수 있는 하나 이상의 NF1 SSD들(예를 들어, 데이터 스토리지 장치(250)는 2개 SSD들을 포함)을 수용할 수 있다. 데이터 스토리지 장치(250)에 포함된 SSD들 각각의 최대 I/O 성능은 550k IOPs이다. 이 경우, 데이터 스토리지 시스템(100B)의 예상되는 최대 I/O 성능은 2450만 IOPs(또는 24,500,000 IOPs; 각 SSD 당 48 곱하기 550k IOPs)이다. 일 실시 예에 따라, 데이터 스토리지 시스템(100B)은 1개 SSD를 갖는 하나 이상의 데이터 스토리지 장치들(150)을 또한 수용할 수 있다. 다시 말해서, 도 1a의 데이터 스토리지 장치(150)는 (M.2 커넥터를 갖는) 데이터 스토리지 시스템(100B)과 역으로 호환되지만, 도 1a의 데이터 스토리지 시스템(100A)에서 적절히 기능하는 도 1b에서 도시된 데이터 스토리지 장치(250)를 갖도록 새로운 제어 로직 설계가 필요할 수 있다. 새로운 제어 로직은 SSD 인터포저 카드 내에서 또는 데이터 스토리지 장치(250) 내부적으로 구현될 수 있다(예를 들어, 도 3의 제어 로직 장치(351)). 1개 SSD는 1개 PCIe RP(root port; 루트 포트)를 요구한다. 데이터 스토리지 장치(250) 내의 2개 SSD들은 2개 PCIe 루트 포트들을 요구할 것이고, 그리고 이것들은 새로운 제어 로직에 의해 구현될 필요가 있다. 이 경우, 데이터 스토리지 시스템(100B)의 최대 성능은 2640만 최대 IOPs(26,400,000 IOPs, 즉, 24개 eSSD 250, 및 eSSD 250 내 2개 SSD들 각각 당 550k IOPs)에서 3360만 최대 IOPS(33,600,000 IOPS, 즉, 48개 eSSD 250, 및 eSSD 250 내 2개 SSD들 각각 당 700k IOPs)로 증가할 수 있다. 시스템 구성 및 목표 비용에 따라, 상이한 I/O 성능은 상이한 데이터 스토리지 시스템들 및 상이한 데이터 스토리지 장치들을 선택함으로써 달성될 수 있다.

[0046] 도 2는 일 실시 예에 따라, 데이터 스토리지 시스템의 미드플레인 내에 포함된 예시적인 CPLD의 블록도를 도시한다. 데이터 스토리지 시스템(200)은 2개 스위치보드들(201a, 201b), 미드플레인(251), 및 각 커넥터(281, 예를 들어, U.2 커넥터, M.2 커넥터)를 통해 미드플레인(251)에 부착된 복수의 데이터 스토리지 장치들(250)을 포함한다. 분전반(267)은 PMBus(266)를 통해 CPLD(252)의 레지스터에 접근할 수 있다. I2C 또는 슬롯 멀티플렉서(263)는 BMC(203)를 활성화하는데 사용되어 한번에 하나씩 각 eSSD 250과 통신한다. BMC 멀티플렉서(262)는 시스템에 하나 이상의 BMC들(203)이 존재할 때 사용된다. 스위치보드들(201) 각각은 BMC(203) 및 도 1a, 1b에서 도시된 데이터 스토리지 시스템(100A, 100B)에 관하여 설명된 바와 같이 다른 구성 요소들(미도시)을 갖는다. 도시의 편의를 위해, 도 2는 스위치보드(201) 내에 BMC(203)만을 도시한다. 미드플레인(251)은 CPLD(252) 및 EEPROM(electrically erasable programmable read-only memory) 칩에 저장될 수 있는 VPD(vital product data, 필수 제품 데이터; 265)를 포함한다. VPD(265)는 스위치보드(201)의 제조 업체에 의해 설정된 제품-특정 데이터를 포함할 수 있다.

[0047] 2개 스위치보드들(201a, 201b)이 서로의 위에 배치되는 것(예를 들어, 도 1b) 대신에 나란히 배치되면(예를 들어, 도 1a), 스위치보드들(201) 각각과 미드플레인(251) 사이의 인터페이스 영역이 절반으로 감소하므로, 스위치보드들(201) 각각과 미드플레인(251) 사이의 다수의 인터페이스 핀들을 갖는 커넥터를 배치하기에 충분한 물리적 공간이 없을 수 있다.

[0048] BMC(203)는 CPLD(252)의 레지스터 공간에 접근하여 부착된 eSSD들 250 각각의 구성을 설정하고, 상태를 읽고, 또는 행동들(또는 동작들)을 수행할 수 있다. CPLD(252)의 레지스터는 구성 레지스터(271), 활성 BMC ID(272), 슬롯 ID(273), 및 PRSNT(present; 존재) 레지스터(274)를 포함하나, 이에 한정되지 않는다. 활성 BMC ID(272)는 2개 스위치보드들(201a, 201b)의 BMC들(203a, 203b) 중 활성 BMC의 식별자를 나타낸다. 슬롯 ID(273)는 데이터 스토리지 시스템의 드라이버 슬롯들, 예를 들어, 1-24, 각각에 대한 식별자를 나타낸다. PRSNT 레지스터(274)는 슬롯 ID(273)에 의해 식별된 것으로서 해당 슬롯 내 각 eSSD 250의 존재/부재의 상태를 나타낸다.

[0049] 시스템이 부팅되거나 리셋 이후에, CPLD(252)는 제 2 스위치보드(201b)가 섀시에 존재하는지 그렇지 않은지 관계없이 제 1 스위치보드(201a) 상의 BMC(203a)를 활성 BMC로서 CPLD(252)에 의해 디폴트(default)로 지정한다. 활성 BMC(203a)가 파워 다운되거나 고장나면, 다른 BMC(203b)가 활성 BMC 역할을 가질 수 있다. 디폴트 BMC(203a)가 고장에서 복구되면, BMC(203a)는 BMC(203b)로부터 활성 역할을 넘겨 받는다.

[0050] 활성 BMC는 활성 BMC 및 타깃 eSSD 250 사이의 SMBus/I2C 연결을 갖는 특정한 적절한 슬롯 ID(273)를 갖는 레

지스터 상태를 읽거나 써서 각 eSSD 250을 구성할 수 있다. 연결이 수립된 이후에, BMC(203)는 명령을 타깃 eSSD 250의 센서 어드레스로 전송한다. BMC(203)는 또한 드라이브 및 슬롯 상태 레지스터들을 읽을 수 있고 그리고 타깃 eSSD에 대한 의도된 행동들(또는 동작들)을 수행할 수 있다.

[0051] 도 3은 일 실시 예에 따라, 2개 소형 폼팩터 SSD들을 포함하는 예시적인 데이터 스토리지 장치를 도시한다. 데이터 스토리지 장치(350)는 제어 로직 장치(351, 예를 들면, FPGA 또는 ASIC), 각 SSD(예를 들어, NF1 SSD)의 표면적을 제공하기 위한 2개 상부 구성 요소 영역들(353a, 353b), 및 제어 로직 장치(351)의 프로세서 서브시스템(processor subsystem; PS) 및 프로그래머블 로직(programmable logic; PL) 각각에 연결되는 메모리들(352a, 352b)를 포함한다. 제어 로직 장치(351)는 다른 구성 요소들, 예를 들어, 전압 레귤레이터(voltage regulator), I2C 스위치와 함께 데이터 스토리지 장치(350)의 PCB의 일 측에 배치될 수 있다. 제어 로직 장치(351)는 커넥터(예를 들어, U.2 커넥터, M.2 커넥터)를 통해 도 1b의 스위치보드(101) 상의 PCIe 스위치(예를 들어, PCIe 스위치(106))와 통신할 수 있고 그리고 따라서 데이터 스토리지 장치(350)에 포함된 2개 소형 폼팩터 SSD들을 구성할 수 있고 동작시킬 수 있다. 일 실시 예에 따라, 데이터 스토리지 장치(350)는 HA 모드 또는 non-HA 모드에서 동작하도록 구성될 수 있다. PCIe RP가 2개 X4 장치들 또는 DualPort_En# 핀이 표명된(asserted) 것을 감지하면, PCIe EP#2 및 이더넷 포트들이 활성화된다. 또한, 데이터 스토리지 장치(350)는 도 1a의 데이터 스토리지 장치(150)로서 핫 스왑이 가능할 수 있다(예를 들어, 파워 오프없이 삽입되거나 제거될 수 있음). 도 3에서, 도시된 데이터 스토리지 장치(350)의 블록 배치는 각각의 SSD들의 상부 구성 요소 영역들(353)이 제어 로직 장치(351) 및 메모리들(353)(그리고 다른 구성 요소들)로부터 이격되어 배치되는 것을 도시한다. 그러나, 물리적 배치에서, 데이터 스토리지 장치(350)에 포함되는 2개 SSD들의 상부 구성 요소 영역들(353)은 부분적으로 또는 완전히 제어 로직 장치(351)의 영역과 겹쳐서 데이터 스토리지 장치(350)가 차지하는 공간을 줄일 수 있다.

[0052] 일 실시 예에 있어서, 제어 로직 장치(351)는 파워-손실(power-loss; 또는 전원 손실) 보호를 제공한다. 일반적인 파워-오프 주기 동안, 호스트 컴퓨터는 시간을 할당하여 데이터 스토리지 장치(350)로 대기 명령을 전송함으로써 데이터 무결성을 보존한다. 예상치 못한 파워 손실의 경우, 데이터 스토리지 장치(350)의 내부 버퍼들(예를 들면, DRAM) 내의 캐시(cache)된 데이터가 손실될 수 있다. 캐시된 데이터의 손실은 유저가 섀시로부터 데이터 스토리지 장치(350)를 뽑는 경우뿐만 아니라 예상치 못한 정전(power outage)으로 발생할 수 있다. 제어 로직 장치(351)는 예상치 못한 파워 셧다운(power shutdown; 전력 차단)으로 인한 데이터 손실을 방지할 수 있다. 고장 또는 핫 스왑의 감지 시, 데이터 스토리지 장치(350)에 포함된 SSD들은 즉시 그 안에(예를 들어, 탄탈륨 커패시터) 저장된 에너지를 이용하여 버퍼 내 캐시된 데이터를 플래시 메모리로 전송하는데 충분한 시간을 제공하여 데이터의 손실 방지를 보장한다. 파워 손실 보호를 위해 버퍼로부터 플래시 메모리로의 데이터 전송의 타이밍 및 대기 시간(latency)은 데이터 스토리지 장치(350)의 유형에 따라 변할 수 있다.

[0053] 도 4는 일 실시 예에 따라, 1U 섀시에서 구현되는 예시적인 데이터 스토리지 시스템의 블록도를 도시한다. 1U 섀시는 스위치보드(401) 및 미드플레인(451)을 포함할 수 있다. 스위치보드(401)는 CPU(402, 예를 들어, X86 프로세서), BMC(403), PCIe 스위치(406), 이더넷 스위치 컨트롤러(407), 관리 LAN 포트(416), 업링크 포트들(414), 및 미드플레인(451)을 통해 하나 이상의 eSSD들(450)과 인터페이스하는 고속 커넥터(415)를 포함할 수 있다. 스위치보드(401)는 도 1b의 스위치보드들(111a, 111b)과 동일할 수 있다. 도 1b의 동일한 스위치보드들(111a, 111b)이 재사용될 수 있으나 미드플레인(451)은 1U 섀시에서 기능하기 위해 새로운 제어 로직(예를 들어, FPGA 및 ASIC)을 필요로 할 수 있다.

[0054] 일 실시 예에 따라, 미드플레인(451)은 하나 이상의 종단 SSD 컨트롤러(455)를 포함한다. 각 종단 SSD 컨트롤러(455)는 최대 4개 데이터 스토리지 장치들(450, 예를 들어 eSSD들)을 지원하고 제어할 수 있다.

[0055] 일 실시 예에 따라, 미드플레인(451)은 부착된 데이터 스토리지 장치들(450)을 파워 온/오프하기 위해 BMC(403)를 위한 보호 메커니즘과의 통신을 제공하도록 구성된 IC2 스위치(452)를 포함한다. BMC(403)는 각 그룹 내 데이터 스토리지 장치들(450)이 동일한 센서 어드레스를 갖는 제약으로 인해 각 데이터 스토리지 장치(450)와 한번에 하나씩 통신할 수 있다. NF1 폼팩터를 갖는 데이터 스토리지 장치(450)는 종단 SSD 컨트롤러의 각 드라이브 슬롯에 부착될 수 있다. 각 종단 SSD 컨트롤러(455)는 4개 데이터 스토리지 장치들(450)의 엮인(또는 결합된) 그룹을 파워 온/오프할 수 있다.

[0056] 도 5는 일 실시 예에 따라, 도 4의 데이터 스토리지 시스템(400) 내에서 이용되는 예시적인 데이터 스토리지 장치를 도시한다. 데이터 스토리지 장치(450)는 커넥터(예를 들어, U.2 커넥터, M.2 커넥터)를 통해 도 4의 종단 SSD 컨트롤러(455)에게 부착되기에 적합한 NF1 폼팩터(M.3 폼팩터 또는 NGSFF로 지칭될 수도 있음)를 가질 수 있다. 데이터 스토리지 장치(450)는 상부 구성 요소 영역(553)에 탑재되는(부착되는) 하나 이상의 플래시 장치

들(예를 들어, SSD들)과 통신하도록 구성되는 제어 로직 장치(551, 예를 들면, ASIC)을 포함할 수 있다.

[0057] 일 실시 예에 있어서, 랙-장착형 데이터 스토리지 시스템은 포함한다: 하나 이상의 스위치보드들을 포함하는 섀시; 하나 이상의 스위치보드들과 인터페이스하는 미드플레인; 그리고 커넥터를 이용하여 미드플레인과 탈착 가능하게 연결된 하나 이상의 데이터 스토리지 장치들. 하나 이상의 데이터 스토리지 장치들 중 적어도 하나의 데이터 스토리지 장치는 미드플레인과 인터페이스하는 로직 장치를 포함한다. 로직 장치는 미드플레인과 해당 데이터 스토리지 장치의 장치-특정 인터페이스를 제공한다. 적어도 하나의 데이터 스토리지 장치는 커넥터의 핀 상의 신호에 기초하여 제 1 프로토콜에 따라 상기 로직 장치를 이용하여 구성되고, 그리고 적어도 하나의 데이터 스토리지 장치는 로직 장치를 이용하여 커넥터의 핀 상의 신호의 변화에 기초하여 제 2 프로토콜에 따라 재구성 가능하다.

[0058] 로직 장치는 FPGA(field-programmable gate array) 또는 ASIC(application-specific integrated circuit)일 수 있다.

[0059] 커넥터는 U.2 커넥터일 수 있다.

[0060] 커넥터는 M.2 커넥터일 수 있다..

[0061] 하나 이상의 데이터 스토리지 장치들은 NF1(new form factor 1) 표준과 호환 가능한 적어도 하나의 NF1 SSD(solid-state drive)를 포함할 수 있다.

[0062] 하나 이상의 데이터 스토리지 장치들은 적어도 하나의 NVMe SSD 및 NVMe-oF SSD를 포함할 수 있다.

[0063] 핀은 커넥터의 공급자에 의해 정의된 예약된 핀일 수 있다.

[0064] 섀시는 2U 섀시이고, 랙-장착형 데이터 스토리지 시스템은 2U 섀시에서 서로의 위에 배치되는 제 1 스위치보드 및 제 2 스위치보드를 포함할 수 있다.

[0065] 미드플레인은 24개 SSD들을 삽입하는 구동 장치 공간을 가질 수 있다.

[0066] 랙-장착형 데이터 스토리지 시스템은 1U 섀시 내의 1개 스위치보드를 포함할 수 있다.

[0067] 랙-장착형 데이터 스토리지 시스템은 2U 섀시에서 나란히 배치되는 제 1 스위치보드 및 제 2 스위치보드를 포함할 수 있다.

[0068] 미드플레인은 하나 이상의 데이터 스토리지 장치들의 존재를 감지하고 하나 이상의 데이터 스토리지 장치들의 동작을 관리하는 하나 이상의 스위치보드들의 BMC(baseboard management board)와 인터페이스를 제공하기 위한 프로그래머블 로직 장치를 포함할 수 있다.

[0069] 미드플레인은 48개 SSD들을 삽입하는 구동 장치 공간을 가질 수 있다.

[0070] 미드플레인은 HA(high-availability) 모드 또는 non-HA 모드에서 동작하도록 구성 가능한 하나 이상의 데이터 스토리지 장치들을 구성할 수 있다.

[0071] 다른 실시 예에 있어서, 프로그래머블 로직 장치는 포함한다: 레지스터; 하나 이상의 BMC(baseboard management board) 멀티플렉서들; 및 슬롯 멀티플렉서. 프로그래머블 로직 장치는 데이터 스토리지 시스템의 미드플레인에 집적, 구현, 또는 통합되고, 그리고 미드플레인은 하나 이상의 데이터 스토리지 장치들을 삽입하기 위한 복수의 커넥터를 포함하는 구동 장치 공간을 포함한다. 프로그래머블 로직 장치는 하나 이상의 BMC 멀티플렉서들의 각각의 BMC 멀티플렉서들을 이용하여 하나 이상의 스위치보드들의 하나 이상의 BMC들에게 인터페이스를 제공하여 하나 이상의 데이터 스토리지 장치들의 동작을 관리한다. 하나 이상의 BMC 멀티플렉서들 각각은 슬롯 멀티플렉서와 연결되어 하나 이상의 스위치보드들의 BMC들이 하나 이상의 데이터 스토리지 장치들과 동시에 통신 가능하게 한다. 프로그래머블 로직 장치는 하나 이상의 BMC들 각각이 레지스터에 접근하도록 함으로써 제 1 프로토콜에 따라 하나 이상의 데이터 스토리지 장치들로 장치-특정 인터페이스를 제공한다. 프로그래머블 로직 장치는 적어도 하나의 데이터 스토리지 장치에 대응하는 복수의 커넥터들의 커넥터의 핀의 변화에 기초하여 제 2 프로토콜에 따라 하나 이상의 데이터 스토리지 장치들 중 적어도 하나의 데이터 스토리지 장치를 재구성 가능하다.

[0072] 프로그래머블 로직 장치는 데이터 스토리지 시스템의 하나 이상의 스위치보드들과 인터페이스할 수 있다..

[0073] 프로그래머블 로직 장치는 FPGA(field-programmable gate array)일 수 있다.

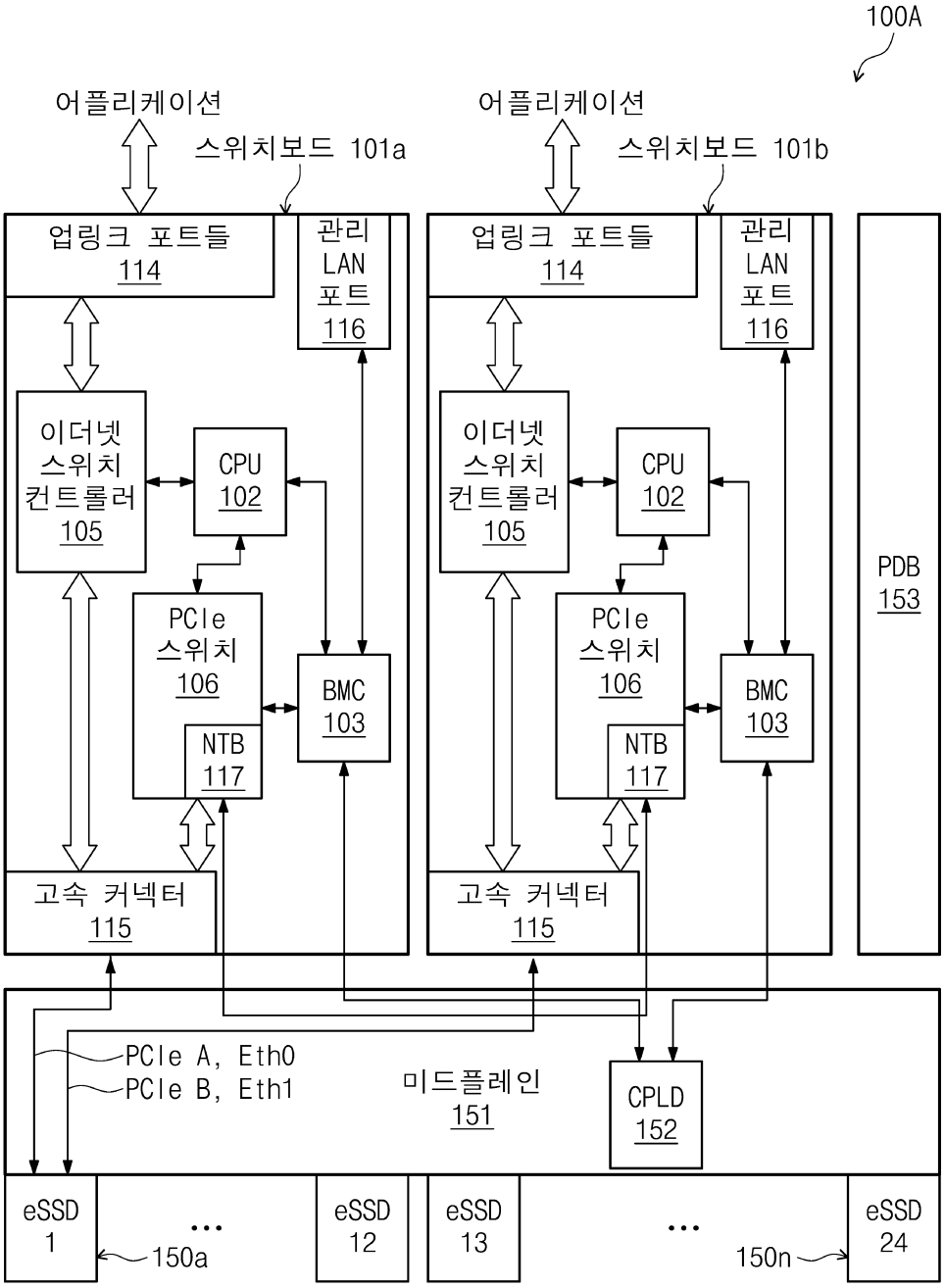
[0074] 레지스터는 구성 레지스터, 활성 BMC ID, 슬롯 ID, 및 존재 레지스터를 포함할 수 있다.

[0075] 핀은 상기 커넥터의 공급자에 의해 정의된 예약된 핀일 수 있다.

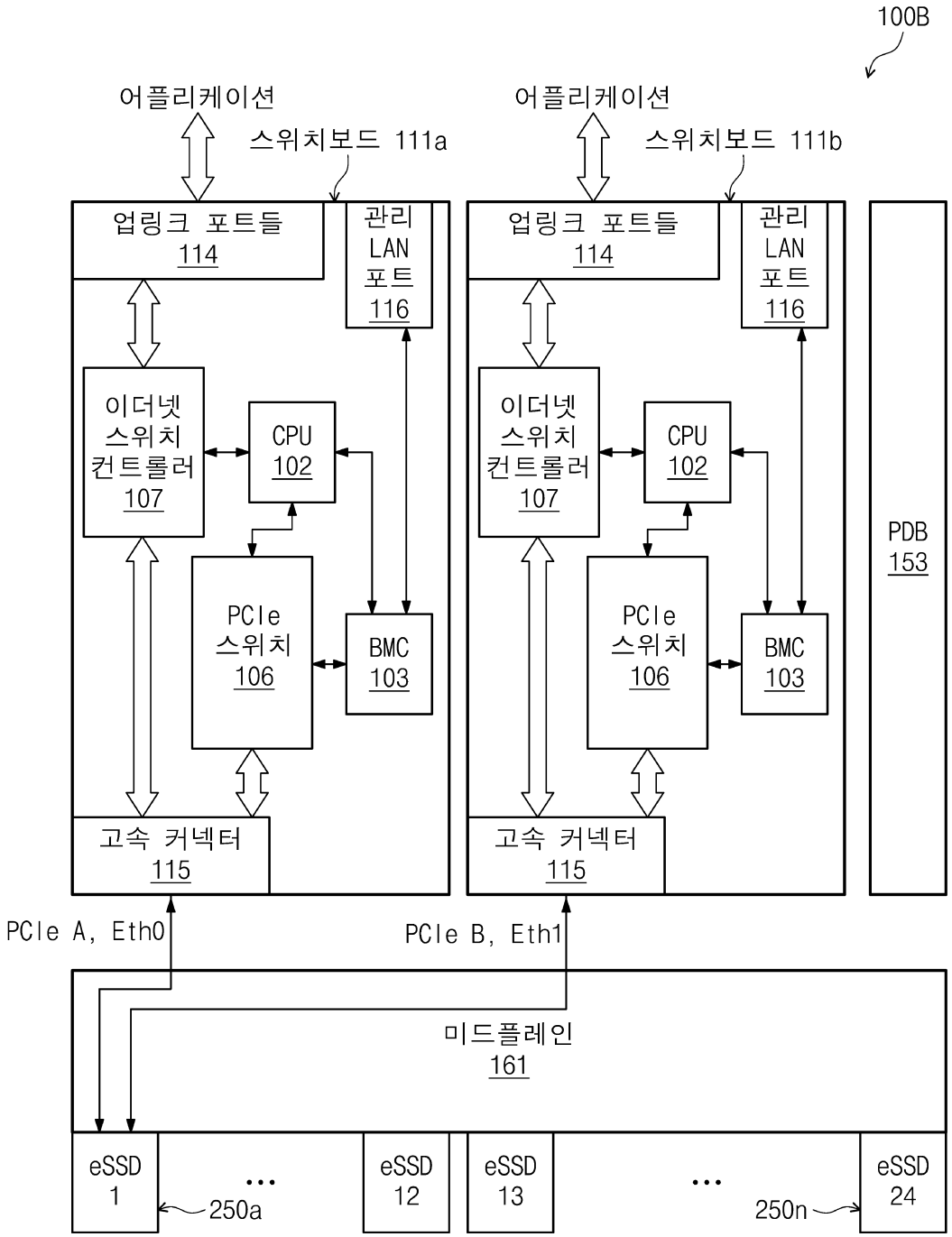
[0076] 전술한 예시적인 실시 예들은 다수의 솔리드-스테이트 드라이브들을 지원하기 위한 모듈러 시스템 아키텍처를 제공하기 위한 시스템 및 방법을 구현하는 다양한 실시 예들을 설명하기 위해 본 명세서에서 설명되었다. 개시된 예시적인 실시 예들로부터 다양한 수정들 및 이탈들은 당업자에게 발생할 것이다. 본 발명의 범위 내에 있는 것으로 의도된 주제는 이하의 청구항들에서 제공된다.

도면

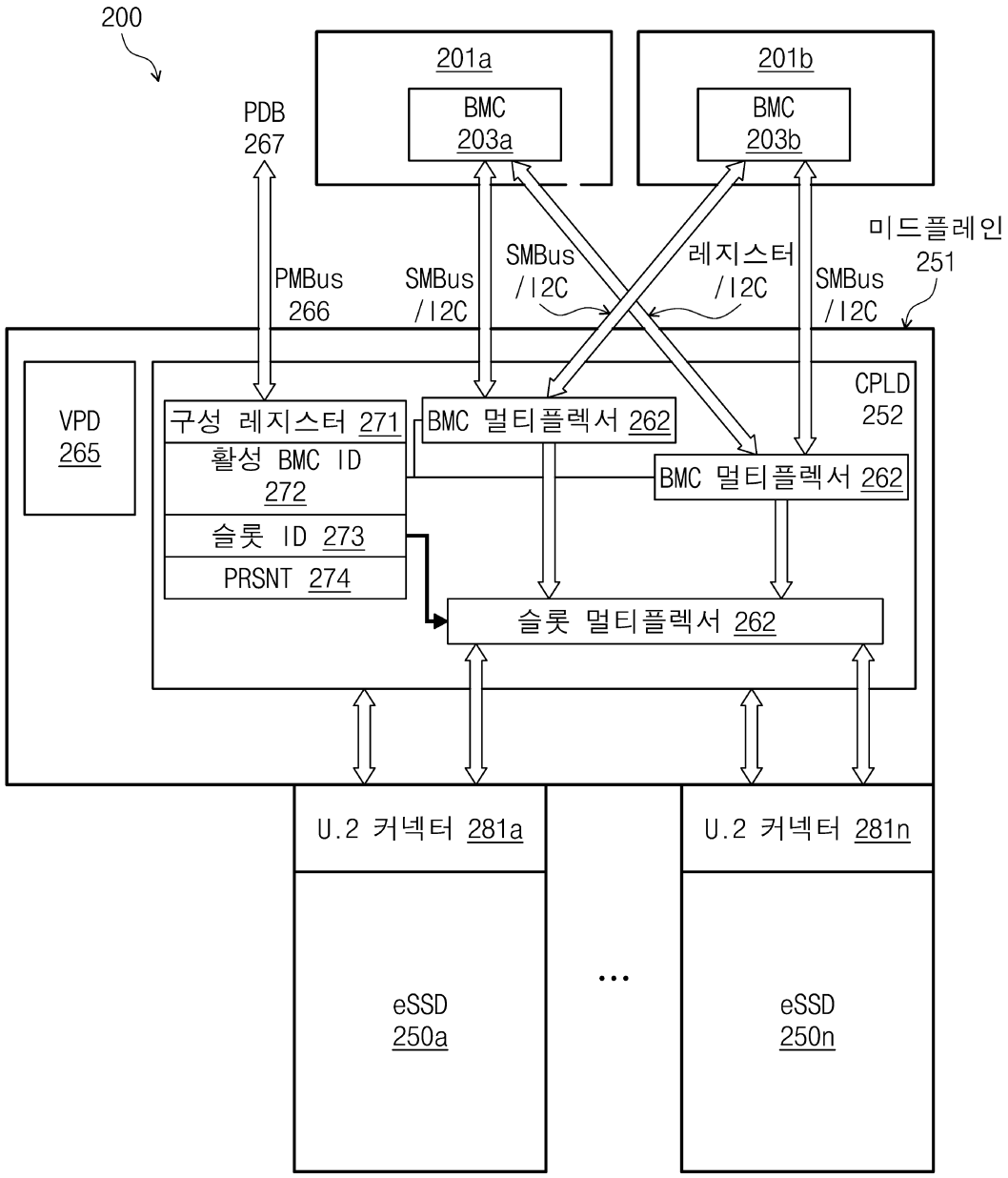
도면1a



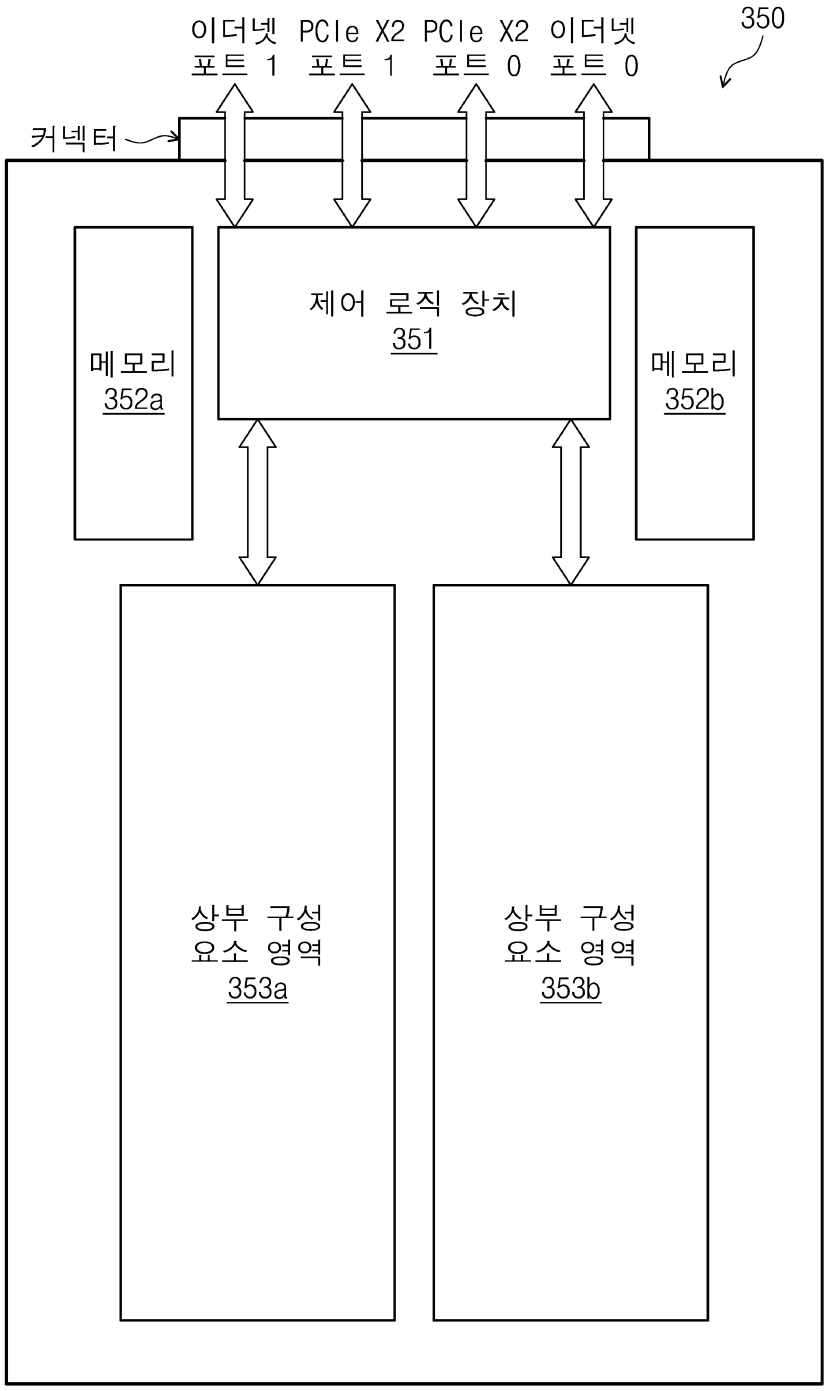
도면1b


100B
어플리케이션
스위치보드 111a
어플리케이션
스위치보드 111b
업링크 포트들
114
관리
LAN
포트
116
이더넷
스위치
컨트롤러
107
CPU
102
PCIe
스위치
106
BMC
103
고속 커넥터
115
업링크 포트들
114
관리
LAN
포트
116
이더넷
스위치
컨트롤러
107
CPU
102
PCIe
스위치
106
BMC
103
고속 커넥터
115
PDB
153
PCIe A, Eth0
PCIe B, Eth1
미드플레인
161
eSSD
1
250a
eSSD
12
eSSD
13
250n
eSSD
24

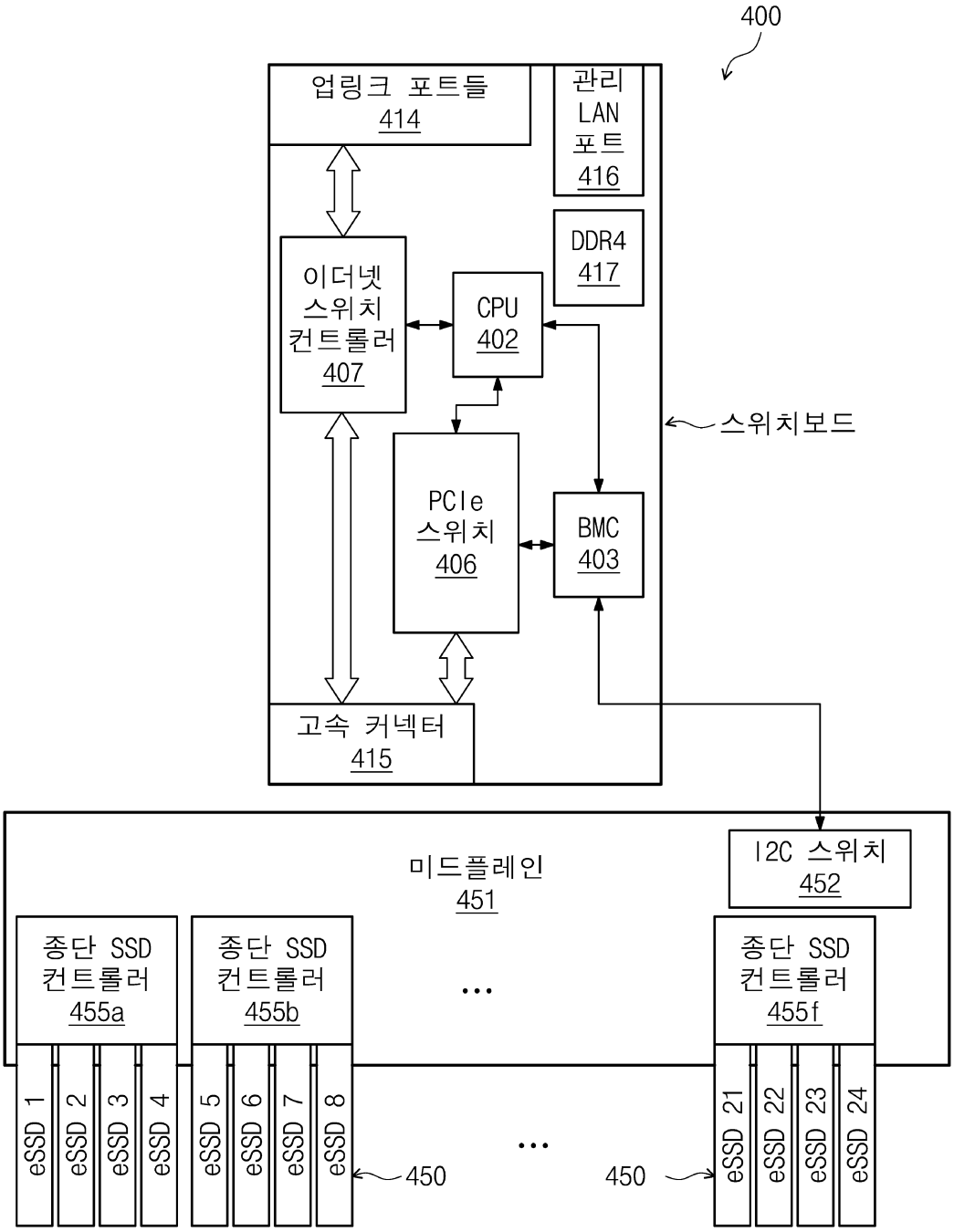
도면2


200
201a
BMC
203a
201b
BMC
203b
PDB
267
PMBus
266
SMBus
/I2C
SMBus
/I2C
레지스터
/I2C
SMBus
/I2C
미드플레인
251
CPLD
252
VPD
265
구성 레지스터 271
활성 BMC ID
272
슬롯 ID 273
PRSNT 274
BMC 멀티플렉서 262
BMC 멀티플렉서 262
슬롯 멀티플렉서 262
U.2 커넥터 281a
U.2 커넥터 281n
eSSD
250a
eSSD
250n

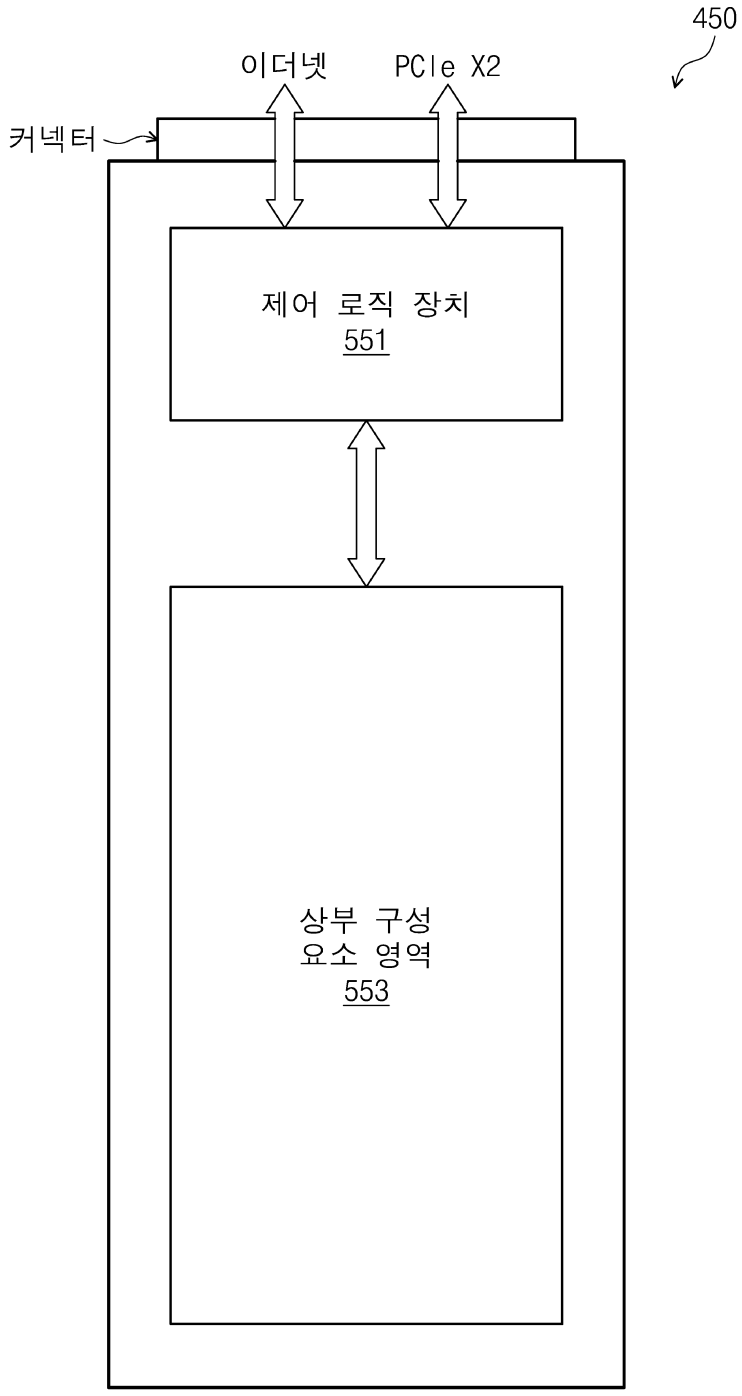
도면3


이더넷 포트 1
PCIe X2 포트 1
PCIe X2 포트 0
이더넷 포트 0
350
커넥터
메모리 352a
제어 로직 장치 351
메모리 352b
상부 구성 요소 영역 353a
상부 구성 요소 영역 353b

도면4


400
업링크 포트들
414
관리
LAN
포트
416
DDR4
417
이더넷
스위치
컨트롤러
407
CPU
402
스위치보드
PCIe
스위치
406
BMC
403
고속 커넥터
415
I2C 스위치
452
미드플레인
451
종단 SSD
컨트롤러
455a
종단 SSD
컨트롤러
455b
종단 SSD
컨트롤러
455f
eSSD 1
eSSD 2
eSSD 3
eSSD 4
eSSD 5
eSSD 6
eSSD 7
eSSD 8
eSSD 21
eSSD 22
eSSD 23
eSSD 24
450
450

도면5



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 8

【변경전】

제 1 항에 있어서,

상기 섀시는 2U 새시이고, 그리고

상기 스토리지 시스템은 상기 2U 섀시에서 서로의 위에 배치되는 제 1 스위치보드 및 제 2 스위치보드를 포함하는 스토리지 시스템.

【변경후】

제 1 항에 있어서,

상기 섀시는 2U 섀시이고, 그리고

상기 스토리지 시스템은 상기 2U 섀시에서 서로의 위에 배치되는 제 1 스위치보드 및 제 2 스위치보드를 포함하는 스토리지 시스템.