(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2025년04월23일
(11) 등록번호 10-2799275
(24) 등록일자 2025년04월17일

(51) 국제특허분류(Int. Cl.)
G06F 15/16 (2018.01) *G06F 9/54* (2018.01)
(52) CPC특허분류
G06F 15/16 (2013.01)
G06F 9/546 (2013.01)
(21) 출원번호 10-2017-0025700
(22) 출원일자 2017년02월27일
심사청구일자 2022년02월18일
(65) 공개번호 10-2017-0117310
(43) 공개일자 2017년10월23일
(30) 우선권주장
62/322,035 2016년04월13일 미국(US)
15/209,566 2016년07월13일 미국(US)
(56) 선행기술조사문헌
KR1020100085564 A*
(뒷면에 계속)
전체 청구항 수 : 총 20 항

(73) 특허권자
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
골라푸디, 벤카타 바누 프라케쉬
미국 캘리포니아주 95051 산타클라라 홈스테드 로드 3455 아파트 60
자쿨라, 비자야
미국 캘리포니아주 95136 산호세 밸리코어 서클 309
(74) 대리인
특허법인 고려

심사관 : 임재우

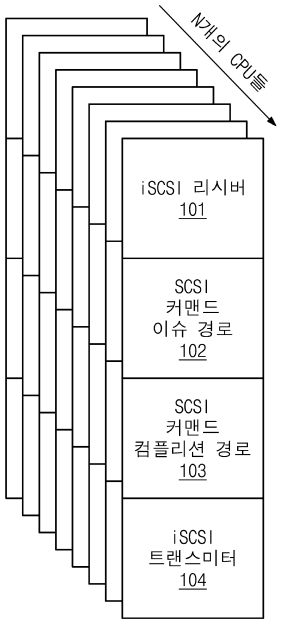
(54) 발명의 명칭 **고성능의 락 없는 스케일러블 타깃을 위한 시스템 및 방법**

(57) 요 약

데이터 스토리지 시스템의 타깃에 저장된 데이터를 액세스하기 위한 방법은, 제 1 그룹의 CPU들 상에서 복수의 전송 스레드들을 실행하는 단계를 포함할 수 있다. 상기 복수의 전송 스레드들 각각은 커맨드 리시버 경로 및 커맨드 트랜스미터 경로를 포함할 수 있다. 상기 방법은 제 2 그룹의 CPU들 상에서 복수의 데이터 경로 스레드들을

(뒷면에 계속)

대 표 도 - 도1



실행하는 단계를 포함할 수 있다. 상기 복수의 데이터 경로 스레드들 각각은 커맨드 이슈 경로 및 커맨드 컴플리션 경로를 포함할 수 있다. 상기 방법은 전송 스레드의 커맨드 리시버 경로를 사용하여 I/O 커맨드 이슈 큐에 I/O 커맨드를 배치하고, 데이터 경로 스레드의 커맨드 이슈 경로를 사용하여 상기 I/O 커맨드를 처리하는 단계, 그리고 상기 데이터 경로 스레드의 상기 커맨드 컴플리션 경로를 사용하여 I/O 컴플리션 큐에 I/O 컴플리션 통지를 배치하고, 상기 전송 스레드의 상기 커맨드 트랜스미터 경로를 사용하여 상기 I/O 컴플리션 통지를 처리하는 단계를 포함할 수 있다.

(52) CPC특허분류
G06F 2213/0036 (2013.01)

(56) 선행기술조사문헌
KR1020150033507 A*
KR101399833 B1
KR1020140131978 A
KR1020150023845 A
KR1020150000413 A
*는 심사관에 의하여 인용된 문헌

명 세 서

청구범위

청구항 1

복수의 CPU들 및 소프트웨어 모듈들 세트를 포함하는 데이터 스토리지 시스템의 타깃에 저장된 데이터를 액세스하기 위한 커맨드들을 상기 소프트웨어 모듈들 세트에 의해 수신하는 단계;

상기 복수의 CPU들 중 제 1 그룹의 CPU들 상에서 복수의 전송 스레드들을 실행하는 단계로써, 상기 복수의 전송 스레드들 각각은 커맨드 리시버 경로 및 커맨드 트랜스미터 경로를 포함하는 것;

상기 복수의 CPU들 중 제 2 그룹의 CPU들 상에서 복수의 데이터 경로 스레드들을 실행하는 단계로써, 상기 복수의 데이터 경로 스레드들 각각은 커맨드 이슈 경로 및 커맨드 컴플리션 경로를 포함하는 것;

전송 스레드의 상기 커맨드 리시버 경로를 사용하여 I/O 커맨드 이슈 큐에 I/O 커맨드를 배치하고, 데이터 경로 스레드의 상기 커맨드 이슈 경로를 사용하여 상기 I/O 커맨드를 처리하는 단계; 그리고

상기 데이터 경로 스레드의 상기 커맨드 컴플리션 경로를 사용하여 I/O 컴플리션 큐에 I/O 컴플리션 통지를 배치하고, 상기 전송 스레드의 상기 커맨드 트랜스미터 경로를 사용하여 상기 I/O 컴플리션 통지를 처리하는 단계를 포함하되,

상기 I/O 커맨드 이슈 큐는 제 1 어레이의 큐들을 포함하고, 상기 제 1 어레이의 큐들 각각은 상기 제 1 그룹의 CPU들의 각각의 CPU에 대응하고,

상기 I/O 컴플리션 큐는 제 2 어레이의 큐들을 포함하고, 상기 제 2 어레이의 큐들 각각은 상기 제 2 그룹의 CPU들의 각각의 CPU에 대응하는 방법.

청구항 2

제 1 항에 있어서,

상기 전송 스레드는 리시버 및 트랜스미터를 포함하고, 상기 리시버는 I/O 커맨드들을 상기 데이터 경로에 전송하는 방법.

청구항 3

제 2 항에 있어서,

상기 데이터 경로 스레드는 상기 I/O 커맨드를 상기 타깃으로 전송하고 상기 타깃으로부터 상태 및 데이터 중 적어도 하나를 수신하고, 상기 전송 스레드의 상기 트랜스미터로 상기 상태 및 상기 데이터 중 상기 적어도 하나를 전송하는 방법.

청구항 4

제 1 항에 있어서,

상기 제 1 어레이의 큐들은 상기 데이터 경로 스레드들에 할당된 상기 제 2 그룹의 CPU들에 대응하는 제 1 복수의 노드들을 포함하는 방법.

청구항 5

제 4 항에 있어서,

상기 제 1 복수의 노드들은 헤더 노드, 테일 노드, 및 상기 제 1 어레이의 큐들의 큐를 가리키는 현재의 노드를 포함하되,

상기 현재의 노드부터 상기 테일 노드까지의 노드들은 소비자에 의해 소유되고, 상기 제 1 어레이의 큐들 중 잔존하는 노드들은 생산자에 의해 소유되는 방법.

청구항 6

제 5 항에 있어서,

상기 생산자는 이니시에이터이고, 상기 소비자는 상기 타깃인 방법.

청구항 7

제 6 항에 있어서,

상기 소비자는 상기 데이터 경로 스레드들 각각에 특유하는 스레드 식별자를 이용하여 상기 제 1 어레이의 큐들로의 배타적인 액세스를 획득하는 방법.

청구항 8

제 1 항에 있어서,

상기 제 2 어레이의 큐들은 상기 전송 스레드들에 할당된 상기 제 1 그룹의 CPU들에 대응하는 제 2 복수의 노드들을 포함하는 방법.

청구항 9

제 8 항에 있어서,

상기 제 2 복수의 노드들은 헤더 노드, 테일 노드, 및 상기 제 2 어레이의 큐들의 큐를 가리키는 현재의 노드를 포함하되,

상기 현재의 노드부터 상기 테일 노드까지의 노드들은 소비자에 의해 소유되고, 상기 제 2 어레이의 큐들 중 잔존하는 노드들은 생산자에 의해 소유되는 방법.

청구항 10

제 9 항에 있어서,

상기 생산자는 상기 타깃이고 상기 소비자는 이니시에이터인 방법.

청구항 11

제 10 항에 있어서,

상기 소비자는 상기 전송 스레드들 각각에 특유하는 스레드 식별자를 이용하여 상기 제 2 어레이의 큐들로의 배타적인 액세스를 획득하는 방법.

청구항 12

제 1 항에 있어서,

상기 I/O 커맨드 이슈 큐와 상기 I/O 컴플리션 큐 각각은 MPMC (multi-producer multi-consumer) 락 없는 큐이고,

하나 또는 그 이상의 전송 스레드들로부터의 작업 요청들은 특정한 데이터 경로 스레드에 도달하고, 개별적인 전송 스레드로부터의 작업 요청들은 하나 또는 그 이상의 데이터 경로 스레드들에 도달하고,

하나 또는 그 이상의 데이터 경로 스레드들로부터의 작업 요청들은 개별적인 전송 스레드에 도달하고, 개별적인 데이터 경로 스레드로부터의 작업 요청들은 하나 도는 그 이상의 전송 스레드들에 도달하는 방법.

청구항 13

복수의 CPU들;

복수의 타깃들; 그리고

I/O 흐름들과 I/O 커맨드 이슈 큐를 처리하기 위한 소프트웨어 모듈들 세트, 및 I/O 컴플리션 큐를 저장하는 메모리를 포함하되,

상기 소프트웨어 모듈들 세트는:

복수의 CPU들을 포함하는 데이터 스토리지 시스템의 타깃에 저장된 데이터를 액세스하기 위한 커맨드들을 수신하고;

상기 복수의 CPU들의 제 1 그룹의 CPU들 중 제 1 CPU 상에서 복수의 전송 스레드들을 실행하되, 상기 복수의 전송 스레드들 각각은 커맨드 리시버 경로 및 커맨드 트랜스미터 경로를 포함하고; 그리고

상기 복수의 CPU들의 제 2 그룹의 CPU들 중 제 2 CPU 상에서 복수의 데이터 경로 스레드들을 실행하되, 상기 복수의 데이터 경로 스레드들 각각은 커맨드 이슈 경로 및 커맨드 컴플리션 경로를 포함하고;

전송 스레드의 상기 커맨드 리시버 경로는 I/O 커맨드를 상기 I/O 커맨드 이슈 큐에 배치하고, 데이터 경로 스레드의 상기 커맨드 이슈 경로는 상기 I/O 커맨드를 처리하고,

상기 데이터 경로 스레드의 상기 커맨드 컴플리션 경로는 I/O 컴플리션 통지를 상기 I/O 컴플리션 큐에 배치하고, 상기 전송 스레드의 상기 커맨드 트랜스미터 경로는 상기 I/O 컴플리션 통지를 처리하고,

상기 I/O 커맨드 이슈 큐는 제 1 어레이의 큐들을 포함하되, 상기 제 1 어레이의 큐들 각각은 상기 제 1 그룹의 CPU들의 각각의 CPU에 대응하고, 그리고

상기 I/O 컴플리션 큐는 제 2 어레이의 큐들을 포함하되, 상기 제 2 어레이의 큐들 각각은 상기 제 2 그룹의 CPU들의 각각의 CPU에 대응하는 데이터 스토리지 시스템.

청구항 14

제 13 항에 있어서,

상기 전송 스레드는 리시버와 트랜스미터를 포함하고,

상기 리시버는 I/O 커맨드를 상기 데이터 경로에 전송하는 데이터 스토리지 시스템.

청구항 15

제 13 항에 있어서,

상기 데이터 경로 스레드는 상기 I/O 커맨드를 상기 타깃으로 전송하고 상기 타깃으로부터 상태 및 데이터 중 적어도 하나를 수신하고, 상기 전송 스레드의 상기 트랜스미터로 상기 상태 및 상기 데이터 중 상기 적어도 하나를 전송하는 데이터 스토리지 시스템.

청구항 16

제 13 항에 있어서,

상기 제 1 어레이의 큐들은 상기 데이터 경로 스레드들에 할당된 상기 제 2 그룹의 CPU들에 대응하는 제 1 복수의 노드들을 포함하는 데이터 스토리지 시스템.

청구항 17

제 16 항에 있어서,

상기 제 1 복수의 노드들은 헤더 노드, 테일 노드, 및 상기 제 1 어레이의 큐들의 큐를 가리키는 현재의 노드를 포함하되,

상기 현재의 노드부터 상기 테일 노드까지의 노드들은 소비자에 의해 소유되고, 상기 제 1 어레이의 큐들 중 잔존하는 노드들은 생산자에 의해 소유되는 데이터 스토리지 시스템.

청구항 18

제 17 항에 있어서,

상기 생산자는 이니시에이터이고, 상기 소비자는 상기 타깃인 데이터 스토리지 시스템.

청구항 19

제 18 항에 있어서,

상기 소비자는 상기 데이터 경로 스레드들 각각에 특유하는 스레드 식별자를 이용하여 상기 제 1 어레이의 큐들로의 배타적인 액세스를 획득하는 데이터 스토리지 시스템.

청구항 20

제 13 항에 있어서,

상기 제 2 어레이의 큐들은 상기 전송 스레드들에 할당된 상기 제 1 그룹의 CPU들에 대응하는 제 2 복수의 노드들을 포함하는 데이터 스토리지 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 데이터 스토리지 시스템에 관한 것으로, 좀 더 상세하게는, 고성능의 락 없는 스케일러블 타깃을 제공하기 위한 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 일반적인 SAN (storage area network)에서, 타깃은 지속적인 데이터 스토리지 공간들(예를 들어, 논리적 유닛 넘버(logical unit number; LUN), 명칭 공간(namespace))을 광섬유 커넥션 또는 스위칭 네트워크를 통하여 하나 또는 그 이상의 이니시에이터(initiator)들에게 노출시킨다. 이니시에이터는 인터페이스 세션(예를 들어, SCSI (small computer system interface) 세션)을 개시하고 커맨드(예를 들어, SCSI 커맨드)를 전송하는 엔드포인트(endpoint)를 일컫는다. 타깃은 이니시에이터의 작업 요청들을 기다리고 입/출력(I/O) 동작들을 수행하는 엔드포인트를 일컫는다. 일반적인 컴퓨터 아키텍처에서, 이니시에이터는 클라이언트로 일컬어지며, 타깃은 서버로 일컬어진다. 싱글 타깃은 복수의 이니시에이터들을 제공할 수 있으며 이니시에티더들에게 하나 또는 그 이상의 LUN들을 제공할 수 있다.

[0003] 타깃은 상호 협정된 SAN 프로토콜을 통하여 하나 또는 그 이상의 이니시에이터들과 통신할 수 있다. 예로써, SAN 프로토콜은 FCP (Fibre Channel Protocol), pSCSI (parallel SCSI), iSCSI (Internet small computer system interface), HyperSCSI, 파이버 채널 (Fibre Channel), ATA (Advanced Technology Attachment), SATA (Serial ATA), AoE (ATA over Ethernet), InfiniBand, 및 NVMe (Non-Volatile Memory Express) over Fabrics를 포함할 수 있으나, 이에 한정되지 않는다. SAN 프로토콜은 이니시에이터가 타깃으로 I/O 커맨드들을 전송하도록 허용한다. 데이터 센터의 데이터 스토리지 장치들은 스토리지 영역 네트워크를 통해 물리적 및/또는 논리적으로 분산될 수 있는 복수의 타깃들을 포함한다. SAN 프로토콜은, 데이터 스토리지 장치들이 국부적으로 부착된 것으로 보이게 하는 착각을 호스트에 제공하면서, 데이터 스토리지 장치들을 타깃들의 어레이들로 통합한다.

[0004] 스토리지 타깃들은 일반적으로 복수의 커넥션들을 통하여 백엔드(backend) LUN들을 복수의 이니시에이터들에게 노출시킬 수 있다. 각각의 이니이시에이터는 하나 또는 그 이상의 커넥션들을 타깃에게 개방할 수 있으며, 타깃에 있는 하나 또는 그 이상의 LUN들을 액세스할 수 있다. 입/출력(I/O) 흐름의 관점으로부터, 데이터 경로에서 확립된 복수의 동기점(synchronization point)들은, 특히, 복수의 세션들이 동시에 복수의 LUN들에 액세스할 때, I/O 성능에 영향을 미칠 수 있다.

[0005] 프론트엔드 I/O 스택(frontend I/O stack)에서의 데이터 경로는 전송 프로토콜 계층(layer)과 SCSI 코어 계층으로 나뉠 수 있다. 전송 프로토콜 계층에서의 처리는 커넥션-단위 기반(per-connection based)인 반면, SCSI 코어 계층에서의 처리는 LUN-단위 기반(per-LUN based)이다. 전송 프로토콜 계층에서의 처리는 I/O 커맨드들을 특정한 커넥션에 이슈(issue)하는 것과 그 특정한 커넥션에서 I/O 커맨드들을 완료하는 것을 포함한다. SCSI 코어 계층에서의 처리는 I/O 커맨드들을 특정한 LUN에 이슈하는 것과 특정한 LUN에 이슈된 I/O 커맨드들을 완료하는 것을 포함한다. 만일 전체의 I/O 경로가 전송 프로토콜 계층의 스레드 컨텍스트(thread context)에서 실행된다면, 다양한 동기점들은 LUN 레벨뿐만 아니라 커넥션 레벨에서도 확립될 수 있으므로, 따라서 전체적인 I/O 성능에 영향을 미칠 것이다. 이것은, SCSI 코어 계층에서의 I/O 컨텍스트는 LUN-특유함에 반해, 전송 프로토콜 계층에서의 I/O 컨텍스트는 커넥션-특유하기 때문이다.

발명의 내용

해결하려는 과제

[0006] 본 발명의 목적은 락 없는 스케일러블 타깃을 위한 시스템 및 방법을 제공함에 있다.

과제의 해결 수단

[0007] 본 발명의 실시 예에 따른 방법은: 복수의 CPU들을 포함하는 데이터 스토리지 시스템의 타깃에 저장된 데이터를 액세스하기 위한 커맨드들을 수신하는 단계; 상기 복수의 CPU들 중 제 1 그룹의 CPU들 상에서 복수의 전송 스레드들을 실행하는 단계로써, 상기 복수의 전송 스레드들 각각은 커맨드 리시버 경로 및 커맨드 트랜스미터 경로를 포함하는 것; 상기 복수의 CPU들 중 제 2 그룹의 CPU들 상에서 복수의 데이터 경로 스레드들을 실행하는 단계로써, 상기 복수의 데이터 경로 스레드들 각각은 커맨드 이슈 경로 및 커맨드 컴플리션 경로를 포함하는 것; 전송 스레드의 커맨드 리시버 경로를 사용하여 I/O 커맨드 이슈 큐에 I/O 커맨드를 배치하고, 데이터 경로 스레드의 커맨드 이슈 경로를 사용하여 상기 I/O 커맨드를 처리하는 단계; 그리고 상기 데이터 경로 스레드의 상기 커맨드 컴플리션 경로를 사용하여 I/O 컴플리션 큐에 I/O 컴플리션 통지를 배치하고, 상기 전송 스레드의 상기 커맨드 트랜스미터 경로를 사용하여 상기 I/O 컴플리션 통지를 처리하는 단계를 포함할 수 있다. 상기 I/O 커맨드 이슈 큐는 제 1 어레이의 큐들을 포함하고, 상기 제 1 어레이의 큐들 각각은 상기 제 1 그룹의 CPU들의 개별적인 CPU에 대응할 수 있다. 상기 I/O 컴플리션 큐는 제 2 어레이의 큐들을 포함하고, 상기 제 2 어레이의 큐들 각각은 상기 제 2 그룹의 CPU들의 개별적인 CPU에 대응할 수 있다.

[0008] 본 발명의 실시 예에 따른 데이터 스토리지 시스템은: 복수의 CPU들; 복수의 타깃들; 그리고 I/O 흐름들과 I/O 커맨드 이슈 큐를 처리하기 위한 소프트웨어 모듈들 세트, 및 I/O 컴플리션 큐를 저장하는 메모리를 포함할 수 있다. 상기 소프트웨어 모듈들 세트는: 복수의 CPU들을 포함하는 데이터 스토리지 시스템의 타깃에 저장된 데이터를 액세스하기 위한 커맨드들을 수신하고; 상기 복수의 CPU들의 제 1 그룹의 CPU들 중 제 1 CPU 상에서 복수의 전송 스레드들을 실행하되, 상기 복수의 전송 스레드들 각각은 커맨드 리시버 경로 및 커맨드 트랜스미터 경로를 포함하고; 그리고 상기 복수의 CPU들의 제 2 그룹의 CPU들 중 제 2 CPU 상에서 복수의 데이터 경로 스레드들을 실행하되, 상기 복수의 데이터 경로 스레드들 각각은 커맨드 이슈 경로 및 커맨드 컴플리션 경로를 포함할 수 있다. 전송 스레드의 상기 커맨드 리시버 경로는 상기 I/O 커맨드를 상기 I/O 커맨드 이슈 큐에 배치하고, 데이터 경로 스레드의 상기 커맨드 이슈 경로는 상기 I/O 커맨드를 처리할 수 있다. 상기 데이터 경로 스레드의 상기 커맨드 컴플리션 경로는 I/O 컴플리션 통지를 상기 I/O 컴플리션 큐에 배치하고, 상기 전송 스레드의 상기 커맨드 트랜스미터 경로는 상기 I/O 컴플리션 통지를 처리할 수 있다. 상기 I/O 커맨드 이슈 큐는 제 1 어레이의 큐들을 포함하되, 상기 제 1 어레이의 큐들 각각은 상기 제 1 그룹의 CPU들의 각각의 CPU에 대응할 수 있다. 상기 I/O 컴플리션 큐는 제 2 어레이의 큐들을 포함하되, 상기 제 2 어레이의 큐들 각각은 상기 제 2 그룹의 CPU들의 각각의 CPU에 대응할 수 있다.

[0009] 구현 및 사상들의 다양한 조합의 다양한 신규 사항을 포함하는, 앞의 그리고 다른 바람직한 특징들이 첨부된 도면을 참조하여 설명될 것이며 청구항에서 보다 구체적으로 설명될 것이다. 본 명세서에서 설명되는 시스템 및 방법은 단지 도시의 수단으로 보여주기 위한 것이며 한정하기 위한 것이 아님을 이해해야 한다. 본 발명이 속하는 기술 분야에서 통상의 지식을 지닌 자에게 잘 이해될 것과 같이, 본 명세서에서 설명되는 원리들 및 특징들은 본 발명의 기술 사상을 벗어나지 않는 범위 내에서 다양하고 수 많은 실시 예들에서 채택될 수 있다.

발명의 효과

[0010] 본 발명에 따르면 락 없는 스케일러블 타깃을 위한 시스템 및 방법을 제공할 수 있다.

도면의 간단한 설명

[0011] 본 명세서의 일부로써 포함되는 첨부된 도면은, 앞서 설명된 일반적인 설명 및 본 명세서에서 설명되는 원리를 설명하기 위해 이하 제공되는 바람직한 실시 예의 상세한 설명과 함께, 바람직한 실시 예를 도시한다.

도 1은 본 발명의 실시 예에 따른 스레딩 모델에서의 모놀리식(monolithic) 데이터 경로를 예시적으로 보여주는 도면이다.

도 2는 본 발명의 실시 예에 따른 예시적인 스레딩 모델의 블록도이다.

도 3은 본 발명의 실시 예에 따른 예시적인 나뉘어진 스레딩 모델을 보여주는 도면이다.

도 4는 본 발명의 실시 예에 따른 예시적인 스레딩 모델의 개략도를 보여주는 도면이다.

도 5는 본 발명의 실시 예에 따른 SPSC 락 없는 큐(single-producer single-consumer lockless queue) 를 보여주는 블록도이다.

도 6은 본 발명의 실시 예에 따른 예시적인 MPMC 락 없는 큐(multi-producer multi-consumer lockless queue)를 보여주는 블록도이다.

도 7은 본 발명의 실시 예에 따른 예시적인 스레딩 모델을 보여주는 블록도이다.

도 8은 본 발명의 다른 실시 예에 따른 예시적인 MPMC 락 없는 큐를 보여주는 블록도이다.

도 9는 본 발명의 또 다른 실시 예에 따른 예시적인 MPMC 락 없는 큐를 보여주는 블록도이다.

도면들은 반드시 스케일에 맞게 도시된 것은 아니며, 도시의 목적을 위해 유사한 구조들 또는 기능들의 구성 요소들은 일반적으로 유사한 참조 번호들에 의해 표현된다. 도면들은 본 명세서에서 설명되는 다양한 실시 예들의 상세한 설명을 가능하게 하기 위해 의도된 것이다. 도면들은 본 명세서에 개시된 것들의 교시의 모든 국면을 보여주지는 않는다.

발명을 실시하기 위한 구체적인 내용

[0012] 본 명세서에서 개시된 특징들 및 교시들 각각은 고성능 락 없는 스케일러블 타깃(lockless scalable target)을 제공하기 위한 다른 특징들 및 교시들과 함께 또는 분리되어 사용될 수 있다. 이러한 추가적인 특징들 및 교시들이 분리되거나 결합되어 이용되는 대표적인 예들은 첨부된 도면들을 참조하여 상세히 설명된다. 이러한 상세한 설명은 단지 본 교시들의 측면들을 실시하기 위한 기술 분야에서 상세한 지식을 가진 자를 교시하기 위한 것이고, 청구항들의 범위를 제한하지 않는다. 그러므로, 상세한 설명에 상술된 특징들의 조합들은 넓은 의미에서 교시를 실시할 필요가 없을 수도 있고, 대신에 본 교시들 특히 대표적인 실시 예들을 설명하기 위해 단지 교시된다.

[0013] 아래의 설명에서, 설명의 목적으로만 특정 명칭이 본 발명의 완전한 이해를 제공하기 위해 설명된다. 그러나, 이러한 특정 세부 사항들은 본 발명의 사상을 실시하기 위해 필요하지 않는 것은 통상의 기술자에게 명백할 것이다.

[0014] 상세한 설명의 몇몇 부분들은 알고리즘 및 컴퓨터 메모리 내 데이터 비트들에 대한 동작들의 심볼 표현의 측면에서 제공된다. 이들 알고리즘 설명들 및 표현들은 다른 분야의 통상의 기술자들에게 작업의 실체를 효과적으로 전달하기 위해, 데이터 처리 분야의 통상의 기술자들에 의해 사용된다. 여기에서 알고리즘은 일반적으로 소망하는 결과에 이르는 단계들에 대한 일관성 있는 순서일 수 있다. 단계들은 물리적 양의 물리적 조작이 필요한 것들이다. 일반적으로, 필수적이진 않지만, 이러한 양들은 저장, 전달, 결합, 비교, 그리고 다르게 조작될 수 있는 전기 또는 자기 신호의 형태를 취한다. 이러한 신호들을 비트들, 값들, 요소들, 심볼들, 특징들, 용어들, 숫자들 등으로 지칭하는 것이 주로 공통적인 사용의 이유로 때때로 편리하게 입증되었다.

[0015] 그러나, 이들 및 유사한 용어들 모두는 적절한 물리량과 연관되며 단지 이러한 양에 적용되는 편리한 라벨이라는 것을 명심해야 한다. 구체적으로는 아래의 설명으로부터 명백한 바와 같이, 설명에서, 이러한 “처리”, “컴퓨팅”, “계산”, “결정”, “표시” 등과 같은 용어를 사용하는 논의는 컴퓨터 시스템 또는 컴퓨터 레지스터 및 메모리들 내에서 물리적(전기적) 양으로서 나타나는 데이터를 컴퓨터 시스템 메모리들 또는 레지스터들 또는 다른 정보 스토리지, 전송 또는 표시 장치들 내에서 물리적 양으로서 나타나는 유사한 다른 데이터로 조작 및 변형하는 유사한 전자 컴퓨팅 장치의 활동 및 과정을 나타내는 것으로 이해된다.

[0016] 본 명세서에서 설명된 알고리즘은 본질적으로 임의의 특정 컴퓨터 또는 다른 장치들과 관련된 것이 아니다. 다양한 일반적인 목적의 시스템들, 컴퓨터 서버들, 또는 개인용 컴퓨터들은 본 명세서의 교시에 따른 프로그램과 함께 사용될 수 있거나 요구된 방법 단계들을 수행하기 위한 보다 특수화된 장치를 구성하는 것이 편리할 수 있다. 이러한 다양한 시스템을 위해 요구되는 구조는 이하의 설명에서 나타날 것이다. 다양한 프로그래밍 언어들이 본 명세서에서 기재된 바와 같이 발명의 교시를 구현하는데 사용될 수도 있는 것을 이해할 수 있을 것이다.

[0017] 게다가, 대표적인 예들에 대한 다양한 특징들 그리고 종속항들은 본 발명의 교시에 대한 유용한 추가적인 실시 예들을 제공하기 위해 명시적이지 않은 그리고 열거되지 않은 방식으로 결합될 수 있다. 또한 모든 값의 범위 또는 독립체들의 그룹들의 암시들은 모든 가능한 중간 값 또는 당해 발명을 제한하는 목적뿐만 아니라 본래의 개시 목적을 위한 중간 독립체들을 개시하는 것이 주목된다. 또한, 명시적 기준 및 도면에 도시된 구성 요소들

의 형상은 본 명세서에서 실시되는 방식을 이해할 수 있도록 설계되지만, 치수 및 실시 예에 나타난 형상에 한정되지 않는 것을 유의한다.

[0018] 도 1은 본 발명의 실시 예에 따른 모놀리식(monolithic) 스레딩 모델에서의 데이터 경로를 예시적으로 보여주는 도면이다. 데이터 경로는 iSCSI 리시버 경로 (iSCSI receiver path, 101), SCSI 커맨드 이슈 경로 (command issue path, 102), SCSI 커맨드 컴플리션 경로 (command completion path, 103), 및 iSCSI 트랜스미터 경로 (transmitter path, 104)의 모놀리식 시퀀스를 포함한다. 스토리지 시스템은 복수의 중앙 처리 장치(central processing unit; CPU) (예를 들어, N개의 CPU들) 들을 포함할 수 있으며, 스토리지 시스템의 각각의 CPU는 데이터 경로의 라이프사이클 동안 각각의 데이터 경로를 확립하고, 처리하고, 관리하고, 그리고 완료하는 것을 담당할 수 있다. 데이터 경로들(101 내지 104) 중에, 데이터 경로들(101, 104)은 이니시에이터와 SCSI 타깃 사이의 전송 프로토콜(즉, iSCSI)에서 확립될 수 있으며, 커넥션 단위에 기초할 수 있다. 예를 들어, 로그인/로그아웃, 버퍼 관리, 및 작업 관리 처리(task management handling)는 커넥션 단위 기반의 세션/커넥션 레벨에서 실행될 수 있다. I/O 커맨드들은 세션/커넥션 레벨에서 큐잉(queue)될 수 있으며 추적(track)될 수 있다. 데이터 경로들(101 내지 104) 중에, 데이터 경로들(102, 103)은 LUN 레벨에서 SCSI 타깃과 LUN들 사이에서 확립될 수 있다. 예를 들어, 타깃 스토리지 장치로의 그리고 타깃 스토리지 장치로부터의 SCSI 커맨드들, 및 에러 처리는 LUN 레벨에서 추적될 수 있다. 도 1에 도시된 모놀리식 스레드 모델에 따르면, 타깃 시스템의 이용 가능한 CPU 리소스들은 공유되어 전송 프로토콜 및 SCSI 코어 레벨 프로토콜 모두를 구동할 수 있다. 모놀리식 스레딩 모델은 캐시 지역성(cache locality)을 최대화하지 않는다.

[0019] 본 발명은 개별적인 스레드들에서 독립적으로 전송 프로토콜들과 SCSI 코어 프로토콜들을 나누고 처리하는 신규한 스레딩 모델(threading model)을 제공한다. 본 스레딩 모델은 전송 프로토콜 스레드와 SCSI 코어 스레드 사이에서 락 없는 큐 설계(lockless queue design)를 채용한다. 락 없는 큐 설계는, 전송 프로토콜들 및 SCSI 코어 프로토콜들에 대한 스레드들을 쪼개고 독립적으로 실행함으로써, I/O 성능을 향상시킬 수 있다. 스토리지 시스템의 CPU 리소스들은 전송 프로토콜 계층과 SCSI 코어 계층 사이에서 분배된다. 전송 프로토콜 스레드들은, 전송 프로토콜 계층에서의 스레드들 실행을 처리하기 위해 할당된, CPU들에서 오직 스케줄링 된다. SCSI 코어 스레드들은, SCSI 코어 계층에서의 스레드들 실행을 처리하기 위해 할당된, CPU 들에서 오직 스케줄링 된다. 전송 프로토콜 스레드들은 수신 경로 및 송신 경로를 처리할 수 있다. SCSI 코어 스레드들은 특정한 LUN에 대한 I/O 요청들 및 그 특정한 LUN에 대한 I/O 컴플리션(completion) 을 처리할 수 있다.

[0020] 본 발명의 시스템 및 방법은 복수의 LUN들 및 그들로의 커넥션들에 대해 고 확장성(scalability)을 제공할 수 있다. 본 발명의 시스템 및 방법은 캐시 지역성을 최대화하기 위해 전송 프로토콜 스레드들 및 LUN 스레드들을 더 분리할 수 있다. 게다가, 락 없는 큐 설계는, 복수의 LUN들이 복수의 커넥션들을 통하여 액세스 되었을 때, 락 충돌(lock contention)을 제거한다. 비록 본 발명이 iSCSI 타깃에 대하여 설명된다 하더라도, 본 발명은 시스템 리소스들을 효율적으로 이용하는 것과 동기화 병목을 피하기 위한 고성능의 타깃 I/O 성능을 제공하는 것의 이점을 이용할 수 있는 어떠한 SAN 프로토콜(예를 들어, FCP, pSCSI, iSCSI, HyperSCSI, Fibre Channel, ATA, SATA, AoE, InfiniBand, 및 NVMe over Fabrics)에 적용될 수 있음이 이해되어야 한다. 예를 들어, NVMe over Fabrics 프로토콜에서, SCSI 코어 계층에 상당하는 코어 계층은 NVMe 큐잉 인터페이스 (NVMe Queuing interface) 및 커맨드 세트들로 일컬어질 수 있다.

[0021] 도 2는 본 발명의 실시 예에 따른 예시적인 스레딩 모델에서의 데이터 경로를 보여주는 블록도이다. 데이터 경로(200)는 커넥션-특유의 경로(251)와 LUN-특유의 경로(252)로 분리될 수 있다. 커넥션-특유의 경로(251)는 리시버(201) 및 트랜스미터(211)를 포함할 수 있다. 이니시에이터(클라이언트)의 리시버(201)는 호스트로부터 명령어(instruction)를 수신하고, iSCSI 커맨드(202) 및 연관된 데이터(203)를 생성하고, 타깃과의 커넥션을 확립하고, 및 SCSI I/O 모듈(204)과 작업 관리자(task management; TM) I/O 모듈(205)을 통하여 타깃으로 iSCSI 커맨드(202) 및 데이터(203)를 전송할 수 있다.

[0022] LUN-특유의 경로(252)는 읽기, 쓰기, 및 트림(trim)과 같은 SCSI 커맨드들과 관련된 데이터를 저장하기 위한 데이터 커맨드 디스크립터 블록들(data command descriptor blocks; CDBs)(221), 그리고 문의(inquiry), 읽기, 및 용량과 같은 SCSI 커맨드들을 저장하기 위한 제어 커맨드 디스크립터 블록들(control CDBs)(222)을 포함할 수 있다. LUN-특유의 경로(252)는 SCSI 관리 커맨드들(예를 들어, 중단, LUN 리셋)을 저장하기 위한 작업 관리 I/O 블록(223), 그리고 상태/데이터 블록(224)을 더 포함할 수 있다. 호스트로부터 수신된 TM 커맨드들은 작업 관리 I/O 블록(223)에 저장될 수 있다. 타깃으로부터 수신된 타깃의 상태 및 관련 데이터는 상태/데이터 블록(224)에 저장될 수 있다. 제어 커맨드 디스크립터 블록들(222)은 특정한 제어 커맨드에 대한 상태 및 데이터를 업데이트하기 위해 상태/데이터 블록(224)을 직접 액세스할 수 있다. I/O 컴플리션과 같은 타깃에 대한 상태/데

이터 정보는 커넥션-특유의 경로(251)의 트랜스미터(211)로 되돌려 보내질 수 있다.

[0023] 도 3은 본 발명의 실시 예에 따른 예시적인 나뉘어진 스레딩 모델을 보여주는 도면이다. 타깃 시스템은 워크 로드(work load)의 유형에 기초하여 iSCSI 전송 프로토콜들과 SCSI 커맨드들 사이에서의 처리를 위해 분배된 복수의 CPU들을 포함할 수 있다. 커넥션-특유의 스레드(즉, 전송 프로토콜에서의 iSCSI)는 이니시에이터와 SCSI 타깃 사이의 iSCSI리시버 경로(301) 및 iSCSI 트랜스미터 경로(304)를 포함할 수 있다. 커넥션-특유의 스레드는 커넥션 단위에 기초할 수 있다. SCSI 커맨드 이슈 경로(302) 및 SCSI 커맨드 컴플리션 경로(303)는 LUN 레벨에서 확립된다. 타깃 시스템의 복수의 이용 가능한 CPU들 중에서, M개의 CPU들이 데이터 경로를 확립하고, 이니시에이터와 SCSI 타깃 사이의 SCSI 커맨드들을 전송하고, 그리고 전송 프로토콜 계층에서, 확립된 데이터 경로를 완료하기 위해 할당될 수 있다. 반면, N 개의 CPU들이 SCSI 타깃과 LUN들 사이의 SCSI 커맨드들을 처리하고, 관리하고, 다루기 위해 할당될 수 있다.

[0024] 이니시에이터와 SCSI 타깃 사이의 iSCSI 프로토콜은 커넥션마다 확립된 커넥션-특유의 스레드를 실행할 수 있으며, iSCSI 리시버 경로(301)와 iSCSI 트랜스미터 경로(304)의 시퀀스를 포함할 수 있다. 커넥션-특유의 스레드는 하나 또는 그 이상의 M 개의 CPU들에 할당된다. iSICSI 리시버 경로(301)는 SCSI 커맨드 이슈 경로(302)(즉, SCSI 리시버)를 포함하는 SCSI I/O 요청들을 큐잉(queue) 한다. SCSI 커맨드가 완료된 후, SCSI 타깃은 SCSI 커맨드 컴플리션 경로(303)(즉, SCSI 트랜스미터)를 큐잉 한다. LUN 단위에 기초하여 확립된 SCSI 커맨드 이슈 경로(302) 및 SCSI 커맨드 컴플리션 경로(303)는 N개의 CPU들에 할당된다. SCSI 커맨드들이 완료된 후, SCSI 타깃은 이니시에이터와 SCSI 타깃 사이의 이전에 확립된 전송 커넥션 상의 이니시에이터로 I/O 컴플리션 (예를 들어, iSCSI 트랜스미터 경로(304))를 큐잉 할 수 있다. 마지막으로, I/O 컴플리션은 이니시에이터와 SCSI 타깃 사이의 커넥션-특유의 스레드를 확립한 CPU(들)에 의해 처리될 수 있다.

[0025] 도 4는 본 발명의 실시 예에 따른 예시적인 스레딩 모델을 보여주는 도면이다. 본 스레딩 모델은 iSCSI 리시버 경로(401), iSCSI 트랜스미터 경로(404), SCSI 커맨드 이슈 경로(402), 및 SCSI 커맨드 컴플리션 경로(403)를 포함하는 데이터 경로를 제공한다. I/O 커맨드 이슈 경로에서, iSCSI 리시버 경로(401)는 I/O 커맨드 이슈 큐(410)를 사용하여 I/O 커맨드들을 SCSI 커맨드 이슈 경로(402)에 배치시킬 수 있다. I/O 커맨드 리턴 경로에서, SCSI 커맨드 컴플리션 경로(403)는 I/O 컴플리션 큐(411)를 사용하여 I/O 컴플리션들을 iSCSI 트랜스미터 경로(404)에 배치시킬 수 있다. 일 실시 예에 따르면, I/O 커맨드 이슈 큐(410) 및 I/O 컴플리션 큐(411)는, 복수의 커넥션들로부터의 I/O 커맨드들이 하나의 LUN에 다다를 수 있고 그리고 하나의 커넥션으로부터의 I/O 커맨드들이 복수의 LUN들에 다다를 수 있는, MPMC (multi-producer and multi-consumer) 락 없는 큐들일 수 있다.

[0026] 본 발명의 일 실시 예에 따르면, 큐들의 문맥에서, 생산자(producer)는 이니시에이터를 일컬을 수 있으며, 소비자(consumer)는 타깃을 일컬을 수 있다. 어떤 실시 예들에서는, 생산자는 타깃을 일컬을 수 있으며, 소비자는 이니시에이터를 일컬을 수 있다. 예를 들어, iSCSI 리시버 경로(401)와 iSCSI 트랜스미터 경로(404)는 생산자에 의해 소유될 수 있고, SCSI 커맨드 이슈 경로(402)와 SCSI 커맨드 컴플리션 경로(403)는 소비자에 의해 소유될 수 있다. 다른 실시 예에서, SCSI 커맨드 이슈 경로(402)와 SCSI 커맨드 컴플리션 경로(403)는 생산자에 의해 소유될 수 있고, iSCSI 리시버 경로(401)와 iSCSI 트랜스미터 경로(404)는 소비자에 의해 소유될 수 있다.

[0027] MPMC 큐에서, 생산자 작업들 및 소비자 작업들은 복수의 스레드들에서 실행될 수 있다. 예를 들어, 생산자 작업들은 n개의 스레드들에서 실행될 수 있으며, 소비자 작업들은 m개의 스레드들에서 실행될 수 있다. 특정한 생산자 스레드를 담당하는 복수의 생산자들이 있을 수 있으며, 특정한 소비자 스레드를 담당하는 복수의 소비자들이 있을 수 있다. I/O 이슈 경로에서, 전송 프로토콜 계층은 SCSI 코어 계층으로의 작업 요청들을 생성할 수 있다. 이 경우, 전송 프로토콜 계층은 생산자이고, SCSI 코어 계층은 소비자이다. 반면, I/O 컴플리션 경로에서, SCSI 코어 계층은 전송 프로토콜 계층으로의 작업 요청들을 생성할 수 있다. 이 경우, SCSI 코어 계층은 생산자이고, 전송 프로토콜 계층은 소비자이다. 커넥션 단위인 전송 프로토콜 계층에 의해 생성된 작업 요청들은 복수의 LUN들로 전달될 수 있다. 유사하게, LUN 단위인 SCSI 코어 계층에 의해 생성된 작업 요청들은 복수의 커넥션들로 전달될 수 있다. 전송 프로토콜 계층과 SCSI 코어 계층 사이의 통신은 일반적으로 락(lock)을 요구하는 동기점들을 포함할 수 있다. 본 발명의 실시 예에 따른 시스템은 전송 프로토콜 계층 및 SCSI 코어 계층이 락 없는 방식으로 액세스될 수 있도록 한다.

[0028] 도 5는 본 발명의 실시 예에 따른 SPSC (single-producer single-consumer) 락 없는 큐를 보여주는 블록도이다. 락 없는 큐의 각각의 노드는 데이터 컨테이너(data container) 및 포인터, 그리고 링크드 리스트(linked list)에 연결된 일련의 노드들을 포함할 수 있다. 일 실시 예에 따르면, 도 4의 I/O 커맨드 이슈 큐(410) 및 I/O 컴플리션 큐(411) 각각은 도 5에 도시된 락 없는 큐를 포함할 수 있다. 락 없는 큐는 헤드 노드

(501) 및 테일 노드(503)를 포함할 수 있다. 만일 큐에 단지 하나의 노드만 있다면, 헤드 노드(501) 및 테일 노드(503)는 동일할 수 있다. 현재의 노드(502)는 리스트의 시작점을 의미할 수 있는데, 시작점부터 소비자는 리스트 순회(traversal)를 시작할 수 있으며 노드를 소비할 수 있다. 어떤 실시 예들에서, 현재의 노드(502)는 (생산자에 의해 소유되는 헤드 노드(501)에 반대되는) 소비자 시작 또는 소비자 헤드로 일컬어질 수 있다.

[0029] 도 5에 도시된 링크드 리스트의 각각의 노드는 데이터 컨테이너(511) 및 다음의 노드를 가리키는 포인터(512)를 포함할 수 있다. 생산자는 새로운 노드(예를 들어, 새로운 I/O 커맨드에 대응하는 노드)를 생성할 수 있으며, 새롭게 생성된 노드를 큐의 테일 노드(503)에 연결할 수 있으며, 테일 노드(503)의 포인터를 업데이트하여 새로운 노드를 가리킬 수 있다. 이러한 방법에서, 새로운 노드는 기존의 큐에 더해질 수 있다. 유사하게, 생산자는 소비된 노드들을 헤드 노드(501)로부터 현재의 노드(502)로 해방시킬 수 있다. 시스템 리소스들이 이용 가능할 때, 생산자는, 소비자에 의한 노드 소비 처리로부터 독립적으로, 소비된 노드들을 해방시킬 수 있다. 이러한 의미에서, 생산자에 의한, 소비된 노드를 해방하는 과정은 여유로운 삭제(lazy delete)라 일컬어질 수 있다. 헤드 노드(501)부터 현재의 노드(502) 앞의 노드까지의 노드들은 생산자에 의해 소유되는 반면, 현재의 노드(502)부터 테일 노드(503)까지의 노드들은 소비자에 의해 소유될 수 있다. 소비자는 현재의 노드(502)부터 테일 노드(503)까지 리스트를 순회할 수 있으며, 현재의 노드(502)에 아이템을 소비할 수 있으며, 뒤따르는 노드로 현재의 포인터를 업데이트할 수 있다. 만일 현재의 포인터가 테일 노드(503)를 가리킨다면, 소비자는 노드들을 소비하지 않는다.

[0030] 락 없는 큐는, 제어 정보를 보유하여 락 없는 리스트를 관리하는 제어 구조를 제공할 수 있다. 생산자는 첫 번째 그리고 마지막 포인터들을 소유할 수 있다. 생산자에 의해 새로운 노드가 락 없는 리스트에 더해질 때, 마지막 포인터는 업데이트될 수 있다. 생산자가 소비된 노드들을 삭제할 때, 첫 번째 포인터는 업데이트될 수 있다. 소비자가 현재의 노드(502)로부터 테일 노드(503)까지 리스트를 순회하므로, 현재의 포인터는 소비자에 의해 업데이트 될 수 있다. 락 없는 큐의 현재의 포인터의 제어 정보와 소유권은 락 없는 큐의 제어 구조를 사용하여 생산자와 소비자 사이에서 막힘없이(seamlessly) 교환되기 때문에, 본 발명의 락 없는 큐는 데이터 경로에서 동기점들 및 락들을 위한 필요성을 제거할 수 있고, 특히, 복수의 세션들이 동시에 복수의 LUN들을 액세스할 때, I/O 성능을 향상시킬 수 있다.

[0031] 도 6은 본 발명의 실시 예에 따른 예시적인 MPMC 락 없는 큐(multi-producer multi-consumer lockless queue)를 보여주는 블록도이다. MPMC 락 없는 큐는 N개의 헤드 노드들(헤드1 내지 헤드N)의 제어 어레이(601)를 포함할 수 있다. 여기서, N은 생산자들이 실행할 예정인 스레드들의 개수이다. 생산자들이 N개의 스레드들에서 실행할 예정인 반면, 소비자들은 M개의 스레드들에서 실행할 예정이다. 하나의 생산자가 복수의 스레드들을 실행할 수 있기 때문에, 생산자들의 개수와 스레드들의 개수(N)는 다를 수 있다. 유사하게, 하나의 소비자가 복수의 스레드들을 실행할 수 있고 복수의 생산자들에 의해 생산된 노드들을 소비할 수 있기 때문에, 소비자들의 개수와 스레드들의 개수(M)는 다를 수 있다. 제어 어레이(601)에서, 생산자들은 노드들을 생성할 수 있고, 소비자들은 노드들을 소비할 수 있다. N개의 스레드들 각각에 대응하는 제어 구조는 SPSC 락 없는 큐를 유지할 수 있다. 소비자들은 N개의 스레드들의 제어 어레이(601)를 유지할 수 있다.

[0032] 일 실시 예에 따르면, MPMC 락 없는 큐는 스레드 식별자(ID)를 이용하여 스레드로의 배타적인 액세스를 제공할 수 있다. 예를 들어, 주어진 스레드 상에서 실행하는 생산자는 대응하는 큐 안으로 새로운 노드를 생성할 필요가 있을 수 있다. 새로운 노드가 더해질 때, 생산자는, 자신의 스레드 식별자(ID)를 제어 어레이(601)에 인덱싱함으로써, 큐에 배타적으로 액세스할 수 있다. 각각의 생산자는 큐로의 배타적인 액세스를 획득하여 새로운 노드를 생성하기 때문에, 복수의 생산자들 사이에 경합이 없다. 주어진 스레드 상에서 실행하는 소비자는 복수의 생산자들에 속하는 복수의 노드들을 소비할 수 있다. 유사하게, 각각의 소비자는 제어 어레이(601)에 있는 큐들에 배타적으로 액세스 하기 때문에, 복수의 소비자들 사이에 경합이 없다.

[0033] 도 7은 본 발명의 실시 예에 따른 예시적인 스레딩 모델을 보여주는 블록도이다. 본 스레딩 모델은 복수의 iSCSI 커넥션 스레드들(701) 및 복수의 SCSI LUN 스레드들(751)을 포함한다. 설명의 목적을 위해, iSCSI 커넥션 스레드들(701)은 3개의 CPU들에서 실행하고, SCSI LUN 스레드들(751)은 2개의 CPU들에서 실행한다. iSCSI 커넥션 스레드들(701) 각각은, SCSI LUN 스레드들(751)을 실행하는 CPU들 각각에 대응하는 2개의 노드들(즉, 헤드 1a, 헤드 2a)을 포함한다. SCSI LUN 스레드들(751) 각각은, iSCSI 커넥션 스레드들(701)을 실행하는 CPU들 각각에 대응하는 3개의 노드들(즉, 헤드1b, 헤드 2b, 헤드 3b)을 포함한다. 비록 본 실시 예는 iSCSI 커넥션 스레드들(701) 각각을 각각 실행하기 위해 할당된 3개의 CPU들, 및 SCSI LUN 스레드들(751) 각각을 각각 실행하기 위해 할당된 2개의 CPU들을 포함하는 것으로 도시되었으나, 본 발명의 사상을 벗어나지 않는 범위 내에서, 본 스레딩 모델은 다양한 수의 iSCSI 커넥션 스레드들 및 SCSI LUN 스레드들에 적용될 수 있는 것임을 이해해야 한

다. 각각의 iSCSI 커넥션 스레드(CT1, CT2, CT3)는 각각의 CPU(CPU 1, CPU 2, 및 CPU 3)에 할당될 수 있다. 유사하게, 각각의 SCSI LUN 스레드(LT1, LT2)는 각각의 CPU(CPU4, CPU5)에 할당될 수 있다. 6개의 iSCSI 커넥션들(C1 내지 C6)은 iSCSI 스레드들(CT1, CT2, 및 CT3)에 의해 서비스를 받을 수 있다. 4개의 LUN들(L1 내지 L4)은 SCSI LUN 스레드들(LT1, LT2)에 의해 서비스를 받을 수 있다.

[0034] 생산자 커넥션 스레드들(예를 들어, iSCSI 커넥션 스레드(701))은 각각의 CPU-ID에 의해 인덱싱 되는 LUN-단위의 큐로 직접 (처리될 예정인) I/O 커맨드들을 생성할 수 있다. SCSI LUN 스레드들(751)은, 각각의 CPU-ID에 의해 인덱싱 됨으로써 iSCSI 커넥션 스레드들(701)에 의해 생성되는 I/O 커맨드들에 따라, 커넥션-단위의 큐에 직접 (완료된) I/O 커맨드들을 생성할 수 있다. 소비자 커넥션 스레드들(예를 들어, SCSI LUN 스레드들)은 해당 커넥션에 속하는 I/O 커맨드들을 소비할 수 있다. SCSI LUN 스레드들(751)은 해당 LUN에 속하는 I/O 커맨드들을 소비할 수 있다. iSCSI 커넥션 스레드들(701)(생산자)는 I/O 커맨드들을 개별적인 SCSI LUN 스레드들(751)(소비자)에 직접 이슈할 수 있으며, SCSI LUN 스레드들(751)(생산자)는 I/O 컴플리션들을 개별적인 iSCSI 커넥션 스레드들(701)(소비자)에 직접 이슈할 수 있다. 각각의 I/O 커맨드는 특정한 스레드를 실행하는 프로세서에 특유한 CPU-ID에 의해 확인될 수 있으며, 그 결과, 독립적으로 실행되는 스레드들에서 동기점들 또는 락들을 위한 필요성을 제거할 수 있다.

[0035] 도 8은 본 발명의 다른 실시 예에 따른 예시적인 MPMC 락 없는 큐를 보여주는 블록도이다. I/O 이슈 경로에서, 생산자는 전송 프로토콜 계층일 수 있으며, 소비자는 SCSI 코어 계층일 수 있다. 전송 프로토콜 스레드(예를 들어, 도 3의 iSCSI 리시버 경로(301))는 N개의 CPU들의 LUN 어레이(801)를 액세스할 수 있다. 여기서, N은 전송 프로토콜 스레드들을 위해 할당된 CPU들의 개수이다. 어레이 인덱스는 전송 프로토콜 스레드가 실행중인 CPU 번호(1, 2 내지 N)를 일컫는다. 어레이 인덱스들 각각은 락 없는 싱글 링크드 리스트(lockless single linked list)를 포함하는데, 락 없는 싱글 링크드 리스트에서 전송 프로토콜 스레드는 특정한 LUN에 이슈된 I/O 커맨드들을 위한 작업 요청을 생성한다. SCSI 코어 스레드(예를 들어, 도 3의 SCSI 커맨드 이슈 경로(302))는, 전송 프로토콜 스레드들에 의해 큐잉 되는 작업 엔트리들을 처리하는, 대응하는 SCSI 코어 CPU에서 실행할 수 있다. SCSI 코어 스레드는 대응하는 SCSI 코어 CPU에서 실행하는 스레드에 있는 현재의 포인터를 업데이트할 수 있다. SCSI 모듈들(LUN들)은 도 8에 도시된 것과 같은 데이터 구조를 가질 수 있고, iSCSI 모듈들(전송 프로토콜 계층)에 인터페이스를 제공하여 I/O 요청들을 생성하고 배치할 수 있다.

[0036] 도 9는 본 발명의 또 다른 실시 예에 따른 예시적인 MPMC 락 없는 큐를 보여주는 블록도이다. I/O 컴플리션 경로에서, 생산자는SCSI 코어 계층일 수 있으며, 소비자는 전송 프로토콜 계층일 수 있다. LUN 스레드(예를 들어, 도 3의 SCSI 커맨드 컴플리션 경로(303))는 M개의 CPU들의 커넥션-단위의 어레이(901)를 액세스할 수 있다. 여기서, M은 SCSI 코어 스레드들을 위해 할당된 CPU들의 개수이다. 어레이 인덱스는 SCSI 코어 스레드가 실행중인 CPU 번호(1 내지 M)를 일컫는다. 어레이 인덱스들 각각은 락 없는 싱글 링크드 리스트를 포함하는데, 락 없는 싱글 링크드 리스트에서 SCSI 코어 스레드는 특정한 커넥션에 대한 완료된 I/O 커맨드들을 위한 작업 요청을 생성한다. 대응하는 CPU에서 실행하는 전송 프로토콜 스레드(예를 들어, 도 3의 iSCSI 트랜스미터 경로(304))는 모든 SCSI 코어 스레드들에 의해 큐잉 되는 I/O 커맨드들을 처리할 수 있다. iSCSI 스레드는 대응하는 전송 프로토콜 CPU에서 실행하는 iSCSI 스레드에 있는 현재의 포인터를 업데이트할 수 있다. iSCSI 모듈들(iSCSI 커넥션들)은 도 9에 도시된 것과 같은 데이터 구조를 가질 수 있고, SCSI 모듈들(SCSI 코어 계층)에 인터페이스를 제공하여 I/O 컴플리션들을 생성하고 배치할 수 있다.

[0037] 본 발명의 시스템 및 방법은 전송 처리와 데이터 (또는 코어) 처리를 분리할 수 있다. 본 발명의 시스템 및 방법은 전송 처리와 데이터 처리와 함께 내재하는 동기화 쟁점들 해결하는 MPMC (multi-producer-multi-consumer) 락 없는 설계를 구현한다. 본 발명의 시스템 및 방법은 전송 처리와 데이터 처리 사이의 리소스(예를 들어, CPU) 공유를 제공하여, 전송 계층과 코어 계층 각각의 리소스들에 대한 서로 다른 그리고 다양한 요구들을 수용한다. I/O 쓰기(IOW) 및 CPU 리소스들의 비율은 시스템 사양에 따라 다양할 수 있다.

[0038] 본 발명의 실시 예에 따른 방법은: 복수의 CPU들을 포함하는 데이터 스토리지 시스템의 타깃에 저장된 데이터를 액세스하기 위한 커맨드들을 수신하는 단계; 상기 복수의 CPU들 중 제 1 그룹의 CPU들에 대한 복수의 전송 스레드들을 실행하는 단계로써, 상기 복수의 전송 스레드들 각각은 커맨드 리시버 경로 및 커맨드 트랜스미터 경로를 포함하는 것; 상기 복수의 CPU들 중 제 2 그룹의 CPU들에 대한 복수의 데이터 경로 스레드들을 실행하는 단계로써, 상기 복수의 데이터 경로 스레드들 각각은 커맨드 이슈 경로 및 커맨드 컴플리션 경로를 포함하는 것; 전송 스레드의 커맨드 리시버 경로를 사용하여 I/O 커맨드 이슈 큐에 I/O 커맨드를 배치하고, 데이터 경로 스레드의 커맨드 이슈 경로를 사용하여 상기 I/O 커맨드를 처리하는 단계; 그리고 상기 데이터 경로 스레드의 상기 커맨드 컴플리션 경로를 사용하여 I/O 컴플리션 큐에 I/O 컴플리션 통지를 배치하고, 상기 전송 스레드의 상기

커맨드 트랜스미터 경로를 사용하여 상기 I/O 컴플리션 통지를 처리하는 단계를 포함할 수 있다. 상기 I/O 커맨드 이슈 큐는 제 1 어레이의 큐들을 포함하고, 상기 제 1 어레이의 큐들 각각은 상기 제 1 그룹의 CPU들의 개별적인 CPU에 대응할 수 있다. 상기 I/O 컴플리션 큐는 제 2 어레이의 큐들을 포함하고, 상기 제 2 어레이의 큐들 각각은 상기 제 2 그룹의 CPU들의 개별적인 CPU에 대응할 수 있다.

[0039] 상기 전송 스레드는 리시버 및 트랜스미터를 포함하고, 상기 리시버는 I/O 커맨드들을 상기 데이터 경로에 전송할 수 있다.

[0040] 상기 데이터 경로 스레드는 상기 I/O 커맨드를 상기 타깃으로 전송하고 상기 타깃으로부터 상태 및 데이터 중 적어도 하나를 수신하고, 상기 전송 스레드의 상기 트랜스미터로 상기 상태 및 상기 데이터 중 상기 적어도 하나를 전송할 수 있다.

[0041] 상기 제 1 어레이의 큐들은 상기 데이터 경로 스레드들에 할당된 상기 제 2 그룹의 CPU들에 대응하는 제 1 복수의 노드들을 포함할 수 있다.

[0042] 상기 제 1 복수의 노드들은 헤더 노드, 테일 노드, 및 상기 제 1 어레이의 큐들의 큐를 가리키는 현재의 노드를 포함하되, 상기 현재의 노드부터 상기 테일 노드까지의 노드들은 소비자에 의해 소유되고, 상기 제 1 어레이의 큐들 중 잔존하는 노드들은 생산자에 의해 소유될 수 있다.

[0043] 상기 생산자는 상기 이니시에이터이고, 상기 소비자는 상기 타깃일 수 있다.

[0044] 상기 소비자는 상기 데이터 경로 스레드들 각각에 특유하는 스레드 식별자를 이용하여 상기 큐로의 배타적인 액세스를 획득할 수 있다.

[0045] 상기 제 2 어레이의 큐들은 상기 전송 스레드들에 할당된 상기 제 1 그룹의 CPU들에 대응하는 제 2 복수의 노드들을 포함할 수 있다.

[0046] 상기 제 2 복수의 노드들은 헤더 노드, 테일 노드, 및 상기 제 2 어레이의 큐들의 큐를 가리키는 현재의 노드를 포함하되, 상기 현재의 노드부터 상기 테일 노드까지의 노드들은 소비자에 의해 소유되고, 상기 제 2 어레이의 큐들 중 잔존하는 노드들은 생산자에 의해 소유될 수 있다.

[0047] 상기 생산자는 상기 타깃이고 상기 소비자는 상기 이니시에이터일 수 있다.

[0048] 상기 소비자는 상기 전송 스레드들 각각에 특유하는 스레드 식별자를 이용하여 상기 큐로의 배타적인 액세스를 획득할 수 있다.

[0049] 상기 I/O 커맨드 이슈 큐와 상기 I/O 컴플리션 큐 각각은 MPMC (multi-producer multi-consumer) 락 없는 큐일 수 있다. 하나 또는 그 이상의 전송 스레드들로부터의 작업 요청들은 특정한 데이터 경로 스레드에 도달하고, 개별적인 전송 스레드로부터의 작업 요청들은 하나 또는 그 이상의 데이터 경로 스레드들에 도달할 수 있다. 유사하게, 하나 또는 그 이상의 데이터 경로 스레드들로부터의 작업 요청들은 개별적인 전송 스레드에 도달하고, 개별적인 데이터 경로 스레드로부터의 작업 요청들은 하나 도는 그 이상의 전송 스레드들에 도달할 수 있다.

[0050] 본 발명의 다른 실시 예에 따른 데이터 스토리지 시스템은: 복수의 CPU들; 복수의 타깃들; 그리고 I/O 흐름들과 I/O 커맨드 이슈 큐를 처리하기 위한 소프트웨어 모듈들 세트, 및 I/O 컴플리션 큐를 저장하는 메모리를 포함할 수 있다. 상기 소프트웨어 모듈들 세트는: 복수의 CPU들을 포함하는 데이터 스토리지 시스템의 타깃에 저장된 데이터를 액세스하기 위한 커맨드들을 수신하고; 상기 복수의 CPU들의 제 1 그룹의 CPU들 중 제 1 CPU에 대한 복수의 전송 스레드들을 실행하되, 상기 복수의 전송 스레드들 각각은 커맨드 리시버 경로 및 커맨드 트랜스미터 경로를 포함하고; 그리고 상기 복수의 CPU들의 제 2 그룹의 CPU들 중 제 2 CPU에 대한 복수의 데이터 경로 스레드들을 실행하되, 상기 복수의 데이터 경로 스레드들 각각은 커맨드 이슈 경로 및 커맨드 컴플리션 경로를 포함할 수 있다. 전송 스레드의 상기 커맨드 리시버 경로는 상기 I/O 커맨드를 상기 I/O 커맨드 이슈 큐에 배치하고, 데이터 경로 스레드의 상기 커맨드 이슈 경로는 상기 I/O 커맨드를 처리할 수 있다. 상기 데이터 경로 스레드의 상기 커맨드 컴플리션 경로는 I/O 컴플리션 통지를 상기 I/O 컴플리션 큐에 배치하고, 상기 전송 스레드의 상기 커맨드 트랜스미터 경로는 상기 I/O 컴플리션 통지를 처리할 수 있다. 상기 I/O 커맨드 이슈 큐는 제 1 어레이의 큐들을 포함하되, 상기 제 1 어레이의 큐들 각각은 상기 제 1 그룹의 CPU들의 각각의 CPU에 대응할 수 있다. 상기 I/O 컴플리션 큐는 제 2 어레이의 큐들을 포함하되, 상기 제 2 어레이의 큐들 각각은 상기 제 2 그룹의 CPU들의 각각의 CPU에 대응할 수 있다.

[0051] 상기 전송 스레드는 리시버와 트랜스미터를 포함하고, 상기 리시버는 I/O 커맨드를 상기 데이터 경로에 전송할

수 있다.

[0052] 상기 데이터 경로 스레드는 상기 I/O 커맨드를 상기 타깃으로 전송하고 상기 타깃으로부터 상태 및 데이터 중 적어도 하나를 수신하고, 상기 전송 스레드의 상기 트랜스미터로 상기 상태 및 상기 데이터 중 상기 적어도 하나를 전송할 수 있다.

[0053] 상기 제 1 어레이의 큐들은 상기 데이터 경로 스레드들에 할당된 상기 제 2 그룹의 CPU들에 대응하는 제 1 복수의 노드들을 포함할 수 있다.

[0054] 상기 제 1 복수의 노드들은 헤더 노드, 테일 노드, 및 상기 제 1 어레이의 큐들의 큐를 가리키는 현재의 노드를 포함하되, 상기 현재의 노드부터 상기 테일 노드까지의 노드들은 소비자에 의해 소유되고, 상기 제 1 어레이의 큐들 중 잔존하는 노드들은 생산자에 의해 소유될 수 있다.

[0055] 상기 생산자는 상기 이니시에이터이고, 상기 소비자는 상기 타깃일 수 있다.

[0056] 상기 소비자는 상기 데이터 경로 스레드들 각각에 특유하는 스레드 식별자를 이용하여 상기 큐로의 배타적인 액세스를 획득할 수 있다.

[0057] 상기 제 2 어레이의 큐들은 상기 전송 스레드들에 할당된 상기 제 1 그룹의 CPU들에 대응하는 제 2 복수의 노드들을 포함할 수 있다.

[0058] 상기 제 2 복수의 노드들은 헤더 노드, 테일 노드, 및 상기 제 2 어레이의 큐들의 큐를 가리키는 현재의 노드를 포함하되, 상기 현재의 노드부터 상기 테일 노드까지의 노드들은 소비자에 의해 소유되고, 상기 제 2 어레이의 큐들 중 잔존하는 노드들은 생산자에 의해 소유될 수 있다.

[0059] 상기 생산자는 상기 타깃이고 상기 소비자는 상기 이니시에이터일 수 있다.

[0060] 상기 소비자는 상기 전송 스레드들 각각에 특유하는 스레드 식별자를 이용하여 상기 큐로의 배타적인 액세스를 획득할 수 있다.

[0061] 상기 I/O 커맨드 이슈 큐와 상기 I/O 컴플리션 큐 각각은 MPMC (multi-producer multi-consumer) 락 없는 큐일 수 있다. 하나 또는 그 이상의 전송 스레드들로부터의 작업 요청들은 특정한 데이터 경로 스레드에 도달하고, 개별적인 전송 스레드로부터의 작업 요청들은 하나 또는 그 이상의 데이터 경로 스레드들에 도달할 수 있다. 유사하게, 하나 또는 그 이상의 데이터 경로 스레드들로부터의 작업 요청들은 개별적인 전송 스레드에 도달하고, 개별적인 데이터 경로 스레드로부터의 작업 요청들은 하나 도는 그 이상의 전송 스레드들에 도달할 수 있다.

[0062] 이상 설명된 예시적인 실시 예들이 고성능의 락 없는 스케일러블 타깃을 제공하기 위한 시스템 및 방법의 구현의 다양한 실시 예들을 보여주기 위해 설명되었다. 본 발명이 속하는 기술 분야에서 통상의 지식을 지닌 자들은 개시된 예시적인 실시 예들을 다양하게 변형하거나 변경할 수 있을 것이다. 본 개시의 범위 내에 있는 것으로 의도된 주제는 후술될 청구항들에서 설명될 것이다.

부호의 설명

[0063] 101: iSCSI 리시버 102: SCSI 커맨드 이슈 경로

103: SCSI 커맨드 컴플리션 경로 104: iSCSI 트랜스미터

201: 리시버 202: 커맨드

203: 데이터 204: SCSI I/O

205: TM I/O 211: 트랜스미터

221: 데이터 SDB들 222: 제어 SDB들

223: TM I/O들 224: 상태/데이터

251: 커넥션-특유의 경로 252: LUN-특유의 경로

301: iSCSI 리시버 302: SCSI 커맨드 이슈 경로

303: SCSI 커맨드 컴플리션 경로 304: iSCSI 트랜스미터

401: iSCSI 리시버 402: SCSI 커맨드 이슈 경로

403: SCSI 커맨드 컴플리션 경로 404: iSCSI 트랜스미터

410; I/O 이슈 큐 411; I/O 컴플리션 큐

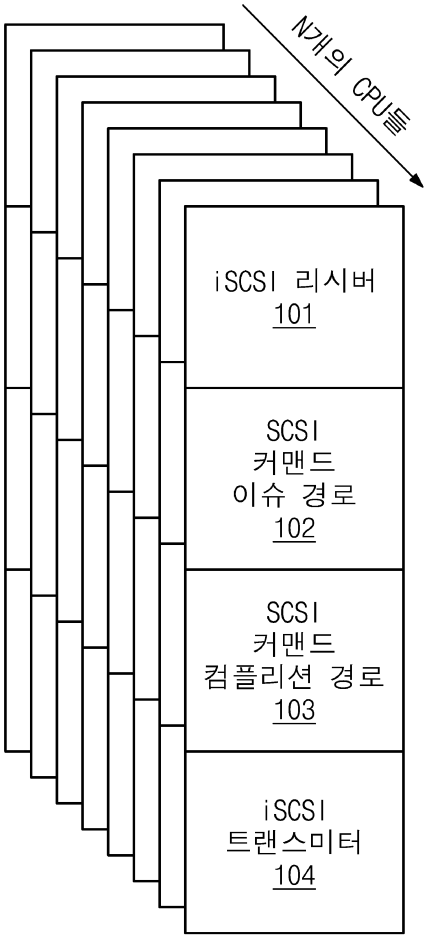
501: 헤드 502: 현재

503: 테일 701: iSCSI 커넥션 스레드들

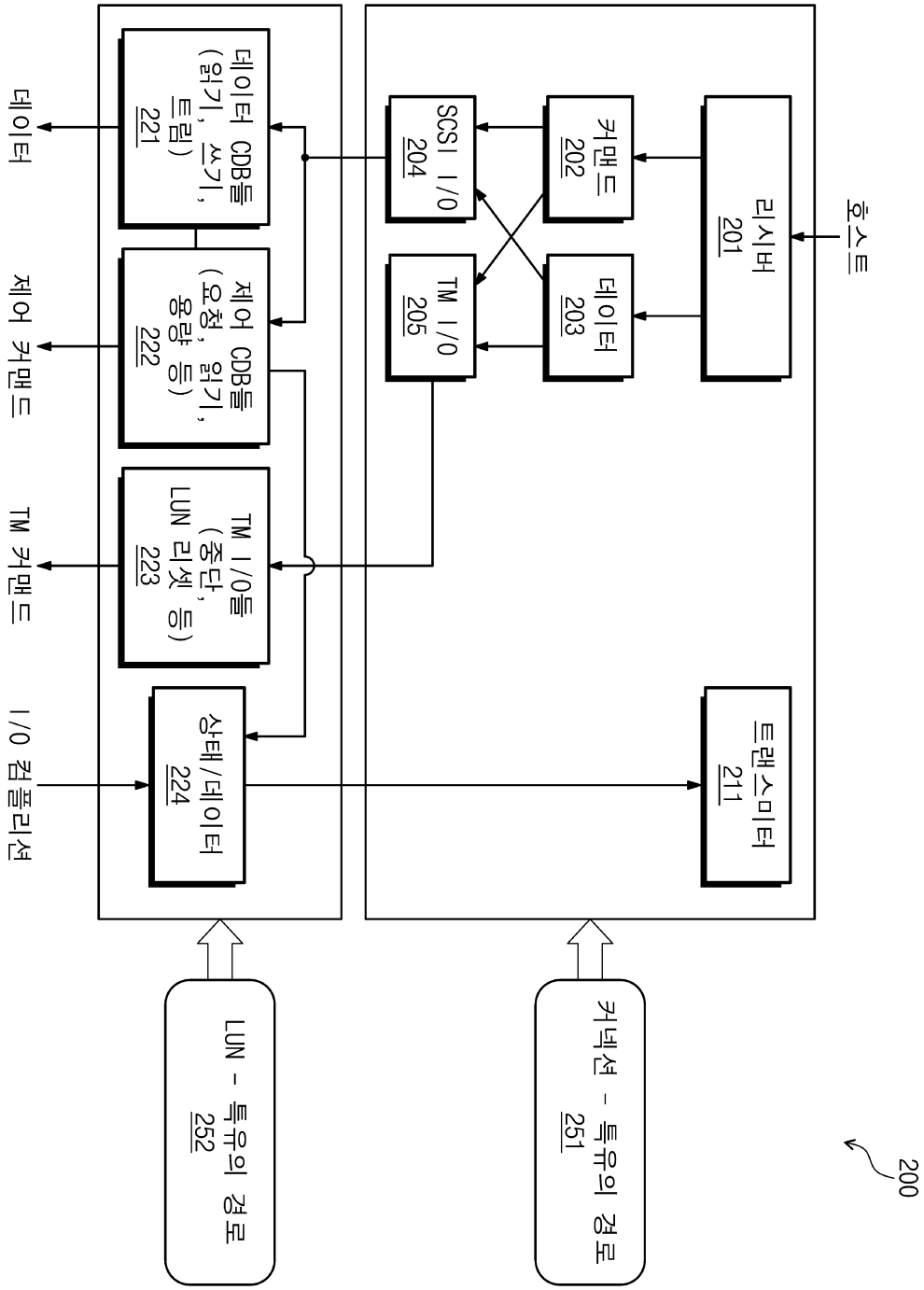
751: SCSI LUN 스레드들

도면

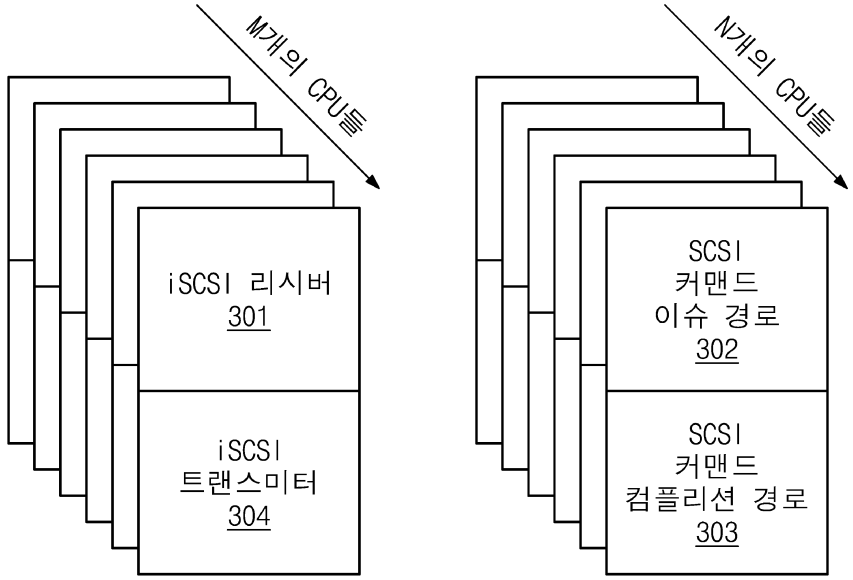
도면1



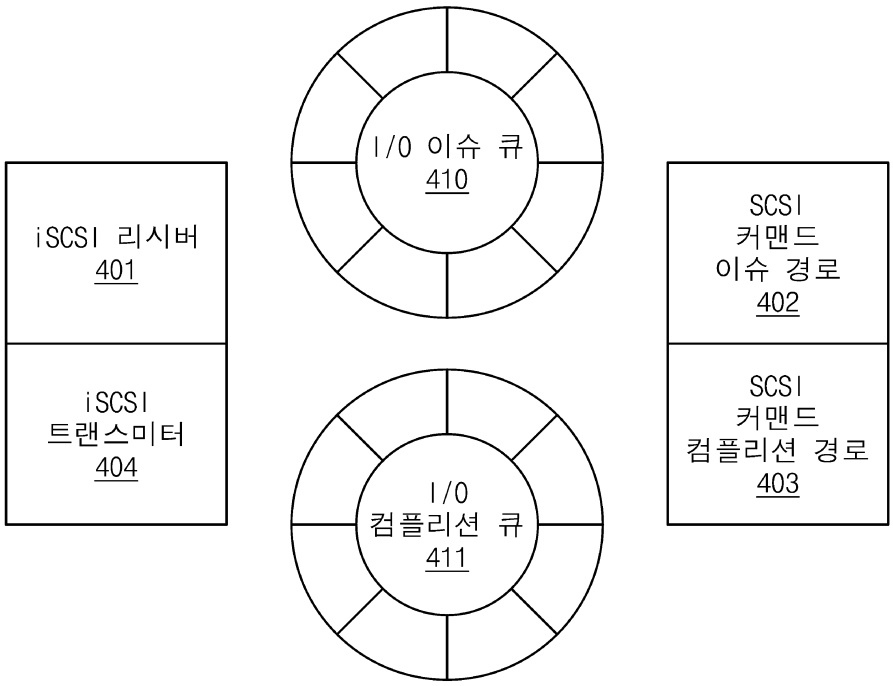
도면2


호스트
리시버 201
커맨드 202
데이터 203
SCSI I/O 204
TM I/O 205
트랜스미터 211
데이터 CDB들 (읽기, 쓰기, 트림) 221
제어 CDB들 (요청, 읽기, 용량 등) 222
TM I/O들 (중단, LUN 리셋 등) 223
상태/데이터 224
데이터
제어 커맨드
TM 커맨드
I/O 컴플리션
커넥션 - 특유의 경로 251
LUN - 특유의 경로 252
200

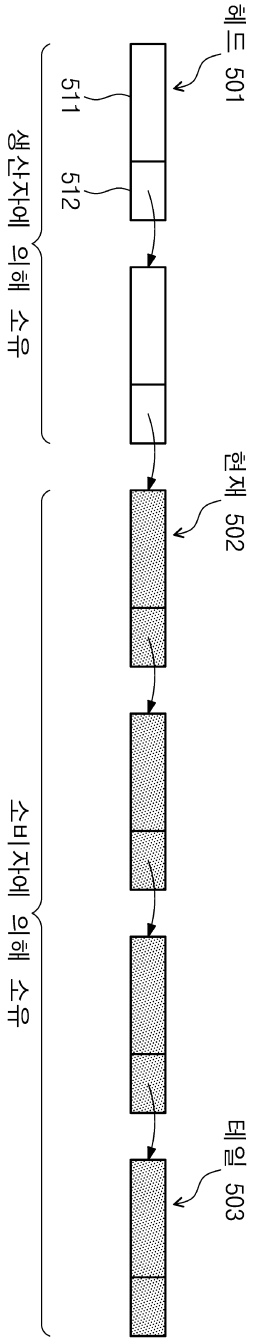
도면3


M개의 CPU들
N개의 CPU들
iSCSI 리시버
301
iSCSI
트랜스미터
304
SCSI
커맨드
이슈 경로
302
SCSI
커맨드
컴플리션 경로
303

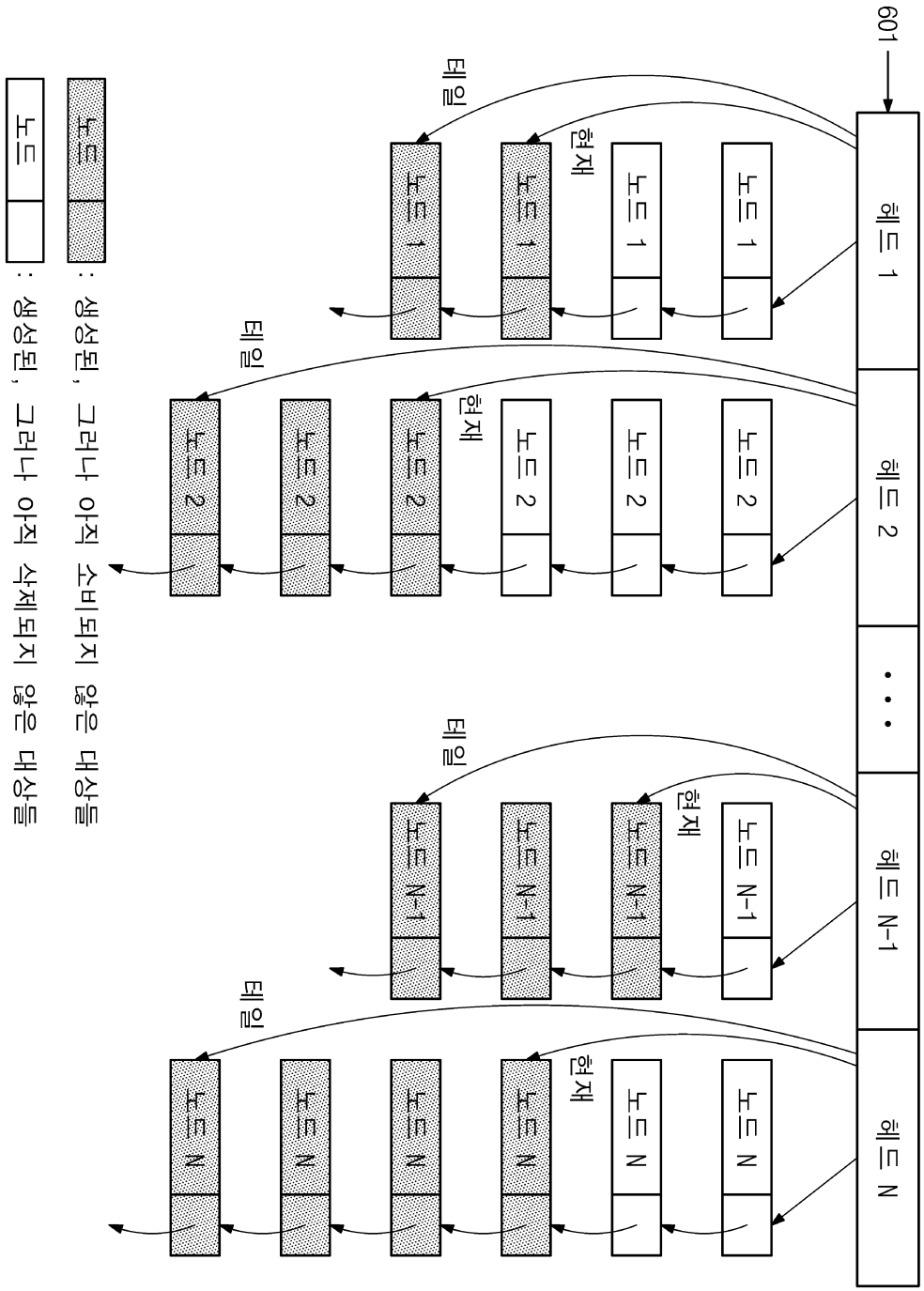
도면4


I/O 이슈 큐
410
iSCSI 리시버
401
iSCSI
트랜스미터
404
I/O
컴플리션 큐
411
SCSI
커맨드
이슈 경로
402
SCSI
커맨드
컴플리션 경로
403

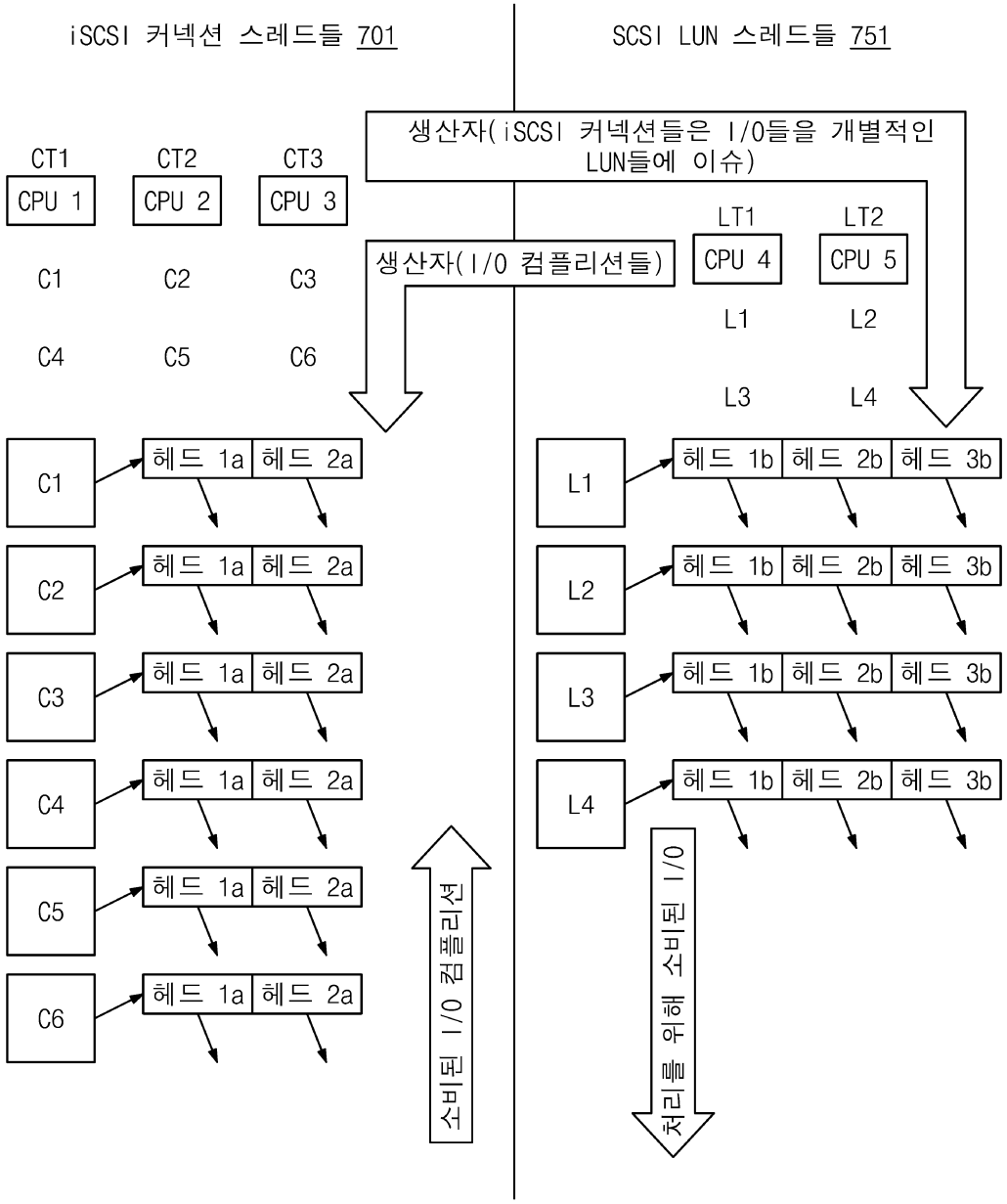
도면5



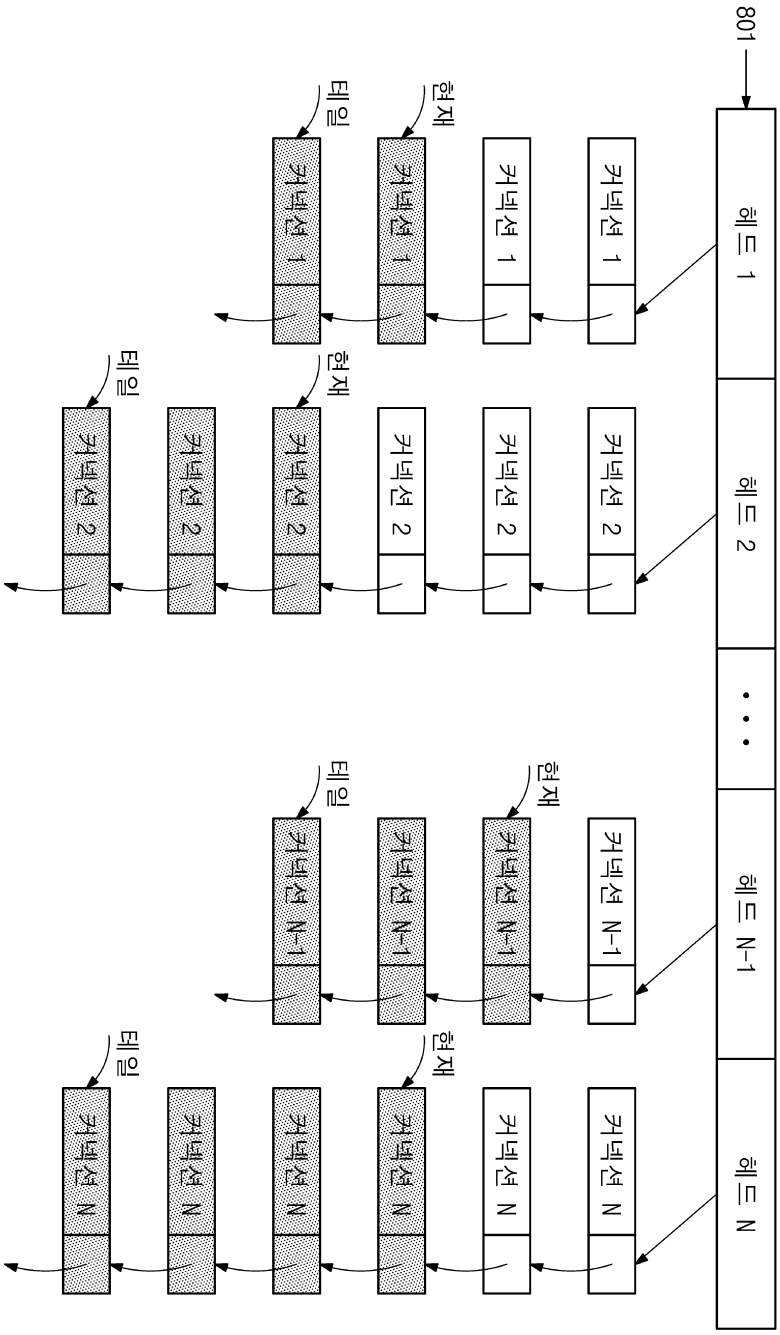
도면6


601
헤드 1
헤드 2
• • •
헤드 N-1
헤드 N
노드 1
노드 2
노드 N-1
노드 N
현재
테일
노드 1
노드 2
노드 N-1
노드 N
: 생성된, 그러나 아직 소비되지 않은 대상들
: 생성되, 그러나 아직 삭제되지 않은 대상들

도면7


iSCSI 커넥션 스레드들 701
SCSI LUN 스레드들 751
생산자(iSCSI 커넥션들은 I/O들을 개별적인 LUN들에 이슈)
CT1
CT2
CT3
CPU 1
CPU 2
CPU 3
C1
C2
C3
C4
C5
C6
생산자(I/O 컴플리션들)
LT1
LT2
CPU 4
CPU 5
L1
L2
L3
L4
헤드 1a
헤드 2a
헤드 1b
헤드 2b
헤드 3b
소비된 I/O 컴플리션
처리를 위해 소비된 I/O

도면8



도면9