(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2012년11월12일
(11) 등록번호 10-1200594
(24) 등록일자 2012년11월06일

(51) 국제특허분류(Int. Cl.)
G06F 15/16 (2006.01) *G06F 15/167* (2006.01)
G06F 12/02 (2006.01)
(21) 출원번호 10-2009-0112567
(22) 출원일자 2009년11월20일
심사청구일자 2012년04월24일
(65) 공개번호 10-2010-0069566
(43) 공개일자 2010년06월24일
(30) 우선권주장
12/316,713 2008년12월15일 미국(US)
(뒷면에 계속)
(56) 선행기술조사문헌
US20060235977 A1
US20070192554 A1
전체 청구항 수 : 총 7 항

(73) 특허권자
엘에스아이 코포레이션
미합중국 캘리포니아 95035, 바버 레인 밀피타스 1621
(72) 발명자
즈위슬러 로스 이
미국 콜로라도주 80026 라파예트 게이트웨이 서클 732
스프라이 앤드류 제이
미국 캔자스주 67205 위치타 노스 쇼어 서클 2642 엔
(뒷면에 계속)
(74) 대리인
제일특허법인, 김원준

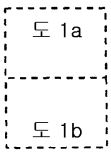
심사관 : 이상헌

(54) 발명의 명칭 **방법 및 스토리지 클러스터**

(57) 요 약

데이터 세트를 어드레싱하고, 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 또는 적어도 데이터 세트의 제 1 서브세트 및 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나를 어드레싱하기 위한 커맨드가 제 1 데이터 스토리지 시스템으로 발행된다. 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 또는 적어도 데이터 세트의 제 1 서브세트 및 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나가 액세스된다. 데이터 세트를 어드레싱하고, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답을 어드레싱하기 위한 커맨드가 제 2 데이터 스토리지 시스템으로 발행되며, 제 2 데이터 스토리지 시스템은 적어도 데이터 세트의 제 2 서브세트를 포함한다. 데이터 세트의 제 2 서브세트와, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답이 액세스된다.

대 표 도 - 도1



(72) 발명자

프레딘 제럴드 제이

미국 캔자스주 67205 위치타 노스 쉬포드 2401

깁슨 케네스 제이

미국 콜로라도주 80026 라파예트 레이크 메도우 드라이브 2534

(30) 우선권주장

12/316,778 2008년12월15일 미국(US)

12/553,558 2009년09월03일 미국(US)

61/215,304 2009년05월04일 미국(US)

특허청구의 범위

청구항 1

데이터 세트를 어드레싱하고, 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴(referral)을 포함하는 제 1 리퍼럴 응답 또는 적어도 상기 데이터 세트의 제 1 서브세트 및 적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답 중 적어도 하나를 어드레싱하기 위한 커맨드를 제 1 데이터 스토리지 시스템으로 발행하는 단계와,

적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답 또는 적어도 상기 데이터 세트의 상기 제 1 서브세트 및 적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답 중 적어도 하나에 액세스하는 단계와,

상기 데이터 세트를 어드레싱하고, 상기 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답을 어드레싱하기 위한 커맨드를 상기 제 2 데이터 스토리지 시스템으로 발행하는 단계 - 상기 제 2 데이터 스토리지 시스템은 적어도 상기 데이터 세트의 제 2 서브세트를 포함함 - 와,

상기 데이터 세트의 상기 제 2 서브세트에 액세스하고, 상기 제 1 데이터 스토리지 시스템과 상기 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 상기 제 2 리퍼럴 응답에 액세스하는 단계를 포함하되,

상기 제 1 데이터 스토리지 시스템에 대한 리퍼럴은,

상기 제 1 데이터 스토리지 시스템의 포트와 연관된 제 1 포트 식별자와,

상기 데이터 세트의 상기 제 1 서브세트의 제 1 데이터 오프셋과,

상기 제 1 데이터 스토리지 시스템에 대한 제 1 데이터 길이를 더 포함하고,

상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴은,

상기 제 2 데이터 스토리지 시스템의 포트와 연관된 제 2 포트 식별자와,

상기 데이터 세트의 상기 제 2 서브세트의 제 2 데이터 오프셋과,

상기 제 2 데이터 스토리지 시스템에 대한 제 2 데이터 길이를 더 포함하며,

상기 제 3 데이터 스토리지 시스템에 대한 리퍼럴은,

상기 제 3 데이터 스토리지 시스템의 포트와 연관된 제 3 포트 식별자와,

상기 데이터 세트의 제 3 서브세트의 제 3 데이터 오프셋과,

상기 제 3 데이터 스토리지 시스템에 대한 제 3 데이터 길이를 더 포함하고,

상기 제 1 데이터 스토리지 시스템에 대한 제 1 데이터 길이는, 상기 데이터 세트의 상기 제 1 서브세트에 대한 데이터 길이와, 상기 제 1 데이터 스토리지 시스템에 대한 모든 디센던트(descendant) 리퍼럴 응답 내의 모든 디센던트 데이터 스토리지 시스템 상의 상기 데이터 세트의 모든 서브세트에 대한 데이터 길이의 합이고,

상기 제 2 데이터 스토리지 시스템에 대한 제 2 데이터 길이는, 상기 데이터 세트의 상기 제 2 서브세트에 대한 데이터 길이와, 상기 제 2 데이터 스토리지 시스템에 대한 모든 디센던트 리퍼럴 응답 내의 모든 디센던트 데이터 스토리지 시스템 상의 상기 데이터 세트의 모든 서브세트에 대한 데이터 길이의 합이며,

상기 제 3 데이터 스토리지 시스템에 대한 제 3 데이터 길이는, 상기 데이터 세트의 상기 제 3 서브세트에 대한 데이터 길이 및 상기 제 3 데이터 스토리지 시스템에 대한 모든 디센던트 리퍼럴 응답 내의 모든 디센던트 데이터 스토리지 시스템 상의 상기 데이터 세트의 모든 서브세트에 대한 데이터 길이의 합인

방법.

청구항 2

제 1 항에 있어서,

상기 제 1 데이터 스토리지 시스템으로 커맨드를 발행하는 단계는 SCSI(Small Computer System Interface) 입력/출력(I/O) 요청 커맨드를 통해 상기 제 1 데이터 스토리지 시스템으로 커맨드를 발행하는 단계를 더 포함하고,

상기 제 2 데이터 스토리지 시스템으로 커맨드를 발행하는 단계는 SCSI I/O 요청 커맨드를 통해 상기 제 2 데이터 스토리지 시스템으로 커맨드를 발행하는 단계를 더 포함하되,

상기 제 1 데이터 스토리지 시스템은 SCSI 스토리지 시스템이고, 상기 제 2 데이터 스토리지 시스템은 SCSI 스토리지 시스템이며, 상기 제 3 데이터 스토리지 시스템은 SCSI 스토리지 시스템인

방법.

청구항 3

제 1 항에 있어서,

상기 제 1 데이터 스토리지 시스템에 대한 리퍼럴은 제 1 포트 식별자를 더 포함하되, 상기 제 1 포트 식별자는 상기 제 1 데이터 스토리지 시스템이 액세스할 수 있는 상기 제 1 데이터 스토리지 시스템의 포트를 식별하고,

상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴은 제 2 포트 식별자를 더 포함하되, 상기 제 2 포트 식별자는 상기 제 2 데이터 스토리지 시스템이 액세스할 수 있는 상기 제 2 데이터 스토리지 시스템의 포트를 식별하며,

상기 제 3 데이터 스토리지 시스템에 대한 리퍼럴은 제 3 포트 식별자를 더 포함하되, 상기 제 3 포트 식별자는 상기 제 3 데이터 스토리지 시스템이 액세스할 수 있는 상기 제 3 데이터 스토리지 시스템의 포트를 식별하는

방법.

청구항 4

데이터 세트를 어드레싱하고, 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴(referral)을 포함하는 제 1 리퍼럴 응답 또는 적어도 상기 데이터 세트의 제 1 서브세트 및 적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답 중 적어도 하나를 어드레싱하기 위한 커맨드를 제 1 데이터 스토리지 시스템으로 발행하는 단계와,

적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답 또는 적어도 상기 데이터 세트의 상기 제 1 서브세트 및 적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답 중 적어도 하나에 액세스하는 단계와,

상기 데이터 세트를 어드레싱하고, 상기 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답을 어드레싱하기 위한 커맨드를 상기 제 2 데이터 스토리지 시스템으로 발행하는 단계 - 상기 제 2 데이터 스토리지 시스템은 적어도 상기 데이터 세트의 제 2 서브세트를 포함함 - 와,

상기 데이터 세트의 상기 제 2 서브세트에 액세스하고, 상기 제 1 데이터 스토리지 시스템과 상기 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 상기 제 2 리퍼럴 응답에 액세스하는 단계를 포함하되,

적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답 또는 적어도 상기 데이터 세트의 상기 제 1 서브세트 및 적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답 중 적어도 하나에 액세스하는 상기 단계는, 상기 제 1 데이터 스토리지 시스템에 대한 상태 및 적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 상기 제 1 리퍼럴 응답을 수신하는 단계를 더 포함하고,

상기 데이터 세트의 상기 제 2 서브세트에 액세스하는 단계는, 상기 제 2 데이터 스토리지 시스템에 대한 상태 및 상기 제 1 데이터 스토리지 시스템 또는 상기 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 상기 제 2 리퍼럴 응답을 수신하는 단계를 더 포함하며,

상기 데이터 세트의 상기 제 1 서브세트에 대한 액세스는 상기 제 1 데이터 스토리지 시스템에 대한 상태를 전달하는 상기 제 1 데이터 스토리지 시스템을 통해 완료되고,

상기 데이터 세트의 상기 제 2 서브세트에 대한 액세스는 상기 제 2 데이터 스토리지 시스템에 대한 상태를 전

달하는 상기 제 2 데이터 스토리지 시스템을 통해 완료되는

방법.

청구항 5

스토리지 클러스터로서,

제 1 리퍼럴 응답 생성 또는 데이터 세트의 제 1 서브세트 저장과 상기 제 1 리퍼럴 응답 생성 중 적어도 하나를 수행하는 제 1 데이터 스토리지 시스템과,

상기 데이터 세트의 제 2 서브세트 저장 및 제 2 리퍼럴 응답 생성을 수행하는 제 2 데이터 스토리지 시스템을 포함하되,

상기 제 1 리퍼럴 응답은 적어도 상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하고, 상기 제 2 리퍼럴 응답은 상기 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하고,

상기 제 1 데이터 스토리지 시스템에 대한 리퍼럴은,

상기 제 1 데이터 스토리지 시스템의 포트와 연관된 제 1 포트 식별자와,

상기 데이터 세트의 상기 제 1 서브세트의 제 1 데이터 오프셋과,

상기 제 1 데이터 스토리지 시스템에 대한 제 1 데이터 길이를 더 포함하고,

상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴은,

상기 제 2 데이터 스토리지 시스템의 포트와 연관된 제 2 포트 식별자와,

상기 데이터 세트의 상기 제 2 서브세트의 제 2 데이터 오프셋과,

상기 제 2 데이터 스토리지 시스템에 대한 제 2 데이터 길이를 더 포함하며,

상기 제 3 데이터 스토리지 시스템에 대한 리퍼럴은,

상기 제 3 데이터 스토리지 시스템의 포트와 연관된 제 3 포트 식별자와,

상기 데이터 세트의 제 3 서브세트의 제 3 데이터 오프셋과,

상기 제 3 데이터 스토리지 시스템에 대한 제 3 데이터 길이를 더 포함하고,

상기 제 1 데이터 스토리지 시스템에 대한 제 1 데이터 길이는, 상기 데이터 세트의 상기 제 1 서브세트에 대한 데이터 길이와, 상기 제 1 데이터 스토리지 시스템에 대한 모든 디센던트 리퍼럴 응답 내의 모든 디센던트 데이터 스토리지 시스템 상의 상기 데이터 세트의 모든 서브세트에 대한 데이터 길이의 합이고,

상기 제 2 데이터 스토리지 시스템에 대한 제 2 데이터 길이는, 상기 데이터 세트의 상기 제 2 서브세트에 대한 데이터 길이와, 상기 제 2 데이터 스토리지 시스템에 대한 모든 디센던트 리퍼럴 응답 내의 모든 디센던트 데이터 스토리지 시스템 상의 상기 데이터 세트의 모든 서브세트에 대한 데이터 길이의 합이며,

상기 제 3 데이터 스토리지 시스템에 대한 제 3 데이터 길이는, 상기 데이터 세트의 상기 제 3 서브세트에 대한 데이터 길이와, 상기 제 3 데이터 스토리지 시스템에 대한 모든 디센던트 리퍼럴 응답 내의 모든 디센던트 데이터 스토리지 시스템 상의 상기 데이터 세트의 모든 서브세트에 대한 데이터 길이의 합인

스토리지 클러스터.

청구항 6

제 5 항에 있어서,

상기 제 1 데이터 스토리지 시스템은 SCSI 스토리지 시스템이고,

상기 제 2 데이터 스토리지 시스템은 SCSI 스토리지 시스템이며,

상기 제 3 데이터 스토리지 시스템은 SCSI 스토리지 시스템인

스토리지 클러스터.

청구항 7

제 5 항에 있어서,

상기 제 1 데이터 스토리지 시스템에 대한 리퍼럴은 제 1 포트 식별자를 더 포함하되, 상기 제 1 포트 식별자는 상기 제 1 데이터 스토리지 시스템이 액세스할 수 있는 상기 제 1 데이터 스토리지 시스템의 포트를 식별하고,

상기 제 2 데이터 스토리지 시스템에 대한 리퍼럴은 제 2 포트 식별자를 더 포함하되, 상기 제 2 포트 식별자는 상기 제 2 데이터 스토리지 시스템이 액세스할 수 있는 상기 제 2 데이터 스토리지 시스템의 포트를 식별하며,

상기 제 3 데이터 스토리지 시스템에 대한 리퍼럴은 제 3 포트 식별자를 더 포함하되, 상기 제 3 포트 식별자는 상기 제 3 데이터 스토리지 시스템이 액세스할 수 있는 상기 제 3 데이터 스토리지 시스템의 포트를 식별하는

스토리지 클러스터.

청구항 8

삭제

청구항 9

삭제

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

청구항 14

삭제

청구항 15

삭제

청구항 16

삭제

청구항 17

삭제

청구항 18

삭제

청구항 19

삭제

명 세 서

발명의 상세한 설명

기 술 분 야

[0001] 본 명세서는 일반적으로는 네트워크화된(networked) 스토리지 분야에 관한 것으로, 보다 구체적으로는 개시자 시스템(an initiator system)과 블록 스토리지 클러스터 내의 다수의 포트 사이에 임의의 개수의 SCSI 리퍼럴(Small Computer System Interface referral)을 제공하는 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 블록 스토리지 클러스터링(block storage clustering)과 네트워크화된 스토리지가, 복수의 스토리지 디바이스에 스패닝하는(spanning) 데이터에 액세스하기 위한 시스템/방법을 제공할 수 있다.

발명의 내용

과제 해결수단

[0003] 본 발명의 방법은, 데이터 세트를 어드레싱하고, 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 또는 적어도 데이터 세트의 제 1 서브세트 및 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나를 어드레싱하기 위한 커맨드를 제 1 데이터 스토리지 시스템으로 발행하는 단계와, 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 또는 적어도 데이터 세트의 제 1 서브세트 및 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나에 액세스하는 단계와, 데이터 세트를 어드레싱하고, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답을 어드레싱하기 위한 커맨드를 제 2 데이터 스토리지 시스템으로 발행하는 단계로서, 제 2 데이터 스토리지 시스템은 적어도 데이터 세트의 제 2 서브세트를 포함하는 단계와, 데이터 세트의 제 2 서브세트와, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답에 액세스하는 단계를 포함하지만, 이것으로 제한되지는 않는다.

[0004] 본 발명의 스토리지 클러스터는, 제 1 리퍼럴 응답 생성 또는 데이터 세트의 제 1 서브세트 저장과, 제 1 리퍼럴 응답 생성 중 적어도 하나를 수행하는 제 1 데이터 스토리지 시스템과, 데이터 세트의 제 2 서브세트 저장 및 제 2 리퍼럴 응답 생성을 수행하는 제 2 데이터 스토리지 시스템을 포함하되, 제 1 리퍼럴 응답은 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하고, 제 2 리퍼럴 응답은 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하지만, 이것으로 제한되지는 않는다.

[0005] 본 발명의 시스템은, 데이터 세트를 어드레싱하고, 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 또는 적어도 데이터 세트의 제 1 서브세트 및 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나를 어드레싱하기 위한 커맨드를 제 1 데이터 스토리지 시스템으로 발행하는 수단과, 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 또는 적어도 데이터 세트의 제 1 서브세트 및 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나에 액세스하는 수단과, 데이터 세트를 어드레싱하고, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답을 어드레싱하기 위한 커맨드를 제 2 데이터 스토리지 시스템으로 발행하는 수단으로서, 제 2 데이터 스토리지 시스템은 적어도 데이터 세트의 제 2 서브세트를 포함하는 수단과, 데이터 세트의 제 2 서브세트와, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답에 액세스하는 수단을 포함하지만, 이것으로 제한되지는 않는다.

[0006] 전술된 일반적인 설명과 아래의 상세한 설명 모두 예시적이고 설명을 위한 것으로 본 발명을 제한하고자 함이 아님을 이해할 것이다. 본 명세서에 포함되어 본 명세서의 일부를 구성하는 첨부된 도면은 본 발명의 청구 사항을 도시한다. 본 명세서의 설명과 도면은 함께 본 발명의 원리를 설명한다.

발명의 실시를 위한 구체적인 내용

[0007] 본 발명의 다수의 장점들이 첨부된 도면을 참조로 하여 당업자에게 보다 잘 이해될 수 있다.

[0008] 첨부된 도면에 도시된 바와 같은, 개시된 본 발명의 청구 대상에 대한 세부사항이 아래에서 기술될 것이다.

[0009] 도 1을 일반적으로 참조하면, 블록 스토리지 프로토콜을 통해 액세스 가능한 네트워크화된 스토리지 구현물/시스템이 도시되었다. 예를 들어, 블록 스토리지 프로토콜은 Fibre Channel, iSCSI, 또는 SAS(Serial Attached SCSI)와 같은 네트워크 가능 미디어 수송부 상에서 구현되는 SCSI(Small Computer System Interface) 프로토콜일 수 있다. 본 발명은 SCSI RDMA 프로토콜(SRP)과 같은 다른 블록 스토리지 프로토콜 내에 추가로 구현될 수 있다. 본 발명의 시스템/방법은 블록 스토리지 프로토콜을 확장하여, 블록 스토리지 클러스터를 형성하는 복수의 이산적인 공동(cooperating) 스토리지 시스템에 걸쳐 공통 논리 블록 어드레스 공간을 갖는 데이터의 분포를 허용한다.

[0010] 네트워크화된 스토리지 구현물/시스템/스토리지 하부구조(100)는 애플리케이션 시스템/서버(102)를 포함할 수 있다. 애플리케이션 시스템(102)은 하나 이상의 애플리케이션(104)을 실행할 수 있다. 애플리케이션 시스템(102) 상에서 실행되는 애플리케이션(104)은 네트워크(예컨대, 스토리지 영역 네트워크(SAN)(110))에 걸쳐/네트워크를 통해/네트워크를 사용하여 하나 이상의 스토리지 시스템(106-108) 상에 저장된 스토리지 리소스에 액세스할 수 있다. 애플리케이션은 스토리지 리소스/네트워크 스토리지 리소스에 액세스하기 위해 (일반화의 손실 없이) 서버의 운영 시스템(114)의 블록 스토리지 프로토콜 스택(예컨대, SCSI 프로토콜 스택)(112)을 사용할 수 있다. 서버(102)의 운영 시스템(114)은 일반성의 손실 없이 가상화된 환경에서 가상 장치 내에서 실행될 수 있거나 또는 서버 하드웨어 상에서 직접 실행될 수 있다.

[0011] 본 발명의 현재 실시예에서, 서버(102)의 SCSI 프로토콜 스택(112)은 로컬의(예컨대, 서버상의) 또는 원거리의(예컨대, 네트워크 상의) 스토리지 리소스를 애플리케이션(104)에게 블록 스토리지 디바이스/논리 유닛/SCSI 논리 유닛으로서 제공할 수 있다. 각 논리 유닛/SCSI 논리 유닛은 고유한 논리 블록 어드레스 공간을 구비할 수 있다. 원거리의 스토리지 리소스/원거리의 스토리지 디바이스(106-108)는, 블록 스토리지 프로토콜이 맵핑되는 네트워크 미디어 수송 프로토콜을 실행할 수 있는 서버(102) 및/또는 스토리지 시스템(106-108)의 하나 이상의 SAN 어답터(116)에 의해 액세스될 수 있다. 예를 들어, SCSI 프로토콜은 구현가능한 다양한 유형의 네트워크 수송을 통해 맵핑될 수 있다. 일반화의 손실 없이, SAN 어답터(116) 및 그들의 미디어 수송 프로토콜 층은 물리적 또는 가상의 네트워크 어답터일 수 있다.

[0012] 본 발명의 예시적인 실시예에서, 스토리지 영역 네트워크(110)는 포트 레벨 어드레싱을 허용하는 임의의 네트워크 미디어 및 수송 프로토콜로부터 구현될 수 있다(예컨대, Fibre Channel, EtherNet, InfiniBand 및 SAS). 미디어 수송 층 프로토콜은 엔드-포인트(end-point)로부터 네트워크 패브릭(fabric)(110)을 가로지르는 엔드-포인트 포트로의 프로토콜 패킷의 모든 라우팅을 조정할 수 있다. 일반성의 손실 없이, 네트워크(110)는 단일 패브릭 또는 복수의 리던던트(redundant) 패브릭으로서 구현될 수 있다. 바람직하게는, 애플리케이션 시스템(들)/서버(들)(102) 상의 네트워크 포트는 스토리지 시스템(들)(106-108) 상의 네트워크 포트에 물리적으로 도달할 수 있다.

[0013] 본 발명의 다른 실시예에서, 스토리지 시스템(들)(106-108)은 네트워크화된 부착형(attached) 스토리지 디바이스(들)일 수 있다. 예를 들어, 스토리지 시스템(106-108)은 자신들의 로컬 디바이스를 SAN(110) 상에서 볼 수 있도록 하는 범용 컴퓨터, 전용 스토리지 어레이 또는 네트워크화된 디스크 드라이브일 수 있다. 스토리지 시스템의 스토리지 리소스는 미디어 수송 프로토콜 층을 실행하는 SAN 포트를 통해 액세스될 수 있다. SCSI 층은 스토리지 포트로서 스토리지 네트워크와 통신하기 위해 상기 SAN 포트를 사용할 수 있다. 각 스토리지 시스템(106-108)은 자신의 물리적 스토리지 디바이스에게 데이터 보호 또는 블록 추상화(abstraction)를 적용할 수 있는 로컬 블록 가상화 층(118-120)을 포함할 수 있다. 예를 들어, RAID(Redundant Array of Inexpensive Disk)와 같은 데이터 보호는 전용 네트워크 스토리지 시스템 상에서 사용될 수 있다. 각각의 스토리지 시스템(106-108)은 자신이 네트워크(110)로 익스포트(export)시키는 실질적인 부착형 물리적 스토리지 디바이스(126)에 액세스할 수 있는 내부 블록 프로토콜 스택(122-124)을 더 포함할 수 있다.

[0014] 애플리케이션 서버/애플리케이션 시스템/서버(102)에서 이용가능한 스토리지의 양은, 보다 많은 스토리지 디바이스(126)를 개별적인 스토리지 시스템(106-108)에 추가하거나, 또는 추가의 스토리지 시스템(106-108)을 스토리지 영역 네트워크로 추가함으로써 확장될 수 있다. 추가적인 스토리지 디바이스(126)가 개별적인 스토리지 시

스템(106-108)에 추가되었을 때, 스토리지 시스템(106-108) 내의 로컬 블록 가상화 층(118-120)은 복수의 물리적 디스크(126)로부터 복수의 스토리지 시스템(106-108)에 걸쳐 보다 많은 가상 볼륨(virtual volume)(128, 130)을 생성하도록 클러스터 블록 가상화 층(132)에 의해 사용될 수 있다. 이것은 가상 볼륨(128, 130)의 단일 논리 블록 어드레스 공간을 보존할 수 있지만, 일부 포인트에서는, 개별적인 스토리지 시스템(들)(106-108) 상의 물리적 부착 포인트의 개수가 고갈될 수 있으며, 그에 따라 총 용량 확대에 대한 제한을 발생시킬 수 있다. 스토리지 시스템이 스토리지 영역 네트워크로 추가되었을 때, 애플리케이션에서 이용가능한 총 스토리지는 단일 스토리지 시스템의 물리적 제한을 넘어 증가할 수 있다. 그러나, 복수의 스토리지 시스템(106-108)에 의해 제공된 스토리지는 애플리케이션 서버(들)(102)에 의한 사용을 위해 공통 논리 블록 어드레스 공간으로 결합되어야 할 수도 있다.

[0015] 복수의 네트워크 부착형 스토리지 시스템(106-108) 상의 스토리지 리소스(126)로부터 단일 네임 공간/공통 논리 블록 어드레스 공간을 생성하기 위해 다수의 기술이 사용될 수 있다. 예를 들어, 이러한 기술은 클러스터링된 파일 시스템 또는 객체 스토리지 프로토콜과 같은 서로 다른 스토리지 프로토콜을 사용할 수 있다. 블록 스토리지 클러스터 집합은, 블록 클러스터 집합이 복수의 리던던트 SAN 패브릭(110)의 각각에 있는 클러스터 블록 가상화 디바이스에 의해 제공될 수 있도록, 스토리지 네트워크(110)로 추가될 수 있다. 클러스터 블록 가상화 디바이스는 네트워크 스토리지 시스템(들)과 애플리케이션 시스템(들) 사이에 위치할 수 있다. 클러스터 블록 가상화 디바이스는 네트워크 스토리지 시스템/스토리지 시스템에 의해 익스포트된 블록 스토리지 논리 유닛을 임포트(import)할 수 있으며, 가상 볼륨을 생성함으로써 추가적인 블록 가상화 층을 생성할 수 있다. 그 다음 클러스터 블록 가상화 디바이스는 가상 볼륨을 논리 유닛으로서 애플리케이션 시스템(들)으로 익스포트할 수 있다. 애플리케이션 시스템은 스토리지 시스템에 의해 익스포트된 논리 유닛을 찾거나 액세스하지 않으며, 오히려 가상 볼륨/클러스터 가상 볼륨을 찾는다. 클러스터 물리적 구조 발견, 가상화 맵핑 및 관리는 클러스터 가상 관리자에 의해 제공될 수 있다. 클러스터 가상 관리자는 SAN의 에지 또는 에지 상의 임의의 위치의 리던던트 디바이스의 개별적인 쌍에서 존재할 수 있다. 일반화의 손실 없이, 블록 스토리지 클러스터 집합 기능부는 클러스터 블록 가상화 디바이스/블록 클러스터 가상화 디바이스에 걸쳐 분포될 수 있다.

[0016] 이와 달리, 블록 스토리지 클러스터 집합/블록 클러스터 집합은 애플리케이션 시스템(들)(102)(애플리케이션 시스템 집합)으로 추가될 수 있다. 예를 들어, 블록 클러스터 집합은 애플리케이션 시스템의 블록 스토리지 프로토콜 스택으로 추가된 추가적인 추상화 층에 의해 제공될 수 있다. 다수의 옵션이 애플리케이션 시스템 상에 상기 추상화 층을 배치하기 위해 구현될 수 있다. 블록 가상화 층은 스토리지 시스템(들)에 의해 익스포트된 논리 유닛을 마스킹하거나 숨길 수 있고, 블록 스토리지 프로토콜 스택 내의 블록 가상화 층 위의 층들에게 가상 볼륨을 제공할 수 있다. 블록 스토리지 클러스터 집합이 스토리지 영역 네트워크(네트워크 집합)로 추가되었을 때와는 달리, 블록 스토리지 클러스터 집합을 애플리케이션 시스템에 추가할 때, 스토리지 시스템(들)에 의해 익스포트된 논리 유닛은 애플리케이션 시스템(들)/서버(들)로 액세스가능하다. 블록 가상화 층은 애플리케이션 시스템(들)/서버(들) 상에서 실행되는 애플리케이션으로부터 상기 논리 유닛으로의 액세스를 숨길 수 있다. 네트워크 집합과 같이, 블록 클러스터 집합이 애플리케이션 시스템(들)에 추가되었을 때, 클러스터 가상화 관리자 기능부가 클러스터 내의 스토리지 리소스를 발견하고 애플리케이션 서버(들)에 걸쳐 가상화 맵핑을 분배하기 위해 제공될 수 있다. 이러한 관리 접근법의 하나의 변경은, 가상 볼륨이 애플리케이션 서버에 걸쳐 공유되는 방지할 수 있도록 각 서버 내의 개별적인 클러스터 가상화 구성을 갖는 것을 포함할 수 있다. 이와 달리, 가상 볼륨의 공유를 제공하기 위해, 클러스터-와이드 가상화 관리자가 요구될 수도 있다.

[0017] 본 발명의 예시적인 실시예에서(도 1에 도시된 바와 같음), 블록 스토리지 클러스터 집합은 스토리지 시스템(들)(106-108)(스토리지 시스템 집합)에 추가될 수 있다. 블록 클러스터 집합은 적어도 하나의 스토리지 시스템(106-108)의 블록 프로토콜 스택(122-124)으로 추가되는 클러스터 블록 가상화 층(들)(132)에 의해 제공될 수 있다. 클러스터 블록 가상화 층(132)은 로컬 및 원거리의 스토리지 시스템 상의 스토리지 디바이스(126)들을 가상 볼륨(128, 130)으로 결합할 수 있다. 클러스터 내의 각 스토리지 시스템(106-108) 상의 스토리지 디바이스(126)는, 클러스터 블록 가상화 층(132)에 의한 가상 볼륨(128, 130)의 생성을 허용하도록 하나 이상의 다른 스토리지 시스템에 의해 검출가능하거나 또는 보여질 수 있다(예컨대, 스토리지 시스템(106)의 스토리지 디바이스는 스토리지 시스템(107, 108)에게 보일 수 있고, 스토리지 시스템(107)의 스토리지 디바이스는 스토리지 시스템(들)(106, 108)에 의해 보일 수 있으며, 스토리지 시스템(108)의 스토리지 디바이스는 스토리지 시스템(106, 107)에게 보여질 수 있음). 다수의 스토리지 시스템 집합 구현에서, 오직 가상 볼륨(128, 130)만이 클러스터 블록 가상화 층(132)에 의해 스토리지 영역 네트워크(110) 상에서 애플리케이션 시스템(들)(102)으로 익스포트된다. 일부 네트워크화된 스토리지 구현에서, 스토리지 시스템(106-108 중 하나)에 도달하여 하나 이상의 서로 다른 스토리지 시스템 상의 데이터를 요청하는 입력/출력(I/O) 요청은, I/O 요청을 만족시키기 위해 올바

른 스토리지 시스템(들)으로 포워딩될 수 있다. 다수의 기술이 프록시 I/O 및 커맨드 포워딩과 같은 I/O 리디렉션(redirection)을 수행하도록 구현될 수 있다. 전술된 다른 블록 스토리지 클러스터 기술들에서와 같이, 스토리지 시스템 집합에서, 개별적인 클러스터 가상화 관리자 기능부(134)가 스토리지 하부구조 내의 스토리지 시스템(106-108) 중 적어도 하나에 존재할 것이 요구될 수 있다. 일반화의 손실 없이, 상기 클러스터 가상화 관리자 기능부(134)는 클러스터 내의 스토리지 시스템(106-108)에 걸쳐 분포될 수 있으며, 그에 따라 저 비용, 저 침략성(invasiveness) 스토리지 관리 기능 구현을 제공할 수 있다.

[0018] 블록 스토리지 볼륨은 복수의 스토리지 시스템(106-108)에 걸쳐 분포될 수 있다. 또한, 애플리케이션 시스템(들)(102)은 클러스터 내의 임의의 스토리지 시스템 상의 데이터에 액세스할 수 있다. 또한, 가상 볼륨(128, 130)은 모든 스토리지 노드/스토리지 시스템(106-108)에 걸친 공통 블록 어드레스 공간을 제공할 수 있다.

[0019] SCSI 리퍼럴(referral) 기술/방법이 도 1에 도시된 구현물/시스템(100)과 같이 네트워크화된 스토리지 구현물/시스템을 이용하기 위해 제공된다. 도 3을 일반적으로 참조하면, 본 발명의 예시적인 실시예에 따른 네트워크화된 스토리지 구현물을 통한 데이터 전송 방법(예컨대, 개시자 시스템/개시자와 블록 스토리지 클러스터 사이의 통신 방법)이 도시되었다. 예를 들어, 이 방법은 아래에 기술되는 바와 같이(그리고 도 2 및 3에 도시된 바와 같이) 스토리지 프로토콜 커맨드와 응답 시퀀스(예컨대, SCSI 커맨드/응답 원거리 절차 호출 모델)를 사용하여 블록 스토리지 클러스터링을 위한 기술을 구현할 수 있다. 본 발명의 현재 실시예에서, 방법(300)은 데이터 세트를 어드레싱하고, 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 또는 적어도 데이터 세트의 제 1 서브세트 및 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나를 어드레싱하기 위한 커맨드를 제 1 데이터 스토리지 시스템으로 발행하는 단계(302)를 포함한다. 예를 들어, 블록 스토리지 클러스터는 적어도 제 1 데이터 스토리지 시스템(106), 제 2 데이터 스토리지 시스템(108) 및 제 3 데이터 스토리지 시스템(107)을 포함할 수 있으며, 각각이 물리적 스토리지 디바이스(들)(126)와 통신 연결되거나 물리적 스토리지 디바이스(들)(126)를 포함한다. 제 1 데이터 스토리지 시스템(106)은 데이터 세트의 제 1 서브세트를 포함할 수 있다. 제 1 리퍼럴 응답은 데이터 세트의 다른 서브세트를 포함하는 다른 데이터 스토리지 시스템(예컨대, 제 2 데이터 스토리지 시스템(108))에 대한 적어도 하나의 리퍼럴을 포함할 수 있다. 커맨드는 스토리지 개시자/개시자 시스템/호스트/서버(102)에 의해서 영역 네트워크(110)를 통해 제 1 데이터 스토리지 시스템(106)(예컨대, 타겟 시스템/타겟)으로 전송될 수 있다. 예시적인 실시예에서, 커맨드는 데이터에 대한 요청(예컨대, 판독 요청)과 같은 I/O 요청일 수 있다. 다른 실시예에서, 타겟은 블록 스토리지 클러스터 내의 임의의 스토리지 시스템일 수 있고, 커맨드는 블록 스토리지 클러스터 내의 임의의 예상된 타겟 스토리지 상의 임의의 포트(예로서, 도 2에 도시된 바와 같은 SCSI Target 0의 포트)를 사용하여 전송될 수 있다. 커맨드는 SCSI 커맨드일 수 있고, 개시자/개시자 시스템(102)은 SCSI 개시자일 수 있으며, 타겟(예로서, 제 1 스토리지 시스템(106))은 SCSI 타겟일 수 있다.

[0020] 추가적인 실시예에서, 커맨드가 스토리지 영역 네트워크(110)/네트워크 수송부 상에서 전송되었을 때, 커맨드는 확립된 개시자 및 타겟의 조합부(예컨대, I_T Nexus) 상에서 전송될 수 있다. SCSI 프로토콜에서, 개시자와 타겟 사이의 I_T Nexus는 개시자 상의 SCSI 포트(예컨대, 서버/애플리케이션 시스템(102)의 SCSI 포트)와 타겟 상의 SCSI 포트(예컨대, 제 1 스토리지 시스템(106)의 SCSI 포트) 사이에서 확립될 수 있다. 복수의 스토리지 시스템을 갖는 블록 스토리지 클러스터는 클러스터 내의 모든 스토리지 시스템들 상의 각 포트에 대한 고유한 포트 식별자를 제공할 수 있다. 또한, 각 SCSI 커맨드는 볼륨의 논리 블록 어드레스 공간 내의 시작 어드레스 및 길이에 의해 전송될 데이터를 식별할 수 있다.

[0021] 예시적인 실시예에서, 방법(300)은 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 또는 적어도 데이터 세트의 제 1 서브세트 및 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나에 액세스하는 단계(304)를 더 포함할 수 있다. 제 1 리퍼럴 응답은 제 1 데이터 스토리지 시스템(106)으로부터 개시자 시스템(102)으로 전송될 수 있다. 본 발명의 현재 실시예에서, 커맨드를 수신하는 스토리지 시스템(예컨대, 타겟 스토리지 시스템) 내에 저장된/상에 존재하는 요청된 데이터의 임의의 서브세트가 개시자로 또는 개시자로부터 이동/전송될 수 있다. 예를 들어, 데이터는 전술된/동일한 I-T Nexus에 대한 일련의 SCSI 데이터 전송 단계들을 통해 타겟(106)과 개시자(102) 사이에서 이동될 수 있다(예컨대, 제 1 스토리지 시스템(106) 상에 저장된 데이터가 애플리케이션 시스템/개시자 시스템(102)으로 전송될 수 있다). 본 발명의 현재 실시예에서, 데이터는 특정한 SCSI 커맨드에 의해 요청된 바와 같이 개시자와 타겟 사이에서 한 방향 또는 양 방향으로 이동할 수 있다. 데이터 세트의 제 1 서브세트에 액세스하는 것은 제 1 서브세트의 전송 완료를 표시하기 위해 개시자 시스템(102)에 의해 제 1 데이터 스토리지 시스템(106)의 상태를 수신

하는 것을 포함할 수 있다. 상태는 표준 SCSI 상태 표시자일 수 있다.

[0022] 리퍼럴 응답은 데이터 요청 내에서 요청된 데이터의 서브세트가, 블록 스토리지 클러스터의 복수의 스토리지 시스템들 내에 포함된 제 1 데이터 스토리지 시스템에 의해 저장되지 않고/제 1 데이터 스토리지 시스템 상에 존재하지 않지만, 제 2 데이터 스토리지 시스템에 의해서는 저장되어/제 2 데이터 스토리지 시스템 상에 존재할 때, 제 1 데이터 스토리지 시스템으로부터 개시자 시스템으로 전송될 수 있다. 리퍼럴 응답은 개시자에게 오리지널 데이터 요청에서 요청된 데이터의 전부가 전송된 것은 아님을 나타내는 표시를 제공할 수 있고, 리퍼럴 응답은 개시자 시스템을 제 2 데이터 스토리지 시스템으로 디렉팅하는 리퍼럴을 제공할 수 있고/있거나, 리퍼럴 응답은 스토리지의 하나 이상의 다른 스토리지 시스템(예컨대, 제 2 스토리지 시스템(108))이 데이터의 일부/데이터의 남아있는 부분을 저장한다는 것을 표시하거나/이를 개시자 시스템에게 알리는 표시자를 제공할 수 있다. 리퍼럴 응답은 제 1 데이터 스토리지 시스템 상에 포함된 데이터에 기초하여 제 1 데이터 스토리지 시스템에 의해 생성될 수 있다. 리퍼럴 응답은 요청된 데이터의 나머지(예컨대, 단계(302)에서 수신된 오리지널 데이터 요청 내에서 요청된 데이터의 나머지)가 위치/저장된 클러스터의 하나 이상의 다른 스토리지 시스템/클러스터 노드(예컨대, 제 2 스토리지 시스템(108)과 같은 것)에 대한 리퍼럴의 리스트를 포함할 수 있다.

[0023] 전술된 바와 같이, 데이터가 오리지널 데이터 요청을 만족시키기 위해 개시자에 의해 획득되어야만 하는 각각의 추가적인 클러스터 노드/스토리지 시스템에 대한 리퍼럴이 존재한다. 본 발명의 현재 실시예에서, 각각의 리퍼럴 리스트의 리퍼럴은 개시자를 리퍼링하는(referring) 각각의 스토리지 시스템/노드에 대해 아래와 같은 정보(도 4에 도시된 바와 같음)를 포함할 수 있다: (예컨대, 오리지널 데이터 요청 내에서 요청된 데이터의 나머지의 적어도 일부를 포함하는 클러스터 노드/스토리지 시스템 상의 포트와 관련된) 포트 식별자; 오프셋(예컨대, 자신의 관련된 데이터 스토리지 시스템/스토리지 노드 상의 데이터의 제 1 바이트의 논리 블록 어드레스); 및 데이터 길이(예컨대, 리퍼럴을 위해 전달될 데이터량). 예를 들어, 데이터 길이는 자신의 관련된 데이터 스토리지 시스템/스토리지 노드 상에서 전송될 서브세트 내의 데이터량일 수 있다. 데이터 길이는 자신의 관련된 데이터 스토리지 시스템/스토리지 노드의 모든 디센던트(descendant) 데이터 스토리지 시스템/저장 노드 상에서 전송될 모든 서브세트 내의 총 데이터량으로 이루어질 수 있다. 제 2 데이터 스토리지 시스템/스토리지 노드는 만약 제 2 데이터 스토리지 시스템/스토리지 노드가 제 1 데이터 스토리지 시스템/스토리지 노드의 리퍼럴 리스트 내에 있다면 제 1 데이터 스토리지 시스템/스토리지 노드의 디센던트일 수 있다. 또한, 제 2 데이터 스토리지 시스템/스토리지 노드의 디센던트는 제 1 데이터 스토리지 시스템/스토리지 노드의 디센던트일 수 있다. 리퍼럴 응답의 부재는 데이터 스토리지 시스템/스토리지 노드가 디센던트 데이터 스토리지 시스템/스토리지 노드를 구비하지 않음을 나타낼 수 있다. 볼륨, 논리 유닛 및 타겟과 같이 리퍼럴을 완료하는 데에 필요한 다른 정보는 SCSI 리퍼럴이 생성된 커맨드의 문맥으로부터 입수가능하다.

[0024] 본 발명의 현재 실시예에서, 방법(300)은 데이터 세트를 어드레싱하고, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답을 어드레싱하기 위해, 제 2 데이터 스토리지 시스템으로 커맨드를 발행하는 단계(306)를 더 포함할 수 있다. 제 2 데이터 스토리지 시스템(108)은 데이터의 세트의 제 2 서브세트를 포함할 수 있다. 제 2 리퍼럴 응답은 데이터의 세트의 추가 서브세트를 포함하는 다른 데이터 스토리지 시스템(예컨대, 제 3 데이터 스토리지 시스템(107))으로의 적어도 하나의 리퍼럴을 포함할 수 있다. 커맨드는 스토리지 영역 네트워크(110)를 통해 개시자/개시자 시스템/호스트/서버(102)에 의해서 제 2 데이터 스토리지 시스템(108)(예컨대, 타겟 시스템/타겟)으로 전송될 수 있다. 예시적인 시스템에서, 커맨드는 데이터에 대한 요청(예컨대, 판독 요청)과 같은 I/O 요청일 수 있다. 다른 실시예에서, 타겟은 블록 스토리지 클러스터 내의 임의의 스토리지 시스템일 수 있으며, 커맨드는 블록 스토리지 클러스터 내의 임의의 예상 타겟 스토리지 시스템 상의 임의의 포트(예컨대, 도 2에 도시된 바와 같은 SCSI Target 1의 포트)를 사용하여 전송될 수 있다. 커맨드는 SCSI 커맨드일 수 있으며, 개시자/개시자 시스템(102)은 SCSI 개시자일 수 있고, 타겟(예컨대, 제 2 스토리지 시스템(108))은 SCSI 타겟일 수 있다.

[0025] 추가의 실시예에서, 스토리지 영역 네트워크(110)/네트워크 수송부 상에서 전송되었을 때, 커맨드는 확립된 개시자와 타겟 조합(예컨대, I-T Nexus) 상에서 전송될 수 있다. SCSI 프로토콜에서, 개시자와 타겟 사이의 I-T Nexus는 개시자 상의 SCSI 포트(예컨대, 서버/애플리케이션 시스템(102)의 SCSI 포트)과 타겟 상의 SCSI 포트(예컨대, 제 2 스토리지 시스템(108)의 SCSI 포트) 사이에서 확립될 수 있다. 복수의 스토리지 시스템을 갖는 블록 스토리지 클러스터는 클러스터 내의 모든 스토리지 시스템들 상의 각 포트에 대한 고유한 포트 식별자를 제공할 수 있다. 또한, 각 SCSI 커맨드는 볼륨의 논리 블록 어드레스 공간 내의 자신의 시작 어드레스 및 길이에 의해 전달될 데이터를 식별할 수 있다.

[0026] 예시적인 실시예에서, 방법(300)은 데이터의 세트의 제 2 서브세트와 제 2 리퍼럴 응답(308)에 액세스하는 단계

를 더 포함할 수 있다. 제 2 리퍼럴 응답은 제 2 데이터 스토리지 시스템(108)으로부터 개시자 시스템(102)으로 전달될 수 있다. 본 발명의 현재 실시예에서, 커맨드를 수신하는 스토리지 시스템(예컨대, 타겟 스토리지 시스템) 내에 저장된/스토리지 시스템 상에 존재하는 요청된 데이터의 임의의 서브세트가 개시자로 또는 개시자로부터 이동될 수 있다. 예를 들어, 데이터는 전술된/동일한 I-T Nexus 상에서의 일련의 SCSI 데이터 전달 단계를 통해 타겟(108)과 개시자(102) 사이에서 이동될 수 있다(예컨대, 제 2 데이터 스토리지 시스템(108) 상에 저장된 데이터는 애플리케이션 시스템/개시자 시스템(102)으로 전달될 수 있다). 본 발명의 현재 실시예에서, 데이터는 특정한 SCSI 커맨드에 의해 요청된 바와 같이 개시자와 타겟 사이의 한쪽 방향 또는 양방향으로 흐를 수 있다. 데이터의 세트의 제 2 서브세트에 액세스하는 것은 제 2 서브세트의 전달 완료를 나타내기 위해 개시자 시스템(102)에 의해 제 2 데이터 스토리지 시스템(108)의 상태를 수신하는 것을 포함할 수 있다. 상태는 표준 SCSI 상태 표시자일 수 있다.

[0027] 리퍼럴 응답은 데이터 요청 내에서 요청된 데이터의 서브세트가 제 2 데이터 스토리지 시스템에 의해 저장되어 제 2 데이터 스토리지 시스템 상에 존재하지는 않지만 블록 스토리지 클러스터의 복수의 스토리지 시스템들 내에 포함된 제 3 데이터 스토리지 시스템에 의해 저장되어 제 3 데이터 스토리지 시스템 상에 존재할 때, 제 2 데이터 스토리지 시스템으로부터 개시자 시스템으로 전달될 수 있다. 리퍼럴 응답은 개시자에게 오리지널 데이터 요청에서 요청된 데이터의 전부가 전송된 것은 아님을 나타내는 표시를 제공할 수 있고, 리퍼럴 응답은 개시자 시스템을 제 3 데이터 스토리지 시스템으로 디렉팅하는 리퍼럴을 제공할 수 있고/있거나, 리퍼럴 응답은 클러스터의 하나 이상의 다른 스토리지 시스템(예컨대, 제 3 스토리지 시스템(107))이 데이터의 일부/데이터의 남아있는 부분을 저장한다는 것을 표시하거나/이를 개시자 시스템에게 알리는 표시자를 제공할 수 있다. 리퍼럴 응답은 제 2 데이터 스토리지 시스템 상에 포함된 데이터에 기초하여 제 2 데이터 스토리지 시스템에 의해 생성될 수 있다. 리퍼럴 응답은 요청된 데이터의 나머지(예컨대, 단계(302)에서 수신된 오리지널 데이터 요청 내에서 요청된 데이터의 나머지)가 위치/저장된 클러스터의 하나 이상의 다른 스토리지 시스템/클러스터 노드(예컨대, 제 3 스토리지 시스템(107)과 같은 것)에 대한 리퍼럴의 리스트를 포함할 수 있다. 다른 실시예에서, 블록 스토리지 프로토콜 개시자(102)는 리퍼럴 리스트 내에 표시된 포트를 사용함으로써 오리지널 요청 내에서 요청된 데이터를 홀딩하는 클러스터 내의 모든 다른 스토리지 시스템으로 개별적인 커맨드를 전송할 수 있다. 상기 리퍼럴에 기초하여 발행된 커맨드에 응답하여 모든 데이터 전송이 완료된 후, 블록 스토리지 프로토콜은 자신의 호출자를 반환함으로써 동작을 완료할 수 있다.

[0028] 도 5-6을 참조하면, 본 발명에 의해 제공되는 데이터 세트 분포의 다양한 리퍼럴 리스트가 도시되었다. 예를 들어, 링크된 리퍼럴 리스트 분포(500)는 총 데이터 길이 160 블록인 4개의 리퍼럴의 최대 리퍼럴 리스트 길이를 가지고, 각 개별적인 데이터 스토리지/클러스터 노드는 10 블록의 데이터 세트의 서브세트를 포함한다. 링크된 리퍼럴 리스트 분포(500)에서, 개시자 시스템(도시되지 않음)은 총 데이터 길이 170 블록에 액세스하기 위해 제 1 데이터 스토리지 시스템(도시되지 않음)으로 커맨드를 발행하고, 10 블록 데이터 길이의 데이터 세트의 서브세트와 데이터 스토리지/클러스터 노드(502/508)에 대한 리퍼럴을 포함하는 리퍼럴 리스트를 수신한다. 리퍼럴(502-506)은 포트 식별자. 데이터 오프셋 및 그들의 리퍼럴들에 의해 참조되는 데이터 서브세트에 대한 데이터 길이를 포함한다. 리퍼럴(508)의 데이터 스토리지/클러스터 노드는 데이터 스토리지/클러스터 노드(510-516)에 대한 리퍼럴을 포함하는 리퍼럴 리스트(화살표(550)에 의해 도시됨)를 구비한다. 데이터 스토리지/리퍼럴(510-532)의 클러스터 노드는 모두 리퍼럴(508)의 데이터 스토리지/클러스터 노드의 디센던트이며, 따라서 리퍼럴(508)의 데이터 길이는 리퍼럴(510-532)의 총 데이터 길이와 리퍼럴(508)의 서브세트 데이터 길이의 합을 포함한다. 유사하게, 리퍼럴(516)의 데이터 스토리지/클러스터 노드는 데이터 스토리지/클러스터 노드(518-524)에 대한 리퍼럴을 포함하는 리퍼럴 리스트(화살표(551)에 의해 도시됨)를 구비한다. 리퍼럴(518-532)의 데이터 스토리지/클러스터 노드는 모두 리퍼럴(516)의 디센던트이다. 데이터 스토리지/클러스터 노드의 리퍼럴(502-506, 510-514, 518-522, 526-532)은 디센던트를 갖지 않는다.

[0029] 리퍼럴 리스트 트리 분포(600)는 총 160 블록의 데이터 길이인 네 개의 리퍼럴의 최대 리퍼럴 리스트 길이를 가지며, 각 개별적인 데이터 스토리지/클러스터 노드는 10 블록의 데이터 세트의 서브세트를 포함한다. 리퍼럴 리스트 트리 분포(600)에서, 개시자 시스템(도시되지 않음)은 총 데이터 길이 170 블록인 데이터 세트에 액세스하기 위한 커맨드를 제 1 데이터 스토리지 시스템(도시되지 않음)으로 발행하며, 데이터 길이가 10 블록인 데이터 세트의 서브세트와 데이터 스토리지/클러스터 노드(602-608)에 대한 리퍼럴을 포함하는 리퍼럴 리스트를 수신한다. 리퍼럴(602-608)은 자신의 리퍼럴과 디센던트에 의해 참조되는 데이터 서브세트에 대한 포트 식별자, 데이터 오프셋 및 데이터 길이를 포함한다. 리퍼럴(602-608)에 대한 데이터 스토리지/클러스터 노드는 모두 리퍼럴 리스트(650-653)를 구비한다. 리퍼럴(608)의 데이터 스토리지/클러스터 노드는 데이터 스토리지/클러스터 노드(628-632)에 대한 리퍼럴을 포함하는 리퍼럴 리스트(화살표(653)에 의해 도시됨)를 구비한다. 리퍼럴(628-

632)의 데이터 스토리지/클러스터 노드는 모두 리퍼럴(608)의 데이터 스토리지/클러스터 노드의 디센던트이며, 따라서 리퍼럴(608)의 데이터 길이는 리퍼럴(628-632)의 총 데이터 길이와 리퍼럴(608)의 서브세트 데이터 길이의 합을 포함한다. 유사하게, 리퍼럴(604)의 데이터 스토리지/클러스터 노드는 데이터 스토리지/클러스터 노드(616-620)에 대한 리퍼럴을 포함하는 리퍼럴 리스트(화살표(651)에 의해 도시됨)를 구비한다. 리퍼럴(616-620)의 데이터 스토리지/클러스터 노드는 모두 리퍼럴(604)의 디센던트이다. 데이터 스토리지/클러스터 노드(610-614, 616-620, 622-626, 628-632)의 리퍼럴은 디센던트를 갖지 않는다.

[0030] 블록 스토리지 클러스터 기술은 다수의 속성(attribute)을 제공할 것이 요구될 수 있다. 예를 들어, 블록 스토리지 프로토콜 타겟은 클러스터 내의 모든 스토리지 시스템(106-108)에 걸쳐 분포될 것이 요구될 수 있다. 또한, 클러스터 내의 모든 스토리지 시스템 상의 모든 포트는 고유한 포트 식별자를 구비할 것이 요구될 수 있다. 또한, 가상 볼륨에 대한 논리 블록 어드레스 공간은 가상 볼륨이 존재하는 모든 스토리지 시스템에 걸쳐 공통적일 것이 요구될 수 있다. 또한, 모든 스토리지 시스템(106-108) 상의 클러스터 블록 가상화 기능부(134)가 클러스터 내의 어떤 스토리지 시스템이 가상 볼륨(128, 130) 내의 데이터의 어드레스 범위를 홀딩하는지 결정할 수 있을 것이 요구될 수 있다.

[0031] 전술된 바와 같이, 본 발명의 방법은 스토리지 시스템(들)(106-108) 상의 블록 가상화를 제공하는 블록 클러스터 내에서 구현될 수 있다. 예시적인 실시예에서, 커맨드 포워딩 또는 프록시 I/O를 이용하지 않는 본 발명의 시스템/방법은, 데이터가 상태 정보 및 SCSI 감지 데이터 내의 리퍼럴의 리스트를 이용하여 로컬 데이터 전송을 완료함으로써 다른 클러스터 노드 상에 존재함을 나타내는 클러스터 블록 가상화(132, 134)를 구현한다. 상태 정보는 SCSI 검사 상태를 포함할 수 있다.

[0032] 다른 실시예에서, SCSI 개시자(102)는 새로운 검사 상태를 검출하고, 각 리퍼럴에 대한 새로운 SCSI 커맨드를 발행하며, 모든 리퍼럴이 완료되었을 때 트랙킹하도록 구성될 수 있다. 개시자(102)는 추가로 복수의 개시자-타겟 넥서스에 걸친 리퍼럴을 통해 검색된 데이터를 축적하도록 구성될 수도 있다.

[0033] 본 발명에서, 개시된 방법은 디바이스에 의해 판독가능한 소프트웨어 또는 장치의 세트로서 구현될 수 있다. 또한, 기술된 방법의 단계들의 특정한 순서 또는 계층은 설명적인 접근을 위한 예시임을 이해할 것이다. 바람직한 설계에 기초하여, 개시된 청구사항의 범주 내에서 이 방법의 단계들의 특정한 순서 또는 계층이 재배열될 수도 있음을 이해할 것이다. 첨부된 방법 청구항은 다양한 단계의 요소들을 샘플 순서로 나타내었으며, 기술된 특정한 순서 또는 계층으로 한정되어야 하는 것은 아니다.

[0034] 본 발명과 그에 수반되는 다수의 장점이 전술된 설명에 의해 이해될 것이며, 개시된 청구 사항으로부터 벗어나지 않고 본 발명의 물질적인 모든 장점을 희생하지 않은 채 구성요소들의 구조 및 배열의 형태로 다양한 변경이 이루어질 수 있음이 명백하다. 기술된 형태는 단지 예시적인 것이며, 아래의 특허청구범위는 이러한 변경을 모두 포괄하고 포함한다.

도면의 간단한 설명

[0035] 도 1은 본 발명에 따른 블록 스토리지 프로토콜을 통해 액세스 가능한 네트워크화된 스토리지 구현/시스템의 블록도,

[0036] 도 2는 본 발명에 따른 또는 본 발명에 의해 구현되는, 리퍼럴을 이용한 SCSI 커맨드/응답 원격 절차 호출의 개략적인 블록도,

[0037] 도 3은 본 발명에 따른 블록 스토리지 클러스터와 개시자 시스템 사이의 통신을 위한 방법을 도시한 순서도,

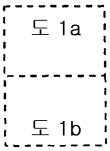
[0038] 도 4는 본 발명의 실시예에 따른 SCSI 리퍼럴 리스트 포맷을 도시한 도면,

[0039] 도 5는 링크된 리퍼럴 리스트 분배를 위한 일련의 SCSI 리퍼럴을 도시한 블록도,

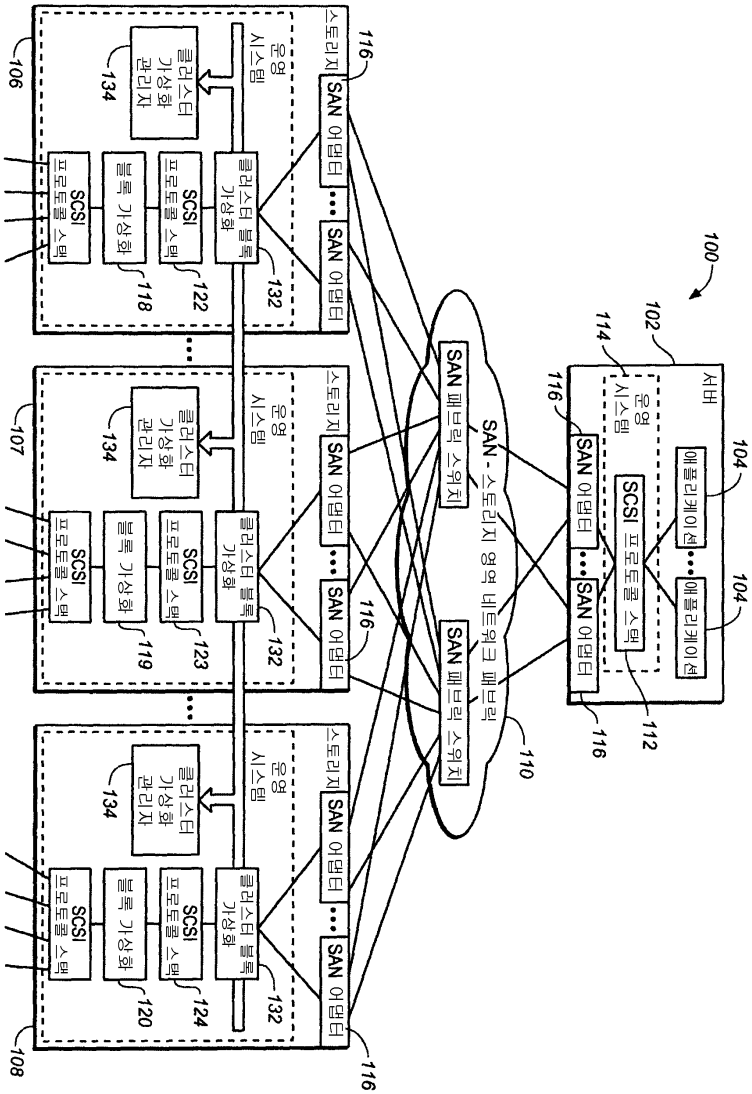
[0040] 도 6은 리퍼럴 리스트 트리 분배를 위한 일련의 SCSI 리퍼럴을 도시한 블록도.

도면

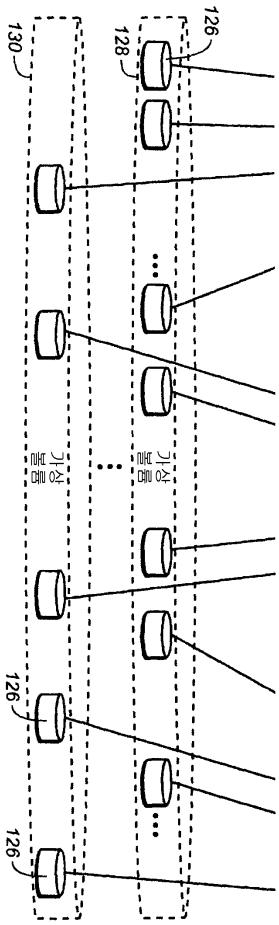
도면1


나 1a
나 1b

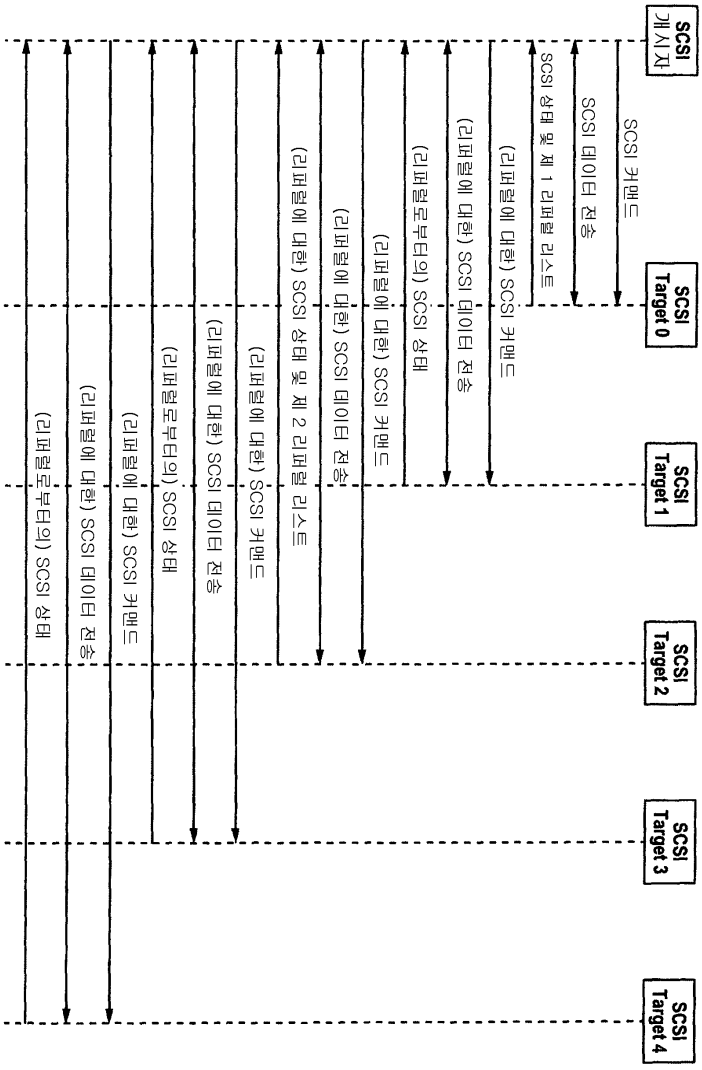
도면1a


100
서버
102
운영 시스템
114
SCSI 프로토콜 스택
112
애플리케이션
104
SAN 어댑터
116
SAN - 스토리지 영역 네트워크 패브릭
110
SAN 패브릭 스위치
스토리지
106
107
108
클러스터 가상화 관리자
134
SCSI 프로토콜 스택
블록 가상화
118
119
120
클러스터 블록 가상화
132
SCSI 프로토콜 스택
122
123
124

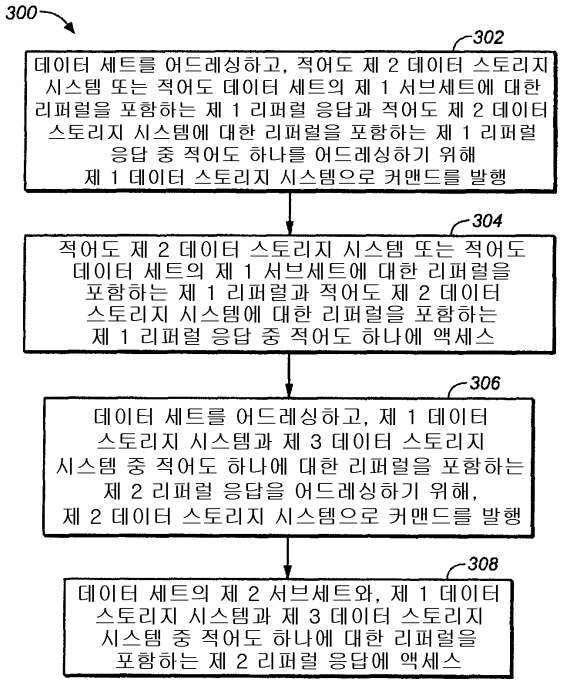
도면1b



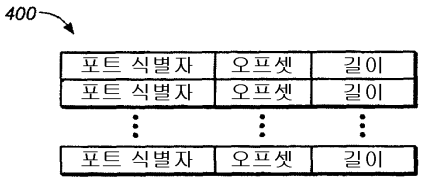
도면2


SCSI 개시자
SCSI Target 0
SCSI Target 1
SCSI Target 2
SCSI Target 3
SCSI Target 4
SCSI 커맨드
SCSI 데이터 전송
SCSI 상태 및 제 1 리퍼럴 리스트
(리퍼럴에 대한) SCSI 커맨드
(리퍼럴에 대한) SCSI 데이터 전송
(리퍼럴로부터의) SCSI 상태
(리퍼럴에 대한) SCSI 커맨드
(리퍼럴에 대한) SCSI 데이터 전송
(리퍼럴에 대한) SCSI 상태 및 제 2 리퍼럴 리스트
(리퍼럴에 대한) SCSI 커맨드
(리퍼럴에 대한) SCSI 데이터 전송
(리퍼럴로부터의) SCSI 상태
(리퍼럴에 대한) SCSI 커맨드
(리퍼럴에 대한) SCSI 데이터 전송
(리퍼럴로부터의) SCSI 상태

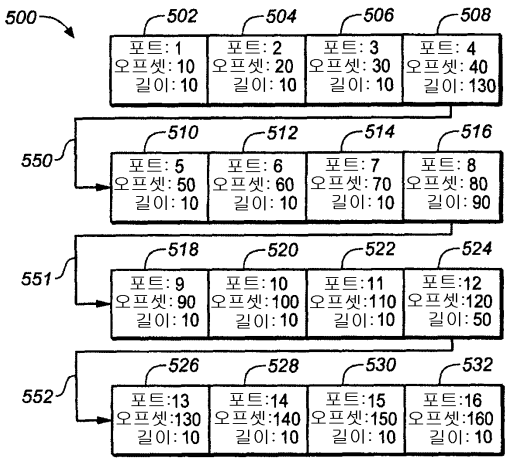
도면3


300
302
데이터 세트를 어드레싱하고, 적어도 제 2 데이터 스토리지 시스템 또는 적어도 데이터 세트의 제 1 서브세트에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답과 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나를 어드레싱하기 위해 제 1 데이터 스토리지 시스템으로 커맨드를 발행
304
적어도 제 2 데이터 스토리지 시스템 또는 적어도 데이터 세트의 제 1 서브세트에 대한 리퍼럴을 포함하는 제 1 리퍼럴과 적어도 제 2 데이터 스토리지 시스템에 대한 리퍼럴을 포함하는 제 1 리퍼럴 응답 중 적어도 하나에 액세스
306
데이터 세트를 어드레싱하고, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답을 어드레싱하기 위해, 제 2 데이터 스토리지 시스템으로 커맨드를 발행
308
데이터 세트의 제 2 서브세트와, 제 1 데이터 스토리지 시스템과 제 3 데이터 스토리지 시스템 중 적어도 하나에 대한 리퍼럴을 포함하는 제 2 리퍼럴 응답에 액세스

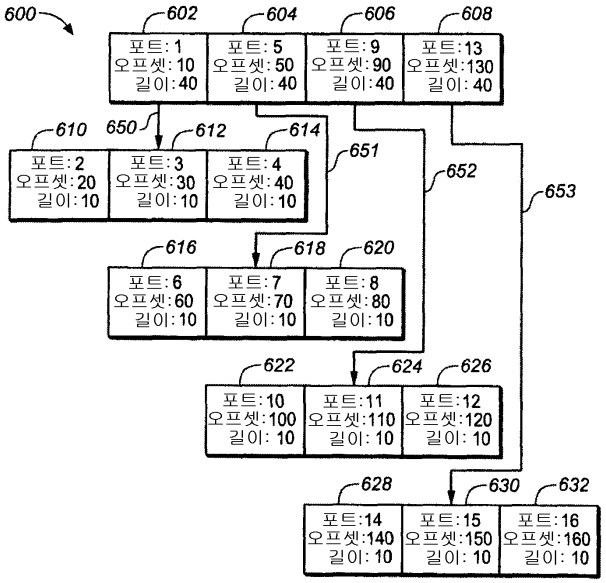
도면4


400
포트 식별자
오프셋
길이
포트 식별자
오프셋
길이
포트 식별자
오프셋
길이

도면5


500
502
504
506
508
포트: 1 오프셋: 10 길이: 10
포트: 2 오프셋: 20 길이: 10
포트: 3 오프셋: 30 길이: 10
포트: 4 오프셋: 40 길이: 130
550
510
512
514
516
포트: 5 오프셋: 50 길이: 10
포트: 6 오프셋: 60 길이: 10
포트: 7 오프셋: 70 길이: 10
포트: 8 오프셋: 80 길이: 90
551
518
520
522
524
포트: 9 오프셋: 90 길이: 10
포트: 10 오프셋: 100 길이: 10
포트: 11 오프셋: 110 길이: 10
포트: 12 오프셋: 120 길이: 50
552
526
528
530
532
포트: 13 오프셋: 130 길이: 10
포트: 14 오프셋: 140 길이: 10
포트: 15 오프셋: 150 길이: 10
포트: 16 오프셋: 160 길이: 10

도면6


600
602
포트:1
오프셋:10
길이:40
604
포트:5
오프셋:50
길이:40
606
포트:9
오프셋:90
길이:40
608
포트:13
오프셋:130
길이:40
610
포트:2
오프셋:20
길이:10
650
612
포트:3
오프셋:30
길이:10
614
포트:4
오프셋:40
길이:10
651
652
653
616
포트:6
오프셋:60
길이:10
618
포트:7
오프셋:70
길이:10
620
포트:8
오프셋:80
길이:10
622
포트:10
오프셋:100
길이:10
624
포트:11
오프셋:110
길이:10
626
포트:12
오프셋:120
길이:10
628
포트:14
오프셋:140
길이:10
630
포트:15
오프셋:150
길이:10
632
포트:16
오프셋:160
길이:10