(19)대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。Int. Cl.
G06F 11/34 (2006.01)
G06F 11/30 (2006.01)

(11) 공개번호 10-2007-0057828
(43) 공개일자 2007년06월07일

(21) 출원번호 10-2007-7005709
(22) 출원일자 2007년03월12일
심사청구일자 없음
번역문 제출일자 2007년03월12일
(86) 국제출원번호 PCT/EP2005/054594
국제출원일자 2005년09월15일
(87) 국제공개번호 WO 2006/037719
국제공개일자 2006년04월13일

(30) 우선권주장 10/959,859 2004년10월05일 미국(US)

(71) 출원인 인터내셔널 비지네스 머신즈 코포레이션
미국 10504 뉴욕주 아몬크 뉴오차드 로드

(72) 발명자 후드 로버트 앨렌
미국 플로리다주 33433 보카 래턴 래고 매르 서클 23265
스튜어트 앨랜
미국 뉴욕주 10583 스캐스데일 리온스 로드 156

(74) 대리인 신정건
송승필

전체 청구항 수 : 총 10 항

(54) 디스크 데이터 기억 시스템의 유지보수비를 결정하기 위한주문형 비용량성 기반 방법, 장치 및 컴퓨터 프로그램

(57) 요약

본 발명의 일 태양에 따르면, 데이터 기억 시스템의 유지보수비를 결정하는 동작을 수행하기 위해 디지털 처리 장치에 의해 실행가능한 기계 판독가능한 명령어의 프로그램을 실체적으로 구체화하는 신호 운반 매체를 제공한다. 상기 동작은 데이터 기억 시스템의 동작 중에 적어도 하나의 데이터 기억 장치를 감시하여 듀티 사이클을 결정하는 동작 및 적어도 부분적으로 상기 결정된 듀티 사이클에 기초하여 유지보수비의 현재 값을 결정하는 동작을 포함한다. 추가의 디스크 드라이브 기반 실시예에서, 상기 동작은 추가적으로 또는 대신으로 데이터 기억 시스템의 디스크 드라이브의 디스크 드라이브 용장 구성을 결정하는 동작을 포함할 수 있다. 상기 동작은 그 다음에, 상기 결정된 듀티 사이클과 임계값을 비교하고, 그 비교 결과에 기초하여 용장 구성 변경 신호를 어서트한다. RAID 레벨을 가진 RAID 구성 실시예의 경우, 상기 어서트하는 동작은 비교의 결과에 기초하여 RAID 레벨 변경 신호를 어서트한다.

대표도

도 2

특허청구의 범위

청구항 1.

데이터 기억 시스템의 유지보수비를 결정하는 동작을 수행하기 위해 디지털 처리 장치에 의해 실행가능한 기계 판독가능한 명령어의 프로그램을 실체적으로 구체화하는 신호 운반 매체에 있어서, 상기 동작이,

데이터 기억 장치 듀티 사이클을 결정하기 위해 데이터 기억 시스템의 동작 중에 적어도 하나의 데이터 기억 장치를 감시하는 동작; 및

적어도 부분적으로 상기 결정된 데이터 기억 장치 듀티 사이클에 기초하여 유지보수비의 현재 값을 결정하는 동작을 포함하는 것인 신호 운반 매체.

청구항 2.

제1항에 있어서, 상기 데이터 기억 장치는 디스크 드라이브로 구성되고, 상기 감시하는 동작은 적어도 다수의 읽기 동작을 샘플링 간격으로 기록하는 동작을 더 포함하는 것인 신호 운반 매체.

청구항 3.

제1항 또는 제2항에 있어서, 상기 데이터 기억 장치는 디스크 드라이브로 구성되고, 상기 감시하는 동작은 적어도 다수의 쓰기 동작을 샘플링 간격으로 기록하는 동작을 더 포함하는 것인 신호 운반 매체.

청구항 4.

데이터 기억 시스템의 디스크 드라이브의 디스크 드라이브 용장 구성을 결정하는 동작을 수행하기 위해 디지털 처리 장치에 의해 실행가능한 기계 판독가능한 명령어의 프로그램을 실체적으로 구체화하는 신호 운반 매체에 있어서, 상기 동작이,

듀티 사이클을 결정하기 위해 데이터 기억 시스템의 동작 중에 상기 디스크 드라이브를 감시하는 동작;

상기 결정된 듀티 사이클을 임계값과 비교하는 동작; 및

상기 비교 결과에 기초하여 용장 구성 변경 신호를 어서트하는 동작을 포함하는 것인 신호 운반 매체.

청구항 5.

디스크 드라이브에 있어서,

적어도 하나의 회전식 디스크와;

적어도 하나의 이동가능한 디스크 읽기/쓰기 헤드와;

최소한 상기 적어도 하나의 읽기/쓰기 헤드에 결합되고, 샘플링 간격으로 디스크 드라이브 듀티 사이클을 묘사하는 유지보수비 결정 함수에 관한 정보를 기록 및 보고하는 감시 장치

를 포함하는 디스크 드라이브.

청구항 6.

디스크 드라이브에 있어서,

적어도 하나의 회전식 디스크와;

적어도 하나의 이동가능한 디스크 읽기/쓰기 헤드와;

최소한 상기 적어도 하나의 읽기/쓰기 헤드에 결합되고, 샘플링 간격으로 디스크 드라이브 듀티 사이클을 묘사하는 RAID 레벨 결정 함수에 관한 정보를 기록 및 보고하는 감시 장치

를 포함하는 디스크 드라이브.

청구항 7.

적어도 하나의 디스크 드라이브에 결합된 디스크 드라이브 제어기에 있어서,

각 디스크 드라이브는 적어도 하나의 회전식 디스크와 적어도 하나의 이동가능한 디스크 읽기/쓰기 헤드를 포함하는 것이고;

상기 디스크 드라이브 제어기는, 최소한 상기 적어도 하나의 읽기/쓰기 헤드에 결합되어 샘플링 간격으로 디스크 드라이브 듀티 사이클을 묘사하는 유지보수비 결정 함수에 관한 정보를 기록 및 보고하는 감시 장치를 포함하는 것인 디스크 드라이브 제어기.

청구항 8.

적어도 하나의 디스크 드라이브에 결합된 디스크 드라이브 제어기에 있어서,

각 디스크 드라이브는 적어도 하나의 회전식 디스크와 적어도 하나의 이동가능한 디스크 읽기/쓰기 헤드를 포함하는 것이고;

상기 디스크 드라이브 제어기는, 최소한 상기 적어도 하나의 읽기/쓰기 헤드에 결합되어 샘플링 간격으로 디스크 드라이브 듀티 사이클을 묘사하는 RAID 레벨 결정 함수에 관한 정보를 기록 및 보고하는 감시 장치를 포함하는 것인 디스크 드라이브 제어기.

청구항 9.

디스크 드라이브 데이터 기억 시스템을 동작시키기 위해 청구되는 유지보수비의 값을 확립하도록 동작하는 유지보수 제공자 시스템에 있어서,

상기 디스크 드라이브 데이터 기억 시스템의 동작을 감시하여 디스크 드라이브의 사용량을 결정하는 감시 장치에 결합하기 위한 인터페이스와;

인터페이스를 통해 감시 장치로부터 수신한 정보에 응답하여 디스크 드라이브 이용율을 결정하고, 적어도 부분적으로 상기 결정된 이용율에 기초하여 유지보수비의 현재 값을 결정하는 프로세서

를 포함하는 유지보수 제공자 시스템.

청구항 10.

디스크 드라이브 기반 데이터 기억 시스템에 있어서,

단위 시간당의 다수의 쓰기 동작, 단위 시간당의 다수의 읽기 동작 및 단위 시간당의 다수의 시크 동작 중의 적어도 하나를 포함하는 디스크 드라이브의 사용량 정보를 결정하기 위해서 데이터 기억 시스템의 동작 중에 상기 디스크 드라이브를 감시하는 수단과;

상기 데이터 기억 시스템의 유지보수비를 결정하는 수단에 결합하기 위한 인터페이스 수단

을 포함하고,

상기 인터페이스 수단은 사용량 정보를 상기 결정하는 수단에 보고하는 것인 디스크 드라이브 기반 데이터 기억 시스템.

명세서

기술분야

본 발명은 일반적으로 디스크 기반(disk-based) 데이터 기억 시스템 및 방법에 관한 것으로서, 보다 구체적으로는 복수의 디스크 드라이브로 구성된 데이터 기억 시스템 및 그러한 데이터 기억 시스템의 유지보수비(maintenance fee)를 결정하고 청구하는 기술에 관한 것이다.

배경기술

데이터 기억 시스템에서, 독립 기억 장치의 어레이는 RAID(버클리에 있는 캘리포니아 대학의 연구자에 의해 최초에 'Redundant Array of Inexpensive Disks'라고 불리었던 Redundant Array of Independent Disks B)라고 알려진 기술을 이용하여 단일의 가상 기억 장치로서 동작하도록 구성될 수 있다. 이 명세서에서 '디스크(disk)'는 가끔 '디스크 드라이브(disk drive)'의 축약어로서 사용된다.

RAID 기억 시스템은 독립 기억 장치의 어레이 및 적어도 하나의 RAID 제어기를 포함한다. RAID 제어기는 독립 기억 장치의 어레이의 가상화 모습(virtualized view)을 제공하고, RAID 기억 시스템과 함께 동작하도록 구성된 컴퓨터 시스템은 마치 RAID 기억 시스템의 독립 기억 장치의 어레이가 단일 기억 장치인 것처럼 입력 및 출력(I/O) 동작을 수행할 수 있다. 따라서, 기억 장치의 어레이는 기억 소자들의 순차적인 리스트와 함께 단일 가상 기억 장치로서 나타난다. 기억 소자들은 기억 블록이라고 공통적으로 알려져 있고, 데이터 블록에 저장된 데이터는 데이터 블록이라고 알려져 있다. I/O 동작(예를 들면, 읽기와 쓰기)은 가상 기억 장치 내의 하나 이상의 기억 블록들을 참조해서 권한이 부여된다. I/O 동작이 가상 기억 장치에서 수행될 때, RAID 제어기는 I/O 동작을 독립 기억 장치의 어레이에 맵한다. 기억 장치의 어레이를 가상화하고 I/O 동작을 맵하기 위해, RAID 제어기는 당업계에 잘 알려져 있는 표준 RAID 기술을 이용할 수 있다. 이러한 기술들의 일부가 이하에서 간단히 설명된다.

RAID 제어기는 데이터 블록들을 독립 기억 장치의 어레이 전반에 분산시킨다. 이것을 달성하기 위한 한가지 방법은 스트리핑(striping)이라고 알려진 기술을 사용하는 것이다. 스트리핑은 기억 장치 전반에 데이터 블록들을 라운드 로빈(round-robin) 방식으로 분산시키는 것을 수반한다. RAID 기억 시스템에 데이터 블록들을 저장할 때, 스트립이라고 알려져 있는 다수의 데이터 블록들이 각 기억 장치에 저장된다. 스트립의 크기는 특수한 RAID 구현예에 의해 결정될 수도 있고, 또는 구성가능하게 될 수도 있다. 제1 기억 장치에 저장된 제1 스트립 및 후속 기억 장치에 저장된 후속 스트립을 포함한 스트립의 행(row)은 스트라이프(stripe)라고 알려져 있다. 스트라이프의 크기는 스트라이프를 구성하는 모든 스트립의 총합 크기(total size)이다.

이러한 방법으로 데이터 블록을 저장하기 위해 다중의 독립 기억 장치를 사용하는 경우, 다중 기억 장치가 I/O 동작 중에 병렬로 동작하기 때문에 단일 기억 장치에 비하여 고성능의 I/O 동작을 제공한다. 성능 개선은 RAID 기술의 중요한 이익 중의 하나이다. 하드 디스크 드라이브 성능은 하드 디스크 드라이브가 전형적인 컴퓨터의 최저속 내부 구성 요소들 중의 일부이기 때문에 컴퓨터 시스템에서 중요하다.

일부 하드 디스크 드라이브는 신뢰도가 낮은 것으로 알려져 있고, 아직도 하드 디스크 드라이브 신뢰도는 회복할 수 없는 데이터 손실의 심각한 결과(또는 데이터의 일시적 접근곤란성) 때문에 중요하다. 전형적인 RAID 기억 시스템의 중요 목적은 신뢰성있는 데이터 저장을 제공하는 것이다.

신뢰도를 제공하는 하나의 기술은 독립 기억 장치의 어레이에 데이터와 함께 체크 정보의 저장을 수반한다. 체크 정보는 기억 장치의 어레이에서 단일의 기억 장치가 고장으로 되는 것과 같은 단일의 고장점 때문에 읽기불능으로 된 데이터의 표시를 가능하게 하는 용장성 정보(redundant information)이다. 읽기불능 데이터는 읽기가능 데이터와 용장성 체크 정보의 조합으로부터 재발생된다. 체크 정보는 스트라이프 내의 단일 스트립을 점유하는 '패리티' 데이터로서 기록되고, 스트라이프 내의 모든 데이터 스트립에 배타적 OR(XOR) 논리 연산자를 적용함으로써 계산되어진다. 예를 들면, 데이터 스트립 A, B 및 C를 포함하는 스트라이프는 A XOR B XOR C로서 계산된 관련 패리티 스트립을 가질 것이다. 기억 시스템에서 단일 고장점이 발생한 경우, 패리티 스트립은 액세스불능 데이터 스트립을 재발생하기 위해 사용된다. 데이터 스트립 A, B, C와 패리티를 포함한 스트라이프가 4개의 독립 기억 장치(W, X, Y, Z)에 각각 저장되어 있고 기억 장치 X가 고장이면, 기억 장치 X에 저장된 스트립 B가 액세스불능으로 될 것이다. 스트립 B는 나머지 데이터 스트립과 패리티 스트립으로부터 XOR 연산을 통하여 계산될 수 있다. 이 복구성 연산은 'A XOR C XOR 패리티 = B'이다. 이것은 임의의 하나의 손실된 스트립(A, B 또는 C)을 생성하기 위한 XOR 동작의 가역 특성(reversible nature)을 나타낸다. 물론, 손실 데이터가 패리티 정보인 경우에도 앞의 XOR 동작이 반복될 수 있다.

스트리핑(병렬 동작의 성능 이익을 위한 것) 및 패리티(용장성을 위한 것) 외에, 일부 RAID 솔루션에서 사용되는 다른 용장성 기술은 미러링(mirroring)이다. 미러링을 사용하는 RAID 시스템에서, 시스템의 모든 데이터는 2 개의 하드 디스크 드라이브에 동시에 기록된다. 이것은 이중화 데이터를 내포한 디스크 중 어느 하나의 고장을 보호하고 (다른 디스크가 고장인 경우에도 데이터가 하나의 디스크에서 사용할 준비가 되어 있기 때문에) 디스크 고장으로부터 비교적 빠른 복구를 가능하게 한다. 이러한 장점은 (디스크 공간의 절반이 이중화 데이터를 기억하기 위해 사용되기 때문에) 비용이 증가하는 단점과 균형이 맞아야 한다. 이중화는 RAID 제어기뿐만 아니라 디스크 드라이브 B를 이중화하여 제어기의 고장뿐만 아니라 디스크 드라이브 고장을 보호하는 미러링의 확장이다.

다른 RAID 구현예는 상기 기술의 다른 조합을 이용한다. 많은 표준화 RAID 방법이 단일 RAID "레벨들" 0-7로서 식별되고, "내포(nested)" RAID 레벨이 예를 들면 아래와 같이 또한 규정된다.

RAID 1은 결함 허용성(fault tolerance)을 위해 미러링(또는 이중화)를 사용하고; 한편 RAID 0은 패리티없는 블록 레벨 스트리핑을 사용한다. 즉, 용장성이 없고 따라서 다른 RAID 레벨의 결함 허용성이 없으며, 그래서 그 비용에 비하여 좋은 성능을 갖는다. RAID 0은 전형적으로 비임계치 데이터(즉, 드물게 변화하고 규칙적으로 백업되는 데이터)에 대하여 사용되고, 고속 및 저비용이 신뢰성보다 더 중요한 경우에 사용된다.

RAID 3과 RAID 7은 패리티가 있는 바이트 레벨 스트리핑을 사용한다.

RAID 4, RAID 5 및 RAID 6은 패리티가 있는 블록 레벨 스트리핑을 사용한다. RAID 5는 어레이 내의 모든 드라이브에 대하여 분산 패리티 알고리즘, 쓰기 데이터 및 패리티 블록을 사용한다(이것은 쓰기 성능을 약간 개선하고 RAID 4의 전용 패리티 드라이브와 비교해서 개선된 병렬 계산(parallelism)을 가능하게 한다). 결함 허용성은 임의의 정해진 데이터 블록의 패리티 정보가 데이터 자체를 저장하기 위해 사용되는 드라이브로부터 분리되어 드라이브에 저장되는 것을 보장함으로써 RAID 5에서 유지된다. RAID 5는 좋은 성능, 좋은 결함 허용성 및 고용량 및 저장 효율을 가지며, 쓰기 주도형(write-intensive)이 아닌 트랜잭션 처리 및 다른 애플리케이션과 같은 애플리케이션의 임의의 단일 RAID 레벨의 최상의 타협(compromise)으로 생각되고 있다.

위에서 설명한 단일 RAID 레벨 외에, 성능을 더욱 개선하기 위해 내포(nested) RAID 레벨이 또한 사용된다. 예를 들면, 고성능 RAID 0의 특징들은 결함 허용성을 또한 제공하기 위해 1, 3 또는 5 등의 용장성 RAID 레벨의 특징들과 함께 내포 구성으로 결합될 수 있다.

RAID 01은 2 개의 스트라이프 세트의 미러 구성이고, RAID 10은 다수의 미러 세트에 걸친 스트라이프이다. RAID 01과 RAID 10은 둘 다 고성능 및 좋은 결함 허용성(대부분의 사용에서)을 가진 대형 어레이를 생성할 수 있다.

RAID 15 어레이는 구성 요소로서 다수의 미러 쌍을 이용하여 패리티가 있는 스트라이프 세트를 생성함으로써 형성될 수 있다. 이와 유사하게, RAID 51은 전체 RAID 5 어레이를 미러링함으로써 생성되고, RAID 5 어레이 중 어느 하나의 각 멤버는 디스크 드라이브의 미러(RAID 1) 쌍으로서 저장된다. 데이터의 2 개의 카피는 추가적인 보호를 위해 물리적으로 다

른 장소에 위치될 수 있다. 훌륭한 결함 허용성 및 가용성은 이 방법으로 패리티의 용장성 방법과 미러링을 결합함으로써 달성될 수 있다. 예를 들면, 8 개의 드라이브 RAID 15 어레이는 임의의 3 개 드라이브의 동시 고장을 허용할 수 있다. 단일 디스크 고장 후에, 데이터는 여전히 단일 디스크 드라이브로부터 판독될 수 있고, RAID 5는 더 복잡한 재구축(rebuild)을 요구할 것이다.

일 예로서, RAID 5는 단일 드라이브 에러가 보정될 수 있게 한다. 예시적인 14 드라이브 RAID 5 시스템에서, 데이터를 저장하는 12 개의 드라이브가 있을 수 있는데, 그 중 하나의 드라이브는 패리티를 저장하기 위한 것이고, 하나의 스페어 드라이브는 RAID 재구축 동작 중에 고장 드라이브의 정보를 이주시키기 위한 것이다. 그러나, 14 드라이브 RAID 10 시스템은 6 개의 데이터 드라이브와 하나의 패리티 드라이브로 이루어진 2 개의 세트로 분할된다. 그 결과, 동일한 저장 용량 필요조건을 충족하기 위하여, RAID 5와 비교할 때, 약 2배인 물리적 디스크의 수를 필요로 한다는 것을 알 수 있다(정확한 비율은 2N/(N+ 1)이고, 여기에서 N은 RAID 5 어레이의 데이터 디스크의 수이다). RAID 5 구성의 시스템은 RAID 10 시스템에 이주될 수 있지만, 일반적으로 그 역으로는 이주될 수 없다고 알려져 있다.

비교적 저비용의 병렬 ATA(Advanced Technical Attachment) 디스크 드라이브는 또한 가끔 IDE(Integrated Drive Electronics) 드라이브라고도 부르며, 직렬 ATA, 즉 SATA 디스크 드라이브는 고객 개인용 컴퓨터(PC) 설비(데스크탑과 랩탑 둘 다)에서 수년간 폭넓게 사용되어 왔다. 그러나, 적어도 부분적으로는, 이들 디스크 드라이브의 데이터 저장 용량의 폭발적 증가에 응답하여, 위에서 간단히 설명한 것과 같은 RAID 기반 기억 시스템을 포함한 대규모의 개방 및 기업 레벨 디스크 기반 기억 시스템에서 ATA 및/또는 SATA 드라이브를 이용하기 위하여 개발중에 있는 추세이다.

이 추세의 결과로서 야기되는 문제점은, ATA 및 SATA 드라이브의 고유 신뢰도 및 결과적인 평균 고장 간격(MTBF)이 전통적으로 대규모의 기업 분류 디스크 기억 시스템에서 사용되었던 디스크 드라이브의 다른 유형에 대하여 크게 낮아질 수 있기 때문에, 신뢰도와 관련이 있다. 한가지 결과는 고장율 및 디스크 기억 시스템 제조업자의 후속되는 유지보수비가 전통적인 방식의 것, 즉 종래의 시스템에서 사용자에게 청구되는 유지보수비는 전형적으로 사용된 총 데이터 저장 용량의 함수로서 결정하는 방식의 것보다 더 커질 수 있다는 것이다.

미국 특허 제5,828,583호에는 ATA 디스크 드라이브, 및 디스크 드라이브의 임박한 고장(imminent failure)을 예측하기 위하여 동작 중에 특정의 속성들을 감시하는 것에 대하여 개시되어 있다.

미국 특허 제5,371,882호에는 용장 그룹을 가진 폼팩터(form factor)가 큰 디스크 드라이브 메모리에서 사용된 공유의 스페어 디스크 드라이브의 풀(pool)이 디스크 드라이브 고장 데이터의 기록 및 스페어 디스크 드라이브 고갈 목표 일자에 대한 과거 고장 사건의 추론에 의해 언제 고갈될 것인지를 예측하는 기술에 대하여 개시되어 있다.

미국 특허 제6,411,943 B1에는 그 57칼럼 46행부터 58칼럼 10행까지에 고객 대신 읽기 또는 쓰기된 시간량 및/또는 가상 디스크 기억 용량에 기초하여 고객에게 청구서를 보내는 온라인 서비스에 대하여 개시되어 있다.

발명의 상세한 설명

본 발명의 양호한 실시예에 따르면, 전술한 및 기타의 문제점들이 극복되고 다른 장점들이 실현된다.

본 발명의 일 태양에 따르면, 데이터 기억 시스템의 유지보수비를 결정하는 동작을 수행하기 위해 디지털 처리 장치에 의해 실행가능한 기계 판독가능한 명령어의 프로그램을 실체적으로 구체화하는 신호 운반 매체를 제공한다. 상기 동작은 데이터 기억 장치 듀티 사이클을 결정하기 위해 데이터 기억 시스템의 동작 중에 적어도 하나의 데이터 기억 장치를 감시하는 동작 및 적어도 부분적으로 상기 결정된 데이터 기억 장치 듀티 사이클에 기초하여 유지보수비의 현재 값을 결정하는 동작을 포함한다.

본 발명의 일 태양에 따르면, 데이터 기억 시스템의 디스크 드라이브의 디스크 드라이브 용장 구성을 결정하는 동작을 수행하기 위해 디지털 처리 장치에 의해 실행가능한 기계 판독가능한 명령어의 프로그램을 실체적으로 구체화하는 신호 운반 매체를 제공한다. 상기 동작은 듀티 사이클을 결정하기 위해 데이터 기억 시스템의 동작 중에 디스크 드라이브를 감시하는 동작, 결정된 듀티 사이클을 임계치와 비교하는 동작 및 비교 결과에 따라 용장 구성 변경 신호를 어서트(assert)하는 동작을 포함한다. 디스크 드라이브가 RAID 레벨을 가진 RAID 구성에서 동작하는 실시예의 경우, 상기 어서트하는 동작은 비교의 결과에 따라 RAID 레벨 변경 신호를 어서트한다.

본 발명의 다른 태양에 따르면, 적어도 하나의 회전식 디스크, 적어도 하나의 이동가능한 디스크 읽기/쓰기 헤드, 및 최소한 상기 적어도 하나의 읽기/쓰기 헤드에 결합되고 샘플링 간격에서 디스크 드라이브 듀티 사이클을 묘사하는 유지보수비 결정 함수 정보를 기록 및 보고하는 감시 장치(monitor)를 구비한 디스크 드라이브를 제공한다.

본 발명의 다른 태양에 따르면, 감시 장치가 샘플링 간격에서 디스크 드라이브 듀티 사이클을 묘사하는 RAID 레벨 결정 함수 정보를 기록 및 보고하기 위하여 최소한 상기 적어도 하나의 읽기/쓰기 헤드에 결합된 디스크 드라이브를 제공한다.

본 발명의 다른 태양에 따르면, 적어도 하나의 디스크 드라이브에 결합된 디스트 드라이브 제어기를 제공하고, 각 디스크 드라이브는 적어도 하나의 회전식 디스크와 적어도 하나의 이동가능한 디스크 읽기/쓰기 헤드를 포함한다. 디스크 드라이브 제어기는 유지보수비 결정 함수 및/또는 RAID 레벨 결정 함수에 대하여 샘플링 간격에서 디스크 드라이브 듀티 사이클을 묘사하는 정보를 기록 및 보고하기 위하여 최소한 상기 적어도 하나의 읽기/쓰기 헤드에 결합된 감시 장치를 포함한다.

본 발명의 또 다른 태양에 따르면, 디스크 드라이브 데이터 기억 시스템을 동작시키기 위해 청구되는 유지보수비의 값을 확립하도록 동작하는 유지보수 제공자 시스템이 제공된다. 유지보수 제공자 시스템은 디스크 드라이브 데이터 기억 시스템의 동작을 감시하여 디스크 드라이브의 사용을 결정하는 감시 장치에 결합하기 위한 인터페이스를 포함하고; 또한 인터페이스를 통해 감시 장치로부터 수신한 정보에 응답하여 디스크 드라이브 이용율을 결정하고, 적어도 부분적으로 상기 결정된 이용율에 기초하여 유지보수비의 현재 값을 결정하는 프로세서를 포함한다.

본 발명의 또 다른 태양에 따르면, 단위 시간당의 다수의 쓰기 동작, 단위 시간당의 다수의 읽기 동작 및 단위 시간당의 다수의 시크(seek) 동작 중의 적어도 하나를 포함하는 디스크 드라이브의 사용량 정보를 결정하기 위해 데이터 기억 시스템의 동작 중에 디스크 드라이브를 감시하는 수단; 및 데이터 기억 시스템의 유지보수비를 결정하는 수단에 결합하기 위한 인터페이스 수단을 포함하는 디스크 드라이브 기반 데이터 기억 시스템이 제공되고, 상기 인터페이스 수단은 사용량 정보를 상기 결정하는 수단에 보고한다.

RAID 레벨을 가진 RAID 구성에서 동작되는 복수의 디스크 드라이브가 있는 비제한적인 실시예에서, 디스크 드라이브 기반 데이터 기억 시스템은 사용량 정보를 임계치와 비교하는 수단 및 비교 결과에 따라 RAID 레벨 변경 신호를 어서트하는 수단을 더 포함한다.

본 발명에 따른 교시의 전술한 및 다른 태양은 첨부 도면을 참조하여 이하의 양호한 실시예에 대한 상세한 설명에서 더욱 명백하게 될 것이다.

실시예

도 1을 참조하면, 본 발명의 실시예에 따라 구성되고 동작되는 디스크 기반 데이터 기억 시스템(10)의 블록도가 도시되어 있다. 이 비제한적인 예의 디스크 드라이브 시스템(10)에서는 데이터 경로(12A)를 통하여 적어도 하나의 호스트 어댑터(14)에 결합된 메인프레임 또는 임의의 다른 적당한 유형의 컴퓨터와 같은 적어도 하나의 호스트(12)가 있는 것으로 가정한다. 적어도 하나의 호스트 어댑터(14)는 데이터 경로(14A)를 통해 데이터 캐시(16)에 결합되고, 이 데이터 캐시(16)는 데이터 경로(16A)를 통해 적어도 하나의 디스크 어댑터(18)에 결합된다. 적어도 하나의 디스크 드라이브, 그러나 전형적으로는 복수의 디스크 드라이브(20)는 데이터 경로(18A)를 통해 디스크 어댑터(18)에 결합된다. 디스크 드라이브(20)는 RAID 형식으로 조직될 수 있다. 예를 들면, 디스크 드라이브(20)는 RAID 제어기(20A)의 방향하에 RAID 5 구성으로 조직되고 동작할 수 있다(편리를 위해 디스크 어댑터(18)와 연합된 것으로 도시되어 있다). 이 설명을 위해, 그러나 본 발명의 실시에 제한을 주지 않는 예로서, 디스크 드라이브(20)는 ATA 또는 SATA형 디스크 드라이브일 수 있다.

이러한 본 발명의 태양에 따르면, 시스템(10)에는 디스크 사용량 감시 장치(22)가 제공된다. 디스크 사용량 감시 장치(22)는 디스크 드라이브(20)에 직접 또는 간접적으로 결합되고, 시간에 따른 디스크 동작(disk activity)의 기록을 유지한다. 유지보수 측량치(metric) 또는 디스크 사용량 측량치 또는 디스크 이용율로서 생각될 수 있는 디스크 동작은 단위 시간당의 다수의 디스크 쓰기 동작(예를 들면, 5분 동안 또는 반시간 동안의 다수의 쓰기 동작), 및/또는 단위 시간당의 다수의 디스크 읽기 동작, 및/또는 단위 시간당의 다수의 디스크 시크 동작을 포함할 수 있다. 만일 관심 대상 측량치가 디스크 시크 동작이면, 디스크 드라이브 시크 이벤트의 구성 요소는 디스크 헤드가 이동하도록 요구된(예를 들면, 트랙에서 측정된) 실제 거리일 수 있다. 디스크 드라이브의 유형에 따라서, 디스크 동작은 디스크 드라이브가 단위 시간당 스핀업 및/또는 스핀다운하는 횟수를 또한 포함할 수 있다.

이러한 각종의 예시적 유형의 디스크 동작은 디스크 드라이브(20)의 기계적 동작을 반영하고, 그에 따라서 단위 시간당 디스크 드라이브 사용량, 또는 디스크 드라이브 듀티 사이클의 표시이며, 일부의 경우에는 단위 시간당의 입출력(I/O) 이용량의 표시임을 알 수 있다. 단위 시간당 디스크 드라이브의 사용량이 더 많아지면, 즉 디스크 드라이브(20)의 듀티 사이클이 더 커지면, 고장의 가능성이 더 커지고, 따라서 MTBF가 그에 비례하여 감소되는 것으로 예상된다. 이러한 각종의 유지보수 측량치는 종래의 유지보수 관련 인자의 경우에서와 마찬가지로, 디스크 드라이브 시스템(10)의 기사용(used) 기억 용량을 직접 표시하지 않는다는 것을 또한 알 수 있다.

감시 장치(22)는 디스크 어댑터(18) 또는 RAID 제어기(20A)의 레벨에서 물리적으로 위치될 수 있고, 또는 특히 각 디스크 드라이브가 내장형 디스크 드라이브 제어기(IDE 드라이브용)를 포함하고 있다면 디스크 드라이브(20) 전체에 분산될 수 있으며, 또는 더 낮은 선호도로서 캐시(16)의 레벨에서 물리적 디스크 드라이브(20)로부터 더 멀리 위치될 수 있다는 점이 주목된다. 상기 후자의 경우에, 감시 장치(22)는 캐시 히트 대 미스뿐만 아니라 캐시 재기록(writeback)만을 인식할 수 있고, 그에 따라서 이들 및 다른 캐시 동작을 물리적 디스크 드라이브(20)의 실제 사용량과 상관시킬 수 있다. 감시 장치(22)는 하드웨어로 구현될 수 있지만, 소프트웨어 또는 펌웨어로 구현되는 것이 가장 바람직하고, 디스크의 헤드가 움직일 때마다, 및/또는 디스크 읽기 동작이 수행될 때마다, 및/또는 디스크 쓰기 동작이 수행될 때마다 디스크 드라이브 제어기 또는 어떤 다른 로직에 의해 실행되는 루틴으로서 구체화될 수 있다. 감시 장치(22)는 또한 예를 들면 디스크 시크 또는 읽기 또는 쓰기 동작이 수행될 때마다 펌웨어에 의해 증분되는 적어도 하나의 하드웨어 카운터를 제공함으로써 하드웨어와 소프트웨어의 조합으로서 구현될 수 있다. 가장 바람직한 실시예에서, 하드웨어 카운터는 소프트웨어에 의해 관리되는 하나의 메모리 위치 또는 다수의 메모리 위치로서 간단히 구현될 것이다.

디스크 드라이브(20)의 감시 결과를 보고하기 위해, 링크(22A)가 원격으로 위치된 또는 함께 위치된 유지보수 제공자(24)와의 통신을 위해 제공되는 것이 바람직하다. 유지보수 제공자(24)는 시스템(10)의 전부 또는 일부의 제조자가 될 수도 있고, 또는 유지보수 제공자(24)는 시스템(10)에 대하여 유지보수 서비스만을 제공하기로 계약한 제3자가 될 수도 있다. 링크(22A)는 전용 접속일 수도 있고, 또는 근거리 통신망(LAN)을 이용함으로써, 또는 인터넷을 포함한 광역 통신망을 이용함으로써 구현될 수 있으며, 또한 TCP/IP를 통하여 및/또는 무선 통신 프로토콜을 포함한 다른 프로토콜을 통하여 통신할 수도 있다. 일반적으로, 링크(22A)는 디스크 사용량 감시 장치(22)의 출력을 매뉴얼 수단을 포함한 유지보수 제공자(24)에게 운반하는 임의의 수단을 나타내는 것으로 볼 수 있다(예를 들면, 디스크 드라이브 유지보수 측량 데이터가 수록된 디스켓을 주기적으로 메일링할 수 있다).

감시 장치(22)가 디스크 드라이브(20)를 통하여 분산된 경우에, 도 4 참조를 참조하면, 디스크 사용량 데이터를 각 디스크 드라이브(20)의 로컬 감시 장치(LM)(22B)로부터 수집하고 드라이브 사용량 통계를 링크(22A)를 통해 유지보수 제공자(24)에게 회송하는 공통 엔티티(CE)(22C)가 예를 들면 RAID 제어기(20A)에 하나씩 있을 수 있다. 상기 공통 엔티티(22C)는, 이 설명의 목적상, 분산된 디스크 사용량 감시 장치(22)의 I/O 인터페이스를 형성하기 위한 것으로 생각할 수 있다.

감시 장치(22)로부터 출력된 디스크 사용량 데이터가 감시 장치(22)로부터 유지보수 제공자(24)에게 주기적으로 전송될 수 있지만, 디스크 사용량 데이터 출력은 유지보수 제공자(24)에 의해 감시 장치(24)로부터 주기적으로 추출되는 것이 더 바람직하다. 전송 또는 추출 주기의 길이, 즉 디스크 사용량 측량치 보고 빈도는 고정되어 있을 수도 있고, 또는 시스템(10)의 동작 중에 링크(22A)를 통해 유지보수 제공자(24)에 의해 프로그램 및 변경 가능하게 될 수도 있다.

도 2와 도 3을 참조하면, 유지보수 제공자(24)는 적당한 링크 인터페이스(I/F)(24B)에 의해 링크(22A)에 결합된 데이터 프로세서(24A)를 포함할 수 있다. 수신된 디스크 사용량 데이터는 디스크 사용량 측량치 기억 장치 또는 메모리(24C)에 주기적으로 저장되고(도 3의 블록 3A), 디스크 드라이브(20)의 실제 사용량, 즉 이용율을 반영하는 추세(예를 들면, 몇 가지 비제한적인 예로서, 시간별 추세, 일별 추세, 주별 추세, 월별 추세 또는 분기별 추세 등)를 얻기 위해 N 시간 주기(N 샘플 주기)동안 누적되는 것이 바람직하다(도 3의 블록 3B). 그 다음에, 이용율은 디스크 드라이브(20)의 실제 사용량에 기초하는 유지보수비 청구 금액을 반영하는 결과를 얻기 위하여 유지보수비 표(24D)의 포인터로서 사용될 수 있다(도 3의 블록 3C). 대안적으로, 이용율은 각 청구서 발송 주기의 유지보수 청구 금액을 동적으로 계산하기 위해 데이터 프로세서(24A)에 의해 풀어지는 공식에서 명시적으로 사용될 수 있다(도 3의 블록 3D). 양쪽 유형의 동작을 수행하는 것, 예를 들면, 표(24D)로부터 값을 얻기 위해 이용율을 이용하는 것 및 그 다음에 상기 얻어진 값을 유지보수비를 명시적으로 계산하기 위해 사용하는 것(또는 그 반대)은 본 발명의 범위에 또한 포함된다. 사용자에게는 청구된 금액을 유효화하기 위해 디스크 사용량 측량치 기억 장치(24C)의 저장된 기록에 기초하여 상세한 보고서가 제공될 수 있다는 것에 주목하여야 한다.

여기에서 사용하는 용어 이용율 및 이용 측량치는 디스크 드라이브(20)의 듀티 사이클에 관련된 것이고, 듀티 사이클은 디스크 드라이브(20)의 사용량의 함수이며, 본질적으로 디스크 드라이브(20)에 저장된 데이터의 양이나 디스크 드라이브(20)의 기억 용량(전체 용량 또는 잔류 용량)이 아니라는 점에 주목하여야 한다.

이러한 본 발명의 실시예의 사용 결과로서, 디스크 드라이브(20)를 조금만 사용하는(즉, 적은 듀티 사이클 또는 이용율을 가진) 사용자는 디스크 드라이브(20)를 더 많이 사용하는(즉, 큰 듀티 사이클 또는 이용율을 가진) 사용자와는 다른 금액이 청구될 수 있다. 만일 디스크 시크 동작(및 시크 거리)이 유지보수 제공자(24)에 의해 수신된 디스크 사용량 측량치의 일부를 형성하면, 다른 사용자들은 그들이 가장 공통적으로 사용하는 디스크 I/O 동작의 유형에 따라서 다르게 청구될 수 있다. 예를 들면, 일부 데이터 베이스 애플리케이션에서와 같이 엣지에서 엣지까지(edge-to-edge)의 디스크 시크를 많이 하는 디스크 동작을 필요로 하는 사용자는 (아카이빙 동작에서와 같이) 더 순차적인 읽기 및/또는 쓰기를 수행하는 사용자와는 다르게 청구될 수 있다. 동일한 관점에서, 동일한 사용자는 사용자가 최근에 시스템(10)을 사용한 유형에 따라서 다른 때에 다르게 청구될 수도 있다. 예를 들어서 RAID 유형의 디스크 드라이브 시스템에서, RAID 5 시스템은 패리티 계산을 위해 읽기-수정하기-쓰기 동작을 사용하고, 읽기-수정하기-쓰기 계산은 각각의 쓰기 동작에 대하여 수행된다. 따라서, 이 유형의 동작은 데이터 프로세서(24A)에 의해 계산된 이용율의 값에 직접적으로 영향을 줄 수 있고, 또한 다른 사용자들 간에는 다를 수 있다.

감시 장치(22)의 사용은 선택적일 수 있고, 일정한 사용자에게는 감시 장치(22)를 설치하여 사용하는 것에 대하여 예를 들면 디스카운트를 적용하는 것과 같은 인센티브가 제공될 수 있는 점이 지적된다. 또한, 감시 장치(22)를 사용하면, 유지보수 제공자(24)가 디스크 드라이브(20)에 저장된 데이터의 양 대신에, 사용자의 실제 사용량에 기초하여 디스크 드라이브(20)의 MTBF와 직접 상관될 수 있는 유지보수비를 정해진 사용자에게 청구할 수 있기 때문에, 유지보수 제공자(24)에게 또한 장점이 있다는 점이 지적된다. 이 유형의 주문형, 동적 유지보수비 결정은 고성능의 기업 레벨 디스크 드라이브 기반 기억 시스템을 구성하기 위해 종래에 사용되었던 다른 유형의 디스크 드라이브보다 열등한 MTBF 특성을 적어도 현재 전형적으로 나타내는 ATA 및 SATA 유형의 디스크 드라이브를 사용할 때 특히 중요할 수 있다. 그러나, 본 발명의 교시는 임의 유형의 디스크 드라이브 및 디스크 드라이브 기술을 사용하는 데이터 기억 시스템뿐만 아니라, 기억 장치 신뢰도가 테이프 드라이브 기반 시스템에서와 같이 발생되는 다른 유형의 데이터 기억 시스템과 함께 사용될 수 있다.

일반적으로, 본 발명의 이 실시예는 적어도 하나의 데이터 기억 장치에서 동작할 수 있지만, 바람직하게는 디스크 장치와 같은 데이터 기억 장치의 집합에서 동작하고, 데이터 기억 장치 듀티 사이클을 결정하기 위해 데이터 기억 시스템의 동작 중에 적어도 하나의 데이터 기억 장치를 감시하는 동작, 및 적어도 부분적으로는 상기 결정된 데이터 기억 장치 듀티 사이클에 기초하여 유지보수비의 현재 값을 결정하는 동작을 포함한다. 이러한 각종 동작은 메모리 또는 메모리들에 저장되어(더 일반적으로는 고정식 또는 이동식 디스크를 포함한 신호 운반 매체에 저장되어) 시스템(10)의 일부를 포함하거나 시스템(10)에 결합된 하나 이상의 데이터 프로세서에 의해 실행되는 컴퓨터 코드 또는 프로그램 명령어에 의해 수행될 수 있다.

본 발명의 추가의 태양은 감시 장치(22)로부터 RAID 제어기(20A)까지 그려진 점선 신호 라인(23)으로 도 1 및 도 2에 도시되어 있다. 디스크 드라이브의 듀티 사이클이 증가함에 따라 디스크 드라이브 고장의 가능성이 높아진다. 도 5를 또한 참조하면, 드라이브 사용량 데이터 및 감시 장치(20)에 의해 수집된 추세를 감시함으로써(도 5의 블록 5A), 특정 듀티 사이클 임계치에서 감시 장치(22)는 RAID 레벨 변경이 행하여지는 것을(예를 들면, RAID 5로부터 RAID 10으로) 자동으로 지시하도록(권유하도록) RAID 제어기(20A)에 대한 신호 라인(23) 상에 RAID 레벨 변경 신호를 어서트할 수 있다. 일반적으로, RAID 레벨 변경 신호는 현재의 RAID 레벨보다 단일 또는 다중 디스크 드라이브 고장의 허용도가 더 큰 RAID 레벨로의 변경을 야기 또는 권유하도록 어서트된다. 환경에 따라서, RAID 제어기(20A)는 현재의 RAID 레벨로부터 다른 RAID 레벨로(예를 들면, RAID 5로부터 RAID 10으로) 온더플라이 이주(on-the-fly migration)를 시작함으로써 응답할 수 있다. 대안적으로, RAID 제어기(20A)[또는 감시 장치(22)]는 RAID 레벨 변경 신호의 수신을 사용자에게 통지할 수 있고, 이것에 의해 일부 경우 하드웨어 업그레이드 및/또는 재구성을 요구하는 권유를 승인 또는 거부하기 위한 기회를 사용자에게 제공한다. 대안적으로, RAID 레벨 변경 신호는 유지보수 제공자(24)의 데이터 프로세서(24A)에 의해 발생되어 감시 장치(22)를 통해 RAID 제어기(20A)에 통지될 수 있고, 또는 적당한 사용자 인터페이스를 통해 사용자에게 직접 통지될 수도 있다. 만일 승인되면, 디스크 드라이브(20)의 RAID 레벨은 디스크 드라이브 고장 허용도가 더 큰 것으로 알려진 RAID 레벨로 변경된다(도 5의 단계 5C).

프로세스는 또한 역으로 작용할 수 있고, 만일 디스크 사용량이 어떤 시간 주기 동안에 동일하거나 다른 임계치 아래로 떨어지면, RAID 레벨 변경 신호(또는 다른 신호)의 어서트는 더 낮은 RAID 레벨로의 변경이 가능하다는 것을 표시할 수 있다. 이 경우, 사용자는 더 낮은 RAID 레벨을 동작시킴으로써 디스크 드라이브(20)의 전체 기억 용량을 증가시킬 수 있고, 단일 드라이브 또는 다중 드라이브 고장을 경험할 가능성은 디스크 사용량의 감소에 의해 감소되는 것으로 추정된다.

일반적으로, 본 발명의 이 실시예는 특정 디스크 드라이브 용장 구성을 가진 디스크 드라이브의 집합에서 동작하고, 데이터 기억 시스템(10)의 동작 중에 디스크 드라이브(20)를 감시하여 듀티 사이클을 결정하는 동작; 결정된 듀티 사이클을 임계값과 비교하는 동작; 및 비교 결과에 기초하여 용장 구성 변경 신호를 어서트하는 동작을 수행한다. 양호한 실시예에서,

용장 구성은 임의의 주어진 시간에 비교 결과에 기초하여 변경 또는 적어도 검토를 받는 특정 RAID 레벨을 가진 RAID 구성이다. 앞에서와 같이, 이들 각종 동작은 메모리 또는 메모리들에 저장되어(더 일반적으로는 고정식 또는 이동식 디스크를 포함한 신호 운반 매체에 저장되어) 시스템(10)의 일부를 포함하거나 시스템(10)에 결합된 하나 이상의 데이터 프로세서에 의해 실행되는 컴퓨터 코드 또는 프로그램 명령어에 의해 수행될 수 있다.

RAID 레벨 변경(도 5)과 관련된 이들 실시예는 유지보수비의 동적 결정(도 3)과 관련된 실시예와 함께 사용될 수 있고, 또는 이들 각종 실시예들의 어느 하나는 다른 것과 독립적으로 사용될 수 있다. 일 예로서, 도 5의 논리 흐름도에 표시된 기능성은 디스크 드라이브의 듀티 사이클을 결정하기 위해 감시 장치(22)의 기능성과 함께 RAID 제어기(20A)에 통합될 수 있다. 본 발명의 이 대안적 실시예에서, 도 2에 표시된 유지보수 제공자(24)의 기능성은 제공될 수도 있고 제공되지 않을 수도 있다.

전술한 설명은 비제한적인 예로서 제공된 것이고, 본 발명을 실시하기 위해 발명자가 현재 예상하고 있는 최상의 방법 및 장치의 완전하고 유익한 설명을 제공한 것이다. 그러나, 당업자라면, 첨부 도면 및 청구범위와 함께 읽을 때, 위의 설명에 비추어서 각종 수정 및 변형이 가능할 것이다. 일부 예로서, 다른 유사한 또는 등가 유형의 디스크 드라이브, 디스크 드라이브 구성, 데이터 기억 시스템 아키텍쳐 및 데이터 기억 장치를 사용하는 것이 당업자에 의해 시도될 수 있다. 그러나, 그러한 및 유사한 본 발명의 모든 수정예는 본 발명의 실시예의 범위에 포함된다.

또한, 본 발명의 일부 특징들은 대응하는 다른 특징들의 사용없이 이익을 위해 사용될 수 있다. 그래서, 전술한 설명은 본 발명의 원리, 교시 및 실시예를 단순히 설명하는 것으로 이해하여야 하며, 본 발명을 제한하는 것으로 해석되어서는 안된다.

도면의 간단한 설명

도 1은 본 발명의 실시예에 따른 디스크 사용량 감시 장치를 가진 디스크 기반 데이터 기억 시스템의 블록도이다.

도 2는 도 1에 도시된 유지보수 제공자의 간단한 블록도이다.

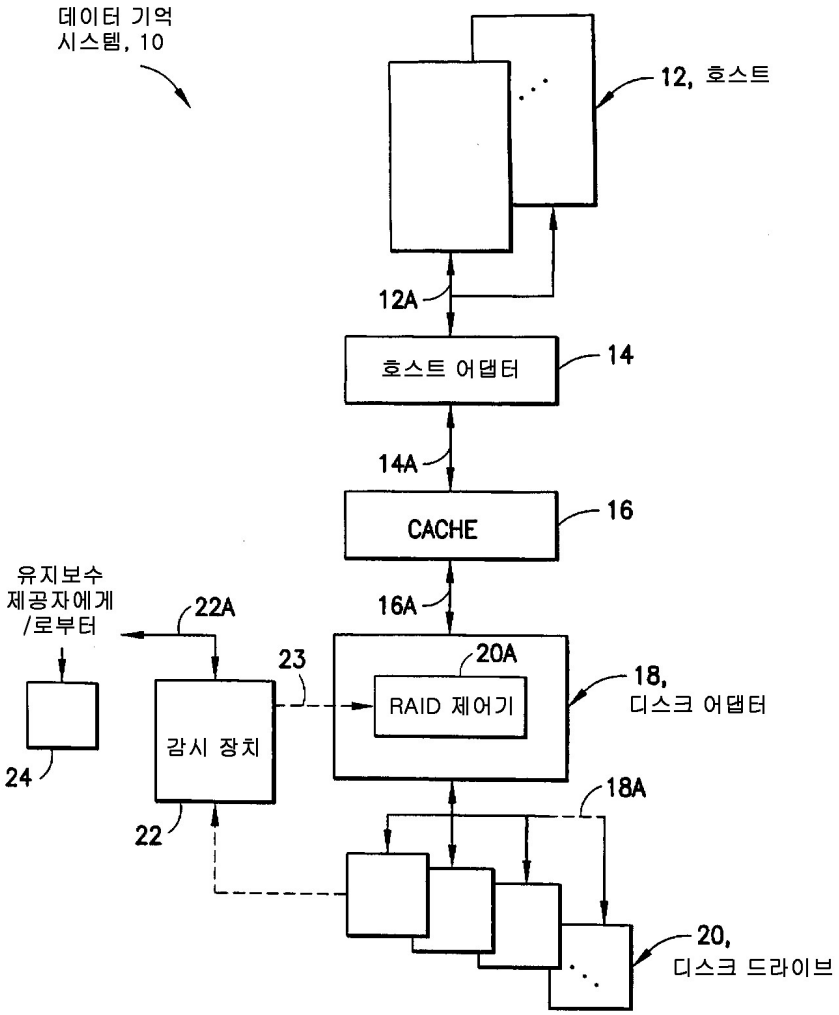
도 3은 유지보수 제공자(24)의 동작을 설명하기 위한 논리 흐름도이다.

도 4는 도 1의 디스크 사용량 감시 장치의 분산된 실시예를 나타낸 간단한 블록도이다.

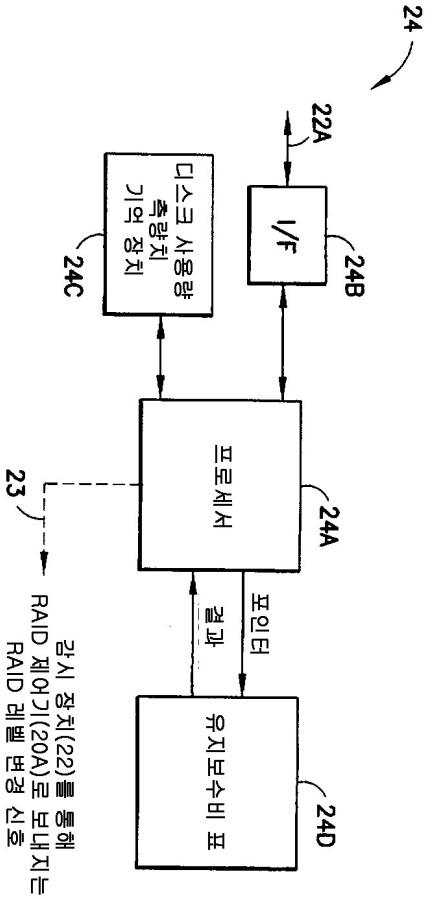
도 5는 본 발명의 실시예에 따른 RAID 레벨 변경 함수의 동작을 설명하기 위한 논리 흐름도이다.

도면

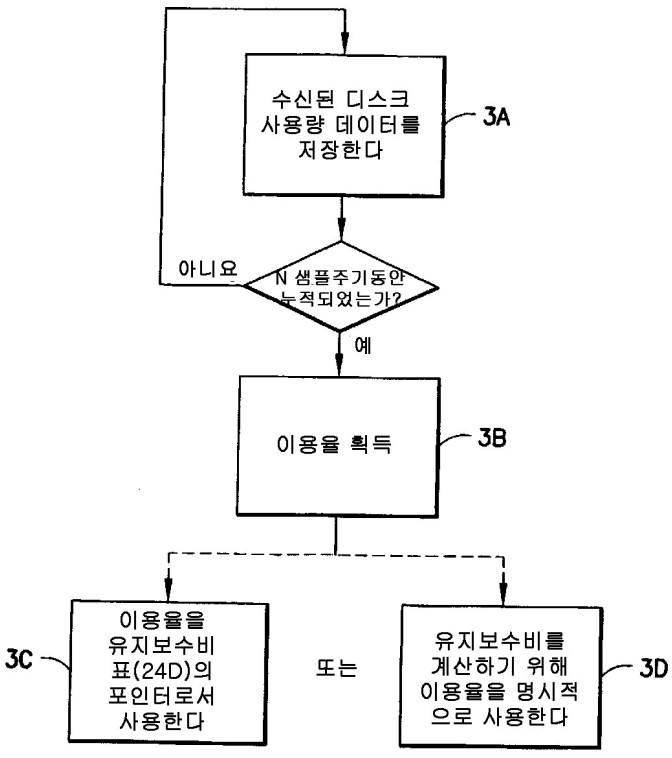
도면1


데이터 기억
시스템, 10
12, 호스트
12A
호스트 어댑터
14
14A
CACHE
16
16A
유지보수
제공자에게
/로부터
22A
23
20A
RAID 제어기
18,
디스크 어댑터
24
감시 장치
22
18A
20,
디스크 드라이브

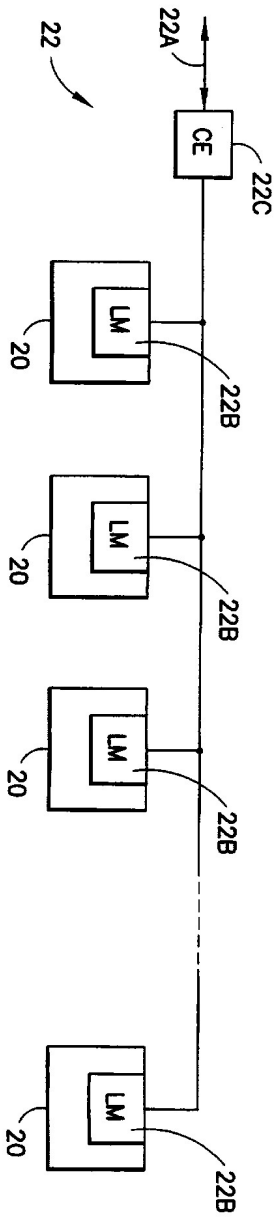
도면2


24
22A
I/F
24B
디스크 사용량 측량치 기억 장치
24C
프로세서
24A
23
포인터
결과
유지보수비 표
24D
감시 장치(22)를 통해 RAID 제어기(20A)로 보내지는 RAID 레벨 변경 신호

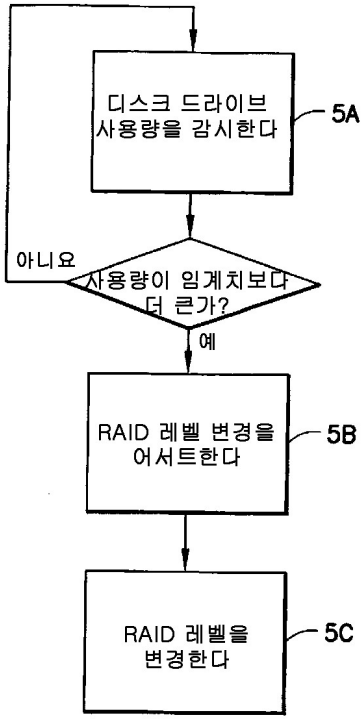
도면3


수신된 디스크 사용량 데이터를 저장한다
3A
아니요
N 샘플주기동안 누적되었는가?
예
이용율 획득
3B
이용율을 유지보수비 표(24D)의 포인터로서 사용한다
3C
또는
유지보수비를 계산하기 위해 이용율을 명시적으로 사용한다
3D

도면4



도면5


디스크 드라이브 사용량을 감시한다
5A
아니요
사용량이 임계치보다 더 큰가?
예
RAID 레벨 변경을 어서트한다
5B
RAID 레벨을 변경한다
5C