



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2015년09월15일

(11) 등록번호 10-1553028

(24) 등록일자 2015년09월08일

(51) 국제특허분류(Int. Cl.)

G06F 12/06 (2006.01)

(21) 출원번호 10-2014-0107590

(22) 출원일자 2014년08월19일

심사청구일자 2014년08월19일

(56) 선행기술조사문헌

[논문] 유영준 외 3명. "비율 기반 대표 해시 기법을 이용한 파일 유사도 평가 시스템". 한국정보기술학회논문지 12(04) pp.81-88, 2014.4.*

[카탈로그] "데이터 중복 제거 기술에 대한 배경 지식 : 기술백서". Quantum. 2014.7.*

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

한림대학교 산학협력단

강원도 춘천시 한림대학길 1, 한림대학교(옥천동)

(72) 발명자

고영웅

강원도 춘천시 춘주로 174, 107동 1501호 (퇴계동, 그린타운아파트)

(74) 대리인

특허법인세신

전체 청구항 수 : 총 6 항

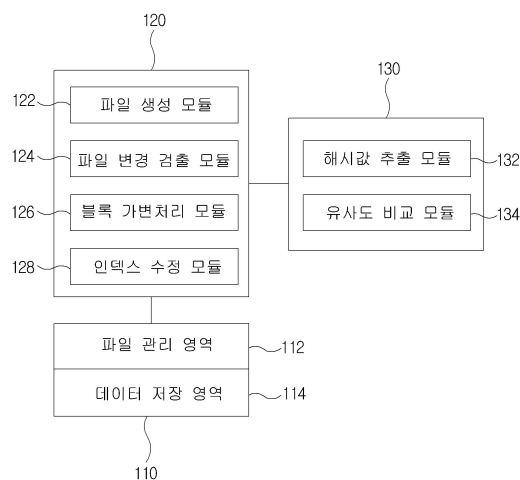
심사관 : 임정묵

(54) 발명의 명칭 고정 블록에 기반을 둔 가변 처리 파일 시스템 및 파일 가변 처리 방법

(57) 요약

본 발명은 고정 블록에 기반을 둔 가변 처리 파일 시스템 및 파일 가변 처리 방법에 관한 것으로, 본 발명에 따른 고정 블록에 기반을 둔 가변 처리 파일 시스템은 파일 데이터를 저장하기 위한 불휘발성 메모리부와, 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 복수의 고정 블록들로 파일 데이터를 저장하며, 불휘발성 메모리부에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터가 변경되는 경우에는 변경된 블록을 가변 블록으로 처리하여 저장하는 파일 관리부를 포함함으로써, 파일 데이터의 변경에 따른 파일 저장을 빠르게 수행하게 할 수 있다.

대표도 - 도1



명세서

청구범위

청구항 1

파일 데이터를 저장하기 위한 불휘발성 메모리부와, 및

상기 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 복수의 고정 블록들로 파일 데이터를 저장하며, 상기 불휘발성 메모리부에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터가 변경되는 경우에는 변경된 블록을 가변 블록으로 처리하여 저장하는 파일 관리부를 포함하고,

상기 불휘발성 메모리부는 파일 관리 영역과 데이터 저장 영역을 포함하고,

상기 파일 관리부는 상기 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 상기 파일 관리 영역에 고정 블록 정보를 저장하고 상기 고정 블록 정보에 할당된 데이터 저장 영역에 파일 데이터를 저장하며, 상기 불휘발성 메모리부에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터가 변경되는 경우에는 상기 파일 관리 영역에 변경된 블록에 대한 블록 가변 정보를 더 저장하고 변경된 블록에 대해서만 해당하는 파일 데이터를 저장하는 것을 특징으로 하는 고정 블록에 기반을 둔 가변 처리 파일 시스템.

청구항 2

삭제

청구항 3

제1항에 있어서,

상기 고정 블록 정보는 각 블록에 대한 상기 데이터 저장 영역의 위치 정보를 포함하며,

상기 블록 가변 정보는 변경된 블록의 개수와 변경된 블록에 저장되는 파일 데이터의 바이트수를 포함하는 것을 특징으로 하는 고정 블록에 기반을 둔 가변 처리 파일 시스템.

청구항 4

제1항 또는 제3항에 있어서,

상기 파일 관리부에 의해 상기 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 저장될 하나의 파일 데이터를 블록 크기보다 작은 소정 단위의 크기로 구분하여 복수의 해시값들을 추출하고, 추출된 복수의 해시값들과 상기 불휘발성 메모리부에 저장된 파일 데이터들과 관련된 해시값들을 이용하여 파일 유사도를 측정하는 파일 유사도 처리부를 더 포함하는 것을 특징으로 고정 블록에 기반을 둔 가변 처리 파일 시스템.

청구항 5

제4항에 있어서,

상기 파일 관리부는 상기 파일 유사도 처리부에서 측정된 파일 유사도를 이용하여 중복 블록들에 대해서는 인덱스를 생성하고 바이트-시프트가 발생하지 않도록 중복 블록들의 앞부분에 존재하는 일부 파일 데이터에 대해서는 블록 가변으로 처리하는 것을 특징으로 하는 고정 블록에 기반을 둔 가변 처리 파일 시스템.

청구항 6

불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 복수의 고정 블록들로 파일 데이터를 저장하는 단계와, 및

상기 불휘발성 메모리부에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터가 변경되는 경우에는 변경된 블록을 가변 블록으로 처리하여 저장하는 단계를 포함하고,

상기 불휘발성 메모리부는 파일 관리 영역과 데이터 저장 영역을 포함하고,

상기 복수의 고정 블록들로 파일 데이터를 저장하는 단계는 상기 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 상기 파일 관리 영역에 고정 블록 정보를 저장하고 상기 고정 블록 정보에 할당된 데이터 저장 영역에 파일 데이터를 저장하고,

상기 변경된 블록을 가변 블록으로 처리하여 저장하는 단계는 상기 불휘발성 메모리부에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터가 변경되는 경우에는 상기 파일 관리 영역에 변경된 블록에 대한 블록 가변 정보를 더 저장하고 변경된 블록에 대해서만 해당하는 파일 데이터를 저장하는 것을 특징으로 하는 고정 블록에 기반을 둔 파일 가변 처리 방법.

청구항 7

삭제

청구항 8

제6항에 있어서,

상기 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 저장될 하나의 파일 데이터를 블록 크기보다 작은 소정 단위의 크기로 구분하여 복수의 해시값들을 추출하는 단계와,

상기 복수의 해시값들을 추출하는 단계에서 추출된 복수의 해시값들과 상기 불휘발성 메모리부에 저장된 파일 데이터들과 관련된 해시값들을 이용하여 파일 유사도를 측정하는 단계와, 및

상기 파일 유사도를 측정하는 단계에서 측정된 파일 유사도를 이용하여 중복 블록들에 대해서는 인덱스를 생성하고 바이트-시프트가 발생하지 않도록 중복 블록들의 앞부분에 존재하는 일부 파일 데이터에 대해서는 블록 가변으로 처리하는 단계를 더 포함하는 것을 특징으로 하는 고정 블록에 기반을 둔 파일 가변 처리 방법.

발명의 설명

기술 분야

[0001]

본 발명은 고정 블록에 기반을 둔 가변 처리 파일 시스템 및 파일 가변 처리 방법에 관한 것으로, 고정 크기의 블록에다 가변 크기의 파일 데이터의 저장을 가능케 하는 고정 블록에 기반을 둔 가변 처리 파일 시스템 및 파일 가변 처리 방법에 관한 것이다.

배경 기술

[0002]

파일 시스템은 컴퓨터 시스템에서 운영 체제, 다양한 실행 파일 그리고 사용자의 데이터를 저장하는 역할을 담당하고 있다. 사용자 영역에서 파일은 바이트 스트림(byte stream)으로 간주되며 파일 핸들을 통해 파일 데이터를 읽거나 쓰는 작업을 수행하게 된다.

[0003]

사용자 영역에서 파일에 대한 읽기/쓰기 요청은 커널 영역의 파일 시스템으로 전달되며 최종적으로는 블록 디바이스에 저장된 데이터 블록에 대한 읽기/쓰기 요청으로 변환한다. 따라서 사용자 영역에서 처리되는 바이트 스트림이 디바이스 수준에서는 블록 단위로 변환되므로 파일에 대한 오퍼레이션에 있어서 다양한 문제점을 유발하고 있다.

[0004]

예를 들어, 파일의 일부 내용을 사용자 영역의 메모리 버퍼로 읽어온 후에, 버퍼의 중간에 한 바이트의 데이터를 추가하거나 삭제하는 작업을 수행하는 경우를 설명한다. 이 경우 사용자 영역에서는 변경하고자 하는 영역에 파일 오프셋을 이동한 후 새로운 바이트를 삽입하거나 삭제하면서 데이터 변경을 쉽게 처리할 수 있다. 하지만, 이와 같은 변경 사항을 파일 시스템에 저장하는 경우에는 새로운 문제를 야기하게 된다. 메모리 버퍼 상에 존재하는 바이트 스트림은 최종적으로 디스크의 블록에 저장되어야 하는데, 한 바이트의 데이터 변경은 연속되는 뒷부분의 모든 데이터 블록의 변경을 요구하게 된다.

[0005]

이러한 문제는 바이트 단위의 삽입 또는 삭제뿐만 아니라 블록 단위로 데이터의 삽입 또는 삭제가 발생이 되는 경우에도 동일하게 발생한다. 예를 들어, 데스크톱 컴퓨터에서 널리 사용되고 있는 Winzip이나 WinRAR와 같은 응용 프로그램은 다수의 서브 파일(sub-file)이 하나의 파일을 구성하는 구조를 가지고 있으며, 각 서브 파일은 블록 경계(block-aligned) 구조를 가진다. 따라서 서브 파일을 삭제하거나 삽입하는 과정은 연속된 블록을 삽입하거나 삭제하는 작업에 해당된다. 하지만, 대부분의 파일 시스템은 이와 같은 블록 단위의 데이터 삽입 또는

삭제에 대한 연산을 고려하지 않고 일반 파일과 동일하게 파일 수정 작업을 처리한다.

발명의 내용

해결하려는 과제

[0006] 상술한 문제점을 해결하기 위해, 본 발명은 고정 크기의 블록에다 가변 크기의 파일 데이터의 저장을 가능케 하는 고정 블록에 기반을 둔 가변 처리 파일 시스템 및 파일 가변 처리 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0007] 상술한 목적을 달성하기 위해, 본 발명의 일실시예에 따른 고정 블록에 기반을 둔 가변 처리 파일 시스템은 파일 데이터를 저장하기 위한 불휘발성 메모리부와, 상기 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 복수의 고정 블록들로 파일 데이터를 저장하며, 상기 불휘발성 메모리부에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터가 변경되는 경우에는 변경된 블록을 가변 블록으로 처리하여 저장하는 파일 관리부를 제공한다.

[0008] 상기 불휘발성 메모리부는 파일 관리 영역과 데이터 저장 영역을 포함하고, 상기 파일 관리부는 상기 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 상기 파일 관리 영역에 고정 블록 정보를 저장하고 상기 고정 블록 정보에 할당된 데이터 저장 영역에 파일 데이터를 저장하며, 상기 불휘발성 메모리부에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터가 변경되는 경우에는 상기 파일 관리 영역에 변경된 블록에 대한 블록 가변 정보를 더 저장하고 변경된 블록에 대해서만 해당하는 파일 데이터를 저장할 수 있다.

[0009] 상기 고정 블록 정보는 각 블록에 대한 상기 데이터 저장 영역의 위치 정보를 포함하며, 상기 블록 가변 정보는 변경된 블록의 개수와 변경된 블록에 저장되는 파일 데이터의 바이트수를 포함할 수 있다.

[0010] 상기 고정 블록에 기반을 둔 가변 처리 파일 시스템은 상기 파일 관리부에 의해 상기 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 저장된 하나의 파일 데이터를 블록 크기보다 작은 소정 단위의 크기로 구분하여 복수의 해시값들을 추출하고, 추출된 복수의 해시값들과 상기 불휘발성 메모리부에 저장된 파일 데이터들과 관련된 해시값들을 이용하여 파일 유사도를 측정하는 파일 유사도 처리부를 더 포함할 수 있다.

[0011] 상기 파일 관리부는 상기 파일 유사도 처리부에서 측정된 파일 유사도를 이용하여 중복 블록들에 대해서는 인덱스를 생성하고 바이트-시프트가 발생하지 않도록 중복 블록들의 앞부분에 존재하는 일부 파일 데이터에 대해서는 블록 가변으로 처리할 수 있다.

[0012] 본 발명의 다른 실시예에 따른 고정 블록에 기반을 둔 파일 가변 처리 방법은 불휘발성 메모리부에 최초로 파일을 저장하는 경우에는 복수의 고정 블록들로 파일 데이터를 저장하는 단계와, 상기 불휘발성 메모리부에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터가 변경되는 경우에는 변경된 블록을 가변 블록으로 처리하여 저장하는 단계를 제공함으로써, 상술한 목적을 달성할 수 있다.

발명의 효과

[0013] 상술한 구성에 의해, 본 발명은 고정 크기의 블록에다 가변 크기의 파일 데이터의 저장을 가능케 함으로써, 파일 데이터의 변경에 따른 파일 저장을 빠르게 수행하게 할 수 있다.

[0014] 또한, 본 발명은 저장 매체의 데이터 영역에 대한 기록을 최소화함으로써 반도체 타입의 불휘발성 메모리인 플래시 메모리 등의 수명을 연장할 수 있다.

도면의 간단한 설명

[0015] 도 1은 본 발명의 일실시예에 따른 고정 블록에 기반을 둔 가변 처리 파일 시스템의 전체적인 블록도를 도시하는 도면이다.

도 2a 내지 도 2c는 본 발명에 따른 고정 블록에 기반을 둔 가변 처리를 설명하기 위한 도면이다.

도 3은 본 발명에 따른 메타데이터의 구현예를 도시하는 도면이다.

도 4a 및 도 4b는 본 발명에 따른 고정 블록에 기반을 둔 가변 처리 파일 시스템에서의 중복 데이터 처리를 설명하기 위한 도면이다.

도 5는 도 1에 도시된 파일 유사도 처리부를 좀 더 상세하게 도시하는 도면이다.

도 6은 본 발명의 다른 실시예에 따른 고정 블록에 기반을 둔 파일 가변 처리 방법의 흐름도를 도시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0016] 이하, 첨부된 도면을 참조하여 본 발명에 따른 고정 블록에 기반을 둔 가변 처리 파일 시스템 및 파일 가변 처리 방법의 바람직한 실시예를 설명한다. 참고로, 아래에서 본 발명을 설명함에 있어서, 본 발명의 구성요소를 지칭하는 용어들은 각각의 구성 요소들의 기능을 고려하여 명명된 것이므로, 본 발명의 기술적 구성요소를 한정하는 의미로 이해되어서는 안 될 것이다.
- [0017] 도 1은 본 발명의 일실시예에 따른 고정 블록에 기반을 둔 가변 처리 파일 시스템의 전체적인 블록도를 도시하는 도면이다.
- [0018] 도 1에 도시된 바와 같이, 고정 블록에 기반을 둔 가변 처리 파일 시스템은 불휘발성 메모리부(110), 파일 관리부(120) 및 파일 유사도 처리부(130)를 포함한다.
- [0019] 불휘발성 메모리부(110)에는 파일 데이터를 저장하기 위해 파일 관리 영역(112)과 데이터 저장 영역(114)이 구분되어 있다. 여기서 불휘발성 메모리부(110)는 반도체 메모리인 플래시 메모리로 한정하기 위해 사용한 것이 아니며 하드 디스크 등을 포함한다.
- [0020] 파일 관리 영역(112)에는 고정 길이의 블록 정보와 관련된 고정 블록 정보와 고정 길이의 블록에 가변 길이의 정보를 저장하기 위한 블록 가변 정보가 저장되며, 데이터 저장 영역(114)에는 파일 관리 영역(112)의 고정 블록 정보와 블록 가변 정보에 따라 파일 데이터가 저장된다. 한편, 고정 블록 정보에는 각 블록에 대한 데이터 저장 영역(114)의 위치 정보가 포함되며, 블록 가변 정보에는 변경된 블록의 개수와 변경된 블록에 저장되는 파일 데이터의 바이트수가 포함될 수 있다.
- [0021] 파일 관리부(120)는 불휘발성 메모리부(110)에 저장되는 파일 데이터를 관리하기 위해, 파일 생성 모듈(122), 파일 변경 검출 모듈(124), 블록 가변 처리 모듈(126) 및 인덱스 수정 모듈(128)을 포함한다.
- [0022] 파일 생성 모듈(122)은 불휘발성 메모리부에 최초로 파일을 저장하는 경우 복수의 고정 블록들로 파일 데이터를 저장한다. 또한, 파일 생성 모듈(122)은 불휘발성 메모리부에 최초로 파일을 저장하는 경우 파일 관리 영역(112)에 고정 블록 정보를 저장하고 고정 블록 정보에 할당된 데이터 저장 영역(114)에 파일 데이터를 저장할 수 있다.
- [0023] 파일 변경 검출 모듈(124)은 불휘발성 메모리부(110)에 저장된 파일 데이터가 독출된 후에 해당 파일 데이터에 변경이 있는지를 검출한다. 예를 들어 파일 변경 검출 모듈(124)은 사용자가 파일 데이터를 읽은 후 해당 파일 데이터의 소정 블록에 해당하는 바이트에 대한 삭제 등을 수행하였는지를 검출한다.
- [0024] 블록 가변 처리 모듈(126)은 파일 변경 검출 모듈(124)이 불휘발성 메모리부(110)에 저장된 복수의 고정 블록들 중 적어도 하나의 블록에 저장된 파일 데이터의 변경을 검출하면, 저장시 변경된 블록을 가변 블록으로 처리하여 저장한다. 또한, 블록 가변 처리 모듈(126)은 저장시 파일 관리 영역(112)에 변경된 블록에 대한 블록 가변 정보를 더 저장하고 변경된 블록에 대해서만 해당하는 파일 데이터를 저장할 수 있다. 또한 블록 가변 처리 모듈(126)은 인덱스 수정 모듈(128)에서 중복 블록들에 대해 인덱스 정보를 변경하고자 하는 경우 중복 블록들의 앞부분에 존재하는 일부 파일 데이터에 대해서는 블록 가변으로 처리할 수 있다.
- [0025] 인덱스 수정 모듈(128)은 파일 유사도 처리부(130)에서 측정된 파일 유사도를 이용하여 파일 데이터가 동일한 중복 블록들에 대해서는 데이터 저장 영역(114)에 해당 중복 블록을 저장하지 않고 파일 관리 영역(112)에다 저장된 중복 블록의 데이터 저장 영역(114)의 위치 정보로 인덱스를 수정하여 저장한다.
- [0026] 파일 유사도 처리부(130)는 불휘발성 메모리부(110)에 최초로 저장될 파일 데이터와 불휘발성 메모리부(110)에 이미 저장되어 있는 파일 데이터들 사이에 유사도를 판단하며, 이를 위해 해시값 추출 모듈(132) 및 유사도 비교 모듈(132)을 포함한다.
- [0027] 해시값 추출 모듈(132)은 파일 관리부(120)가 불휘발성 메모리부(110)에 최초로 파일을 저장하는 경우 저장될 하나의 파일 데이터를 블록 크기보다 작은 소정 단위의 크기로 구분하여 각 소정 단위의 크기마다 해시값을 추출한다.

- [0028] 유사도 비교 모듈(132)은 해시값 추출 모듈(132)에서 추출된 복수의 해시값들과 불휘발성 메모리부(110)에 저장된 파일 데이터들과 관련된 해시값들을 비교하여 동일한 해시값이 있는지를 확인한다.
- [0029] 이하에서는 본 발명의 실시예들을 보다 상세하게 설명한다.
- [0030] 도 2a 내지 도 2c는 본 발명에 따른 고정 블록에 기반을 둔 가변 처리를 설명하기 위한 도면이다.
- [0031] 도 2a에는 종래의 파일 시스템의 개념도가 도시되어 있다. 파일은 바이트 스트림으로 실제 불휘발성 메모리부(110)에 저장될 때에는 청킹의 단계를 거쳐서 저장되게 된다. 이 때, INODE는 파일의 정보와 함께 각 블록의 위치를 기록하는 정보를 가진다. 여기서 각 블록의 고정 크기는 예를 들면 512바이트일 수 있다.
- [0032] 도 2b에는 저장된 파일 데이터를 수정하는 종래의 파일 시스템의 개념도가 도시되어 있다. 도 2b에 도시된 바와 같이, "K"가 삭제되면 청킹 단계에서 3번 블록부터 내부 파일 데이터에 변화가 있기 때문에 불휘발성 메모리부(110)에 3번 블록부터 변경되어 저장되게 된다. 이 경우 파일 시스템에 따라서 기존 블록을 무효화하고 새로운 블록을 할당받아 파일 데이터가 저장될 수도 있고, 기존의 블록을 덮어쓰는 형태로 저장될 수 있다.
- [0033] 도 2c에는 본 발명에 따른 고정 블록에 기반을 둔 가변을 이용하여 파일 데이터를 수정하는 개념도가 도시되어 있다. 도 2c에 도시된 바와 같이, "K"가 삭제되면 청킹 단계에서 3번 블록에 512바이트보다 작은 크기의 파일 데이터 크기를 담게 된다. 그리고 INODE에서는 변경된 내용에 대해서 블록 인덱스를 수정하게 된다. 이와 같이 본 발명에 따라 파일 데이터를 저장하게 되는 경우에는 데이터 블록에서 3번째 블록만 변경이 되고 나머지 이후에 있는 블록에는 변경이 발생하지 않게 된다.
- [0034] 도 3은 본 발명에 따른 메타데이터의 구현예를 도시하는 도면이다.
- [0035] INODE에는 파일 입출력을 위한 기본 정보가 저장되게 된다. 본 발명에서는 기존의 INODE를 확장하여 고정 블록에 기반을 두어 블록의 가변 처리를 지원할 수 있도록 구현한다. 도 3에 도시된 바와 같이, 본 발명에 따른 메타데이터 정보는 기존의 고정 블록에 대한 정보들(B0, B1, B2 및 B3)뿐만 아니라 고정 블록의 가변 처리와 관련하여 2가지 정보가 추가된다. VB(variable block)에는 몇 개의 블록 가변을 가지고 있는지가 기록된다. VB가 1이면, 파일 데이터에서 블록의 가변 처리에 의해 변경이 발생한 블록이 1 블록임을 나타낸다. 그리고 VBI(variable block index)는 블록에 대한 가변 처리가 발생한 블록의 인덱스를 가리키는데, 도 3에서는 V1, V2, V3 및 V4로 도시되어 있다.
- [0036] 도 3에 도시된 바와 같이 3번째 블록에서 파일 데이터의 변경으로 16바이트가 삭제되었다고 가정한다. 이 경우 53번에 위치한 블록에는 496바이트만 저장되고 나머지 16바이트는 의미가 없는 데이터가 저장된다. 이를 파일 시스템에서 참조하기 위해서는 INODE에 해당 정보가 기술되어야 한다. 고정 블록에 대한 가변 처리가 1개 생성되었기 때문에 VB에 1이 저장된다. 그리고 V1에는 53번 블록이 고정 블록에 대한 가변 처리가 있는 블록임을 알려주고 또한 53번 블록에는 496바이트가 기록되어 있음에 대한 정보를 저장한다.
- [0037] 도 3에서는 고정 블록에 대한 가변 처리를 최대 4개로 제한하였지만, 필요에 따라서 최대 개수를 증가시킬 수 있다. 하지만 최대 개수 이상의 고정 블록에 대한 가변 처리가 필요한 경우에는 강제로 가비지 컬렉션을 수행하여 고정 블록에 대한 가변 처리를 모두 제거하는 과정을 거칠 수 있다.
- [0038] 도 4a 및 도 4b는 본 발명에 따른 고정 블록에 기반을 둔 가변 처리 파일 시스템에서의 중복 데이터 처리를 설명하기 위한 도면이다.
- [0039] 도 4a에는 종래의 파일 시스템에서의 중복 데이터 처리 방식이 도시되어 있다. 도 4a에 도시된 바와 같이, 새로운 파일이 생성되는 경우 종래의 파일 시스템에서는 중복된 영역이 4개 블록이나 존재함에도 불구하고 중복을 제거할 수 없다. 왜냐하면 중복 데이터가 존재하지만, 바이트-시프트 현상으로 블록 단위별로 공유가 불가능하기 때문이다.
- [0040] 도 4b에는 본 발명에 따른 고정 블록에 기반을 둔 가변 처리를 이용하여 중복 데이터를 처리하는 방식이 도시되어 있다. 도 4b에 도시된 바와 같이, 본 발명으로 구현된 고정 블록에 기반을 둔 가변 처리 파일 시스템을 사용하는 경우에는 중복 블록을 기준으로 청킹을 수행하여 4개의 블록에 대해서 중복을 찾아낼 수 있으며, 이 중복된 4개의 블록에 대해서는 새롭게 불휘발성 메모리에 저장하지 않고 기존 파일의 블록으로 인덱스를 수정하는 것으로 불휘발성 메모리에의 저장을 대신할 수 있다.
- [0041] 파일 유사도는 파일간의 유사한 정도를 나타내는 수치로, 유사도 정보를 이용하여 효율적으로 중복 데이터를 제거할 수 있다. 중복 제거는 새로운 파일을 저장할 때 다른 파일에서 동일한 부분을 찾아 동일한 부분은 저장하

지 않고 서로 다른 부분에 대해서만 저장을 하는 기법이다. 이에 파일의 유사도 정보는 높은 정확도가 요구된다.

[0042] 도 5는 도 1에 도시된 파일 유사도 처리부를 좀 더 상세하게 도시하는 도면이다.

[0043] 도 5에 도시된 바와 같이, 파일 유사도 처리부(130)는 해시값 추출 모듈 및 유사도 비교 모듈(132)을 포함하며, 파일 데이터의 유사도 처리를 위해 불휘발성 메모리부(110)인 HDD(510)와, 새로운 파일 데이터가 생성되어 저장되어 있는 휘발성 메모리와 관련된 파일 생성 메모리(520)와, 및 파일에 대한 해시값이 저장되어 있는 휘발성 메모리와 관련된 해시값 저장 메모리(530)에 연결되어 있다. 여기서 파일 생성 메모리(520)와 해시값 저장 메모리(530)는 별개의 메모리일 수 있지만 하나의 메모리로 이루어질 수 있다.

[0044] 파일 관리부(120)의 파일 생성 모듈(122)이 파일 생성 메모리(520)에 저장되어 있는 파일을 HDD(510)에 저장하고 하면, 해시값 추출 모듈(132)은 파일 생성 메모리(520)에 저장되어 있는 파일 데이터를 읽어 해시값을 추출한다. 이 경우 해시값 추출 모듈(132)은 파일 데이터를 블록 크기보다 작은 소정 단위의 크기로 구분하여 각 소정 단위의 크기마다 해시값을 추출할 수 있다. 해시값 추출 모듈(132)에서 추출된 해시값들은 해시값 저장 메모리(530)에 저장될 수 있다.

[0045] 유사도 비교 모듈(132)은 해시값 추출 모듈(132)에서 추출된 복수의 해시값들과 해시값 저장 메모리(530)에 저장된 각 파일 데이터와 관련된 해시값들을 비교하여 동일한 해시값이 있는지를 확인한다. 해시값 저장 메모리(530)에 저장된 각 파일 데이터와 관련된 해시값들은 HDD(510)로 읽어 올 수 있다.

[0046] 이러한 파일 유사도 처리를 위해 SDHASH(similarity digest hashing) 기법 또는 SSDEEP 기법을 이용할 수 있다.

[0047] SDHASH는 라빈 핑거프린트(rabin fingerprint)를 이용한 임의의 청크 크기를 선택하는 대신에 새넨 엔트로피를 이용하여 확률적인 특징점(feature)을 선택하는데, 특징점 선택이 데이터의 특정 부분에 치우치지 않고 전체적으로 고른 선택이 가능하여 신뢰성이 향상된다. 선택된 특징점들은 블룸 필터를 이용하여 통계적 기반의 파일 유사도를 측정한다. 64바이트 크기로 된 특징점들은 주어진 임계값 이상의 인기도 점수가 포함된 특징점을 선택하여 블룸 필터 생성에 사용한다. 또한, 선택된 특징점 중에서 100 이하 990 초과 값을 가지는 특징점은 제외하여 블룸 필터의 긍정 오류율을 낮춘다.

[0048] SDHASH에서 블룸 필터는 256바이트 크기로 고정되어 생성된다. 생성된 블룸 필터 1개당 128개의 특징점 값들이 할당된다. 특징점들이 128개를 초과하면 새로운 블룸 필터가 추가되어 특징점 값들이 할당된다. 각 특징점들은 SHA-1 해시 함수를 이용하여 160비트 크기의 해시값으로 계산된다. SDHASH의 비교 과정은 생성된 모든 블룸 필터들을 하나씩 비교하여 전체 블룸 필터 비교 값들의 평균으로 파일 유사도를 측정한다.

[0049] SSDEEP는 2개의 해시 방법을 이용하여 파일 유사도 측정을 수행한다. 롤링 해시(rolling hash) 함수는 파일 내용을 한 바이트씩 이동하며 해시값을 생성한다. 생성된 롤링 해시값을 이용하여 파일 내부의 경계점인 트리거 위치를 계산한다. 트리거 위치가 발견되면 파일의 시작 위치부터 트리거 위치까지 하나의 청크로 처리하고 해당 청크는 MD5 해시 함수를 이용하여 청크 해시값을 생성한다. 생성된 청크 해시값은 Base64 인코딩을 사용하여 비트 스트링으로 변환되고 변환된 비트 스트링의 마지막 6비트 값을 추출한 후 각 비트 스트링의 비트 값을 연결하여 파일 다이제스트를 생성한다.

[0050] 파일 유사도 처리부(130)는 상술한 기법으로 처리된 각 파일의 유사도 값을 INODE에 스트링 형태로 유지할 수 있다. 따라서 파일 유사도 처리부(130)는 새로운 파일이 유입되거나 생성이 되는 경우에, 해시 스트링의 비교를 통하여 가장 유사한 파일을 찾아낼 수 있으며, 파일 관리부(120)는 원본 파일과 유입되는 파일 간의 중복 블록을 찾아내어 공유시키는 작업을 진행할 수 있다.

[0051] 한편, 파일 유사도 처리부(130)는 고정 블록에 기반을 두어 파일 유사도를 고정 분할 기법으로 처리하면 중복 블록의 앞부분에 존재하는 데이터에 대해서는 가변 분할 기법을 적용할 수 있다. 이에 의해 유사 파일에 대해서 효과적인 파일 데이터 저장이 가능하고 입출력 오버헤드를 상당히 감소시킬 수 있다.

[0052] 도 6은 본 발명의 다른 실시예에 따른 고정 블록에 기반을 둔 파일 가변 처리 방법의 흐름도를 도시하는 도면이다.

[0053] 파일 관리부(120)가 파일을 불휘발성 메모리에 저장하고자 하는 경우 파일 생성 모듈(122)은 해당 파일이 불휘발성 메모리에 최초로 저장되는 파일인지를 확인한다(S602). 파일 관리부(120)가 파일을 불휘발성 메모리에 최초로 저장하는 것이라면, 해시값 추출 모듈(132)은 저장하고자 하는 하나의 파일 데이터를 블록 크기보다 작은 소

정 단위의 크기로 구분하여 복수의 해시값들을 추출한다(S604).

[0054] 유사도 비교 모듈(132)은 해시값 추출 모듈(132)에서 추출된 복수의 해시값들과 불휘발성 메모리부(110)에 저장된 파일 데이터들과 관련된 해시값들을 비교하여 파일 유사도를 측정한다(S606). 인덱스 수정 모듈(128)은 유사도 비교 모듈(132)에서 측정된 파일 유사도를 이용하여 중복 블록들에 대해서는 인덱스를 생성한다(S608). 그리고 블록 가변 처리 모듈(126)은 바이트-시프트가 발생하지 않도록 중복 블록들의 앞부분에 존재하는 일부 파일 데이터에 대해서는 블록 가변으로 처리한다(S610).

[0055] 파일 변경 검출 블록은 불휘발성 메모리에 저장하고자 하는 파일에 변경이 있는지를 검출한다(S612). 파일에 변경이 있으면, 블록 가변 처리 모듈(126)은 파일 관리 영역(112)에 변경된 블록에 대한 블록 가변 정보를 더 저장하고(S614) 변경된 블록에 대해서만 해당하는 파일 데이터를 저장한다(S616).

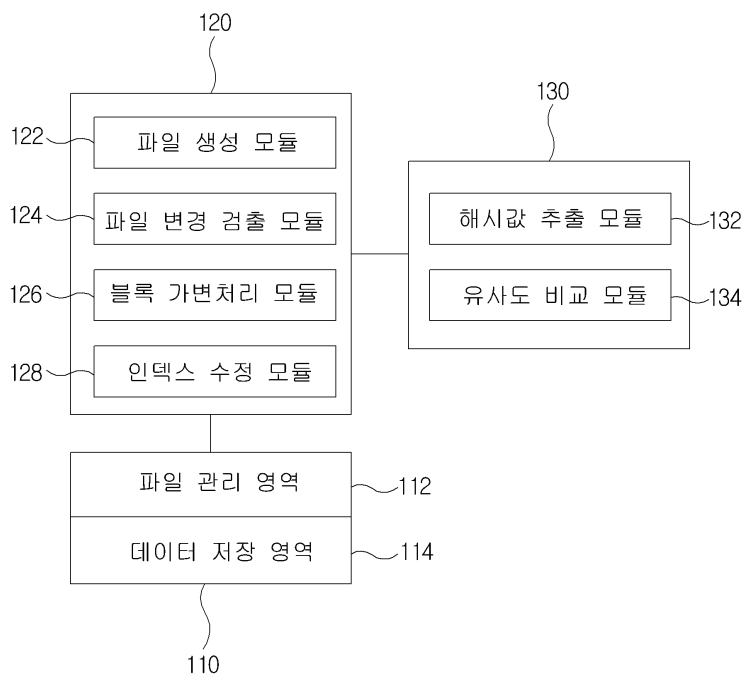
[0056] 본 발명의 보호 범위는 이하 특허청구범위에 의하여 해석되어야 마땅할 것이다. 또한, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것인 바, 본 발명과 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

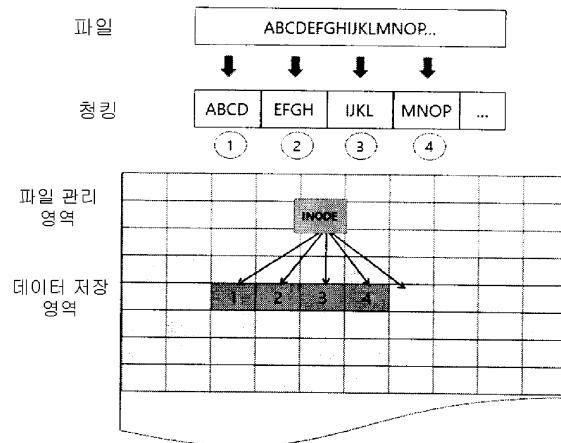
[0057]	110: 불휘발성 메모리부	112: 파일 관리 영역
	114: 데이터 저장 영역	120: 파일 관리부
	122: 파일 생성 모듈	124: 파일 변경 검출 모듈
	126: 블록 가변 처리 모듈	128: 인덱스 수정 모듈
	130: 파일 유사도 처리부	132: 해시값 추출 모듈
	134: 유사도 비교 모듈	510: HDD
	530: 파일 생성 메모리	530: 해시값 저장 메모리

도면

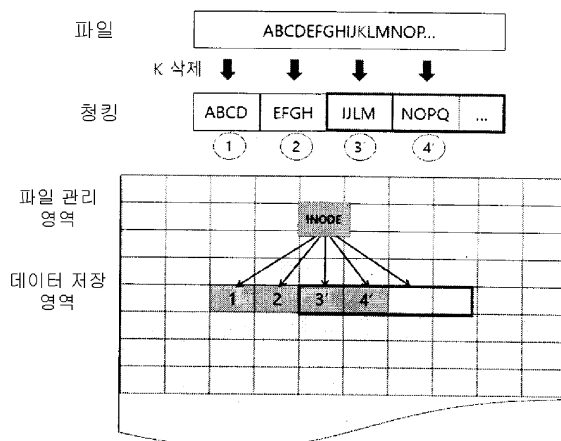
도면1



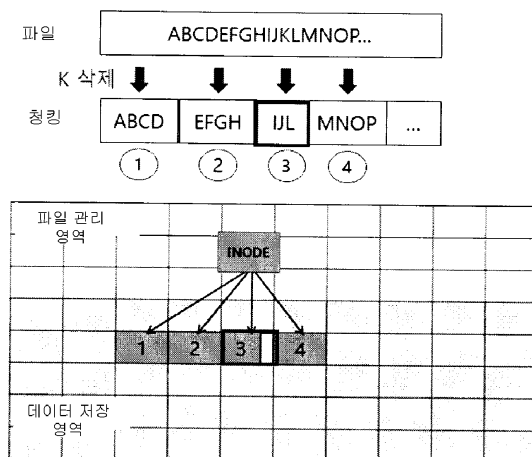
도면2a



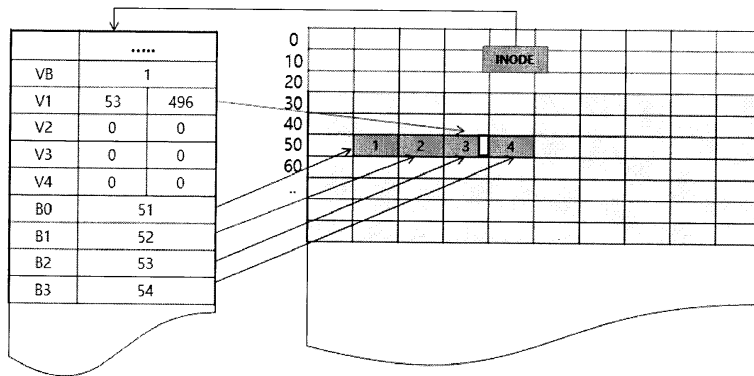
도면2b



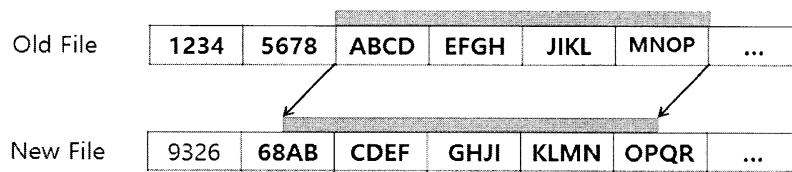
도면2c



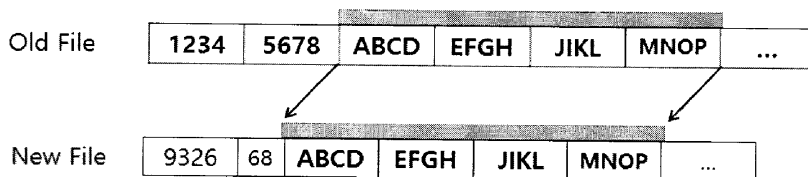
도면3



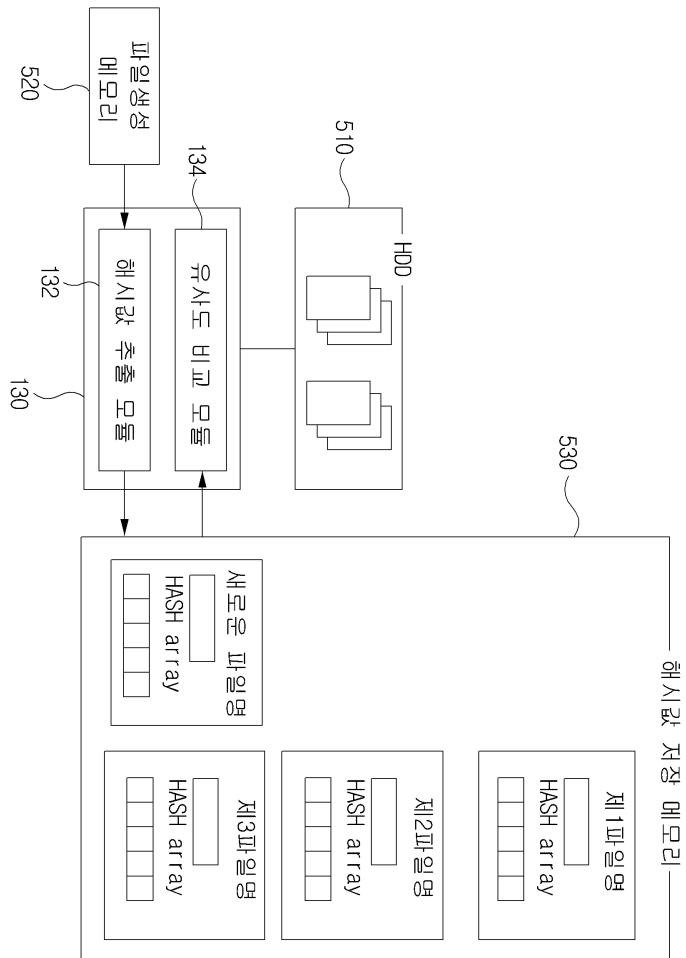
도면4a



도면4b



도면5



도면6

