3차 저장장치를 부착한 MIDAS-Ⅲ에서의 고장 회복*

김수희° 김영성 문찬호 강현철 김 준†
중 앙 대 학 교 컴 퓨 터 공 학 과
†한국전자통신연구원 인터넷서비스 연구부
{shkim, yskim, moonch, hckang}@rose.cse.cau.ac.kr
†june@bada1.etri.re.kr

Recovery in MIDAS-Ⅲ with Tertiary Storage

Soohee Kim° Yungsung Kim Chanho Moon Hyunchul Kang June Kim†
Department of Computer Science and Engineering, Chung-Ang University
†Internet Service Department, Electronics and Telecommunications Research Institute

요 약

MIDAS-Ⅲ는 한국전자통신연구원에서 개발한 바다-DBMS의 저장 시스템이다. MIDAS-Ⅲ는 대용량 멀티미디어 데이타 서버의 저장 시스템으로 기능하도록 3차 저장장치를 부착하여 MIDAS-Ⅲ/TS로 확장 구현되었다. 3차 저장장치를 관리하기 위하여, MIDAS-Ⅲ/TS는 3차 저장장치 관리 볼륨을 새롭게 구성하였다. 본 논문에서는 먼저 MIDAS-Ⅲ/TS의 구조와 확장 전 MIDAS-Ⅲ의 고장 회복 유틸리티인 restart에 대해 살펴본다. 그리고, MIDAS-Ⅲ/TS의 고장 회복 유틸리티인 restart -t의 구현을 위하여, MIDAS-Ⅲ가 3차 저장장치를 부착함에 따라 새롭게 발생할 수 있는 고장의 유형과 고장 회복에 있어서의 문제점에 대해 기술하고 회복 방법을 제시한다.

1. 서론

대용량 멀티미디어 데이타를 저장하기 위해서 기존 DBMS의 보조 기억 장치인 디스크만을 이용하는 데에는 한계가 있다. 이에 성능은 떨어지지만 가격이 싼 3차 저장장치의 도입이 요청되고 있다[1][2][3]. 한국전자통신연구원에서 개발한 바다-DBMS의 저장 시스템인 MIDAS-Ⅲ는 3차 저장장치를 부착하여 대용량 멀티미디어 데이타 서버의 저장 시스템으로 기능하도록 MIDAS-Ⅲ/TS(Tertiary Storage)로 확장되었다. MIDAS-Ⅲ/TS의 3차 저장장치는 광 디스크 쥬크박스 또는 테이프 라이브러리와 같이 하나 또는 그 이상의 저장매체(platter)로 구성되며, 저장매체로는 재기록이 가능한(rewritable) 저장매체와, WORM (write-once-read-many)도 사용할 수 있다. MIDAS-Ⅲ/TS는 3차 저장장치를 관리하기 위하여 기존의 볼륨 이외에 3차 저장장치 관리 볼륨을 두어 3차 저장장치를 효율적으로 관리하도록 하였다.

일반적으로 DBMS에서 고장의 종류에는 크게 트랜잭션 고장(transaction failure), 시스템 고장(system crash), 매체 고장(media failure)이 있는데, MIDAS-Ⅲ에서는 고장 회복을 위해 media, restoredb, restorelog, restart 등의 유틸리티를 제공하며, 이중 시스템 고장의 회복을 위해 restart 유틸리티가 사용된다.

MIDAS-Ⅲ가 3차 저장장치를 부착하여 MIDAS-Ⅲ/TS로 확장되면서 새로운 고장의 유형이 나타날 수 있다. 이에 따라 본 논문에서는, MIDAS-Ⅲ의 고장 회복 유틸리티인 restart를 확장하여 MIDAS-Ⅲ/TS의 고장 회복 유틸리티인 restart -t(t는 tertiary를 의미)를 구현하기 위해, 새로운 고장의 유형을 파악하며 회복 방법을 제시한다.

본 논문의 구성은 다음과 같다. 2절에서는 MIDAS-Ⅲ/TS와 3차 저장장치 관리 볼륨의 구조에 대해서 기술한다. 3절에서는 확장 이전의 MIDAS-Ⅲ의 고장 회복에 대해서 설명한다. 4절에서는 MIDAS-Ⅲ에 3차 저장장치를 부착하면서 새롭게 생겨난 고장의 유형과 고장 회복에 있어서의 문제점을 파악하고, 그에 따른 회복 방법을 제시한다. 5절에서는 결론 및 향후 연구를 기술한다.

* 본 논문은 한국전자통신연구원에서 수행중인 "인터넷 멀티미디어 문서 DBMS 개발" 사업의 위탁 결과 일부입니다.

2. MIDAS-Ⅲ/TS와 3차 저장장치 관리 볼륨

MIDAS-Ⅲ는 데이타를 저장하기 위하여 볼륨을 사용한다. 볼륨은 디스크 파티션의 연속된 디스크 영역 또는 유닉스 화일인 세그먼트들로 구성되는 논리적인 저장 구조이다. 세그먼트는 물리적으로 연속된 페이지들의 집합인 익스텐트로 구성된다. 또한 MIDAS-Ⅲ의 화일은 익스텐트들의 집합으로 구성된다.

MIDAS-Ⅲ의 볼륨으로는, MIDAS-Ⅲ의 유틸리티인 volumeinit을 수행하여 처음 생성되는 볼륨인 루트 볼륨과 createvolume을 수행하여 추가로 생성되는 일반 디스크 볼륨이 있으며, MIDAS-Ⅲ/TS로 확장되면서, 3차 저장장치를 관리하기 위한 3차 저장장치 관리 볼륨이 추가되었다.

MIDAS-Ⅲ/TS의 3차 저장장치는 여러 개의 저장매체로 구성되는데, MIDAS-Ⅲ/TS의 유틸리티인 loadplatter를 이용하여 3차 저장장치로 저장매체의 삽입이나 ejectplatter를 이용하여 3차 저장장치로부터 저장매체의 제거가 가능하다. MIDAS-Ⅲ/TS에서 3차 저장장치의 각 저장매체는 데이타의 붙여쓰기 및 읽기만(append and read-only) 가능하다. 3차 저장장치에 저장된 데이타를 읽기 위해서는 일단 데이타가 속한 익스텐트를 디스크로 캐쉬해 와야 하는데, 이를 위해 3차 저장장치 관리 볼륨의 첫번째 세그먼트를 디스크 캐쉬 영역으로 사용한다. MIDAS-Ⅲ/TS에서 디스크로부터 3차 저장장치로의 데이타 이동은 화일 단위로 이루어지며, 3차 저장장치에서 디스크로의 데이타 캐쉬는 익스텐트 단위로 이루어진다. MIDAS-Ⅲ/TS의 대략적인 구조는 그림 1과 같다.

MIDAS-Ⅲ/TS에서는 화일을 읽기 전용 여부 및 3차 저장장치로의 이동 여부에 따라, 수정 가능 화일, 읽기 전용 화일, 3차 저장장치 화일의 세 가지 타입으로 분류한다. MIDAS-Ⅲ/TS에서 화일을 디스크로부터 3차 저장장치로 이동시키기 위해서는, 먼저 3차 저장장치로 이동시키고자 하는 화일을 읽기 전용으로 표시하여 디스크에서 읽기 전용으로 사용하다가, 사용자가 원하는 시점에 3차 저장장치로 이동

시킬 수 있다. 3차 저장장치 관리 볼륨에는 화일 사상 카탈로그 화일이 있어서, 화일이 3차 저장장치로 이동할 때, 화일 식별자, 디스크에서의 화일의 위치, 3차 저장장치에서의 저장매체 번호, 그리고 저장매체 내에서의 위치 등의 화일 사상 정보를 기록한다. 따라서 원하는 페이지가 속하는 화일이 3차 저장장치로 이동되었을 경우에, 화일 사상 정보를 이용하여, 원하는 페이지를 포함하고 있는 3차 저장장치의 해당 익스텐트를 디스크로 캐쉬해 올 수 있다. 3차 저장장치 관리 볼륨에는 디스크 캐쉬 정보를 관리하는 익스텐트 사상 정보 페이지가 있어서 캐쉬해 온 익스텐트의 정보와 디스크 캐쉬에 저장된 위치 등의 정보를 저장한다.

또한 3차 저장장치 관리 볼륨에는 3차 저장장치의 상태를 저장하고 있는 3차 저장장치 상태 정보 페이지가 있어, 3차 저장장치에 부착되어 있는 저장매체 정보를 관리한다.

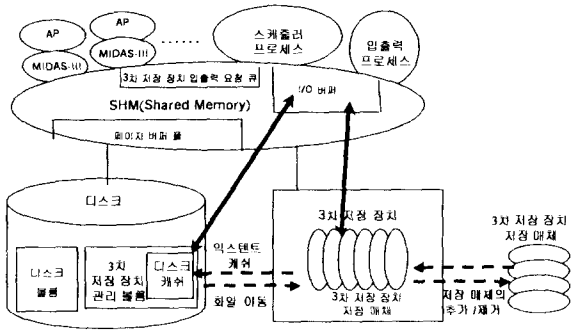

그림 1 MIDAS-III/TS 의 구조

3. MIDAS-III의 고장 회복

MIDAS-III에서 고장 회복은 media, restoredb, restorelog, restart 등의 유틸리티 실행을 통해 수행된다. 이중 restart 유틸리티는 시스템 고장에 대해 고장 회복을 수행한다. MIDAS-III의 회복 관리기 모듈은 ARIES(Algorithm for Recovery and Isolation Exploiting Semantics)[4] 알고리즘을 채택하고 있다. MIDAS-III의 회복 관리기 모듈은, 시스템의 정상적인 수행시에는, 데이타베이스에 변화를 일으키는 연산이 수행되었을 때 변화된 부분의 이전 내용(before image)과 변경된 내용(after image)으로 로그 레코드를 만들어 메모리 내의 로그 버퍼에 저장한다. 로그 버퍼가 꽉 차거나 여러 경우로 인해 로그 버퍼의 로그 레코드가 로그 화일에 저장된다.

MIDAS-III의 고장 회복 기능은 부분적 철회(partial rollback)를 지원하고 페이지 기반의 회복을 쓰며 재수행/복귀(redo/undo)는 물리적(physical)으로 수행한다. 또한 퍼지 검사점(fuzzy checkpoint)을 지원한다. 재시동시에는 restart 유틸리티를 실행시켜, 분석(analysis pass), 재수행(redo pass), 복귀(undo pass)의 단계를 수행하여 시스템을 회복한다.

MIDAS-III의 로그 레코드는 로그 레코드 타입에 따라 약 80여개의 종류가 지원되고 있다.

4. MIDAS-III/TS의 고장 회복

MIDAS-III가 3차 저장장치를 부착하여 MIDAS-III/TS로 확장됨에 따라 새로운 고장의 유형이 생겨나게 되었다. 이에 따라 MIDAS-III가 제공하는 고장 회복 유틸리티인 restart를 확장한 MIDAS-III/TS의 고장회복 유틸리티, restart -t의 구현이 필요하다.

본 절에서는 MIDAS-III/TS의 고장 회복 유틸리티 restart -t의 구현을 위해서 MIDAS-III/TS에서 발생할 수 있는 새로운 고장의 유형과 그에 따른 문제점을 파악하고, 회복 방법을 제시한다.

4.1. 새로운 고장 유형

MIDAS-III에 3차 저장장치를 부착하면서 여러 가지 상황에서 새로운 고장이 발생할 수 있다. 대표적인 고장의 유형은 다음과 같다

- 디스크에서 3차 저장장치로의 화일 이동시 고장
- 3차 저장장치에서 디스크로의 익스텐트 캐쉬시 고장
- 3차 저장장치로의 저장매체의 삽입 및 3차 저장장치로부터의 저장매체 제거시 고장

4.2. 고장 유형에 따른 고장 회복에서의 문제점 및 회복 방법

본 절에서는 4.1절에서 기술한 대표적인 고장의 유형에 대해서 고장 회복에서의 문제점과 회복 방법을 기술한다.

4.2.1. 디스크로부터 3차 저장장치로의 화일 이동시 고장

(1) 화일 이동 작업시 고장 회복의 문제점

MIDAS-III/TS에서는 화일을 3차 저장장치로 이동하기 위하여 다음과 같은 작업을 수행한다.

① 화일의 이동시 요구되는 익스텐트 수와 저장 매체의 사용 가능한 익스텐트 수를 계산하여 화일의 이동 가능 여부를 확인한다.
② 화일의 타입을 3차 저장장치 화일로 수정한다.
③ 화일을 이동하고, 화일 이동에 따른 화일 사상 카탈로그 정보를 작성한다.
④ 화일의 이동이 끝나면, 디스크의 익스텐트를 해제시킨다.

위와 같은 화일 이동 작업에서의 고장 발생시 고장 회복은 3차 저장장치의 저장매체를 화일 이동 이전의 상태로 복구해야 한다. 그러나 MIDAS-III에 부착된 3차 저장장치의 저장매체는 데이타의 붙여쓰기 및 읽기만(append and read-only) 가능하므로 화일 이동 작업의 철회가 불가능하다.

(2) 회복 방법

고장의 회복을 위해서는 (1)에서의 문제점이 우선 해결되어야 한다. 따라서, 다음과 같이 문제점을 해결한다.

- ③의 화일 이동 시점에서 화일 사상 정보만을 작성하고, 실제 화일의 이동은 트랜잭션 승인 시점에 수행한다.
- 실제 화일의 이동을 트랜잭션 승인 시점에 수행할 때, 화일 이동 작업이 요청된 시점과 트랜잭션의 승인 시점 사이에 3차 저장장치로 이동할 화일은 3차 저장장치 화일이지만, 디스크에 존재한다. 따라서, 화일 이동 작업이 요청된 시점과 트랜잭션의 승인 시점 사이에 이동시킬 화일에 속한 데이타를 읽으려 할 경우, 화일을 3차 저장장치가 아닌 디스크에서 찾을 수 있도록, 새로운 화일 타입을 정의한다.

고장 회복을 위해 화일 이동 작업 순서에 따라 다음과 같은 로그 레코드를 작성하고 로그 화일에 기록한다.

ⓐ 화일 이동 작업에서 화일의 타입을 3차 저장장치 화일로 수정시, 화일 타입에 대한 로그 레코드
ⓑ 화일 이동 작업에서 화일 사상 정보 작성시 화일 사상 정보에 대한 로그 레코드
ⓒ 화일 이동 작업을 포함한 트랜잭션 승인 시점에서, 실제 화일 이동의 시작을 표시하는 로그 레코드
ⓓ 화일 이동 작업을 포함한 트랜잭션 승인 시점에서, 실제 화일 이동의 끝을 표시하는 로그 레코드

고장 회복시에 로그 화일에 기록되어 있는 로그 레코드를 바탕으로 다음과 같이 복구한다.

- **ⓑ까지의 로그 레코드가 존재하는 경우**: ⓐ와 ⓑ의 로그 레코드

를 이용해 화일 타입과 화일 사상 카탈로그 화일을 이전 상태로 복구한다.

- **ⓒ까지의 로그 레코드가 존재하는 경우**: 화일 이동 중에 일어난 고장이므로, ⓐ와 ⓑ의 로그 레코드를 이용하여 화일 타입과 화일 사상 카탈로그 화일을 이전 상태로 복구하며, 저장매체의 쓰고 있던 익스텐트의 남은 부분을 *dummy* 값으로 채우고, 3차 저장장치상태 정보 페이지를 수정하여, 저장매체의 상태와 3차 저장장치의 저장매체 상태 정보가 일치하도록 한다.
- **ⓓ까지의 로그 레코드가 존재하나, 해당 트랜잭션의 승인을 나타내는 로그 레코드가 존재하지 않는 경우**: 3차 저장장치 상태 정보 페이지를 수정하여, 저장매체의 상태와 저장매체의 상태 정보가 일치하도록 한다.

4.2.2 3차 저장장치에서 디스크로의 익스텐트 캐쉬시 고장

(1) 익스텐트 캐쉬 작업시 고장 회복의 문제점

원하는 데이타가 속한 페이지가 3차 저장장치에 저장된 화일에 속하는 경우, 3차 저장장치로부터 원하는 데이타가 있는 페이지를 읽기 위한 과정은 다음과 같다.

① 화일 사상 카탈로그 정보를 이용해 3차 저장 장치에서 해당 페이지가 속한 익스텐트의 위치를 확인한다.
② ①에서의 익스텐트 위치 정보를 가지고, 현재 디스크 캐쉬에 해당 익스텐트가 존재하는지 확인한다.
③ 해당 익스텐트가 디스크 캐쉬에 존재할 경우, ⑥으로 이동하고 존재하지 않을 경우 디스크 캐쉬에 익스텐트를 캐쉬할 공간이 있는지 확인한다. .
④ 해당 익스텐트를 캐쉬할 공간이 부족할 경우, LRU 알고리즘에 따라 교체할 익스텐트를 결정해, 익스텐트 정보를 free로 표시한다.
⑤ 필요한 익스텐트를 디스크로 캐쉬하고 익스텐트 정보를 새롭게 디스크로 캐쉬해 온 익스텐트 정보로 수정한다.
⑥ 익스텐트로부터 필요한 페이지를 읽는다

위와 같은 익스텐트 캐쉬 과정에서 3차 저장 장치로부터 익스텐트를 디스크로 캐쉬하는 중이나, 3차 저장 장치로부터 익스텐트를 디스크로 캐쉬하는 작업이 끝난 이후에 고장이 발생하여, 디스크 캐쉬 작업을 복귀(undo)해야 하는 경우에는, 3차 저장장치로부터 이전의 익스텐트를 다시 디스크로 캐쉬해 와서 복구해야 한다. 그러나 이와 같은 고장 회복은 3차 저장장치와 디스크 사이의 데이타 이동 시간을 고려해 볼 때, 고장 회복이 오래 걸린다는 문제점이 있다.

(2) 회복 방법

3차 저장장치로부터 디스크로의 익스텐트 캐쉬 작업은 3차 저장장치에 이미 존재하는 데이타를 디스크로 캐쉬하는 것이므로, 디스크 캐쉬의 상태를 새로운 익스텐트를 캐쉬하기 이전 상태로 복구하지 않고, 익스텐트 사상 정보와 디스크 캐쉬에 있는 익스텐트의 내용이 일치하도록 고장 회복을 한다.
3차 저장 장치로부터 디스크로의 익스텐트 캐쉬 과정의 순서에 따라 다음과 같이 로그 레코드를 작성하고 로그 화일에 기록한다.

ⓐ 익스텐트 캐쉬 작업에서 ④의 수행시, 교체될 익스텐트 정보를 위한 로그 레코드
ⓑ 익스텐트 캐쉬 작업에서 ⑤의 수행 이전에, 익스텐트를 캐쉬하는 과정의 시작을 나타내는 로그 레코드
ⓒ 익스텐트 캐쉬 작업에서 ⑤의 수행 완료 후에, 익스텐트를 캐쉬하는 과정의 끝을 나타내는 로그레코드와 새롭게 디스크로 캐쉬된 익스텐트의 정보를 위한 로그 레코드

고장 회복시에 로그 화일에 기록되어 있는 로그 레코드의 내용을 바탕으로, 다음과 같이 복구한다.

- **ⓐ까지의 로그 레코드가 존재하는 경우**: 3차 저장 장치로부터 디스크로의 새로운 익스텐트의 캐쉬가 일어나지 않았으므로, 디스크 캐쉬에는 교체하려 했던 익스텐트가 남아있다. 따라서 ⓐ의 로그 레코드를 이용하여, 익스텐트 사상 정보 페이지를 복구한다.
- **ⓑ까지의 로그 레코드가 존재하는 경우**: 3차 저장장치에서 디스크로 캐쉬해 오려는 익스텐트의 데이타가 일부 디스크로 옮겨졌으므로, 익스텐트 사상 정보 페이지의 디스크 캐쉬에 대한 익스텐트 정보를 free 상태로 표시한다.
- **ⓒ까지의 로그 레코드가 존재하는 경우**: 3차 저장장치에서 디스크로 캐쉬해 오려는 익스텐트가 전부 옮겨진 상태이므로, ⓒ에서 새롭게 캐쉬한 익스텐트의 정보를 저장하고 있는 로그 레코드를 이용해 익스텐트 사상 정보 페이지 내용과 디스크 캐쉬의 익스텐트 상태가 일치하도록 한다.

4.2.3 3차 저장장치로의 저장매체의 삽입 및 3차 저장장치로부터의 저장매체 제거시 고장

3차 저장장치로의 저장매체의 삽입 또는 3차 저장장치로부터의 저장매체 제거는 각각 loadplatter와 ejectplatter 유틸리티에 의해 이루어진다. 3차 저장장치로의 저장매체의 삽입 또는 3차 저장장치로부터의 저장매체 제거시 발생한 고장에 대비하기 위해서 다음과 같은 로그 레코드가 필요하다.

- 저장매체의 삽입이나 저장매체의 제거 이전의 3차 저장장치 상태 정보 페이지에 대한 로그 레코드

이들 레코드는 loadplatter 또는 ejectplatter 유틸리티의 실행 직전에 로그 화일에 기록한다. 이에 따른 고장 회복 방법은 다음과 같다.

- 3차 저장장치 상태 정보 페이지의 내용이 3차 저장장치 상태와 일치하도록 3차 저장장치 상태 페이지의 내용을 복구한다.

5. 결론 및 향후 연구

확장 전 MIDAS-III의 시스템 고장 회복은 restart 유틸리티에 의해 수행된다. MIDAS-III에 3차 저장장치를 부착하여 MIDAS-III/TS로 확장됨에 따라 새로운 고장 유형이 생겨나게 되었다. 이에 본 논문에서는 MIDAS-III/TS의 고장 회복 유틸리티인 restart -t를 구현하기 위하여, MIDAS-III/TS에서 발생할 수 있는 대표적인 고장의 유형을 파악하고, 그에 따른 회복 방법을 제시하였다.
MIDAS-III/TS는 Solaris 2.7 환경에서, HP사의 광 디스크 쥬크박스인 HP SureStore 320ex를 부착하여 구현되었으며, 고장 회복 유틸리티인 restart -t의 구현은 본 논문에서 제안한 고장 회복 방법을 바탕으로 수행되고 있다.

6. 참고 문헌

[1] M. Carey et al., "Tapes Hold Data, Too: Challenges of Tuples on Tertiary Store," Proc. ACM SIGMOD Int'l Conf., 1993, pp. 413-417.
[2] J. Yu and D. Dewitt, "Query Pre-execution and Batching in Paradise: A Two-Pronged Approach to the Efficient Processing of Queries on Tape-Resident Data Sets," Proc. Int'l Conf. on Scientific and Statistical Database Management, 1997.
[3] S. Sarawagi, "Query Processing in Tertiary Memory Databases," Proc. Int'l Conf. on VLDB, 1995, pp. 585-596.
[4] C. Mohan et al., "ARIES: A Transaction Recovery Method Supporting Fine-Granularity Locking and Partial Rollbacks Using Write-Ahead Logging," ACM Trans. on Database Systems, Vol. 17, No. 1, 1992, pp. 94-162.