논•문 17-1-5

지역성에 기반한 RAID 5의 2단계 디스크 캐시

(Two-Level Disk Cache of RAID 5 based on Locality)

허정호*, 장태무*

*동국대학교 컴퓨터공학과

*{hjh99, jtm}@dgu.ac.kr

요 약

RAID 시스템에서 디스크 캐시는 시스템 성능 향상에 중요한 요소 중 하나이다. 2단계 캐시는 1단계 캐시에 비해 우수한 성능을 보이고 시간적, 공간적 지역성에도 효율적이다. 제안된 캐시 시스템은 2 단계로 구성되어 1단계 캐시는 작은 블록 크기로 구성되어 세트 연관 사상 방식을 이용하고 2단계 캐시는 큰 블록 크기로 구성되어 전 연관 사상 방식을 사용한다. 본 논문에서는 특히 대용량 디스크 캐시에서 디스크 입출력 시간을 향상시키고 효율적으로 일관성을 유지할 수 있는 디스크 제어기 상에 위치하는 RAID 5 디스크 캐시의 모델을 제시하여 적중률과 서비스 시간 면에서 본 논문에서 제시한 2단계 캐시 구조가 향상되었음을 보이고자 한다

1. 서론

최근 프로세서의 처리속도는 매년 40-100%씩 증가되는데 비해 기계적인 부품을 가진 자기 디스크의 성능 향상은 7%정도에 불과하다. 따라서 중앙처리장치 및 주기억장치와 디스크 입출력 속도의 차이를 극복하기 위한 방법으로 자기 디스크의 신뢰도와 성능을 증가시키는 RAID(Redundant Arrays of Inexpensive Disks)가 널리 사용되어 왔다[1]. 디스크 캐시는 디스크 접근을 줄이는 방안으로 널리 쓰이며 DRAM(Dynamic Random Access Memory)의 가격은 매 10년 만에 100 분의 1로 떨어질 정도로 저렴해

지는 추세이므로[2] 대용량의 캐시를 사용하는 경향이 높다. 본 논문에서는 대용량화된 디스크 캐시에서 적절한 구현 방안이 되고 성능도 개선할 수 있는 시간적 및 공간적 지역성을 활용한 RAID 5의 2단계 디스크 캐시 모델을 제안하고자 한다.

2단계 디스크 캐시는 각 단계별 캐시 간에 데이터 일관성 유지를 위한 쓰기정책과 포함관계 특성에 따라 다른 동작을 하게 되는데 본 논문에서는 두 캐시의 내용이 포함관계를 어느 정도 갖는 모델인 NVM(Non-Volatile Memory)모델 1과 포함관계를 갖지 않는 모델인 NVM 모델 2로 구분하였다. 본 논문에서는 2단계 디스크 캐시의 예로 리디움 밧데리(Lithium Battery)를 이용한 비 휘발성(Non-Volatile) 기억소자를 1단계 캐시로 사용하고 휘발성(Volatile) 기억소자를 2단계 캐시로 사용하는 모델의 예를 들고 운영방법을 제시하여 비 휘발성 기억소자를 사용하는 통상적인 디스크 캐시 제어기에 비해 캐시 적중률을 높여 수행속도를 개선시키는 결과를 얻고자 한다.

2. 관련 연구

작은 블록으로 구성된 전형적인 1단계 캐시[3]나 cached RAID[4]는 LRU 정책을 사용하고 캐시 부재시 직접 디스크 배열로부터 부재된 블록을 캐시로 가져오는데 이는 구현하기 쉬운 장점이 있는 반면 선인출이 부족하여 부재율이 높다. 그러므로 캐시의 성능을 향상시키기 위한 다양한 구조의 2단계 캐시 모델들이 연구되었다.

Split temporal/spatial cache(STS)(5)는 2단계 계층 구조이며 선인출 기법을 이용한 일반적인 블록 크기를 지원하는 공간적 캐시와 워드 단위로 된 시간적 캐시로 구성되어 컴파일러에 의해 데이터의 지역성이 분류되어 각각 저장된다.

Selective cache(6)는 작은 블록을 지원하는 시간적 캐시와 큰 블록을 지원하는 공간적 캐시, 지역성 예측 테이블로 구성되어 만약 요청된 데이터가 두 캐시에 존재하지 않을 경우 접근 실패가 발생되어 프로세서는 접근 실패를 처리할 동안 지연된다. 이때 요청된 데이터는 지역성 예측 테이블의 정보로부터 판단되어 시간적 캐시와 공간적 캐시에 각각 저장된다. 시간적, 공간적 지역성의 판단이 되지 못할 경우에는 캐시에 저장되지 못하고 지역성 예측 테이블에 그 정보만을 저장한다.

Victim cache(7)는 주 캐시와 victim 버퍼를 동시에 참조하는 형태로 victim 버퍼에서 적중이 되면 그 데이터는 프로세서로 보내거나 쓰기 동작이 일어나고 주 캐시인 직접 사상 캐시와 블록 교체가 일어나 추가적인 한 사이클의 동작이 필요하다. 두 캐시에서 모두 접근 실패가 일어나면 디스크로부터 필요한 블록을 인출하여 주 캐시에 저장하고 동시에 victim 버퍼에 저장된다. 이 동작은 캐시 제어기가 접근 실패를 수행하는 동안 처리할 수 있으므로 추가적인 시간 지연은 발생하지 않는다. 이는 주 캐시의 단점인 충돌에 의한 접근 실패를 줄여 시스템 전체 접근 적중률을 높일 수 있으나 데이터 교체량이 많고 높은 성능을 얻기 위해서는 큰 블록 크기를 제공해야 한다.

Dual data cache(8)는 2개의 캐시로 구성되어 시간적 지역성과 공간적 지역성을 분리하고 소규모 블록으로 된 직접 사상 방식과 큰

규모 블록으로 된 전 연관 사상 방식이다. 복수 개의 이웃한 작은 블록들을 선인출하여 공간적 지역성을 높이고 과거에 선택적으로 공간 버퍼에 참조된 작은 블록들로 시간적 지역성을 높인다. 이는 일반적인 직접 사상 캐시에 비해 4배의 더 작은 규모의 캐시 크기로 같은 효과를 얻을 수 있어 1단계 캐시에 비해 더 나은 성능을 가져온다.

HP 7200 assist cache(9)는 시간 간격과 선인출 기법을 사용하여 컴파일러의 지원으로 시간적 지역성을 효과적으로 이용하고자 한 캐시로 시간적 지역성으로만 판단되는 경우 직접 사상 캐시에 저장하고 같은 계층에 완전 연관 캐시로 구성되어 두 캐시를 동시 접근하여 한 사이클 동안 접근 적중, 실패를 결정하게 된다. 접근 실패가 발생하거나 선인출 기법을 사용하여 블록을 버퍼에 저장한다.

NTS caches(10)는 완전 연관 버퍼와 직접 사상 캐시로 구성되어 시간적 지역성과 비시간적 지역성을 가진 블록을 구분하여 시간적 지역성을 가진 데이터를 효과적으로 사용하고자 하며 캐시 오염을 방지하기 위한 방법이다.

위에서 제시된 여러 가지 2단계 캐시는 1단계 캐시보다는 성능이 우수하지만 본 논문에서 제안된 캐시와의 차이점은 연관도와 블록 크기이다. (5, 6)은 다른 블록 크기를 지원하지만 같은 연관도를 갖고, (7, 9, 10)은 같은 블록 크기를 지원하지만 다른 연관도를 갖는다. 본 논문에서 제안된 캐시는 다른 연관도와 다른 블록 크기를 이용하여 시간적, 공간적 지역성을 효과적으로 이용할 수 있어 캐시의 성능을 향상시킨다. 또한 서로 다른 특성으로는 (5)는 공간적 지역성을 이용하기 위해 선인출 기법을 이용하고 (6)은 요청한 데이터가 어떤 지역성을 가지는지를 판단하기 위한 지역성 예측 테이블을 활용한다. (10)은 시간적 지역성을 효율적으로 이용하기 위한 구조이며 추가적으로 시간적 지역성을 예측하는 장치를 갖고 있다. 이러한 (5, 6, 10)은 지역성을 결정하는 방법과 한 가지 지역성만을 효과적으로 반영하기 위한 구조이므로 본 논문의 시간적, 공간적 지역성을 높이는 구조와는 구별된다. (8)은 같은 계층에 위치한 이중 캐시 구조로 본 논문에서 제안된 2단계 캐시와는 계층 구조가 다르며 본 논문에서는 RAID 시스템에 알맞는 응용환경과 특성을 고려하여 RAID 시스템의 성능을 개선시키고자 하였다.

3. RAID에서의 2단계 캐시의 필요성

RAID에서의 캐시는 응용 프로그램에서 메모리 접근이 임의로 이루어지지 않고 지역성[11]이 존재한다는 측면에서 의미가 크다. 즉 시간적 지역성과 공간적 지역성의 면에서 L1 캐시는 시간적 지역성을, L2 캐시는 공간적 지역성을 높일 수 있기 때문이다. 이 지역성은 RAID 시스템에서 캐시의 존재를 필수적으로 만들어주는데 캐시는 읽기/쓰기의 버퍼처럼 사용되고 저속으로 실행되는 대용량 디스크 배열의 접근시간의 감소를 가져와 응용 프로그램의 수행속도를 개선시킬 수 있기 때문이다. 응용환경과 구성모델의 분석에 따라 RAID의 캐시에 전형적인 메모리 캐시를 구별하여 적용해야 하는 점은 다음과 같다.

첫째, 용량의 증가가 비교적 용이하다.

전형적인 메모리 캐시는 제조원가나 구성

의 제약으로 인해 대규모로 구성하기 어렵지만 RAID의 캐시는 SDRAM(Synchronous DRAM)으로 구성되어 용량의 증가가 비교적 용이하다. 예를 들어 L1 캐시는 리디움 밧데리를 이용하여 비휘발성 기억소자로 구성하고 L2 캐시는 휘발성 기억소자로 구성한다면 저가이면서 고성능의 2단계 캐시가 구성될 수 있다.

둘째, RAID 시스템의 공간적 지역성을 높일 수 있다.

대부분의 컴퓨터 시스템은 메모리에 세그먼트나 페이지에 기초한 알고리즘을 사용하는데 논리적으로 연속적인 어느 파일이 메모리에서는 물리적으로 연속되지 않는다. 전형적인 하드웨어 캐시는 주로 시간적 지역성에만 집중되지만 RAID의 캐시는 디스크 배열로부터 호스트(Host)로 데이터를 저장할 수 있는 실제적인 버퍼이며 대부분의 운영체제에서 파일은 트랙이나 버퍼 단위로 저장된다. 또한 대부분의 응용프로그램은 벤치마킹 결과 50% 이상의 인접 데이터를 사용하므로 시간적 지역성뿐만 아니라 L2 캐시를 이용하여 공간적 지역성 또한 높일 수 있다 [12].

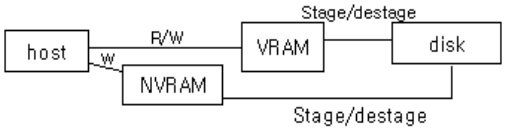
셋째, RAID의 캐시는 RAID 제어기에 내장된 프로세서에 의해 운영되므로 알고리즘이 호스트의 프로세서에 영향을 주지 않는다 [13].

전형적인 하드웨어 캐시의 사상 알고리즘과 주소변환은 호스트의 프로세서에 의해 운영되며 알고리즘의 복잡성을 가져온다. RAID 시스템의 캐시는 RAID 제어기의 내장된 프로세서에 의해 운영되므로 알고리즘의 복잡성이 호스트의 프로세서에 영향을 주지 않는다.

위와 같은 이유로 메모리 캐시에 RAID 디스크 캐시를 그대로 적용하기는 어려우므로 RAID 디스크에 해당되는 캐시 설계가 필요하다. 즉, L1 캐시는 시간적 지역성을, L2 캐시는 공간적 지역성을 높여 주므로 RAID에서 2단계로 캐시 설계를 하여 이를 읽기/ 쓰기의 버퍼처럼 사용하여 저속으로 실행되는 대용량 디스크 배열의 접근시간의 감소를 가져와 수행 속도를 개선한다.

4. 전형적인 디스크 캐시 모델

디스크 캐시에서 사용되는 2단계 캐시 모델은 (그림 1)과 같으며 1단계 캐시에서는 쓰기 시에 이용되는 비휘발성 기억소자를, 2단계 캐시에서는 읽기/쓰기에 이용되는 휘발성 기억소자를 사용한다.



(그림 1) NVRAM을 이용한 전형적인 디스크 캐시

(그림 1)에서 1단계 캐시는 쓰기 시에만 이용되며 읽기는 하지 않고 1단계 캐시에 기록된 자료가 디스크로 destage될 때는 해당 블록의 내용이 분리되거나 1단계 캐시가 찼을 때 그리고 일정시간 동안 1단계 캐시를 사용하지 않을 때이다. 2단계 캐시에서의 탐색과 비교 결과에 의해 디스크로 읽기, 쓰기 각각에 대한 적중, 부재 시의 동작은 다음과 같다.

1)읽기 적중: VRAM에서 데이터를 읽는다

2)읽기 부재: 디스크에서 요청된 블록이 VRAM으로 전송되어 host 측으로 가고 트랙의 나머지 부분이 VRAM으로 stage된다.

3)쓰기 적중:4가지 모드가 있다

첫째, DFW(DASD fast write): NVRAM과 VRAM에 모두 쓰기

둘째, CFW(Cache fast write): VRAM에만 쓰기

셋째, DC(dual copy):신뢰성을 높이기 위해 디스크 상의 두 군데에 이중 복사

넷째, DCFW(dual copy with fast write): DC 모드이면서 NVRAM에도 쓰기

DFW와 CFW의 모드 선택은 디스크 일관성의 중요도와 NVRAM의 용량에 의해 결정할 수 있다.

4)쓰기 부재: 쓰기 동작하기 위한 그 블록이 포함된 디스크 트랙의 나머지 부분이 stage되고 쓰기 적중으로 감

이 모델은 stage, destage는 비동기적으로 이루어지므로 해당 읽기, 쓰기 동작에는 영향을 주지 않으나 요청된 읽기, 쓰기 동작이 디스크 캐시가 사용 중이므로 대기하여야 하고 디스크 캐시에 접근이 이루어지고 있는 동안에도 stage/destage는 이루어질 수 없으므로 디스크 캐시의 성능과 관계가 있다. 또한 디스크 일관성의 중요도로 볼 때 쓰기 적중 시 DFW로 동작한다면 NVRAM과 VRAM에 있는 자료가 중복되어 고가의 NVRAM을 데이터 백업용으로만 사용하게 되며 쓰기의 비율이 점차 증가하여 읽기/쓰기가 비슷해져가는 최근의 추세로 볼 때 1차 캐시를 충분히 활용하지 못하는 결과가 된다. 또한 쓰기가 많은 경우 NVRAM의 용량이 충분하지 않다면 destage가 자주 발생되고 NVRAM을 사용하여 동일한 주소에 대한 여러 번의 쓰기 요청을 한번의 디스크 동작으로 처리할 수 있는 효과가 적어져 30초 이내에 40% 정도의 수정자료가 새로 지워진다는 연구 자료[14]로 볼 때 빈번한 디스크 접근은 디스크 캐시의 성능을 저하시킨다.

본 논문에서는 2단계 캐시의 한 모델로 프로세서에 가까운 1단계 캐시는 NVRAM이고 디스크에 가까운 2단계 캐시는 VRAM인 모델을 2가지 제시하고 각각의 운영방법을 다르게 하여 전형적인 2단계 디스크 캐시와 비교하여 본 논문에서 제시한 2단계 디스크 캐시 모델의 성능이 우수함을 입증하고자 한다.

본 논문에서 제시하는 2가지 모델을 각각 NVM 모델 1, NVM모델 2라 정의한다. NVM 모델 1은 NVRAM의 내용을 VRAM이 어느 정도는 포함하고 있으며 1단계 캐시의 읽기/쓰기와 2단계 캐시와 디스크와의 stage/destage가 병렬성을 가지므로 디스크의 stage/destage의 오버헤드가 적다. NVM 모델 2는 NVM 모델 1의 장점과 NVRAM의 내용을 VRAM이 포함하지 않으므로 해서 생기는 NVRAM만큼의 용량 증가 효과를 얻고자 한다.

5. RAID 5에서의 2단계 디스크 캐시 모델

본 논문에서는 NVRAM(Non-Volatile RAM)을 디스크 캐시에 이용하여 L1 캐시에서 읽기/쓰기에 이용하고 L2 캐시는 휘발성 기억소자로 구성된 2가지 디스크 캐시 모델을 제안한다. 각 단계 캐시의 용량은

NVRAM을 사용하는 통상적인 시스템에서 NVRAM 및 휘발성 기억소자의 캐시의 용량과 유사하게 구성하며 L1 캐시의 읽기/쓰기와 L2 캐시와 디스크 사이의 동작은 병렬성이 있다. L1 캐시에 선택적으로 캐싱된 작은 블록은 생명 주기가 증가되어 시간적 지역성이 높아져 캐시 부재가 발생할 때마다 L2 캐시에 있는 다수의 이웃한 소규모 블록들을 선인출하여 공간적 지역성을 높인다. 본 논문에서 제안된 2가지 모델의 주요 특성은 다음과 같다.

첫째, L1 캐시는 세트 연관 사상 알고리즘을 사용하고 L2 캐시는 완전 연관 사상 알고리즘을 사용한다.

작은 블록 크기의 L1 캐시는 세트 연관 사상 알고리즘을 사용하여 부재율을 낮추고 스래싱 효과(thrashing effect)를 줄인다. 그리하여 직접 사상 방법에 비해 시스템의 효율이 개선되고 전 연관 사상 방법에 비해 탐색 시간 및 주기를 줄이고 시스템 속도를 높인다. L2 캐시에서는 비교적 큰 규모의 소수의 데이터 블록이 전 연관 사상 방법을 택하는 공간적 버퍼이다. 그리하여 충돌 부재(conflict miss)를 최소화하고 탐색에 대한 시간낭비를 줄인다.

둘째, L1 캐시는 LRU 교체 알고리즘을 사용하고 L2 캐시는 LFU 교체 알고리즘을 사용한다.

L1 캐시는 시간적 지역성을 고려하여 LRU정책을 사용한다. L2 캐시는 이미 L1 캐시에서 실행 프로그램이 가진 시간적 지역성을 고려하였으므로 L2 캐시로 넘어오는 요청은 시간적 지역성이 줄어들게 되어 LRU 정책의 의미가 적다. 그러므로 L2 캐시에서는 LFU 정책도 고려하여 디스크와 내용이 일치하는 것에 한하여 교체 알고리즘을 적용한다. 셋째, L1 캐시와 L2 캐시의 라인 크기를 다르게 운영한다.

L1 캐시는 작은 단위인 섹터나 그보다 더 작은 단위로 하고 L2 캐시는 비교적 큰 단위인 트랙 단위로 한다. 라인의 크기는 캐시의 적중률과 관계가 밀접한데 L2 캐시의 적중률을 높이기 위해서는 L1 캐시의 적중률을 높여야 한다. L1 캐시의 속도가 L2 캐시에 비해 빠르고 L2 캐시는 디스크 동작을 해야 하므로 디스크와의 회전 위치 유실 지연을 막으려면 L2 캐시는 트랙단위여야 한다. L1 캐시의 라인크기는 섹터로 하여 단편화를 방지하고 입출력 요청 단위가 클 경우에도 L2 캐시가 L1 캐시의 내용을 포함하는 트랙을 유지해야 하므로 L2 캐시의 공간적 지역성을 높여야 한다.

넷째, 쓰기 정책은 write back을 사용하여 L2 캐시를 디스크 동작에 전담시키고, write through는 L1 캐시가 비 휘발성 기억소자이므로 고려하지 않는다.

2단계 캐시의 개념적인 구조는 (그림 2)와 같으며 L2 캐시는 작은 블록들이 특정한 큰 블록에 속해 있는 형태로 특정한 큰 블록이 교체될 때까지 존재한다[13].

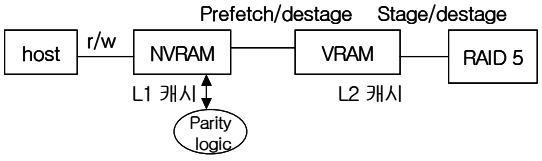
(그림 2) 2단계 캐시의 구조

또한 작은 블록이 포함된 큰 블록이 교체될 때 뿐 아니라 전에 접근되었던 여러 개의 작

은 블록들이 L1 캐시로 선택적으로 이동한다. L1 캐시를 작은 블록 사이즈로 하면 시간적 지역성을 가진 데이터가 주어진 캐시 공간에서 L1의 블록 개수를 더 증가시킨다.

5.1 NVM 모델 1

본 논문에서 제시하는 RAID 5에서의 2단계 디스크 캐시 NVM 모델 1은 (그림 3)과 같고 L1 캐시에서 읽기/쓰기하며 동작원리는 다음과 같다.



(그림 3) NVM 모델 1

(1) 읽기/쓰기 모두 L1 캐시에서 하고 읽기 부재 시에는 L2 캐시에 접근하여 적중하면 L1 캐시로 옮겨와 접근한다.

(2) 쓰기 동작은 L1 캐시에 접근 후 완료된다. L2 캐시 부재인 경우 L2 캐시로의 stage동작이 비동기적으로 수행된다. 이후 수정된 라인은 L2 캐시로 옮겨진 후 디스크로 destage된다. 쓰기 동작 이후의 stage와 destage는 비동기적인 동작이며 이후의 읽기/쓰기가 L1 캐시에 적중하였을 때의 대기시간을 줄일 수 있다.

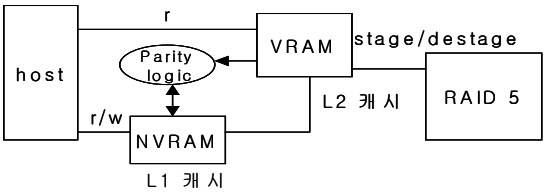
(3) L2 캐시는 L1 캐시의 내용을 어느 정도는 포함하지만 항상 내용이 일치할 필요는 없고 L1 캐시가 수정된 상태에서 destage동작을 지연시켜도 안정성은 유지된다.

(4) L1 캐시는 공간적 지역성을 감안하여 순차적으로 여러 라인을 읽어오고 L2 캐시는 탐색 거리가 최소인 한 트랙을 선인출 한다.

(그림 3)에서 패리티 회로(parity logic)는 L1 캐시에 쓰기 시에 패리티 갱신이 이루어지는 회로이며[15] 일반적인 RAID 5 쓰기 방법인 읽기-수정-쓰기(Read-Modify-Write) 방식으로 동작하고 독자적인 패리티 엔진에서 이루어진다.

5.2 NVM 모델 2

RAID 5에서의 2단계 디스크 캐시 NVM 모델 2는 (그림 4)와 같고 L1 캐시는 읽기/쓰기, L2 캐시는 읽기에 이용되며 동작원리는 다음과 같다.



(그림 4) NVM 모델 2

(1) 읽기는 L1 캐시, L2 캐시에서 모두 이루어진다. 우선 L1 캐시를 탐색하여 적중하면 읽고 아니면 L2 캐시에 접근한다. 따라서 L1 캐시 부재 시에 L2 캐시에서 L1 캐시에 읽어오는 오버헤드를 감소시킨다. 또한 L1 캐시에서의 읽기 부재 시에 두 캐시에 동시에 존재하는 라인의 수를 줄이는 효과가 있다.

(2) 쓰기동작은 L1 캐시에 접근 후에 완료된다. L1 캐시에서 적중이든 부재든 L2 캐시를 점검하여 부재 시에 L2 캐시로 stage하

는 동작이 따라야 한다. 이후 수정된 라인은 L2 캐시로 옮겨지며 다시 디스크로 destage 된다. 쓰기 동작 이후의 stage/destage는 비동기적인 동작이며 이후의 읽기/쓰기가 L1 캐시에서 적중하였을 때의 대기 시간을 줄일 수 있다.

(3) L2 캐시는 수정된 라인이 디스크에 아직 옮겨지지 않은 경우에는 L1 캐시의 내용을 포함한다. 그 이외의 경우에는 L1 캐시의 내용을 포함하지 않는다.

패리티 회로(parity logic)는 (그림 3)과 같이 읽기-수정-쓰기(Read-Modify-Write)방식으로 독자적인 패리티 엔진에서 이루어진다.

6. RAID 5에서의 2단계 디스크 캐시 모델의 장점

2단계 디스크 캐시 모델의 공통적인 장점은 stage/destage 등의 디스크 동작과 읽기/쓰기 적중 시의 캐시 동작들의 병렬처리가 가능하다는 점이다. 따라서 입출력 요청 자체의 빈도수가 많거나 읽기 부재, 쓰기 동작 등으로 인한 stage/destage 동작이 많아서 디스크 접근이 빈번한 환경에서 매우 효율적이라고 할 수 있다. 본 논문에서 제안한 RAID 5에서의 2단계 디스크 캐시 모델의 장점은 다음과 같이 정리할 수 있다.

첫째, 디스크 캐시 전체를 고가의 반도체 기억장치 부품으로 구성하지 않더라도 성능이 높고 대용량인 저가의 디스크 캐시를 제공할 수 있다.

둘째, 디스크 동작과 각 단계의 읽기 적중, 쓰기 등의 캐시 동작들이 병렬처리가 가능하다.

셋째, 각 단계의 캐시들이 디스크 캐시 전체의 용량으로 사용 가능하다.

넷째, 각 단계별 캐시라인의 크기와 교체 알고리즘, 사상 방식을 다르게 적용하여 캐시 적중률을 높일 수 있다..

다섯째, RAID 5 디스크를 이용하여 패리티 정보를 분산 저장하여 디스크 쓰기 동작을 병렬로 수행하고 신뢰도를 높이다. 또한 캐시에 데이터는 물론 패리티를 함께 적재하여 패리티 연산 시 패리티를 디스크에서 캐시로 가져오기 위한 추가적인 디스크 접근을 방지하여 RAID 5의 '작은 쓰기 문제'(Small Write Problem)를 완화한다.

7. 성능분석

(그림 1)과 같은 전형적인 디스크 캐시 모델에서 평균 적중률을 H, 적중 시의 서비스 시간을 T_h, 부재 시의 서비스 시간을 T_m이라 한다면 평균 입출력 서비스 시간은 다음과 같다.

$$T=H*T_h+(1-H)*T_m \quad (1)$$

$$T_h=oh+T_{ph}+B/X_c \quad (2)$$

$$T_m=om+T_{pm}+sk+LAT+RPS+B/X_d \quad (3)$$

(단 H:적중률, 읽기 적중 및 쓰기의 전체 요청 횟수에 대한 비율

서비스 시간: 디스크 제어기에 입출력 요청이 들어와서 비동기적 동작을 제외한 요청에 대한 처리가 완료될 때까지의 소요 시간

X_c:캐시에서 읽어 오는 속도

X_d:디스크에서 읽어 오는 속도

oh, om:버스 프로토콜 시간 및 적중 또는 부재 시의 캐시에서의 오버헤드

waiting in queue	w ch

T_{ph}, T_{pm}:적중 및 부재 시의 대기시간 (캐시 또는 디스크가 사용 중일 때 기다리는 시간)

B:평균적으로 읽어오는 블록의 크기

sk:평균 디스크 탐색 시간

LAT:평균 디스크 회전 지연 시간 (rotational latency)

RPS:평균 회전 위치 유실 지연 시간)

(그림 5)는 디스크 접근 시간에 대한 세부적인 구성 요소이다.

waiting in queue	waiting for channel	seek	rotational latency	RPS miss delay	data transfer

(그림 5) 디스크 접근 시간 구성 요소

(그림 5)의 구체적인 의미는 다음과 같고 RPS 지연 시간과 큐 대기시간은 시스템의 부하에 좌우된다.

*큐 대기시간(waiting in queue):디스크와 채널이 사용 가능해질 때까지 대기하는 시간

*탐색 시간(seek time):원하는 자료가 저장되어 있는 실린더에 판독/기록 헤드를 위치시키는 데 걸리는 시간

*회전 지연 시간(rotational latency time): 판독/기록 헤드가 원하는 트랙에 도달한 후 그 트랙 내에서 다시 원하는 목적 섹터까지 도달하는 데 걸리는 시간

*RPS 부재 지연 시간(RPS miss delay): 통신 경로(communication path)가 설정될 때까지 디스크가 한 회전하는 데 걸리는 시간

*자료 전송 시간(data transfer):디스크와 채널 혹은 주 기억 장치의 버퍼 간에 자료를 전달하는 데 걸리는 시간

*그 외 채널이나 제어기 오버헤드 시간 (overhead time)

NVM 모델 1에서 1단계와 2단계 캐시 전체에 관한 적중률이 (그림 1)의 전형적인 디스크 캐시에서의 적중률과 같다고 가정하면 T_h는 다음과 같다.

$$T_h=oh+B/X_c+(1-H_1)(T_{ph}+B/X_{12}) \quad (4)$$

(단 H_1=1단계 캐시에서의 적중률, X_{12}:캐시 사이의 자료 전송 속도)

대기시간 T_{ph}는 stage/destage의 빈도수와 소요시간, 그리고 디스크에 대한 요청 빈도에 따라 달라진다. stage/destage에 소요되는 시간 T_{sd}는 다음과 같다.

$$T_{sd}=sk+LAT+RPS+B/X_d) \quad (5)$$

위 식에서 T_{ph}에 관계 있는 부분은 B/X_d이며 RPS부분은 stage/destage 동작 중 캐시 적중이 일어날 때 커지게 되며 (3)식의 T_{pm} 부분을 길게 한다. 만일 디스크에 대한 요청이 지수분포를 갖고 읽기/쓰기 부재가 발생하여 stage 동작을 하고 있을 때 다음의 요청이 캐시에 적중한다고 가정하면 대기시간 T_{ph}는 다음과 같다.

$$T_{ph}=\int_a^b \frac{1}{b} * e^{-x/\beta}(T_{sd}-x)dx \text{(단 } a=T_{sd}-B/X_d, b=T_{sd}) \quad (6)$$

(6)을 scsi 접속회로를 가진 실제 디스크 시스템[16]에서 다음의 여러 값을 얻고, 한 트랙을 모두 stage하고 회전 위치 유실 지연시간을 고려하지 않는다고 가정하여 (7)의 값을 얻고, 이들을 식(6)에 대입하면,

X_c=4MB/sec, X_d=1.628MB/sec, oh=om=2ms, b=512bytes, sk=12.5ms, LAT=0, RPS=0, B/X_d=14.74ms, 평균요청간격 ß=100ms (7)

T_{ph}=1ms 이다.

위의 가정에서 destage는 고려하지 않았기 때문에 실제로는 더 큰 값일 것이라고 판단된다. (4)에서 H_1이 0.7정도이고 120ns의 사이클 시간을 갖는 SDRAM을 사용한다고 가정하면 $(1-H_1)B/X_{12}$는 0.03ms이므로 0.67ms 만큼 캐시 적중시의 서비스 시간이 단축됨을 알 수 있다. T_{pm}도 고려한다면 성능향상은 더욱 뚜렷하다고 판단된다.

8. 결론

본 논문에서는 시간적, 공간적 지역성을 고려하여 2단계 캐시로 캐시 적중률을 높이고 L1 캐시에 비 휘발성 기억소자를 이용하여 전원이 차단되어 시스템이 오류가 나더라도 디스크 캐시의 신뢰도를 높인다. 또한 쓰기 캐시에 데이터와 패리티를 함께 적재하여 쓰기 시에 패리티의 추가적인 디스크 접근 없이 RAID 5의 '작은 쓰기 문제'를 완화시키고자 하였다. 빠른 접근 시간과 다양한 응용 프로그램에 적합한 시간적, 공간적 지역성을 이용한 캐시 적중률을 높이기 위해서는 단일 캐시 구조만으로는 상반된 특성을 가진 두 가지 지역성을 효율적으로 운영하는데 한계가 있으므로 각각의 지역성을 효과적으로 반영할 수 있는 2단계 디스크 캐시를 이용한 RAID 5 제어기를 구현하기 위한 모델을 제시하여 통상적인 2단계 디스크와 비교하였을 때 성능향상을 기대할 수 있다.

앞으로의 작업으로는 본 논문에서 제시한 방법을 수학적인 이론으로 구체화시켜 분석적 모형을 만들고 여러 가지 다양한 작업 부하를 사용한 시뮬레이션을 통해 비교 연구해 볼 수 있다.

참고문헌

[1] A. K. Sahai, " Performance Aspects of RAID Architectures," IEEE International Conference on Performance Computing and Communications, pp.321-327, 1997.

[2] J. Gray and R. Shenoy, "Rules of Thumb in Data Engineering," Technical Report MS-TR-99-100, MicroSoft Research, 2000.

[3] Jai Menon and Jim Cortney, "The architecture of a fault-tolerant cached RAID controller," Proceedings of the 20th annual international symtpsium on computer architecture, pp.76-86, 1993.

[4] Gao Jun, Wu Zhiming, Jiang Zhiping, "A key technology of design in RAID system," computer Peripherals Review, Vol.24, no. 1, pp.5-8, 2000.

[5] V. Milutinovic, M. Tomasevic, B. Markovic, M. Tremblay, "The Split Temporal/Spatial Cache:Initial Performance Analysis," SCIZZL-5, Mar, 1996.

[6] A Gonzalez, C. Aliagas , M. Matco, "Data cache with multiple caching strategies tuned to different types of locality," Supercomputing '95,

pp.338-347, 1995.

[7] N. P. Jouppi, ""Improving Direct-Mapped Cache Performance by the Addition of a Small Fully Associative Cache and Prefetch Buffers," Proc. 17th ISCA, pp.364-373. May. 1990.

[8] Kil-Whan Lee, Gi-Ho Park, Tack-Don Han, Shin-Dug Kim, "An Effective Selection Mechanism to exploit Spatial Locality in Dual Data Cache International Conference on Computers, Communications and Systems' 98, pp.31-37, Nov. 1998.

[9] G. Kurpanchek et al, Pa-7200:A Pa-RISC Processor with Integrated High Performance MP Bus Inteface, COMPCON Digest of Papers, Feb. pp.375-382. 1994.

[10] Jude A. Rivers, Edward S. Davidson, Reducing Conflicts in Direct-Mapped Caches with a Temporality-Based Design," Proceedings of the 1996 International Conference on Parallel Processing, Vol. 1, pp.151-162. Aug. 1996.

[11] Dai Mei-e, "Analysis for organization fashions and character of cache in high property computer system," Microelectronics & Computer, No.5, pp.15-18. 2000.

[12] Jung-Hoon Lee, Jang-Soo Lee and Shin-Duk Kim, "A new cache architecture based on temporal and spaial locality," Journal of Systems Architecture, pp. 1451-1467, 2000.

[13] Chen Yun, Yang Genke, Wu Zhiming, "The Application of Two-Level Cache in RAID System," Proceedings of the 4th World Congress on Intelligent Control and Automation, June, 2002.

[14] M. Alonso, V. Santonja, "A New Destage Algorithm for Disk Cache: DOME," Euromicro '99.

[15] A. Varma, Q. Jacobson, "Destage Algorithms for Disk Arrays with Nonvolatile Caches," IEEE Transactions on Computers, Feb. 1998.

[16] Western Digital, WD-SC8320 Technical Reference Manual, WD0097S8/89 ver 1.0.

허정호

1980~1984 동국대학교 학사(전산)

1984~1987 한양대학교 석사(전산)

1999~2001 동국대학교 박사수료(컴퓨터 공학)

관심분야: 입출력 시스템, 컴퓨터 구조, 병렬처리, 디스크 캐시, RAID 등

장 태무

1977년 서울대학교
전자공학과 졸업(학사)
1979년 한국 과학기술원
전산학과 졸업(이학 석사)
1995년 서울대학교
컴퓨터 공학과 졸업(공학박사)
1979년~1981년 한국 전자 기술 연구소
연구원
1998년 University of Southeastern
Lousiana 교환교수
1981년~현재 동국대학교 컴퓨터.
멀티미디어 공학과 교수

관심분야: 분산 및 병렬 처리, 컴퓨터
구조, 입출력 시스템 등