

2단계 디스크 캐시를 이용한 RAID 5 제어기 설계

허정호⁰, 장태무 {hjh99⁰, jtm}@dgu.ac.kr
동국대학교 컴퓨터공학과

Designing a RAID 5 Controller with Two-Level Disk Cache

Jung-ho Huh, Tae-Mu Chang
Dept. of Computer Engineering, Dongguk University

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 적중률을 향상시켜 수행속도를 개선시키고자 한다.

1. 서 론

최근 프로세서의 처리속도는 매년 40-100%씩 증가되는데 비해 기계적인 부품을 가진 자기 디스크의 성능 향상은 7%정도에 불과하다. 따라서 중앙처리장치 및 주 기억장치와 디스크 입출력 속도의 차이를 극복하기 위한 방법으로 자기 디스크의 신뢰도와 성능을 증가시키는 RAID(Redundant Arrays of Inexpensive Disks)가 널리 사용되어 왔다[1, 2]. 디스크 캐시는 디스크 접근을 줄이는 방안으로 널리 쓰이며 DRAM(Dynamic Random Access Memory)의 가격은 매 10년 만에 100분의 1로 떨어질 정도로 저렴해지는 추세이므로[3] 디스크 시스템에서 대용량의 캐시를 사용하는 경향이 높다. 본 논문에서는 대용량화된 디스크 캐시에서 적절한 구현 방안이 되고 성능도 개선할 수 있는 2단계 디스크 캐시를 이용한 RAID 5 제어기의 모델을 제시하고자 한다.

2단계 디스크 캐시는 각 단계별 캐시 간에 데이터 일관성 유지 위한 쓰기정책과 포함관계 특성에 따라 다른 동작을 갖게 되는데 본 논문에서는 두 캐시의 내용이 어느 정도 포함관계를 갖는 모델(NVM 모델 1)과 포함관계를 갖지 않는 모델(NVM 모델 2)로 구분하였다.

본 논문에서는 2단계 디스크 캐시의 예로 리튬 배터리(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. RAID에서 2단계 캐시의 필요성

RAID에서의 캐시는 응용 프로그램에서 메모리 접근이 임의로 이루어지지 않고 지역성의 원리가[4] 크다. 응용환경과 구성 모델의 분석에 따라 RAID의 캐시에 전형적인 메모리 캐시를 구별하여 적용해야 하는 점은 다음과 같다.

첫째, 용량 증가가 비교적 용이하다.

메모리 캐시에 비해 RAID의 캐시는 구성의 제약이 비교적 덜하다.

둘째, RAID 시스템의 공간적 지역성을 높일 수 있다.

세그먼트(segment)나 페이지(page)에 기초한 알고리즘을 사용하는 전형적인 메모리는 논리적으로 연속적인 파일이 메모리에서는 물리적으로 연속되지 않는다. 전형적인 메모리 캐시는 주로 시간적 지역성에만 집중되지만 RAID의 캐시는 디스크 배열로부터 호스트(host)로 데이터를 저장할 수 있는 실제적인 버퍼이며 운영체제에서 파일은 트랙이나 버퍼 단위로 저장된다. 또한 대부분의 응용프로그램이 벤치마킹 결과 50% 이상의 인접 데이터를 사용하므로 시간적 지역성뿐만 아니라 L2 캐시를 이용하여 공간적 지역성 또한 높일 수 있다[4].

셋째, RAID의 캐시는 RAID 제어기에 내장된 프로세서에 의해 운영되므로 알고리즘이 호스트의 프로세서에 영향을 주지 않는다[5].

위와 같은 세 가지 이유로 메모리 캐시에 RAID 디스크 캐시를 적용하기가 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 지속적으로 실행되는 대용량 디스크 배열의 접근시간의 감소를 가져와 수행 속도를 개선한다.

3. 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 선인출(prefetching)하여 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 부재율(Miss Ratio)을 낮추고 스래싱 효과(Thrashing Effect)를 줄인다. 직접 사상 방법에 비해 시스템의 효율이 개선되고 전 연관 사상 방법에 비해 탐색 시간 및 탐색 주기를 줄이고 시스템 속도를 높인다. L2 캐시에서는 비교적 큰 규모의 소수의 데이터 블록이 전 연관 사상 방법을 택하는 공간적 버퍼이다[5]. 그리하여 충돌 부재(Conflict Miss)를 최소화하고 탐색에 대한 시간낭비를 줄인다.

둘째, 교체 알고리즘은 L1 캐시는 시간적 지역성을 고려한 LRU 정책을 사용하고 L2 캐시는 디스크와 내용이 일치하는 것에 한하여 LFU 정책을 사용한다.

셋째, 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 트랙 단위여야 한다. L1 캐시의 라인크기는 섹터로 하여 단편화를 방지하고 입출력 요청 단위가 클 경우에도 L2 캐시가 L1 캐시의 내용을 포함하는 트랙을 유지한다.

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

2단계 캐시의 개념적인 형태는 (그림 1)과 같으며 L2 캐시는 작은 블록들이 특정된 큰 블록에 속해 있는 형태로 그 블록이 교체될 때까지 존재한다[5].

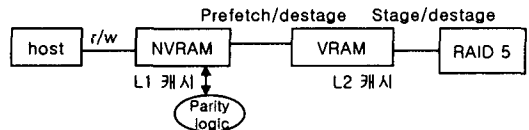


(그림 1) 2단계 캐시의 구조

또한 작은 블록이 포함된 큰 블록이 교체될 때 뿐 아니라 전에 접근되었던 여러 개의 작은 블록들이 L1 캐시로 선택적으로 이동한다. L1 캐시를 작은 블록사이므로 하면 시간적 지역성을 가진 데이터가 주어진 캐시 공간에서 L1 캐시의 블록 개수를 더 증가시킨다.

3.1 NVM 모델 1

본 논문에서 제시하는 RAID 5에서의 2단계 디스크 캐시 NVM(Non-Volatile Memory) 모델 1은 (그림 2)와 같이 L1 캐시에서 읽기, 쓰기하며 L1 캐시의 내용을 L2 캐시가 일부 포함한다.



(그림 2) RAID 5에서의 2단계 디스크 캐시 모델 1

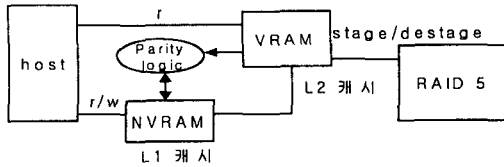
이 모델은 L1 캐시에서 쓰기가 발생할 때마다 L2 캐시로 자료 전송은 하지 않고 수정 발생 사실만 알린다. 이는 L2 캐시에서 발생하는 디스크와의 stage/destage 동작의 독립성을 높이고자 하는 방법이다. 두 캐시는 모두 write back으로 운영되고 L2 캐시를 디스크 동작에 전담시키고 2개 이상의 프로세서가 디스크를 공유하여 읽기/쓰기 동작을 모두 하는 경우에 사용할 수 있으며 두 캐시가 공통적으로 dirty상태일 때는 L1은 L2과 내용이 다르고 L2는 디스크와 내용이 다르다는 것을 나타낸다. NVM 모델 1의 세부적인 동작 특성은 다음과 같다.

- (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 캐시부재인 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 포함하지만 항상 내용이 일치할 필요는 없고 L1 캐시가 수정된 상태에서 destage동작을 지연시켜도 안정성은 유지된다.
- (4) L1 캐시는 공간지역성을 감안하여 순차적으로 여러 라인을 읽어오고 L2 캐시는 탐색 거리가 최소인 한 트랙을 선인출한다.

(그림 2)에서 패리티 회로(parity logic)는 L1 캐시에 쓰기 시에 패리티 갱신이 이루어지는 회로이며[6] 일반적인 RAID 5 쓰기 방법인 읽기-수정-쓰기(Read-Modify-Write)방식으로 동작하고 독자적인 패리티 엔진에서 이루어진다.

3.2 NVM 모델 2

RAID 5에서의 2단계 디스크 캐시 NVM 모델 2는 (그림 3)과 같이 L1 캐시는 읽기, 쓰기, L2 캐시는 읽기에 이용된다. 이 모델은 L2 캐시가 L1의 내용을 포함하지 않는 형태로 전체 캐시의 용량이 L1 캐시들의 합만큼 커질 수 있다. 이는 호스트 측에서 L1, L2 캐시에 모두 접근할 수 있고 읽기 부재 시에 동일한 내용



(그림 3) RAID 5에서의 2단계 디스크 캐시 모델 2

이 두 캐시에 중복되어 존재하지 않는다. stage/destage 동작을 L1 캐시의 동작과 분리하기 위하여 L1과 디스크는 연결하지 않고 호스트 측에 위치한 L1 캐시가 디스크 측에 위치한 L2 캐시에 비해 비교적 빠르므로 L1 캐시에서 읽기 적중하는 것이 좋다. 이 모델은 L2 캐시의 속도에 따라 성능이 좌우되며 세부적인 동작원리는 다음과 같다.

- (1) 읽기는 L1 캐시, L2 캐시에서 모두 이루어진다. 우선 L1 캐시를 탐색하여 적중하면 읽고 아니면 L2 캐시에 접근한다. 따라서 L1 캐시 부재 시 L2 캐시에서 L1 캐시에 읽어오는 오버헤드(overhead)를 감소시킨다. 또한 L1 캐시에서의 읽기 부재 시에 두 캐시에 동시에 존재하는 라인의 수가 줄어든다.
- (2) 쓰기동작은 L1 캐시에 접근 후에 완료된다. L1 캐시에서 적중이든 부재든 L2 캐시를 점검하여 부재 시 L2 캐시로 stage한다. 수정된 라인은 L2 캐시로 옮겨지며 다시 디스크로 destage된다. 쓰기 동작 이후의 stage, destage는 비동기적인 동작이며 이후의 읽기, 쓰기가 L1 캐시에서 적중하였을 때의 대기 시간을 줄일 수 있다.
- (3) L2 캐시는 수정된 라인이 destage되지 않은 경우에는 L1 캐시의 내용을 포함하고 그 이외에는 포함하지 않는다.
- (4) L1 캐시는 선인출하지 않아도 되며 L2 캐시는 NVM 모델 1에서와 같이 탐색 거리가 최소인 한 트랙을 선인출한다.

패리티 회로(parity logic)는 (그림 2)와 같이 읽기-수정-쓰기(Read-Modify-Write)방식으로 독자적인 패리티 엔진에서 이루어진다.

4. RAID 5에서의 2단계 디스크 캐시 모델의 장점

본 논문에서 제안한 RAID 5에서의 2단계 디스크 캐시 모델인 NVM 모델 1과 NVM 모델 2의 장점을 정리하면 다음과 같다.

- 첫째, 디스크 캐시 전체를 고가의 반도체 기억장치 부품으로 구성하지 않더라도 성능이 높고 대용량인 저가의 디스크 캐시를 제공할 수 있다.
- 둘째, stage/destage 등의 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 저장하

여 디스크 쓰기 동작을 병렬로 수행하고 신뢰도를 높이며, 캐시에 패리티를 함께 적재하여 '작은 쓰기 문제'(Small Write Problem)를 완화한다.

5.결 론

본 논문에서는 시간적, 공간적 지역성의 원리를 이용하여 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 체계를 구현하기 위한 모델을 제시하여 통상적인 2단계 디스크와 비교하였을 때 성능향상을 기대할 수 있다.

앞으로의 작업으로는 본 논문에서 제시한 각 모델들의 L1, L2 캐시 간에 일관성을 유지하기 위한 오버헤드 문제와 NVM 모델 2가 캐시의 용량이 커짐에 따라 성능 또한 나아진다고 단정할 수는 없으므로[7] 이들을 수학적 이론으로 구체화시켜 분석적 모형을 만들고 여러 가지 다양한 작업 부하를 사용한 시뮬레이션 을 통해 비교 연구해 볼 수 있다.

참고 문헌

- [1] S. Chen and D. Towsley, "A Performance Evaluation of RAID Architectures," IEEE Transactions on Computers, Vol.4, No.10, pp.1116-1129, 1996.
- [2] A. K. Sahai, " Performance Aspects of RAID Architectures," IEEE International Conference on Performance Computing and Communications, pp.321-327, 1997.
- [3] J. Gray and R. Shenoy, "Rules of Thumb in Data Engineering," Technical Report MS-TR-99-100, MicroSoft Research, 2000.
- [4] 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.
- [5] 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.
- [6] Anujan Varma and Quinn Jacobson, "Destage Algorithms for Disk Arrays with Nonvolatile Caches," IEEE Transactions on Computers, Feb. 1998.
- [7] R. Karedla, J. S. Love, B. G. Wherry, "Caching Strategies to Imporve Disk System Performance," IEEE Computer, pp. 38-46, Mar. 1994.