

## 경량의 재사용 거리 측정 기법을 이용한 디스크 캐싱 기법

손영재<sup>0</sup>, 정석현\*, 길진욱\*, 강민재\*\*, 노동진\*

<sup>0</sup>숭실대학교 융합소프트웨어학과,

\*숭실대학교 지능시스템학과,

\*\*숭실대학교 융합특성화자유전공학부

e-mail: karit@ssu.ac.kr<sup>0</sup>, {beewol, gil6399, dnoh}@ssu.ac.kr\*, minjaekang@ssu.ac.kr\*\*

## Disk Caching Scheme Using Lightweight Reuse Distance Measurement Scheme

Youngjae Son<sup>0</sup>, Seok Hyun Cheong\*, Gun Wook Gil\*, Minjae Kang\*\*, Dong Kun Noh\*

<sup>0</sup>Dept. of Software Convergence, Soong-sil University,

\*Dept. of Intelligent Systems, Soong-sil University,

\*\*School of Convergence Specialization, Soong-sil University

### ● 요약 ●

응용프로그램의 응답성을 향상시키기 위해서는 저장장치 시스템의 데이터 처리 능력이 중요하다. 한편, 차세대 메모리(NVDIMM)는 DRAM과 SSD 중간 정도의 성능 특성과 저장 용량을 갖는다. NVDIMM을 저장장치의 캐시로 사용함으로써 메모리와 저장장치의 격차는 많이 줄게 된다. 본 논문에서는 경량의 재사용 거리 측정 기법을 이용하여 효율적으로 디스크를 캐싱하는 기법을 제안한다. 제안 기법은 경량의 재사용 거리 측정 기법을 바탕으로 계산된 CFD(Computational Fluid Dynamics)값에 따라 디스크 캐시에 해당 데이터 적재 여부를 결정한다. 결과적으로 제안 기법을 적용한 디스크 캐시를 운용함에 따라 캐시의 히트율을 향상시켰다.

**키워드:** 재사용거리(reuse distance), 비휘발성 메모리 모듈(NVDIMM), 디스크 캐시(disk cache)

### I. Introduction

현대의 컴퓨터 시스템에서는 다량의 데이터를 빠르게 처리하여 사용자의 대기시간을 최소화하는 것이 중요하다. 이때 저장장치의 데이터 처리 능력은 응용 프로그램의 응답성에 큰 부분을 차지하고 있다.

한편, SSD는 바이트 단위의 접근이 되지 않아 CPU에서 데이터를 처리하려 할 때 필요 이상의 읽기/쓰기가 실행되며, 쓰기 횟수에도 제약이 존재한다. 그 때문에 일부 고성능 시스템에선 DRAM을 저장장치로 사용하고 있으나, 휘발성과 비용적인 측면 때문에 대용량 저장장치로 적합하지 않다. 반면에 차세대 메모리는 바이트 단위의 접근이 가능하며, 비휘발성의 특징을 갖고, PRAM과 Optane(3D XPoint) 등은 DRAM과 SSD 중간 정도의 성능과 저장 용량을 갖는다[1]. 그러므로 차세대 메모리를 디스크 캐시로 사용하면 저장장치와 메모리의 격차를 줄일 수 있다. 주목할 점은 디스크 캐시의 크기가 워드크드의 워킹 셋 크기를 커버할 수 있는지에 따라 캐시의 성능에 큰 영향을 주기 때문에, 어떤 데이터를 적재할지가 중요하다. 이에, 제안 기법에서는 데이터의 재사용 거리를 기준으로 캐싱할 데이터를 결정한다.

한편, 기존의 재사용 거리의 연구는 데이터를 예측하여 적절한 캐시 사이즈를 도출하는 방식으로 사용되었다[2]. 하지만 현실적으로 캐시의 크기는 대부분 고정되어 있기 때문에 고정된 캐시 크기를 사용하는 환경에서 재사용 거리를 분석하고, 어떤 데이터를 캐시에 적재할 것인지 판단하는 것이 더 중요하다.

본 논문에서는 경량의 재사용 거리 측정 기법과 이를 활용한 디스크 캐싱 기법을 제안한다. 재사용 거리가 짧아 재접근 될 가능성이 높은 데이터를 디스크 캐시에 올리며, 캐싱 되어도 재접근 될 가능성이 낮은 데이터는 캐시에 올리지 않고 우회한다.

### II. The Proposed Scheme

본 장에서는 경량의 재사용 거리 측정 기법 기반 디스크 캐싱 기법에 대해서 설명한다.

1. Reuse Distance

기존의 재사용 거리 측정 기법은 데이터 접근 기록 재탐색을 통해 측정하기 때문에 부하가 크다. 반면에 경량의 재사용 거리 측정 기법은 데이터마다 기록되어있는 마지막 접근 시점을 통해 측정되기 때문에, 상대적으로 작은 부하로 측정이 가능하다.

2. Disk caching based on Reuse Distance

Fig. 1는 제안 기법의 전반적인 동작 과정이다. 제안 기법은 데이터가 접근될 때마다 각 데이터의 재사용 거리를 측정한다. 측정된 재사용 거리에 따라 데이터 접근 빈도를 조정하여, 접근된 데이터를 디스크 캐시에 적재할지를 결정한다.

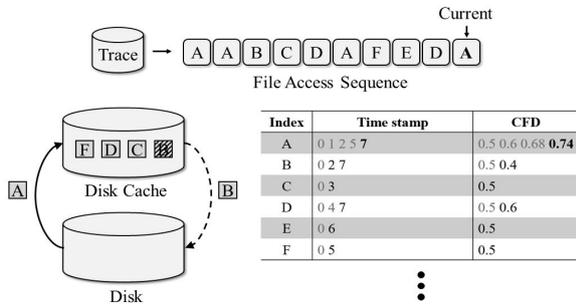


Fig. 1. Overview of the proposed scheme

제안 기법은 데이터가 접근될 때마다 데이터 접근 빈도를 나타내는  $CFD_i$ 와  $Timestamp_i$ 를 갱신한다.

여기서  $ReuseDistance$ 는 Equation 1과 같이 나타낼 수 있다.  $History$ 는 모든 데이터에 대해 접근된 횟수이고,  $Timestamp_i$ 는 데이터  $i$ 가 접근되었을 시점을 나타내는 값이다.

$$ReuseDistance = History - Timestamp_i \quad (1)$$

측정된  $ReuseDistance$ 가 Equation 2를 만족하면, 해당 데이터는 다시 접근될 가능성이 낮은 데이터를 뜻한다. 그러므로  $CFD_i$ 를 낮추어 캐시에 적재될 확률을 줄인다.

$$ReuseDistance > CacheSize \quad (2)$$

반대로, Equation 3을 만족하면 해당 데이터는 캐시에 적재하기에 충분히 가치 있는 데이터이다. 그러므로  $CFD_i$ 를 높여 캐시에 적재될 확률을 높여준다.

$$ReuseDistance \leq CacheSize \quad (3)$$

$ReuseDistance$ 에 의해 조정된  $CFD_i$ 가 Equation 4를 만족하면, 접근이 빈번한 데이터로 판단하여 디스크 캐시에 적재시킨다.

$$CFD_i > Threshold \quad (4)$$

III. Experiments

본 연구에서 제안한 기법의 성능 분석을 시뮬레이션으로 진행하였다. 워크로드는 OLTP 어플리케이션 I/O의 WebSearch를 연구에 맞게 수정하였다[3]. 그리고 디스크 캐시 사이즈는 128GB로 고정하여 제안 기법을 적용한 결과 히트율이 약 11.4% 향상되었다.

IV. Conclusions

본 논문에서는 디스크 캐시에 경량의 재사용 거리 측정 기법을 바탕으로 CFD값에 따라 데이터 적재 여부를 판단한다. CFD값이 작을 경우, 캐시에서 재접근 될 가능성이 낮다. 따라서 CFD값이 큰 데이터를 캐시에 적재하는 기법을 제안하였다. 본 기법을 적용한 결과 캐시 히트율이 향상됨을 알 수 있었다.

ACKNOWLEDGEMENT

본 연구는 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. NRF-2018R1D1A1B07045972).

REFERENCES

[1] Arthur Sainio. 2016. NVDIMM: Changes are Here So What's Next?. In In-Memory Computing Summit 2016.  
 [2] C. Ding and Y. Zhong. "Reuse distance analysis," Technical Report TR741, Dept. of Computer Science, University of Rochester, 2001.  
 [3] UMass TraceRepository, <http://traces.cs.umass.edu/index.php/Storage/Storage>.