JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 7, pp. 9-16, July 2022
https://doi.org/10.9708/jksci.2022.27.07.009

Designing a low-power L1 cache system using aggressive data of frequent reference patterns

Bo-Sung Jung*, Jung-Hoon Lee*

*University Lecturer, Dept. of Control&Instrument Engineering, Gyeongsang National University, Jinju, Korea
*Professor, Dept. of Control&Instrument Engineering, Gyeongsang National University, Jinju, Korea

[Abstract]

Today, with the advent of the 4th industrial revolution, IoT (Internet of Things) systems are advancing rapidly. For this reason, a various application with high-performance and large-capacity are emerging. Therefore, there is a need for low-power and high-performance memory for computing systems with these applications. In this paper, we propose an effective structure for the L1 cache memory, which consumes the most energy in the computing system. The proposed cache system is largely composed of two parts, the L1 main cache and the buffer cache. The main cache is 2 banks, and each bank consists of a 2-way set association. When the L1 cache hits, the data is copied into buffer cache according to the proposed algorithm. According to simulation, the proposed L1 cache system improved the performance of energy delay products by about 65% compared to the existing 4-way set associative cache memory.

▸ **Key words**: low-power system, L1 cache memory, Memory Characteristics, buffer system, replacement policy

[요 약]

오늘날, 4차산업혁명의 도래와 함께 사물인터넷(Internet of Things (IoT)) 시스템이 빠르게 발전하고 있다. 이러한 이유로, 고성능 및 대용량의 다양한 애플리케이션이 등장하고 있다. 따라서, 이러한 애플리케이션을 가지는 컴퓨팅 시스템을 위한 저전력 및 고성능 메모리가 필요하다. 본 논문에서는 컴퓨팅 시스템에서 가장 많은 에너지 소비가 발생하는 L1 캐시 메모리에 대한 효과적인 구조를 제안하였다. 제안된 캐시 시스템은 크게 L1 메인 캐시와 버퍼캐시로 구성되어 진다. 메인 캐시는 2-뱅크 시스템으로, 각 뱅크는 2-웨이 연관사상으로 구성된다. L1캐시에서 접근 성공이 발생하면 제안된 알고리즘에 따라 데이터가 버퍼캐시에 복사가 된다. 시뮬레이션 결과에 따르면, 제안된 L1 캐시 시스템은 기존 4웨이 연관사상 캐시 메모리에 비해 에너지-지연에서 약65%의 성능 향상을 보였다.

▸ **주제어**: 저전력 시스템, L1 캐시 메모리, 메모리 특성, 버퍼 시스템, 교체 정책

• First Author: Bo-Sung Jung, Corresponding Author: Jung-Hoon Lee
*Bo-Sung Jung (blueking80@gnu.ac.kr), Dept. of Control&Instrument Engineering, Gyeongsang National University
*Jung-Hoon Lee (leejh@gnu.ac.kr), Dept. of Control&Instrument Engineering, Gyeongsang National University
• Received: 2022. 04. 26, Revised: 2022. 07. 07, Accepted: 2022. 07. 13.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

오늘날, 4차산업혁명의 등장으로 IoT(Internet of Things) 시스템의 급속한 발전으로 대용량의 다양한 애플리케이션의 등장과 함께 컴퓨팅 시스템을 위한 저전력 및 고성능 메모리가 요구된다[1]. 따라서 현재 컴퓨팅 시스템은 멀티 코어(multi-core) 및 멀티 스레딩(multi-threading)의 제조 기술발전으로 더 작은 트랜지스터의 개발로 CPU 처리 능력이 급속도로 증가하고 있다[2].

메모리 시스템은 대부분의 컴퓨팅 시스템에서 가장 전력이 많이 소모되는 구성 요소 중 하나이다. 디스플레이 장치를 제외하고 메모리 시스템은 데스크탑 및 서버 환경에서 전력 소비 측면에서 프로세서 다음으로 두 번째 높은 에너지를 소비하고 있다[3][4]. 또한 스마트폰 및 태블릿 PC와 같은 모바일 장치 시장은 지속적으로 성장하고 있으며[5], 기업들은 모바일 장치의 속도를 높이기 위해 노력하고 있다. 최근에는 모바일 장치의 옥타 코어 CPU [6]의 벤치마크 점수[7]가 intel core i7- 1065(4cores), intel core 15-9600K(6cores)과 같은 일반 컴퓨터의 CPU와 비슷한 성능을 보이고있다. 따라서 스마트 폰 및 태블릿 PC와 같은 주기적인 전원 공급이 필요한 장치의 경우, 메모리 시스템의 전력 소비가 더욱 중요한 문제로 자리 잡고 있다.

실제, 메모리 시스템은 컴퓨팅 시스템의 다른 구성 요소와 달리 지속적으로 전력을 소비하고 메모리 용량이 계속 증가하기 때문에 대부분의 컴퓨팅 환경에서 에너지 소비는 주요 문제로 여겨지고 있다[8].

현재까지, 캐시 메모리 시스템은 프로세서와 주 메모리 간의 성능 차이로 인한 대기 시간(latency) 및 에너지 소비(Energy Consumption)를 줄이는 효과적인 메커니즘으로 그 중요성이 점점 증가하고 있다[9].

하지만, 캐시 메모리 시스템의 성능과 에너지 소비는 각 애플리케이션 및 단일 애플리케이션의 단계에 따라 크게 달라진다. 캐시 메모리가 너무 크면 성능향상을 이룰 수 있지만, 높은 에너지 소비를 가져오며, 반대로 에너지 소비를 줄이기 위해 작은 용량의 캐시 메모리를 사용하면 빈번한 블록 교체로 성능저하의 원인이 된다. 이에, 현재까지 캐시 메모리의 성능향상과 에너지 소비를 줄이기 위해 버퍼 시스템 혹은 비휘발성 메모리를 이용한 하이브리드 등 많은 연구가 이루어지고 있다[9][13][14][15].

본 논문에서는 캐시 시스템에서 가장 빈번하게 발생하는 L1 캐시 메모리의 에너지 소비를 줄이는 효과적인 구조를 제안하였다. 제안된 캐시 메모리 시스템은 L1 주 캐시와 버퍼캐시로 크게 2부분으로 구성된다. 주 캐시 메모리 시스템은 2뱅크로 각 뱅크는 2-웨이 집합 연관사상으로 구성된다. 그리고 각 뱅크는 하나의 버퍼캐시를 가진다. 그리고 버퍼캐시는 효과적으로 데이터 저장을 위해 완전연관 사상(fully set-associative)으로 구성되며, L1 주 캐시로부터 접근 성공이 발생하면, 본 논문에서 제안된 접근 알고리즘으로 데이터를 복사하고 저장하게 된다.

II. Preliminaries

1. Related works

1.1 Cache Memory

오늘날, 캐시 메모리는 프로세서와 메인 메모리 사이의 접근에 대한 병목현상을 줄이는 가장 대표적인 메커니즘을 자리 잡고 있다. 일반적으로 캐시 메모리는 빠른 접근시간과 데이터 손실을 줄이기 위해 SRAM이 사용되며, 이는 고속의 프로세서와 저속의 메인 메모리 사이의 속도 차이를 개선함으로 컴퓨팅 시스템의 성능에 큰 영향을 주고 있다.

또한 고성능 컴퓨팅 시스템에서 메인 메모리 시스템이 차지하는 에너지 소비는 전체의 약 40%을 소비할 정도로 높은 에너지 소비를 가진다[16][17]. 캐시 메모리 시스템은 메인 메모리의 접근을 효과적으로 줄임으로 구동 시간의 성능을 효과적으로 개선 할 뿐 아니라 전체 시스템 에너지 소비를 줄일 수 있는 효과적인 메커니즘이다.

현재, 4차산업혁명의 등장과 함께 개인용 IoT를 위한 대용량의 다양한 애플리케이션이 등장하고 있다. 결과적으로, 컴퓨팅 시스템을 위한 고성능 및 저전력에 관한 연구가 지속적으로 이루어지고 있다.

현재, 캐시 메모리의 성능 및 에너지 소비를 위한 효과적인 방법으로 대표적으로 버퍼 시스템 운용과 데이터 교체 정책이 대표적이다. 이와 더불어 차세대 메모리로 주목받고 있는 STT-RAM(Spin Torque Transfer RAM), M-RAM(Magnetic RAM)등과 같은 비휘발성 메모리를 함께 운용하는 하이브리드 메모리 시스템 역시 현재 캐시 메모리의 성능과 에너지 소비를 줄이기 위해 급속히 연구가 진행되고 있다.

1.2 Related work for low-power cache memory

Pack[10]은 현재 애플리케이션에 대한 기존 캐시 메모리의 교체 정책의 실효성을 재평가하였다.

Jo[11]는 기존 4-웨이 집합 연관사상 캐시 메모리의 선택적 접근을 위해 Look up 테이블을 이용하여, 선택적 메모리 접근으로 에너지 소비를 줄였다. 하지만, 캐시 메모

리 접근을 위해 Look up 테이블의 접근과 접근 실패시 추가적인 지연시간 및 에너지 소비를 발생시킬 수 있다.

Navarro[12]는 L1 데이터 캐시와 희생 캐시를 최상의 조합을 위해 휴리스틱 기법(Heuristics)을 제안하였다. 이 연구에서는 휴리스틱 기법을 통해 캐시 메모리에 효과적인 희생 캐시를 추가하여 프로세스의 유연성과 캐시 메모리의 에너지 소비를 줄였다. 하지만, 휴리스틱을 위해 희생 캐시 메모리의 캐시 플러시에 주의해야 하는 단점과 종료시 그 결과를 하위 캐시 혹은 메모리에 쓰기 작업이 필요 함으로 추가적인 시간과 에너지 소비가 요구된다.

Lee[8]는 전체 메모리 시스템의 공간을 고려하여 버퍼 캐시에 대해 접근 가능한 데이터를 저장하는 방법으로 기존 메모리 시스템의 에너지 소비를 효과적으로 줄였다. 하지만, 전체 메모리 시스템을 고려한다는 것은 추가적인 접근시간을 요구하게 된다.

Jung[9], Palangappa[13], Ahn[14], Imani[15]의 연구는 현재 비휘발성, 저전력 그리고 SRAM과 비슷한 접근 속도를 보장하는 차세대 메모리로 주목받고 있는 STT-RAM 혹은 M-RAM등을 이용하여 저전력, 고성능 하이브리드 캐시 메모리 시스템을 제안하였다.

Jung[9]의 경우, 기본적으로 N-웨이 집합 연관사상 캐시의 쓰기 연산과 읽기 연산을 고려하여 기존 SRAM과 차세대 메모리(STT-RAM)의 효과적인 구성 비율을 제안하였다. 하지만, Jung[9]이 제안한 캐시 메모리 시스템은 L2 캐시로 L1 캐시의 접근 빈도가 낮다. 따라서 접근 빈도가 높은 L1 캐시에 적합하지 못한 구조이다.

Palangappa[13]는 쓰기 연산이 발생할 데이터를 예측하여 차세대 메모리인 STT-RAM의 쓰기 연산에 대한 단점을 SRAM에 대체하는 방법으로 에너지 소비 및 쓰기 연산 시간을 단축하였다. 하지만, 데이터의 특성을 고려한다면, 이러한 쓰기 연산 데이터를 예측하는 것은 적용되는 애플리케이션 마다 다른 특성을 보이기 때문에 정확도가 매우 낮아진다.

Ahn[14]은 누설 전력이 높은 L2 캐시에 STT-RAM을 응용하여 전체 에너지 소비를 줄였으며, Imani[15]은 역시 STT-RAM의 단점인 쓰기 연산에 대한 접근 값에 의해 SRAM과 혼용되는 하이브리드 메모리를 제안하였다.

실제 비휘발성의 차세대 메모리와 기존 SRAM을 혼용한 캐시 메모리 시스템 구조는 저전력에 효과적이다. 하지만, 비휘발성 차세대 메모리는 쓰기 횟수 제한 및 쓰기 연산에 대한 높은 에너지 소비 및 긴 수행 시간을 가진다는 단점이 있다. 즉 위에서 제안된 하이브리드 캐시 메모리 시스템은 빈번한 접근이 발생하는 L1 캐시 시스템에 적합하지 못한 구조이다.

III. The Proposed Scheme

1. Proposal Motives and Methods

캐시 메모리 시스템은 컴퓨팅 시스템의 프로세서와 메모리간 접근시간과 에너지 소비를 줄일 수 있는 효과적인 메커니즘이다. 또한, 현재 스마트폰과 같은 고성능 개인 단말기의 등장과 함께 대용량의 다양한 애플리케이션의 등장으로 컴퓨팅 시스템의 에너지 소비는 더욱 중요한 문제로 인식 되어지고 있다.

캐시 메모리 시스템은 일반적으로 작은 용량의 L1 캐시와 함께 L2, L3 캐시 메모리로 구성된다. 작은 용량을 가지는 L1 캐시는 L2 및 L3 캐시 메모리에 비해 빈번한 접근이 이루어진다. 따라서 캐시 메모리 시스템에서 에너지 소비를 줄이는 가장 효과적인 방법은 빈번하게 접근이 발생하는 L1 캐시의 에너지 소비를 줄이는 것이 효과적인 방법이다. 따라서 본 논문에서는 컴퓨팅 시스템의 에너지 소비를 줄이기 위해서 새로운 저전력 L1 캐시를 제안하였다.

본 논문에서 제안된 저전력 L1 캐시는 효과적인 에너지 소비를 줄이기 위해 크게 완전연관 사상을 가지는 버퍼캐시와 주 캐시 메모리인 L1 캐시로 구성된다. 제안된 버퍼캐시는 접근 성공률이 가장 높은 완전연관 사상으로 구성되며, 에너지 소비를 줄이기 위해 작은 용량을 가진다. 그리고 L1 캐시 시스템은 2개의 뱅크로 구성되며, 각 뱅크는 2-웨이 집합 연관사상(2 way-set associative)로 이루어진다. 또한 L1 캐시의 각 뱅크는 하나의 버퍼캐시를 가진다. 본 논문에서 제안된 L1 캐시 시스템은 그림 1과 같다.

주 캐시에 작은 용량의 버퍼캐시를 사용하는 것은 에너지 소비에 효과적인 방법이지만, 빈번한 데이터 복사는 오히려 에너지 소비를 유발할 수 있다. 따라서 본 논문에서 제안된 L1 캐시 시스템에서는 주 캐시 메모리에서 빈번하게 접근 가능한 데이터를 선별하여 버퍼캐시에 복사하게 된다.

따라서 본 논문에서 참조 가능한 데이터만을 버퍼캐시에 복사하게 된다. 본 논문에서는 이러한 데이터 선택을 위해 L1 캐시의 주 캐시 메모리에서 데이터가 접근 성공한 횟수를 이용하였다. 이를 위해 제안된 L1 캐시에서는 주 캐시 메모리에서 데이터의 접근 성공한 평균 횟수(Hit_Average bit, HA_bit)를 가지는 상태 비트와 버퍼캐시에 데이터 복사가 발생 여부를 나타내는 상태 비트(Copy, C_bit)를 가진다. 그리고 버퍼캐시 역시 HA_bit를 가지며, 버퍼캐시의 HA_bit는 주 캐시 메모리로부터 계승되어 진다.

만약 L1 캐시의 주 캐시 메모리에서 접근 성공이 발생하면, 그 블록의 HA_bit의 값과 버퍼캐시의 HA_bit를 비

교하게 된다. 만약 주 캐시 메모리의 HA_bit값이 버퍼캐시의 블록 중 하나의 값보다 크다면 그 블록은 버퍼캐시로 복사하게 된다. 이때, 주 캐시 메모리의 데이터를 효과적으로 관리하기 위해 접근 성공이 발생시, HA_bit 값이 감소하게 된다. 이러한 동작은 버퍼캐시 역시 동일 하게 작동한다. 하지만, 버퍼캐시로부터 추출된 데이터는(First In First Out, FIFO 동작) 주 캐시 메모리에 HA_bit값을 계승하지 않는다.

그리고 주 캐시 메모리에서 추출되는 값의 HA_bit 값은 하위 캐시 메모리에 저장된다. 만약, 하위 계층으로부터 데이터가 요청되면, 요청된 데이터와 함께 HA_bit 값이 주 캐시 메모리에 저장된다. 그리고 L2 캐시로부터 HA_bit값이 갱신된 것을 나타내기 위해 추가적인 상태 비트(Update bit, U_bit)를 가진다. 만약 L2 캐시로부터 요청된 데이터의 HA_bit값이 존재한다면, L2 캐시로부터 주 캐시에 데이터가 저장되고, U_bit가 '1'로 갱신된다. HA_bit의 값이 존재한다는 것은 이전에 한번은 L1 캐시에 데이터가 요청되었다는 의미하며, 또한 L1 캐시에서 데이터의 수명을 의미하기도 한다. 반면 L2 캐시의 HA_bit값이 '0'이면 L1 캐시의 U_bit는 '0'으로 갱신된다. HA_bit가 '0'의 값은 이전에 한번도 L1 캐시에서 요청되지 못한 데이터를 의미한다. 만약 L1 캐시에서 접근 성공한다면, U_bit가 '1'이면, HA_bit값이 감소한다. 이러한 이유는 HA_bit는 이전 데이터가 L1 캐시에서 접근 수명을 나타내기 때문에, 이를 기준으로 L1 캐시에서의 데이터 수명을 예상할 수 있기 때문이다. 반면, U_bit가 '0'이면, 현재 데이터가 L1 캐시에서 처음 요청된 상태이기 때문에, L1 캐시에서 접근 수명을 측정해야 한다. 따라서 단순히 HA_bit만 증가하게 된다.

그리고 본 논문에서 제안된 버퍼캐시와 주 캐시는 순차적으로 접근이 일어난다. 버퍼캐시는 완전연관 사상 캐시로 접근시 높은 에너지 소비를 발생 된다. 따라서 저전력을 위해 본 논문에서는 버퍼캐시에 접근시 최하위 1비트를 우선 확인하여, 버퍼캐시의 접근 블록을 선택적으로 결정하게 된다. 그리고 주 캐시의 뱅크 설정은 태그(Tag)의 최하위 1비트를 이용하여 주 캐시의 뱅크를 선택하였다.

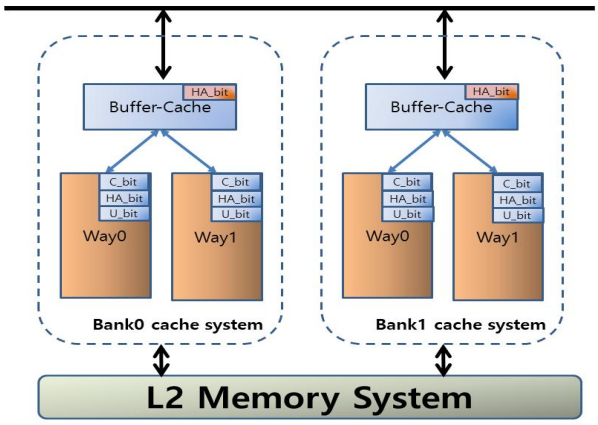

Fig. 1. The proposed L1 cache system.

2. Proposed Cache System Management

2.1 Hit of the main cache memory access :

그림 2는 현재 주 캐시 메모리와 버퍼캐시의 상태를 나타내고 있으며, 뱅크와 2-웨이 집합 연관사상에 대한 동작은 고려하지 않았다. 블록 B의 경우, U_bit가 '1'이므로 L2 캐시로부터 HA_bit가 갱신되었으며, C_bit가 '0'이므로 버퍼캐시에 데이터가 복사되지 않은 상태이다. 블록 D와 A는 U_bit가 '0'으로 현재 주 캐시 메모리에 처음 요청된 블록이다. 그리고 블록 A의 C_bit가 '1'로써, 버퍼캐시에 복사되었으며, 이때 주 캐시의 HA_bit 역시 버퍼캐시에 갱신이 되었다.

만약, CPU에서 블록 A를 요청했다면, 버퍼캐시에서 접근 성공이 발생하며, HA_bit 값이 감소하게 된다. 이때, HA_bit 값은 다음의 식(1)의 평균값으로 구해진다.

L1 main cache

Block	State bits		
....	HA_bit	C_bit	S_bit
B	3	0	1
...	...	...	...
D	3	0	0
A	4	1	0

buffer cache

Block	State bits
....	HA_bit
A	4

Fig. 2. The state of proposed cache system

$$HA\_bit=((HA\_bit*L1\_set) - 1) / L1\_set; \quad (1)$$

(L1_set = L1 주 캐시 메모리 블록 수)

만약, CPU로부터 블록 D가 요청되었다면 블록 D의 HA_bit와 버퍼캐시의 블록 A(버퍼캐시의 블록이 하나라고 가정)의 HA_bit를 비교하게 된다. 블록 A의 HA_bit가 더 크기 때문에, 요청된 블록 D는 버퍼캐시에 저장되지 않는

다. 그리고 블록 D는 U_bit가 '0'이므로, 주 캐시에 처음 요청된 블록이다. 따라서 블록 D의 주 캐시에서 효과적인 접근을 파악하기 위해 HA_bit만 증가하게 된다.

반면, CPU로부터 요청된 L1 캐시의 어떤 데이터가 블록 A보다 HA_bit 값이 크다면, 그 데이터는 버퍼캐시에 복사된다. 이때 버퍼캐시의 블록 A가 쓰기 요청이 발생한 상태라면, 주 캐시 메모리의 블록 A에 쓰기 요청이 발생하지만, HA_bit는 갱신되지 않는다. 즉 버퍼캐시의 HA_bit값은 단순히 주 캐시와의 HA_bit 값의 비교로만 사용된다.

2.2 Miss of the cache memory system:

만약, CPU로부터 요청된 블록이 L1 캐시 시스템에 존재하지 않는다며, L2 캐시로부터 블록을 요청하게 된다. 이때, L2 캐시로부터 요청된 블록은 L1 주 캐시에 저장되며, 만약 L2 캐시에서 HA_bit 값이 존재한다면, 요청된 블록의 U_bit가 '1'로 갱신이 된다. 그렇지 않다면, U_bit는 '0'으로 갱신된다. 예로, 그림 2에서 블록 C가 CPU로부터 요청되었다면(블록 D에 저장된다고 가정), L1 캐시 시스템은 접근 실패가 발생하고 L2 캐시로부터 블록 C를 요청하게 된다. 이때 블록 C의 최하위 태그 비트를 이용하여 제안된 캐시 시스템의 주 캐시의 뱅크를 선택하게 된다. 만약 선택된 뱅크의 그리고 만약 L2 캐시의 블록 C가 '0'이 아닌 HA_bit 값을 가진다면, U_bit가 '1'로 갱신된다. 이때, 블록 D는 L1 캐시에 처음 요청된 데이터이기 때문에 HA_bit 값은 L2 캐시에 저장된다. 즉 제안된 캐시 시스템을 위해 L2 캐시 역시 HA_bit를 위한 상태 비트를 가진다. 그림 3은 CPU로부터 블록 C가 요청된 상태를 나타내며, 이때 블록 C는 L1 캐시 시스템으로부터 한번 이상 요청된 블록으로 HA_bit 값을 가지며, L1 주 캐시의 U_bit가 갱신된다.

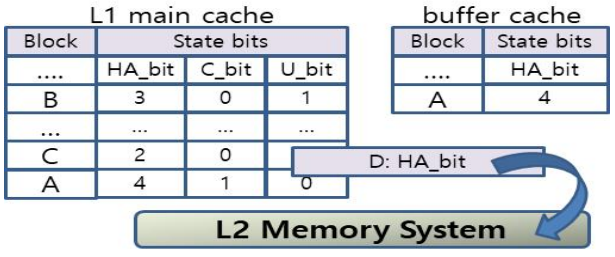

Fig. 3. The state of proposed cache system by access miss

3. Performance evaluation

본 논문에서는 제안된 L1 캐시 시스템의 성능평가를 위하여 미디어 벤치마크의 실행되는 동안 메모리에 접근하는 트레이스 파일을 추출하여 사용하였다. 트레이스 파일은 Ubuntu 시스템에서 valgrind의 Cachegrind[18]를 수정하여 L1 캐시로부터 접근 실패 후 메모리에 접근하는 1억 개의 데이터의 주소를 모니터링을 하였다. 모니터링된 주소는 Visual sutdio에서 본 논문에서 제안된 구조 및 알고리즘과 4-웨이 집합 연관사상 캐시를 구현하여 캐시 성능을 측정하였다. 성능평가를 위한 제안된 저전력 캐시 시스템의 파라미터는 표. 1과 같다.

Table 1. cache system's parameters.

	value
Processor	3.09 Ghz
Buffer cache	32byte line size 128Byte(Full set associative)
L1 cache	32byte line size D:32Kbyte(2bank,2-way)
L2 cache	Private, 8way, 64byte line_size 1Mbyte *8-way

본 논문에서는 제안된 캐시 시스템의 성능을 평가는 오직 데이터 캐시만을 평가하였다. 데이터 캐시는 순차적인 접근이 우월한 명령어 캐시 메모리와 달리 접근 데이터가 불확실하다. 따라서 명령어 캐시 메모리에 비해 높은 접근 실패율과 인하여 높은 에너지 소비를 가진다.

본 논문에서 제안된 캐시 시스템의 성능을 평가를 위해 기존 4-웨이 집합 연관사상 캐시와 Jo[11]을 비교 캐시로 사용하였다. 두 캐시 모두 32Kbyte의 캐시 크기와 32byte의 블록 크기를 가진다. 성능평가를 위해 평균 메모리 접근시간을 사용하였다. 그림 4는 제안된 캐시 메모리 시스템과 비교 캐시들의 평균 메모리 접근시간을 보여주고 있다.

시뮬레이션 결과, 제안된 L1 캐시는 4-웨이 집합 연관사상 캐시 메모리에 비해 약 2%, Jo[11]에 비해 약 6%의 성능향상을 보였다. 실험 결과 제안된 캐시 시스템 및 비교 캐시에서 비슷한 접근 실패율을 보였다. 하지만, Jo[11]의 경우 직접 연관사상 캐시에서 좋은 성능을 보였지만, 이를 위해 2-웨이 집합 연관사상 캐시로부터 교체 동작이 발생으로 높은 추가 접근시간이 발생하였다. 그리고 4-웨이 집한 연관사상 캐시는 가장 좋은 접근 실패율을 보였지만, 데이터 선택을 위한 4개의 블록 접근에서 높은 접근시간을 소비하였다. 반면, 제안된 캐시 시스템을 4-웨이 집합 연관사상 캐시중 2웨이 접근과 완전 집합 연관사상 캐시에 선택적 접근으로 비교 캐시에 비해 낮은 접근시간이 보장되었다.

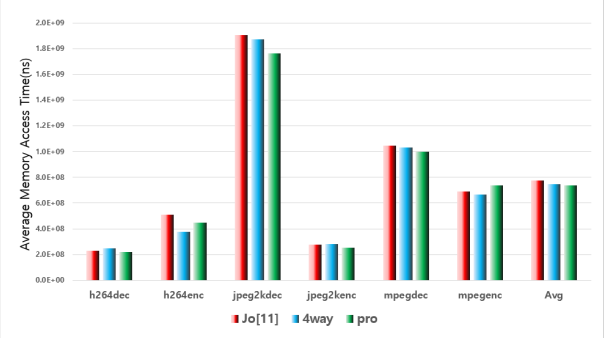

Fig. 4. Average Memory Access Time.

그 이유는 제안된 캐시 메모리 시스템에서 주 캐시는 선택적으로 하나의 16Kbyte 2-웨이 집합 연관사상 캐시에만 접근이 발생하게 된다. 또한 제안된 버퍼캐시에 데이터를 교체가 아닌 복사만 된다. 더욱이 제안된 작은 용량의 버퍼는 선택적 블록에만 접근하게 되고, 시뮬레이션 결과 버퍼 접근시 4개의 블록중 약 50%가 하나의 블록에만 접근이 발생하였다. 실제 시뮬레이션 결과 제안된 버퍼캐시의 데이터 저장 방법과 단순히 주 캐시 메모리에서 접근 성공이 발생시 데이터 복사하는 방법을 비교했을 경우, 제안된 버퍼캐시 알고리즘이 더 효과적인 것을 확인할 수 있었다.

그림 5는 제안된 캐시 메모리 시스템에서 L1 주 캐시로부터 버퍼캐시에 저장되는 데이터 당 평균 접근 성공 횟수를 나타내고 있다.

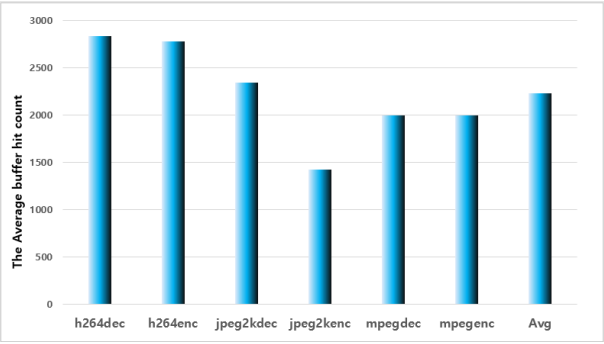

Fig. 5. Number of hit per buffer cache update number

시뮬레이션 결과, 제안된 캐시 시스템의 버퍼는 평균 하나의 데이터가 2000번 이상 접근 성공이 발생하였다. 이는 제안된 캐시 시스템의 버퍼는 하위 메모리에서 한번 이상 요청된 데이터에서 높은 참조 가능성을 가진 데이터만 저장하게 된다. 이에 참조 가능성이 높은 데이터를 빠른 메모리 접근시간과 저전력을 가지는 버퍼에서 높은 접근율을 이룰 수 있었다. 결과적으로 제안된 버퍼캐시에서 효과적으로 데이터를 관리 할 수 있었다.

본 논문의 주목적은 현재 대용량의 다양한 애플리케이션의 등장으로 인한 컴퓨팅 시스템을 위한 저전력 L1 캐시 메모리 시스템을 설계하는 것이다. 따라서 우리는 제안된 캐시 메모리 시스템과 4-웨이 집합 연관사상 캐시에 대한 에너지 소비를 측정하였다. 에너지 소비는 Cacti 6.5를 이용하여 각 캐시의 에너지 소비를 모니터링을 하였다. 모니터링된 값은 본 논문에서 설계된 캐시 시스템의 메모리에 대입하여 전체 에너지 소비를 모니터링을 하였다[19].

그림 6은 캐시 메모리 시스템들의 에너지 소비를 보여주고 있다. 시뮬레이션 결과, 제안된 캐시 메모리 시스템은 기존의 4-웨이 집합 연관사상 캐시에 비해 약 70%의, Jo[11]에 비해 약 25%의 에너지 소비를 줄였다. 이러한 결과는, 앞서 언급한 내용과 같이 제안된 L1 캐시는 작은 버퍼와 뱅크 시스템을 바탕으로 2-웨이 집합 연관사상 캐시에 접근된다. 각 캐시 구조의 시뮬레이션에 따르면, 4-웨이 집합 연관사상 캐시는 버퍼캐시에 비해 약 3배, 제안된 주 캐시 시스템에 비해 2배 정도의 높은 에너지 소비를 보였다. 그리고 Jo[11]는 제안된 캐시 시스템에 비해 약 1.5배의 높은 에너지를 소비하였다. Jo[11]의 경우 낮은 에너지 소비를 하는 직접 연관사상 캐시에 대부분 접근 성공이 발생했지만, 2-웨이 집합 연관사상 캐시와 데이터 교체 동작으로 높은 에너지 소비가 발생하였다. 반면 제안된 캐시 메모리 시스템은 참조 가능성이 높은 선택적 데이터를 작은 용량의 버퍼에서 접근이 발생하였으며, 주 캐시 역시 2-웨이 집합 연관사상 캐시만 접근이 발생한다. 이러한 이유로 제안된 캐시 시스템이 효과적인 에너지 성능향상을 이룰 수 있었다.

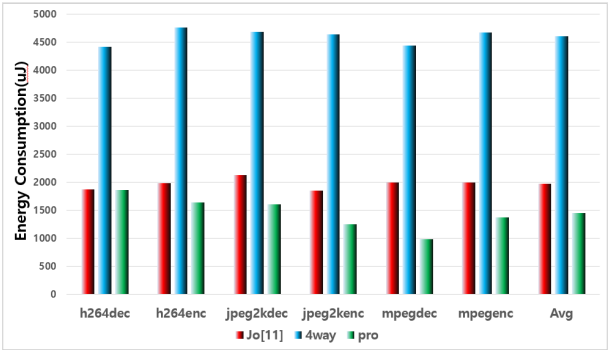

Fig. 6. Energy Consumption(uJ)

Energy-Delay-Product는 에너지 소비*지연시간(평균 메모리 접근시간)을 나타내는 성능지표로써 메모리 시스템의 에너지 소비와 평균 메모리 접근 시간의 trade-off 현상에 대해 효과적으로 평가할 수 있는 지표이다. 따라서

본 논문에서 Energy-Delay-Product로 캐시 시스템의 최종 성능평가를 하였다. 그림 7은 Energy-Delay-Product를 나타내고 있다.

시뮬레이션 결과, 제안된 L1 캐시 시스템은 4-웨이 집합 연관사상 캐시에 비해 약 65%의 성능향상을 이루었다. 또한 Jo[11]과 비해 약 23%의 성능향상을 이루었다. 이는 실제로, 제안된 L1 캐시가 그림 4에서 보듯이 비슷한 메모리 접근시간에도 불구하고, 그림 6의 에너지 소비에서 본 논문에서 제안된 L1 캐시 시스템이 비교 캐시 메모리들에 비해 효과적으로 에너지 소비를 줄였기 때문이다. 결과적으로, 제안된 메모리 시스템의 Energy*delay 역시 좋은 성능을 보였다.

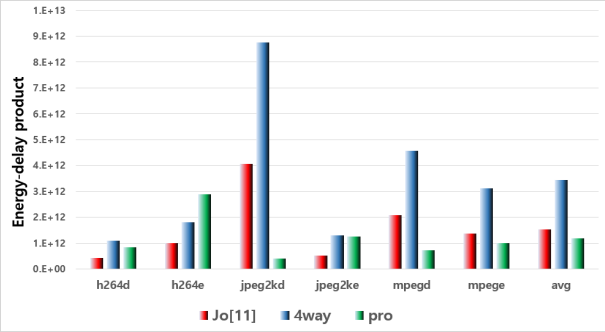

Fig. 7. Energy*Delay Product

IV. Conclusions

오늘날 대용량 애플리케이션의 등장과 IoT 시스템의 급속한 발달로 인한 저전력 컴퓨팅 시스템이 요구된다. 컴퓨팅 시스템의 메모리 시스템은 지속적인 접근과 동작으로 전체 시스템에서 두 번째 높은 에너지 소비를 가진다. 특히 L1 캐시는 작은 용량이지만 L2, L3 캐시에 비해 빈번한 접근으로 에너지 소비가 높다.

본 논문에서는 저전력 컴퓨팅 시스템을 위해 저전력 L1 캐시 시스템을 제안하였다. 제안된 L1 캐시 시스템은 버퍼캐시와 L1 주 캐시 메모리로 구성되며, 버퍼캐시는 아주 작은 용량으로 접근 성능을 보장하기 위해 완전 연관사상 캐시로 구성되며, L1 주 캐시 메모리는 저전력을 위해 2뱅크 시스템으로 구성된다. 또한 L1 주 캐시 메모리의 각 뱅크는 메모리의 성능을 보장하기 위해 2-웨이 집합 연관사상 캐시로 구성된다. 또한 L1 주 캐시 메모리의 각 뱅크는 하나의 버퍼캐시를 가진다.

제안된 L1 캐시 메모리 시스템에서 효과적인 버퍼캐시를 운용하기 위해, 본 논문에서는 주 캐시 메모리에서 데이터 접근 성공을 모니터링을 하였으며. 그 결과에 따라 버퍼캐시에 데이터를 운용하였다. 시뮬레이션 결과에 따르면, 단순 접근을 이용한 방법보다 제안된 버퍼캐시 운용 방법이 더 효과적인 것을 확인하였다.

에너지 소비를 위한 미디어 벤치마크 시뮬레이션 결과, 제안된 L1 캐시 메모리 시스템은 기존 4-웨이 집합 연관사상 캐시 메모리에 비해 약 70%, J0[11]에 비해 약 25%의 에너지 감소를 이루었다. 또한 평균 메모리 접근 시간과 에너지 소비에 대한 성능지표인 Energy*delay Product에서도 4-웨이 집합 연관사상 캐시에 비해 65%, Jo[11]에 비해 23%의 성능향상을 보였다. 이는 제안된 L1 캐시가 미디어 벤치마크에서 에너지 소비를 줄이는 효과적인 구조라는 것을 의미한다.

REFERENCES

[1] T.J. Pack, W.Y. Jang, “Large-Scale Last-Level Cache Design Based on Parallel TLC STT-MRAM,” Jounal of KIIT, Vol. 15, No, 12, pp.77-89, 2017.

[2] A. Valero, J. Sahuquillo, S. Petit, P. López, J. Duato, “Design of Hybrid Second-Level Caches,” IEEE Trans.Comput. Vol. 64, Issue. 7, pp.1884-1897, 2015

[3] C. Lefurgy, K. Rajamani, F. et. al., “Energy Management for Commercial Servers,” Computer, Dec. Vol. 36, No. 12, pp.39-48, 2003.

[4] A. Malik, B. Moyer, D. Cermak, “A low power unified cache architecture providing power and performance flexibility,” Proceedings of the 2000 International Symposium on Low Power Electronics and Design, July, pp.241-243, 2000

[5] A. Gutierrez, R. G. Dreslinski, T. F. Wenisch, T. Mudge, A. Saidi, C. Emmons, N. Paver, "Full-system analysis and characterization of interactive Smartphone applications", IEEE International Symposium on Workload Characterization, Nov., No. 11, pp.81-90, 2011.

[6] Wikipedia, Apple A12X Processor, https://en.wikipedia.org.

[7] Geekbench, https://browser.geekbench.com

[8] L. Min, S. Euiseong, J.W. Lee, etc “PABC: Power-Aware Buffer Cache Management for Low Power Consumption,” IEEE TRANSACTION on Computers, Vol. 56, No. 4, pp.488-501, 2007

[9] B.S. Jung, J.H. Lee, “Way-SEt Associative Management for Low Power Hybrid L2 Cache Memroy,” IEMEK J. Embde. Syst. Vol. 13, No. 3, pp.125-131, 2018.

[10] N.E. Pack, J.W. Kimg, T.S. Jeong, “Cache Memory and Replacement Algorithm Implementation and Performance Comparison,” Journal of The Korea Society of Computer +-and Information, Vol. 25, No. 3, pp.11-17, 2020.

[11] O.R. Jo, J.H. Jung, " Design of Cache Memory System for Next Generation CPU," IEMEK J. Embde. Syst. Vol. 11, No. 6, pp.353-359, 2016.

[12] O. Navarro, T. Leiding, M. Hubner, "Configurable cache tuning with a victim cache," 2015 10th International Symposium on ReCoSoC, July, pp.1-6, 2015.

[13] J.W. Ahn, S.G. Yoo, K.Y. Choi, "Prediction Hybrid Cache: An Energy-Efficient STT-RAM Cache Architecture," IEEE TRANSACTIONS ON COMPUTERS, Vol. 64, No. 3, pp.940-951, 2015.

[14] S.P. Pack, S. Gupta, N. Mojumder, et al., "Future cache design using STT-RAMs for improved energy efficiency: Devices, circuits and architecture," in Proc. Design automat, Conf., pp.492-497, 2012.

[15] M. Imani, S. Patil, T. Rosing, "Low Power Data-Aware STT-RAM based Hybrid Cache Architecture," 17th International Symposium on Quality Electronic Design, pp. 88-94, 2016.

[16] C. Lefurgy, K. Rajamani, F. Rawson, W. Felter, M. Kistler,and T. W. Keller. Energy management for commercial servers. Computer, 2003.

[17] G. Dhiman, R. Ayoub, T. Rosing, "PDRAM: A Hybrid PRAM and DRAM Main Memory System," Proceedings of Design Automation Conference, pp. 664-669, 2009.

[18] N. Nethercote and J. Seward, "Valgrind: A Program Supervision Framwork," Elsevier Electonc Notes in Theoretical Computer Science, Vol. 89, No. 2, pp.44-66, 2003.

[19] N. Muralimanohar, R,.BalasubRAMonian, and N. P. Jouppi, "CACTI 6.0: A tool to model large caches," HP Lab., Palo Alto, Ca, USA, Tech. Rep. HPL-2009-85, 2009.

Authors

Bo-Sung Jung received M.S. and Ph.D. degrees from GyeongSang National University in 2008 and 2018 respectively. His research interests include advance computer architecture and next generation memories system, and Non-volatile memory.

Jung-Hoon Lee received the M.S. and Ph.D. degree in Computer Science from Yonsei University, Seoul, Korea, in 2001 and 2004, respectively. He is currently a professor in GyeongSang National University (ERI).

His research interests include advanced computer architectures and next flash memory.