PMPro: 차세대 비 휘발성 메모리 기반 저장장치를 위한 통합된 스토리지 I/O 분석도구

한상욱○, 김지홍
서울대학교 컴퓨터공학부
shanehahn@davinci.snu.ac.kr, jihong@davinci.snu.ac.kr

PMPro: A Fully-Integrated Storage I/O Profiler for Next-Generation Non-Volatile Memory-Based Storage

Sangwook Shane Hahn○, Jihong Kim
Department of Computer Science and Engineering, Seoul National University

요 약

최근 비 휘발성 메모리 소자의 비약적인 기술 발전으로 DRAM의 응답 시간에 가깝고 바이트 단위로 데이터 입출력이 가능한 새로운 차세대 비 휘발성 메모리들이 개발되어 (3D XPoint, Z-NAND 등) 이를 활용한 메인 메모리 및 저장장치에 대한 연구가 활발히 이루어지고 있다. 차세대 비 휘발성 메모리의 적용을 위해 기존의 컴퓨터 시스템에 많은 변화가 발생하고 있는데 기존의 도구들은 블록 기반 저장장치의 동작에 맞추어 있어서 효과적인 스토리지 I/O 분석이 어려운 상황이다. 본 논문에서는 차세대 비 휘발성 메모리 기반 저장장치를 사용하는 리눅스 운영체제를 대상으로 전체 스토리지 I/O 계층에서 각 계층의 I/O 동작을 측정하고, 이를 바탕으로 계층 통합적 분석을 통하여 각 계층간 I/O 동작을 연결하여 스토리지 I/O 성능 분석이 가능한 도구인 Persistent Memory I/O Profiler (PMPro)를 소개한다. 벤치마크 프로그램들을 사용한 검증 실험을 통하여 PMPro가 다양한 상황에서 1% 미만의 동작부하를 가지면서도 유용한 정보들을 정확히 분석할 수 있음을 확인하였다.

1. 서 론

일반적인 컴퓨팅 시스템은 중앙처리장치, 캐시, 메인 메모리, 저장장치의 구조를 이루고 있으며, 메인 메모리로는 빠른 응답속도와 바이트 단위 데이터 입출력이 가능한 DRAM과 같은 휘발성 메모리를 사용하고 저장장치로는 큰 크기의 용량을 제공하고 블록 단위로 데이터 입출력 하는 HDD, SSD와 같은 비 휘발성 메모리를 사용한다. 하지만 최근 비 휘발성 메모리 기술의 발전으로 DRAM처럼 응답시간이 짧고 바이트 단위로 데이터 입출력이 가능한 3D XPoint[1], Z-NAND[2]와 같은 새로운 메모리 소자들이 개발되어 이를 활용한 메인 메모리 및 저장장치에 대한 연구가 활발히 이루어지고 있다.

짧은 응답시간과 바이트 단위 입출력을 지원하는 차세대 비 휘발성 메모리를 메인 메모리 혹은 저장장치로 사용하기 위해서는 블록 기반 저장장치를 위해 고안되어 있는 기존의 운영체제의 구현에 변경이 필요하다. 이를 위해 리눅스 운영체제는 최근 수년간 기존의 SATA 인터페이스 기반의 블록 단위 데이터 입출력 방식 외에 NVMe 인터페이스와 DIMM 인터페이스와 같은 새로운 인터페이스를 활용한 데이터 입출력 방식이 추가되었다. 하지만 기존의 스토리지 I/O 분석도구들은 기존의 블록 기반 저장장치를 위한 I/O 계층을 위해 구현되어 있어[3,4], 새로운 데이터 입출력 방식을 거치는 스토리지 I/O에 대한 분석에는 그 한계가 있다. 또한 기존의 스토리지 I/O 분석 도구들은 특정 계층에서 얻을 수 있는 정보에 국한되어 있기 때문에, 응용 프로그램 수준의 최상위 계층에서 저장장치를 관리하는 최하위 저장장치 계층까지 스토리지 I/O에 관여하는 모든 스토리지 I/O 계층을 수직적으로 아우르는 통합된 분석에 한계가 존재한다.

본 논문에서는 차세대 비 휘발성 메모리 기반 저장장치를 사용하기 위해 고안된 새로운 인터페이스를 사용하는 데이터 입출력 방식이 적용된 리눅스 운영체제에서 스토리지 I/O의 최상위 계층인 응용 프로그램에서 최하위 계층인 저장장치 드라이버까지 스토리지 I/O에 관여하는 모든 스토리지 I/O 계층에서 실시간으로 스토리지 I/O 동작을 관찰 및 분석하는 통합된 스토리지 I/O 분석 도구인 PMPro를 소개한다. PMPro는 기존 스토리지 I/O 계층의 동작을 수정하지 않고, 각 계층에서 수집된 정보만을 바탕으로 계층 간 연결고리를 이용하여 각 계층의 스토리지 I/O를 이어서 계층 통합 분석을 가능하게 한다. 또한 PMPro는 짧은 응답시간을 제공하는 차세대 비 휘발성 메모리 기반 저장장치의 성능에 영향을 주지 않도록 고안되어 다양한 실험 환경에서 1% 미만의 무시할만한 동작 부하가 발생하도록 구현되었다.

본 논문의 구성은 다음과 같다. 2장에서는 차세대 비 휘발성 메모리 기반 저장장치를 사용하는 리눅스 운영체제의 스토리지 I/O 계층에 대해 소개하고, 3장에서는 제안한 통합된 스토리지 I/O 분석도구인 PMPro의 구조와 동작 원리에 대하여 설명한다. 4장에서는 제안한 도구를 사용하여 다양한 시나리오에서 각 스토리지 I/O 계층의 부하를 측정하고, 도구 자체의 동작 부하를 평가한다. 5장에서 기존에 연구된 스토리지 I/O 분석도구들을 소개하고 6장에서 결론을 맺는다.

표 1. 비 휘발성 및 휘발성 메모리의 응답시간

Memory Device	I/O Latency (us)
DRAM	0.2
3D XPoint	7
Z-NAND	12
SSD (NAND Flash Memory)	100
HDD (Magnetic Disk)	100,000

2. 차세대 비 휘발성 메모리 저장장치를 위한 I/O 계층

기존 비 휘발성 메모리로 대표되는 마그네틱 디스크와 낸드 플래시 메모리는 휘발성 메모리 중 가장 널리 쓰이는 DRAM에 비해 매우 긴 응답속도를 보여주기 때문에[5], 해당 메모리들을 메인 메모리나 메인 메모리의 캐시로 사용하려는 시도들은 적었다. 또한 DRAM은 Byte 단위로 데이터 입출력이 가능한데 비해, 기존 비 휘발성 메모리는 블록 단위로만 데이터 입출력이 가능하여, Load, Store 같은 메인 메모리 동작을 수행하기 위해서는 비효율적인 구조를 지녔었다.

하지만 비 휘발성 메모리 기술의 비약적인 발전으로 3D XPoint와 Z-NAND와 같은 차세대 비 휘발성 메모리들이 개발되었고, 이들은 기존 비 휘발성 메모리들보다 월등히 짧은 응답시간과 바이트 단위 데이터 입출력을 지원하는 특징을 지니고 있다. 표 1은 DRAM, 3D XPoint, Z-NAND, SSD, HDD의 응답시간을 나타낸다[2,5]. 표 1에 따르면, 기존 비 휘발성 메모리 기반 저장장치인 SSD, HDD들은 100 ~ 100,000 us의 긴 응답시간을 보여주는 반면, 3D XPoint와 Z-NAND의 응답시간은 각각 7 us와 12 us로써 응답시간이 0.2 us인 DRAM에 보다 더 가까워진 짧은 응답시간을 보여준다.

짧은 응답시간과 바이트 단위 데이터 입출력을 지원하는 차세대 비 휘발성 메모리를 효율적으로 지원하기 위해, 리눅스 운영체제는 수년간 기존 블록 기반 저장장치를 위한 스토리지 I/O 계층을 탈피하여 새로운 스토리지 I/O 계층을 도입하였다. 첫 번째는 높은 대역폭을 보여주는 PCIe 인터페이스를 사용하는 NVMe 인터페이스가 도입되었으며, 두 번째로는 짧은 응답시간을 보여주는 DIMM 인터페이스를 사용하는 NVDIMM 인터페이스가 도입되었다. 그리고 각 인터페이스를 지원하기 위해 그림 1과 같이 Block layer를 우회하거나 NVMe driver가 도입되는 등 스토리지 I/O 계층에 많은 변화가 생겼다. 그러나 이런 스토리지 I/O 계층을 분석하는 도구들은 아직 개발되지 않았으며, 기존의 블록 기반 저장장치를 위한 도구들로는 새로운 데이터 입출력 방식의 스토리지 I/O 동작에 대한 분석에 그 한계가 존재한다.

3. PMPro 설계 및 구현

그림 1는 본 논문에서 제안한 차세대 비 휘발성 메모리 기반 저장장치를 위한 통합된 스토리지 I/O 분석 도구인 PMPro의 전체 구조를 나타낸 그림이다. 리눅스 커널에 구현된 I/O monitor가 각 계층의 스토리지 I/O 동작에 대한 정보를 수집한 뒤, 이를 메인 메모리에 순환 버퍼에 저장하여 PMPro manager에게 전달한다. PMPro manager는 각 계층의

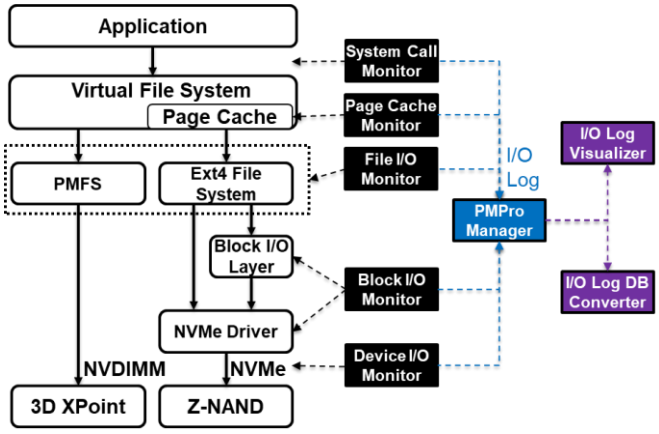

그림 1. PMPro의 전체 구조

스토리지 I/O 정보를 연결하여 특정 스토리지 I/O에 대한 계층 통합적 분석을 가능하게 한다. 또한, PMPro manager는 사용자가 스토리지 I/O에 대한 정보를 다각도에서 분석할 수 있게끔 DB에 입력이 가능한 CSV 파일형식이나 Wave viewer를 사용 가능한 waveform 파일형식으로 출력한다.

I/O Monitor는 크게 5가지 종류의 스토리지 I/O 계층에 구현되었는데, 이는 각 계층에서 얻을 수 있는 스토리지 I/O 정보의 범위에 따라 결정되었다. System call monitor에서는 스토리지 I/O를 유발한 응용 프로그램의 UID, PID, TID 정보와 파일 이름, 크기, 오프셋 정보를 저장한다. Page cache monitor에서는 page cache의 hit과 fault handler의 동작을 분석하며, File I/O monitor에서는 파일 시스템의 정보를 활용하여 파일에 대한 일반적인 정보와 함께 파일의 물리 주소 정보를 바이트 단위로 저장한다. Block I/O monitor과 Device I/O monitor에서는 블록 기반 I/O 구조체로부터 블록 I/O가 데이터 입출력을 수행할 물리 주소 정보를 저장한다. 또한, 각 계층에서 스토리지 I/O 동작에 소요되는 시간도 저장한다. Persistent memory file system (PMFS)와 NVDIMM 인터페이스를 통하여 메모리로 바로 접근하는 경우는 메모리로 I/O 요청을 발생시킨 시간과 완료한 시간을 측정하여 Device I/O 동작으로 간주하였다.

모든 I/O Monitor는 스토리지 I/O 동작에 관여하는 함수들의 파라미터와 파일과 블록 I/O 구조체로부터 스토리지 I/O 정보를 얻는 방식으로 구현되었기 때문에, PMPro의 동작을 위해서 기존 함수들이나 구조체들의 수정이 필요하지 않다.

각 계층에서 얻을 수 있는 스토리지 I/O 정보가 제한적이어서(예를 들면, 디바이스 I/O 구조체와 블록 I/O 구조체로부터 연관된 파일 I/O에 대한 정보를 이끌어내기 힘들다), 특정 스토리지 I/O의 동작을 모든 스토리지 계층에서 통합적으로 분석하기 위해서 PMPro Manager는 각 계층의 스토리지 I/O 정보를 연결하는 작업을 수행한다. PMPro Manager의 디바이스 I/O, 블록 I/O와 파일 I/O에 대한 연결작업은 디바이스 I/O 구조체와 블록 I/O 구조체의 물리주소와 파일 시스템으로부터 취득 가능한 파일이 저장된 물리주소정보를 비교하는 것으로 진행되며, 스토리지 I/O 연결 작업을 통해 응용 프로그램부터 저장장치 드라이버까지 각 계층에서 얻은 정보들 중 특정 스토리지 I/O의 동작을 통합적으로 분석하는 것이 가능해진다.

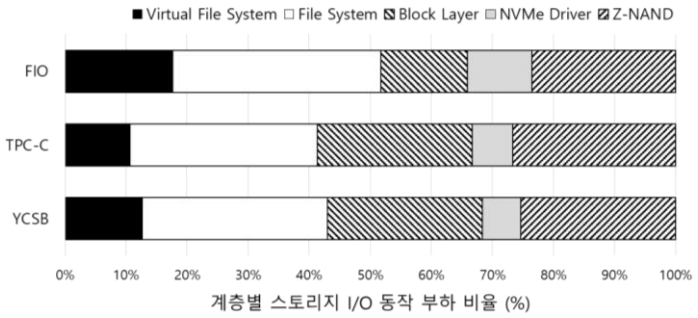

그림 2. Z-NAND에서의 계층별 스토리지 I/O 동작 부하

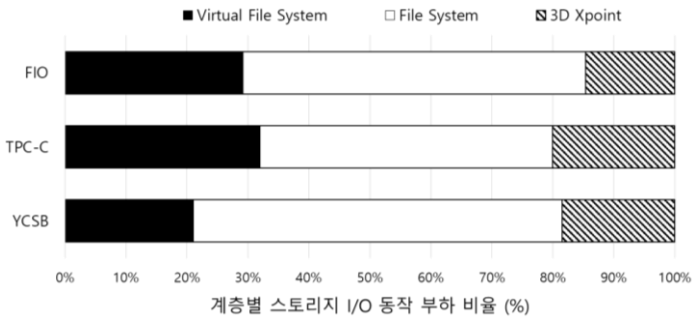

그림 3. 3D XPoint에서의 계층별 스토리지 I/O 동작 부하

4. 실험결과

차세대 비 휘발성 메모리 기반 저장장치를 사용할 때, 현재 구현되어 있는 스토리지 I/O 계층이 저장장치의 성능에 얼마나 영향을 주는지 평가하기 위해, 제안한 스토리지 I/O 분석 도구를 리눅스 커널 버전 4.5 에 구현하여 차세대 비 휘발성 메모리 기반 저장장치를 대상으로 스토리지 I/O 계층에서의 동작 부하를 측정하는 실험을 진행하였다. 3D XPoint와 Z-NAND에 기반한 저장장치는 아직 입수하기 어렵기 때문에, DRAM의 일부를 저장공간으로 설정하여 실험을 진행하였다. 또한 차세대 비 휘발성 메모리 기반 저장장치의 성능을 묘사하기 위해, DRAM에 데이터 입출력을 할 때, 리눅스 커널에서 지원하는 High resolution timer를 사용하여 DRAM으로부터 발생하는 응답시간을 표 1의 3D XPoint와 Z-NAND의 응답시간으로 설정하여 평가하였다.

차세대 비 휘발성 메모리 기반 저장장치를 사용하는 컴퓨팅 시스템의 스토리지 I/O 계층별 부하를 측정하기 위해, 저장장치의 성능을 평가할 때 널리 사용되는 벤치마크 중, FIO[6], TPC-C[7] 그리고 YCSB[8] 벤치마크를 수행하면서 PMPro를 통해 계층별 스토리지 I/O 동작부하를 측정하였다. FIO 벤치마크로는 4 KB random read를 수행하였다.

그림 2는 Z-NAND에 Ext4 파일시스템과 NVMe 인터페이스를 통해 데이터 입출력을 하는 경우, 계층별 스토리지 I/O 동작부하를 측정한 결과이며, 대부분의 경우 파일 시스템과 Block layer에서의 부하가 평균 60% 이상 차지하는 것을 확인할 수 있었다. 그림 3은 3D XPoint에 PMFS 파일시스템과 NVDIMM 인터페이스를 통해 데이터 입출력을 하는 경우, 계층별 스토리지 I/O 동작부하를 측정한 결과이며, 대부분 파일 시스템에서의 부하가 평균 50% 이상 차지하는 것을 확인할 수 있다. PMFS 파일시스템의 경우, Block layer를 거치지 않고 메모리에 데이터 입출력을 직접 수행하기 때문에, Block layer의 동작부하는 존재하지 않았다.

표 2는 바이트 단위로 데이터 입출력이 가능한 차세대 비 휘발성 메모리의 특성으로 인한 실제 데이터 입출력양 차이를 알아보기 위해 벤치마크를 수행하면서 저장장치에 쓰여진 데이터의 총량을 측정한 값을 보여준다. 4 KB 미만의 작은 양의 데이터를 읽거나 쓸 때에도 항상 4 KB 혹은 논리 블록 크기의 배수로 데이터 입출력을 해야 하는 기존의 블록 기반 저장장치 대비 바이트 단위로 데이터 입출력이 가능한 차세대 비 휘발성 메모리 기반 저장장치는 최대 4.24배 더 적은 양의 데이터 쓰기가 발생하는 것을 확인할 수 있었다.

표 2. 저장장치에 인가된 데이터 쓰기양의 비교

Benchmark	*3D XPoint*	*SSD*
TPC-C	169 MB	718 MB
YCSB	27 MB	114 MB

5. 관련 연구

리눅스 커널의 스토리지 I/O를 분석하는 도구는 시스템 콜을 분석하는 strace[3]와 Block layer 계층의 스토리지 I/O를 분석하는 blktrace[4]가 존재한다. 그러나 해당 도구들로는 차세대 비 휘발성 메모리 기반 저장장치를 지원하기 위해 추가된 PMFS 파일시스템 및 NVMe driver와 같은 스토리지 I/O 계층에 대한 분석에 한계가 있다. 계층 통합적인 스토리지 I/O 분석도구로 기존에 AIOPro[9]가 제안되었지만, 이는 안드로이드 스마트폰의 스토리지 I/O 계층을 대상으로 동작하는 분석도구로써 해당 도구로는 차세대 비 휘발성 메모리 기반 저장장치의 스토리지 I/O 동작의 분석에는 한계가 존재한다.

6. 결 론

본 논문에서는 차세대 비 휘발성 메모리 기반 저장장치를 사용할 때 발생하는 스토리지 I/O의 부하를 계층별로 분석하는 통합된 스토리지 I/O 분석도구인 PMPro를 제안하였다. 개발된 도구는 차세대 비 휘발성 메모리 기반 저장장치에 데이터 입출력을 수행할 때, 시스템 콜부터 저장장치 드라이버까지 모든 스토리지 계층 별로 스토리지 I/O 정보를 수집 및 연결하여 각 계층별 부하에 대한 계층 통합적 분석을 가능하게 해준다. 본 도구를 활용하면 차세대 비 휘발성 메모리 기반 저장장치를 위한 스토리지 I/O 성능 최적화 연구에 많은 기여가 있을 것으로 예상된다.

감사의 글

이 연구를 위해 연구장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에 감사 드립니다. 이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015M3C4A7065645). (교신저자: 김지홍)

참 고 문 헌

[1] 3D XPoint. https://www.intel.com/content/www/us/en/architecture-and-technology/intel-optane-technology.html.
[2] Z-NAND. http://www.samsung.com/us/labs/pdfs/collateral/Samsung_Z-NAND_Technology_Brief_v5.pdf.
[3] strace. http://linux.die.net/man/1/strace.
[4] blktrace. http://linux.die.net/man/8/blktrace.
[5] Latency table of memory and storage media. https://www.theregister.co.uk/2016/04/21/storage_approaches_memory_speed_with_xpoint_and_storageclass_memory/.
[6] Flexible I/O Tester. https://github.com/axboe/fio.
[7] TPC-C Benchmark. http://www.tpc.org/tpcc/.
[8] YCSB Benchmark. https://github.com/brianfrankcooper/YCSB/.
[9] S.S. Hahn et al., “AIOPro: A Fully-Integrated Storage I/O Profiler for Android Smartphones,” in *Proc. Korea Computer Congress*, 2016.