인텔 Optane DC Persistent Memory 기반 매니코어 서버에서 비휘발성 메모리 파일시스템의 확장성 평가

김준형, 김영재, 박성용
서강대학교 컴퓨터공학과
{ junehyung, youkim, parksy}@sogang.ac.kr

An Empirical Study of Non-volatile Memory File Systems on Intel Optane DC Persistent Memory based Manycore Servers

June-Hyung Kim, Youngjae Kim, Sungyong Park
Department of Computer Science and Engineering, Sogang University, Seoul, Republic of Korea

요 약

3D-XPoint, PRAM 또는 STT-MRAM 같은 비휘발성 메모리 (Non-volatile Memory)는 기존 Magnetic Disk와 NAND Flash Memory보다 높은 접근 속도와 병렬성을 제공한다. 최근 DRAM과 비휘발성 메모리 기반의 하이브리드 로깅 파일시스템인 NOVA가 등장하였다. NOVA는 높은 I/O 처리율과 데이터 원자성을 보장하는 가장 앞서 있는 비휘발성 메모리 파일시스템으로 평가 받는다. 하지만 매니코어 환경에서 NOVA는 획일적인 락으로 인해 여러 I/O 쓰레드가 같은 파일의 다른 블록들을 읽고 쓰는 경우에 대해서 성능 확장성을 보이지 못한다. 이러한 문제를 해결하기 위해 획일적인 락 대신 범위 기반의 락을 사용하는 NOVA 시스템, pNOVA가 제안되었다. 본 연구에서는 pNOVA의 확장성 분석을 위해 Intel Optane DC Persistent Memory 기반 매니코어 서버에서 실험을 진행하였다. 실험 결과로, N-to-1 쓰기의 경우 쓰레드 수가 증가함에 따라 pNOVA의 성능이 증가하다 특정 쓰레드 수 이후로는 확장성을 보이지 않았으며, N-to-1 읽기의 경우 성능이 선형적으로 증가하지만 실험마다 성능의 편차가 높았다.

1 서 론

최근 프로세서들의 집적도가 높아짐에 따라 하나의 서버가 많은 코어 수를 포함할 수 있다. 이러한 매니코어 환경에서 코어들의 병렬성을 활용하기 위한 고성능 병렬 I/O에 대한 연구가 활발하다 [1]. 병렬 I/O란 쓰레드들이 여러 입출력 작업을 직렬적으로 수행하는 것이 아닌 디스크의 데이터들에 동시에 접근하여 병렬적으로 입출력 동작을 수행하는 것을 의미한다. 최근 매니코어 환경에서 병렬 I/O 성능을 개선하기 위한 연구가 있었다 [2].

차세대 메모리 기술인 비휘발성 메모리 (NVMM, Non-Volatile Main Memory) 기반의 디바이스 [3, 4]는 기존 블락 기반 디바이스에 비해 낮은 읽기/쓰기 지연시간을 제공하므로, 고성능 병렬 I/O 수행을 가능하게 한다. 더욱이 비휘발성 메모리가 메모리 버스 라인에 장착 되고 블록 사이즈보다 작은 단위의 접근을 허용하기 때문에, CPU 코어들의 디스크에 대한 병렬적 접근을 이전보다 개선할 것 이라고 기대된다.

한편, 비휘발성 메모리 기반 파일시스템 [5, 6]들에 관한 연구 역시 활발히 진행되고 있다. 그 중 NOVA 파일시스템 [6]은 높은 I/O 처리율을 보이며 데이터 원자성을 보장하는 가장 앞서있는 메모리 기반 파일시스템이다. 특히 NOVA 파일시스템은 코어 별 메모리 할당 기술, inode 별 로깅 기술 같은 매니코어에 적합한 소프트웨어 디자인을 사용해 비휘발성 메모리의 병렬성을 활용하기 적합하다.

하지만, NOVA 파일시스템이 파일시스템의 확장성을 고려해 디자인 되었음에도 불구하고 I/O를 수행하는 쓰레드 수와 전체 I/O 처리율이 비례하지 못하는 경우가 존재한다. 쓰레드들이 같은 파일의 각각 다른 블록에 병렬적인 읽기/쓰기 동작 (N-to-1)을 수행하는 경우 굉장히 낮은 확장성을 보인다. 이는 NOVA가 하나의 파일에 대한 읽기/쓰기 동작을 수행할 때, 파일 전체에 대하여 락을 통해 동작 간의 발생할 수 있는 일관성 문제를 방지하기 때문이다. 즉, 하나의 파일에 대한 동시적인 I/O 동작들은 락에 의해 직렬화 된다. pNOVA [7]는 이러한 문제점을 개선하기 위한 병렬 파일시스템으로 기존 파일시스템 일관성을 유지하면서 성능을 개선하였다. pNOVA는 비휘발성 메모리 파일시스템에 적합한 범위 기반 락을 구현하여 읽기/쓰기 범위에 따른 파일에 대한 부분적 락킹을 가능하게 했다.

본 연구는 pNOVA의 확장성 평가를 위해 두개의 다른 플랫폼 (인텔의 3D-XPoint [3] 비휘발성 메모리 기반 서버, IBM DRAM 기반 서버)에서 실험하여 성능을 비교 평가를 한다. IBM 서버는 DRAM 기반의 비휘발성 메모리 환경 (배터리를 연결한 DRAM, 또는 NAND 플래시를 연결한 DRAM)을 이용하며 인텔의 3D-XPoint 비휘발성 메모리 기반 서버는 실제 비휘발성 메모리 Optane DC Persistent Memory [8]를 사용해 실험한다.

주요 실험 결과로, N-to-1 쓰기의 경우, I/O를 수행하는 쓰레드의 수가 증가함에 따라 두 서버에서 실험한 pNOVA의 I/O 처리율은 증가하였지만 두 서버 모두 특정 쓰레드 수 이후로는 확장성을 보이지 않았다. 이는 여전히 존재하는 파일시스템의 락 문제 때문이다. N-to-1 읽기의 경우, 두 서버 모두 pNOVA의 성능이 선형적으로 증가하였으나, 비휘발성 메모리 기반 서버의 경우 CPU 소켓 간의 높은 NUMA 거리로 인해 성능 편차가 높았다.

2 인텔 Optane DC Persistent Memory

비휘발성 메모리 관련 선행 연구들은 메모리를 에뮬레이션 하기 위하여 DRAM을 사용하거나, 배터리를 연결한 DRAM 혹은 NAND 플래시를 연결한 DRAM과 같은 환경을 구성하였다. 그러나 최근 실제 비휘발성 메모리가 상용화 되고 밝혀진 바 [9]에 따르면 DRAM과는 많은 특성 차이가 있다. 이 장에서는 DRAM과 인텔의 Optane DC Persistent Memory [8]

이 논문은 2019년도 정부 (과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2014-3-00035, 매니코어 기반 초고성능 스케일러블 OS 기초연구 (차세대 OS 기초연구센터)).

의 일부 차이점에 대해 알아본다. 본 논문에서는 Optane DC Persistent Memory를 간단히 Optane DC로 표기한다.

첫째, Optane DC는 I/O 접근 유형 (읽기/쓰기)에 따라 큰 성능 차이를 보인다. 256GB Optane DC 단일 모듈 기준 256 Byte 읽기의 경우 최대 6.8 GB/s, 쓰기의 경우 최대 2.3 GB/s bandwidth를 보인다. 읽기와 쓰기 동작의 성능 차이가 약 2.95배 정도 있으며 이는 32 GB Micron DDR4 DIMM의 읽기/쓰기 간의 1.3배 성능 차이에 비해 큰 치수이다.

두번째로 Optane DC는 동시에 디바이스에 접근하는 쓰레드 수에 의존적이다. Optane DC 단일 모듈은 접근하는 쓰레드 수에 전혀 확장성을 보이지 못한다. 이에 대한 원인은 프로세서의 메모리 컨트롤러 구조로 추측된다. 인텔은 Optane DC를 지원하기 위한 플랫폼을 구성하고 프로세서의 메모리 컨트롤러를 디자인 하였다. 이 메모리 컨트롤러는 WPQ (write pending queue)를 내장하고 있다. CPU 캐시로부터 메모리 컨트롤러로 전달된 데이터는 WPQ에 우선적으로 저장 되는데, WPQ에 있는 데이터들은 갑작스런 전원 차단에도 일정 시간내에 Optane DC 모듈로 플러시 됨으로써 영구성을 보장 받는다. 프로세서의 메모리 컨트롤러는 WPQ를 Optane DC 모듈마다 관리하기 때문에 쓰레드들의 단일 모듈에 대한 접근은 단일 큐에 대한 경쟁을 유발하고 이는 곧 확장성 문제를 일으킬 것이다.

3 pNOVA 파일 시스템

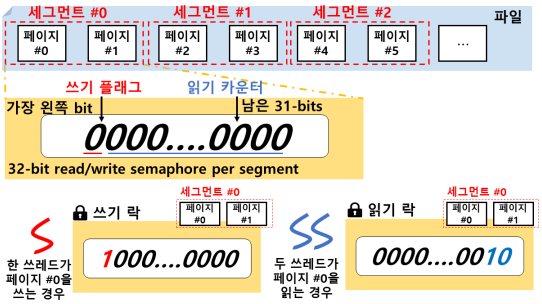


그림 1: pNOVA 세그먼트 범위 기반 락 구조

기존의 NOVA 파일시스템은 읽기/쓰기 동작을 위하여 파일 전체에 획일적인 락을 사용한다. 이러한 획일적인 락은 다음과 같은 두가지 확장성 문제를 유발한다. 첫째, 여러 쓰레드가 한 파일에 동시 읽기를 수행하는 경우 각각의 쓰레드들은 파일의 하나의 참조 카운터를 증가시키기 위해 경쟁을 한다. 이는 단일 참조 카운터 문제로 잦은 캐시 일관성 프로토콜을 일으켜 확장성에 심각한 손상을 일으킨다. 둘째, 여러 쓰레드의 한 파일에 대한 동시 쓰기 동작이 불가하다. 쓰레드들의 쓰기 범위가 다름에도 불구하고 특정 쓰레드가 파일에 대해 쓰기 중이라면 다른 쓰레드들은 블락킹되어 쓰기 동작들이 직렬화 될 수 밖에 없다.

pNOVA는 이러한 획일적인 락을 범위 기반 락으로 대체하였다. pNOVA의 범위 기반 락은 파일을 세그먼트라는 단위로 나누고 각 세그먼트들을 관리하는 락 변수를 두어 읽기/쓰기 범위에 따라 해당 세그먼트들만 락을 한다. 그림 1은 pNOVA의 세그먼트 범위 기반 락의 구조를 보여준다. 파일을 연속적인 페이지들의 그룹으로 생각했을 때, 우선 연속적인 페이지들을 세그먼트라는 단위로 묶는다. (그림에서는 두 페이지들을 하나의 세그먼트로 정의하였다.) 각 세그먼트는 32-bit의 세마포어를 가지는데, 세마포어의 가장 왼쪽 bit는 쓰기 플래그로 사용되며 남은 31-bit는 읽기 카운터로써 사용된다. 쓰레드들은 읽기/쓰기 범위에 따라 해당 세그먼트의 세마포어에 접근한다. 만약 가장 왼쪽의 bit가 1일 경우 해당 세그먼트에 대해 특정 쓰레드가 쓰기 중이므로 이후에 접근하는 읽기/쓰기를 블록킹 한다. 만약 세마포어의 값이 0이 아니면서 가장 왼쪽 bit가 1도 아닌 경우 특정 쓰레드들이 읽기 중이므로 들어오는 읽기 동작에 대해서는 카운터를 증가시키고, 쓰기 동작에 대해서는 블록킹을 한다.

pNOVA의 범위 기반 락은 세그먼트마다 참조 카운터를 가지고 있기 때문에 다른 범위의 읽기 동작에 대해서는 참조 카운터 문제가 발생하지 않는다. 병렬적 쓰기 동작에 대해서는 각각의 쓰레드가 쓰기 범위에 해당하는 세그먼트만 락을 걸어 범위가 다른 경우 서로에 의해 블락킹 되지 않는다.

4 실험 결과

4.1 실험 환경 설정

우리는 실제 비휘발성 메모리를 탑재한 환경 (표 1)과 비휘발성 메모리를 에뮬레이션한 환경 (표 2)에 대해 각각 실험을 진행하였다. 테스트베드2 (표 2)은 리눅스 커널이 지원하는 DRAM 기반의 비휘발성 메모리 에뮬레이션 [10]을 사용하였다.

테스트베드1 (표 1)의 경우 단일 소켓은 6개의 Optane DC 모듈을 탑재하고 있는데, 이 모듈들에 대한 두 가지 설정을 제공한다. 하나는 6개의 단일 모듈들로 인식 될 수 있는 "Non-interleaved" 모드이고 다른 하나는 6개의 모듈들이 인터리빙을 통하여 하나의 모듈로 인식되는 "Interleaved" 모드이다. 본 논문에서는 두 설정에 대해 각각 "-NI", "-I"로 표기한다. 두 테스트베드 모두 하이퍼쓰레딩은 사용하지 않으며 실험에 사용한 pNOVA의 세그먼트 사이즈는 4KB 이다.

파일시스템의 확장성 측정을 위하여 FxMark benchmark [2]를 이용한다. FxMark는 다양한 I/O 패턴의 워크로드를 생성하는 데, 그 중 우리는 쓰레드들이 단일 파일의 각각 다른 블록에 대한 병렬적 읽기/쓰기 동작을 수행하는 워크로드 (DRBM/DWOM)를 이용한다. 특히, 벤치마크 수행 시 I/O를 수행하는 쓰레드와 CPU 코어는 1대1 매핑 된다.

4.2 실험 결과 및 분석

4.2.1 N-to-1 쓰기 패턴 워크로드

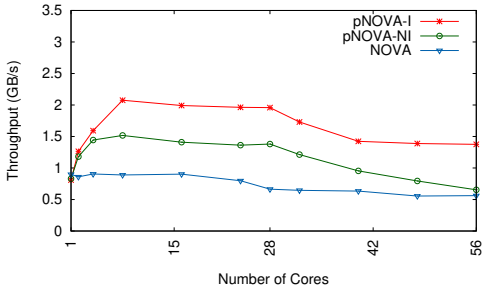
그림 2a의 결과를 보면 쓰기를 하는 쓰레드가 8개인 경우, pNOVA-I 의 I/O 처리율이 기존의 NOVA에 비해 최대 2.3× 증가하였다. pNOVA-NI는 쓰레드 수가 4개일 때 기존의 NOVA에 비해 최대 1.5× 증가하였다. 하지만 pNOVA-I와 pNOVA-NI 모두 코어 수가 증가하면서 성능 확장성은 보이지 않는다. 이에 대한 원인은 I/O 쓰레드 수가 증가함에 따라 WPQ (비휘발성 메모리 컨트롤러 큐)에 들어오는 I/O 요청이 많아지고 Optane DC 모듈이 큐로부터 읽어드리는 속도가 들어오는 속도에 비해 상대적으로 느리기 때문에 큐가 포화 되어 이후의 요청들을 블락킹 했기 때문이라고 추측한다. 또한 두 경우 모두 테스트베드1의 소켓 CPU 코어 수인 28코어

표 1: 테스트베드 1

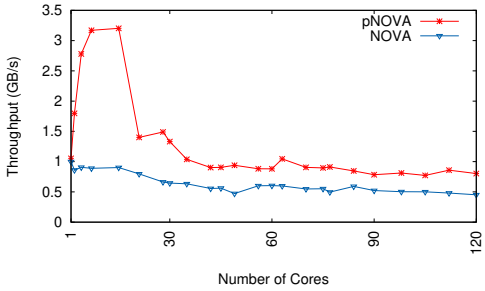
CPU	Intel(R) Xeon(R) 플래티넘 8280M v2 2.70GHz CPU 소켓 (노드) 수 (#): 2 소켓 (노드) 당 코어 수 (#): 28 총 코어 수 (#): 56
메모리	DDR4, 64 GB * 12개 (=768GB)
비휘발성 메모리	128 GB * 12개 (=1.5TB)

표 2: 테스트베드 2

CPU	Intel(R) Xeon(R) CPU E7-8870 v2 2.30GHz CPU 소켓 (노드) 수 (#): 8 소켓 (노드) 당 코어 수 (#): 15 총 코어 수 (#): 120
메모리	DDR3, 96 GB * 8개 (=968GB)
비휘발성 메모리	32 GB (DRAM 기반 에뮬레이션)

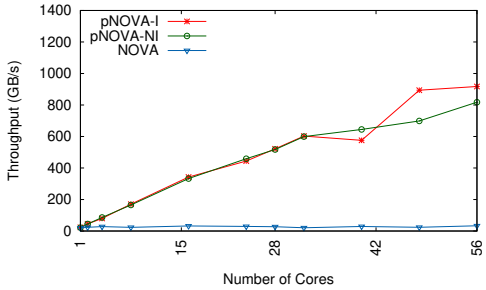


(a) 테스트베드 1

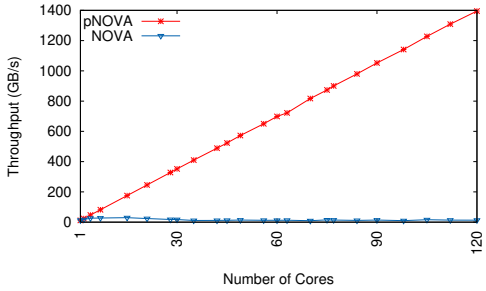


(b) 테스트베드 2

그림 2: DWOM 워크로드 실험 결과



(a) 테스트베드 1



(b) 테스트베드 2

그림 3: DRBM 워크로드 실험 결과

이후로는 파일시스템의 락 오버헤드로 인해 성능이 하락하는 것으로 추정된다. pNOVA-I가 pNOVA-NI에 비해 높은 성능 향상을 보인 주된 이유는 pNOVA-I의 경우 I/O들이 인터리빙 방식으로 6개의 Optane DC 모듈들에 분산 되어 단일 모듈의 처리율보다 나은 성능을 보이기 때문이다.

그림 2b의 DRAM 기반의 실험은 쓰레드가 15개인 경우, pNOVA의 I/O 처리율이 기존의 NOVA에 비해 약 3.5× 증가하였다. Optane DC 기반 실험은 테스트베드1 소켓 CPU 코어 수인 28개까지 확장성을 보이지 못한 반면 DRAM 기반의 실험은 테스트베드2 소켓 CPU 코어 수인 15개까지 확장성을 보였다. DRAM의 메모리 컨트롤러 큐는 뱅크별로 존재하여 WPQ와 같은 블락킹 문제는 발생하지 않았을 것이라고 추측한다. 하지만 DRAM 기반의 실험 역시 소켓 CPU 코어 수인 15개 이후로는 파일시스템의 오버헤드로 인해 성능이 하락하기 시작한다.

4.2.2 N-to-1 읽기 패턴 워크로드

먼저 그림 3a의 결과를 보면 쓰레드 수가 소켓의 CPU 코어 수인 28개일 때까지 증가함에 따라 pNOVA-I, pNOVA-NI 모두 I/O 처리율이 선형적으로 증가하는 것을 알 수 있다. 반면 기존의 NOVA는 단일 참조 카운터 문제로 전혀 확장성을 보이지 못한다.

이어서 그림 3b은 DRAM 기반의 pNOVA 실험을 보여준다. DRAM 기반의 pNOVA 역시 쓰레드 수가 최대 수까지 증가함에 따라 I/O 처리율이 선형적으로 증가하였다. Optane DC 기반의 서버는 쓰레드 수가 28개보다 큰 경우에 대해서 pNOVA-I와 pNOVA-NI 모두 실험마다 성능이 약 33% 정도의 편차를 보였다. 이에 대한 원인은 인텔 Optane DC 서버 플랫폼의 소켓 간의 원격 접속 오버헤드가 IBM 서버에 비해 크기 때문이라고 추측한다. 두 서버의 소켓간의 거리는 각각 21, 11로써 2배 가량 차이난다. (소켓간의 거리는 원격 접속 오버헤드와 비례한다.) 또 DRBM 패턴의 특성상 쓰레드들이 블락을 읽는 동작은 주로 캐시의 효과를 받기 때문에 2장에서 밝힌 DRAM과 비휘발성 메모리의 읽기 성능 차가 결과값에 나타나지 않는다.

5 결론

본 논문에서는 pNOVA 파일시스템의 확장성에 대해 인텔의 3D-XPoint 기반의 비휘발성 메모리와 IBM 서버의 DRAM 기반의 비휘발성 메모리에서 비교 실험하였다. 실제 비휘발성 메모리 위에서의 실험과 DRAM 기반의 실험은 쓰기의 경우 단일 소켓의 CPU 코어 수까지는 성능이 증가하였고, 읽기의 경우 캐시 효과로 인해 성능이 유사하였다.

참고 문헌

[1] X. Wu and A. L. N. Reddy, "Asynchronous I/O support in Linux 2.5," in *LinuxSymposium, Ottawa, Canada*, 2003.

[2] C. Min, S. Kashyap, S. Maass, and T. Kim, "Understanding Many-core Scalability of File Systems," in *Proceedings of the 2016 USENIX Conference on Usenix Annual Technical Conference (ATC)*, pp. 71–85, 2016.

[3] Intel, "Revolutionizing Memory and Storage." https://www.intel.com/content/www/us/en/architecture-and-technology/intel-optane-technology.html, 2017.

[4] S. S. Ameen Akel, Adrian M. Caulfield, "Onyx: A Protoype Phase Change Memory Storage Array," in *3rd USENIX Conference on Hot Topics in Storage and File Systems, HotStorage'11*, 2011.

[5] S. R. Dulloor, S. Kumar, A. Keshavamurthy, P. Lantz, D. Reddy, R. Sankaran, and J. Jackson, "System Software for Persistent Memory," in *Proceedings of the 9th European Conference on Computer Systems (EuroSys)*, pp. 15:1–15:15, 2014.

[6] J. Xu and S. Swanson, "NOVA: A Log-structured File System for Hybrid Volatile/Non-volatile Main Memories," in *Proceedings of the 14th USENIX Conference on File and Storage Technologies (FAST)*, 2016.

[7] Y. K. June-Hyung Kim Jangwoong Kim, Sungyong Park, "pNOVA: Optimizing Shared File I/O Operations of NVM File System on Manycore Servers," in *10th ACM SIGOPS Asia-Pacific Workshop on Systems (APSys '19)*, 2019.

[8] B. Beeler, "Intel optane dc persistent memory module (pmm)." https://www.storagereview.com/node/7416, 2019.

[9] S. S. Jian Yang, Juno Kim, "An Empirical Guide to the Behavior and Use of Scalable Persistent Memory." https://arxiv.org/abs/1908.03583, 2019.

[10] M. Maciejewski, "How to emulate Persistent Memory." https://pmem.io/2016/02/22/pm-emulation.html, 2016.