AI를 위한 파일시스템 병렬 I/O 기술 연구

윤준원[1], 홍태영[1]
[1]한국과학기술정보연구원
jwyoon@kisti.re.kr, tyhong@kisti.re.kr

A Study on Parallel I/O Technology in Filesystem for AI

Junweon Yoon[1], Taeyeong Hong[1]
[1]Dept. of Supercomputing Infrastructure Center, KISTI

요 약

대규모 데이터를 처리하기 위해 머신러닝, 딥러닝과 같은 AI 활용 연구가 일반화되면서 시스템 환경 또한 병렬처리 연산에 강화된 가속기 기반의 이기종 아키텍처로 확산되고 있다. CPU 기반의 계산 환경과 달리 상대적으로 성능이 낮은 수천 개의 산술연산장치(ALU)를 활용해 쓰레드 방식으로 연산을 수행하며, I/O의 특성 또한 대규모의 데이터들이 수많은 연산장치에 전달되기 위한 Small I/O, High-throughput 처리 성능이 애플리케이션에 큰 영향을 끼친다.

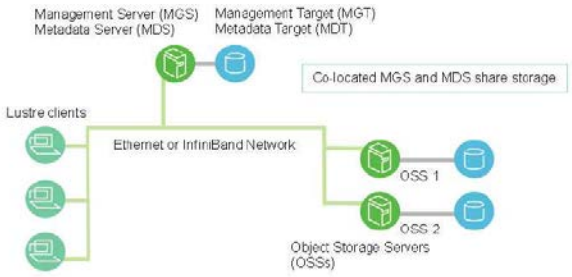
본 논문에서는 병렬 컴퓨팅 환경에 AI 애플리케이션이 접목되면서 요구되는 스토리지, 파일시스템의 환경을 분석하고 나아가 성능 검증을 통해 I/O 특성을 파악하고자 한다.

1. 서론

ML/DL과 같은 AI 분야의 연구가 활성화되면서 GPU와 같은 가속기 기반의 계산자원 도입으로 시스템 성능이 급격하게 증가하였고, 이에 따른 I/O 처리의 요구사항도 늘어나고 있다.[1] 본 연구에서는 AI 환경에 요구되는 I/O의 특성과 고성능 스토리지 및 병렬파일시스템 기술에 대해 살펴보고자 한다. 또한 해당 기술에 대해 벤치마크 프로그램을 이용하여 성능을 검증하고자 한다.

2. AI 기반 병렬파일시스템

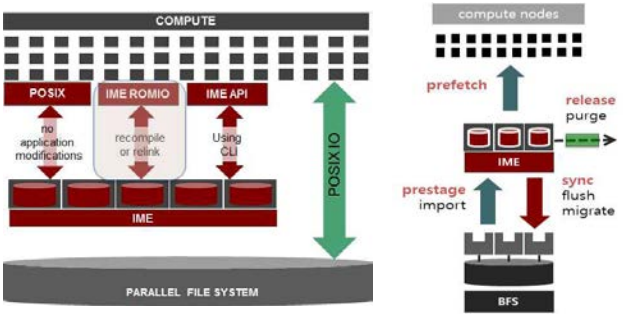
클러스터 환경에서 단일 스토리지의 한정된 I/O 대역폭을 확장하는 방법으로 Network Attached Storage(NAS) 또는 Storage Area Network(SAN) 환경에서 여러 대의 스토리지에 파일을 분산하여 저장하고 각 파일에 위치, 상태 정보를 저장하는 메타데이터를 배치한다. Lustre, GPFS(Spectrum Scale)는 대규모 클러스터 환경에서 사용되는 대표적인 병렬파일시스템으로 계산 시스템에 마운트 되어 스토리지와 연결된 여러개의 인터커넥트 대역폭을 통해 분산되어 I/O를 처리한다. Lustre는 대규모 클러스터 환경을 지원할 수 있는 객체 기반의 파일시스템으로 메타데이터 서버를 통해 데이터의 인덱스 정보를 얻은 후 실제 I/O 처리는 데이터 서버와 직접 통신하게 된다.[2]



(그림 1) Lustre 컴포넌트
<출처: Lustre S/W 2.x Manual>

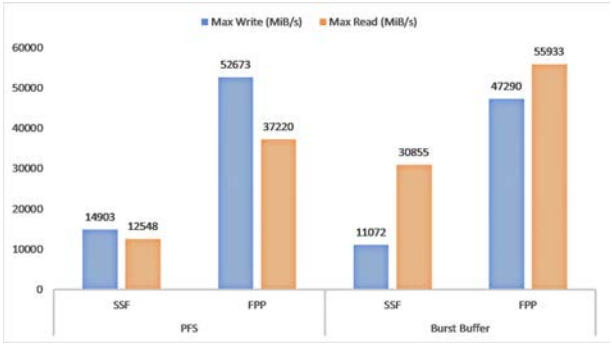
최근 비휘발성의 플래시메모리 기반 NVMe SSD가 일반화되고 가격 경쟁력이 확보됨에 따라 고속의 I/O를 지원하기 위한 스토리지 제품군에 탑재되어 출시되고 있다. 특히, AI 기반의 ML/DL 관련 애플리케이션들이 HPC와 같은 대규모 클러스터 환경에서 수행되면서 수많은 파일 개수 그리고 수 Byte 또는 KB 수준의 작은 파일 처리를 위한 요구사항이 증가하고 있다. 이에 파일 전송의 대역폭(B/W)은 물론 파일의 처리(IOPS) 성능 중요성이 더욱 부각되는 시점이다.

플래시 기반의 스토리지는 여전히 디스크 기반과 비교해 가격 대비 용량, 내구성에 한계가 있어 보다 경제적으로 I/O를 가속할 수 있는 티어링 기반의 스토리지(Storage Tiering) 솔루션들이 적용되고 있다. 대표적인 예로 KISTI 슈퍼컴퓨터 5호기[3]에 도입된 버스트버퍼(Busrt Buffer) IME는 계산노드와 스토리지 중간에 I/O 가속을 위해 NVMe 기반으로 구축된 파일시스템이다[4]. 사용자는 기존 병렬파일시스템에서 데이터를 버스트버퍼에 스테이징한 후 연산 결과를 캐싱하고 다시 병렬파일시스템으로 플러싱 한다.

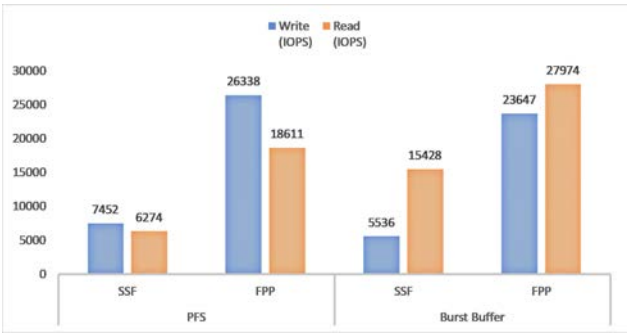


(그림 2) 버스트버퍼 IME 구조 및 데이터 흐름

IME는 Lustre 병렬파일시스템의 캐시 역할을 수행하기 때문에 동일한 메타데이터를 이용하며 별도의 내부 트리를 통해 파일을 관리하고 하드링크를 생성한다. 이는 동시에 대량의 파일을 처리하는 경우 병렬파일시스템 대비 낮은 성능을 가져올 수 있다. IOR은 병렬파일시스템의 I/O 성능을 측정하는 도구로 32개의 클라이언트 노드(노드당 8 프로세스, 총 256 프로세스)를 사용하여 단일 파일 생성(single-shared-file)과 모든 프로세스가 개별 파일을 생성(file-per-process)할 때의 성능을 비교하였다. (그림 3)과 (그림 4)는 병렬파일시스템(PFS)와 버스트버퍼(Burst Buffer)와의 성능 비교를 보여준다[14].



(그림 3) PFS와 Burst Buffer 대역폭 비교



(그림 4) PFS와 Burst Buffer IOPS 비교

(그림 3)에서 파일 쓰기(Write)의 경우 PFS에서 read에 비해 write 성능이 높게 측정된 이유는 커밋(commit)이 메모리에 캐싱된 시점으로 측정되기 때문이다. 한편 Lustre의 분산(striping) 저장 기법이 적용되어 쓰기 성능은 버스트버퍼에 비해 조금 더 높은 대역폭을 나타낸다. 파일 읽기(Read)는 버스트버퍼가 전반적으로 높은 성능을 보인다. (그림 4)의 IOPS 성능을 보면 PFS의 캐시 역할을 위한 메타데이터 처리로 읽기 경우 더 많은 IO 처리량이 필요하다.

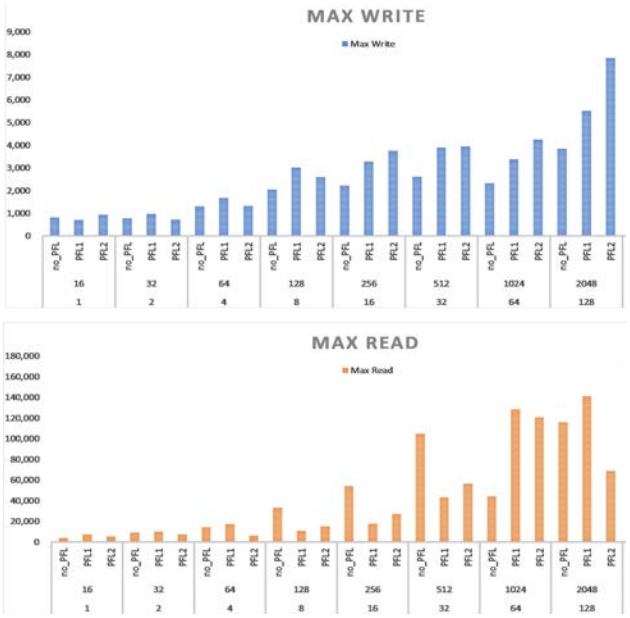
AI 기반 ML/DL 연구 수행을 위해 GPU와 같은 이기종 아키텍처 적용이 확산하고 있다. CPU보다 낮은 클럭의 수 천개 코어를 배치하여 병렬처리 성능을 향상하기 위한 목적이며, 딥러닝의 경우 데이터 병렬화(Data Parallelism)와 모델 병렬화(Model Parallelism)를 이용하여 데이터 셋을 나누어 처리하거나 모델을 나누어 처리하는 방법을 사용하고 있다. 연산 처리의 워크로드를 발생하는 데이터 크기는 수 KB 수준으로 상대적으로 작은 데이터 셋을 가져와 여러 단계에 거쳐 I/O 처리를 수행한다.

NVMe(Non-Volatile Memory Express)는 PCIe 기반 NVMe SSD를 위한 커뮤니케이션 규약으로 기존 AHCI(Advanced Host Controller Interface)를 사용하는 디스크 대비 대량의 데이터를 높은 대역폭으로 전달할 수 있다. NVMe SSD는 최대 64GB/s의 대역폭, 65,535의 명령어 대기열 가지고 있어 기존 SATA SSD의 단일 대기열에 32개 명령어 처리와 비교하면 월등한 성능을 가지고 있어 AI 처리 환경에 필수 요소로 자리매김하고 있다.[5]

병렬파일시스템에는 디스크의 성능, 대규모의 작은 파일의 I/O 처리를 위한 요구사항이 증대되면서 몇가지 새로운 기술들이 적용되고 있다.

- Lustre Progressive File Layout (PFL)
  하나의 파일을 물리적으로 분산된 디스크인 여러 OST(Object Storage Target)로 분할하여 저장함으로써 병목 현상을 제거 하고 I/O 성능 향상시킬 수 있다. PFL은 기존 고정으로 분할하여 저

장하는 방식의 진보된 형태로 파일의 크기에 따라 동적으로 분할 개수(Stripe count)와 크기(Stripe Size)를 조정한다.



(그림 5) PFL의 확장성 성능 테스트

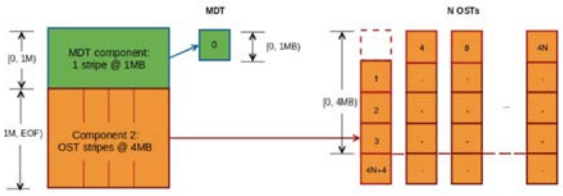
(그림 5)는 1노드~128노드(16ppn)로 확장하면서 단일 파일을 생성(single-shared-file)하는 성능 시험을 수행하였다. PFL이 미적용(no_PFL)과 PFL1, PFL2는 설정후 분할 개수(Stripe count) 차이에 따른 성능이다. <표 1>은 PFL1의 설정이며 PFL2는 1G 이상부터 PFL1에 비해 카운트 개수를 2배로 더 늘렸다. 노드의 개수(파일의 크기)가 증가할수록 PFL 설정한 경우가 더 좋은 성능을 얻을 수 있음을 확인할 수 있다. PFL2의 분할 개수(Stripe count)가 더 많아져 Read의 경우 각 파일의 인덱스 정보를 수집하는 부하가 더 발생할 수 있다. 따라서 실험을 통해 적절한 분할 개수 설정을 얻을 수 있다.

<표 1> PFL1의 동적 Striping 설정

/scratch (PFL)	file < 4M	1	1MB
	4MB < file < 512MB	2	1MB
	512MB < file < 1G	4	1MB
	1G < file < 10G	8	1MB
	10G< file < 100G	16	1MB
	file > 100G	32	1MB

• Data on Metadata (DoM)

Lustre는 상대적으로 작은 파일의 I/O 처리 성능에 취약점을 가지고 있다. DoM 기술은 일반적으로 메타데이터 처리후 디스크에 저장하는 절차와 달리 직접 메타데이터에 저장하여 작은 파일에 대한 I/O 성능을 향상시킬 수 있는 기술이다.



(그림 6) DoM의 파일 배치

<출처: Lustre S/W 2.x Manual>

• Distributed NamespacE (DNE)

DNE는 기존 메타데이터의 저장공간 MDT(MetaData Target)의 단일 구성을 분산하여 복수개로 배치하는 기술이다. AI 환경에서 요구되는 높은 IOPS 처리를 위해 메타데이터를 여러개 배치하여 분산함으로써 용량과 성능 향상을 얻을 수 있다.

3. 결론

시스템 환경 또한 병렬처리 연산에 강화된 가속기 기반의 이기종 아키텍처로 확산되고 있다. CPU 기반의 계산 환경과 달리 상대적으로 성능이 낮은 수천 개의 산술연산장치를 사용하여 병렬성능을 확대해가고 있으며 이에 따른 I/O 특성도 변화한다. 본 연구에서는 병렬파일시스템에서 AI 기반 I/O 특성을 지원하기 위한 기술들을 언급하였다. 이 외에도 계층형 스토리지 기술(Automated Storage Tiering), 메타데이터의 IOPS 성능 향상을 위한 분산 스토리지 기술, 나아가 네트워크(RDMA, RoCE 등) 분야, GPU 직접통신 기술(GPU-Direct Storage) 등 많은 분야에서 AI 연구를 위한 새로운 기술들이 지속해서 발전하고 있다. 향후, 다양한 스토리지, 파일시스템에 관한 지속적인 연구와 성능 검증이 필요하다.

참고문헌

[1] G. Mouzhi, H. Bangui, and B. Buhnova, “Big data for internet of things: a survey”, Future generation computer systems, pp. 601-614, 2018

[2] Lustre Software Release 2.x Available: https://doc.lustre.org/lustre_manual.xhtml

[3] KISTI Supercomputing Center (KSC), Available: http://www.ksc.re.kr/.

[4] Yoon JW, Song US, “A Study on System Performance Verification Methods for Heterogeneous Computing Environments”, The Journal of Digital Contents Society, 23(2), pp. 295-302, 2022.

[5] Kim S, Yang JS, “Optimized I/O determinism for emerging NVM-based NVMe SSD in an enterprise system”, Proceedings of the 55th Annual Design Automation Conference. 2018. pp. 1-6