고성능 I/O 지원을 위한 계층형 스토리지 구현

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

Implementation of Tiering Storage to Support High-Performance I/O

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

요 약

ML/DL과 같은 AI의 연구가 HPC 환경에서 수행되면서 데이터 병렬화, 분산 학습 및 대규모 데이터 세트를 처리를 위한 요구사항이 급격히 증가하였다. 또한, 병렬처리 연산에 특화된 가속기 기반 이기종 아키텍처 환경 변화로 I/O 처리에 고대역폭, 저지연의 스토리지 기술을 필요로 하고 있다.

본 논문에서는 고집적의 병렬 컴퓨팅 환경에 고성능 HPC, AI 애플리케이션을 처리하기 위한 티어링 스토리지 기술을 논한다. 나아가 실제 고성능 NVMe 기반의 플래시 티어링 계층 구성에서 액세스 패턴에 따른 데이터 처리 환경을 구축하고 성능을 검증한다. 이로써 다양한 사용자 어플리케이션의 I/O 패턴을 특성에 맞게 지원할 수 있다.

1. 서론

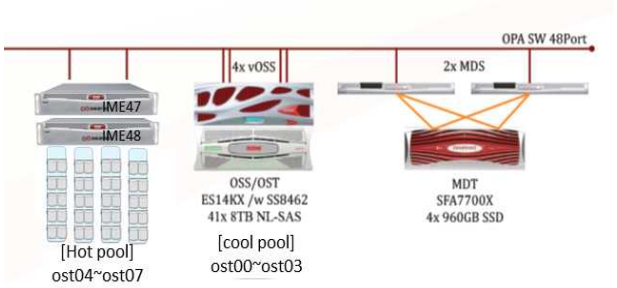
HPC 환경에서 AI 어플리케이션 접목으로 대용량 데이터를 다루고 복잡한 모델 학습 및 추론을 수행을 위해서는 고성능 파일 시스템이 필수 불가결한 요소이다. 이를 위해 NVMe(Non-Volatile Memory Express) 플래시 기반 스토리지가 적용되어 AI 워크로드에 적합한 높은 대역폭과 낮은 지연속도로 I/O 처리를 지원한다[1]. 그러나 용량 대비 고비용의 매체로 다양한 모든 I/O를 처리하기에는 경제적인 부담이 있다. 따라서 NVMe 기반 플래시와 기존 SATA, SAS 기반의 디스크 매체를 함께 사용하여 캐싱 및 티어링 기술을 적용하면 다양한 액세스 패턴에 따른 I/O 처리를 효율적으로 지원할 수 있다.

자동 티어링 스토리지 기술(Automated Tiering Storage)은 데이터의 I/O 패턴에 따라 이주 정책을 정의하고 자주 액세스되는 데이터는 NVMe SSD와 같은 빠른 스토리지에 덜 빈번하고 오랜 기간 저장되는 데이터는 SATA, SAS 기반의 디스크 매체에 배치하여 성능 향상과 비용 절감을 동시에 충족시킨다. 본 연구는 자동 티어링 스토리지를 구축하여 기능과 성능을 검증함으로써 다양한 사용자 어플리케이션의 I/O 요구사항을 수렴하기 위한 지침을 제공한다[2].

2. 계층형 스토리지

계층형 스토리지(Tiering Storage) 기술은 고성능 I/O를 지원하기 위한 기술로 다양한 스토리지 레벨과 계층을 조합하여 데이터 액세스의 효율성을 높이고 데이터 관리를 최적화하는 방법을 제공한다. 고성능 I/O는 대용량 데이터 처리, 복잡한 계산, AI 모델 학습 및 추론과 같은 작업에 핵심요소로 데이터의 빠른 읽기 및 쓰기 속도, 낮은 대기 시간, 확장 가능한 스토리지, 그리고 데이터 관리와 같은 워크로드 요구사항을 갖는다. 계층형 스토리지는 여러 가지 스토리지 레벨을 조합하여 구성하는데 플래시 기반 NVMe 드라이브를 이용한 고속 스토리지, 일반적인 HDD(하드 디스크 드라이브)가 적용된 대용량의 스토리지 그리고 좀 더 확장하면 클라우드 스토리지와 같은 원격 스토리지 등이 포함될 수 있다[3].

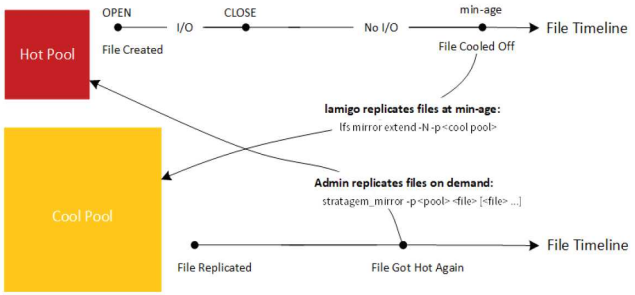
본 연구에서는 플래시 메모리 기반의 NVMe SSD로 구성된 Hot 영역과 HDD 기반의 N-SAS 스토리지 Cold 영역을 계층형 스토리지로 구성하였다. 본 실험을 위해 (그림 1)과 같이 KISTI 슈퍼컴퓨터 5호기의 누리온(Nuirion) 버스트버퍼(NVMe SSD) 스토리지를 활용하여 Hot 영역으로 기존 N-SAS 기반의 디스크는 Cold 영역으로 계층형 스토리지를 구축하였다[4][5].



(그림 1) 누리온 버스트버퍼를 활용한 NVMe SSD 기반 계층형 스토리지 구성

계층형 스토리지는 빈번하게 액세스되는 데이터를 고속(Hot) 스토리지에 저장하고 빈도가 낮게 사용되는 데이터는 대용량 스토리지(Cold)에 저장한다. 또한 용량, 저장 기한 등의 정책을 적용하여 데이터의 빈도에 따라 스토리지 레벨을 조정하고 가장 적합한 스토리지 계층으로 데이터를 이동시킴으로써 최상의 성능을 유지한다. 이렇게 Hot/Cold 영역의 자동화된 데이터 적재(Automated Storage Tiering) 알고리즘과 정책을 적용을 위한 프로그램으로 DDN의 EXAScaler SW를 설치, 적용하였다[6]. 이 자동 적재 알고리즘은 크게 두 가지의 기능을 포함하고 있다. 첫 번째는 Hot에서 Cold 영역으로 데이터를 자동으로 미러링하는 기능으로 일정 시간 후 Cold 영역은 데이터의 복사본을 갖는다. 두 번째는 조건에 따라 Hot 영역의 최적 용량 관리하기 위해 퍼지 기능이다. 이를 통해 상대적으로 공간이 작은 Hot 영역의 용량을 관리할 수 있다.

(그림 2)는 Hot/Cold 영역의 데이터 이동 흐름을 보여준다. 초기 Hot 영역에 저장된 데이터는 Cold 영역에 동일하게 미러링 및 동기화가 발생하며 일정 시간동안 사용되지 않으면 Hot 영역에서 제거된다. 이 후 해당 파일이 재사용되면 Hot 영역에 자동 적재(staging) 되어 해당 영역에서 I/O가 발생시킨다.

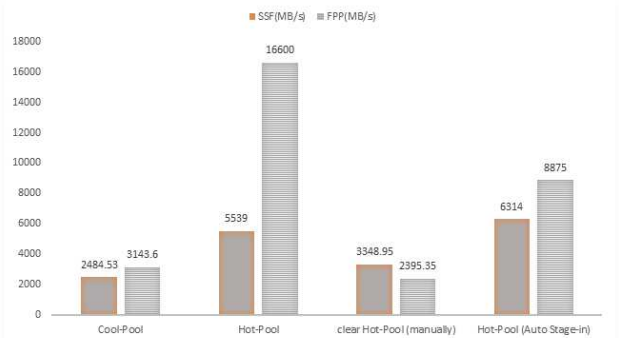


(그림 2) Hot/Cold 영역간 자동화 데이터 적재(Automated Storage Tiering) 흐름

이렇게 구성된 계층형 스토리지에서 각 영역에 대한 데이터 저장 정책은 <표 1>과 같은 명령어를 통해 정의할 수 있다. 초기 Hot/Cold 영역을 설정할 때 해당 공간의 저장소를 nvme_ssd, nsas_hdd와 같이 지정하게 된다. 이 후 대상이 되는 디렉터리(/data_dir)에 속성 파일 크기에 따라 정의하게 되는데 여기에서는 1GB 이하의 파일사이즈 데이터는 Hot(nvm_ssd) 영역에 저장되며 그 이상의 데이터는 Cold(nsas_hdd) 영역에 저장되도록 설정하였다. 또한 Hot/Cold 영역간의 데이터 동기화 옵션을 설정할 수 있는데 comp-flags=prefe 옵션은 지속적으로 데이터의 변경사항이 있는 경우 Hot/Cold 영역에 데이터를 동기화 한다.

<표 1> 계층형 스토리지 속성 정의 명령

```
# lfs setstripe -E 1G -c 1 -p nvme_ssd
  --comp-flags=prefer -E 16G -c 4 -E eof -c
  40 -p nsas_hdd /data_dir
```



(그림 3) Hot/Cold 영역 성능 및 자동 적재 기능 검증

(그림 3)은 Hot/Cold 영역에서 단일파일을 생성(Single-Shared File) 또는 각각 파일을 생성(File per Process)하는 방식에서 읽기 성능을 측정한 결과이다[7]. 플래시기반의 NVMe SSD 풀인 Hot 영역과 N-SAS 기반의 Cool 영역으로 구성된 자동 적재 환경에서 발생하는 I/O 패턴에 따라 성능을 측정하였다. 첫 번째와 두 번째는 각 영역에서 읽기 성능을 각각 측정하였고 세 번째에 일정 시간 후 Hot 영역에 데이터가 제거된 후 다시 데이터가 읽을 때의 성능을 보게 되면(clear Hot-Pool) 처음 Cold 영역에 데이터를 읽어올 때와 유사한 성능을 보인다. 즉, Hot 영역에서 제거된 데이터를 다시 읽을 때는 처음 Cold 영역에서 가져온 후 Hot 영역으로 적재된다. 마지막은 다시 Hot 영역으로 재 적재되었을 때의 성능을 나타낸다.

3. 결론

스토리지의 I/O 성능과 다양한 패턴들에 대한 요구사항이 동시에 급격히 증가하면서 계층형 자동적재 스토리지 기술(Automated Storage Tiering), 메타데이터의 IOPS 성능 향상을 위한 분산 스토리지 기술, 네트워크(RDMA, RoCE 등) 분야, GPU 직접 통신 기술(GPU-Direct Storage) 등과 같은 I/O 성능향상 기술과 NVMe (Non-Volatile Memory Express)와 PMem (Persistent Memory)과 같은 플래시 메모리 기반 매체 적용이 대중화되고 있다. 계층형 스토리지 기술은 고성능 I/O를 지원하기 위해 다양한 스토리지 리소스를 효과적으로 관리하고 활용하는 방법을 제공한다. 이로써 대용량 데이터 처리와 AI 어플리케이션과 같은 요구 사항을 충족시키면서도 비용과 성능을 최적화할 수 있다.

본 연구에서는 계층형 자동적재 스토리지 기술(Automated Storage Tiering)을 실제 구현한 후 기능과 성능을 확인하였다. 해당 사례는 HPC, AI 연구자의 고성능 I/O 처리를 위한 스토리지 구축 사례로 도움이 될 수 있다.

Acknowledgement

본 연구는 2023년도 한국과학기술정보연구원(KISTI) 기본사업 과제로 수행한 것입니다.

참고문헌

[1] D. Milojicic, P. Faraboschi, N. Dube, and D. Roweth, “Future of HPC: Diversifying heterogeneity”, Design, Automation & Test in Europe Conference & Exhibition, pp. 276-281, 2021.

[2] H. Herodotou, E. Kakoulli, “Automating distributed tiered storage management in cluster computing”, arXiv preprint arXiv:1907.02394, 13(1), pp. 43-56, 2019.

[3] 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, pp. 1-6, 2018.

[4] Yoon JW, Song US, “Parallel File System Characteristics and Performance Analysis with NVMe SSD based Cache”, The Journal of Digital Contents Society, 21(1), pp. 157-164, 2021.

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

[6] Turisini, Matteo, Giorgio Amati, Mirko Cestari, "LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI Applications." arXiv preprint arXiv:2307.16885, 2023.

[7] 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.