Overhead Analysis of WAL on RocksDB

성한승[1], 박상현[2]

1) 연세대학교 컴퓨터과학과 석사과정, hssung@yonsei.ac.kr
2) 연세대학교 컴퓨터과학과 교수, 교신저자, sanghyun@yonsei.ac.kr

연세대학교 YONSEI UNIVERSITY

목차

연구 동기

- RocksDB 성능 저하 요소 분석
  - 데이터 지속성 ➡ 로깅
- 성능 향상과 동시에 데이터 지속성 보장
  - NVRAM
- 관련 연구
  - RocksDB의 동기화 성능 비교, 안미진, 오기환, 강운학, 이상원. (2014). 한국정보과학회 학술발표논문집 , 1516-1518.
  - 동기화 옵션에 따른 RocksDB 성능 평가, 박연수, 오기환, 이종백, 강운학, 이상원. (2014). 한국정보과학회 학술발표논문집, , 1731-1733

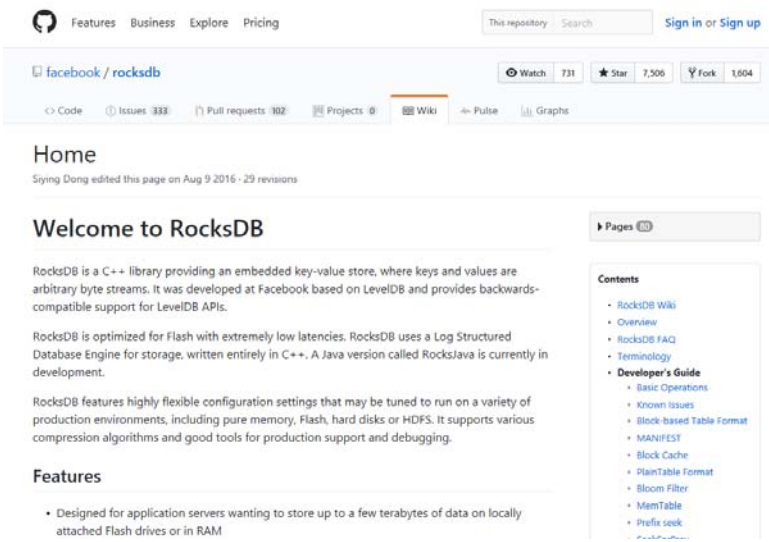
연구 목적

- WAL 오버헤드 분석

- 더 빠른 로깅 디바이스를 통한 성능 개선도 측정
  - 향후 연구에 대한 가능성 확인

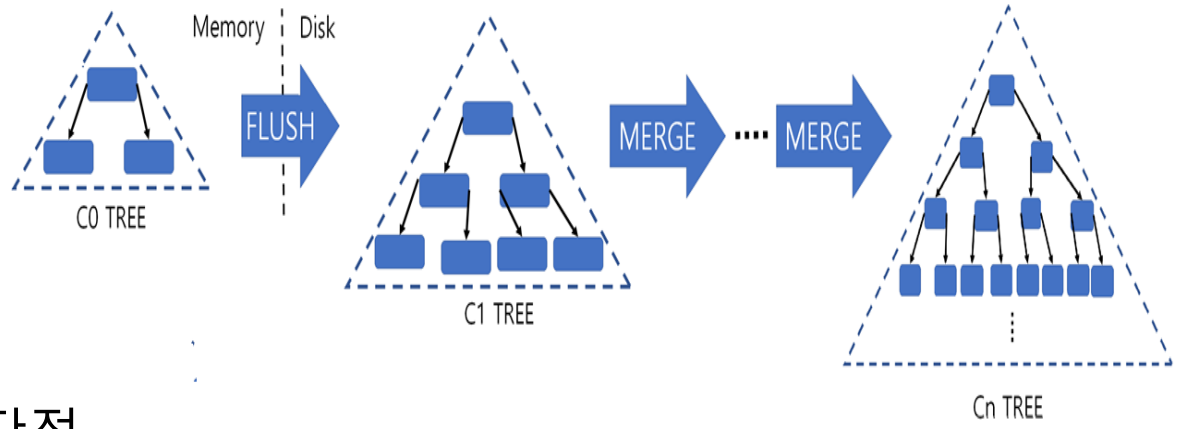
RocksDB

- 소개
  - Facebook 에서 오픈 소스로 공개
  - LevelDB를 기반으로 하는 Key-Value store
  - Storage engine으로 각광 받고 있음
    - MyRocks, MongoRocks에 사용



- 특징
  - Embedded 데이터베이스
  - LSM-Tree 구조를 사용
  - 컬럼 패밀리
  - 블룸 필터
  - 트랜잭션 처리

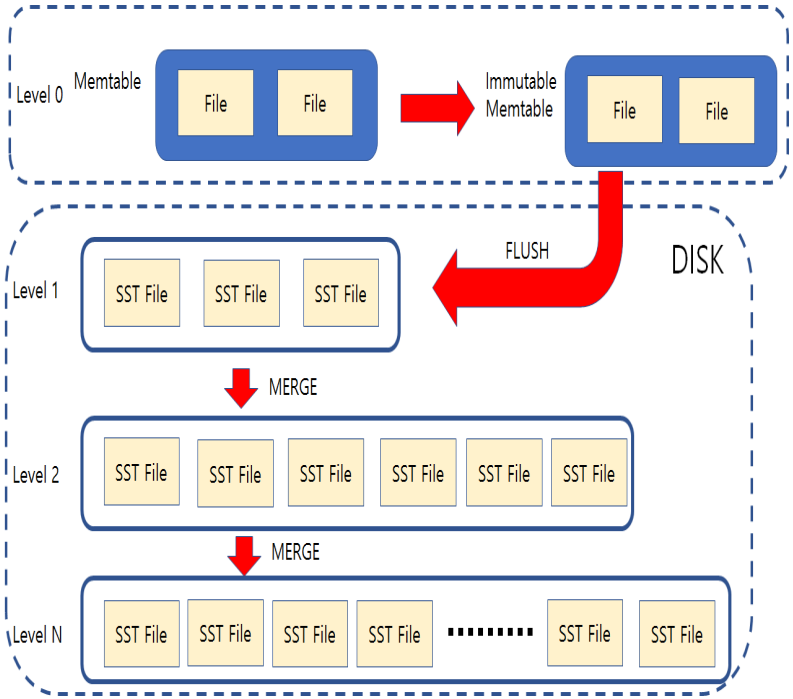
Log Structured Merge-Tree



- 장단점

장점	단점
Sequential Write 유도 디스크를 위한 별도의 로그 필요 없음	Read가 취약 시스템 오류 발생시 메모리 레벨의 데이터 사라짐

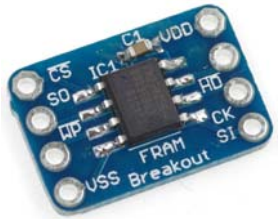
RocksDB Architecture



- 데이터 기록 전에 로그를 먼저 기록
- Memtable
  - 데이터가 기록되는 공간
- Immutable Memtable
  → Read-Only Memtable
- Static Sorted File
  - Key 순서에 따라 정렬된 파일
  - 디스크에 기록된 SST파일은 지워지지 않음
- 상위 레벨의 데이터 - 최신 데이터
- 하위 레벨의 데이터 – 오래된 데이터

NVRAM

- 특징
  - Non volatile
    - 전원이 차단되어도 메모리의 데이터 유지



- 종류
  - 메모리 자체가 비휘발성인 메모리
    - FeRAM, MRAM, FRAM,PRAM
  - 베터리 백업을 이용한 비휘발성 메모리

실험 환경

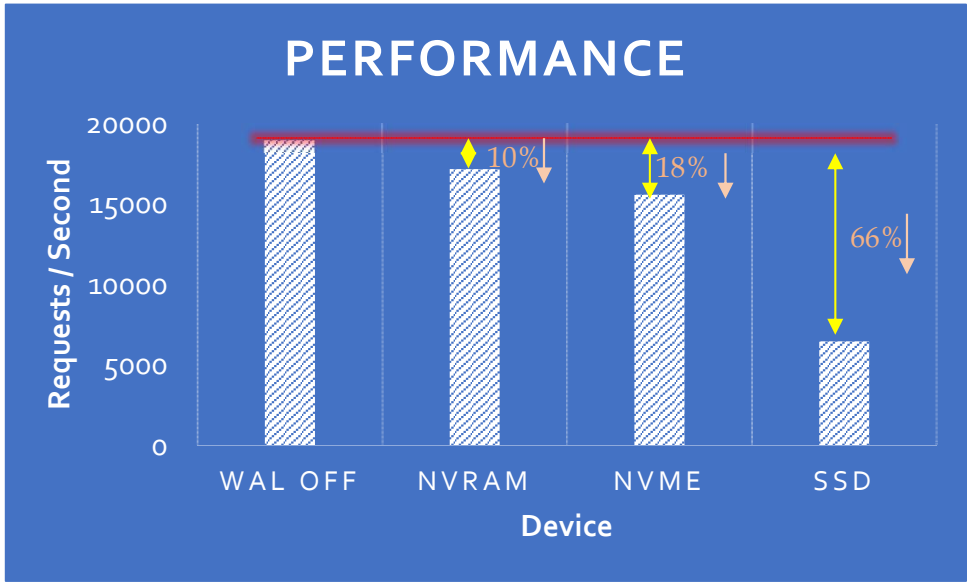
H/W

- OS
  - CentOS 7.3.1611 (x86_64)
- CPU
  - Intel(R) Xeon(R) CPU E5-2620 v2 @2.20GHz
- RAM
  - 64GB
- SSD
  - Crucial MX200 250GB SATA 2.5 256GB * 2
- NVME
  - Intel SSDPEDMD 800G4 DC P3700 800GB
- NVRAM
  - NV1600 Flashtec(TM) NVRAM

Data

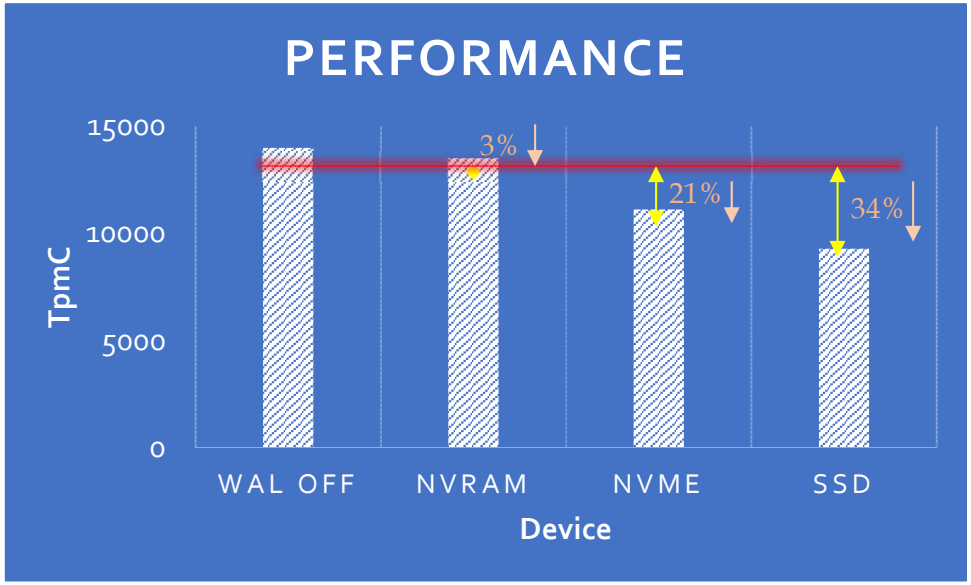
- LinkBench
  - 10,000,000개의 key-value 쌍
  - 3.5GB
- Tpc-C
  - Customer, District, History, Item, New_Order, Order_Line, Orders, Stock, Warehouse 데이터
  - Concurrent Client =10
  - 7GB

실험 결과- Linkbench



Device	Requests/second
SSD	6455
NVMe	15580
NVRAM	17174
WAL Off	18977

실험 결과– Tpc-C

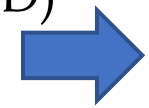


Device	TpmC
SSD	9270
NVMe	11100
NVRAM	13490
WAL Off	13977

연세대학교 YONSEI UNIVERSITY

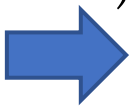
결론

- WAL Overhead (WAL OFF vs WAL on SSD)
  - On Linkbench - 66%
  - On Tpc-C - 34%



성능 저하에 상당한 비중을 차지

- WAL Overhead (WAL OFF vs WAL on NVRAM)
  - On Linkbench - 10%
  - On Tpc-C - 3%



NVRAM을 통한 성능 개선 확인

- 두 벤치마크에서 모두 비슷한 양상을 보임
- 디바이스 별로 WAL 로깅 성능 차이가 발생
  - 디바이스 WRITE 속도
    - SSD < NVMe < NVRAM

디바이스의 WRITE 성능은 WAL 로깅 오버헤드에 중요한 요인

Thank you

연세대학교
YONSEI UNIVERSITY