MyFlashSQL: Open Source Database for Non-Volatile Memory

Sang-Won Lee
New Non-Volatile Storages

- **NAND Flash Storage**
- **PRAM**
- **STT-MRAM**
- **NVDIMM**

**All-flash era in block storage interface**

- “Flash is disk, disk is tape, ..” [Jim Gray]
- Flash SSDs win over HDD even in $/GB [D. Patterson]

**Byte-addressable Non-volatile Storage**

- PRAM, STT-MRAM (?): not viable as of now
- NVDIMM [battery-backed DRAM] + Flash
MySQL on top of NVM

- MySQL (and also other RDBMSs) has been designed assuming HDD as main DB storage

- Thus, it will achieve suboptimal performance when NVMs including flash SSDs are used as faster HDDs
  - Mainly because it does not exploit the characteristics of NVM different from HDDs
  - Page size: 16KB $\rightarrow$ 32 / 8 / 4KB [since 5.6.4]
    - 2.5X performance by changing page size from 16KB to 4KB
    - Using LinkBench, TPC-C [DuraSSD, SIGMOD 2014]
NVM Opportunities for MySQL

- Flash Storage (mainly SSD)
  - Sequential write > random write
  - Intrinsic parallelism
  - FTLs and new semantic storage interfaces

- NVDIMM
  - Byte-addressability, fast durability (e.g. 1us / 64B)

- DB research community lags behind OS community in NVM research (especially using flash SSDs) 😞
MyFlashSQL: On-going works

- Flash as Cache Extension [FaCE]
  - FaCE for Postgres [VLDB 2012a]
  - FaCE for MySQL/InnoDB
    - No double write buffer
    - Both original DB and FaCE at flash SSD

- Preliminary result (20% Cache)
  - 1 SSD + 8 HDDs
    - 1029 TPS vs. 1903 TPS
  - 1 SSD
    - 2419 vs. 3879 TPS
MyFlashSQL: On-going works

- Leverage parallelism
  - Channel/package/die/plane-level: 16, 32, 64, ....
  - Psync [VLDB 2012b]

- E.g. range query
  - read latency: 6 vs. 1

SELECT *
FROM TABLE
WHERE A BETWEEN 19 AND 29

- IDX scan > full scan
  - Selectivity: up to 70%
MyFlashSQL: On-going works

- Orders Table in TPC-H

```sql
SELECT * FROM table FORCE INDEX (idx)
WHERE column_a BETWEEN min AND MAX;
```

SATA SSD: 4.5X

NVMe SSD: 10X
MyFlashSQL: On-going works

- New Flash SSD Interface [SIGMOD 2016]

- NoSQL ForestDB, Journaling/LFS/CoW file systems
- NVMe DSM, Multi-stream[HotStorage ‘14]
MyFlashSQL: Future works

- IPLization of MySQL on NVDIMM
  - In-Page Logging [SIGMOD 2007, ICDE 2011, VLDB 2015]
  - DuraSSD [SIGMOD 2014]: NVcache inside FlashSSDs

- Benefits
  - Write latency, Garbage collection, Life-time
  - Faster write implies better read latency in DB

Benefits

- Write latency, Garbage collection, Life-time
- Faster write implies better read latency in DB
Closing

- Welcome any feedback, research/develop collaboration

  - Sang-Won Lee (swlee@skku.edu)
  - FlashSQL Star Lab.
    - 2.4M/8yrs (funded by IITP)
    - http://flashsql.skku.ac.kr
  - Close research collaboration with Korean Industry