facebook

Web scale MySQL: A short history

Mark Callaghan

Database Infrastructure

What is web scale?

- Transaction processing, Small data
- User facing
- Reliable, available, durable
- Quality of service
- Sharded replica sets

Web scale MySQL at Facebook

Workload

- Secondary indexes
- Simple joins
- Short transactions
- Mix of hot, cold data
- Database >> RAM
- Concurrent

Per second peaks

- 175M QPS
- 140M disk pages read
- 12B rows read

Other

- ~4 ms response time
- Small ops team
- Tens of PB of data
- Availability: many nines

Why MySQL?

- There when I arrived
- We made it better

Define better

- Faster?
- More efficient?
- Easier to scale?
- More reliable?

MySQL in 2005

Software

- MySQL 4.0
- Benchmark: 20,000 QPS
- Weak monitoring
- Not always crash safe
- Oncall was exciting

Hardware

- 4 cores, 1 socket
- 16G RAM
- Disk array, 2000 IOPs

MySQL in 2015

Software

- MySQL 5.6
- Benchmark: 1M QPS
- Excellent monitoring
- Crash safe
- Fast, automated, lossless failover

Hardware

- 40+ cores, 2 sockets
- 256G RAM
- MLC flash

From 2005 to 2015: operations

- Data perf team
- Poor man’s profiler
- Pylander
- Dogpiled
- MPS
- Add covering index
- Query comments
- Ext3 -> XFS
- CFQ -> Deadline
- pmysql
- Bugs found

From 2005 to 2015: engineering

- InnoDB RW lock
- User/table statistics
- Semi-sync replication
- Crash-safe replica
- GTID
- Online schema change
- Online migration
- Async MySQL client
- Group commit
- Replication prefetch
- Flashcache
- InnoDB compression
- Innosim
- Lossless failover
- Linkbench
- Bugs fixed

The future

- RocksDB: NoSQL
- MyRocks: MySQL+RocksDB
- MongoRocks: MongoDB+RocksDB

RocksDB

- Log Structured Merge tree (LSM)
- Derived from LevelDB
- Simple API: Put, Get, Delete, Scan
- Optional support for transactions
- Faster writes, better compression

Thank you

- facebook.com/MySQLatFacebook
- facebook.com/groups/rocksdb.dev
- rocksdb.org
- groups.google.com/forum/#!forum/rocksdb
- smalldatum.blogspot.com
- twitter.com/markcallaghan

facebook