本文是作者在ACMUG 2016 MySQL年会上的演讲内容，版权归作者所有。

中国MySQL用户组（China MySQL User Group）简称ACMUG。
ACMUG是覆盖中国MySQL技术爱好者的一个技术社区，是Oracle User Group Community和MairaDB Foundation共同认可的MySQL技术社区。

我们关注MySQL，MariaDB，以及其他一切周边的开源数据库和开源工具，我们交流使用经验，推广开源技术，为开源贡献力量。

我们是开放社区，欢迎任何关注MySQL及其相关技术的人加入，我愿意跟其他任何技术组织和团体保持沟通和展开合作。

我们期望在我们的活动中大家都能以开心的、轻松的姿态交流技术，分享技术，形成一个良性循环，从而每个人都可以有一份收获。

ACMUG的口号：开源，开放，开心

关注ACMUG公众号，参与社区活动，交流开源技术，分享学习心得，一起共同进步。

Engineering that goes into making Percona Server for MySQL 5.6 & 5.7 different

Colin Charles, Chief Evangelist, Percona Inc.

colin.charles@percona.com / byte@bytebot.net

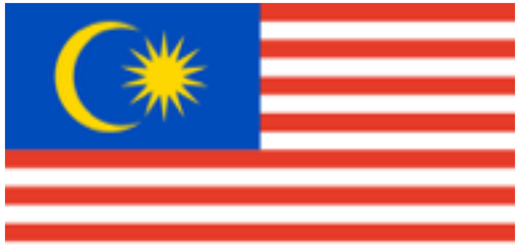
http://bytebot.net/blog/ | @bytebot on Twitter | bytebot on WeChat

ACMUG Meetup, Beijing, China

10 December 2016

PERCONA

whoami



- **Chief Evangelist** (in the **CTO office**), **Percona Inc**
  - Focusing on the MySQL ecosystem (MySQL, Percona Server, MariaDB Server), as well as the MongoDB ecosystem (Percona Server for MongoDB) + **100% open source** tools from Percona like Percona Monitoring & Management, Percona xtrabackup, Percona Toolkit, etc.
- **Founding team of MariaDB Server (2009-2016)**, previously at Monty Program Ab, merged with SkySQL Ab, now MariaDB Corporation
- Formerly MySQL AB (exit: Sun Microsystems)
- Past lives include Fedora Project (FESCO), OpenOffice.org
- MySQL Community Contributor of the Year Award winner 2014

Agenda

- Percona’s Aims
- Developing a branch of MySQL
- Adding features that make a difference
- Custom development
- Tools
- Sustainable engineering going forward

Percona’s Purpose

To Champion Unbiased Open Source Database Solutions

Percona Software Principles

All Percona Software is 100% Free and Open Source

No Restricted “Enterprise” version

No Open Core

No Software Licensing Games

Widely Deployed Open Source Software

PERCONA XtraBackup

2,100,000+ downloads

PERCONA Server for MySQL

3,000,000+ downloads

PERCONA XtraDB Cluster

780,000+ downloads

PERCONA Server for MongoDB

73,000+ downloads (released 9/2015)

PERCONA Toolkit

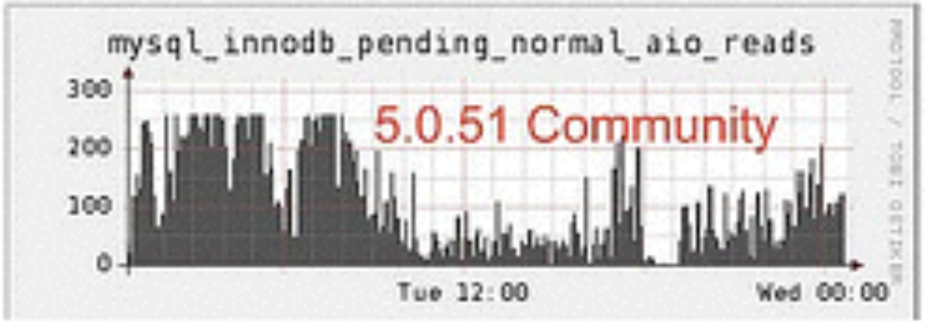
1,000,000+ downloads

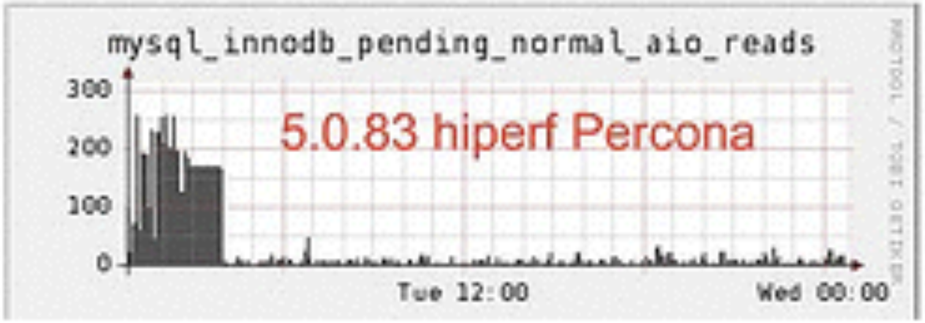
Aims

- **Branch**, not a fork
- Provide **alternatives** to proprietary parts of open source software
- Strong **operations** focus: compatibility, application scalability, high availability, security, observability

History (or how it all began)

- Percona Server: 8 years old — November 2008 (Percona SQL, Percona highperf builds)
- November 2008: Percona Server (patchset ~July)

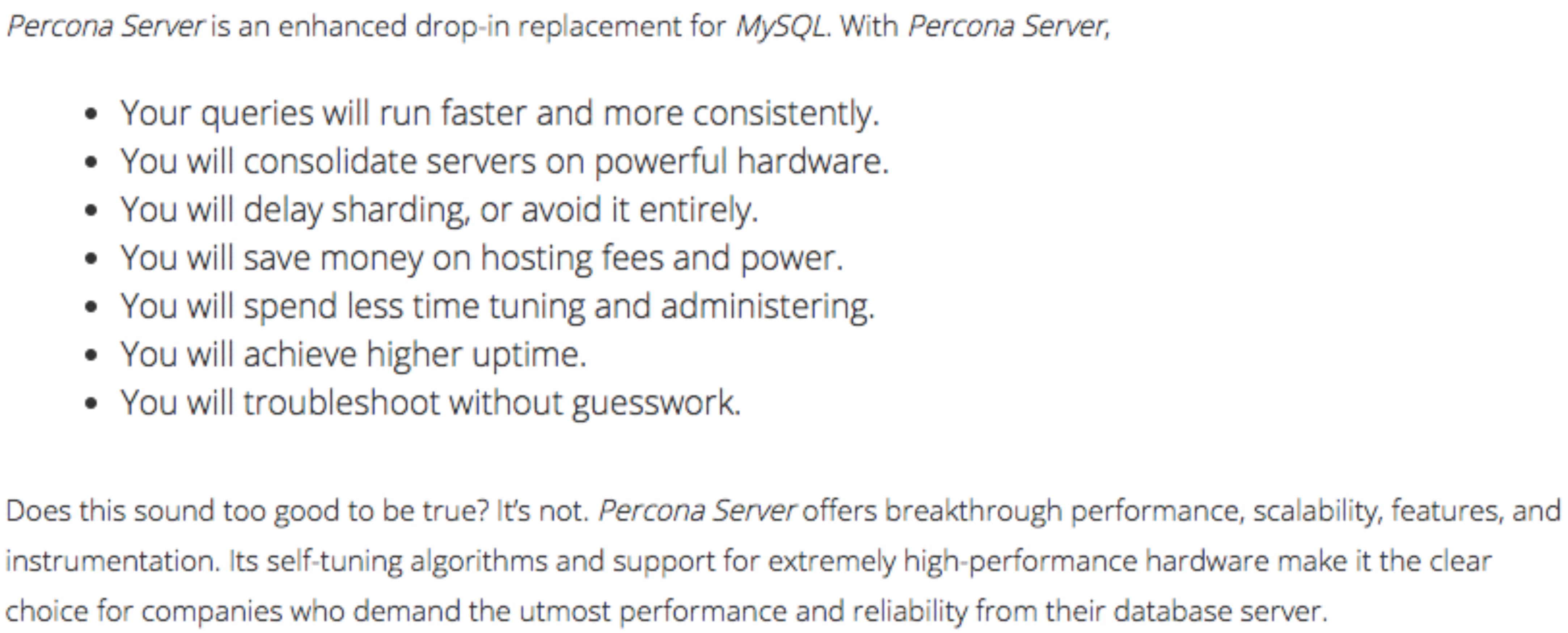




Timeline

- December 2010: MySQL 5.5 GA
- April 2011: Percona Server for MySQL GA
- February 2013: MySQL 5.6 GA
- October 2013: Percona Server for MySQL GA
- October 2015: MySQL 5.7 GA
- February 2016: Percona Server for MySQL GA

Straight from the docs



Percona Server is an enhanced drop-in replacement for *MySQL*. With *Percona Server*,

- Your queries will run faster and more consistently.
- You will consolidate servers on powerful hardware.
- You will delay sharding, or avoid it entirely.
- You will save money on hosting fees and power.
- You will spend less time tuning and administering.
- You will achieve higher uptime.
- You will troubleshoot without guesswork.

Does this sound too good to be true? It's not. *Percona Server* offers breakthrough performance, scalability, features, and instrumentation. Its self-tuning algorithms and support for extremely high-performance hardware make it the clear choice for companies who demand the utmost performance and reliability from their database server.

Software engineering

- Write a **detailed description** to a list linking to a **full specification design document** where people can comment within the document
- Official handover to **quality assurance** (QA)
  - In addition to `pquery` tests, specify what requirements there are
- **Documentation** is made easier thanks to design document and all the discussions that took place on list
- **Blog**, of course!

Development

- Jenkins
- Bazaar (Launchpad) and Git (Github) used
- Internally there is Jira (opening soon)
- Public lists are still more for end user support rather than development

Percona XtraDB - buffer pool, I/O scalability

- `SHOW ENGINE INNODB STATUS` extended output
- Global mutex protecting the buffer pool has been split into several mutexes to decrease contention - good when working set doesn't fit in memory
- Change the log block size, `innodb_flush_method=ALL_O_DIRECT`
    - use O_DIRECT to open data+log files, use fsync() to flush data files but not log/parallel doublewrite files (use when log files are > 8GB in size)

Percona XtraDB - I/O bound, concurrent workloads

- Legacy doublewrite buffer collects all page write requests into a single buffer, when filled, writes it out to disk twice blocking any new write requests. Bottleneck to flusher parallelism.
- Parallel doublewrite buffer
  - Private doublewrite buffers for each buffer pool instance, for each batch flushing mode
  - Only one flusher thread access any shard at a time
  - Each shard is added or flushed independently of the rest

Percona XtraDB - I/O bound, concurrent workloads

- Multi-threaded LRU flusher
    - each buffer pool instance has its own dedicated LRU manager thread performing LRU flushes/evictions to refill the free list of that buffer pool instance
- Priority refill of the buffer pool free list
    - a backoff algorithm to reduce buffer pool free list mutex pressure on empty buffer pool free lists

Percona XtraDB - changed page tracking

- Tracks the pages that have changes written to them according to the redo log
- Speed up incremental backups made with Percona XtraBackup - no need to scan whole data files any longer, sequence number can be used to check presence of required bitmap files
- Changed page tracking is done by a new XtraDB worker thread that reads and parses log records between checkpoints.

Backup locks

- Lightweight alternative to `FLUSH TABLES WITH READ LOCK`
- New MDL lock type to block updates to non-transactional tables and DDL statements for all tables
  - SEs not forced to close tables and tables not removed from table cache; only waits for conflicting statements to complete (in InnoDB, SELECTs or UPDATEs aren't waited for)
- `LOCK TABLES FOR BACKUP, LOCK BINLOG FOR BACKUP, UNLOCK BINLOG`

START TRANSACTION WITH CONSISTENT SNAPSHOT

- Make binary log positions consistent with InnoDB transaction snapshots
  - `binlog_snapshot_position, binlog_snapshot_file` - binlog position corresponding to the state of the database consistent snapshot, irrespective of other transactions committed since snapshot taken
- Obtain logical backups with correct positions without using `FLUSH TABLES WITH READ LOCK`
- … `FROM SESSION <sessionID>`
- `mysqldump —single-transaction —master-data`

NUMA Support

- Avoids the MySQL “swap insanity” problem
- Buffer pool memory bigger than size of node, system would swap allocated memory even if there was available memory on other node. NUMA **interleaving** solves this
- Exists upstream, but has a `flush_caches` variable - flush and purge buffers/caches before starting the server to help ensure NUMA allocation fairness across nodes
- Useful for establishing consistent and predictable behaviour for normal usage/benchmarking

Threadpool

- `thread-handling = pool-of-threads`
- Priority Scheduling
  - tickets given to determine if it goes into a high priority or low priority mode
- Ensures that you never reach the oversubscribe limit

Per query variable statement

- `SET STATEMENT max_execution_time=1000 FOR SELECT name FROM name ORDER BY name;`
- Arrived in Percona Server 5.6, originated as Google Summer of Code 2009 project!

Using HAProxy?

- PROXY protocol implemented
- Allows an intermediate proxying server speaking the proxy protocol between the server and the ultimate client to provide the source client address to the server, which normally would only see the proxying server address instead.

Automated manageability enhancements

- `enforce_storage_engine`
- Utility user
  - user who has system access to do administrative tasks but limited access to user schema

Other features

- Extended `SHOW GRANTS`
- User statistics
- Extended slow query log - microsecond time, additional statistics (for `pt-query-digest`), enable/disable at runtime, logging for slave SQL thread, log rotation/expiration
- Response time distribution plugin - large number of small running queries?

- Kill idle XtraDB transactions
- Disable corrupted XtraDB tables while getting fixed (server continues humming along)
- `SELECT INTO OUTFILE/DUMPFILE` supports UNIX sockets, named pipes
- `max_binlog_files`
- Improved MEMORY and CSV storage engines
- `mysqlbinlog` with SSL options
- Per session server IDs

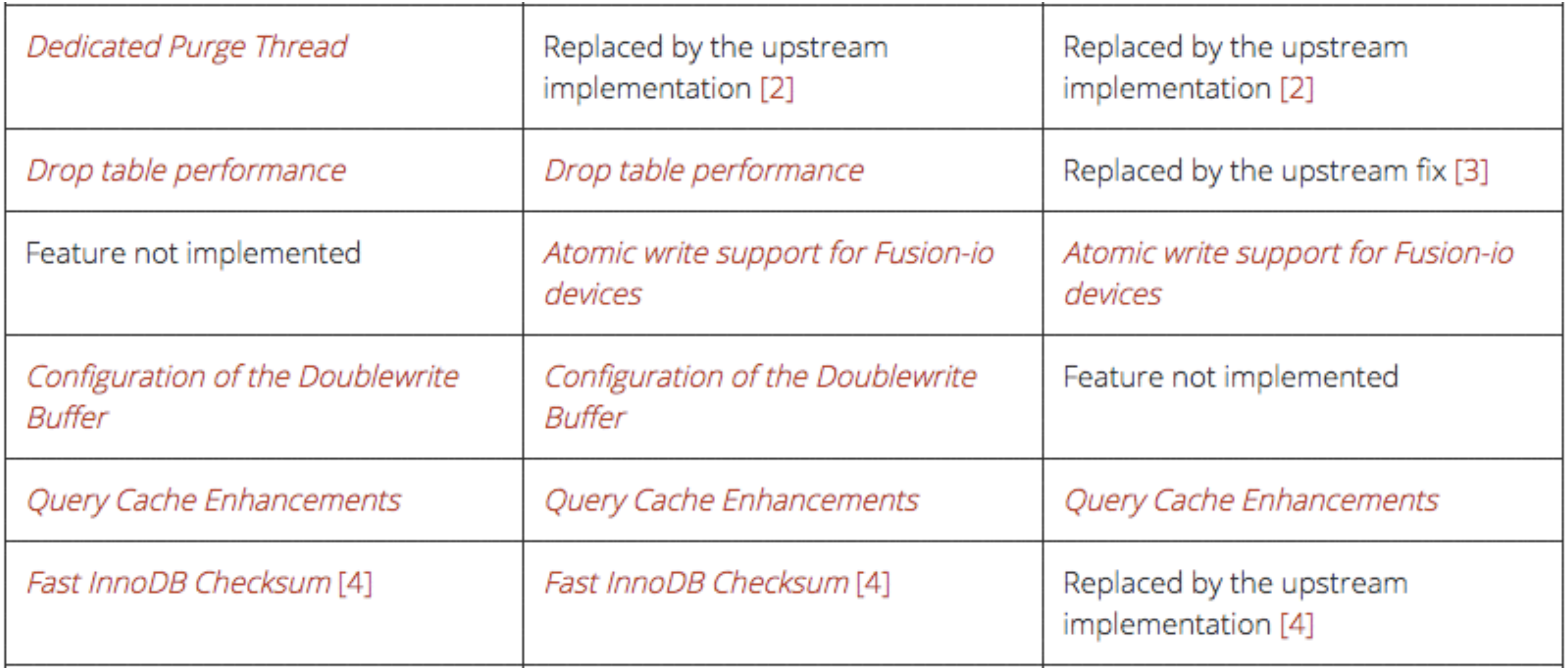
Other Enterprise features made open

- PAM Authentication plugin
  - `auth_pam`, using `dialog.so`
  - `auth_pam_compat`, uses `mysql_clear_password` to be fully MySQL compliant
- Audit plugin
  - Multiple formats (but remain compatible - `mysqlauditgrep`), stream to syslog, filter by user or SQL command type (or database)
  - Performance matters - so strategy can be asynchronous (log memory buffer, do not drop messages if buffer is full), performance, semi-sync (log to file, do not flush and sync every event) or synchronous

Unafraid to remove features

- Scalability metrics plugin deprecated
- Reflex adaptive checkpoint
- Prefer using the **upstream implementation** rather than Percona's own engineering work
  - Better for maintenance

Replaced by upstream fix!



Dedicated Purge Thread	Replaced by the upstream implementation [2]	Replaced by the upstream implementation [2]
Drop table performance	*Drop table performance*	Replaced by the upstream fix [3]
Feature not implemented	*Atomic write support for Fusion-io devices*	*Atomic write support for Fusion-io devices*
Configuration of the Doublewrite Buffer	*Configuration of the Doublewrite Buffer*	Feature not implemented
Query Cache Enhancements	*Query Cache Enhancements*	*Query Cache Enhancements*
Fast InnoDB Checksum [4]	*Fast InnoDB Checksum* [4]	Replaced by the upstream implementation [4]

Some features start life in 5.6 first…

- TLS v1.1/v1.2, disabling v1.0
- HandlerSocket - not going to 5.7
- Online GTID deployment using the “step mode” that Facebook created - not needed in 5.7

Compressed columns

- Data type modifier - compressed on writing, decompressed on read
- XtraDB, and column limitations: BLOB/TEXT/VARCHAR/VARBINARY
- Pre-define a compression dictionary
- https://engineering.pinterest.com/blog/evolving-mysql-compression-part-1
- Thank you Alibaba / Weixiang Zhai

Percona TokuDB

- Integrated into the server
- Percona TokuBackup - no locking of database, intercepts syscalls that write files and duplicates the writes to backup directory
- File layout option
- “This means that within 3-5 years MyRocks could possibly be superior to TokuDB in functionality and performance for all imaginable workloads.”
- https://www.percona.com/blog/2016/11/01/future-tokudb-percona/

Percona XtraDB Cluster with ProxySQL

- Integrated cluster aware load balancer
- Instrumentation with `PERFORMANCE_SCHEMA`
- Full support for data at rest encryption (InnoDB Tablespace Encryption)
- Data is safe by default “strict mode” - features deemed to work incorrectly by upstream, disabled
- Full integration with PMM

MyRocks

- Going from Facebook MySQL 5.6 to Percona Server 5.6 first, then to Percona Server 5.7
- Integrations with Facebook MySQL specific features will be either stripped out and re-implemented/re-integrated with corresponding Percona Server features, or, Facebook MySQL features will be imported into Percona Server as needed
- Goal is to produce a compilable, testable, experimental binaries based on Percona Server 5.6 prior to beginning the port to Percona Server 5.7

MyRocks II

- Q1/2017: Usable branch that compiles by end of Q1/2017, experimental builds by January. Development will continue all of 2017
- `cmake -DWITH_ROCKSDB`
- Identified about 30 compilation issues with Percona Server 5.6
- Working branch currently passes about 80% of MyRocks mtr

MyRocks III - 5.6 remaining

- Differences in group commit implementation. Leads to difference in implementation of `START TRANSACTION WITH CONSISTENT SNAPSHOT` and MyRocks specific `mysqldump` commands
- Different implementations of Read Free Replication and Unique Check Ignore type functionality.
- Different implementations of Crash Safe Slave
- Refactoring compression library linking
- Mixed engine use cases and testing
- Getting to clean mtr runs
- Port to 5.7 will start before 2016 is over

Percona Monitoring & Management

- Built on top of the best of breed open source software tools like Grafana, Consul, Prometheus, and Orchestrator
- Supports MySQL and MongoDB
- Trending and Query Analysis
- http://pmmdemo.percona.com/

Conclusions

- Didn’t cover Percona Server for MongoDB (with MongoRocks)
- Didn’t cover Percona Toolkit (useful since 2011!), Percona XtraBackup (de-facto hot backup solution)
- Didn’t cover the Percona Monitoring plugins (Cacti/Nagios)
- Percona is supporting all software and isn’t stopping (EOL information is published well in advance)
- Percona Server for MySQL development is not going away

Percona is Hiring!

- http://www.percona.com/careers/
- Remote DBAs, C/C++ developers, Go developers, etc.

Percona Live Santa Clara 2017

- https://www.percona.com/live/17/
- April 24-27 2017, Santa Clara, California, USA
- Early bird ends January 2017
- Want a **discount code**? E-mail colin.charles@percona.com

Thank you!

Colin Charles

colin.charles@percona.com / byte@bytebot.net

http://bytebot.net/blog | @bytebot on twitter | bytebot on WeChat

slides: slideshare.net/bytebot