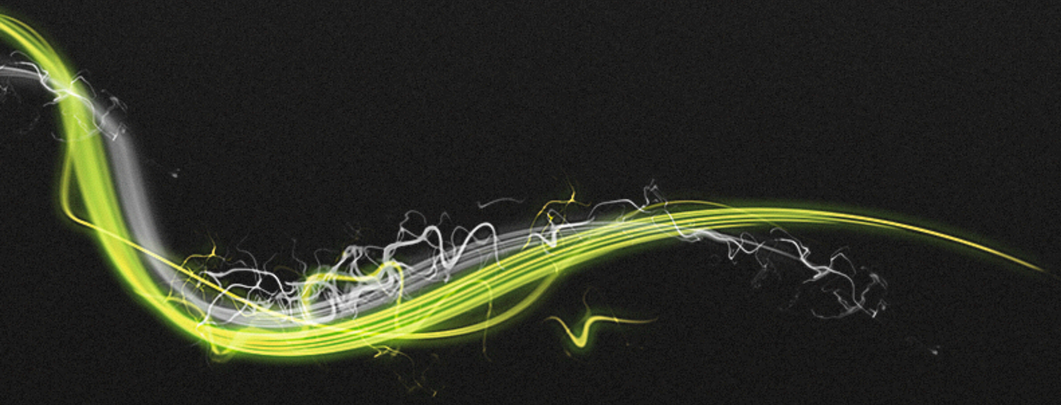
SysEleven
Hosting. Skaliert.

MySQL 5.6 InnoDB

Kristian Köhntopp



InnoDB

SysEleven
Hosting. Skaliert.

- Transactional Engine (ACID).
- Mostly lockless.
- Checksums and Crash Recovery.

Using InnoDB

SysEleven
Hosting. Skaliert.

- CREATE TABLE t (...) ENGINE=InnoDB;
- ALTER TABLE t ENGINE=InnoDB;

- Must not have FULLTEXT, GIS type.
- Must not have compound PK with auto_increment.

Transactions

SysEleven
Hosting. Skaliert.

- **START TRANSACTION** [READ ONLY|READ WRITE|WITH CONSISTENT SNAPSHOT];
  - do stuff
- **COMMIT**;

- Changes visible to others on commit (Isolation).
- Changes undone on **ROLLBACK** (or disconnect).

Transactions

SysEleven
Hosting. Skaliert.

- By Default: **SET AUTOCOMMIT = 1;**
  - ; is an implicit commit.
  - Enter explicit transaction w/ START TRANSACTION even in autocommit mode.
- DDL such as CREATE/ALTER TABLE implies COMMIT.

What a write does

SysEleven
Hosting. Skaliert.

id	d	txn#
1	one	2
2	zwei	1
3	drei	1

Table t

id	d	txn#
1	eins	1

Undo Log

UPDATE t SET d = 'one' WHERE id = 1

What a write does

SysEleven
Hosting. Skaliert.

- The insert pushes the old data into the undo log.
- The new data is put into place.
- On commit, nothing needs to be done.
- On rollback, the old row needs to be copied back.
  - This used to be slow - 1000 to 10000 rows/txn, much improved.

What reads do

SysEleven
Hosting. Skaliert.

- SET TRANSACTION ISOLATION LEVEL …
  - Can be set per connection.
  - Is a reader thing, only.
  - Write into undo log is necessary anyway for rollback.

What reads do

SysEleven
Hosting. Skaliert.

- … READ UNCOMMITTED;
  - Reads from the table, always.
  - Can return 'illegal data' that has never been there, logically.

What reads do

SysEleven
Hosting. Skaliert.

- … READ COMMITTED;
  - Follows roll pointer one deep.

  - Only one write per row at any point in time.
  - If roll pointer exists, it points to comitted data, always.

What reads do

SysEleven
Hosting. Skaliert.

- … REPEATABLE READ;
  - Reader is in a transaction, too:
  - START TRANSACTION … SELECT … SELECT … COMMIT
  - “Follow the roll pointer chain and get the newest version that is older than the start of the readers transaction.”

Shrinking the UNDO log

SysEleven
Hosting. Skaliert.

- For all transactions in the system:
  - Determine the oldest txn#.
  - Purge all records from Undo log that are older than this txn#.

- SHOW ENGINE INNODB STATUS\G

Shrinking the UNDO log

SysEleven
Hosting. Skaliert.

- START TRANSACTION WITH CONSISTENT SNAPSHOT;
  - Go on vacation.
  - What happens?

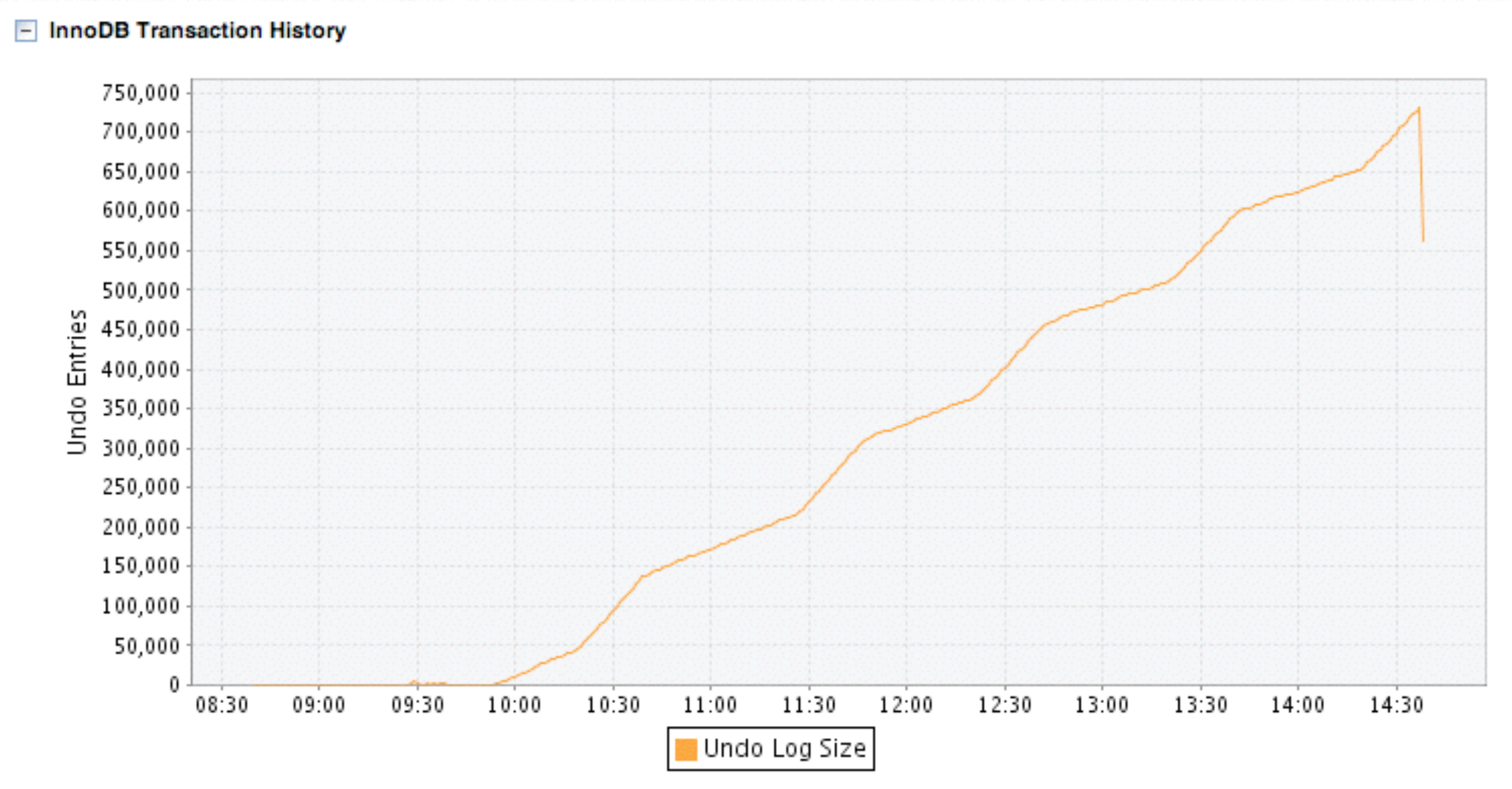
Shrinking the UNDO log

SysEleven
Hosting. Skaliert.

- START TRANSACTION WITH CONSISTENT SNAPSHOT;
  - Go on vacation.
  - What happens?
- wait_timeout / interactive_timeout = 28800;
  - Undo Log always part of ibdata1.
  - Should be set to autoextend.

Shrinking the UNDO log

SysEleven
Hosting. Skaliert.



Shrinking the UNDO log

SysEleven
Hosting. Skaliert.

- mysql> pager grep ACTIVE
  mysql> show engine innodb status\G
  ...
  ---TRANSACTION A90E003AB, ACTIVE **16830** sec, process no **12098**, OS thread id 1749563712
  ...
  mysql> pager less
  mysql> show engine innodb status\G
  ...
  ---TRANSACTION A90E003AB, ACTIVE **16830** sec, process no **12098**, OS thread id 1749563712
  3 lock struct(s), heap size 1248, 2 row lock(s), undo log entries 1
  MySQL thread id **146473882**, query id 4156570244 mc02cronapp-01 192.168.1.10 cron_bp
  Trx read view will not see trx with id >= A90E14DE8, sees < A90E14DE8

Shrinking the UNDO log

SysEleven
Hosting. Skaliert.

- mysql> pager grep cron
  mysql> show processlist;
  …
  | **146473882** | cron_bp | mc02cronapp-01:50154 | bp | Sleep | 16434 |
  …
- ssh mc02cronapp-01 ‘lsof -i -n -P | grep 50154’

- Und dies ist der Prozeß, der gekillt werden muß!

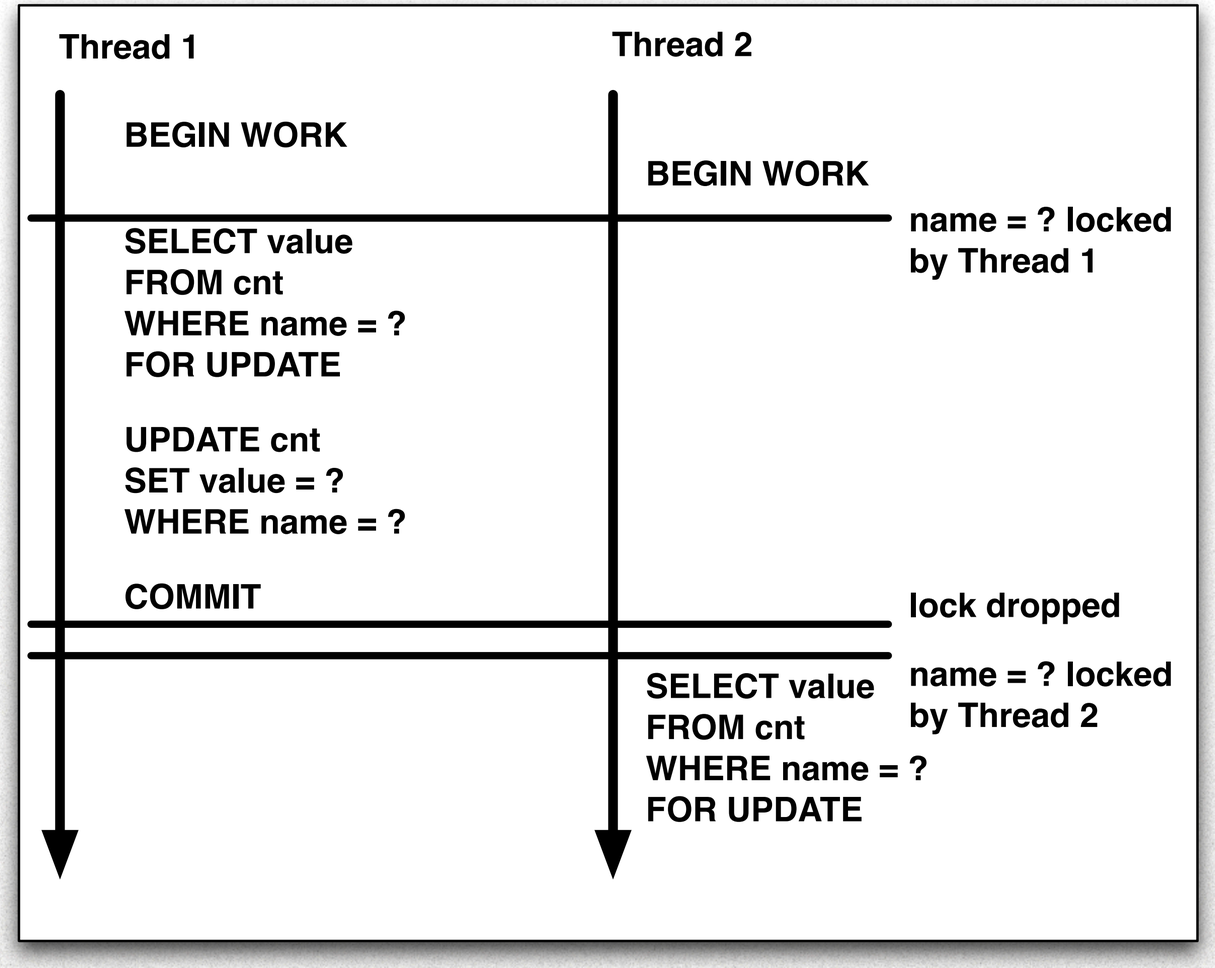
Concurrent writes

SysEleven
Hosting. Skaliert.

- Implement a counter in the application
  - **SELECT value FROM cnt WHERE name = ?**
  - Calculate new value.
  - **UPDATE cnt SET value = ? WHERE name = ?**
- Does not work without locking.

Concurrent writes

SysEleven
Hosting. Skaliert.



Concurrent writes

SysEleven
Hosting. Skaliert.

- Concurrent writes are handled by
  - SELECT … FOR UPDATE
- Is a read statement, locks like a write.
- Locks are taken via the index,
  have a close look at the execution plan.
  - SELECT … WHERE id IN (…) FOR UPDATE

Isolation Level #4

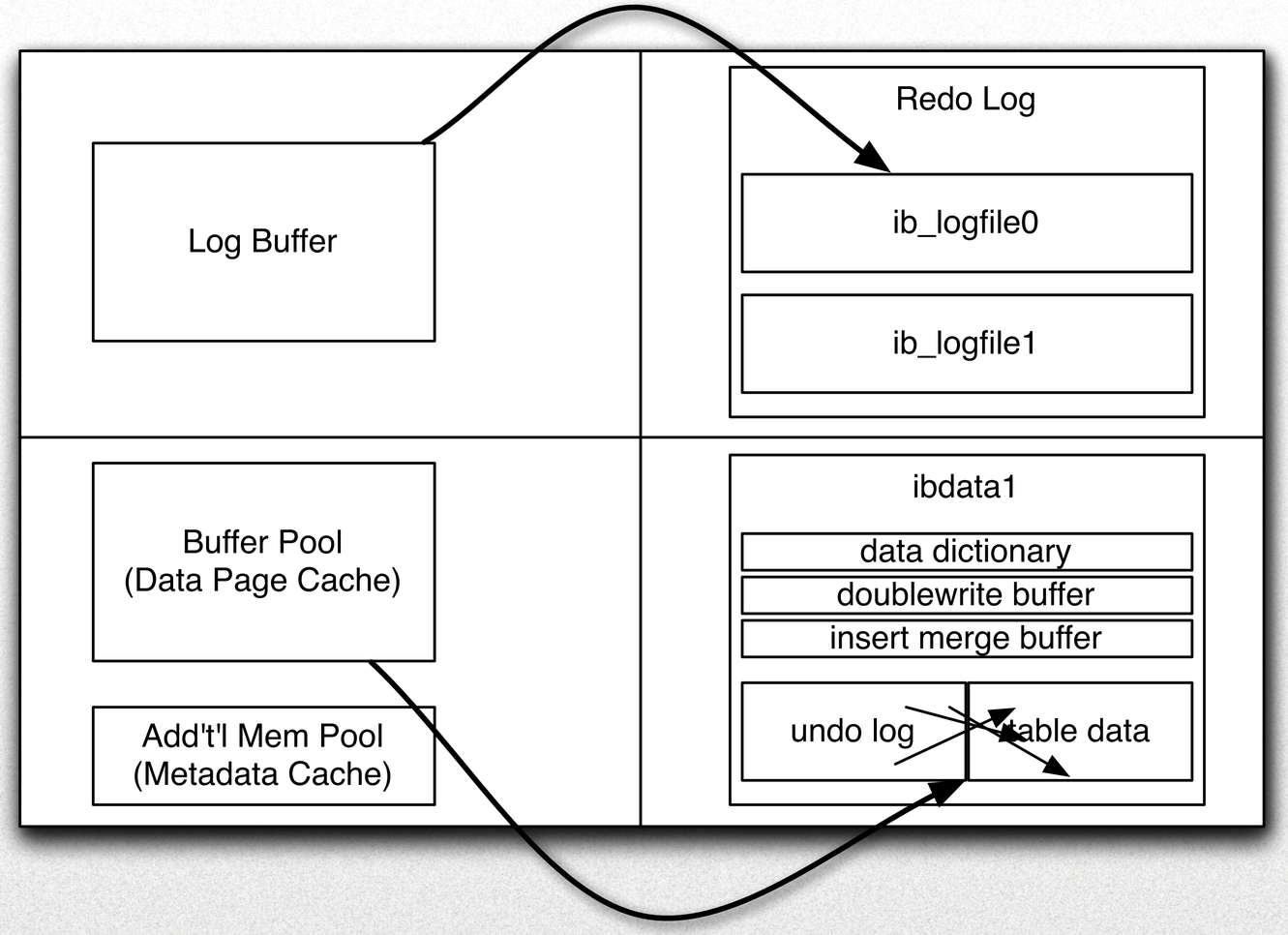
SysEleven
Hosting. Skaliert.

- … SERIALIZEABLE
  - Like an implied SELECT … IN SHARE MODE.

- Never needed:
  - …FOR UPDATE: create X-Lock on every row.
  - … IN SHARE MODE: create S-Lock on every row.

Commit and Checkpoint

SysEleven
Hosting. Skaliert.



Commit and Checkpoint

SysEleven
Hosting. Skaliert.

- Change a record:
  - Load page
  - Modify page, move old data to Undo Log
  - Create Redo Log Record in Log Buffer

Commit and Checkpoint

SysEleven
Hosting. Skaliert.

- On Commit:
  - Write Log Buffer to Redo Log

Commit and Checkpoint

SysEleven
Hosting. Skaliert.

- On checkpoint:
  - Find records from Redo Log
  - Determine dirty pages in Buffer Pool referenced
  - Flush pages to Doublewrite Buffer
  - Free Redo Log
  - Flush pages to tablespaces

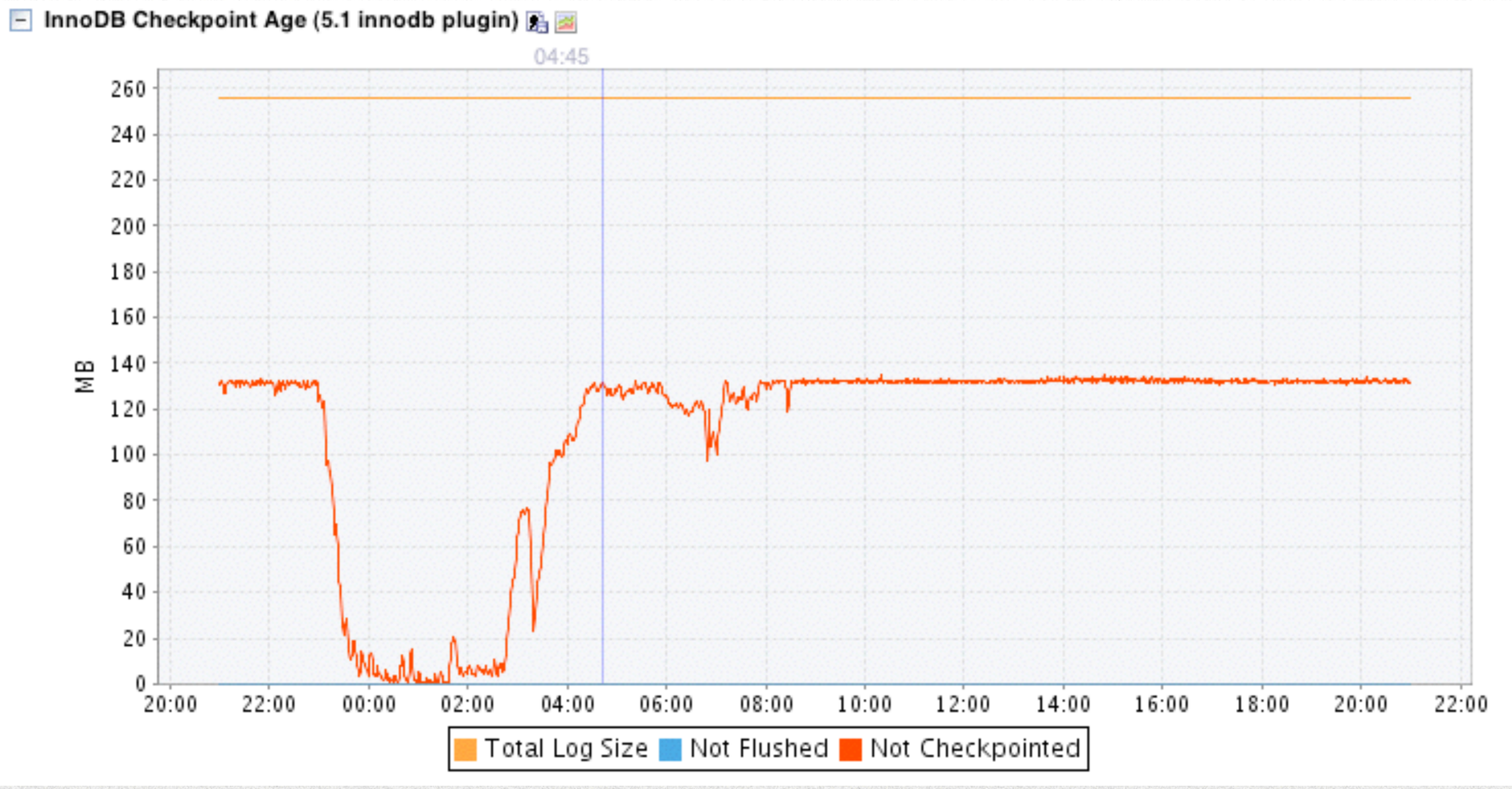
Commit and Checkpoint

SysEleven
Hosting. Skaliert.

- Checkpointing - when?
  - When mysqld feels idle.
  - When innodb_max_dirty_pages_pct exceeded.
  - When Redo Log is full.
- Many improvements in 5.6:
  - Large buffer pools, writes to SSD, bandwidth mgmt.

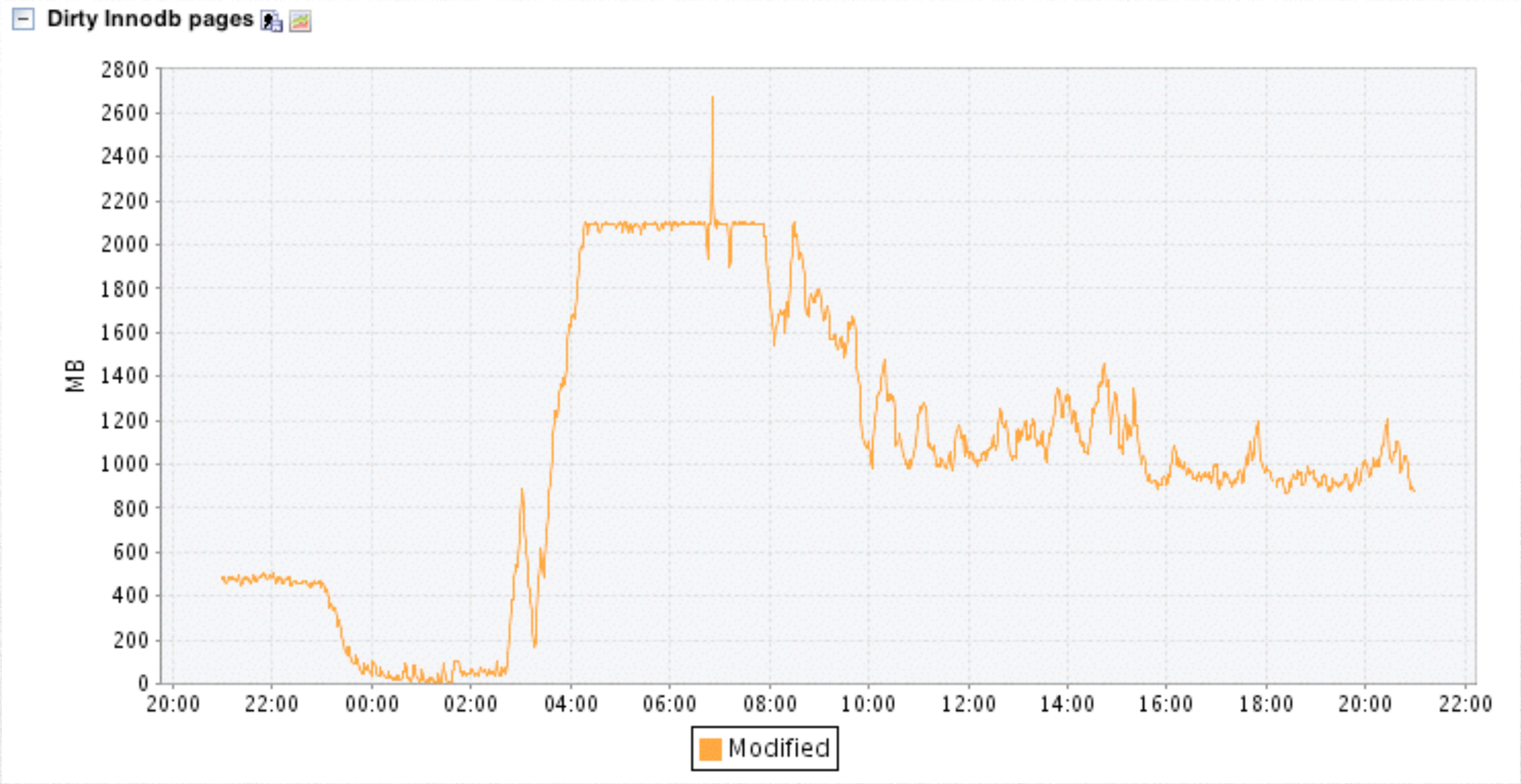
Commit and Checkpoint

SysEleven
Hosting. Skaliert.



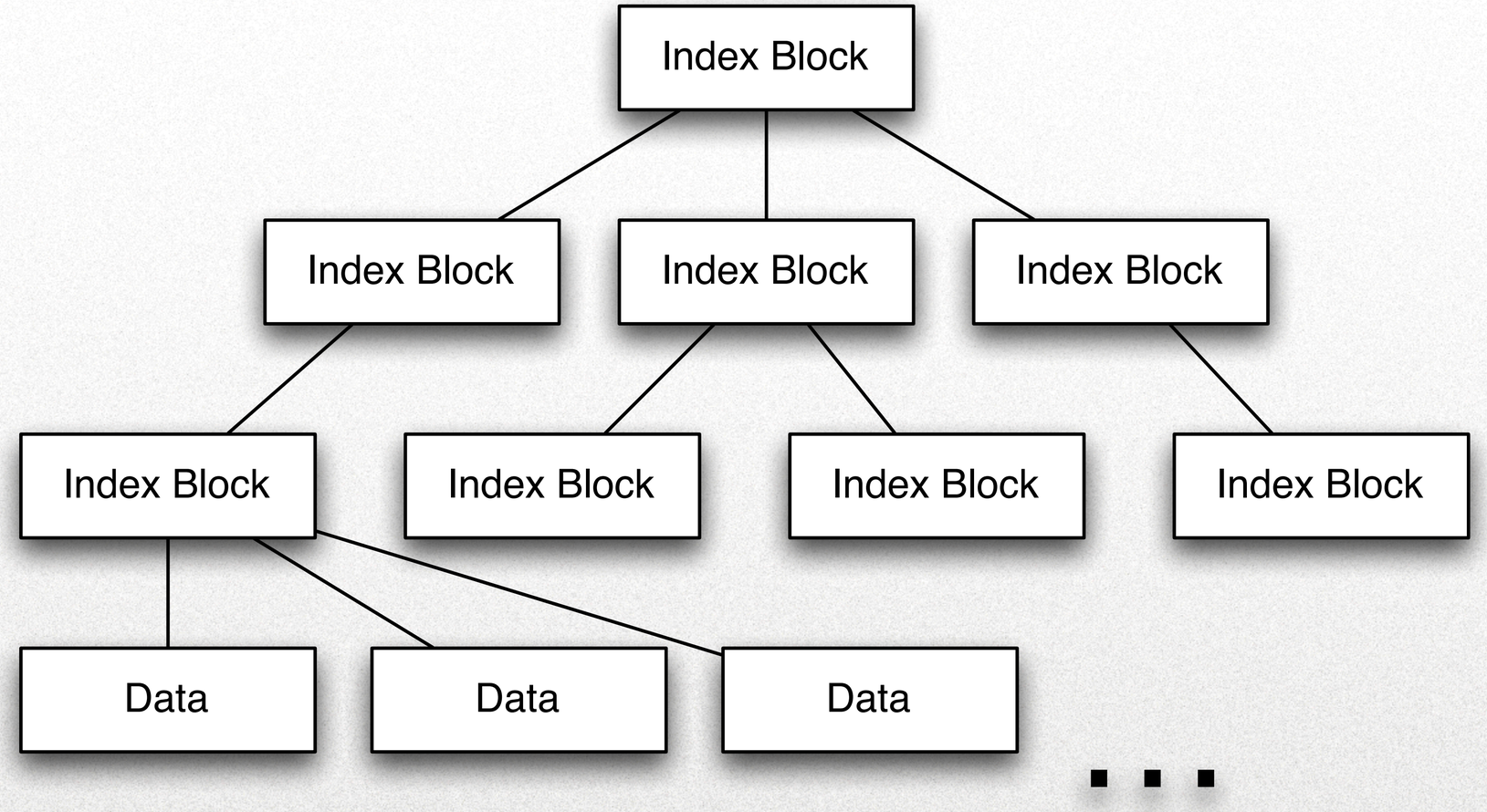
Commit and Checkpoint

SysEleven
Hosting. Skaliert.



Primary Key

SysEleven
Hosting. Skaliert.



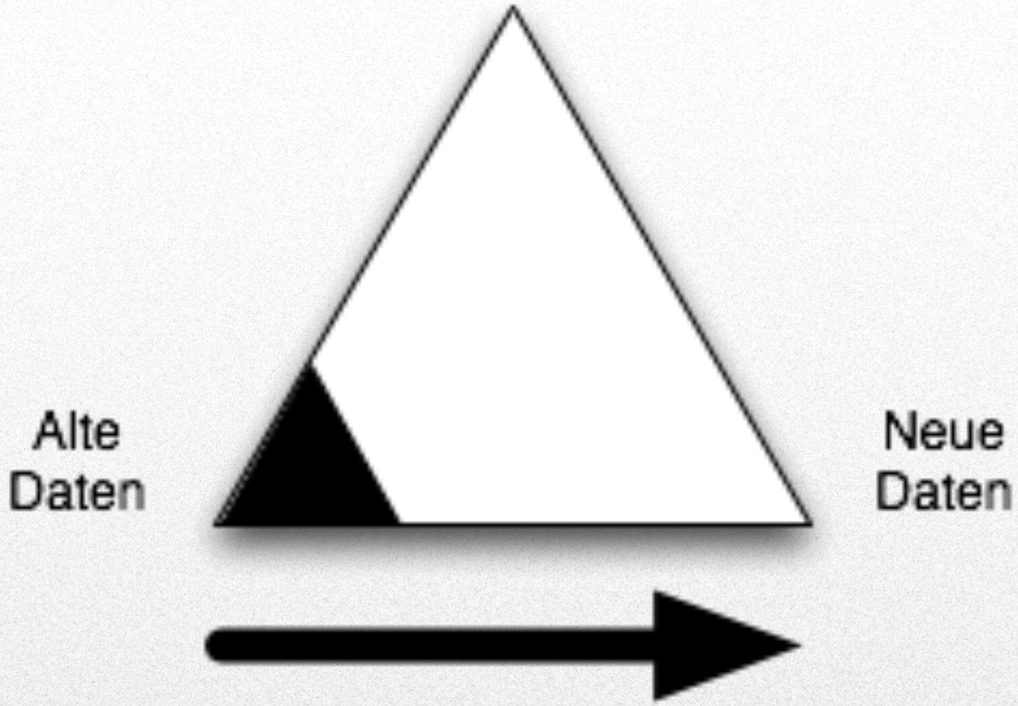
Primary Key

SysEleven
Hosting. Skaliert.

- “There is no MYD”: Data in leaves of PK - data_size is PK.
  - Data is stored in PK Order.
  - auto_increment always inserts at the right end of the table.
  - The tree is kept balanced: Lots of reorders.
  - Change buffer optimization fixes this.

Primary Key and auto_increment

SysEleven
Hosting. Skaliert.

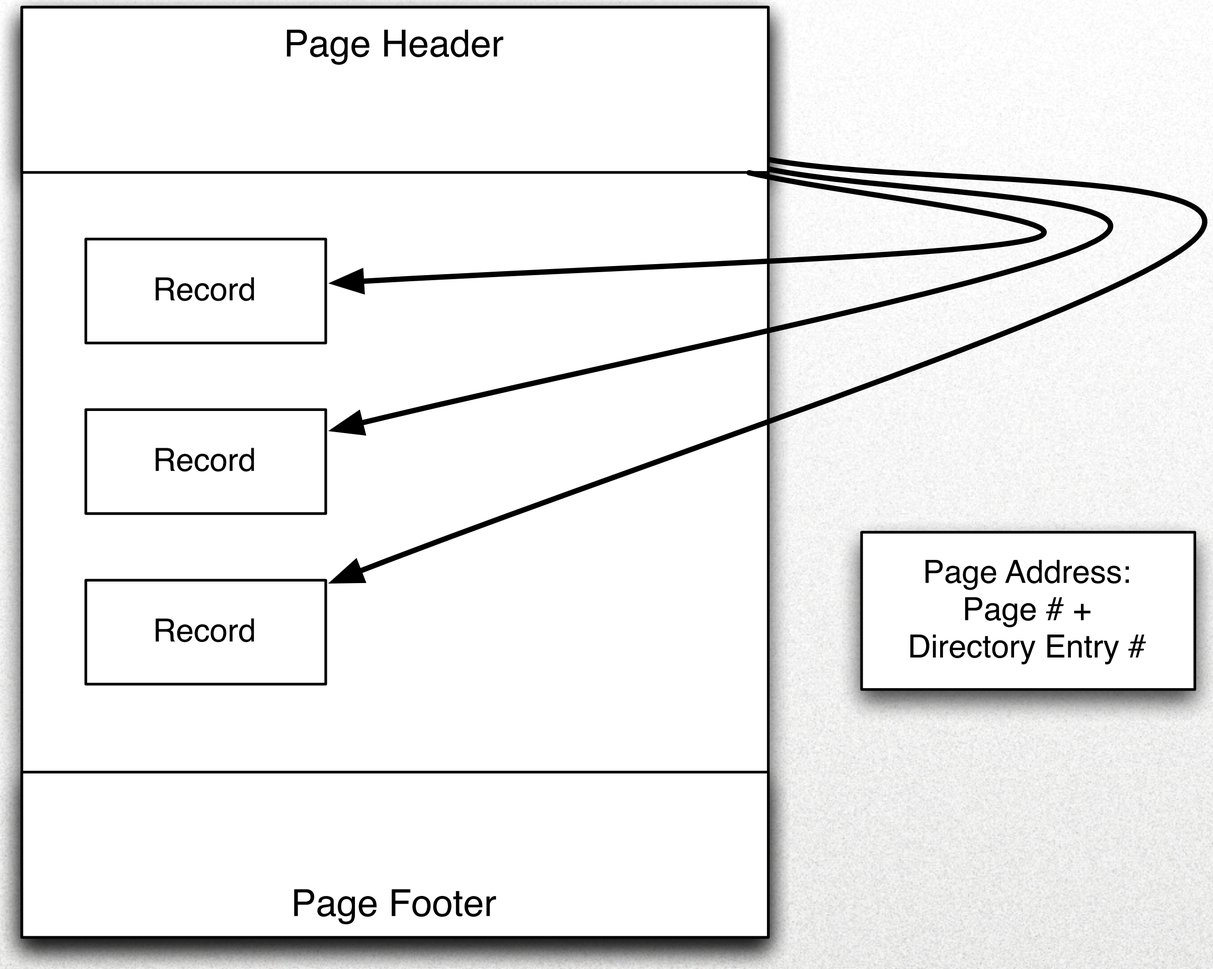


Secondary Indexes

SysEleven
Hosting. Skaliert.

- Secondary Indices use the Primary Key as a row pointer.
  - Size Limit 64T for other reasons.
- Keep the PK short, it is part of every Index.
  - The optimizer can use this.

Page Structure



Page Structure

SysEleven
Hosting. Skaliert.

- All I/O done in 16K pages.
- Page:
  - Page Header w/ Page Directory
  - Many Records + free space
  - Page Footer w/ Checksum

Page Structure

SysEleven
Hosting. Skaliert.

- SK Row Pointer = PK
  - Does not change when record relocates.
- PK Records: Page# + Page Directory Slot#
  - Does not change when record moves around in page.

Page Structure

SysEleven
Hosting. Skaliert.

- Page Structure avoids Index Update Storms
  - Records 8-15KB full (always some empty space)
  - Strings can be grown in place most of the time.
  - Page splits do not affect secondary indices (rebalancing the PK is enough).

Configuration

SysEleven
Hosting. Skaliert.

- innodb_buffer_pool_size = ...
  - As large as possible.
- Needs vm.swappiness = 0 in /etc/sysctl.conf
  - innodb_log_buffer_size = 8..32M
  - innodb_additional_mem_pool_size = 16M

Configuration

SysEleven
Hosting. Skaliert.

- innodb_log_files_in_group = 2
- innodb_log_file_size = <buffer pool/6>
  - Total Limit: 4G (gone in 5.6 and higher)
- innodb_file_per_table = 1
- innodb_data_file_path =ibdata1:10M:autoextend
- innodb_autoextend_increment = 8

Configuration

SysEleven
Hosting. Skaliert.

- innodb_flush_log_at_trx_commit = 0|1|2
- 1 = ACID: Commit writes to OS, OS flushes to disk
- 2 = ACID w/ software failure
  - Commit writes to OS, OS flushes every second
- 0 = not ACID
  - Commit is logical op, write every second

SysEleven
Hosting. Skaliert.

innodb_flush_log_at_trx_commit = 0

“Still a better database than MongoDB”

–Kristian Köhntopp

Monitoring

SysEleven
Hosting. Skaliert.

- SHOW ENGINE INNODB STATUS\G
- SHOW GLOBAL VARIABLES LIKE 'inno%';