Locality of (p)reference

Some thoughts about MySQL consulting issues

Oli Sennhauser
Senior Consultant
osennhauser@mysql.com
ToC

• Locality of (p)reference
• commit_demo.pl (performance test/numbers)
• InnoDB information (discussion)
• RAM disk
• MySQL variables (discussion)
• MyISAM log
• MySQL Visual Explain
Locality of (p)reference

• In theory: We should not care how data are stored internally.
• In practice: It is sometimes good to know!
• Why?
• 2 examples from the last 9 months:
  – wind mills
  – vehicle tracking for parcel delivery
Example 1

• Several 100 wind mills
• 50 measured values per wind mill
• Every 5-15 minutes
• Up to 10 years
• Dozens of GB of data
• Record size up to 2k!

• Search pattern: Give me value x from wind mill #13 in this time range!
Example 2

• Several 100 vehicles
• 24 h/d
• Every 2 min position
• Status/position per vehicle, later per parcel!!!
• Dozens of GB of data
• Record size 400 bytes

• Search pattern: Give me all positions of vehicle #13 from the last 24 hours.
Locality of Reference

• These 2 examples have one behaviour in common:

• Delivery of data is completely different than search pattern.
  – Usually data are delivered sorted by time and also (more or less) retrieved by time.
  – In this cases time has a secondary influence!

• But what happens???
Locality of Reference

• Block size is 16k/4k
• PK is AUTO_INCREMENT

• Synthecial PK are sometimes dangerous!
Locality of Reference

• What to do???

➡ PK on (vehicle_id, ts) for example or
➡ PK on (windmill_id, data, ts)

➡ Can be up to 100 times more efficient (not necessarily faster)

• What about MyISAM?
• What about Falcon? (Mail from Ann can be provided).
commit_demo.pl

• Little ugly script to test your I/O system:
  http://www.shinguz.ch/MySQL/consulting_tools.html

• What it does:
  INSERT -> COMMIT -> INSERT -> COMMIT -> ...

• The idea behind it

• How to call:

```
./commit_demo.pl -u root -c
./commit_demo.pl -u root -i <n>
./commit_demo.pl -u root -c
```

• Does NOT work with 6.0 :-(

• Lit: http://www.shinguz.ch/MySQL/transaction_performance.pdf
commit_demo.pl

trx time over 8 h in a VM/SAN
### commit_demo.pl

<table>
<thead>
<tr>
<th>Time</th>
<th>trx</th>
<th>trx</th>
<th>theorie</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>[s]</td>
<td>[ms/trx]</td>
<td>[trx/s]</td>
<td>[ms/trx]</td>
<td></td>
</tr>
<tr>
<td>12.3</td>
<td>1.2</td>
<td>833</td>
<td>8.3</td>
<td>Laptop, with good disk cache</td>
</tr>
<tr>
<td>27.2</td>
<td>2.7</td>
<td>370</td>
<td>2.5</td>
<td>SAN with up to 2000 I/O per second</td>
</tr>
<tr>
<td>41.0</td>
<td>4.1</td>
<td>244</td>
<td>4.0</td>
<td>Baseline for the following test</td>
</tr>
<tr>
<td>42.2</td>
<td>4.2</td>
<td>238</td>
<td>4.0</td>
<td>With DRBD, 2.5% slower</td>
</tr>
<tr>
<td>157.0</td>
<td>15.7</td>
<td>64</td>
<td>4.0</td>
<td>with XFS, badly configured?</td>
</tr>
<tr>
<td>86.4</td>
<td>8.6</td>
<td>116</td>
<td>8.0</td>
<td>With binary logging</td>
</tr>
<tr>
<td>565.0</td>
<td>56.5</td>
<td>18</td>
<td>8.0</td>
<td>sync_binlog=1!!!</td>
</tr>
</tbody>
</table>
InnoDB information (discussion)

• SHOW GLOBAL STATUS LIKE 'InnoDB%';
• SHOW ENGINE INNODB STATUS\G
• OK. But what does it mean to me?
• Let's start with a rough architecture picture:
InnoDB architecture

• **SHOW STATUS;**

  - Innodb_buffer_pool_%
  - Innodb_data_%
  - Innodb_dblwr_%
  - Innodb_log_%
  - Innodb_os_log_%
  - Innodb_pages_%
  - Innodb_row_lock_%
  - Innodb_rows_%

• **SHOW INNODB STATUS;**

  - SEMAPHORES
  - TRANSACTIONS
  - FILE I/O
  - INSERT BUFFER AND ADAPTIVE HASH INDEX
  - LOG
  - BUFFER POOL AND MEMORY
  - ROW OPERATIONS
### Innodb_buffer_pool%

<table>
<thead>
<tr>
<th>Expression</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Innodb_buffer_pool_pages_dirty</td>
<td>Number of dirty buffer pool pages</td>
</tr>
<tr>
<td>+ Innodb_buffer_pool_pages_clean</td>
<td>Number of clean buffer pool pages</td>
</tr>
<tr>
<td>Innodb_buffer_pool_pages_data</td>
<td>Number of data buffer pool pages</td>
</tr>
<tr>
<td>+ Innodb_buffer_pool_pages_free</td>
<td>Number of free buffer pool pages</td>
</tr>
<tr>
<td>+ Innodb_buffer_pool_pages_misc</td>
<td>Number of miscellaneous buffer pool pages</td>
</tr>
<tr>
<td>= Innodb_buffer_pool_pages_total</td>
<td>Total number of buffer pool pages</td>
</tr>
</tbody>
</table>

---

### BUFFER POOL AND MEMORY

- Total memory allocated: 18415432
- in additional pool allocated: 859008
- Dictionary memory allocated: 20888
- Buffer pool size: 512
- Free buffers: 493
- Database pages: 19
- Modified db pages: 0
- Pending reads: 0
- Pending writes: LRU 0, flush list 0, single page 0
- Pages read: 19, created 0, written 0
- Buffer pool hit rate: 754 / 1000

- 0.32 reads/s, 0.00 creates/s, 0.00 writes/s
Innodb_data%

Innodb_data_fsyncs
Innodb_data_pending_fsyncs
Innodb_data_pending_reads
Innodb_data_pending_writes
Innodb_data_read
Innodb_data_reads
Innodb_data_writes
Innodb_data_written

???
Innodb_dblwr%
Innodb_log%

Innodb_log_waits
Innodb_log_write_requests
Innodb_log_writes

Innodb_os_log_fsyncs
Innodb_os_log_pending_fsyncs
Innodb_os_log_pending_writes
Innodb_os_log_written

LOG
---
Log sequence number 0 46409
Log flushed up to 0 46409
Last checkpoint at 0 46409
0 pending log writes,
0 pending chkp writes
8 log i/o's done,
0.14 log i/o's/second

Pending normal aio reads: 0,
aio writes: 0,
ibuf aio reads: 0,
log i/o's: 0,
sync i/o's: 0
Pending flushes (fsync) log: 0;
buffer pool: 0
Innodb_pages%

- Innodb_page_size
- Innodb_pages_created
- Innodb_pages_read
- Innodb_pages_written

???
Innodb_row_lock%

\[
\text{Innodb_row_lock_current_waits} \\
\text{Innodb_row_lock_time_max} \\
\text{Innodb_row_lock_time_avg} \\
* \text{Innodb_row_lock_waits} \\
\text{------------------------} \\
= \text{Innodb_row_lock_time}
\]

???
<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Innodb_rows_deleted</td>
<td>0</td>
</tr>
<tr>
<td>Innodb_rows_inserted</td>
<td>0</td>
</tr>
<tr>
<td>Innodb_rows_read</td>
<td>0</td>
</tr>
<tr>
<td>Innodb_rows_updated</td>
<td>0</td>
</tr>
</tbody>
</table>

**ROW OPERATIONS**

- 0 queries inside InnoDB,
- 0 queries in queue
- 1 read views open inside InnoDB

Main thread process no. 7924,
id 3004103568,
state: waiting for server activity

Number of rows inserted 0,
updated 0,
deleted 0,
read 0

0.00 inserts/s,
0.00 updates/s,
0.00 deletes/s,
0.00 reads/s
SEMAPHORES

OS WAIT ARRAY INFO:
reservation count 2,
signal count 2
Mutex spin waits 0,
rounds 0,
OS waits 0
RW-shared spins 4,
OS waits 2;
RW-excl spins 1,
OS waits 0
TRANSACTIONS

----------
Trx id counter 0 1280
Purge done for trx's n:o < 0 0
undo n:o < 0 0
History list length 0
Total number of lock structs
in row lock hash table 0

LIST OF TRANSACTIONS
FOR EACH SESSION:
---TRANSACTION 0 0, not started,
process no 7924,
OS thread id 3032894352
MySQL thread id 2,
query id 4 localhost root
show engine innodb status
### FILE I/O

---

I/O thread 0 state:
waiting for i/o request
(insert buffer thread)

I/O thread 1 state:
waiting for i/o request
(log thread)

I/O thread 2 state:
waiting for i/o request
(read thread)

I/O thread 3 state:
waiting for i/o request
(write thread)

Pending normal aio reads: 0,
aio writes: 0,
ibuf aio reads: 0, log i/o's: 0,
sync i/o's: 0

Pending flushes (fsync) log: 0;
buffer pool: 0

25 OS file reads, 3 OS file writes,
3 OS fsyncs
0.42 reads/s, 100433 avg bytes/read,
0.05 writes/s, 0.05 fsyncs/s
INSERT BUFFER AND ADAPTIVE HASH INDEX

Ibuf:  size 1, free list len 0, seg size 2, 0 inserts, 0 merged recs, 0 merges
Hash table size 34679, used cells 0, node heap has 0 buffer(s) 0.00 hash searches/s, 0.31 non-hash searches/s
RAM disks (I)

- ORDER BY, GROUP BY, DISTINCT --> temp tables
  - bigger than:
    - BLOB/TEXT
  - Will be written into:
    - tmpdir = /tmp/
- Can be seen in:
  - Created_tmp_disk_tables = 0
  - Created_tmp_tables = 20

```
tmp_table_size = 32M
max_heap_table_size = 16M
```
RAM disk (II)

• Both counters are increased!

• Solutions?
  – Change your statement/requirements
  – Optimize your Query
  – Reduce size of result set
  – Avoid BLOB/TEXT

• And if you cannot?

--> Use a RAM disk!
RAM disk (III)

- RAM disk is a disk in RAM :-) --> So you need much RAM (8 Gbyte on 32-bit systems?!)
- Can use your SWAP (we do not want that)!
- More info:
  /usr/src/linux/Documentation/filesystems

```
# cat /proc/filesystems
# mount tmpfs -t tmpfs /mnt -o size=100m
# mount
```

- Bug in 5.0.4x!!! :-(
MySQL variables (discussion)

• Customers have very often misconfigured my.cnf

• My postulate: use the DEFAULT and adapt 3 things:
  – key_buffer_size
  – innodb_buffer_pool_size
  – innodb_log_file_size

• That's it! Other changes only after detailed tests!

• What is your opinion?
MyISAM log

- There is a log for MyISAM!
  I did not know that! :-(
- enable in my.cnf

```
log-isam = myisam.log
```

- cat myisam.log: :-(

```
+./mysql/host.MYI+
+./mysql/user.MYI+
+./mysql/db.MYI+
+./mysql/time_zone_leap_second.MYI+
+./mysql/time_zone_name.MYI+
+./mysql/time_zone.MYI+
 +./mysql/time_zone_transition_type.MYI
+./mysql/time_zone_transition.MYI+
+./mysql/tables_priv.MYI+
+./mysql/columns_priv.MYI+
```
MyISAM log

• myisamlog myisam.log

<table>
<thead>
<tr>
<th>Commands</th>
<th>Used count</th>
<th>Errors</th>
<th>Recover errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>open</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>write</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>update</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>close</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>extra</td>
<td>93</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>125</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

• myisamlog -? --> help
• myisamlog -vvv
• myisamlog -i

User time 0.00, System time 0.00
Maximum resident set size 0, Integral resident set size 0
Non-physical pagefaults 519, Physical pagefaults 0, Swaps 0
Blocks in 0 out 0, Messages in 0 out 0, Signals 0
Voluntary context switches 1, Involuntary context switches 8
EXPLAIN
SELECT i.number, l.answer
    FROM poll_item i
    JOIN poll_item_l l ON (l.poll_id = i.poll_id
                        AND l.number = i.number)
WHERE i.poll_id = '4'
    AND l.language_id = '2'
ORDER BY i.number ASC;
MySQL visual explain

- **http://mysqltoolkit.sourceforge.net/**

```
./mysql-visual-explain test.exp

JOIN
+- Filter with WHERE
 | +- Bookmark lookup
 |  +- Table
 |   |  table  1
 |   |  possible_keys PRIMARY
 |   +- Unique index lookup
 |     |  key 1->PRIMARY
 |     |  possible_keys PRIMARY
 |     |  key_len 5
 |     |  ref const,topodb.i.number,const
 |     |  rows 1
+- Filter with WHERE
  +- Index lookup
   |  key i->PRIMARY
   |  possible_keys PRIMARY
   |  key_len 2
   |  ref const
   |  rows 5
```