



InnoDB: Status, Architecture, and Latest Enhancements

O'Reilly MySQL Conference, April 14, 2011


Inaam Rana, Oracle
John Russell, Oracle





Bios

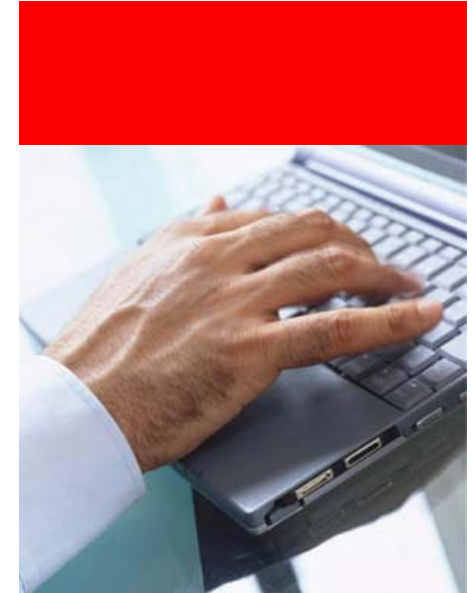
- Inaam Rana (InnoDB / MySQL / Oracle)
 - Crash recovery speedup in 5.5.
 - Native asynchronous I/O for Linux.
 - Page_cleaner thread in 5.6.
- John Russell (InnoDB / MySQL / Oracle)
 - InnoDB Plugin manual for 5.1.
 - InnoDB, Storage Engines, MEM, MEB documentation for 5.5 and 5.6.
 - InnoDB Glossary.



The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

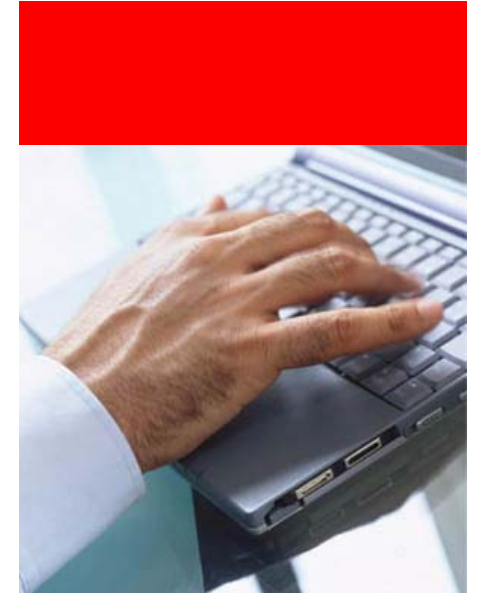
What You Will Learn

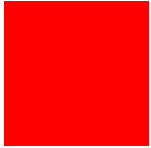
- File-per-table mode
- Barracuda file format
- Table compression
- Off-page storage for BLOBs
- Persistent optimizer statistics
- Multiple buffer pools
- NoSQL API for InnoDB
- Change buffering
- InnoDB-related I_S and P_S tables



Agenda

- Introduction to InnoDB
- InnoDB Architecture
- InnoDB Features in MySQL 5.5 (GA)
- InnoDB Features in MySQL 5.6.2 (beta)
- InnoDB + memcached (labs)
- What's in the plan?
- Q&A





Introduction to InnoDB



What is InnoDB?

- The most popular transactional storage engine for MySQL; default SE in 5.5 and up
- Architected and written by Dr. Heikki Tuuri
- Followed Gray & Reuter's "*Transactions Processing: Concepts & Techniques*"; also modeled on Oracle architecture
- Added unique subsystems/features
 - Doublewrite buffer
 - Change buffering
 - Adaptive hash index
- Many key features for performance and data integrity



InnoDB Key Characteristics



Fast

- Row-level locking - readers and writers on same table
- Multi-version concurrency control
- Efficient B-tree indexing (covering indexes)
- Adaptive hash indexing (automatic)
- Table compression with read/write capability
- Fast DDL operations (CREATE INDEX, TRUNCATE)
- Change buffering; smooth out I/O from DML
- Foreign key indexes for fast joins and RI
- New NoSQL-style interface - memcached API



InnoDB Key Characteristics



Reliable

- ACID-compliant transactions
- Two-phase commit protects data from crashes
- Fast automatic crash recovery
- Doublewrite buffer protects data from H/W errors
- Referential integrity protects cross-table data
- Online backup with MySQL Enterprise Backup
- Well written, well tested code
- Deeper integration in MySQL 5.5 and up; now the default storage engine



InnoDB Key Characteristics



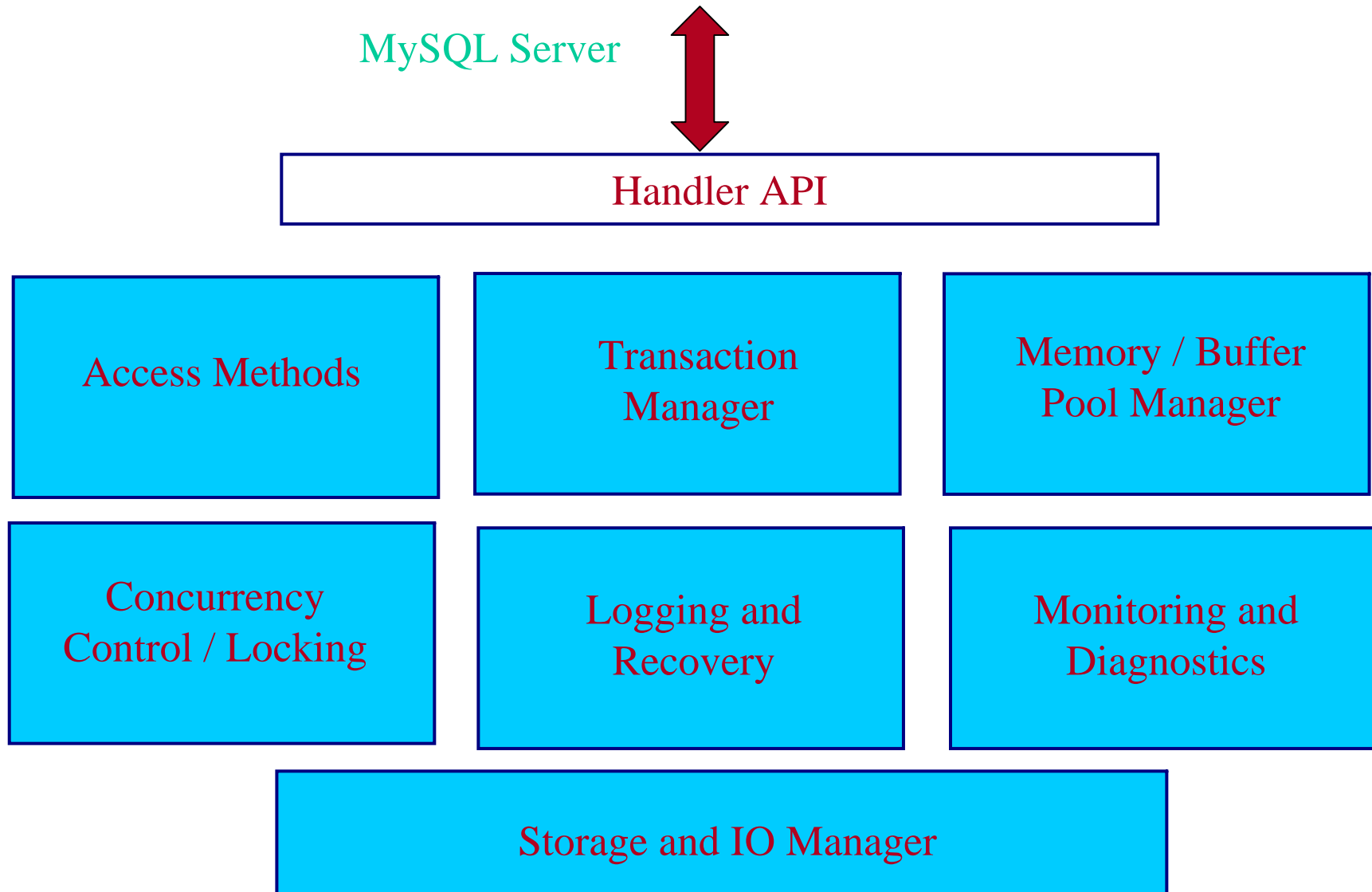
Proven

- Distributed by MySQL since 2001
- Deployed by millions of users
- Wide use in large-scale enterprise customers
- Default MySQL storage engine in 5.5 and up
- The best choice for your most important data

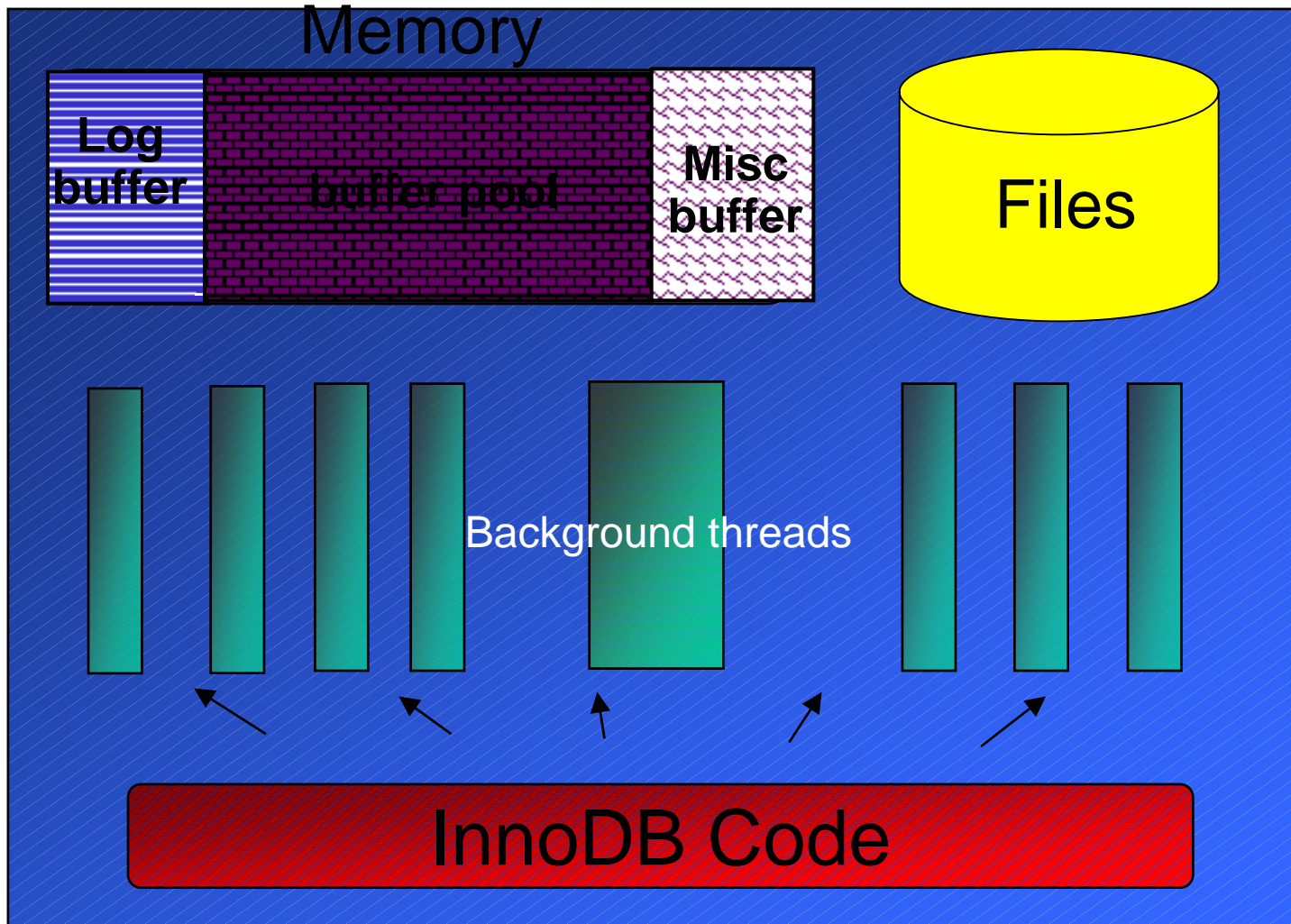


InnoDB Architecture

InnoDB Architecture: Component Model



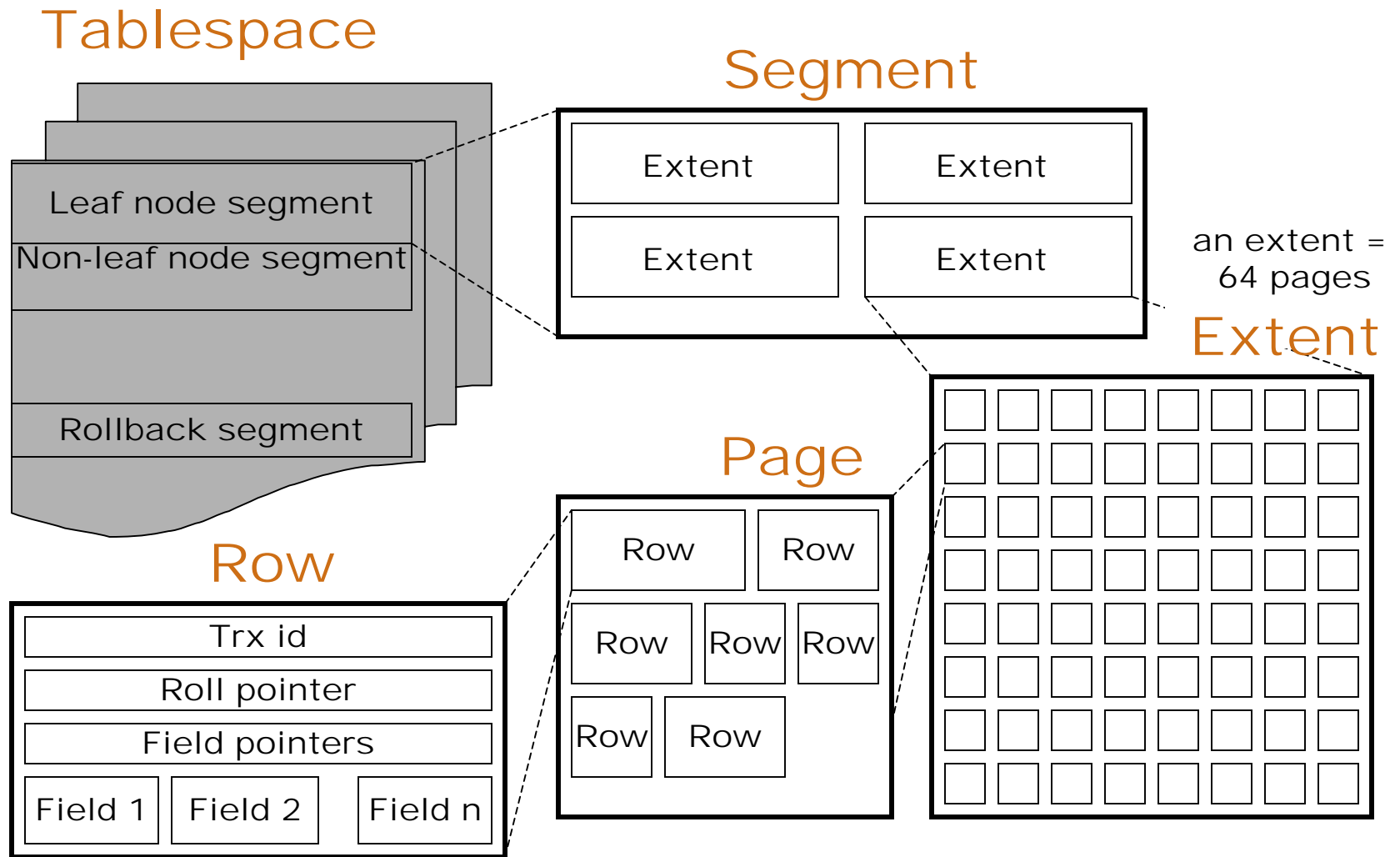
InnoDB Architecture: Runtime Model



Threads:
-master
-IO threads
-monitor
-and more

Buffer Pool:
-data
-index
-undo
-AHI

InnoDB Architecture: Data Storage





InnoDB Features in MySQL 5.5

(GA Release)



MySQL 5.5

- Performance and scalability
 - Multiple buffer pool instances
 - Multiple rollback segments
 - Fast index creation.
 - Improved purge scheduling
 - Improved log sys mutex
 - Separate flush list mutex
 - Extended change buffering with delete buffering and purge buffering
 - Native async I/O support on Linux
 - Windows performance improvements



MySQL 5.5

- Monitoring & Diagnostics
 - Information Schema tables for InnoDB
 - Performance Schema for InnoDB
 - Improved InnoDB transaction reporting
 - Log start and end of InnoDB buffer pool initialization to the error log



MySQL 5.5

- Barracuda file format and related features:
 - Barracuda tables require `innodb_file_per_table` and `innodb_file_format=barracuda`
 - New CREATE TABLE clauses `ROW_FORMAT=DYNAMIC` and `ROW_FORMAT=COMPRESSED`
 - Both new row formats do off-page storage for long columns such as BLOBs
- Compressed tables reduce I/O, at the expense of memory and CPU
 - `KEY_BLOCK_SIZE=1,2,4,8,16`
 - Smaller page size in the buffer pool, fits more rows
 - Lots of monitoring support for compression



InnoDB Features in MySQL 5.6.2

(milestone release)



Scalability: Kernel Mutex Split

- Transaction subsystem
 - Now controlled by `trx_sys_t::lock` (`rw_lock`) and `trx_t::mutex`
- MVCC views
 - Now controlled by `trx_sys_t::read_view_mutex`
- Locking subsystem
 - Now controlled by `lock_sys_t::mutex` and `lock_sys_t::wait_mutex`



Scalability: Multi-Threaded Purge

- 5.1 and earlier, performed by master thread
- Single dedicated purge thread in 5.5
- Multiple dedicated purge threads in 5.6.2
 - innodb_purge_threads [0 – 32]
 - Coordinator thread
 - Potential improvements in multi-table workload



Scalability: Configurable Data Dictionary Cache

- Table definitions are loaded and unloaded in memory based on LRU
 - table-definition-cache defines how many tables are kept open inside InnoDB; soft limit
 - Parent/child tables and system tables exempt
- Transparent to users
- Helps with workloads that have 1000s of tables



Scalability: Buffer Pool Improvements

- `page_cleaner` thread
 - Offload flushing activity from the Master thread
 - Async checkpointing happens in `page_cleaner` instead of query threads
- `rw_locks` for `page_hash`
 - Reduces contention for buffer pool mutex
- Transparent - no user action required.
- Control on-disk size of change buffer relative to buffer pool with `innodb_max_change_buffer_size`: default 25, range 0 - 50.
- Change buffer merge increases as size grows.



Optimizer: Improvements Inside InnoDB

- Support for ICP and MRR
- Persistent Optimizer Stats
 - Accurate: Better sampling algorithm
 - Stable: Same query plan (persistent on disk)
 - Stored in user-visible and user-changeable SQL tables
 - Only ANALYZE command computes new stats
 - `innodb_analyze_is_persistent`,
`innodb_stats_persistent_sample_pages`,
`innodb_stats_transient_sample_pages`



Monitoring and Diagnostics

- Information schema system tables for InnoDB
 - Analogous to InnoDB table monitor output
 - INNODB_SYS_TABLES, INNODB_SYS_TABLESTATS, INNODB_SYS_INDEXES, INNODB_SYS_COLUMNS, INNODB_SYS_FIELDS, INNODB_SYS_FOREIGN, INNODB_SYS_FOREIGN_COLS
- Information schema tables for InnoDB buffer pool
 - Buffer pool stats
 - INNODB_BUFFER_PAGE, INNODB_BUFFER_PAGE_LRU, and INNODB_BUFFER_POOL_STATS
- innodb_print_all_deadlocks: reports details in MySQL server error log
- InnoDB Information_Schema.Metrics Table



INFORMATION_SCHEMA.INNODB_METRICS Table

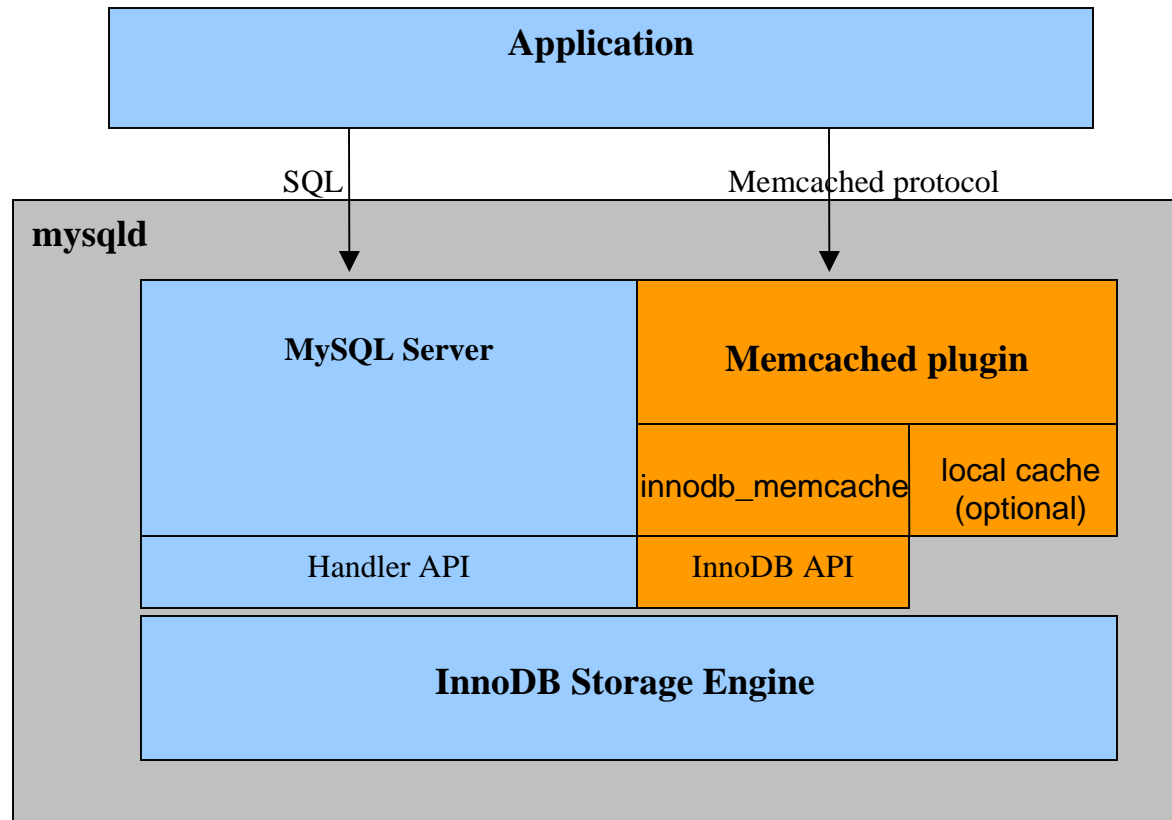
- More elaborate counters than MySQL status variables: can reset, get average, max/min...
- Each row is analogous to a status variable:
 - 176 counters
 - Many counters directly map to a status variable
- Counters grouped into modules, e.g. DML:
 - Can enable/disable/reset individual counters or entire modules
- More details: <http://blogs.innodb.com/wp/2011/04/innodb-metrics-table/>



InnoDB + memcached NoSQL to InnoDB

(labs release)

NoSQL to InnoDB with memcached





NoSQL to InnoDB with memcached

- Architecture:
 - Memcached as a daemon plugin of mysqld: both mysqld and memcached are running in the same process space, with very low latency access to data
 - Put database and caching layers on same tier
 - A single InnoDB table serves as the backing store
 - Map multiple columns to memcached item values



Memcached API - Details

- Memcached is a big in-memory hash table – get/set/modify an item based on a lookup key value
- InnoDB is a durable disk-based data store
- The API “gets” and “sets” items within a single InnoDB table, within memcached cache, or both
- InnoDB has fast PK lookup, easy to map to memcached key lookups
- Fully compatible memcached server – passes “memcapable” tests
- Listens on a port for connections from a regular memcached client
- Use readable/debuggable ASCII protocol, or fast/compact binary protocol
- You can “get” and “set” multiple columns in one request; values are concatenated with a separator
- Choose how frequently to commit
- Download from labs.mysql.com. Setup instructions at blogs.innodb.com:
<http://blogs.innodb.com/wp/2011/04/get-started-with-innodb-memcached-daemon-plugin>



What's in the Plan?



InnoDB Focus

- Performance and scalability
- Online operations
- Monitoring & diagnostics
- Features requested by users, customers



InnoDB Roadmap

- Performance and Scalability
 - Improve thread scheduling
 - Auto-extension of files in background
 - Group commit with `sync_binlog != 0`
 - Changes to LRU flushing
 - Tuning adaptive flushing
 - Fast checksum
 - Increase the max size of redo log files
 - Preload buffer pool



InnoDB Roadmap

- Online DDLs
 - Online add index
 - Online drop index
 - Online index rebuild
- Monitoring and Diagnostics
 - More metrics counters
 - Additional information_schema tables and performance statistics
 - DTrace probes for InnoDB



InnoDB Roadmap

- Optimized for SSD
- Full-text search support for InnoDB
- Lift the limit of index key prefixes (currently 767 bytes)



Resources

- Contact info: inaam.rana@oracle.com,
john.russell@oracle.com
- Blogs: <http://blogs.innodb.com/>
- Docs:
 - 5.6: <http://dev.mysql.com/doc/refman/5.6/en/innodb-storage-engine.html>
 - 5.5: <http://dev.mysql.com/doc/refman/5.5/en/innodb-storage-engine.html>
 - 5.1: <http://dev.mysql.com/doc/innodb-plugin/1.0/en/index.html>
 - Glossary:
<http://dev.mysql.com/doc/refman/5.6/en/glossary.html>



Q & A

QUESTIONS
ANSWERS