



InnoDB: Performance and Scalability Features

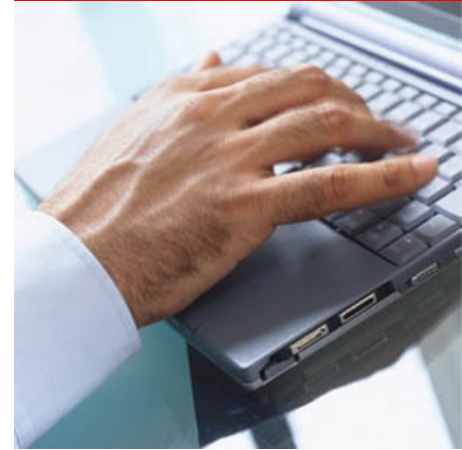
COLLABORATE 11, April 11, 2011

Calvin Sun, Sr. Manager, Oracle



Agenda

- InnoDB Performance Features
- Benchmarks
- Technical Details
- Q&A



InnoDB Performance Features

InnoDB Performance & Scalability Features

- The most popular transactional storage engine for MySQL; architected and written by Dr. Heikki Tuuri
- Followed Gray & Reuter's "*Transactions Processing: Concepts & Techniques*"; also modeled on Oracle architecture
- InnoDB performance and scalability features
 - Row-level locking and MVCC
 - Adaptive hash index
 - Read-ahead
 - Insert buffering
 - Multiple background IO threads
 - Faster locking for improved scalability
 - Using operating system memory allocators
 - Group commit support

InnoDB 1.1 for MySQL 5.5

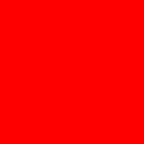
- Performance and scalability
 - Multiple buffer pool instances
 - Multiple rollback segments
 - Improved purge scheduling
 - Extended change buffering with delete buffering and purge buffering
 - Native async I/O support on Linux
 - Improved log sys mutex
 - Separate flush list mutex
 - Windows performance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

InnoDB 1.1 for MySQL 5.5

- Monitoring & Diagnostics
 - Performance schema for InnoDB
 - Improved InnoDB transaction reporting
 - Log start and end of InnoDB buffer pool initialization

InnoDB 1.2 for MySQL 5.6

- Performance and scalability
 - Split the kernel mutex
 - Multi threaded purge
 - Use rw_locks for page_hash
 - Add 'page_cleaner' thread to flush dirty pages
 - Configurable data dictionary cache
- Monitoring & diagnostics
 - InnoDB Information Schema Metrics Table
 - Information schema system tables for InnoDB
 - Information schema table for InnoDB buffer pool

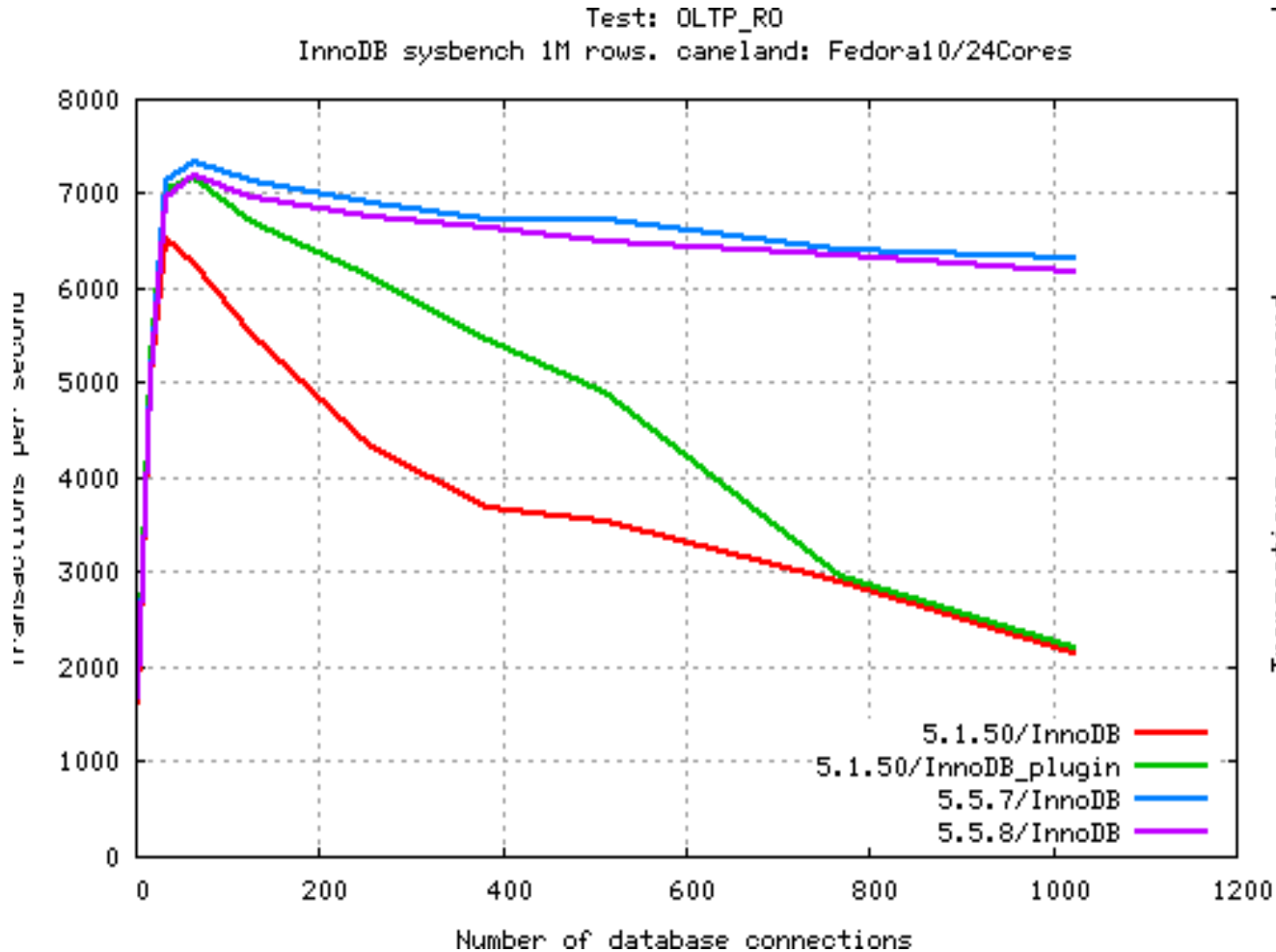


InnoDB Technology Preview on MySQL Labs

NoSQL to InnoDB with memcached

Benchmarks

MySQL 5.5 Benchmarks – Linux



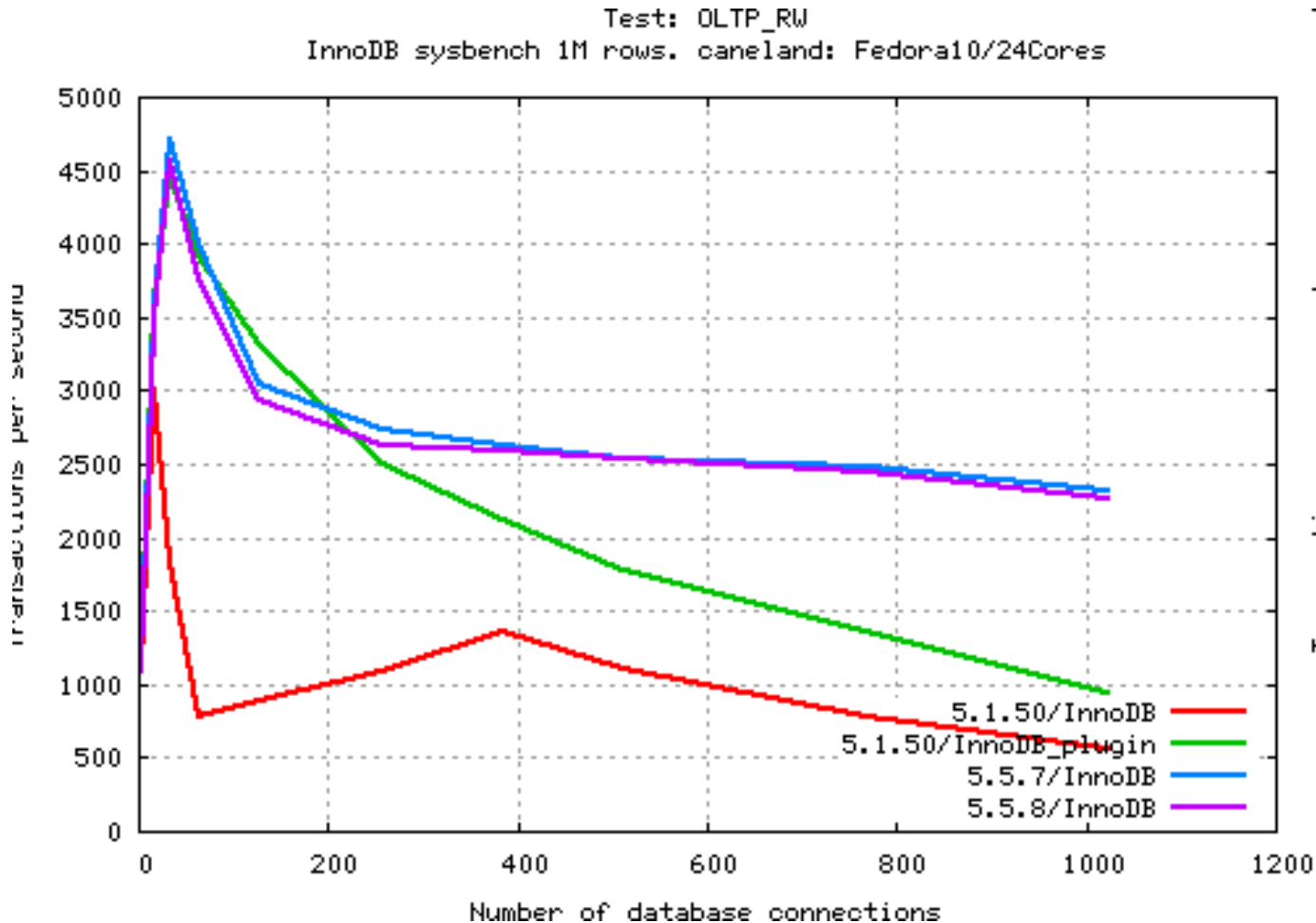
MySQL 5.5.8
(InnoDB 1.1)

MySQL 5.1.50
(InnoDB Plug-in)

MySQL 5.1.50
(InnoDB built-in)

Intel Xeon X7460 x86_64
4 CPU x 6 Cores/CPU
2.66 GHz, 32GB RAM
Fedora 10

MySQL 5.5 Benchmarks – Linux



MySQL 5.5.8
(InnoDB 1.1)

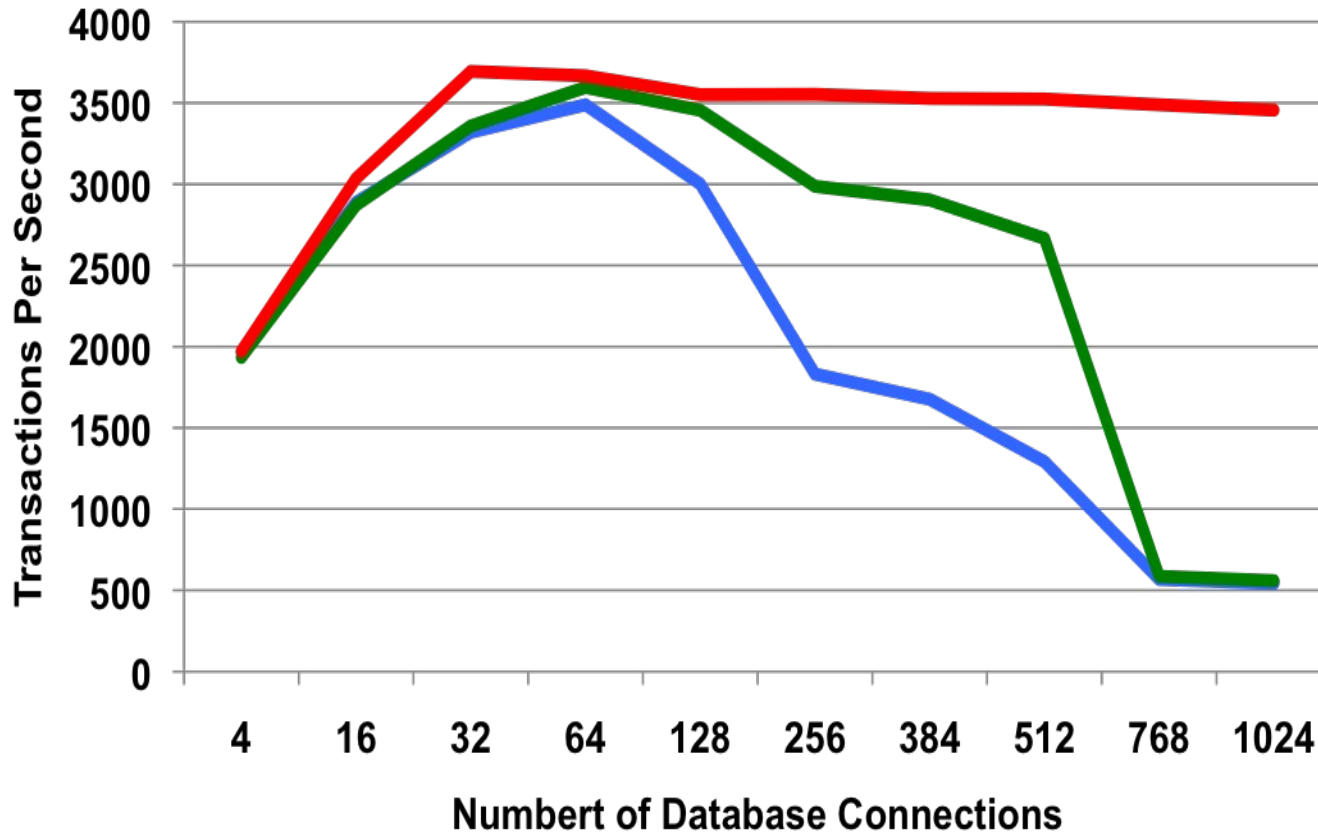
MySQL 5.1.50
(InnoDB Plug-in)

MySQL 5.1.50
(InnoDB built-in)

Intel Xeon X7460 x86_64
4 CPU x 6 Cores/CPU
2.66 GHz, 32GB RAM
Fedora 10

MySQL 5.5 Benchmarks – Windows

MySQL 5.5 vs. 5.1 - Read Only



MySQL 5.5.6

(InnoDB 1.1)

MySQL 5.1.50

(InnoDB Plug-in)

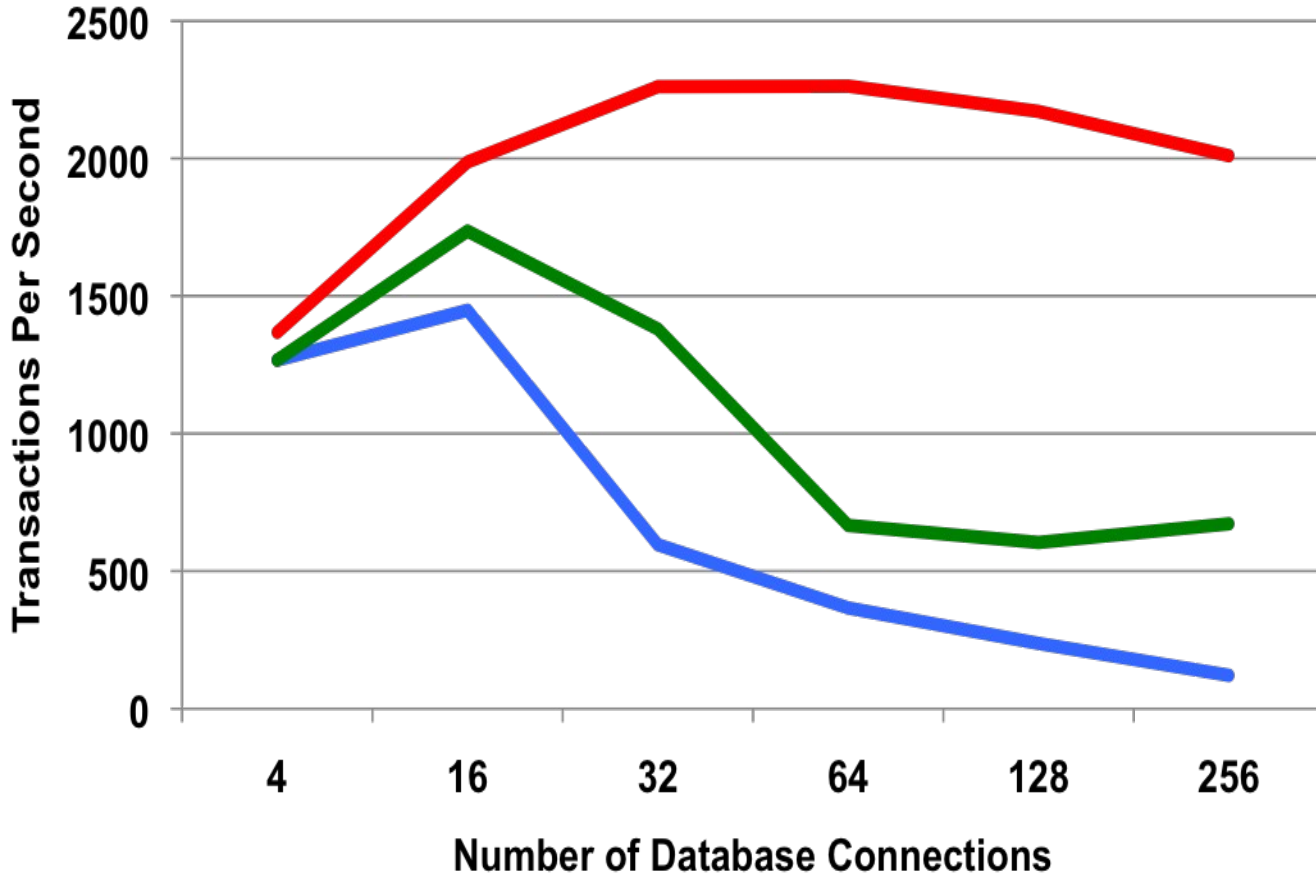
MySQL 5.1.50

(InnoDB built-in)

Intel x86_64
4 CPU x 2 Cores/CPU
3.166 GHz, 8GB RAM
Windows Server 2008

MySQL 5.5 Benchmarks – Windows

MySQL 5.5 vs. 5.1 - Read Write



MySQL 5.5.6

(InnoDB 1.1)

MySQL 5.1.50

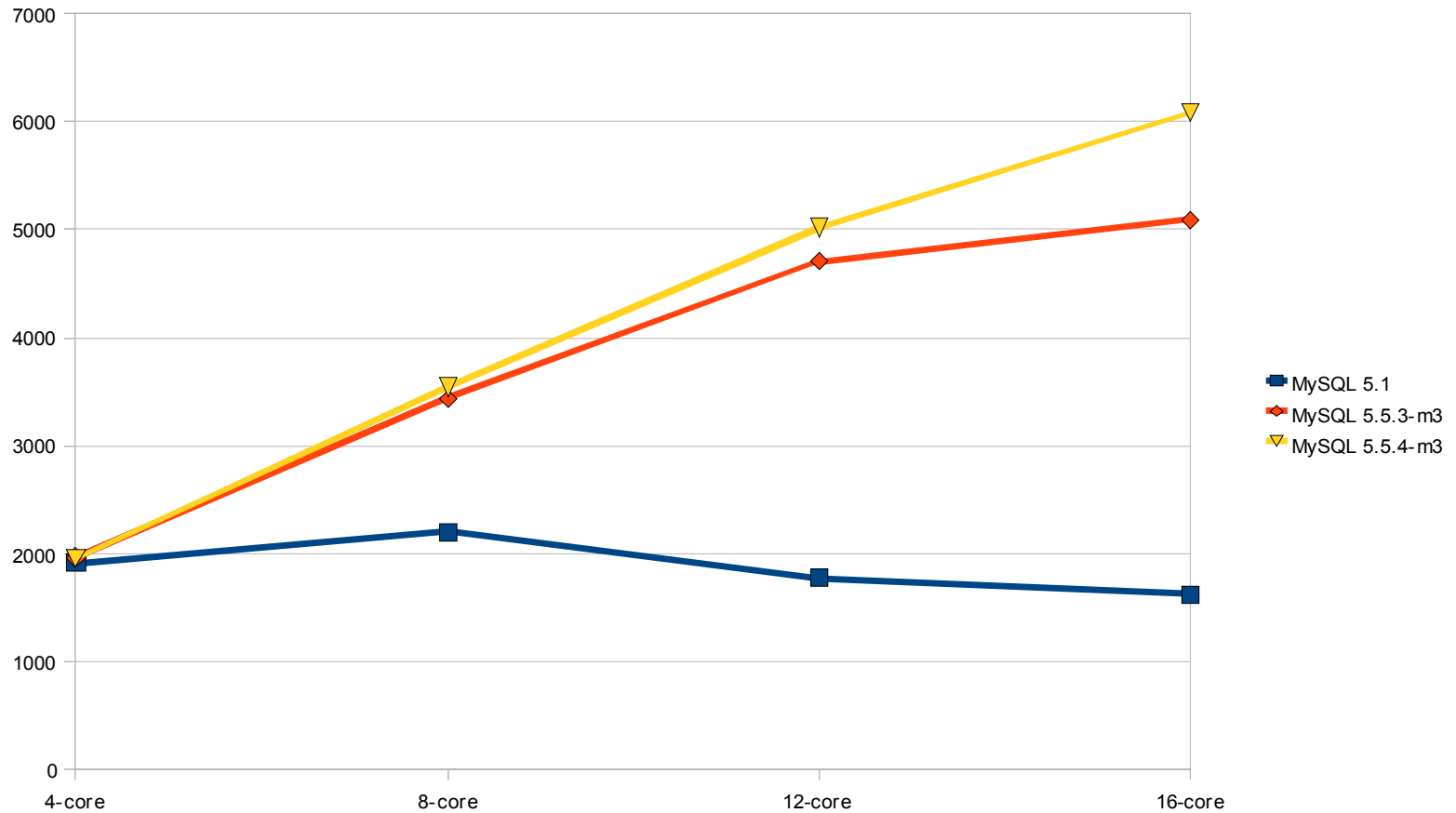
(InnoDB Plug-in)

MySQL 5.1.50

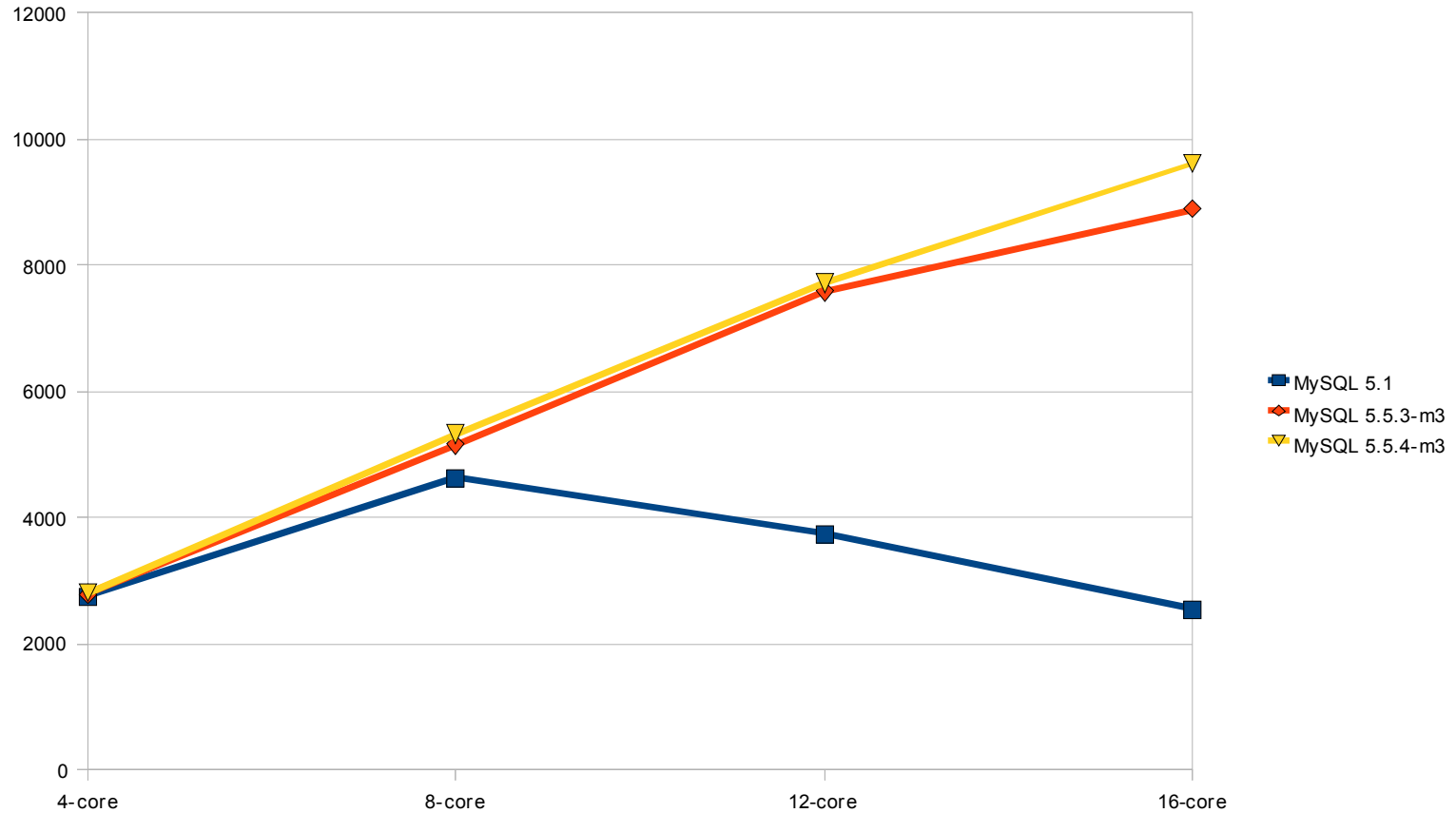
(InnoDB built-in)

Intel x86_64
4 CPU x 2 Cores/CPU
3.166 GHz, 8GB RAM
Windows Server 2008

OLTP RW Scalability (4->16 cores)

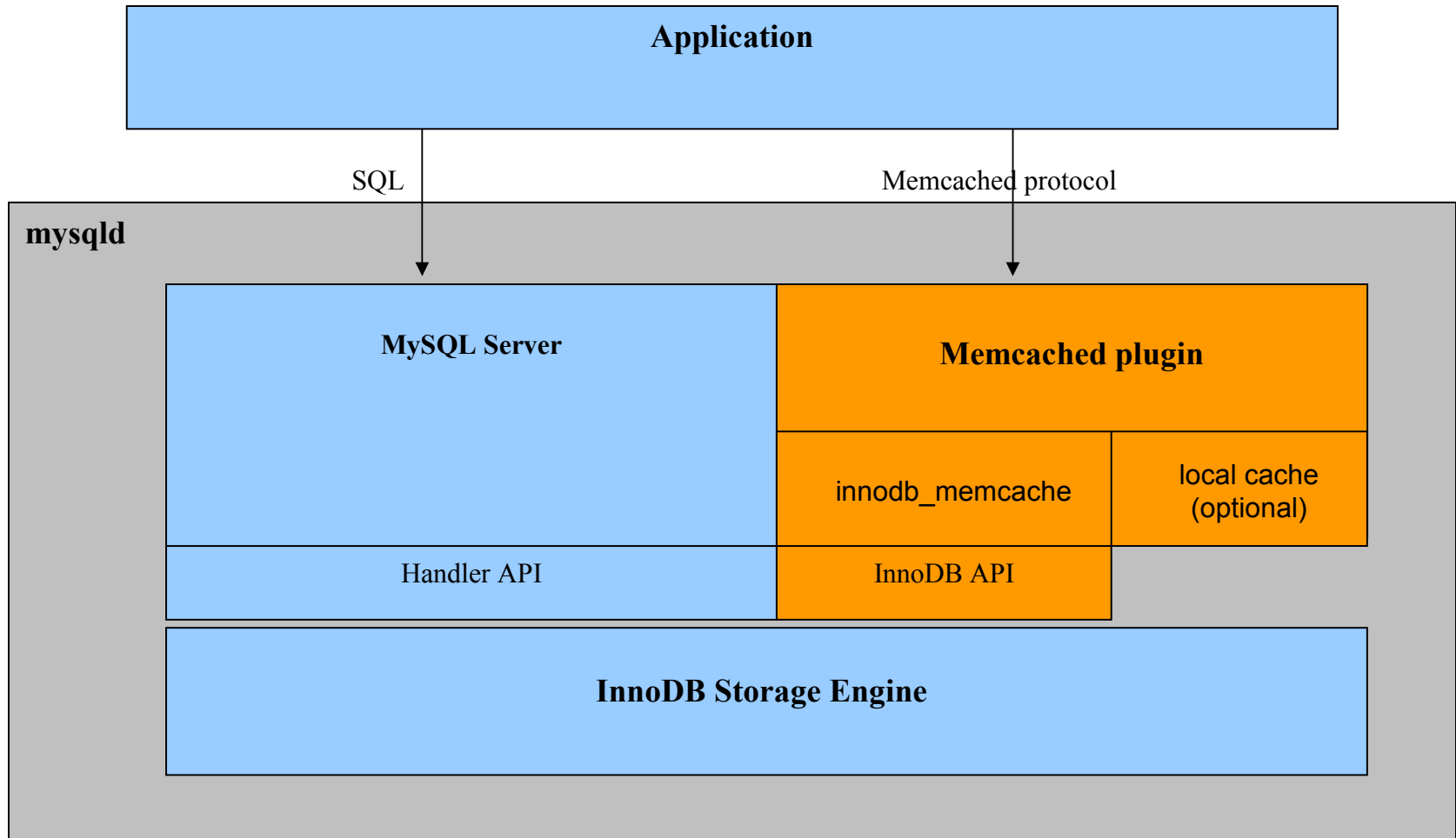


OLTP RO Scalability (4->16 cores)



Technical Details

NoSQL to InnoDB with memcached



NoSQL to InnoDB with memcached

■ Features

- 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
- Memcapable: support both memcached ascii protocol and binary protocol
- Support multiple columns: users can map multiple columns into “value”
- Optional local caching: three options – “innodb-only”, “cache-only”, and “caching”
- Batch operations: batch read/write operations and commit together, resulting in better performance

Multiple Buffer Pool Instances

- Buffer Pool mutex protects many data structures in the Buffer Pool: LRU, Flush List, Free List, Page Hash Table
- It is a hot mutex. In sysbench tests, it is acquired around 700k/sec, held about 50% of time.
- InnoDB Performance schema also confirms this:

<i>EVENT_NAME</i>	<i>COUNT_STAR</i>	<i>SUM_TIMER_WAIT</i>	<i>AVG_TIMER_WAIT</i>
<i>buf_pool_mutex</i>	1925253	264662026992	137468
<i>buffer_block_mutex</i>	720640	80696897622	111979
<i>kernel_mutex</i>	243870	44872951662	184003
<i>purge_sys_mutex</i>	162085	12238011720	75503
<i>trx_undo_mutex</i>	120000	11437183494	95309
<i>rseg_mutex</i>	102167	14382126000	140770
<i>fil_system_mutex</i>	97826	15281074710	156206

Multiple Buffer Pool Instances

- One solution is to split the buffer pool into multiple buffer pool instances
- A configurable number of buffer pool instances means we can tune the system for optimum number of instances as well
- Sysbench RW on 16-cores improves 10%
- Improves Read Only performance as well
- Large improvement on 32-cores
- Splitting out page hash is also available

Multiple Rollback Segments

- Traditionally InnoDB used only one rollback segment which has the limitation of 1023 concurrent transactions
- Accessing the rollback segment mutex is also one of the main causes of performance degradation with many connections
- This feature increases the number of rollback segments to 128; each capable of servicing 1K concurrent transactions. Each rollback segment has its own mutex.
- There is no need to create a new database to take advantage of this feature.
- The number of rollback_segments to use is now configurable via innodb_rollback_segments

Multi-Threaded Purge

■ InnoDB Purge Activity

- Purge activity comes as support for multi-versioning: requires old records to be kept in data and indexes until no transaction will access them anymore
- Purge activities include deleting delete_marked records, deleting entries from secondary indexes, deleting the UNDO record.
- Previously purge activity was performed as part of master thread
- Higher transaction rates means that quite a lot of energy needs to be spent on purging; thus master thread can be blocked to execute only purge activities for a long time
- This leads to master thread not properly flushing dirty pages, not doing checkpoints regularly as it should, leading to very high variance in throughput

Multi-Threaded Purge

- In InnoDB 1.1, a dedicated thread is created for the sole purpose of purging
- But single threaded purge could not keep up
- In InnoDB 1.2, multiple purge threads are implemented, with one purge coordinator thread plus multiple purge worker threads
- This means that master thread will properly handle flush and checkpointing activities; lead to steady performance

Improvements on Flushing

- InnoDB Flushing
 - InnoDB always flush in batches
 - Currently flushing activity happens in either the master thread or in the user threads
- In InnoDB 1.2, we introduce a new background thread, called `page_cleaner` thread, for flushing of dirty pages
 - flushing based on: `max_dirty_ratio`, `adaptive_flushing` heuristic, `async_pre-flushing` when nearing checkpoint age
 - flushing in periods of inactivity
 - flushing at shutdown

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
- Transparent to the users
- Helps with workloads that have 1000s of tables

Optimization Improvements in InnoDB

- Support for MRR/ICP
- Persistent Optimizer Stats
 - More accurate: better sampling algorithm
 - Stable: same query plan (persistent on disk)
 - Stored in user visible and user changeable tables
 - Only ANALYZE command gets new stats
 - `innodb_analyze_is_persistent`,
`innodb_stats_persistent_sample_pages`,
`innodb_stats_transient_sample_pages`

Improvements on Windows

- Implement slow mutex as CriticalSection, i.e. make it fast mutex.
- Take the advantage of condition variables on "new" Windows, which are user mode synchronization primitives and lightweight.
- Significant performance increase
- Also significantly reduces the # of kernel objects used by the InnoDB. With a 4Gb innodb_buffer_pool:
 - # of events: 1199518 --- before the fix
 - # of events: 642 --- after the fix

Monitoring and Diagnostics

- Performance Schema in InnoDB
 - 42 mutexes
 - 10 rwlocks
 - 6 types of threads
 - 3 types of I/O
- InnoDB Metrics Table
 - Counter-based monitoring system
 - Display through information schema tables
 - Lightweight
 - Can be turned on/off/reset
 - Used for both performance and resource usage
 - 16 modules and 170+ counters

Monitoring and Diagnostics

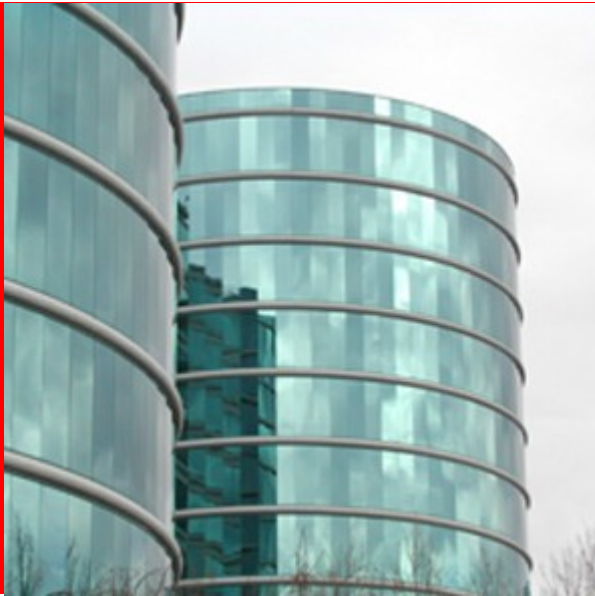
- Information Schema Tables for InnoDB buffer pool:
 - INNODB_BUFFER_PAGE: describes the types of pages in the buffer pool
 - INNODB_BUFFER_POOL_STATS: describes general buffer pool information/statistics
 - INNODB_BUFFER_LRU: describes the types of pages in the buffer pool's LRU list
- Information Schema Tables for InnoDB system tables:
 - INNODB_SYS_TABLES
 - INNODB_SYS_INDEXES
 - INNODB_SYS_COLUMNS
 - INNODB_SYS_FIELDS
 - INNODB_SYS_FOREIGN
 - INNODB_SYS_FOREIGN_COLS
 - INNODB_SYS_TABLESTATS

Resources

Transactions on InnoDB (blogs.innodb.com)

- NoSQL to InnoDB with Memcached
- Get started with InnoDB Memcached Daemon plugin
- MySQL 5.6: InnoDB scalability fix – Kernel mutex removed
- MySQL 5.6: Multi threaded purge
- MySQL 5.6: Data dictionary LRU
- Information Schema for InnoDB System Tables
- Introducing page_cleaner thread in InnoDB
- InnoDB Persistent Statistics at last
- Tips and Tricks for Faster DDL

MySQL Developer Zone (<http://dev.mysql.com/>)



Thanks for attending!

InnoDB: Status, Architecture, and Latest Enhancements

Calvin Sun, 1:00pm -- 2:00pm, April 13, 2011, 307A

Demystified MySQL/InnoDB Performance Tuning

Dimitri Kravtchuk, 10:30am – 11:30am, April 13, 2011, 305B

The logo for Oracle Open World is displayed on a red rectangular background. The word "ORACLE" is written in white, uppercase letters at the top. Below it, the word "OPEN" is written in large, bold, black, uppercase letters, with the "O" and "N" overlapping the "ORACLE" text. At the bottom, the word "WORLD" is written in white, uppercase letters.

ORACLE
OPEN
WORLD

San Francisco 2011

October 2–6

Register and Exhibit Now! oracle.com/openworld



Q & A
QUESTIONS
ANSWERS