PrimeBase

PrimeBase XT

A transactional engine for MySQL

Paul McCullagh
SNAP Innovation GmbH

http://www.primebase.com/xt

Our Company

- SNAP Innovation GmbH was founded in 1996, currently 25 employees.
- Purpose: develop and support PrimeBase database, and related products.
- Originally developed for the P.INK Press publishing system.
- PrimeBase today is primarily a publishing database (BLOBS, full-text index).
- Main activity: support and integration of K4 based on PrimeBase, internet publishing.

Overview

- PrimeBase XT (PBXT) is an MVCC, transactional engine for MySQL.
- An alternative to InnoDB or BDB, or even MyISAM.
- Uniquely designed for high-concurrency, heavy update load.
- Was designed and created for MySQL, not retrofitted.
- PBXT is open source and has been released under Gnu Public License (GPL).

Features

- Pure MVCC: no locking, each transaction gets a snapshot - update conflicts possible.
- File-per-table based - advantages for disk space management - databases independent.
- “Write-once”, reduces work for update significantly.
- Thread conflict is minimized by segmented cache, separate logs.

Motivation

- Customer need: current PrimeBase technology is limited: 2GB tables (without BLOBs), scalability.
- Long term strategy: open source make small companies competitive.
- We believe there is a need for alternatives: we want to give our customers a choice.
- We can bring better, large-data & BLOB handling technology to MySQL.

History

- **Jan. 2005:** Design, planning and work on a “proof-of-concept” implementation.
- **Nov. 2005:** Implementation of the engine for MySQL 4.1.16.
- **29 Mar. 2006:** First source code release.
- **4 Jul. 2006:** Release 0.9.5, a restructured implementation (improved speed of fixed records).
- **Nov. 2006:** Planned Beta release with 5.1 pluggable engine API support.

Future Plans

- Release is planned to coincide with the release of MySQL 5.1 (before the MySQL User Conference).
- Referential integrity.
- Implement CHECK TABLE, improve REPAIR and OPTIMIZE TABLE.
- Implementation of functions required by the MySQL optimizer.
- **Version 2+:** PrimeBase-like BLOB handling, full-text indexing.

Design

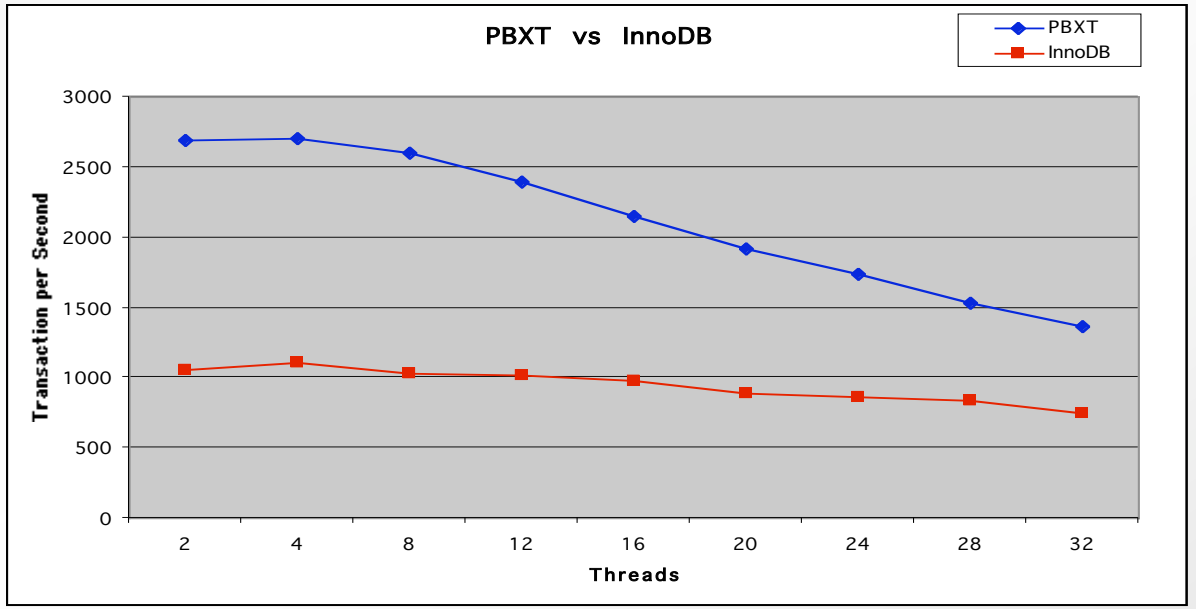
- Data is written sequentially to the logs (records are never updated) - write through.
- Handles are used to locate the data in the logs files.
- Free space is freed by a “garbage collector” thread.
- Commit/rollback/recovery is instantaneous, a “sweeper” thread cleans up after a transaction, removing deleted/rolled-back/old records.

Implementing a Storage Engine

- 4.1 required hacking MySQL, so 5.1 pluggable API is a huge improvement.
- There are still some important aspects missing from the documentation (e.g. building, debugging and running tests).
- Index support is tricky because of the many flags and modes:PREFIX, FIRST, LAST, BEFORE, AFTER, EXACT, etc.
- Transaction support requires code reading, e.g. undocumented: trans_register_ha().

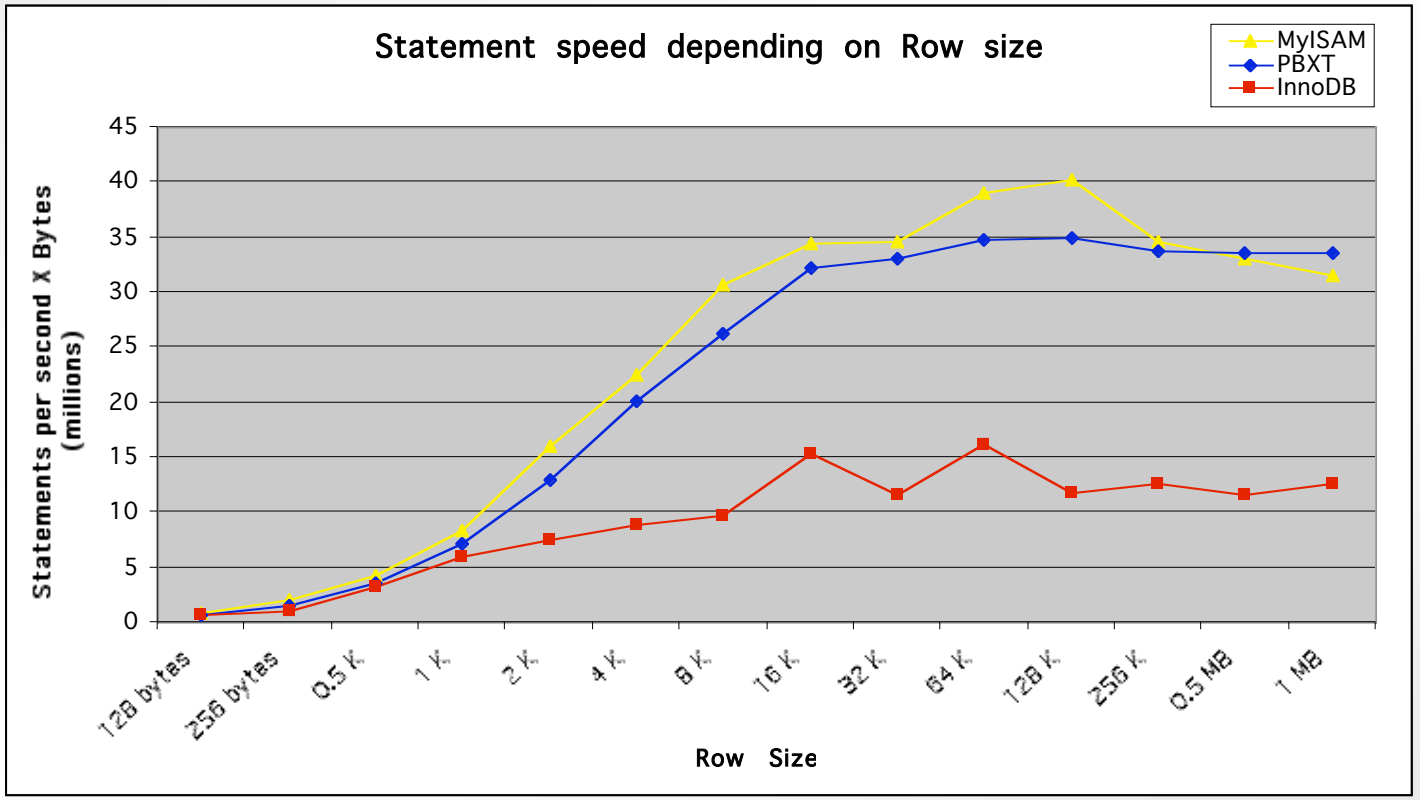
Performance

- SELECT speed is comparable to InnoDB.
- Transactions with INSERT, UPDATE and DELETE over 2 X faster than InnoDB.
  - 2 X Dual core, AMD Opteron, 2.2 GHz, sysbench 0.4.7



Throughput

- This test compares throughput measured in bytes per second.



Questions

- Who is able to test XT or varify our results?
- Are you interested in joining our Beta testing program?
- Do you know of a project where the XT may solve current problems?