

MySQL Performance Landscape



Peter Zaitsev,
MySQL AB

MySQL Users Conference 2006
Santa Clara, CA April 24-27

Introduction

- Very wide choice of platforms
 - Hardware
 - Platforms
 - Distribution Vendors (Linux)
- A lot of possible configuration
 - RAID Levels, Block size
 - FileSystems
 - OS Kernel and Libraries settings
- We will
 - Quantify performance with different configurations
 - Guess why such performance may be observed

About Benchmark

- Using SysBench
 - MultiUser benchmark developed by MySQL High Performance Group
 - Initial Goal – Investigating platform specific performance bugs for MySQL Customers
 - Simple – easy to analyze, but powerful
 - OpenSource <http://sourceforge.net/projects/sysbench>
- Results may vary
 - Do not blindly expect it to be same for your application
 - Do you own benchmarks if you need to be sure

Benchmark Configuration

- Two sizes of Workload
 - 200MB (1.000.000 rows) – CPU bound
 - 1, 16 and 256 clients
 - 20GB (100.000.000 rows) - IO Bound
 - 1, 16 concurrent clients
- Multiple Workloads
 - «Simple» - single row reads
 - «Read Write» - multiple read/write statements
 - «Read Only» - multiple statements, only reads
- Result processing
 - Run 3 times after warmup
 - Best results scores
 - «Peak Performance»

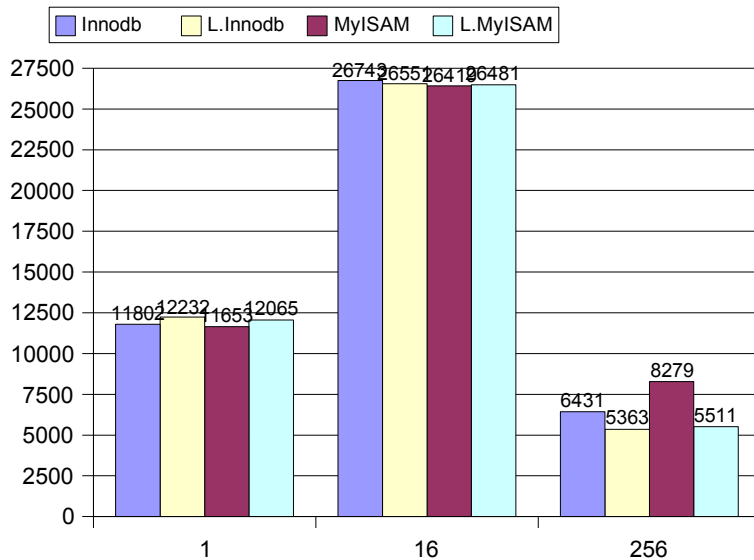
MySQL Configuration

- MySQL 5.0.18 and 5.0.19
- MyISAM and Innodb storage engines
- Optimally configured for each platform
 - No sense to run with «defaults» as results irrelevant

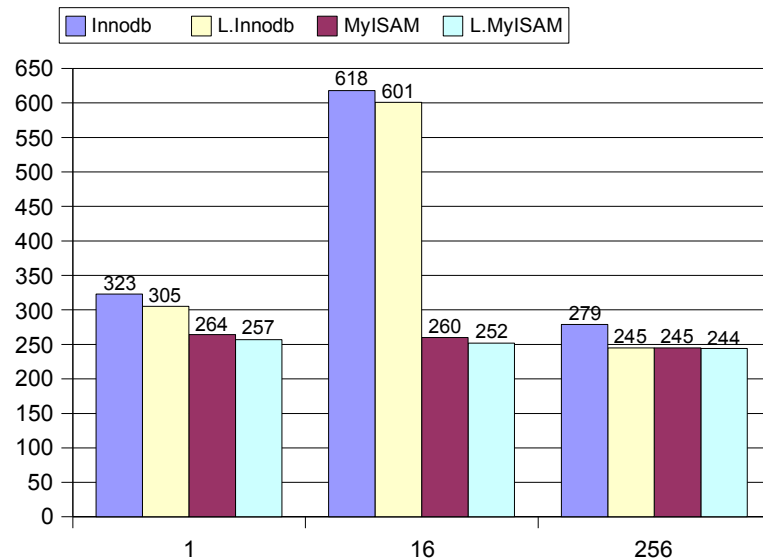
LinuxThreads and NPTL

- CPU Bound workload
- PowerEdge 1425SC, 2*3.0Ghz Xeons
- CentOS 4.2 64bit

Simple



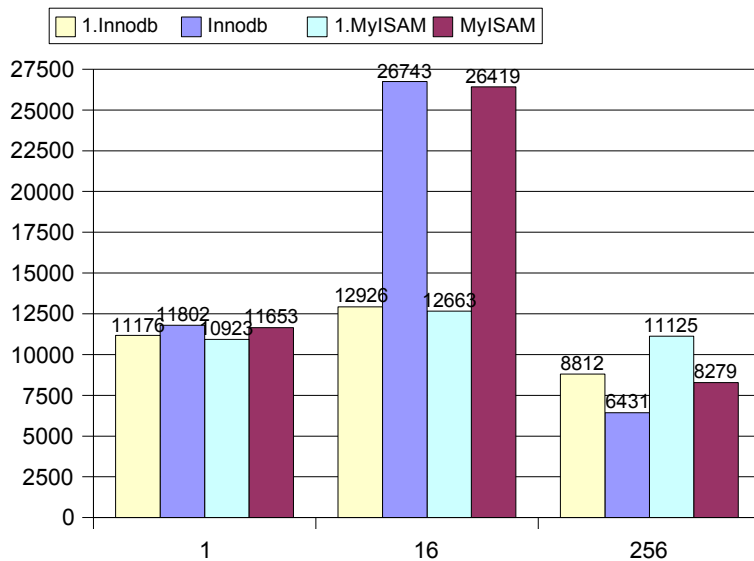
Complex RW



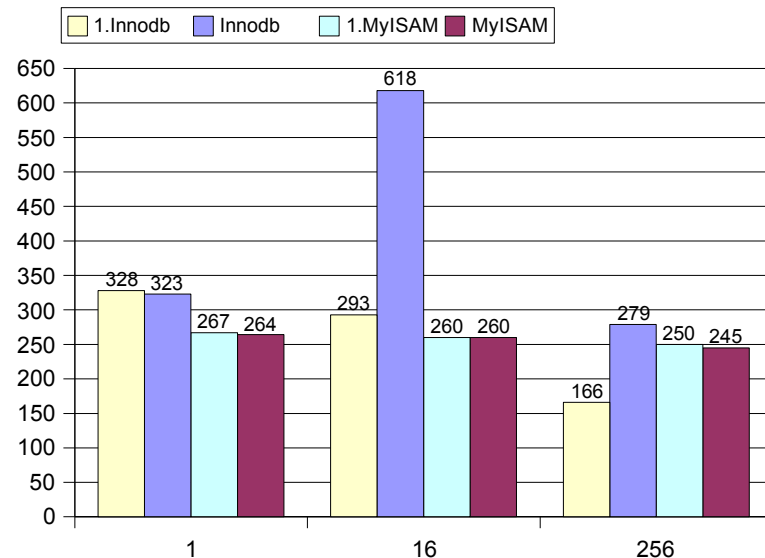
Benefit of Second CPU

- Running Kernel designed for Single CPU
 - Some single CPU optimizations apply
 - Do not use HypperThreading for that CPU

Simple



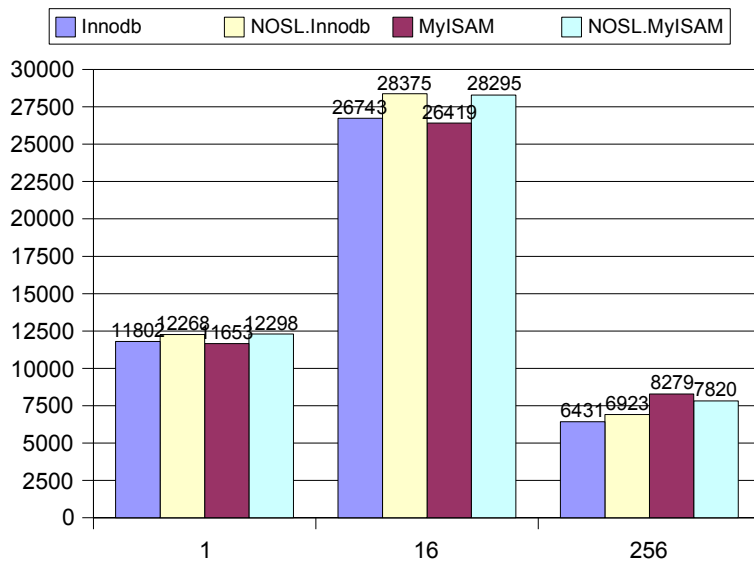
Complex RW



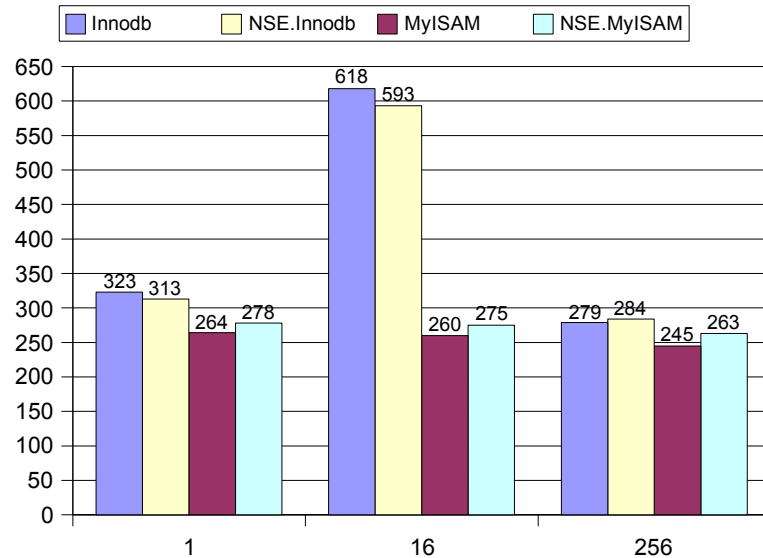
Overhead of SELinux

- SELinux – Security Extensions for Linux
 - Enabled by default on most recent distributions
 - Might not be needed on dedicated MySQL server

Simple



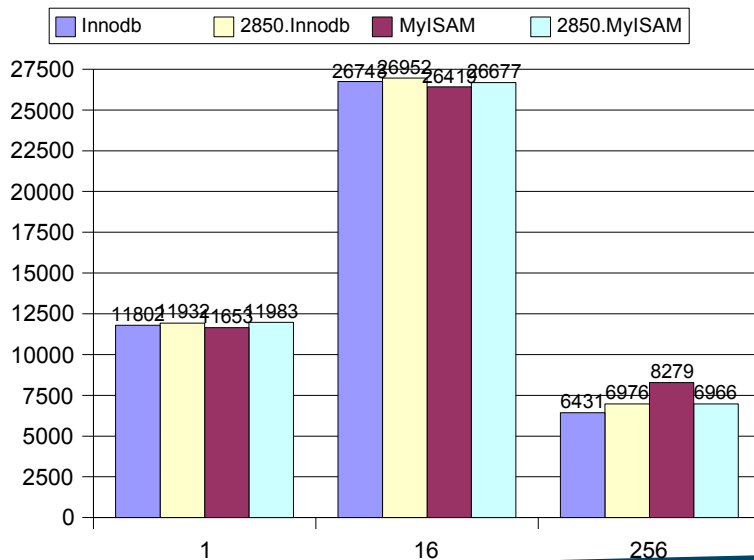
Complex RW



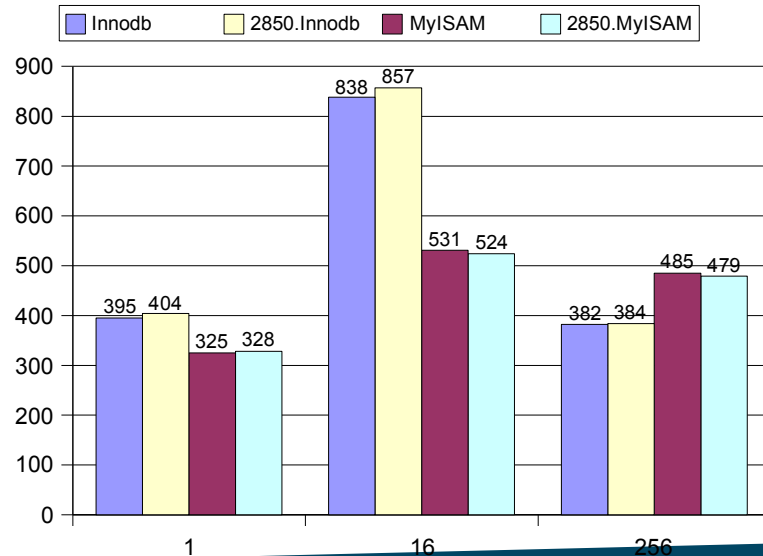
Chipset Differences ?

- Compare PowerEdge 2850 to PowerEdge 1425SC
 - Considered to be system with Higher Performance
 - Use Same CPUs for both of them
 - CPU bound workload – PE 2850 has much better IO

Simple



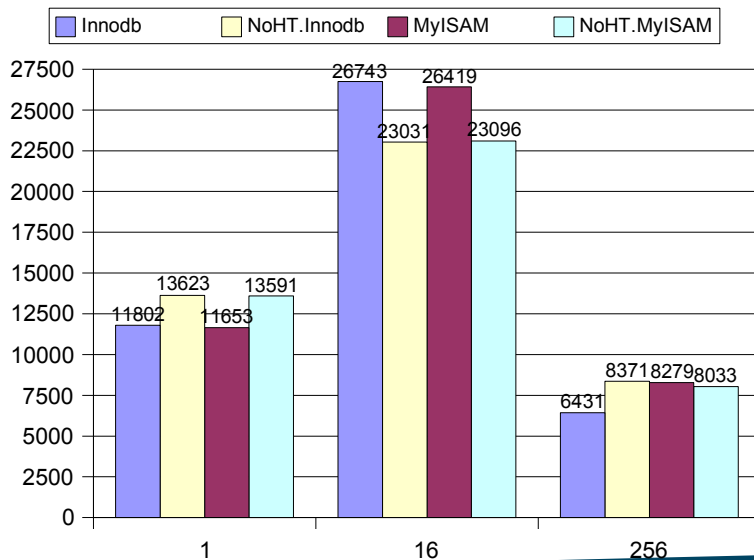
Complex RO



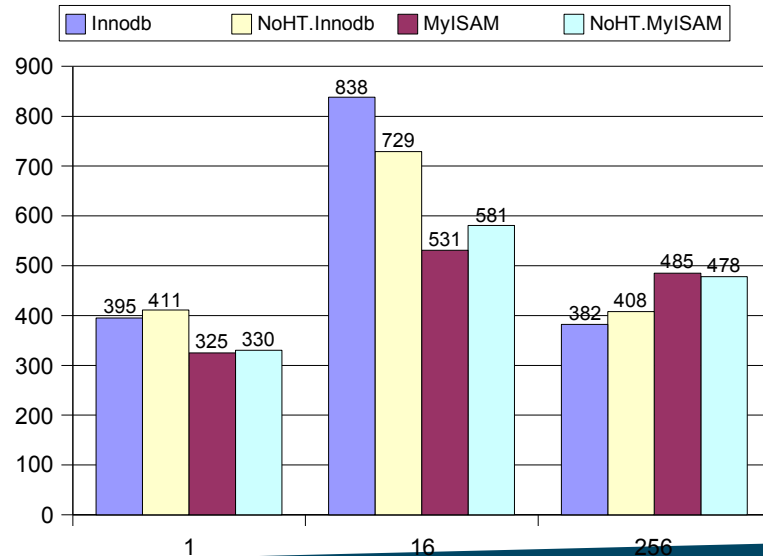
Disable HypperThreading

- HypperThreading allows running multiple threads on the same CPU
 - Using the parts of hardware which are unused
 - Shown as Logical CPU while almost no hardware duplication

Simple



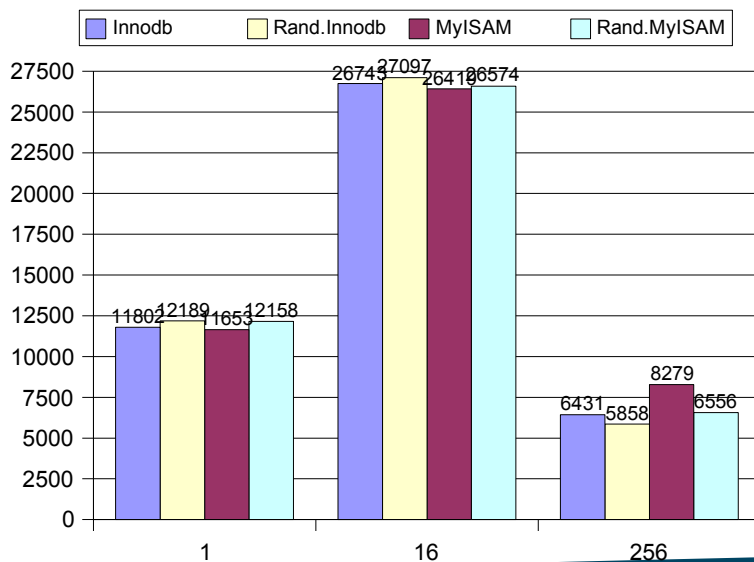
Complex RO



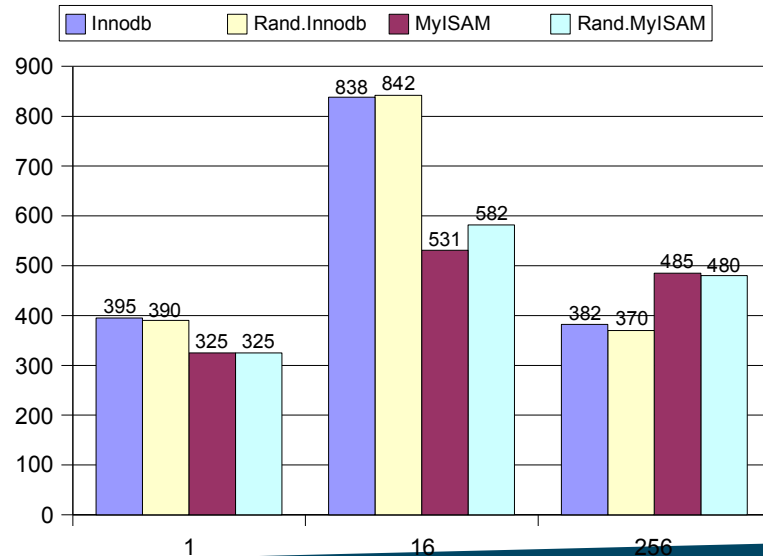
«Random Access» Optimization

- Setting in BIOS to optimize memory for random access
 - Guess different prefetch/caching configuration

Simple



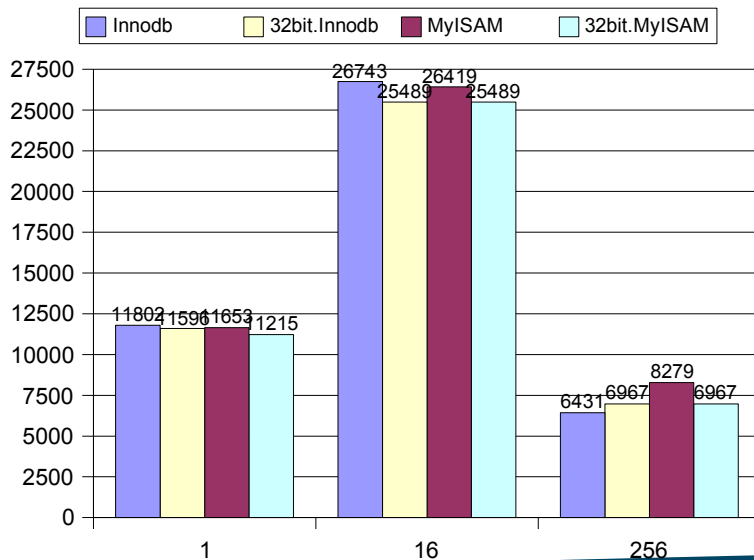
Complex RO



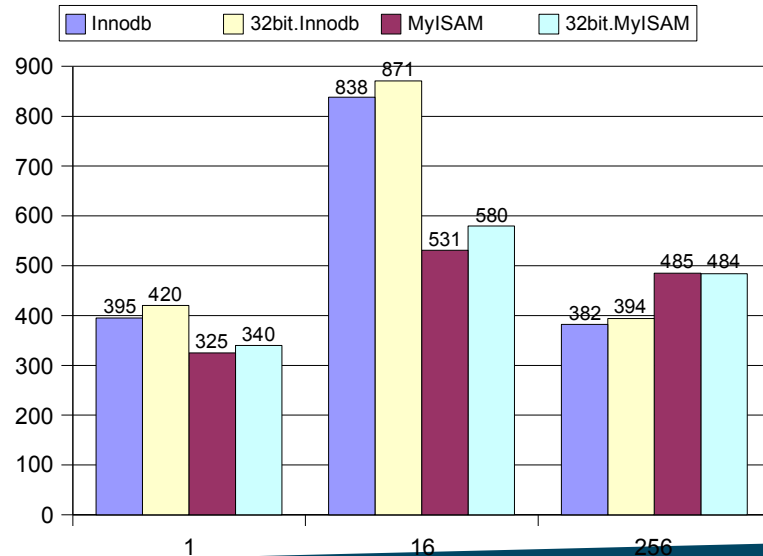
32bit Binary

- x86-64 is good in running 32bit code
- Is performance of 64bit and 32bit binaries any different
 - MySQL official binaries. Newest compilers may expose more differences

Simple



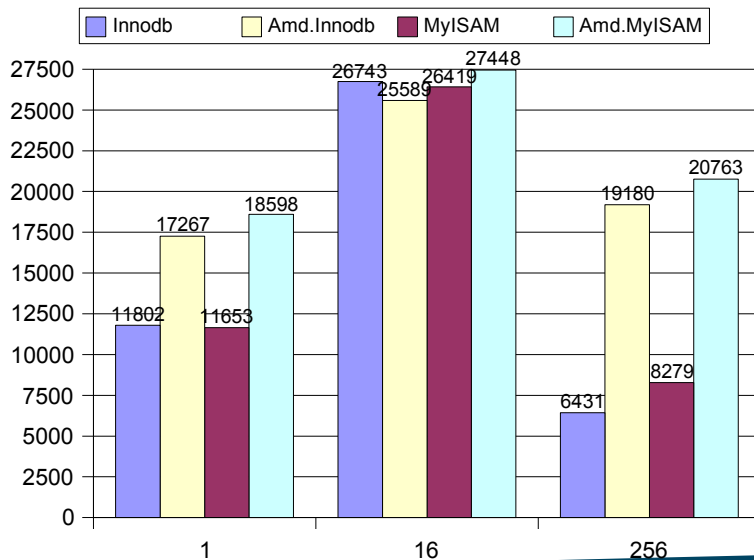
Complex RO



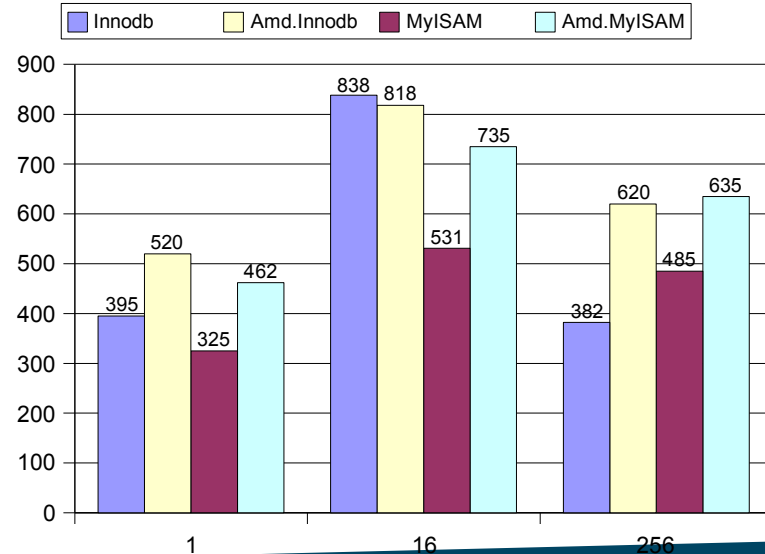
Opteron vs Xeon

- 2*3.0Ghz 2MB Cache Xeon, PE1425SC, CentOS 4.2
- AMD Athlon(tm) 64 X2 Dual Core Processor 3800+, 2Ghz
 - Typically desktop model

Simple



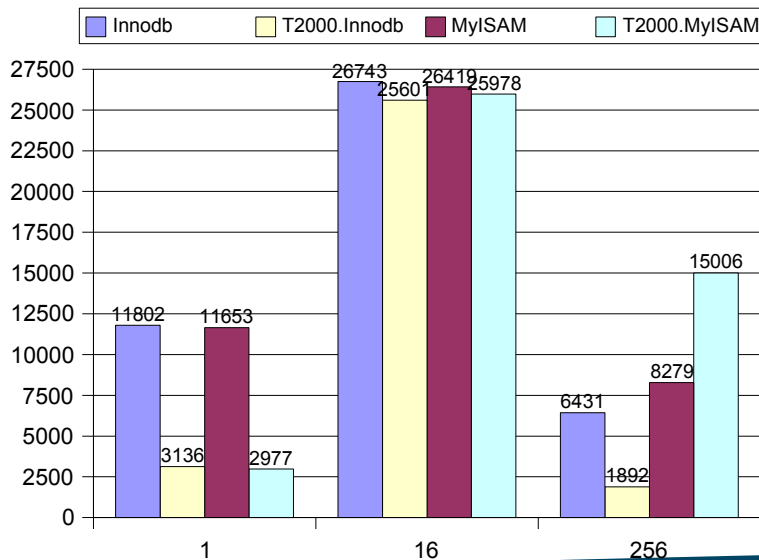
Complex RO



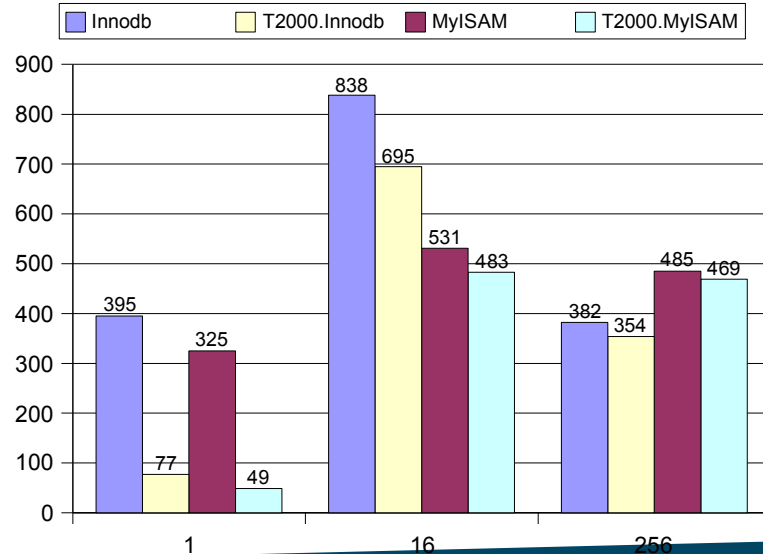
SunFire T2000 vs Xeon

- «Small» version
 - 4 Cores * 4 Threads = 16 Threads
- 2*Xeon 3.0Ghz 2MB Cache (Single Core)

Simple



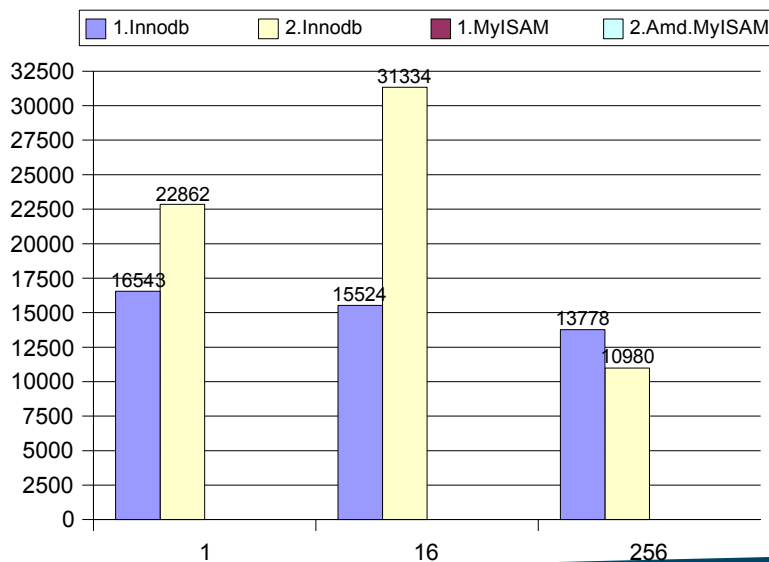
Complex RO



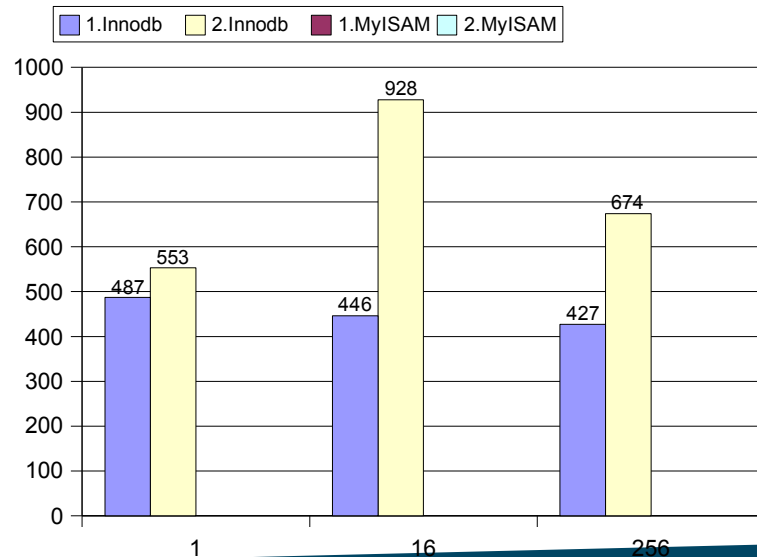
AMD Dual Core

- How much performance does enabling second core give ?
 - Usually results are less than that because Dual Cores come at lesser frequency
- No MyISAM data for single core

Simple



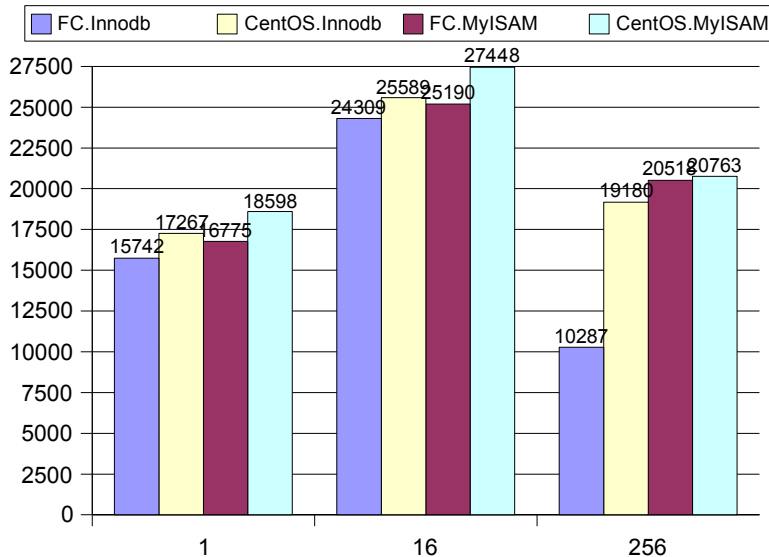
Complex RO



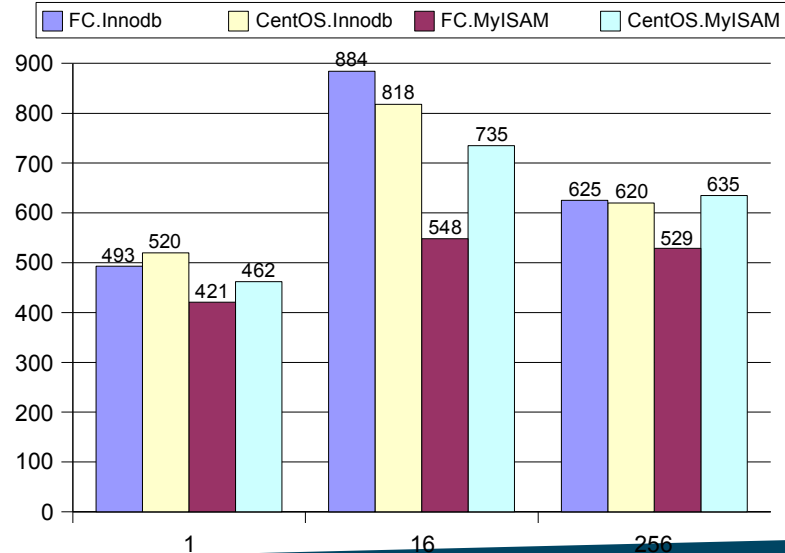
Fedora Core 5 vs CentOS 4.3

- AMD Athlon 64bit
- Newer is not always faster

Simple



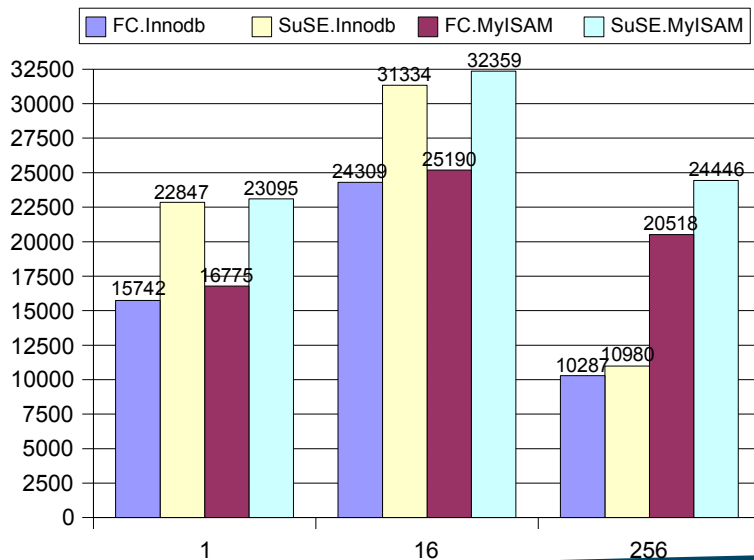
Complex RO



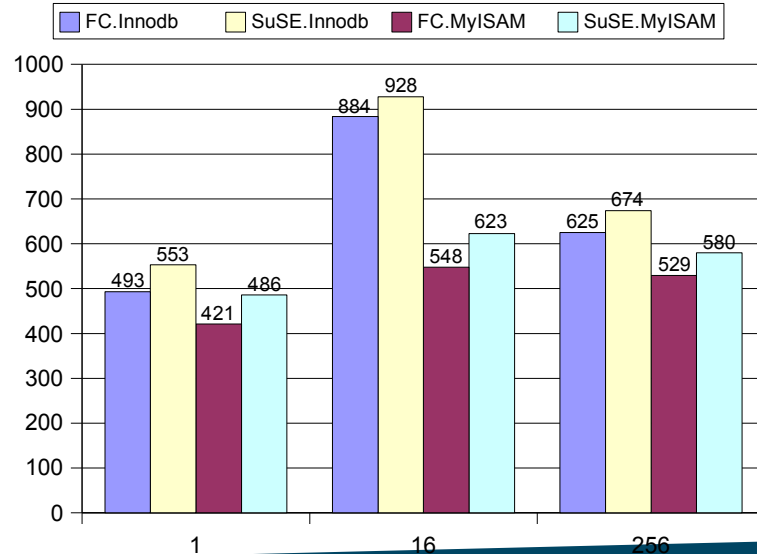
Fedora Core 5 vs SuSE 10

- Comparison on same AMD Athlon 64
- Very surprising to see Fedora Core to be so much slower
 - Or is it SuSE 10 optimized for AMD ?

Simple



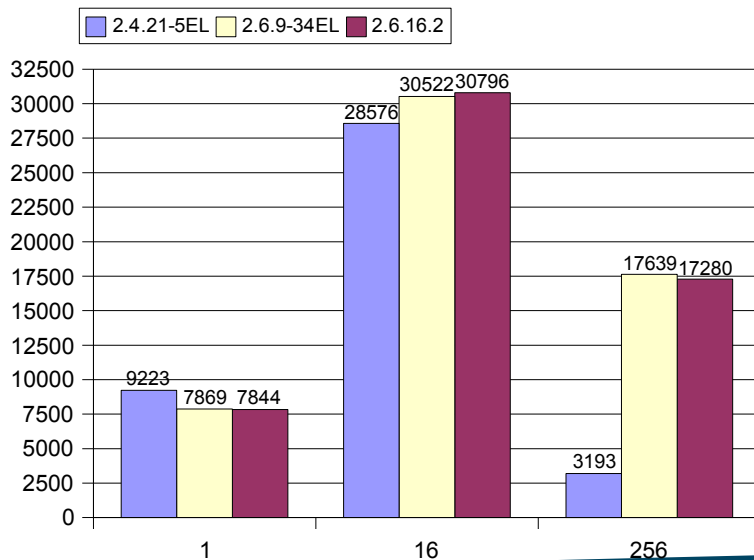
Complex RO



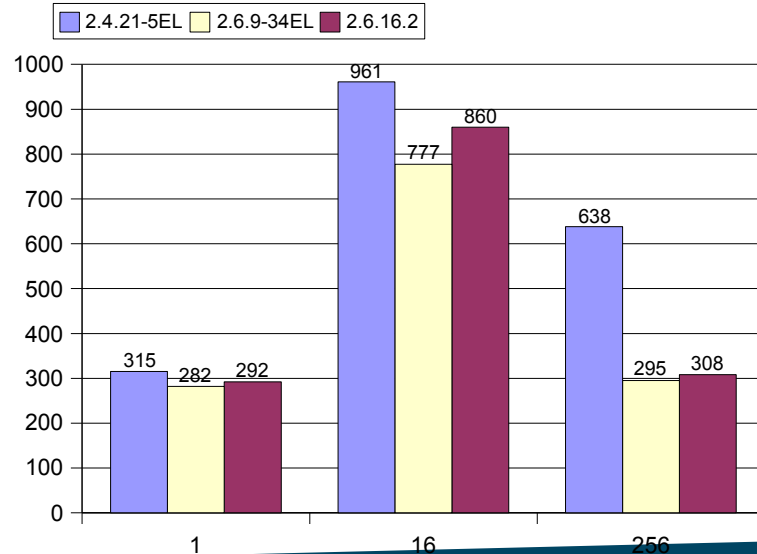
Linux Kernels 2.4 vs 2.6

- Benchmarks run on 4*2.0Ghz Xeon 32bit, Innodb
 - RedHat AS 3.0 Update 2 (2.4.21-5EL)
 - Kernel from RH AS 4.0 (2.6.9-34EL)
 - Latest «Vanilla» Kernel 2.6.16.2

Simple



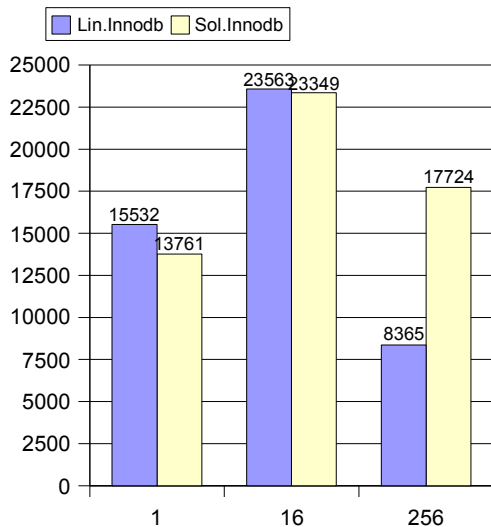
Complex RO



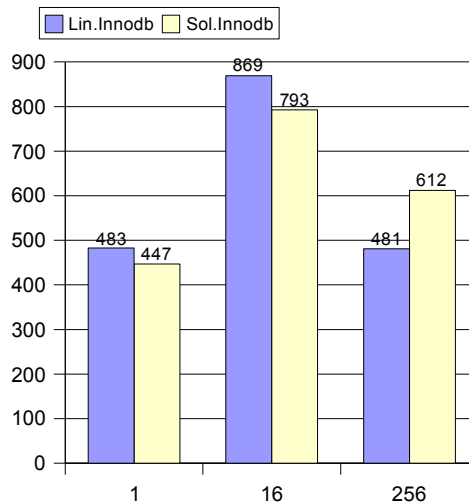
Linux vs Solaris 10

- Dual Opteron 244 (1.8Ghz)
- RedHat AS 4 update 2, Solaris 10 x86-64
- Solaris performance improvement effort is going

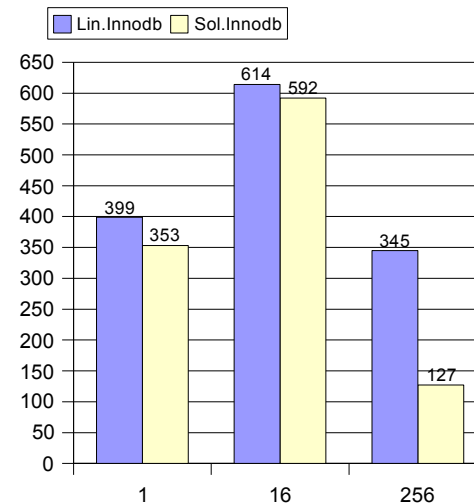
Simple



Complex RO



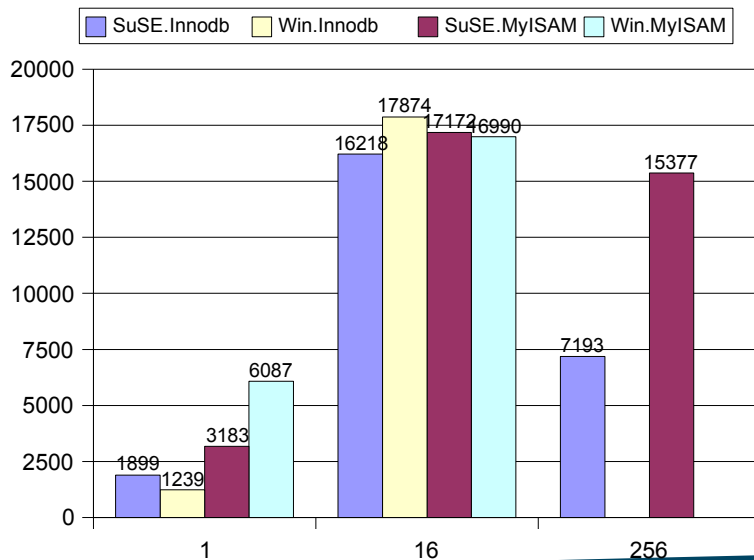
Complex RW



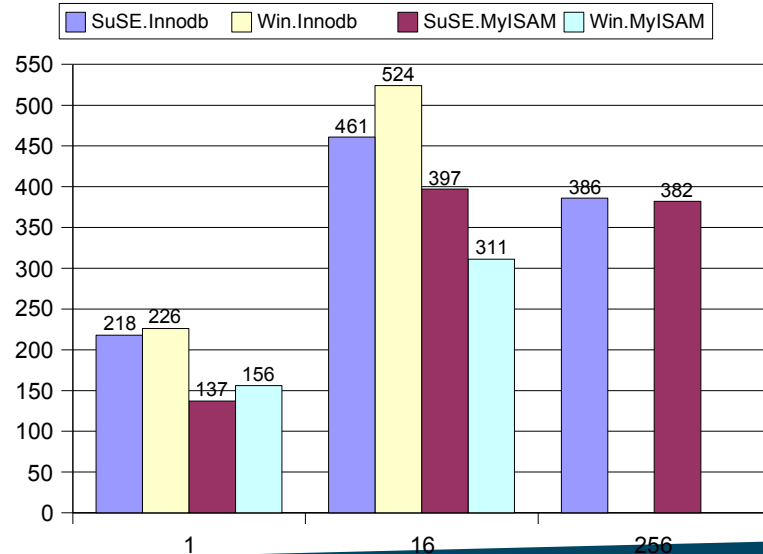
SuSE 9.3 vs Windows

- Dual P4 3.4Ghz, 1GB RAM
- Benchmark Run over 1Gb network
 - SysBench can't be run on Windows
- Results are very supecious. Something is wrong, likely.

Simple



Complex RO



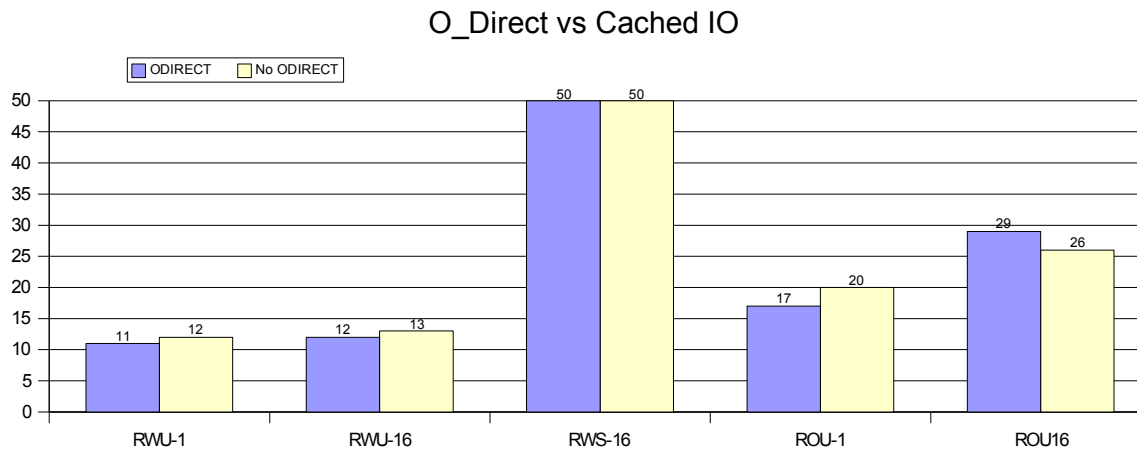
IO Bound Benchmarks

IO Bound Benchmarks

- 100.000.000 rows (about 20GB)
- Only InnoDB Tables
 - Did not have time for MyISAM
- Results are not always stable
 - Fragmentation, background activity etc
- Uniform distribution
 - Even data access. Most hard for disk, bad for caches
- Special Distribution
 - Very Skewed
 - Much better cache rate
 - Close to applications with Mixed CPU/DiskIO load

O_DIRECT PowerEdge 1425SC

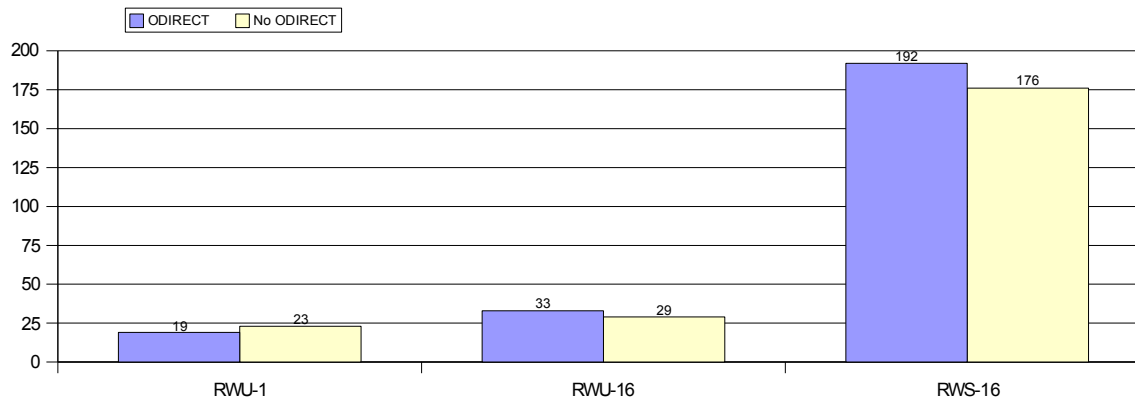
- Dell PowerEdge 1425SC – 4GB RAM
 - 2SATA 7200 Drives in software RAID1
 - ReiserFS
- O_DIRECT is expected to benefit on high end systems
 - This is not the one



O_DIRECT PowerEdge 2850

- Dual Xeon 3.0 Ghz, 8GB RAM, 6*10.000RPM SCSI
 - 4 Drives in RAID5
 - EXT3
- O_DIRECT is better if performance is same
 - Saves your OS cache for other applications

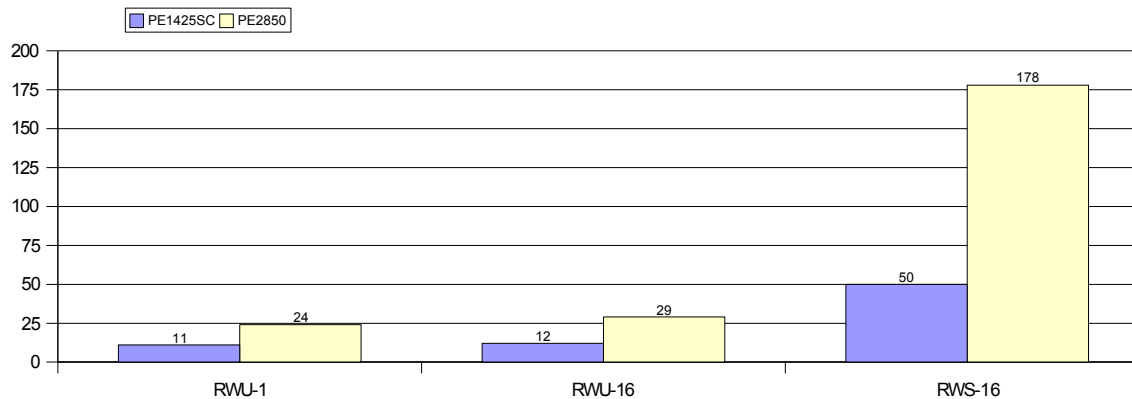
O_Direct vs Cached IO



PE 2850 vs PE 1425SC

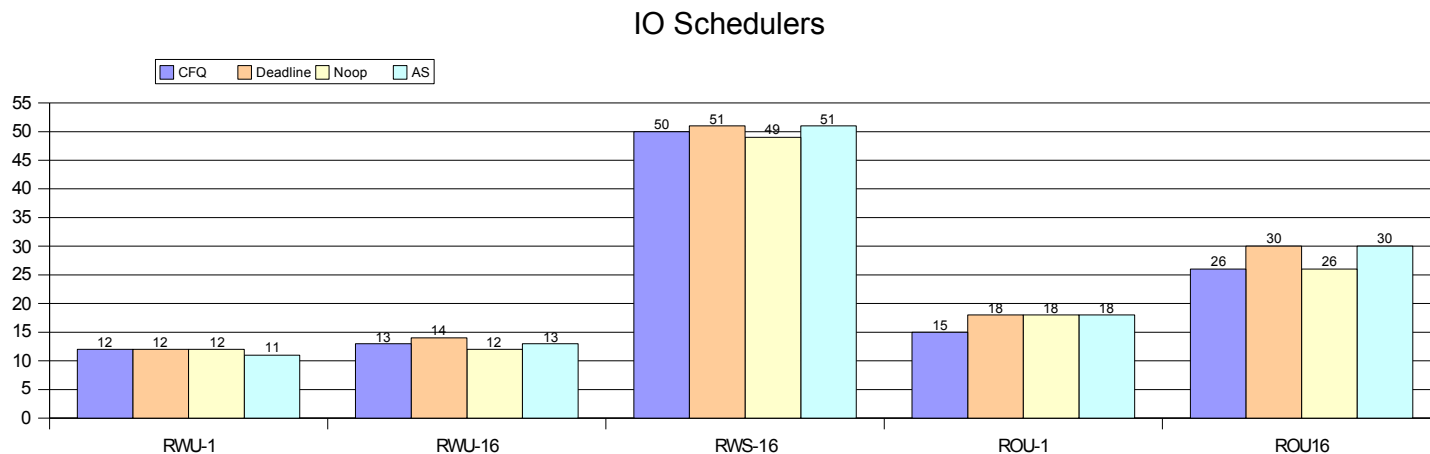
- Using RAID1 on both of them
 - PE 2850 has hardware RAID
 - PE 1425SC uses software one
- Memory sizes 4GB vs 8GB not only disk

Dell Poweredge 2850 and 1425SC



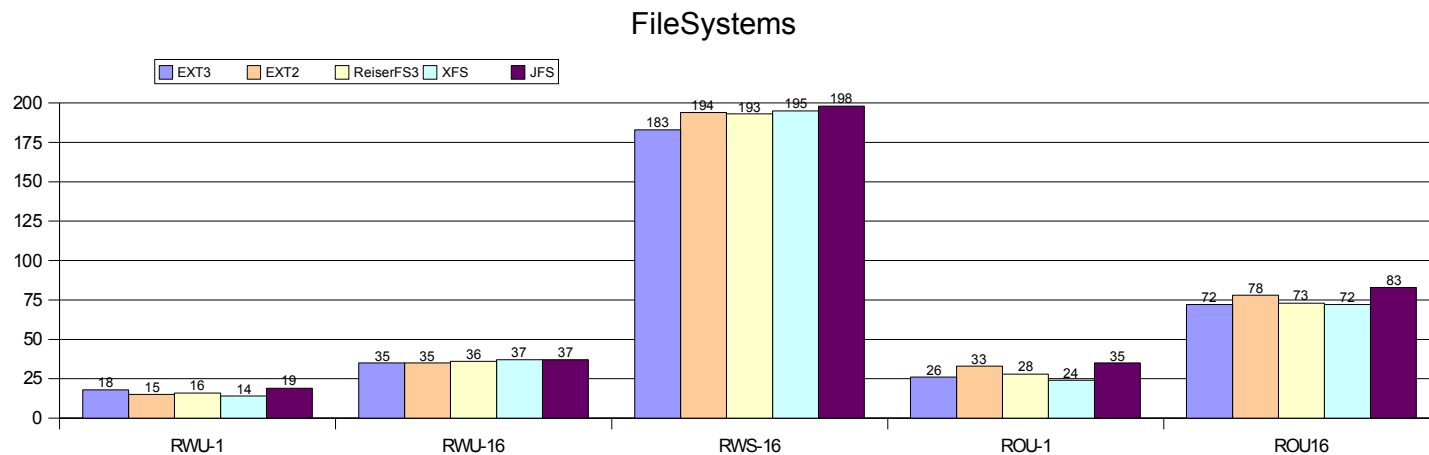
Linux kernel IO Schedulers

- In Linux 2.6 you can select how your disk IO is scheduled
 - Cfq (Completely Fair Queueing) - default for CentOS 4.2
 - Noop (No scheduling)
 - Deadline
 - AS (Anticipatory)



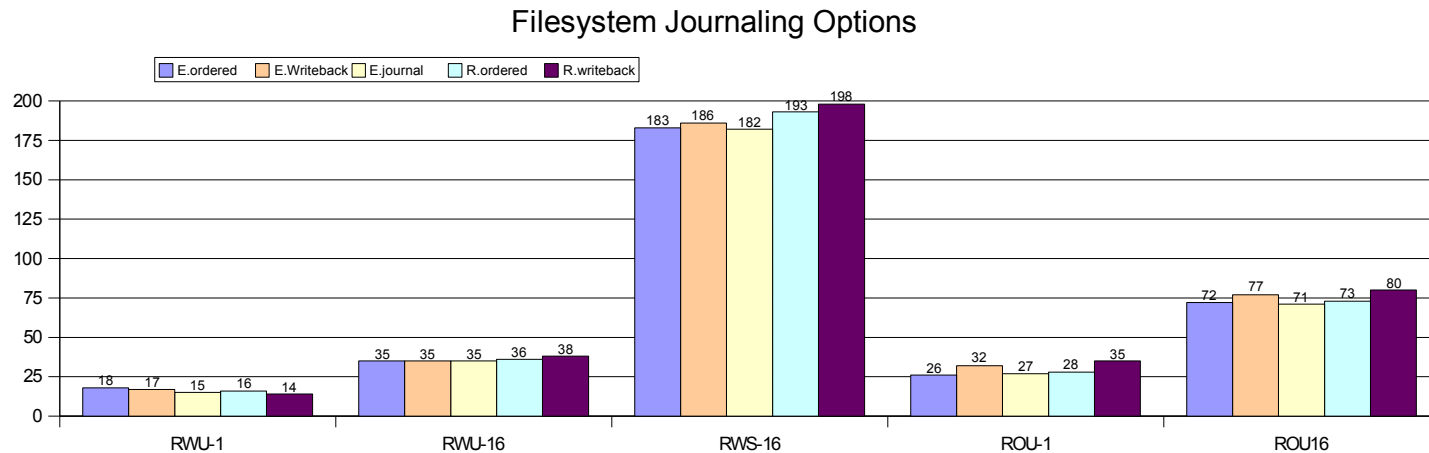
Linux FileSystems

- PowerEdge 2850, 4 Drives in RAID5. O_DIRECT
- Innodb has low requirements about filesystems
 - Single file tablespace, which is preallocated
- JFS is very unexpected winner



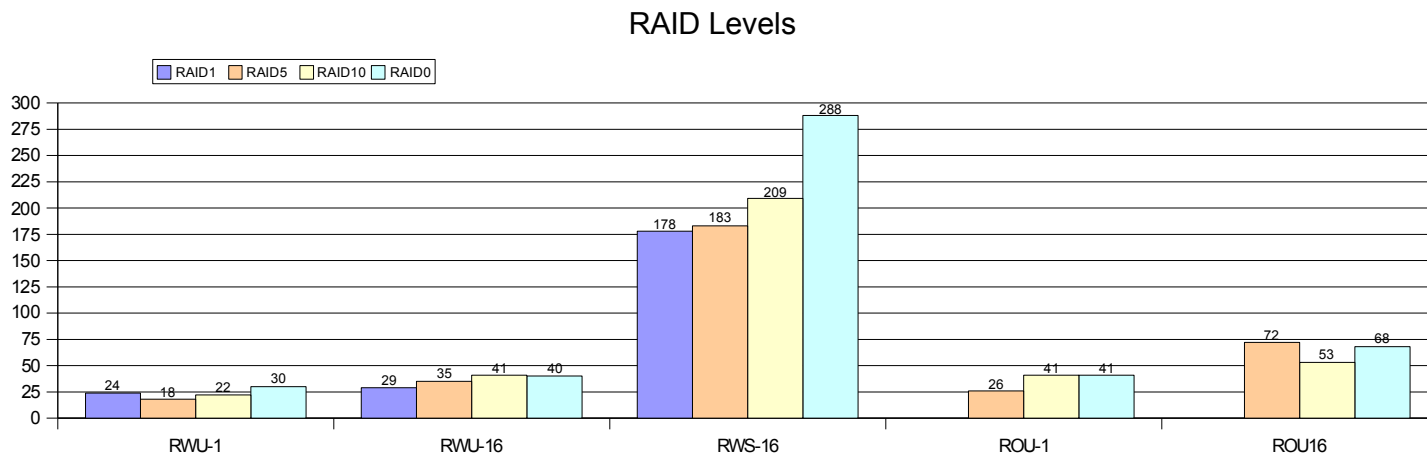
Filesystem Journal Options

- Only tested for EXT3 and ReiserFS
- Writeback gives good boost on ReiserFS
- Especially helps for ReadOnly load. Why ?
 - Might be due to last access time. -o noatime ?



RAID Levels

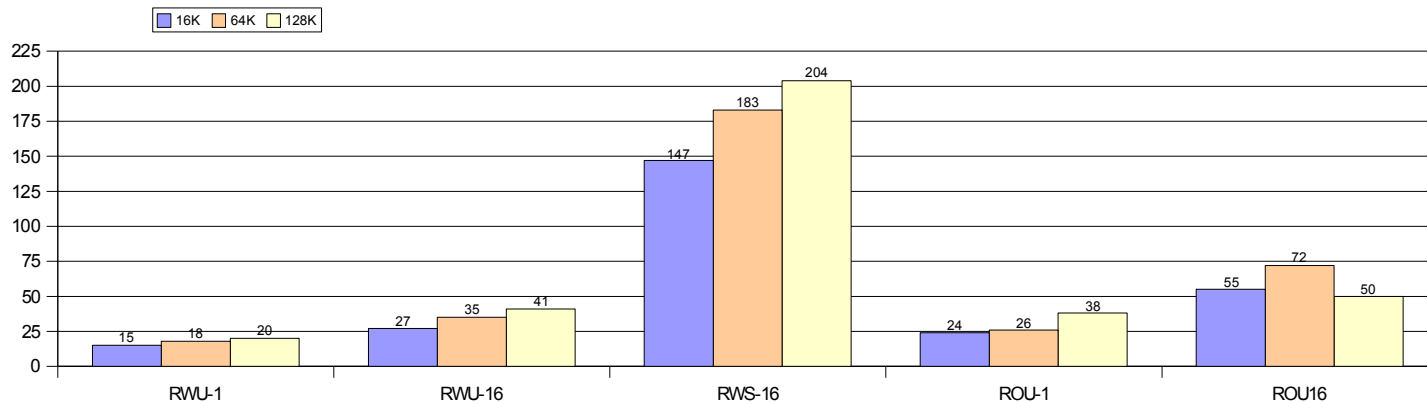
- RAID0 is best performance but insecure
 - Can be used for slaves
- 2 Drive RAID1 may outperform RAID5 for Writes
- RAID10 is good. The difference with RAID5 is a lot controller dependent



RAID Block Sizes

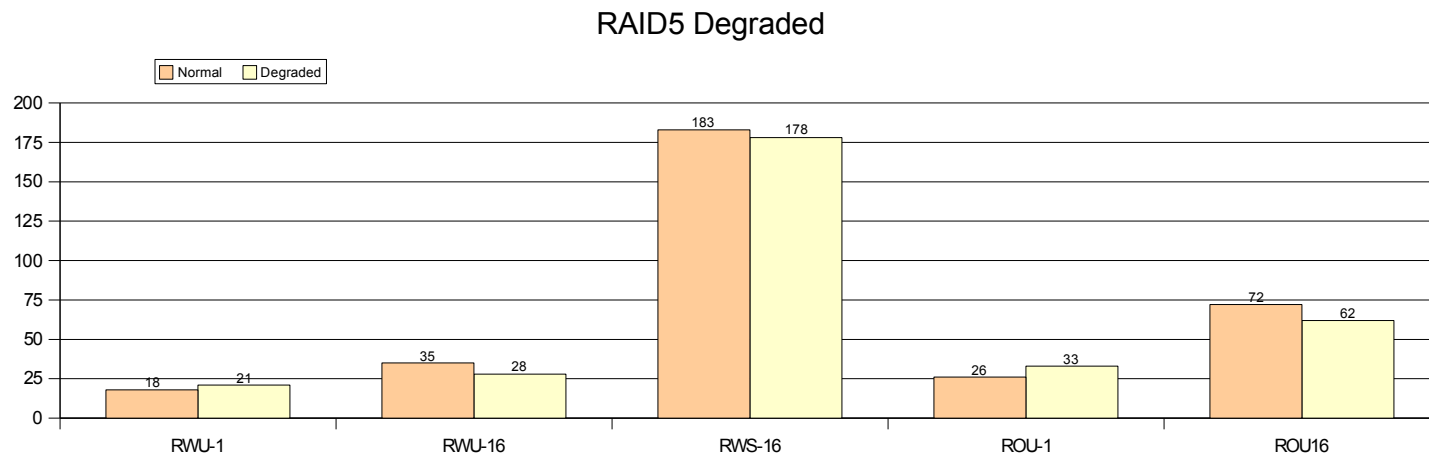
- Innodb page size is 16K
 - So smaller block sizes are worst
- Optimal size a lot depends on RAID controller
- Large block sizes (256K+) are expected to be best for OLTP

RAID5 Block Sizes



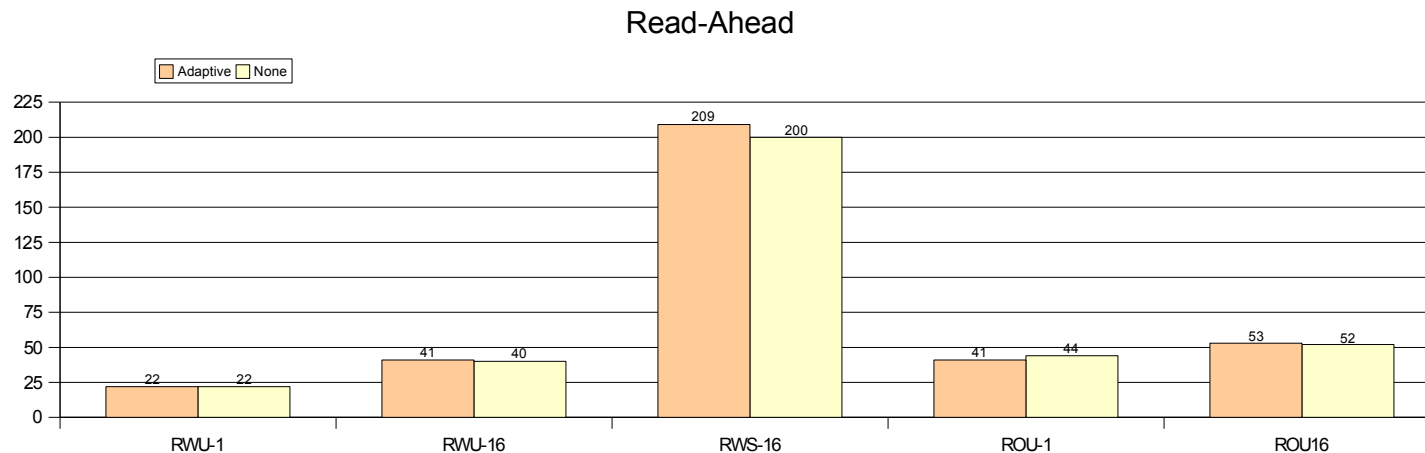
Raid5 Degraded

- Degraded Mode – one of hard drives pulled out
- Why to check in degraded mode ?
 - Because you need to be prepared
 - So your system does not go down because of overload
- Results look really strange but drive was really pulled out



RAID10 Read Ahead

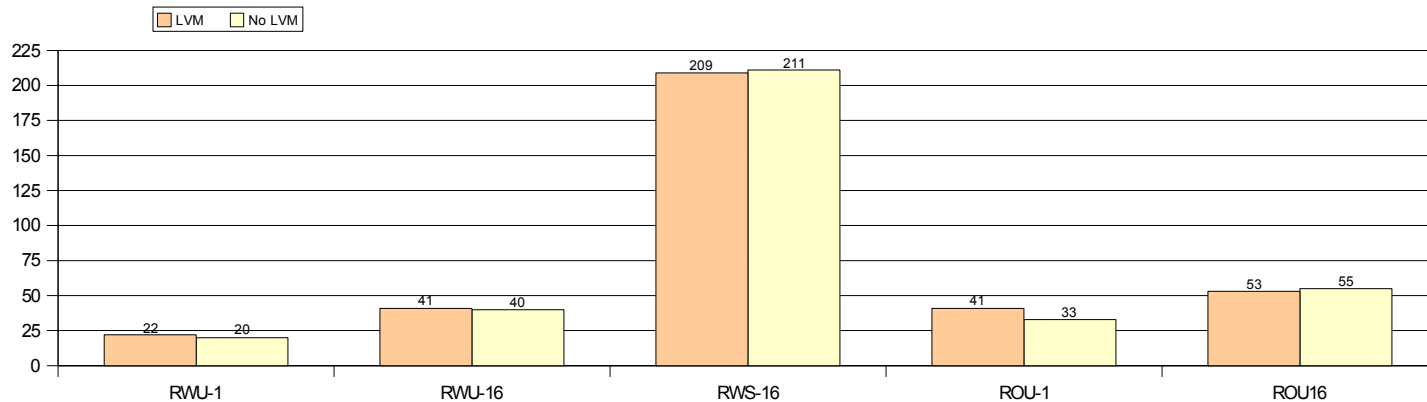
- RAID Read-Ahead Configuration
- Adaptive (default)
 - Read full stripe if more then one access to the stripe



Overhead of LVM

- LVM – Linux Volume Manager
 - Tests are performed with LVM2 by default
- Really good practice
 - Simplifies, speeds up MySQL backup etc
- Normally very low overhead

LVM



Thats all !

- Share you benchmarks experience with me at peter@mysql.com
 - Also send me your questions if you did not have chance to ask them on the conference
- Special Thanks for helping to gather performance data
 - Vadim Tkachanko
 - Aleksey Kishkin
 - Alexey Kopytov
 - CJ Collier