DTCC 2014中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2014

大数据技术探索和价值发现

数据库SSD缓存的过去与现在

@姜承尧

关于我

- 杭州网易研究院技术经理
- MySQL领域Oracle ACE
- 我的书籍
  - MySQL技术内幕：InnoDB存储引擎、SQL编程
  - MySQL内核：InnoDB存储引擎 卷1（2014年5月）
- 联系方式
  - weibo：@姜承尧
  - jiangchengyao@gmail.com
  - www.innomysql.net

Topics

- SSD缓存介绍
- 数据库SSD缓存
- SSD缓存的应用场景
- SSD缓存的过去与现在
- MySQL InnoDB L2 cache

SSD缓存

Cache is everywhere

- CPU L1、L2、L3 cache
- Memory
- Disk cache
- RAID BBU
- SSD/Flash cache

SSD缓存

SSD缓存

- flash cache
- 将SSD作为缓存设备
- 提高整体性能

为什么不直接使用SSD

- 容量因素
- 成本因素

SSD缓存

SSD缓存软件解决方案

- facebook flash cache
- Linux bcached

SSD缓存硬件解决方案

- sTec Enhanced IO
- Mac Fusion Drive
- SSHD（Solid State Hybrid Drive）


充分利用SSD的高随机IOPS特性
write-back cache
缓存热点数据
SSD缓存
write-alloc策略

数据库SSD缓存

SSD缓存在数据库中的应用

- facebook flash cache

SSD缓存案例

- Facebook
- taobao
- Netease

数据库SSD缓存应用场景

适用场景

- 热点数据集中
  - 互联网应用
  - 门户、微薄
  - 架构变动小

不适用场景

- 对相应时间非常敏感的应用
- 写入密集型应用

SSD缓存的过去与未来

过去

- 高大上

现在？？？

SSD缓存并不便宜

- 考虑SSD+HDD的整体成本

应用准则：

- 少量SSD缓存就能达到SSD性能

MySQL InnoDB L2 cache

块设备的SSD缓存

- 一个热点图片被反复读取
- 图片不会发生更新
- 第一次读取即可将图片放入到缓存

数据库SSD缓存

- 数据库5分钟规则
- 数据库已经有自己内存缓存（Buffer Pool）
- 页/块通常需要被更新

MySQL InnoDB L2 cache

数据库SSD缓存的特点

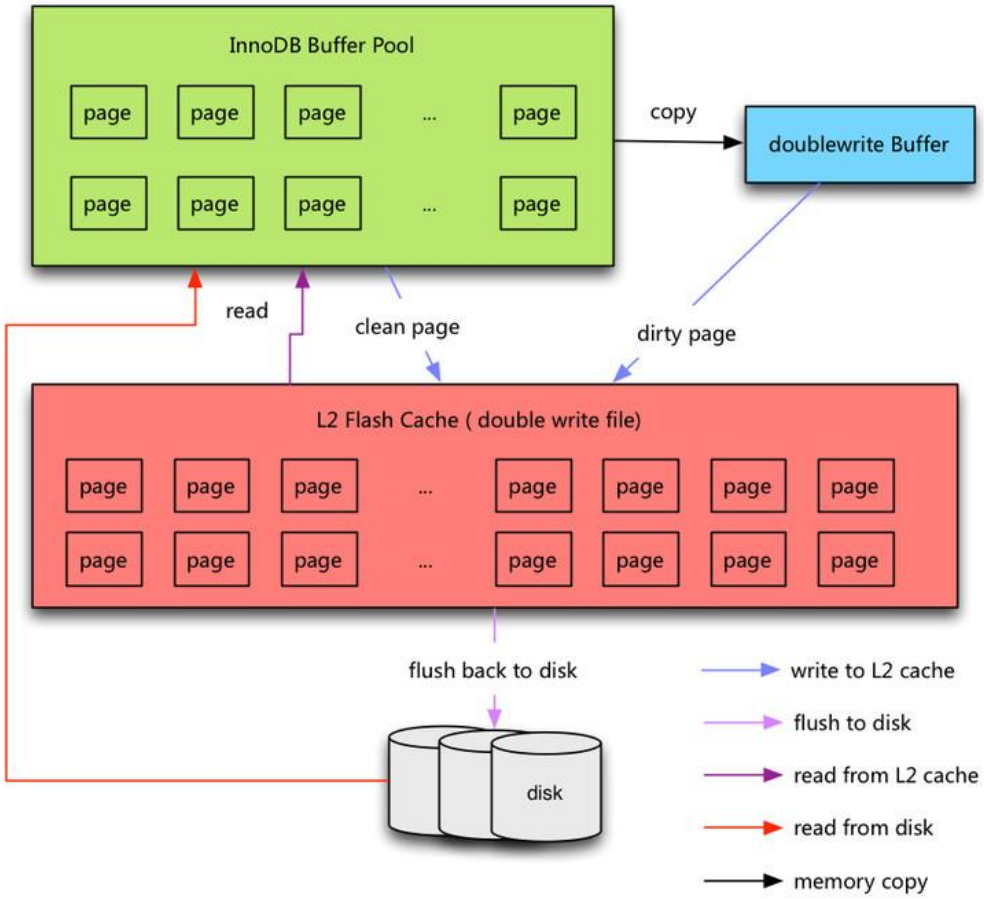
- 一个页/块可能仅被读取1～2次
  - Buffer Pool
- 页不需要在读取时即放入SSD缓存
  - write-alloc => LRU-out
- 缓存真正有用的对象

MySQL InnoDB L2 cache

简介

- 数据库引擎层的SSD缓存
- 针对引擎特性进行优化
- 针对SSD特性进行优化
- write back/write through cache
- L2 cache

MySQL InnoDB L2 cache



MySQL InnoDB L2 cache

特点

- 使用doublewrite做为L2 cache
  - 没有额外的写入开销
- SSD全顺序写入
  - 符合SSD写入优化
- 不使用write-alloc的写入策略
  - LRU-out

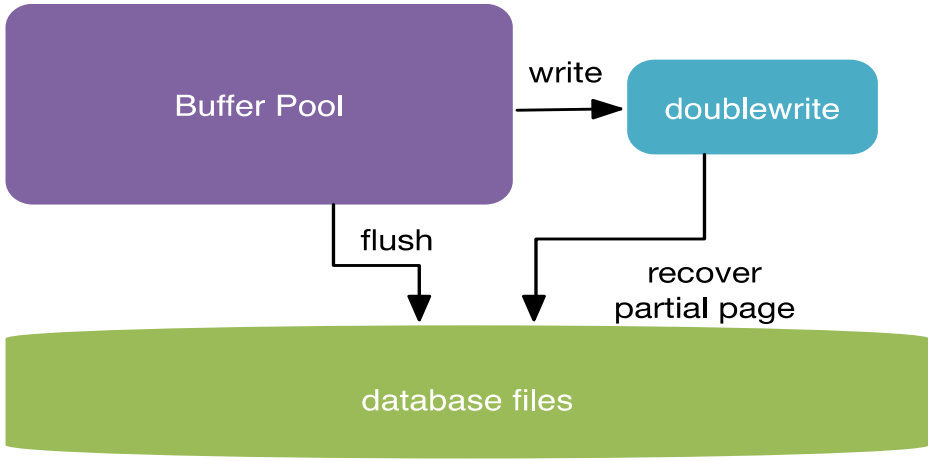
总之

- 性能更优
- 速度更快

MySQL InnoDB L2 cache——doublewrite

doublewrite

- 写入都需先进入doublewrite
- doublewrite 2M
- doublewrite是覆盖写



MySQL InnoDB L2 cache——doublewrite

doublewrite as L2 cache

- doublewrite是SSD设备
- 大小可扩展：100G、200G、400G
- 写入是循环写

MySQL InnoDB L2 cache——顺序写入

优点

- 写入性能提高
- SSD设备寿命提升

缺点

- 缓存的数据变少
- 70% ～ 90%

MySQL InnoDB L2 cache——LRU

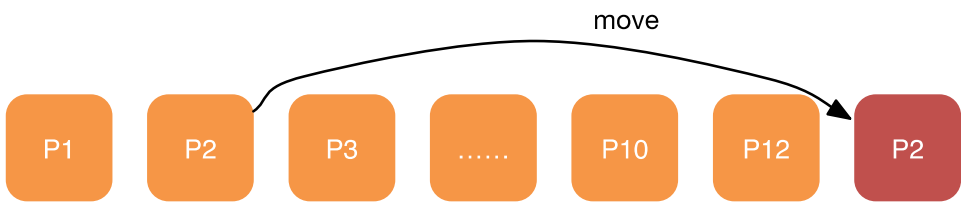
页仅从LRU刷出后放入到L2 cache

- 加快预热速度
- 减轻L2 cache负担
- 符合数据库特性

MySQL InnoDB L2 cache——FIFO

优化的FIFO策略

- 保证写入是完全顺序的
- 通过读取维持热点数据



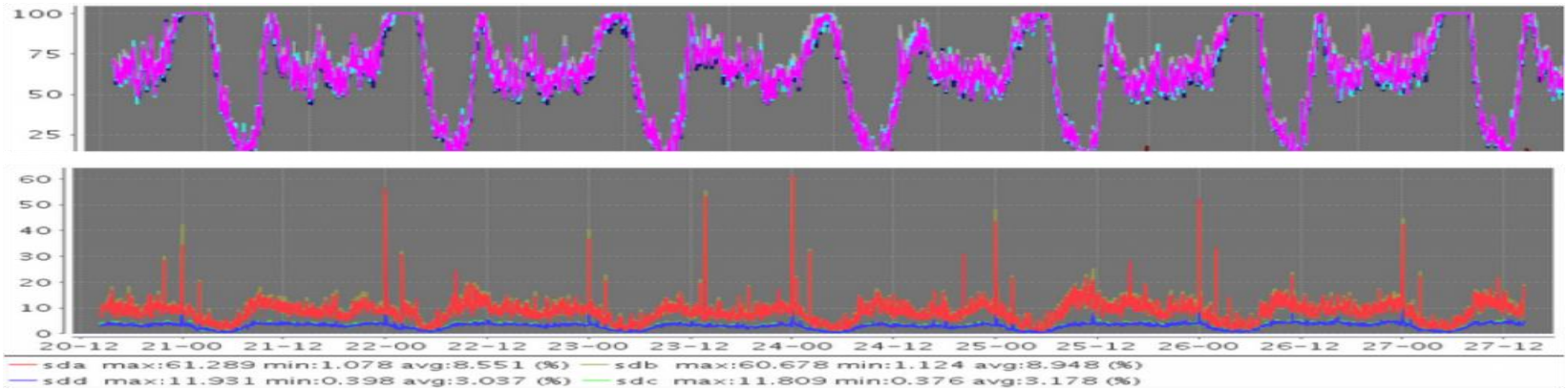
MySQL InnoDB L2 cache——应用

2011年6月正式上线

网易生产环境

- 网易云阅读
- 网易云音乐

SAS 600G => SSD 100G + SATA 2T



MySQL InnoDB L2 cache——2.0

L2 cache 2.0

- 并发优化
- 动态块大小
- 支持压缩功能
- Bug修复

MySQL InnoDB L2 cache——2.0

并发优化

- 锁拆分
  - fc mutex、hash mutex、fc block mutex
- hash mutex => hash rw-lock
- LRU-out mutex优化
- 代码优化
  - fc_write、fc_flush
  - 锁持有时间更短

MySQL InnoDB L2 cache——2.0

压缩

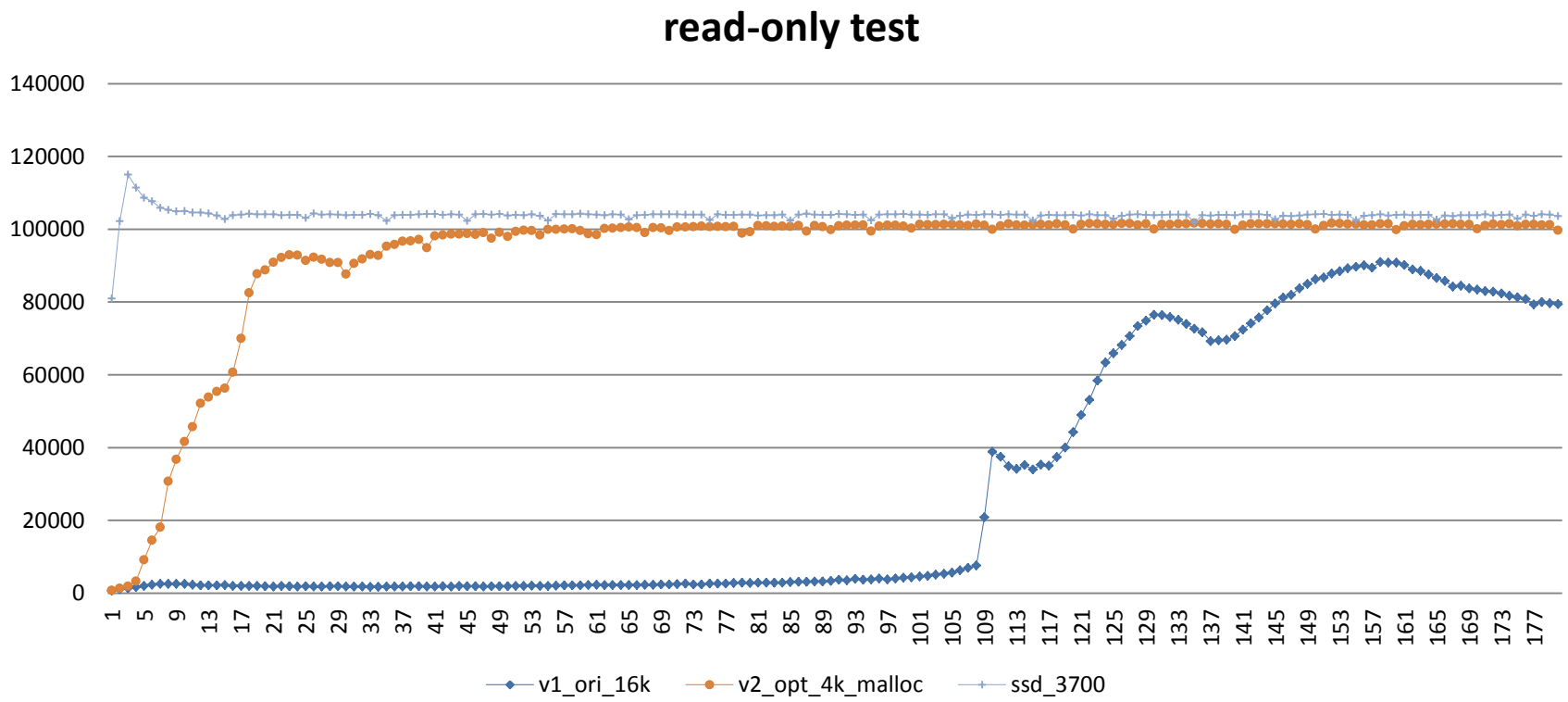
- 使用QuickLZ算法
- CPU开销5%～10%
- cache容量大幅提高30%～80%

MySQL InnoDB L2 cache

动态块大小支持

- 1K、2K、4K、8K、16K进行管理
- 更好的支持InnoDB压缩表
- 支持更细力度的压缩
- 通常建议设置为4K

MySQL InnoDB L2 cache——benchmark



Database: 250G

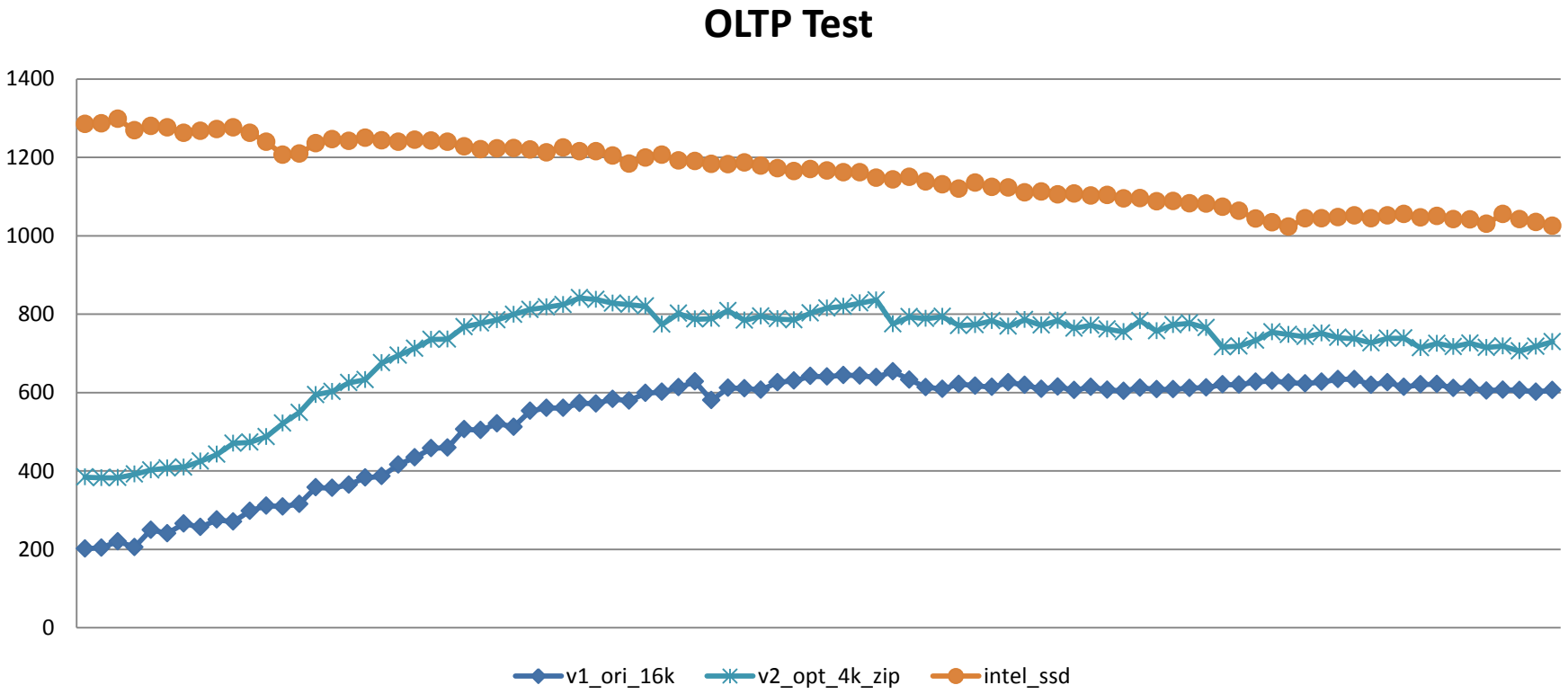
BP: 40G

SSD: 100G

Gain practically 100% SSD performance！！！

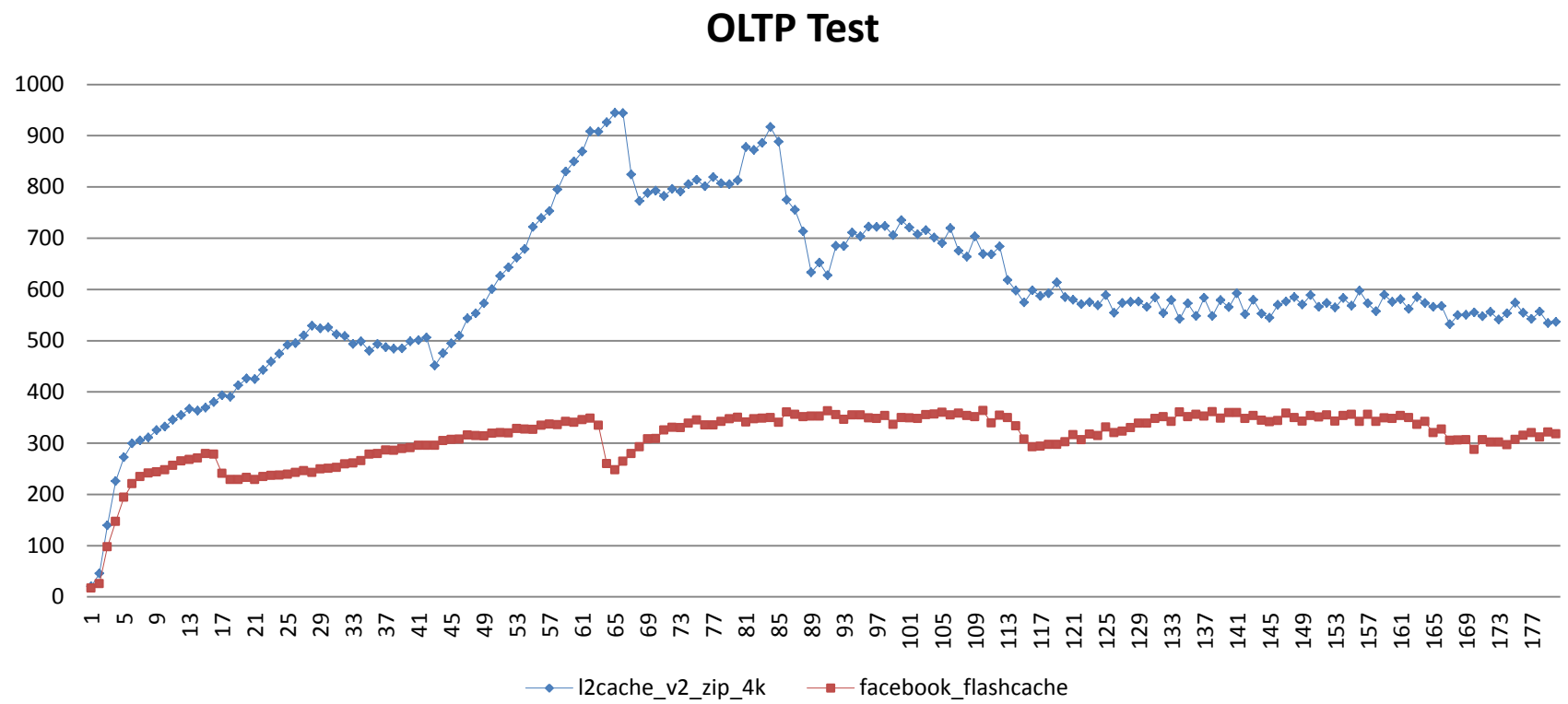
MySQL InnoDB L2 cache——benchmark

- Database: 250G
- BP: 20G
- SSD: 80G
- **Gain 70% SSD performance ！ ！ ！**



MySQL InnoDB L2 cache——benchmark

- L2 cache VS Facebook flash cache



MySQL InnoDB L2 cache

L2 cache 2.0 Now RC

Will be GA in 8th May

Download:

- http://mysql.netease.com/?page_id=38

L2 cache生产环境使用

- jiangchengyao@gmail.com
- 现场技术支持与培训服务
- **7*24 & all free！！！**

Q&A

THANKS