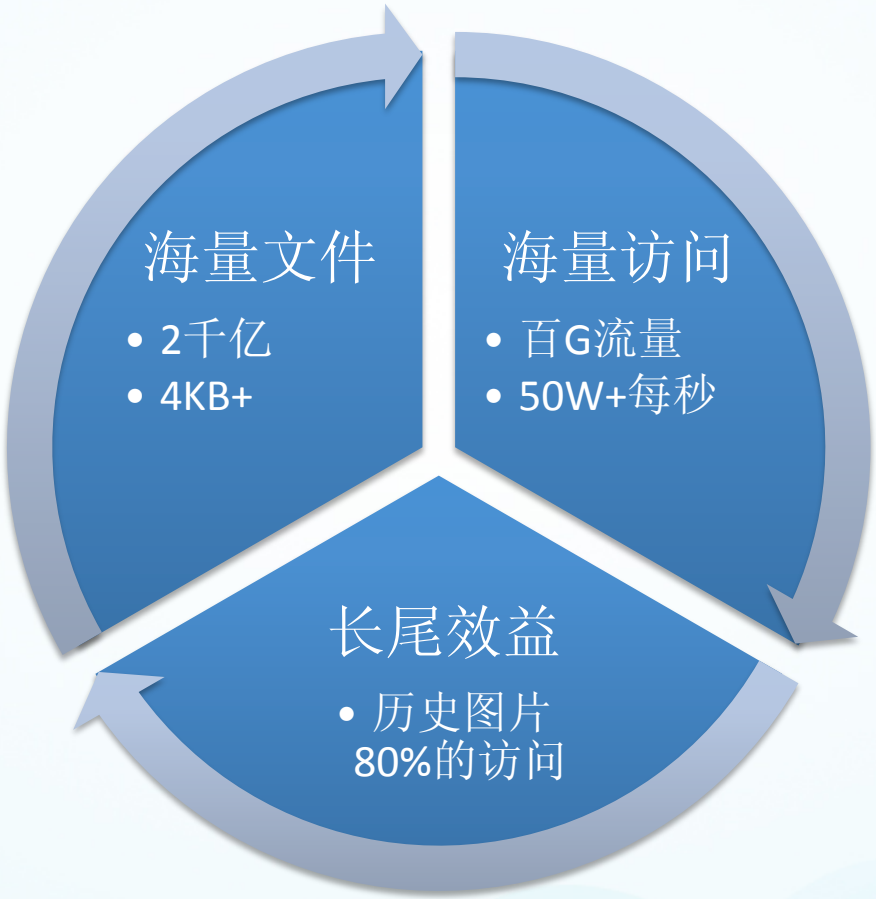
NoSQL在腾讯应用实践

吴悦

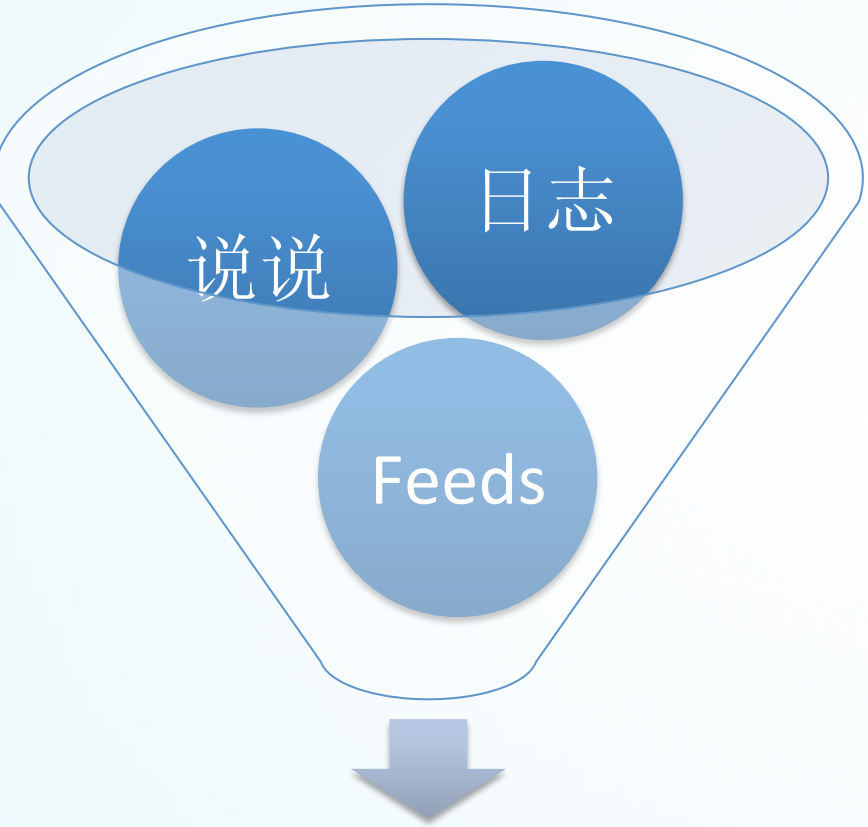
邮箱：joywu@tencent.com

微博：t.qq.com/iwuyue

相册的烦恼



Qzone的特点



近5000亿条记录
每天近5亿次更新

Sns Game的不同

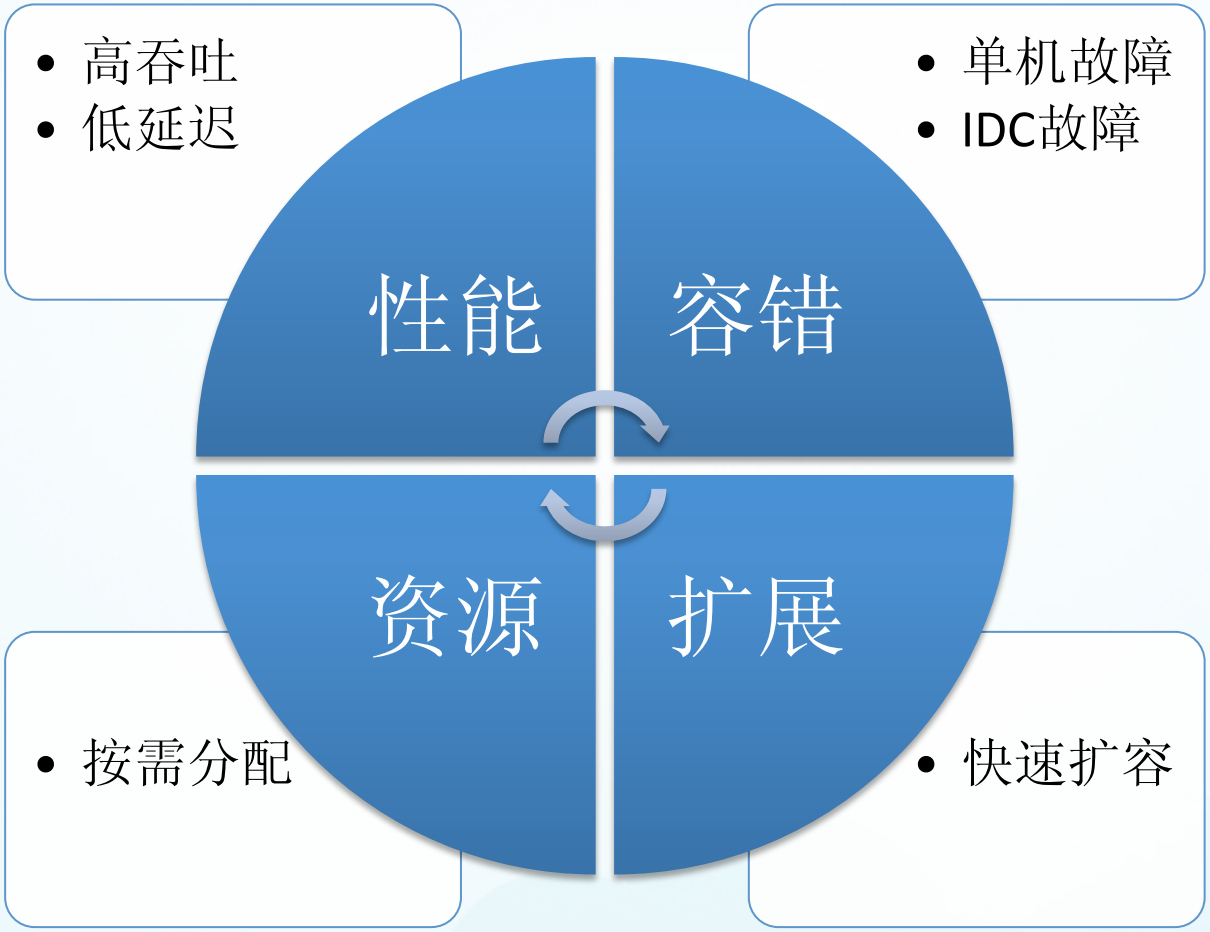
好友互动

交叉访问
更新频繁
放大效应

百W/S
百GB

中小APP的需求

- LAMP（Linux+Apache+**Mysql**+Php）



应对之策

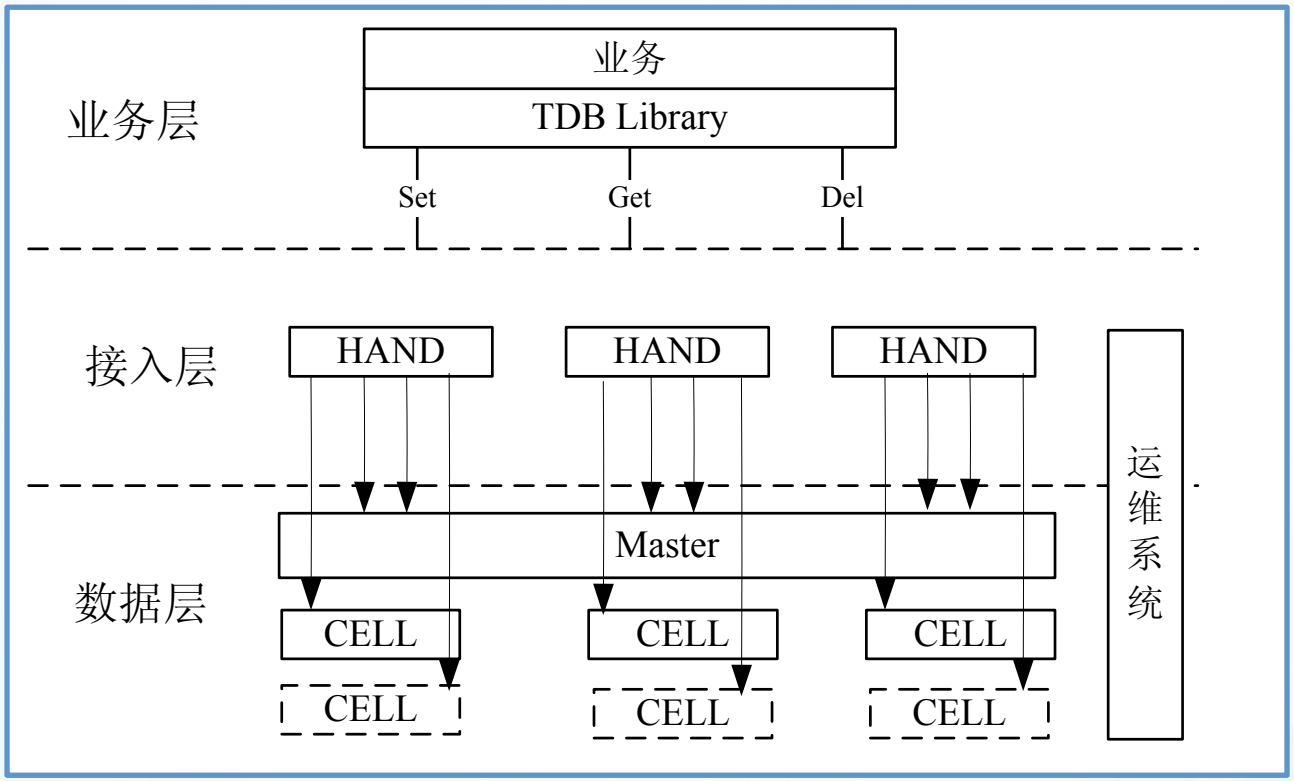
相册的烦恼	• File System - TFS
QZONE 的特点	• NoSQL- TDB/TSSD
SNS Game的不同	• NoSQL– CMEM
第三方APP的需求	• SQL Cluster – CDB

NoSQL-TDB

- 07年研发
- 海量索引管理
- 快速迁移

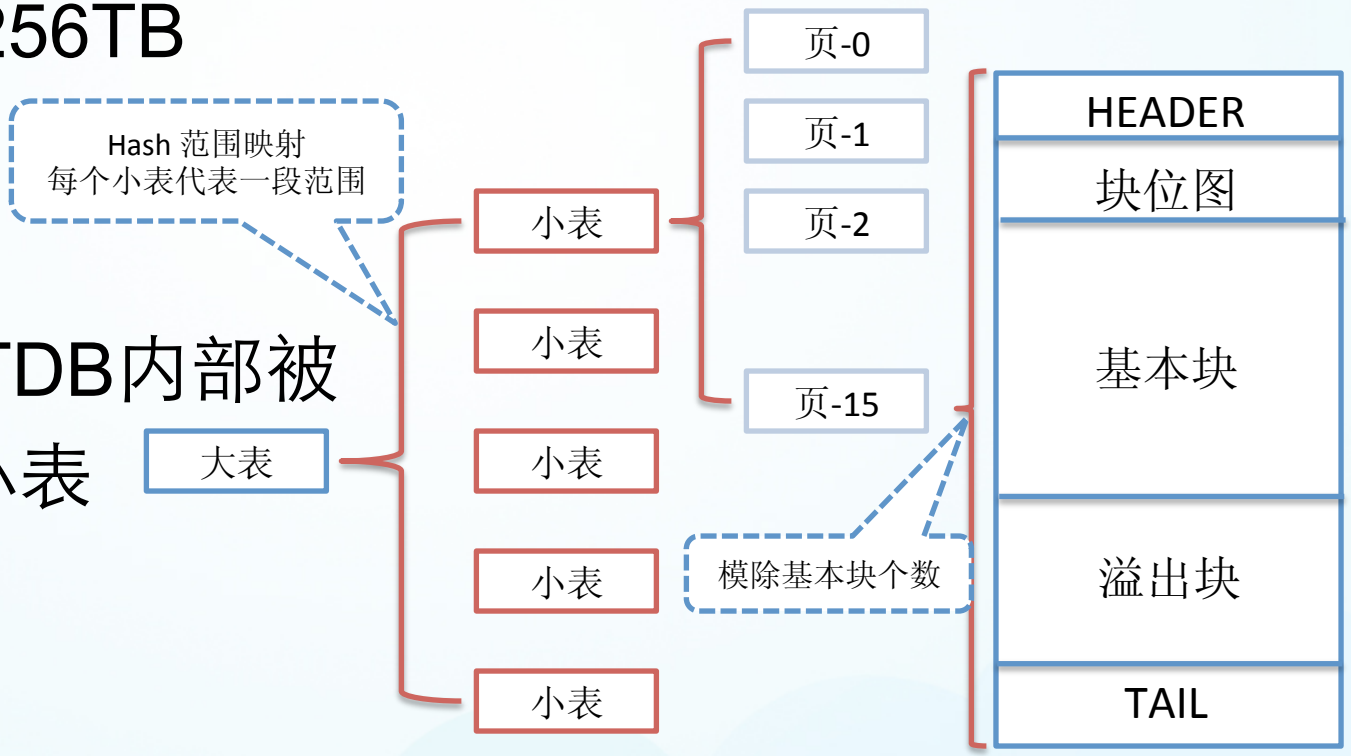
TDB-系统框架

- C.H.M结构



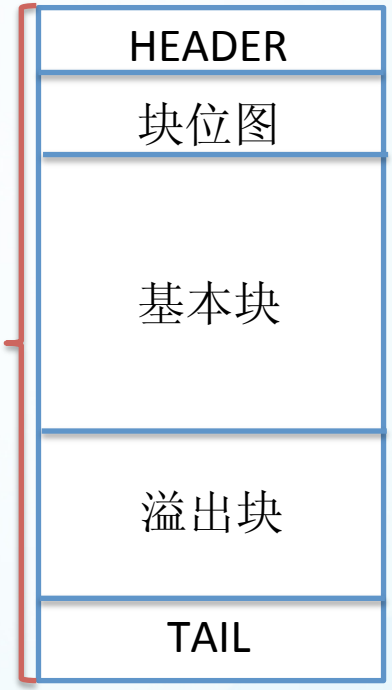
海量索引管理-大表逻辑空间

- 大表
  - 表示一个业务可见的表空间
  - 支持到256TB
- 小表
  - 256MB
  - 大表在TDB内部被分为多个小表



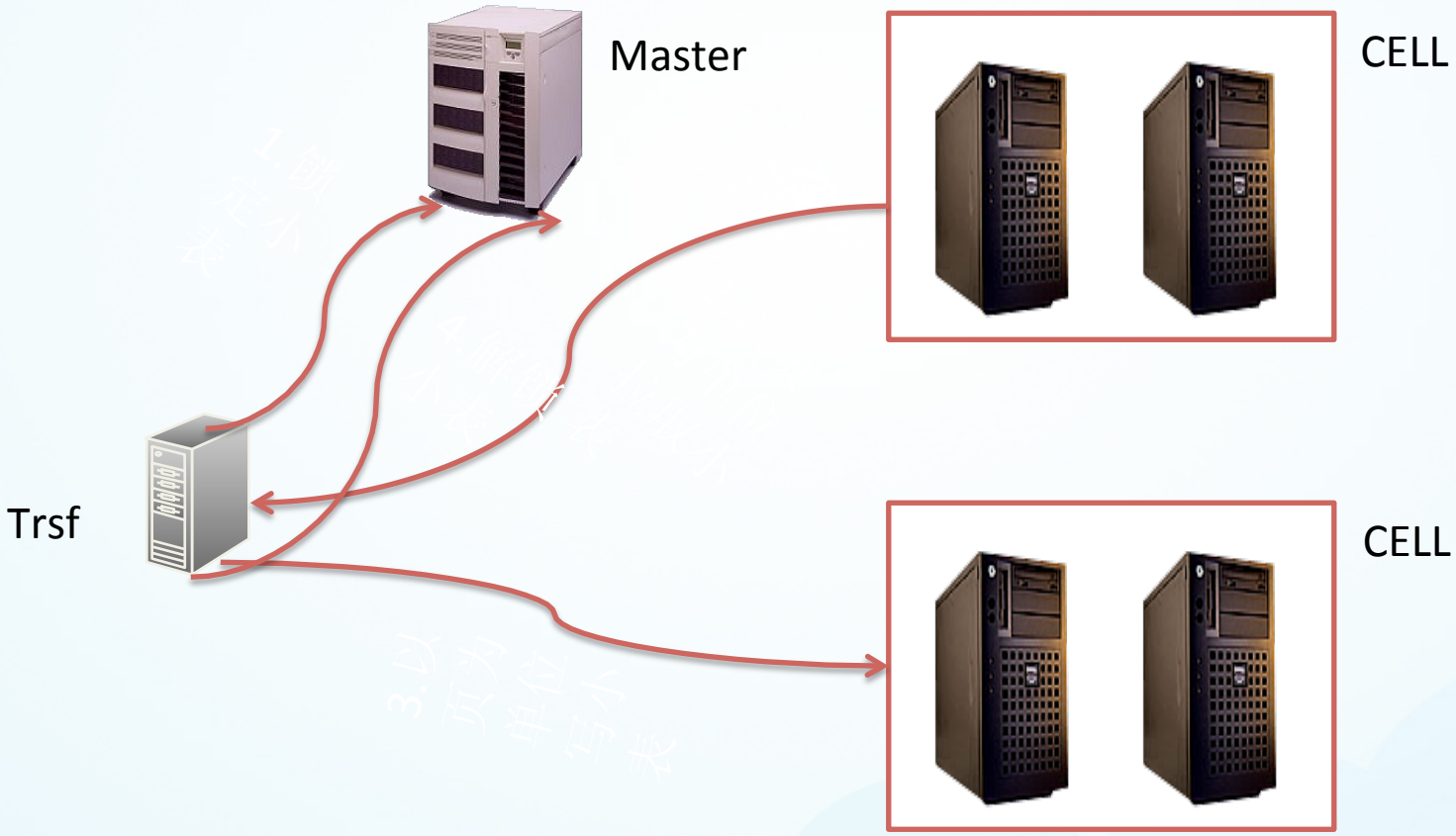
海量索引管理-小表设计

- 小表
  - 分成16个page
  - 每个page是一个hash文件
- 裸设备Direct IO读写
  - 页对应裸设备的特定位置
- 独立写缓存
  - 不能完全相信OS的异步写
  - 可控制写缓存大小
  - 控制写速度



快速搬迁

- 以16MB PAGE为单位进行搬迁



效果

- 索引存储
  - Qzone 5000亿条记录
    - 传统B树设计需500*32G
    - 大表设计需500M
- 空洞
  - 基本无空洞，空间利用率到85%
- IO
  - 80%的记录一次IO

NoSQL - TMEM

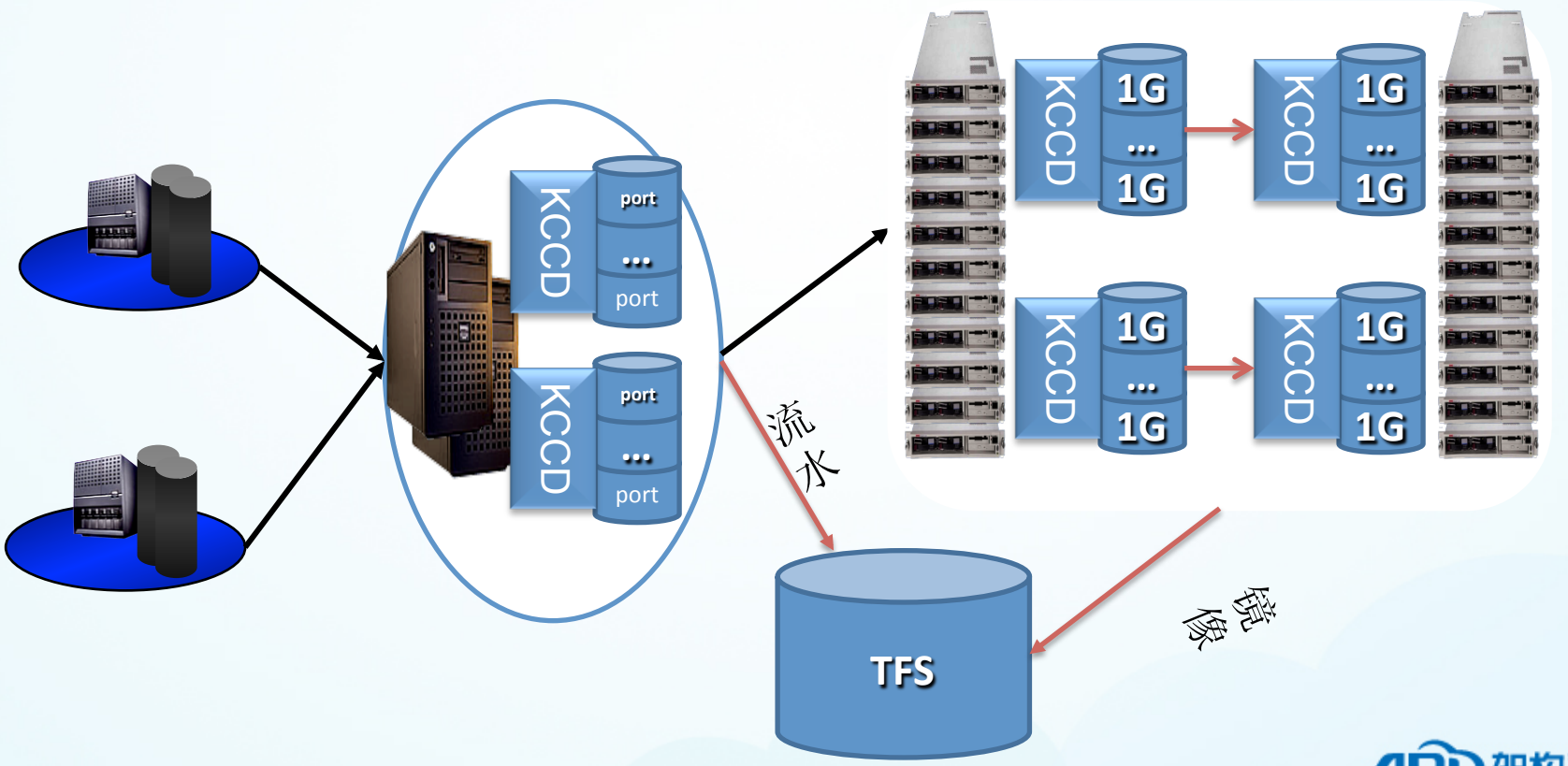
- 09年研发
- 高性能
- 内存可靠性

Memory 特殊性

- 很高的随机读写性能
  - 单机300W + IO/PS
- 易失性存储
  - 掉电后，数据丢失

TMEM-高性能接入+实时备份

- KCCD，解决接入能力
- TFS，做数据流水+镜像



NoSQL-TSSD

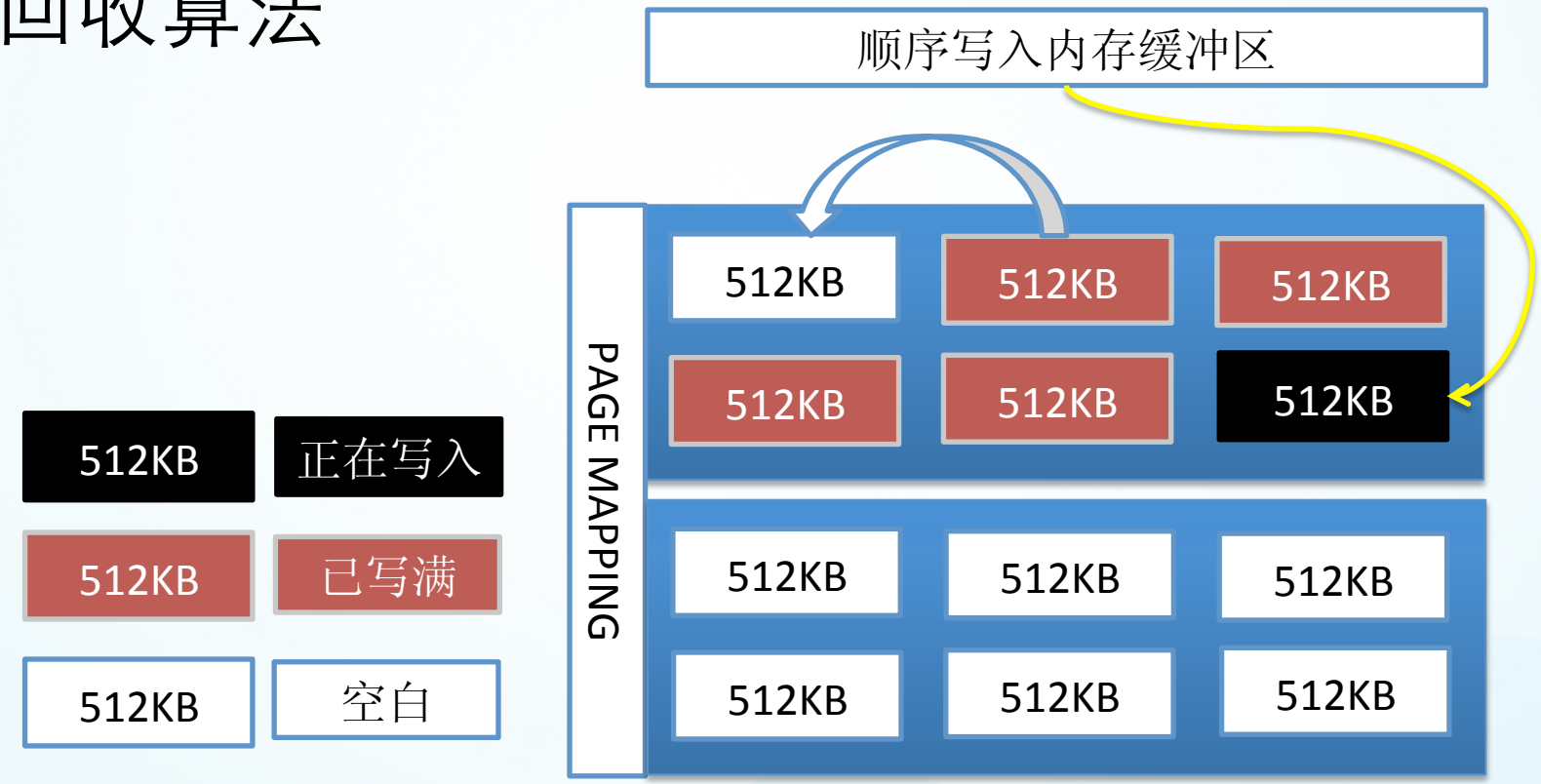
- 10年研发
- 随机写入

SSD 特殊性

- 支持很高的读取IOPS
  - 300GB SSD盘
  - 随机读取4KB，可达30K/S
- 随机写稳定性和寿命不佳
  - 写入放大
  - 300GB的SSD盘
    - 顺序写入：300GB*3000 = 900TB
    - 随机写入：（300GB*3000）（4KB/512KB） = 7TB

TSSD-随机转顺序

- PAGE-MAPPING技术
- 回收算法



阶段小结

SATA (TFS)	SAS (TDB)	SSD (TSSD)	MEM (TMEM)
持久化 大量读，新增 存储100T+	持久化 大量读少量写 存储10+T	持久临时并存 大量读大量写 存储1+T	持久临时并存 大量读大量写 存储10+G
相册，头像，邮件，网盘	Qzone 日志	信息中心、…	农牧场、胡来三国、…

根据IO访问密度选择存储

存储介质	价格($)	容量(GB)	IOPS	每GB IOPS	每GB成本($)
SATA	200	2000	100	1/20	0.10
SAS	370	600	200	1/3	0.62
SSD	1000	600	30000	50	1.67
RAM	440	32	800,000	25,000	13.75

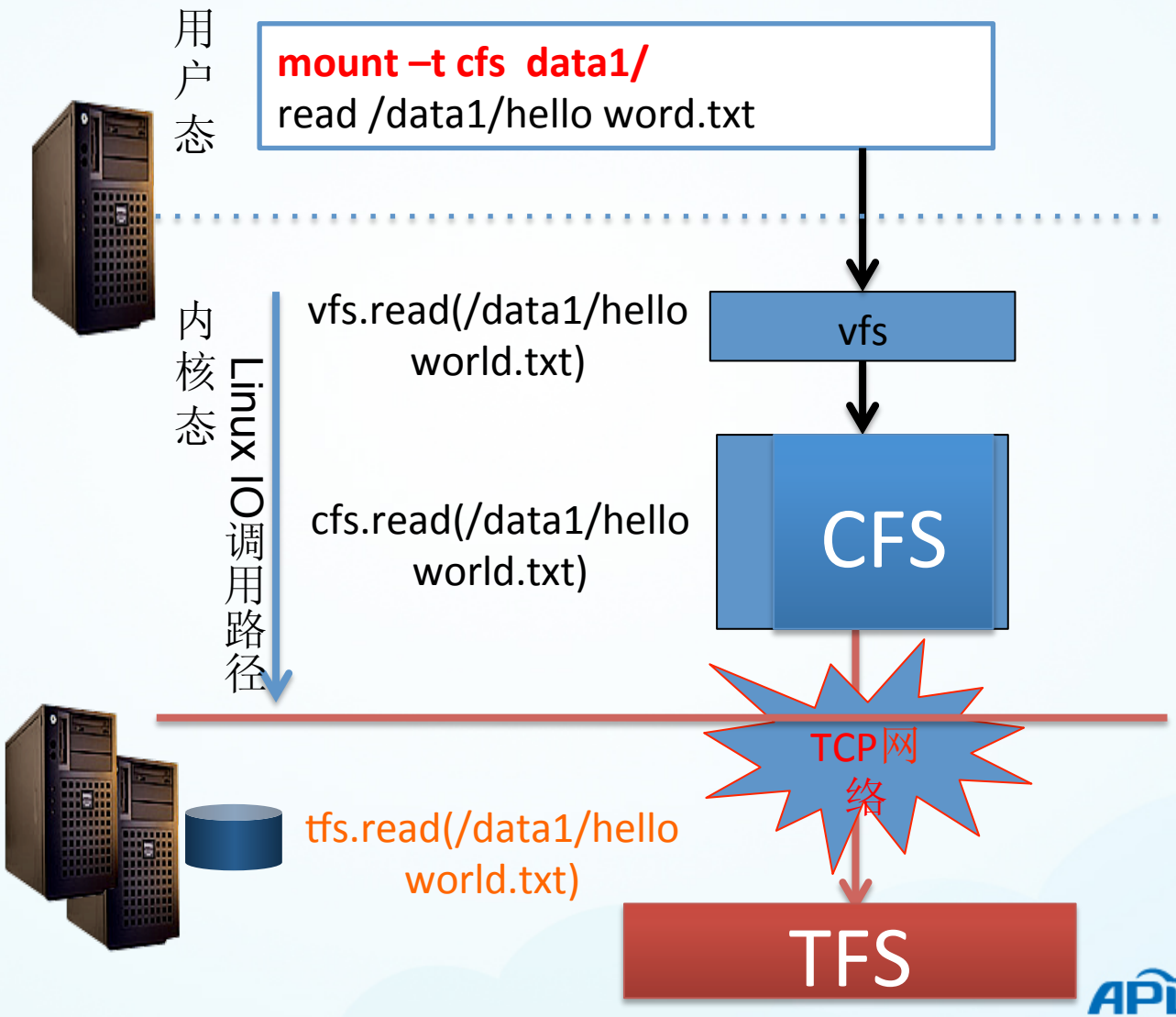
- 计算各种介质适合的IO访问密度

存储介质	每GB IOPS区间	系统
SATA	(0,0.31]	TFS
SAS	(0.31,0.9]	TDB
SSD	(0.9,1664]	TSSD
DRAM	(1664，∞]	TMEM

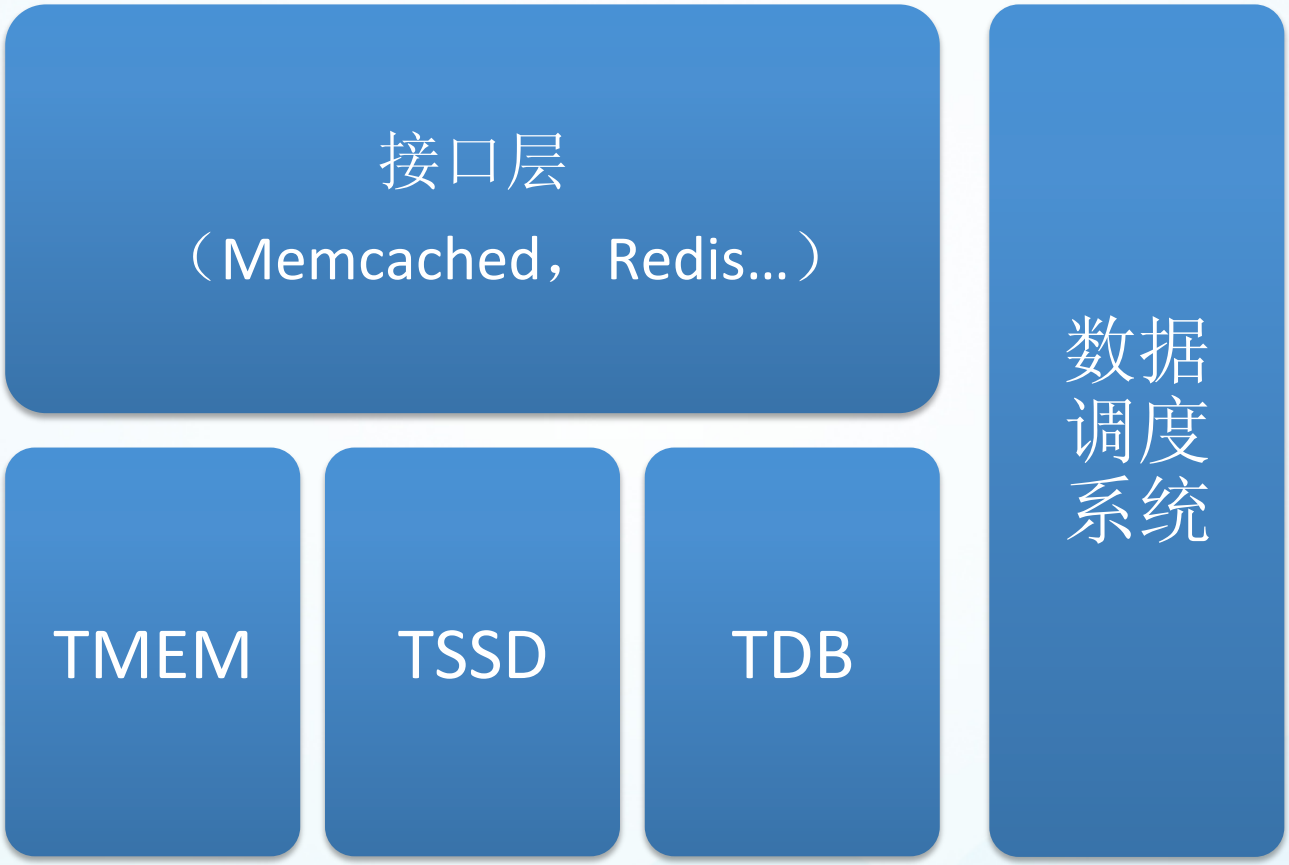
开放的挑战

- 接口的挑战!
- 如何在存储层之上兼容标准接口?
  - Posix
  - Mysql
  - Memcached，redis

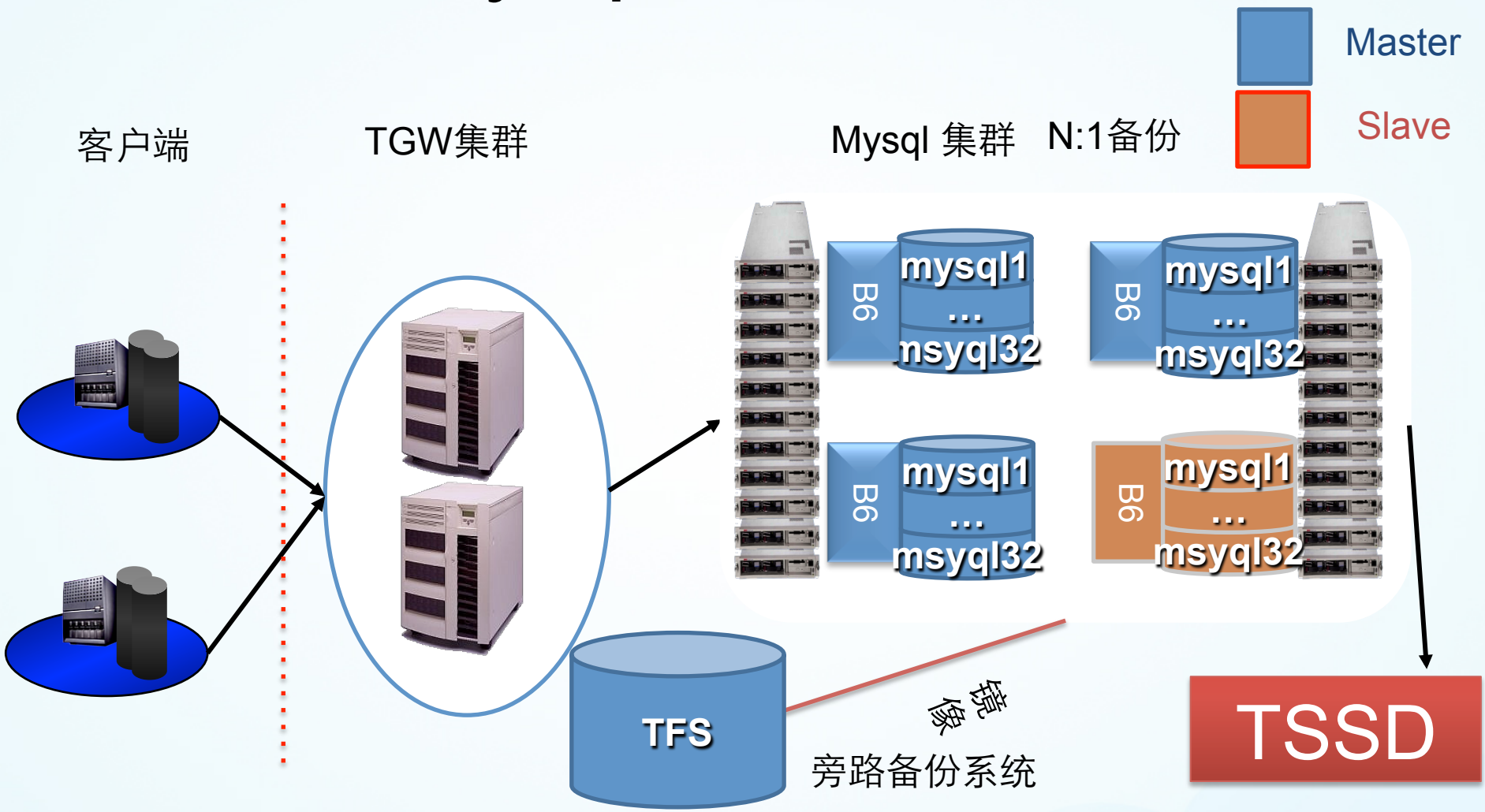
CFS – 像ext3一样使用TFS



CMEM – 像Memcached一样使用NoSQL



CDB – 像Mysql一样使用NoSQL



解决方案总结

应用层

接口层

NoSQL

SQL

POSIX

CMEM

CDB

CFS

存储层

TMEM
TSSD
TDB

TFS

ArchSummit

中国·深圳 2012.08

INTERNATIONAL ARCHITECT SUMMIT

全 球 架 构 师 峰 会

详情请访问：architectsummit.com

- 3天 • 6场主题演讲
- 3场圆桌论坛 • 9场专题会议
- 国内外30余家IT、互联网公司的50多位来自一线的讲师齐聚一堂

主办方：InfoQ 战略合作伙伴：Tencent 腾讯 特别支持：

http://architectsummit.com


QCon
杭州站 · 2012年10月25日~27日
www.qconhangzhou.com（6月启动）
QCon北京站官方网站和资料下载
www.qconbeijing.com