SACC 2014中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2014
发现架构之美

腾讯互娱DB管理平台GCS的迭代路径

tencent dba felixliang
hugefelix@vip.qq.com | felixliang@tencent.com

agenda

- 精品游戏存储的难题
- Services Window自助化
- GCS 1.X – 3.X的演进之路
  - GCS1.0 高可用技术
  - GCS2.0 MySQL分支定制
  - GCS3.0 存储层云化
- GCS 4.X的规划
- Q&A

精品游戏存储的难题


基础DB管理能力
- 基础管理需求多
- 日常变更频繁
痛点 I
透明数据存储能力
- 硬件故障影响面广
- 快速加字段需求大
- Innodb存储优化难
痛点 II
动态运营能力
- 透明扩缩容能力
- 数据化运营能力
痛点 III

精品游戏存储的难题 – 痛点1 运营效率低

- 几个核心数据

250+款游戏(端游+手游)、10000+台服务器、20000+个实例

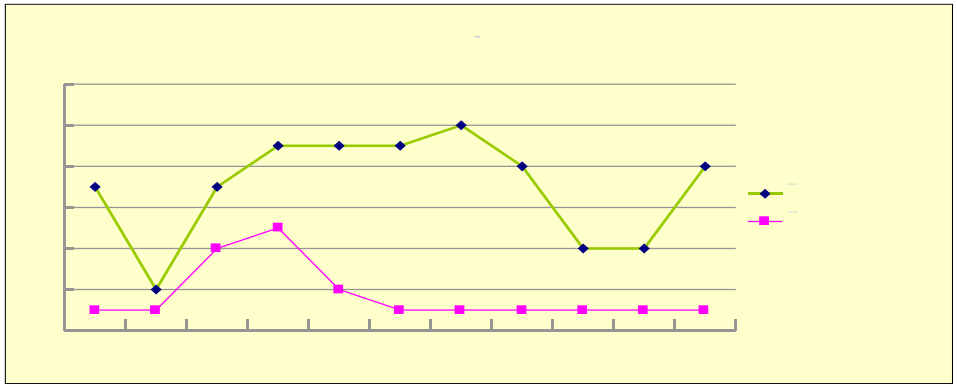
690次SQL变更/月，人均每天支撑2个业务SQL变更，人均管理着500台机器、1000个实例

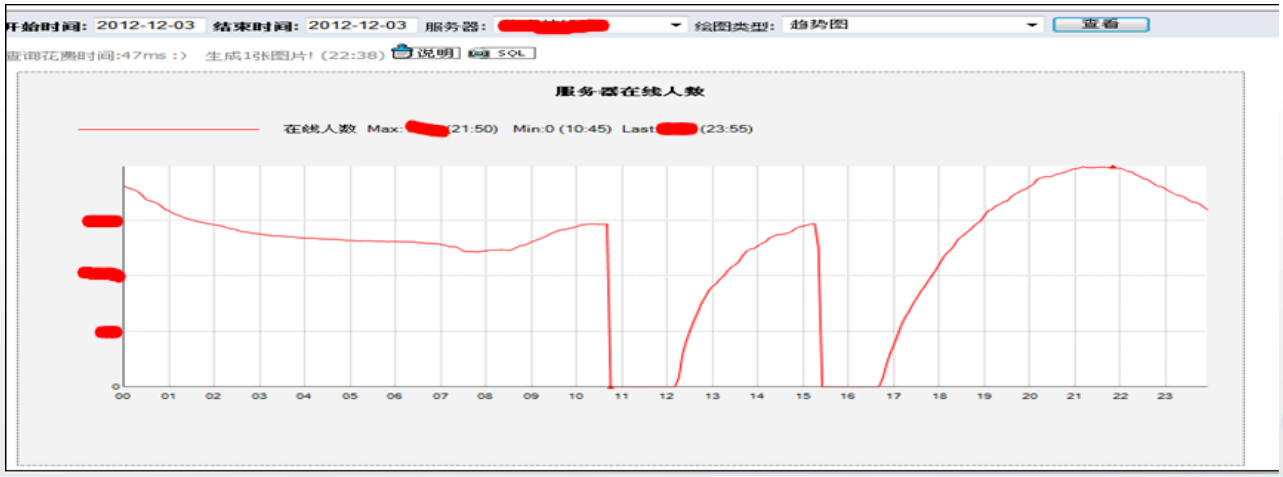
- DBA管理的进程，从进程托管到机器托管

按数据库分类	进程名
oracle	ora_pmon、ora_smon
mysql	mysqld、mysqld_safe、mysql-proxy
sqlserver	sqlservr.exe、SQLAGENT.EXE、sqlbrowser.exe、sqlwriter.exe
mongodb	mongod、mongos
redis	redis-server、nutcracker
Memcache	memcache
tcaplus	tca-server、tca-proxy

精品游戏存储的难题 – 痛点2 玩家体验差

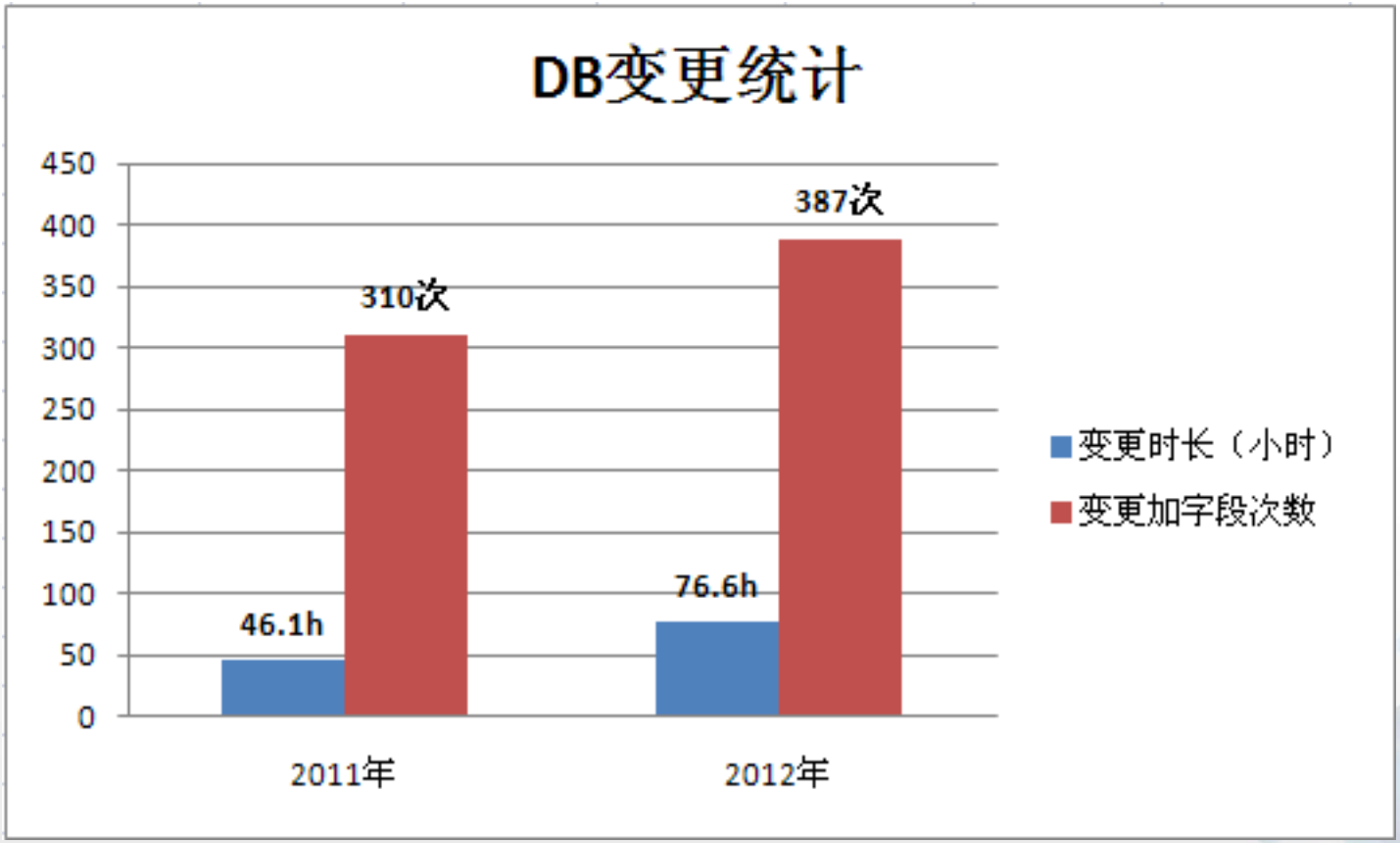
- 硬件故障影响玩家时间长





精品游戏存储的难题 – 痛点2：版本停机时长

- 高星级业务变更加字段停机时间长



精品游戏存储的难题 – 痛点3：成本高

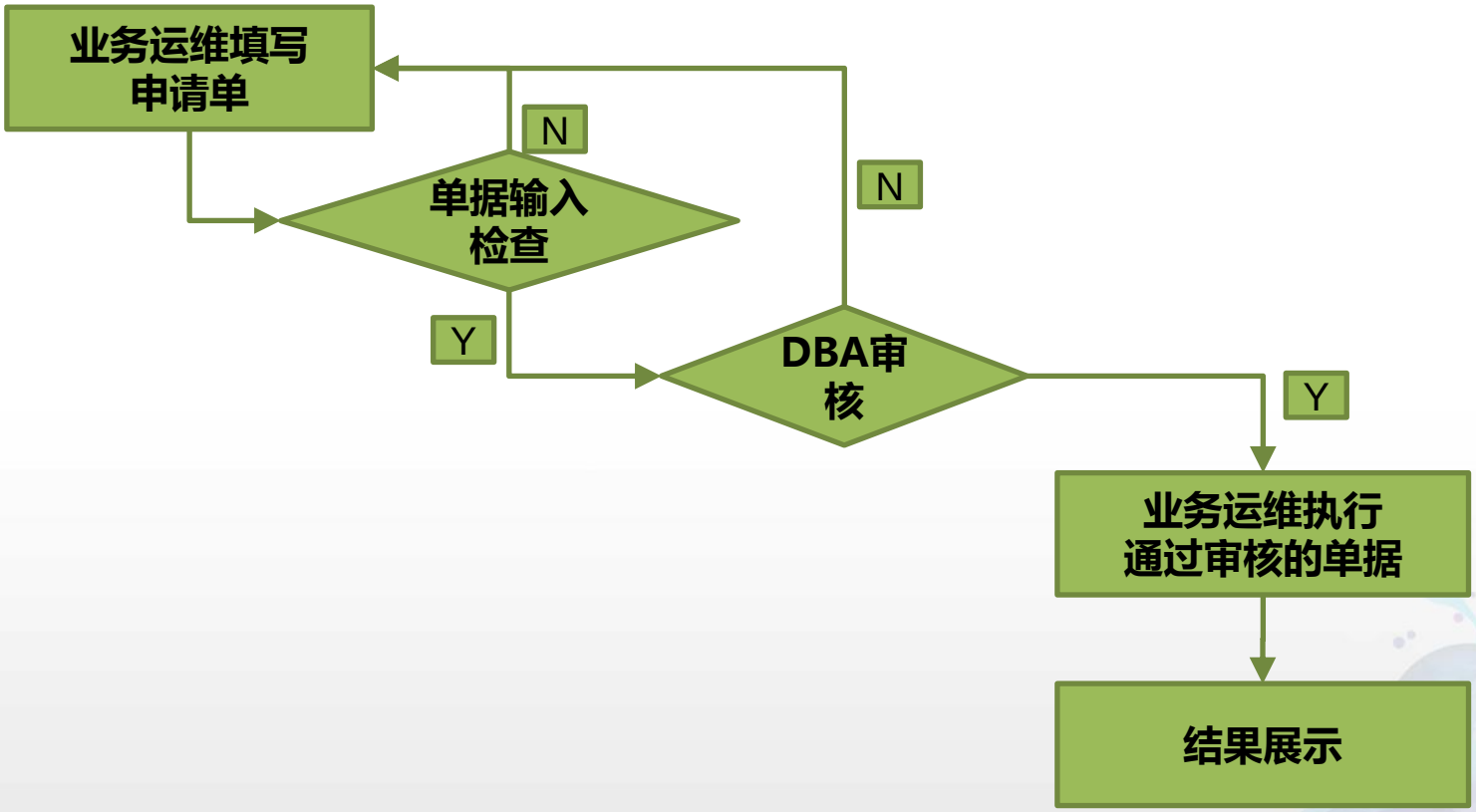
- 2/3机器处在低负载状态
- 不同大区对应DB忙闲不均

Services Window自助化

- 痛点1的应对思路
  - 以统一的Interface管理不同的DB存储类型
  - 提升DB管理效率、释放人力 (90%以上日常需求自助化)

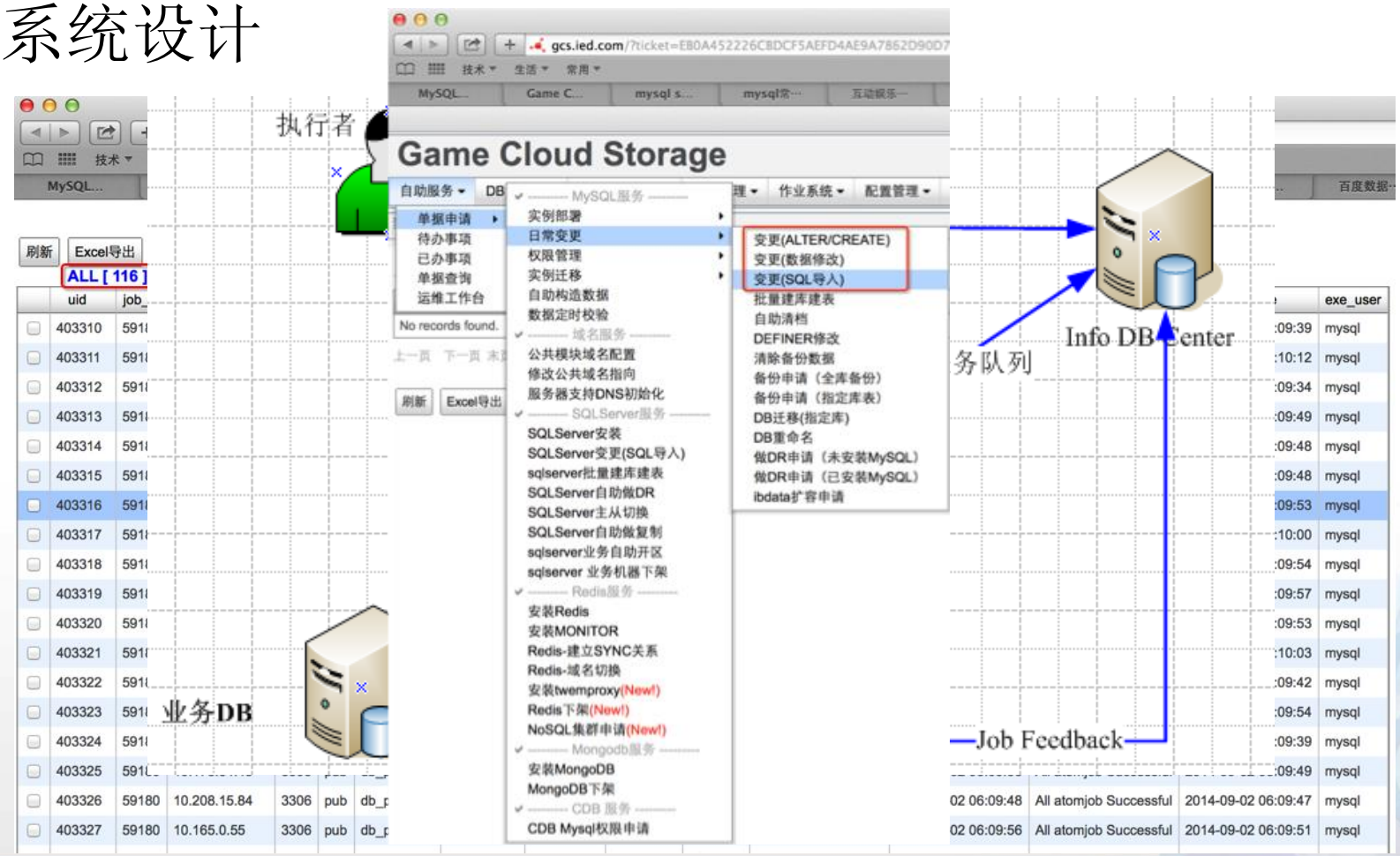
Services Window自助化

- 流程设计



Services Window自助化

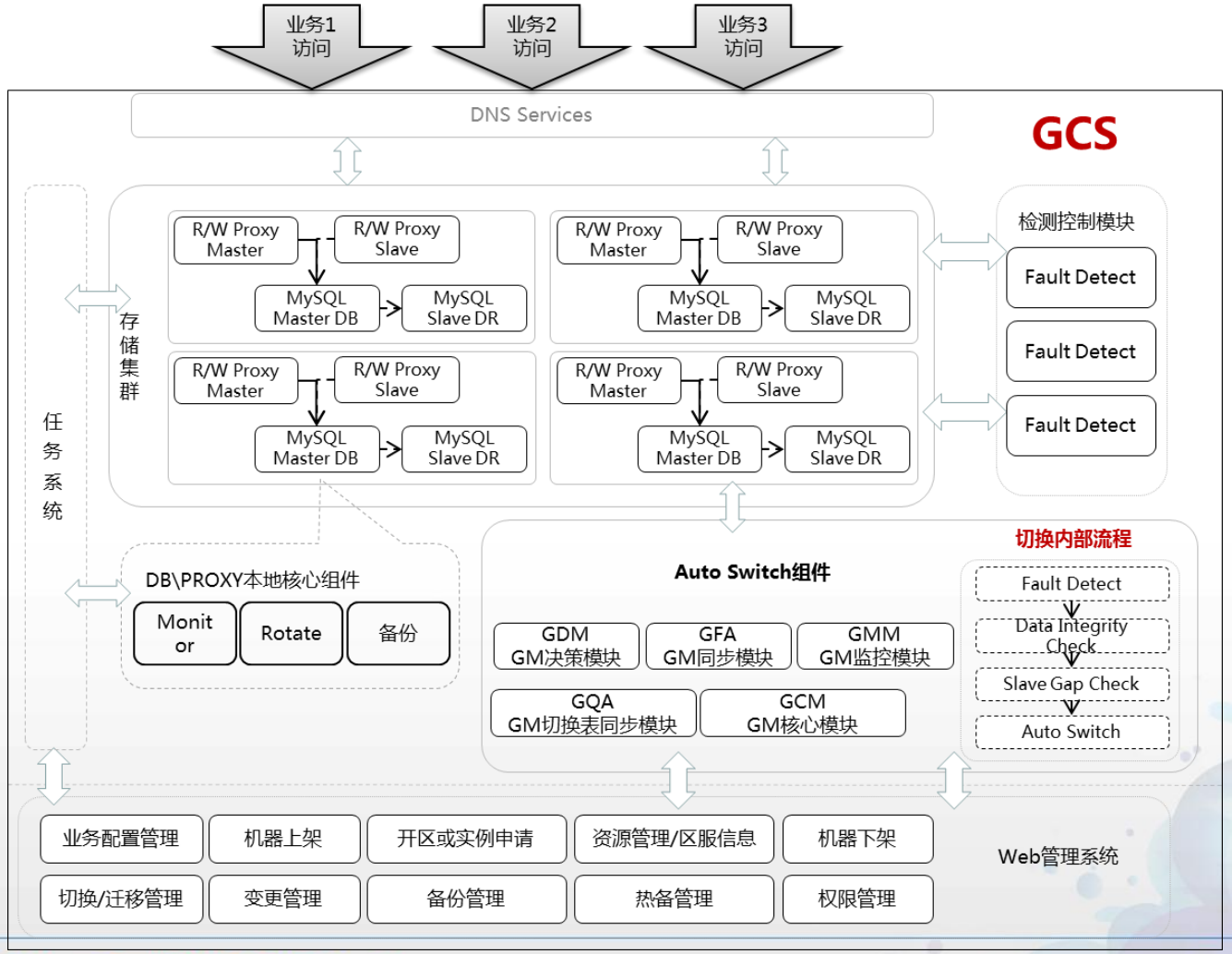
- 系统设计



GCS 1.X – 3.X的演进之路

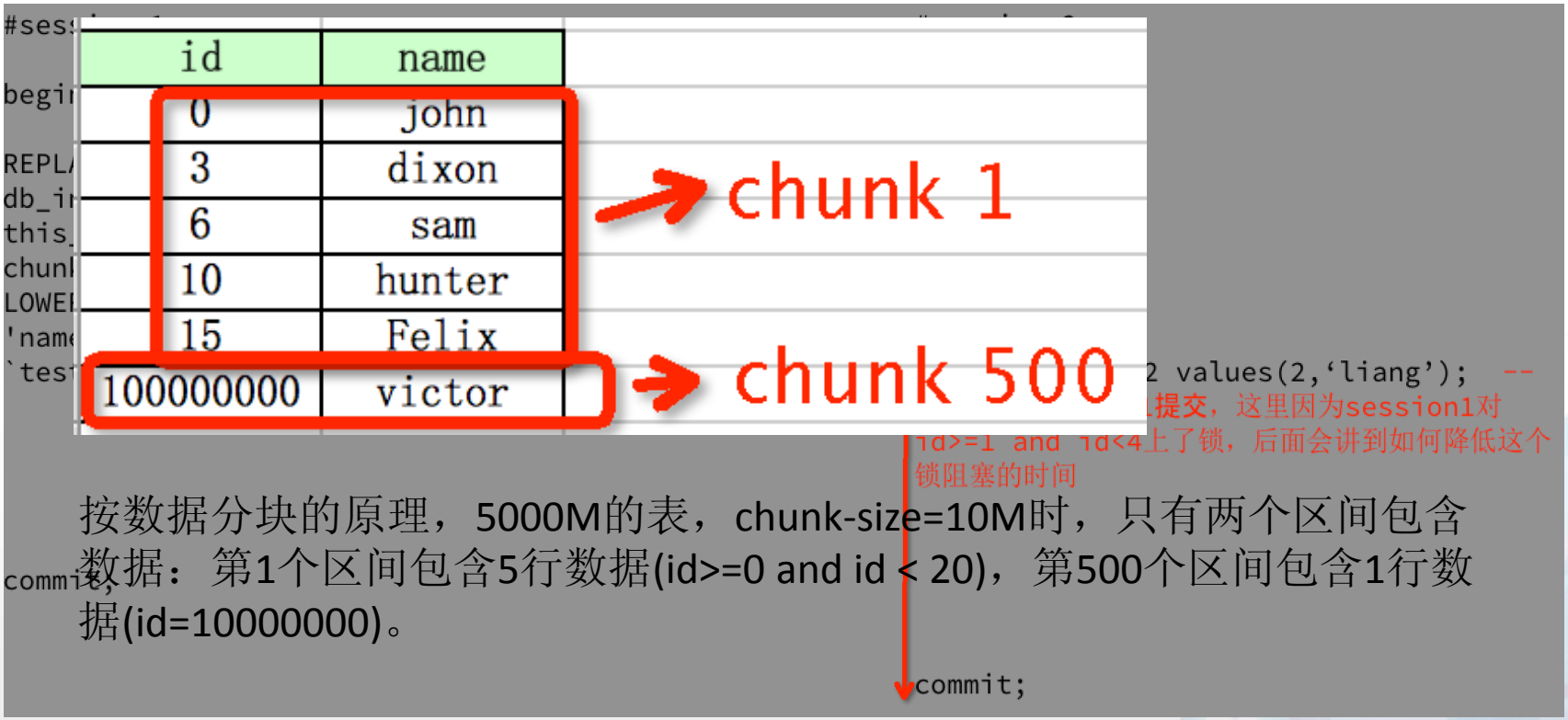
- 痛点2-3的应对思路
  - GCS 1.0 高可用技术
  - GCS 2.0 定制MySQL分支
    - 解决快速加字段问题
    - 解决大字段(blob/text)的压缩问题
    - tmysqlparse语法自动检测工具
  - GCS 3.0 存储云化
    - 解决CPU/MEM/IO的扩展性问题
    - 实现在线扩容及缩容
    - 透明分库分表

GCS 1.X – 3.X的演进之路 – GCS系统架构



GCS 1.X – 3.X的演进之路 – GCS1.0 高可用技术

- 数据切换保护及例行化checksum
  - chunk-size-exact，数据块切分不均在可重复度隔离级别下的“锁数据”问题

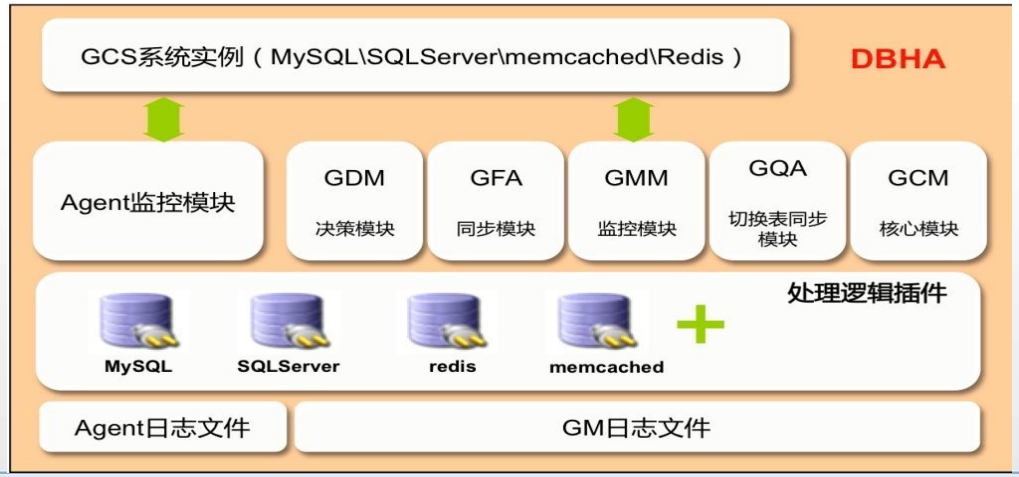


按数据分块的原理，5000M的表，chunk-size=10M时，只有两个区间包含数据：第1个区间包含5行数据(id>=0 and id < 20)，第500个区间包含1行数据(id=10000000)。

GCS 1.X – 3.X的演进之路 – GCS1.0 高可用技术

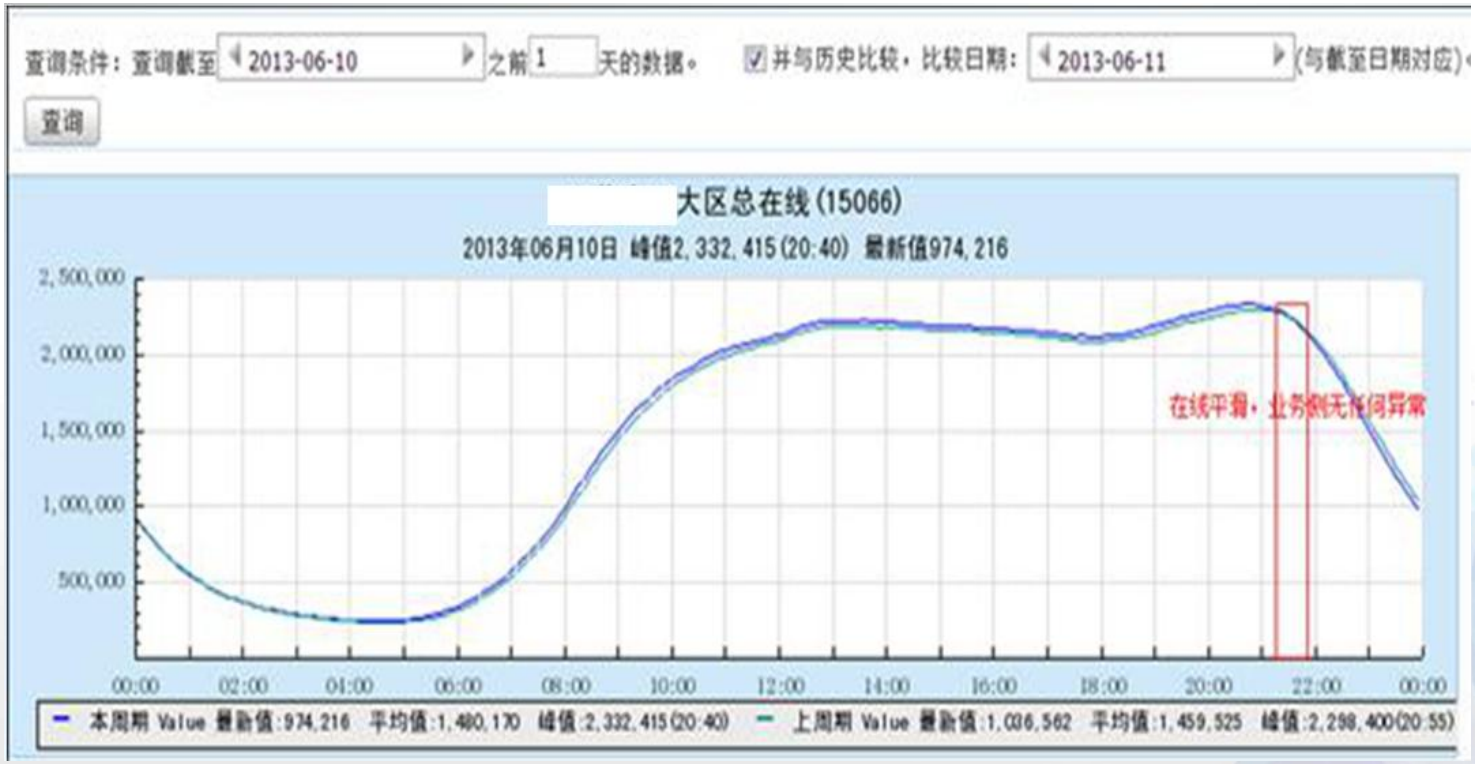
- mysql-proxy admin接口扩展
  - refresh_backends,refresh_users
  - show processlist,refresh_connlog

- 故障探测两段式仲裁及GM中控切换
  - 两个监测点同时认为故障checkmysql、checkssh
  - Double check | Slave Status、Checksum、Time Delay
  - 插件式支持MSSQL、Redis等存储介质



GCS 1.X – 3.X的演进之路 – GCS1.0 高可用技术

- 业务应用效果
  - 52%线上业务接入，涵盖多种类型端游及全部手游
  - 60S内，从故障发生到成功实施切换



GCS 1.X – 3.X的演进之路 – GCS2.0 MySQL分支定制

- TMySQL版本迭代 https://github.com/TencentDBA/TMySQL

版本	主要功能	详细描述	发布时间
TMySQL 1.1	在线加字段	秒级实现，支持MySQL分区表	2012/12/10
TMySQL 1.2	**内存分配优化** 核心BUG修复	Valgrind代码修改定位内存使用过多问题 深入剖析glibc内存碎片问题 **集成tcmalloc作为TMySQL内存管理模块** 修复5个重要mysql bugs，发现并定位15个mysql bugs	2013/3/20
TMySQL 1.3	(In-Place) Upgrade 安全性增强 运营特性增强 **备份、恢复增强**	支持MySQL5.0 → TMySQL的原地快速升级 增加TMySQL客户端程序名审计、密码二次加密 增加Alter Log日志记录 支持跨库表一致性备份 并行数据恢复加速（A5 60% Z3 90%缩短数据导入时间）	2013/6/6
TMySQL 1.4	**SqlParse工具开发** Binlog多线程导入 innodb字段压缩实现	集成语法、语义检查到OSS的变更子系统，提升业务变更效率 Binlog并发导入，缩短业务数据的回档时间 通过配置化的innodb底层字段压缩，提升mysql的cache利用率	2013/11/1

GCS 1.X – 3.X的演进之路 – GCS2.0 MySQL分支定制

- TMySQL在线加字段
  - 1秒以内完成加字段，后期性能损失2%-5%
  - 安装或者升级到TMySQL，并且alter table tbl row_format=GCS;

```
mysql> show table status like 'Operate'\G
*************************** 1. row ***************************
           Name: Operate
         Engine: InnoDB
        Version: 10
     Row_format: Gcs
           Rows: 474355247
 Avg_row_length: 95
    Data_length: 45519437824
Max_data_length: 0
   Index_length: 28483977216
      Data_free: 58720256
 Auto_increment: NULL
    Create_time: NULL
    Update_time: NULL
     Check_time: NULL
      Collation: utf8_general_ci
       Checksum: NULL
 Create_options: partitioned
        Comment: 操作日志
1 row in set (2.95 sec)
```

数据量总大小约74G

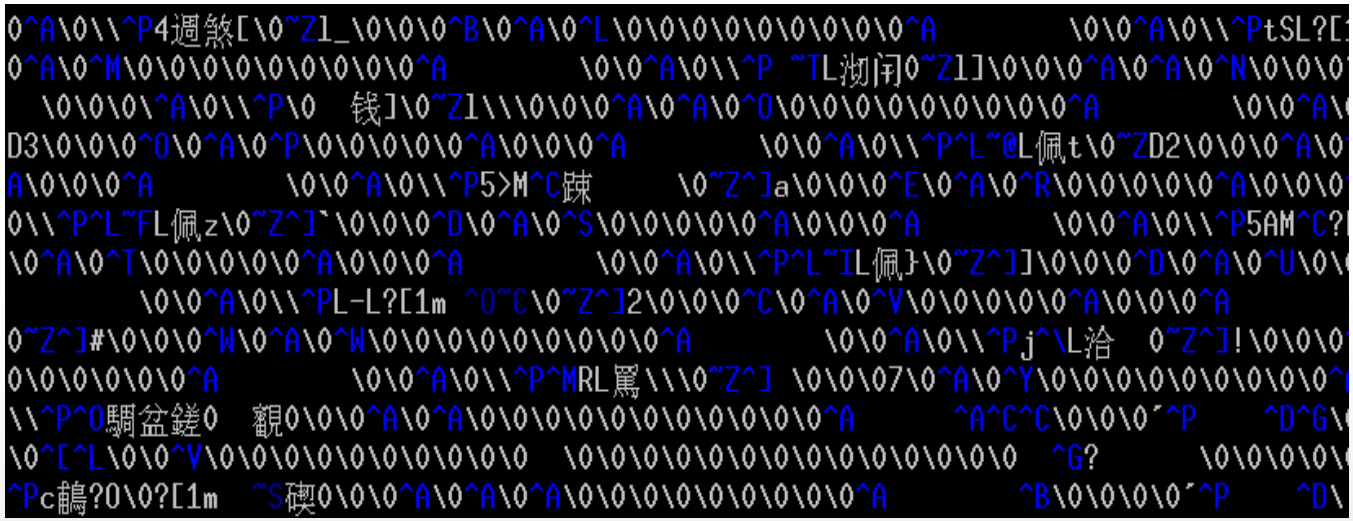
```
mysql> alter table Operate add column (iSource_1 int, iSource_2 int not null default
0);
Query OK, 0 rows affected (0.04 sec)
Records: 0  Duplicates: 0  Warnings: 0

mysql> select iSource_1,iSource_2 from Operate limit 10;
+----------+----------+
| iSource_1 | iSource_2 |
+----------+----------+
|     NULL |         0 |
|     NULL |         0 |
|     NULL |         0 |
|     NULL |         0 |
|     NULL |         0 |
|     NULL |         0 |
|     NULL |         0 |
|     NULL |         0 |
|     NULL |         0 |
|     NULL |         0 |
+----------+----------+
10 rows in set (0.00 sec)
```

仅需0.04秒完成一个74G表的加字段操作

GCS 1.X – 3.X的演进之路 – GCS2.0 MySQL分支定制

- TMySQL Innodb blob/text列压缩 背景
  - 结构体序列化存储
  - 较多C/C++ NULL占位符，序列化 ≠ 压缩
  - DBA推动研发改动几行代码困难



GCS 1.X – 3.X的演进之路 – GCS2.0 MySQL分支定制

- TMySQL Innodb blob/text列压缩 使用及效果
  - 创建表

```
Create table t1 (
  C1 int primary key,
  C2 blob compressed,
  C3 text character set gbk compressed,
  C4 blob
 ) engine = innodb row_format=GCS
```

  - 修改表

```
Alter table t1 change c4 c4 blob compressed.
```

某业务数据，压缩前51G，压缩后7.3G，压缩率达14.3%

GCS 1.X – 3.X的演进之路 – GCS2.0 MySQL分支定制

- TMySQL Innodb blob/text列压缩 性能对比

对比纬度	数据不压缩	row_format=compressed	BLOB列压缩
数据量	51G	24G	7.1G
QPS	1174	1524	3994
IO	100%	100%	30%
CPU	15%	45%	50%

利用空闲的CPU计算能力换取IO能力的提升！

GCS 1.X – 3.X的演进之路 – GCS2.0 MySQL分支定制

- TMySQL Tmysqlparse语法自动检测工具

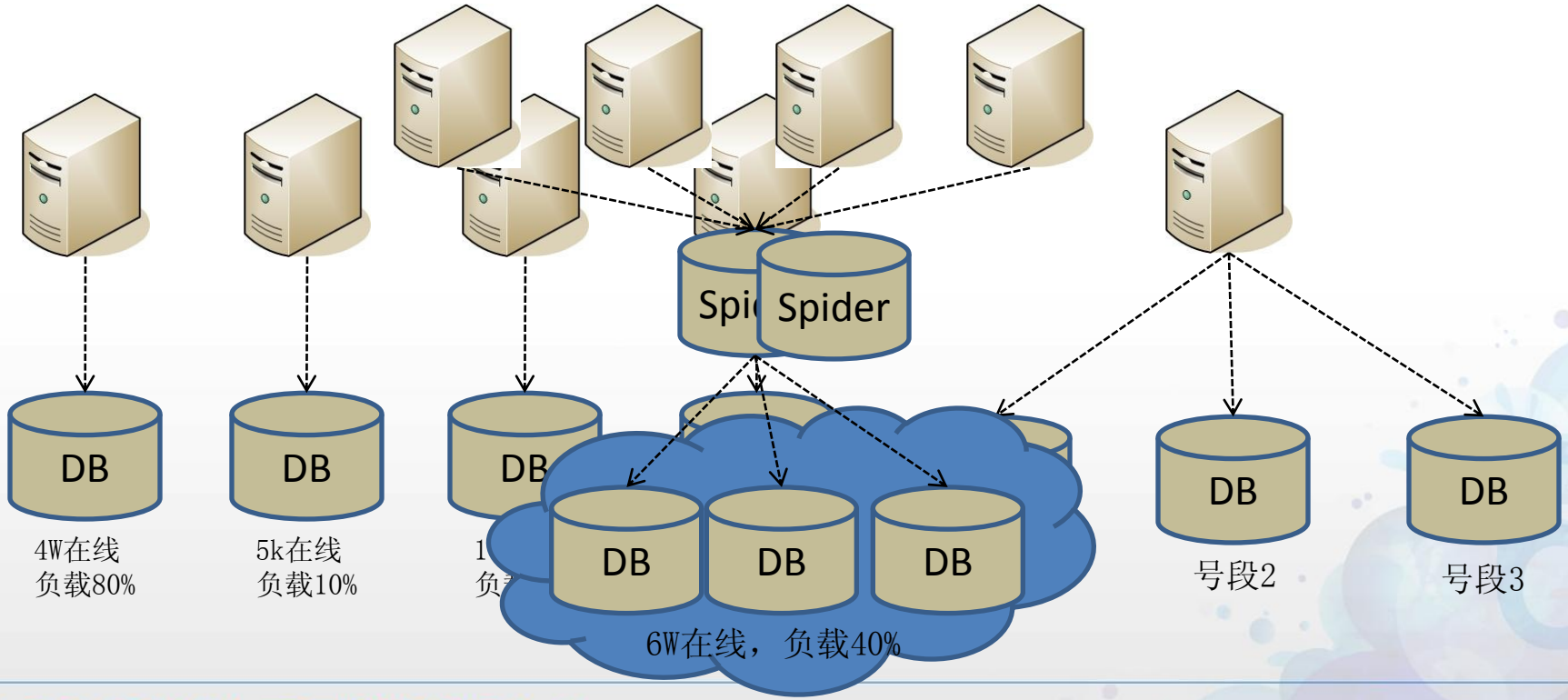
表 1 2013-12-01 至 2014-05-01 语法检测工具检测效果

单据类型	单据数	语法错误数	告警数	错误比
Pkg	2341	96	1221	4%
ddl	261	20	40	7.7%
dml	175	4	51	2.3%
Pkg+ddl+dml	2777	120	1312	4.3%

自 2013-12-01 至 2014-05-01 半年时候，tmysqlparse 总计在 2777 个提单中检测出 120 例语法错误，平均每天大概 1.5 个语法错误的单据被提前发现。

GCS 1.X – 3.X的演进之路 – GCS3.0 存储层云化

- 透明DB中间件-Spider
- CPU/MEM/IO扩展
- 透明分库分表，应用无关性

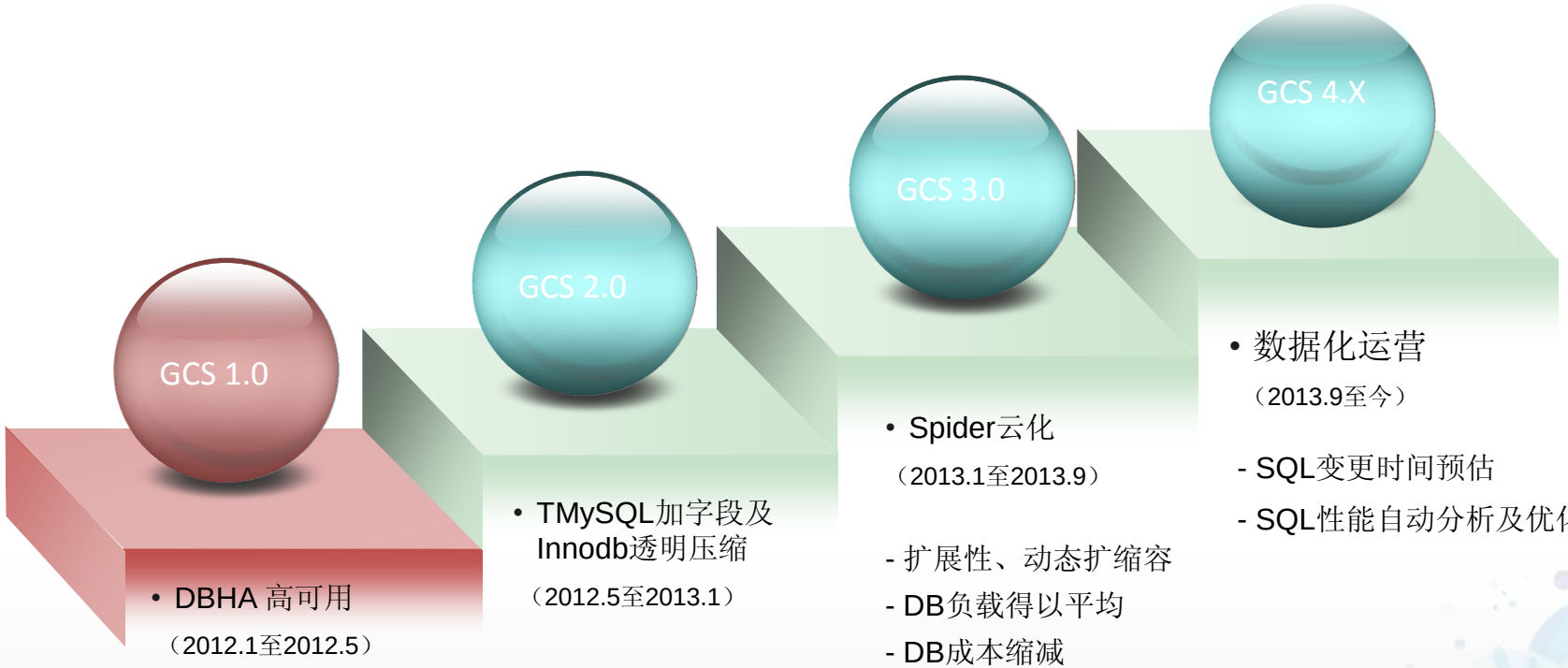


GCS4.X规划

- 从 功能 到 性能
  - TMySQL迁移到MySQL 5.6
  - 取消MySQL的锁粒度

- 数据化运营
  - 数据库优化标准化进而形成有竞争力的产品或服务
  - SQL变更时间自动化预估
    - 需整合现网备份时间数据、实例Schema数据、表信息数据等

GCS的迭代路径



- DBHA 高可用
（2012.1至2012.5）

- mysql-proxy定制
- DNS集群
- 两段式监测点探测及GM切换
- checksum例行化
- 支持Redis及MSSQL

- TMySQL加字段及Innodb透明压缩
（2012.5至2013.1）

- 在线加字段支持分区表
- Innodb大字段透明压缩
- 内存性能问题分析及tcmalloc引入
- 语法自动检测工具

- Spider云化
（2013.1至2013.9）

- 扩展性、动态扩缩容
- DB负载得以平均
- DB成本缩减

- 数据化运营
（2013.9至今）

- SQL变更时间预估
- SQL性能自动分析及优化

Q&A

THANKS

SequeMedia 盛拓传媒 | IT168 www.it168.com | ChinaUnix | ITPUB

附录1：TMySQL在线加字段与业界的对比

在线加字段方案	优势	劣势	谁在使用
Facebook OSC	支持更多类型的DDL，外围实现	触发器实现，性能至少损失20% 对负载高的DB，加字段完成时间不可控 外围管理成本较高	Facebook、新浪、淘宝
MySQL 5.6 DDL Online	支持更多类型的DDL，底层实现	数据需要拷贝，磁盘压力大 GA不足半年，不建议使用	
TMySQL	新增行格式，无需数据拷贝， 只需修改数据字典，立即生效	不是通用的DDL Online方案，但未来会集成到5.6的MySQL版本	腾讯互娱

TMySQL字段扩展达到商业数据库Oracle 11g，MSSQL 2012的能力！

附录2：Spider与业界的对比

云化存储方案	可扩展性	兼容性	成熟度
SPIDER	优 接入层、存储层可自由扩展	良 应用层透明，支持大部分SQL，但不宜过于复杂，事务支持程度有限。 对mysql版本无要求。	良 未release，但已通过基本的压测，待解决问题已基本明确
CDB+CBS	中 存储层(TSSD)可扩展，但CDB本身会成为瓶颈	良 与普通mysql没有差别，理论上支持任意SQL及事务。 仅支持CDB订制的mysql版本	优 已在生产环境中使用
Fabric	良 接入层、存储层可自由扩展，但存在中央节点	中 应用层需要特定访问接口，mysql需要5.6+	中 未release
自制proxy	优 与spider一致	中 支持SQL有限，需要开发支持	差 需重新开发