

ArchSummit

全球架构师峰会（深圳）2014

构建大型云计算平台分布式技术的实践

章文嵩（正明）

ArchSummit · 深圳

2014.7.18

自我介绍

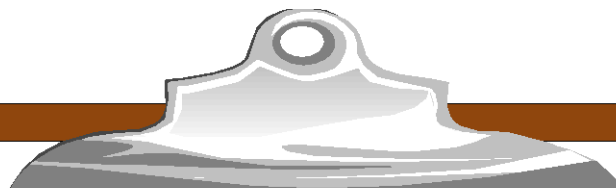
- 章文嵩（正明） 博士
- 阿里高级研究员、核心系统负责人
- LVS开源项目的创始人与主要作者
- 曾为TelTel的首席科学家与联合创始人，国防科技大学副教授、ChinaCluster的联合创始人、Red Hat Kernel Developer



或在来往中搜索wensongzhang加我

议程

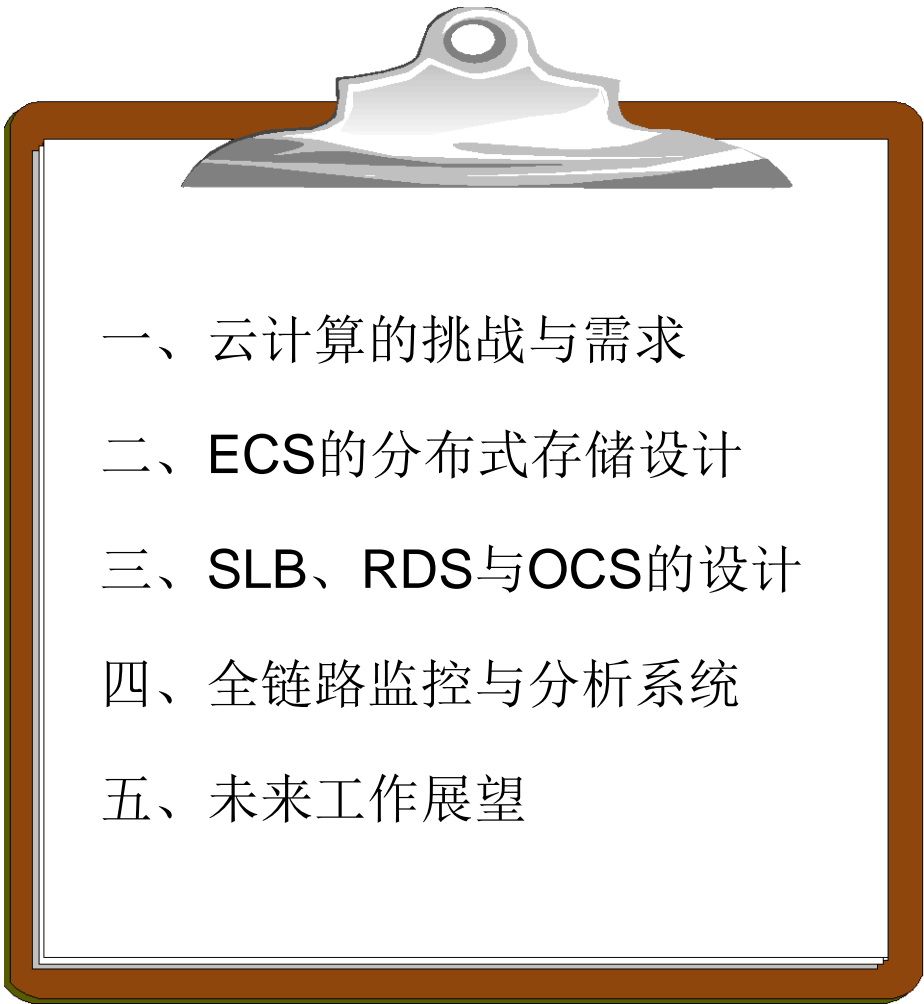


- 
- 一、云计算的挑战与需求
 - 二、ECS的分布式存储设计
 - 三、SLB、RDS与OCS的设计
 - 四、全链路监控与分析系统
 - 五、未来工作展望

- 云计算的挑战
 - 淘宝天猫应用需求 vs 中小网站需求
 - 客户把他们关键的IT系统托付在云平台上
- 对云计算平台的需求
 - 高可靠性
 - 高性能
 - 快速定位问题
 - 安全
 - 低成本

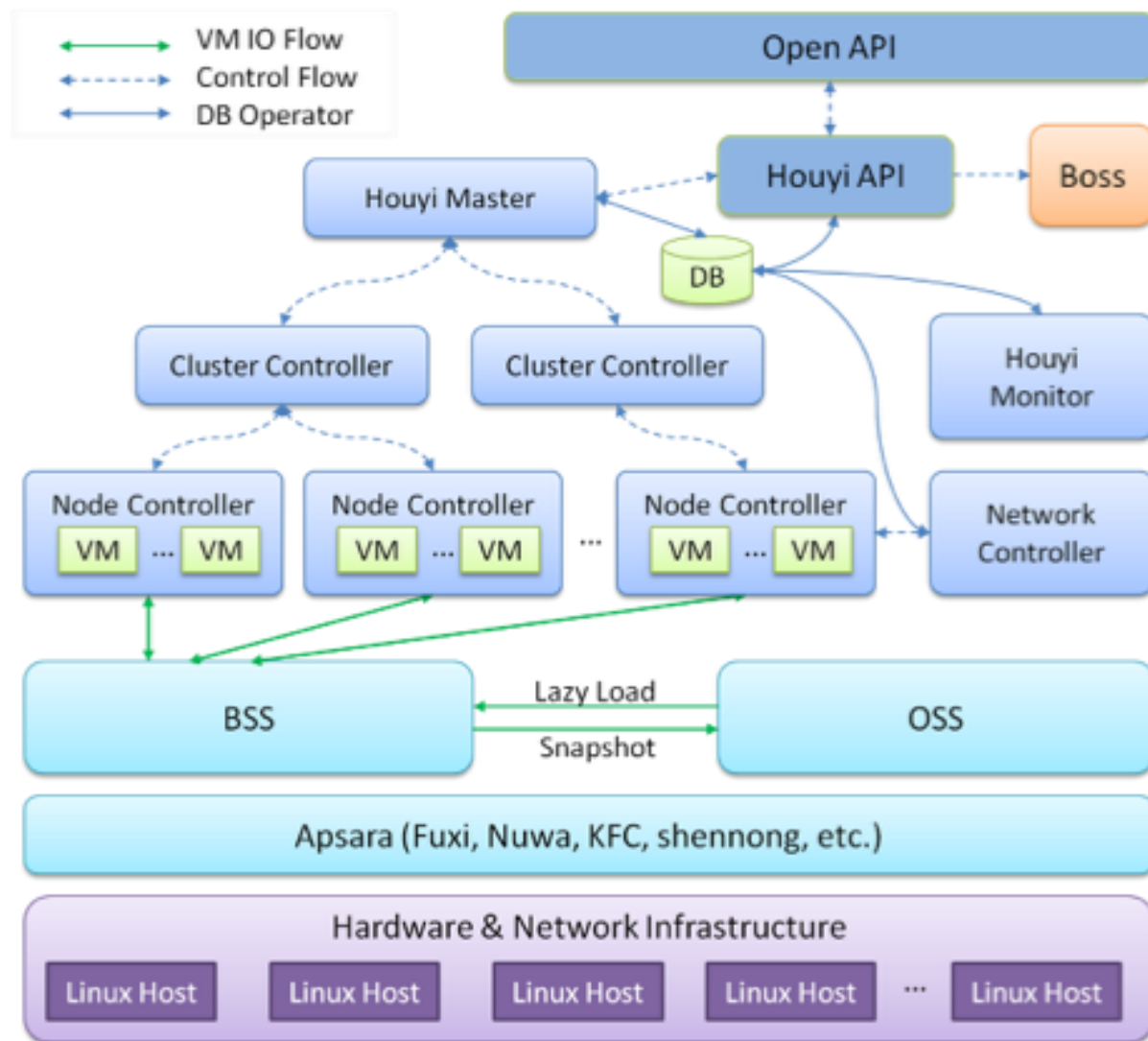
议程



- 
- 一、云计算的挑战与需求
 - 二、ECS的分布式存储设计
 - 三、SLB、RDS与OCS的设计
 - 四、全链路监控与分析系统
 - 五、未来工作展望

云服务器ECS

- 分布式文件存储
- 快照制作
- 快照回滚
- 自定义image
- 故障迁移
- 在线迁移
- 网络组隔离
- 防ARP欺骗
- 自定义防火墙功能
- 支持防DDoS攻击
- 提供流量清洗服务
- 动态升级

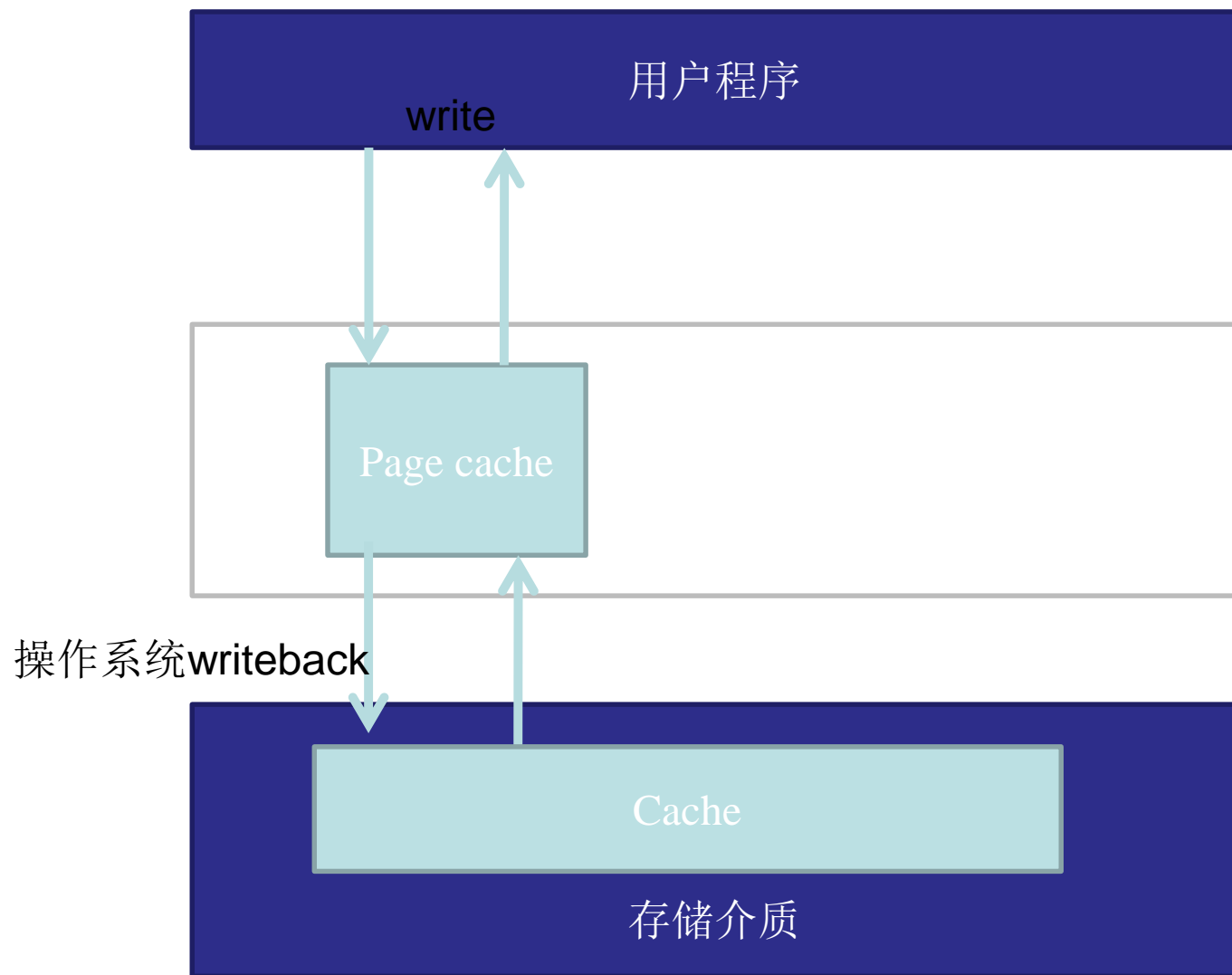


- 为确保数据的高可靠性，实现2-3异步
- 过去的问题
 - 对于任何写，都写入到Chunk Server才算成功，路径长，延时大。实现复杂开销大等。
- 优化思路
 - SSD/SATA混合存储，randwrite-4K-128可达5500 IOPS左右
 - 引入Cache机制，同时实现一样的数据可靠性
 - 多线程事件驱动架构重构TDC和Chunk Server的实现，让一个IO请求在一个线程完成所有工作，避免锁和上下文切换

IO路径上的各层cache

- 应用程序的user cache
 - mysql buffer pool
- 操作系统的缓存
 - linux page cache
- 存储系统的cache
 - 磁盘的缓存

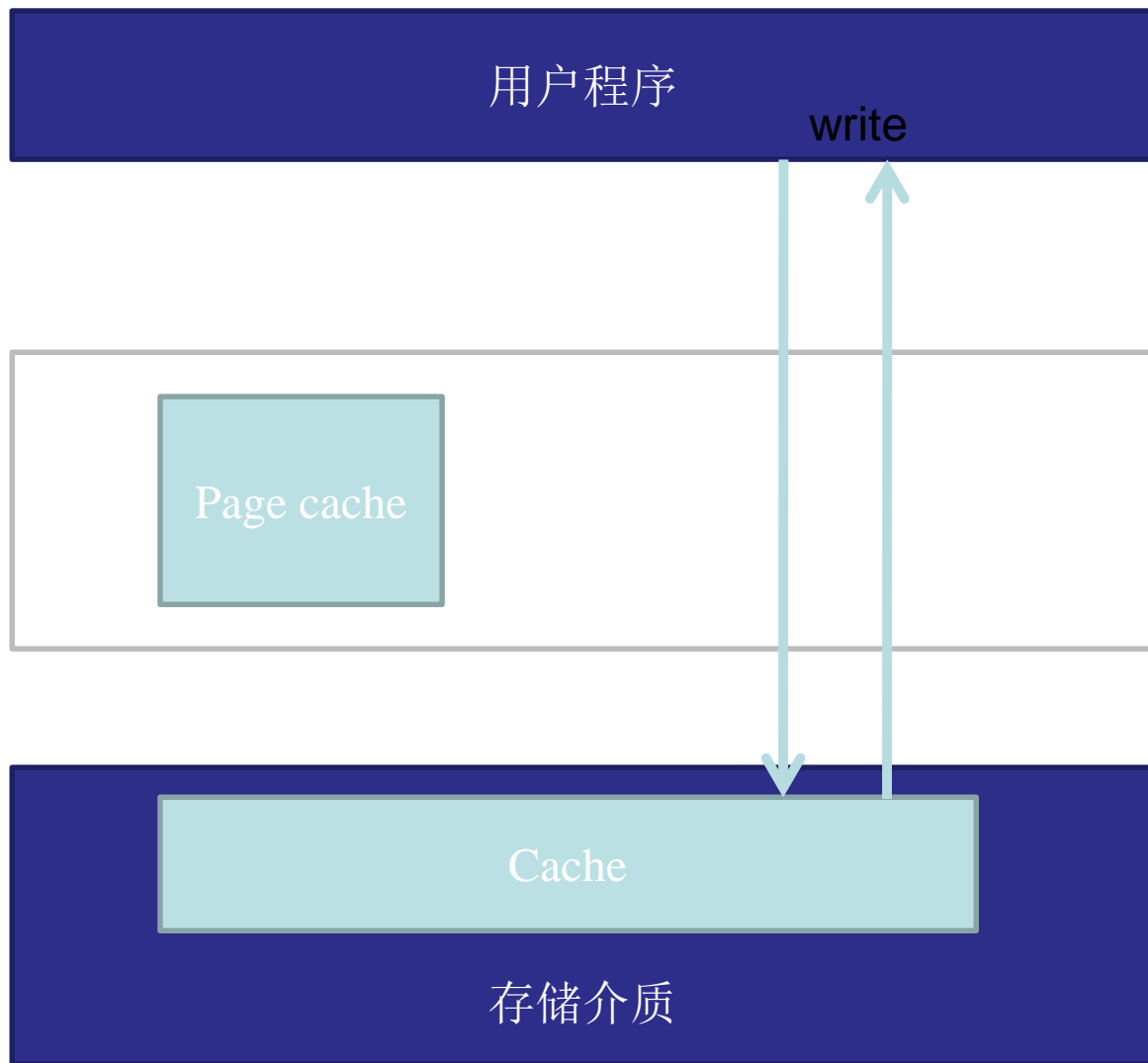
写IO的几种模式 – buffer write



写IO的几种模式 – buffer write

- Buffer write, 应用程序写入到操作系统page cache, 等待操作系统回写到存储介质
 - 优点: 大部分情况下直接写内存, 速度很快
 - 缺点1: 数据完整性无法得到严密保证, 受到操作系统回写到介质的时间和频率影响
 - 缺点2: 小部分写入由于受到系统回写影响会有阻塞, 服务质量没有办法保障

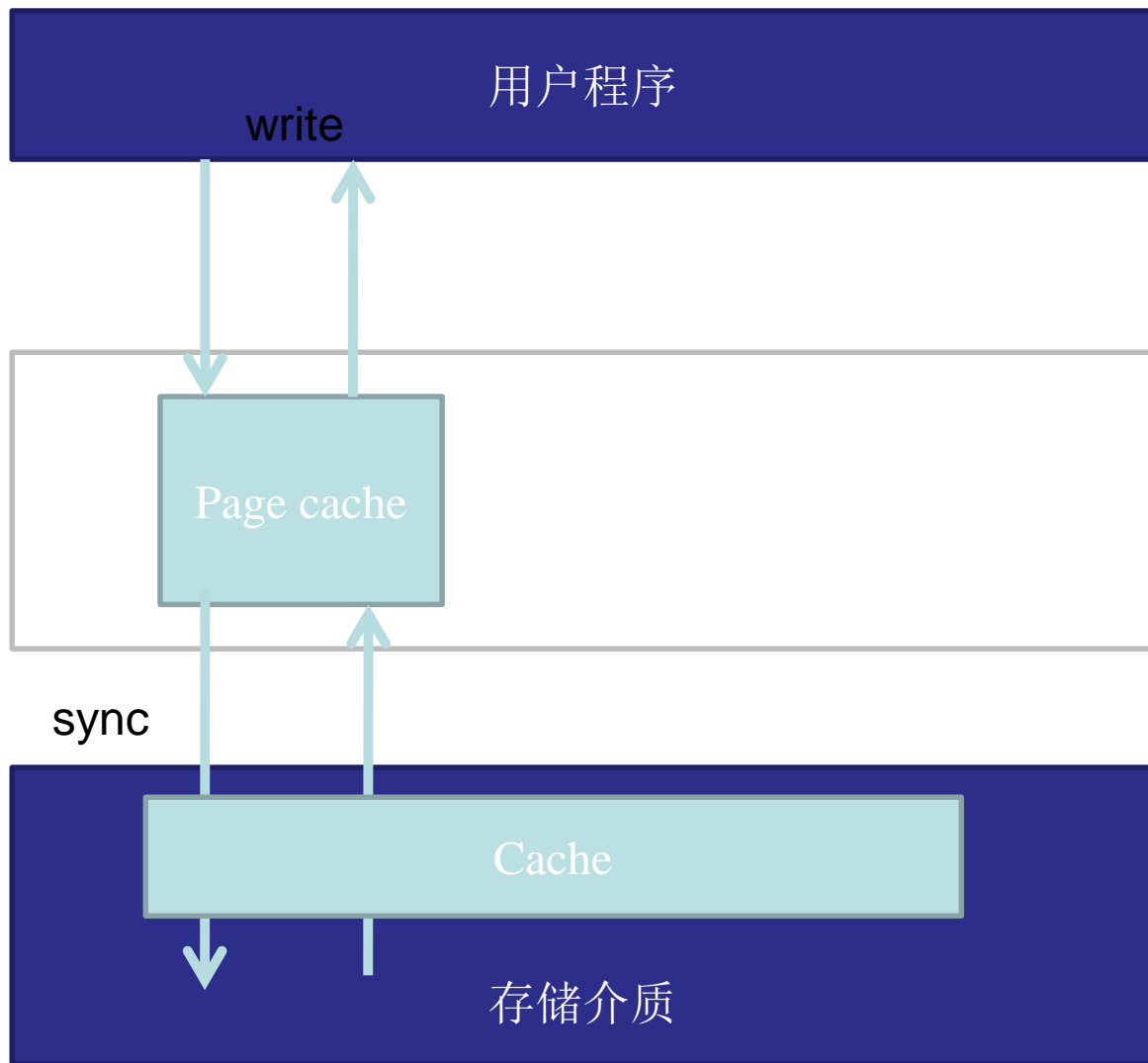
写IO的几种模式 – direct write



写IO的几种模式 – direct write

- Direct write, 应用程序绕过操作系统的page cache, 直接写存储介质, 用户写完介质就返回
- 优点: 规避了page cache的使用, 不受操作系统回写的影响, 安全性考虑稍强
- 缺点: 数据安全性操作系统本身并不保证, 且介质有可能本身有cache, 不能做到绝对安全
- https://ext4.wiki.kernel.org/index.php/Clarifying_Direct_IO%27s_Semantics

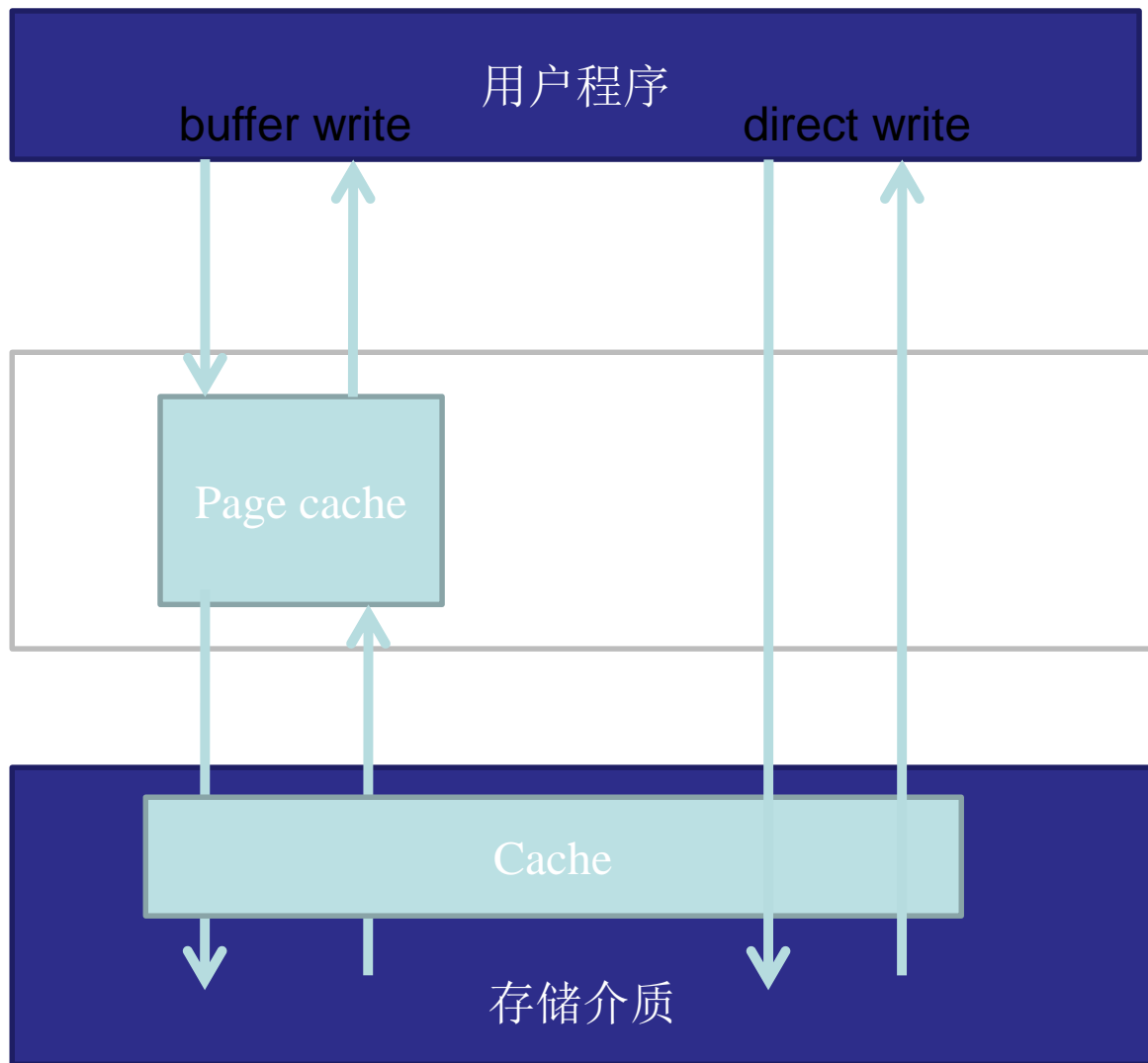
写IO的几种模式 –write+sync



写IO的几种模式 –write+sync

- 写入数据后调用sync/fsync
- 优点： sync返回后数据已经成功写入磁盘介质并足够安全
- 缺点1： 在调用sync前写入的数据有可能丢失
- 缺点2： 随着操作系统内存的使用情况不同， sync等待的时间也会不同
 - <http://www.google.com.hk/#newwindow=1&q=linux+sync+hang&safe=strict>

写IO的几种模式 -O_SYNC

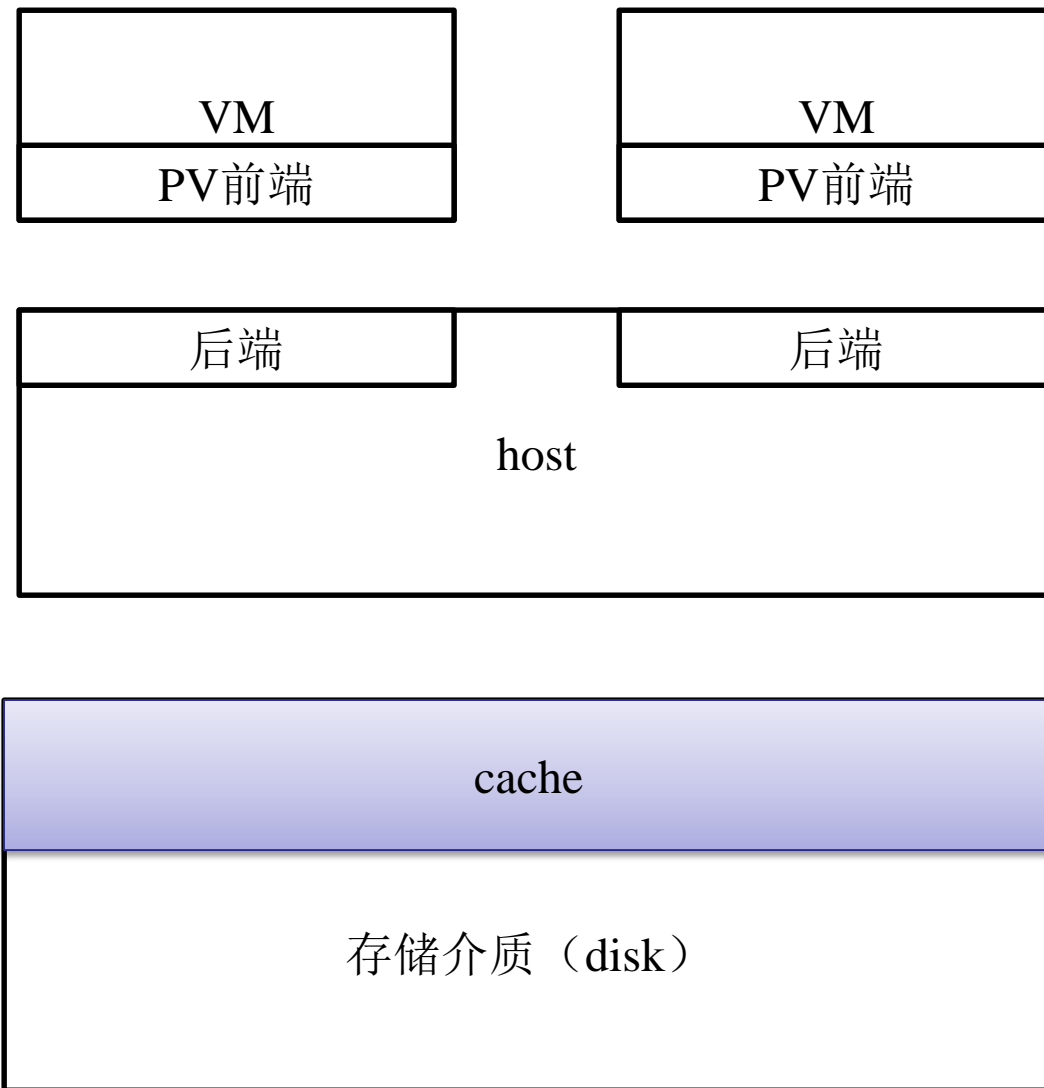


数据安全背后的故事

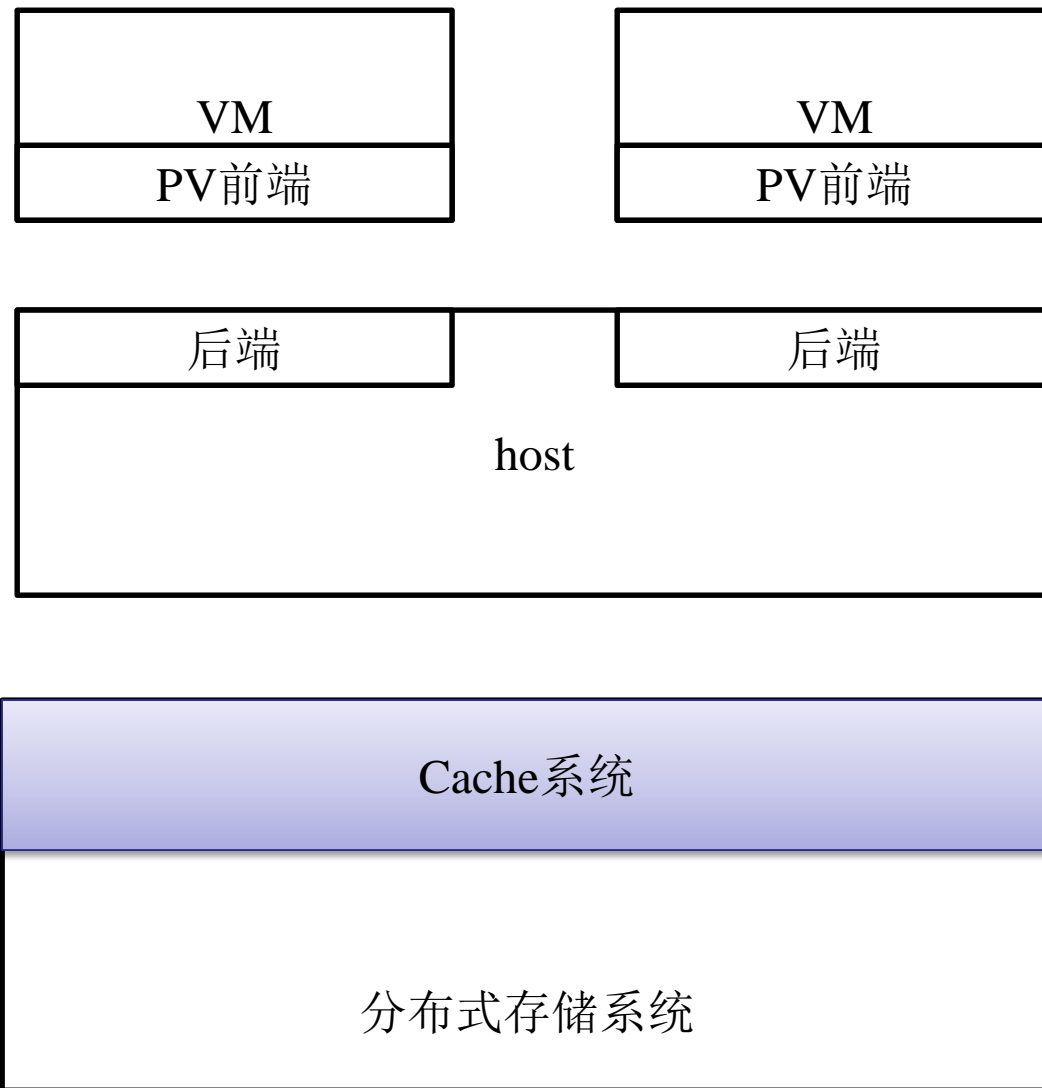
- 用户 sync 或者 `O_SYNC`
- 操作系统将数据写入存储
 - direct write 或者 page cache 回写
- 操作系统指示存储介质将数据写入非易失介质
 - flush 或者 write barrier

```
[t@test]dd if=/dev/zero of=test.img bs=4k count=2000
2000+0 records in
2000+0 records out
8192000 bytes (8.2 MB) copied, 0.038712 s, 212 MB/s
[t@test]dd if=/dev/zero of=test.img bs=4k count=2000 oflag=direct
2000+0 records in
2000+0 records out
8192000 bytes (8.2 MB) copied, 0.119936 s, 68.3 MB/s
[t@test]dd if=/dev/zero of=test.img bs=4k count=2000 oflag=direct, sync
2000+0 records in
2000+0 records out
8192000 bytes (8.2 MB) copied, 31.86 s, 257 kB/s
[t@test]
```

虚拟化中本地磁盘的IO

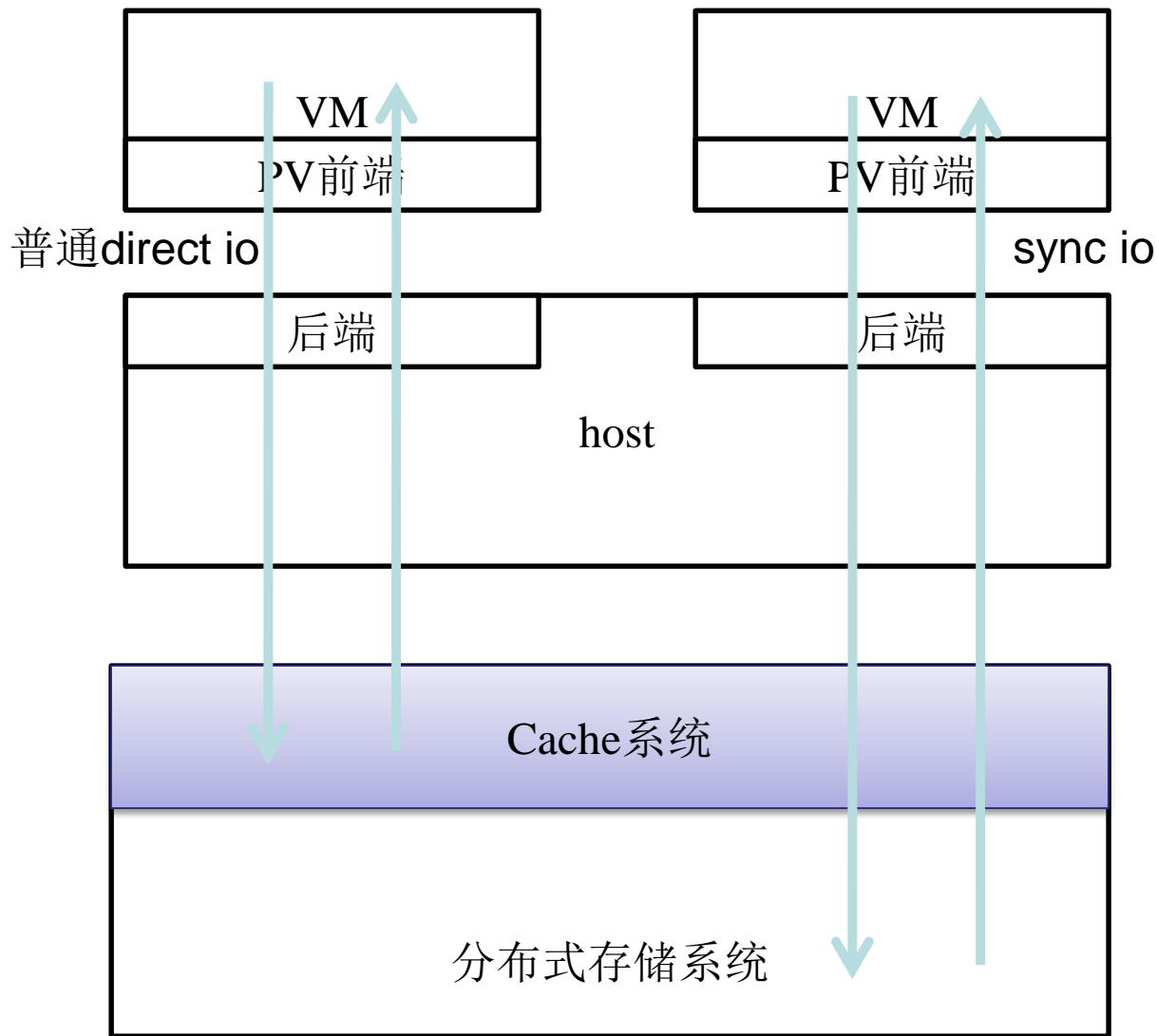


云计算环境中的IO



- 作用1：像磁盘cache一样，加速数据完整性没有要求的写请求
- Q：如何保证VM中应用程序的数据完整性，保证CACHE系统数据符合程序的预期？
- A：保证VM数据完整性语义透传IO全链路
- VM SYNC->PV前端FLUSH->后端->host->cache系统->分布式存储系统

云计算环境中的IO



random write test (direct)

- `./fio -direct=1 -iodepth=1 -rw=randwrite -ioengine=libaio -bs=16k -numjobs=2 -runtime=30 -group_reporting -size=30G -name=/mnt/test30G`

SATA分布式存储

```
6k -numjobs=2 -runtime=30 -group_reporting -size=30G -name=/mnt/test30G
/mnt/test30G: (g=0): rw=randwrite, bs=16K-16K/16K-16K/16K-16K, ioengine=libaio, iodepth=1
/mnt/test30G: (g=0): rw=randwrite, bs=16K-16K/16K-16K/16K-16K, ioengine=libaio, iodepth=1
fio-2.0.14
Starting 2 processes
Jobs: 2 (f=2): [ww] [100.0% done] [0K/4753K/0K /s] [0 /297 /0 iops] [eta 00m:00s]
/mnt/test30G: (groupid=0, jobs=2): err= 0: pid=21868: Mon Jul 14 14:43:05 2014
write: io=109136KB, bw=3636.5KB/s, iops=227, runt= 30015msec
  slat (usec): min=8, max=1620, avg=12.63, stdev=25.82
  clat (msec): min=1, max=113, avg= 8.78, stdev= 7.53
    lat (msec): min=1, max=113, avg= 8.79, stdev= 7.53
  clat percentiles (usec):
    | 1.00th=[ 1720], 5.00th=[ 2704], 10.00th=[ 2768], 20.00th=[ 2864],
    | 30.00th=[ 2960], 40.00th=[ 3088], 50.00th=[ 3504], 60.00th=[12480],
    | 70.00th=[13888], 80.00th=[15168], 90.00th=[16512], 95.00th=[21888],
    | 99.00th=[27008], 99.50th=[28032], 99.90th=[88576], 99.95th=[91648],
    | 99.99th=[114176]
  bw (KB/s) : min= 792, max= 2514, per=49.94%, avg=1815.85, stdev=569.80
  lat (msec) : 2=2.39%, 4=50.48%, 10=4.84%, 20=36.36%, 50=5.79%
  lat (msec) : 100=0.13%, 250=0.01%
cpu          : usr=0.09%, sys=0.13%, ctx=6823, majf=0, minf=39
IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
submit      : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
complete    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
issued     : total=r=0/w=6821/d=0, short=r=0/w=0/d=0

Run status group 0 (all jobs):
WRITE: io=109136KB, aggrb=3636KB/s, minb=3636KB/s, maxb=3636KB/s, mint=30015msec, maxt=30015msec

Disk stats (read/write):
xvdb: ios=0/6827, merge=0/5, ticks=0/60396, in queue=60406, util=99.83%
```


SATA分布式存储+cache

```
/mnt/test30G: (g=0): rw=randwrite, bs=16K-16K/16K-16K/16K-16K, ioengine=libaio, iodepth=1
/mnt/test30G: (g=0): rw=randwrite, bs=16K-16K/16K-16K/16K-16K, ioengine=libaio, iodepth=1
fio-2.0.14
```

Starting 2 processes

```
Jobs: 2 (f=2): [ww] [100.0% done] [0K/10322K/0K /s] [0 /645 /0 iops] [eta 00m:00s]
```

```
/mnt/test30G: (groupid=0, jobs=2): err= 0: pid=21885: Mon Jul 14 14:50:02 2014
```

```
write: io=292448KB, bw=9745.2KB/s, iops=609 , runt= 30010msec
```

```
slat (usec): min=10 , max=32805 , avg=56.94, stdev=774.44
```

```
clat (msec): min=1 , max=106 , avg= 3.22, stdev= 2.41
```

```
lat (msec): min=1 , max=106 , avg= 3.28, stdev= 2.56
```

```
clat percentiles (usec):
```

```
| 1.00th=[ 1608], 5.00th=[ 2192], 10.00th=[ 2704], 20.00th=[ 2800],
| 30.00th=[ 2864], 40.00th=[ 2928], 50.00th=[ 3024], 60.00th=[ 3120],
| 70.00th=[ 3312], 80.00th=[ 3568], 90.00th=[ 3952], 95.00th=[ 4320],
| 99.00th=[ 5024], 99.50th=[ 5472], 99.90th=[13888], 99.95th=[98816],
| 99.99th=[107008]
```

```
bw (KB/s) : min= 1912, max= 5280, per=50.38%, avg=4909.38, stdev=479.39
```

```
lat (msec) : 2=4.27%, 4=86.56%, 10=9.00%, 20=0.10%, 50=0.01%
```

```
lat (msec) : 100=0.01%, 250=0.04%
```

```
cpu : usr=0.31%, sys=0.27%, ctx=18355, majf=0, minf=41
```

```
IO depths : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
```

```
submit : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
```

```
complete : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
```

```
issued : total=r=0/w=18278/d=0, short=r=0/w=0/d=0
```

Run status group 0 (all jobs):

```
WRITE: io=292448KB, aggrbw=9745KB/s, minbw=9745KB/s, maxbw=9745KB/s, mint=30010msec, maxt=30010msec
```

direct test iodepth=8

- `./fio -direct=1 -iodepth=8 -rw=randwrite -ioengine=libaio -bs=16k -numjobs=2 -runtime=30 -group_reporting -size=30G -name=/mnt/test30G`

SATA分布式存储

```
xvdb: ios=871/53936, merge=0/7, ticks=14284/386735, in_queue=401102, util=99.85%
6k -numjobs=2 -runtime=30 -group_reporting -size=30G -name=/mnt/test30Gaio -bs=16
/mnt/test30G: (g=0): rw=randwrite, bs=16K-16K/16K-16K/16K-16K, ioengine=libaio, iodepth=8
/mnt/test30G: (g=0): rw=randwrite, bs=16K-16K/16K-16K/16K-16K, ioengine=libaio, iodepth=8
fio-2.0.14
Starting 2 processes
Jobs: 2 (f=2): [ww] [100.0% done] [0K/28015K/0K /s] [0 /1750 /0 iops] [eta 00m:00s]
/mnt/test30G: (groupid=0, jobs=2): err= 0: pid=21963: Mon Jul 14 18:15:36 2014
write: io=992.0MB, bw=33847KB/s, iops=2115 , runt= 30012msec
  slat (usec): min=6 , max=95592 , avg=57.66, stdev=1182.15
  clat (msec): min=1 , max=214 , avg= 7.50, stdev= 7.04
  lat (msec): min=1 , max=220 , avg= 7.56, stdev= 7.16
  clat percentiles (msec):
    | 1.00th=[ 3], 5.00th=[ 4], 10.00th=[ 5], 20.00th=[ 5],
    | 30.00th=[ 5], 40.00th=[ 6], 50.00th=[ 6], 60.00th=[ 7],
    | 70.00th=[ 7], 80.00th=[ 10], 90.00th=[ 15], 95.00th=[ 16],
    | 99.00th=[ 19], 99.50th=[ 28], 99.90th=[ 111], 99.95th=[ 176],
    | 99.99th=[ 206]
  bw (KB/s) : min= 8562, max=18784, per=50.04%, avg=16935.86, stdev=2081.44
  lat (msec) : 2=0.05%, 4=7.07%, 10=73.41%, 20=18.71%, 50=0.49%
  lat (msec) : 100=0.15%, 250=0.13%
cpu           : usr=0.84%, sys=1.04%, ctx=62313, majf=0, minf=39
IO depths     : 1=0.1%, 2=0.1%, 4=0.1%, 8=100.0%, 16=0.0%, 32=0.0%, >=64=0.0%
submit       : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
complete    : 0=0.0%, 4=100.0%, 8=0.1%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
issued      : total=r=0/w=63488/d=0, short=r=0/w=0/d=0

Run status group 0 (all jobs):
WRITE: io=992.0MB, aggrbw=33846KB/s, minb=33846KB/s, maxb=33846KB/s, mint=30012msec, maxt=30012msec

Disk stats (read/write):
xvdb: ios=182/63529, merge=0/5, ticks=2830/460781, in_queue=463681, util=99.85%
```


SATA分布式存储+cache

```
6k -numjobs=2 -runtime=30 -group_reporting -size=30G -name=/mnt/test30Gaio -bs=16
/mnt/test30G: (g=0): rw=randwrite, bs=16K-16K/16K-16K/16K-16K, ioengine=libaio, iodepth=8
/mnt/test30G: (g=0): rw=randwrite, bs=16K-16K/16K-16K/16K-16K, ioengine=libaio, iodepth=8
fio-2.0.14
Starting 2 processes
Jobs: 2 (f=2): [ww] [100.0% done] [0K/45035K/0K /s] [0 /2814 /0 iops] [eta 00m:00s]
/mnt/test30G: (groupid=0, jobs=2): err= 0: pid=21985: Mon Jul 14 18:25:03 2014
write: io=1361.8MB, bw=46452KB/s, iops=2903, runt= 30004msec
  slat (usec): min=7, max=62898, avg=18.42, stdev=382.54
  clat (msec): min=1, max=63, avg= 5.49, stdev= 1.59
  lat (msec): min=1, max=66, avg= 5.51, stdev= 1.64
  clat percentiles (usec):
    | 1.00th=[ 3024], 5.00th=[ 3664], 10.00th=[ 4016], 20.00th=[ 4448],
    | 30.00th=[ 4704], 40.00th=[ 4960], 50.00th=[ 5216], 60.00th=[ 5536],
    | 70.00th=[ 5920], 80.00th=[ 6368], 90.00th=[ 7200], 95.00th=[ 8032],
    | 99.00th=[ 9792], 99.50th=[10688], 99.90th=[17280], 99.95th=[24448],
    | 99.99th=[43264]
  bw (KB/s) : min=20812, max=26944, per=50.06%, avg=23252.43, stdev=1300.31
  lat (msec) : 2=0.01%, 4=9.63%, 10=89.50%, 20=0.80%, 50=0.06%
  lat (msec) : 100=0.01%
cpu          : usr=1.16%, sys=1.34%, ctx=86611, majf=0, minf=41
IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=100.0%, 16=0.0%, 32=0.0%, >=64=0.0%
submit      : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
complete    : 0=0.0%, 4=100.0%, 8=0.1%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
issued      : total=r=0/w=87109/d=0, short=r=0/w=0/d=0

Run status group 0 (all jobs):
WRITE: io=1361.8MB, aggrbw=46451KB/s, minbw=46451KB/s, maxbw=46451KB/s, mint=30004msec, maxt=30004msec

Disk stats (read/write):
xvdb: ios=28/87212, merge=0/5, ticks=471/474005, in_queue=474519, util=99.84%
```

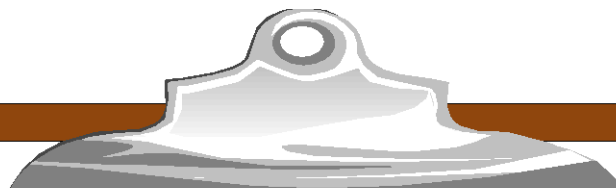
- 作用2：降低分布式存储系统的抖动对上层应用的影响
- Jeff Dean: The Tail at Scale, CACM Feb. 2013
- 如果有1%的概率请求延迟超过1S，并发100个请求，然后等待所有请求返回，延时超过1S的概率为63%

ECS的不同存储选择

- 纯SATA存储集群
 - 高可靠，IOPS能力适中，适合大部分应用
- 混合存储（SATA/SSD）集群
 - 高可靠，高IOPS
- 纯SSD存储集群（预计11月/12月正式推出）
 - 高可靠，超高IOPS
 - 4K随机写，物理机，chunk server 18万IOPS，稳定16.5万
 - TDC做到9万IOPS，万兆用满，消耗6颗HT CPU（需优化）
- SATA本地磁盘（单份）
 - 可靠性低，IOPS低，适合Hadoop/HBase（批量）等
- SSD本地磁盘（单份）
 - 可靠性低，IOPS高，适合Mongodb/HBase（在线）

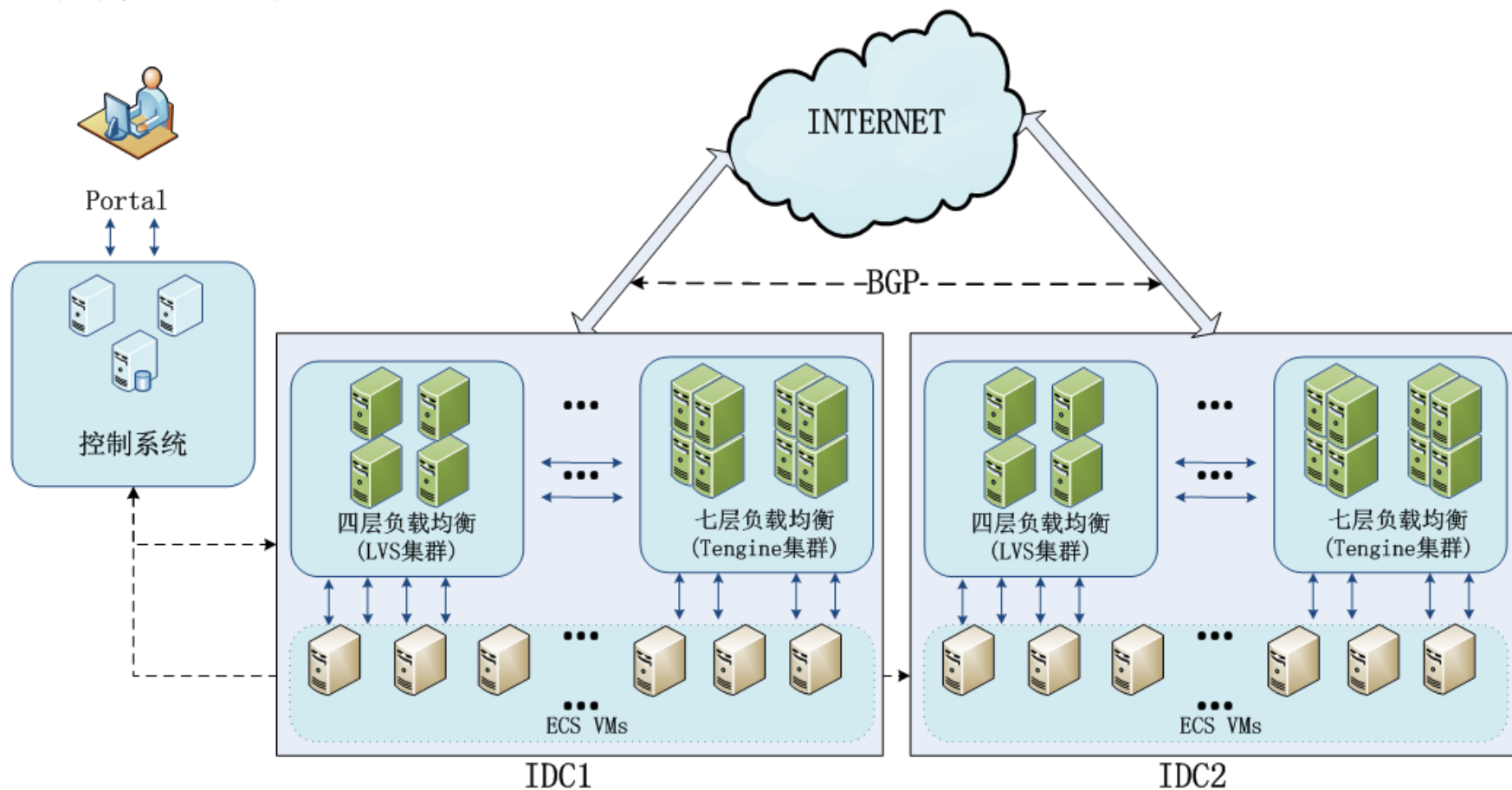
议程



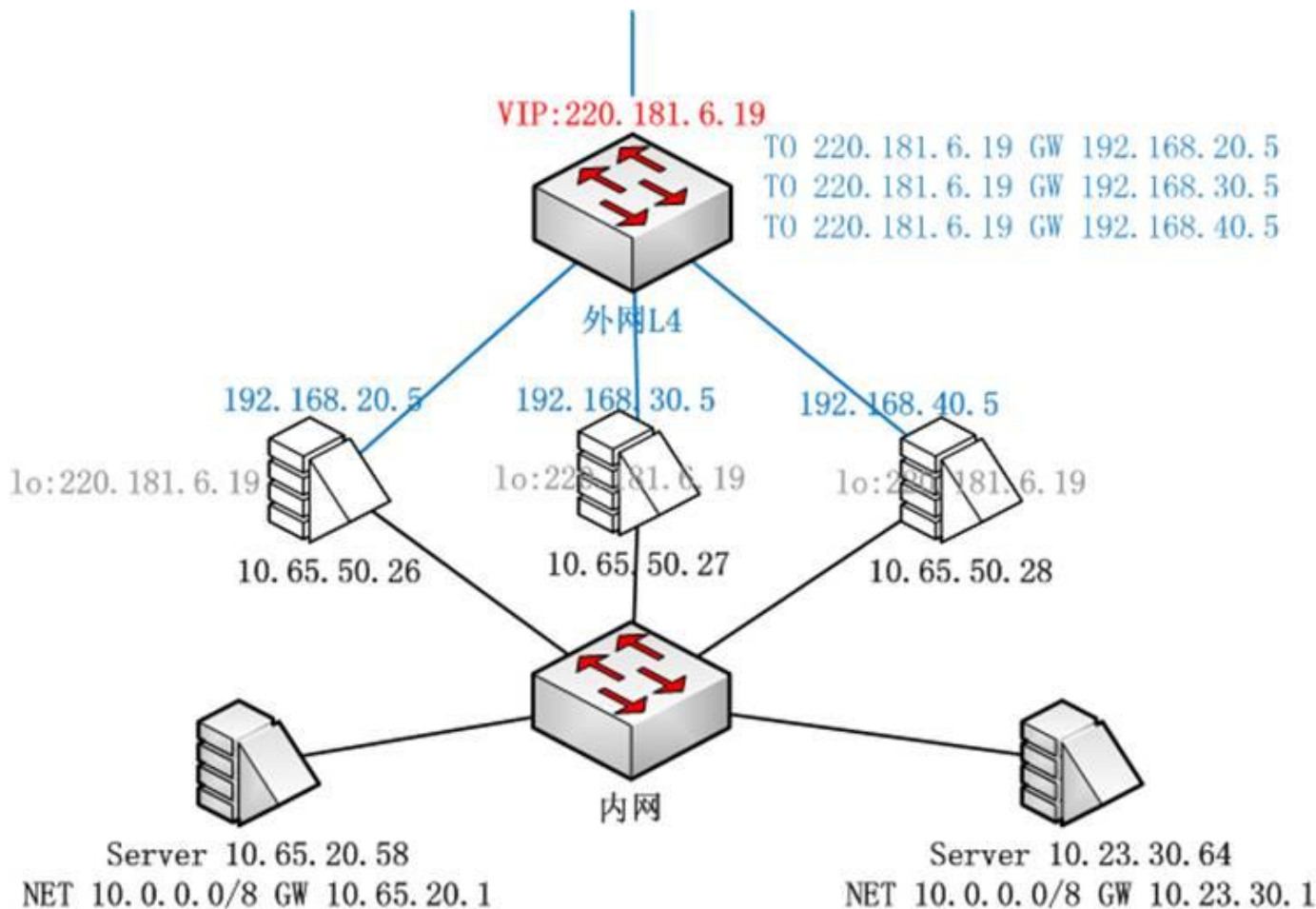
- 
- 一、云计算的挑战与需求
 - 二、ECS的分布式存储设计
 - 三、SLB、RDS与OCS的设计
 - 四、全链路监控与分析系统
 - 五、未来工作展望

SLB的架构设计

SLB集群模块、组件图



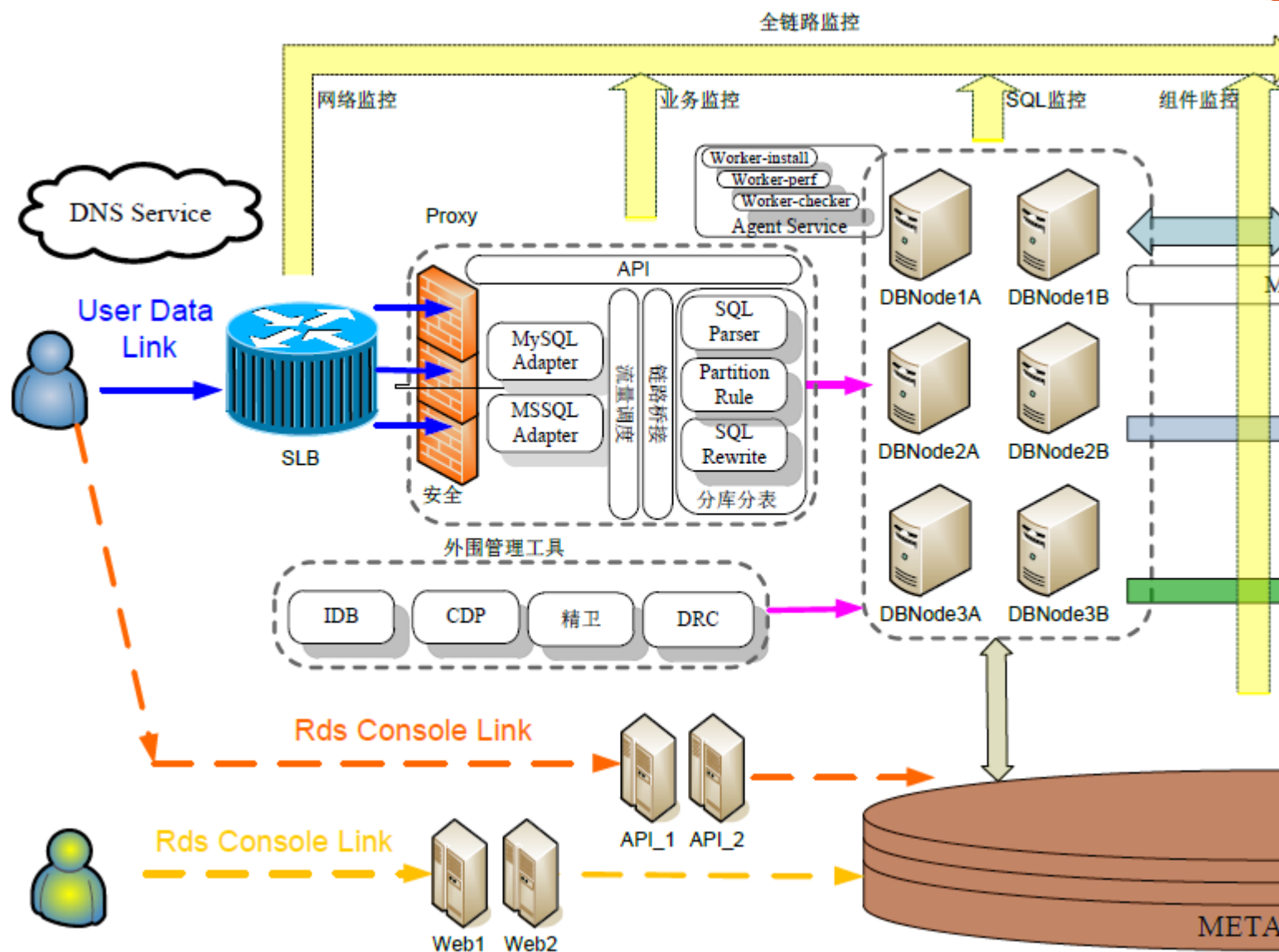
通过OSPF搭建SLB集群



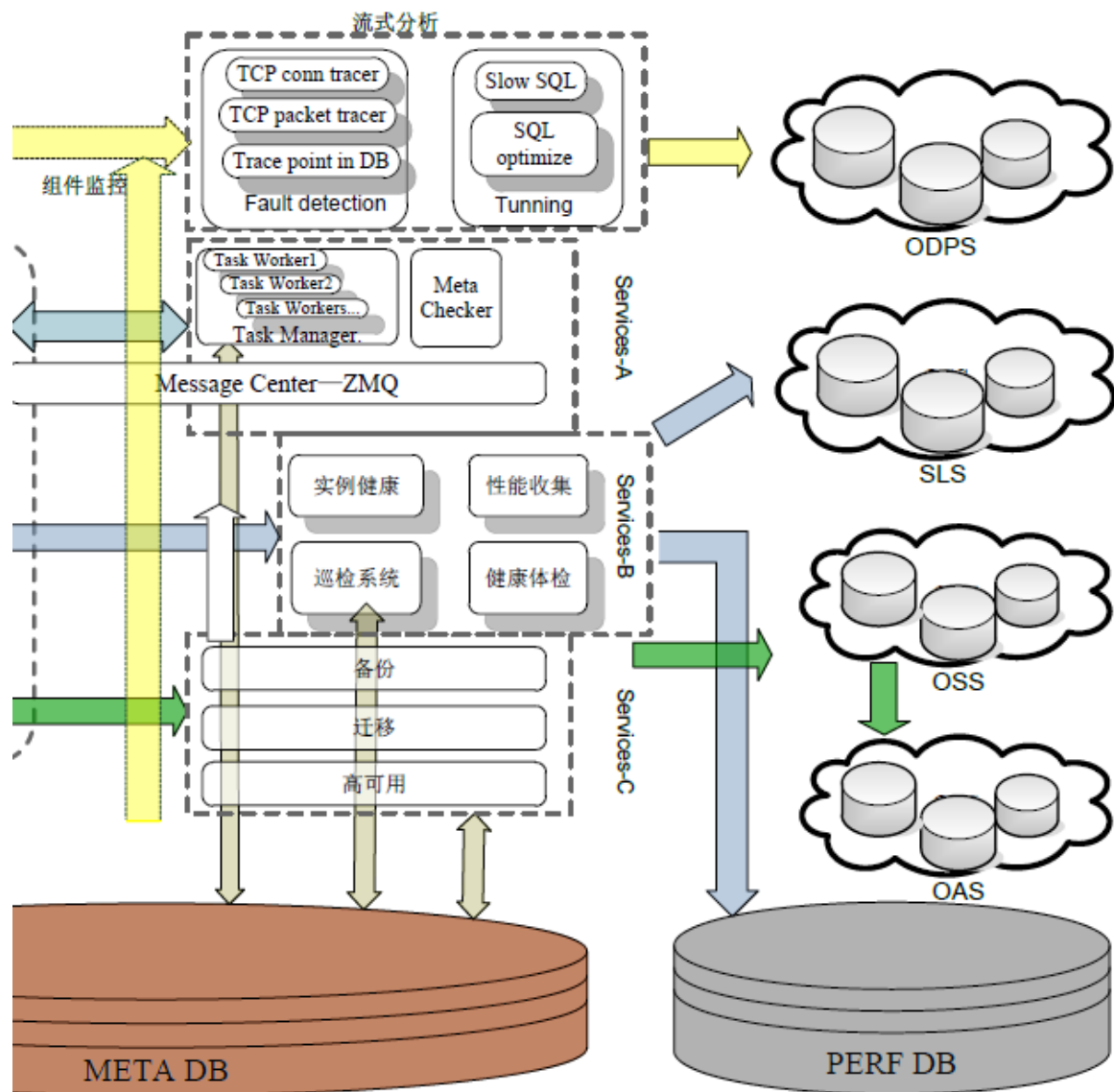
SLB的特点

- 四层负载均衡，采用开源软件LVS，并根据云计算需求对其进行了定制化
 - 12物理核机器，正常转发性能1200万+pps，攻击防御性能万兆线速+
- 七层负载均衡，采用开源软件Tengine
 - 与云盾联动实现七层防攻击
- SLB集群实现非常高的可用性，还通过Anycast做双机房高可用

RDS的架构设计

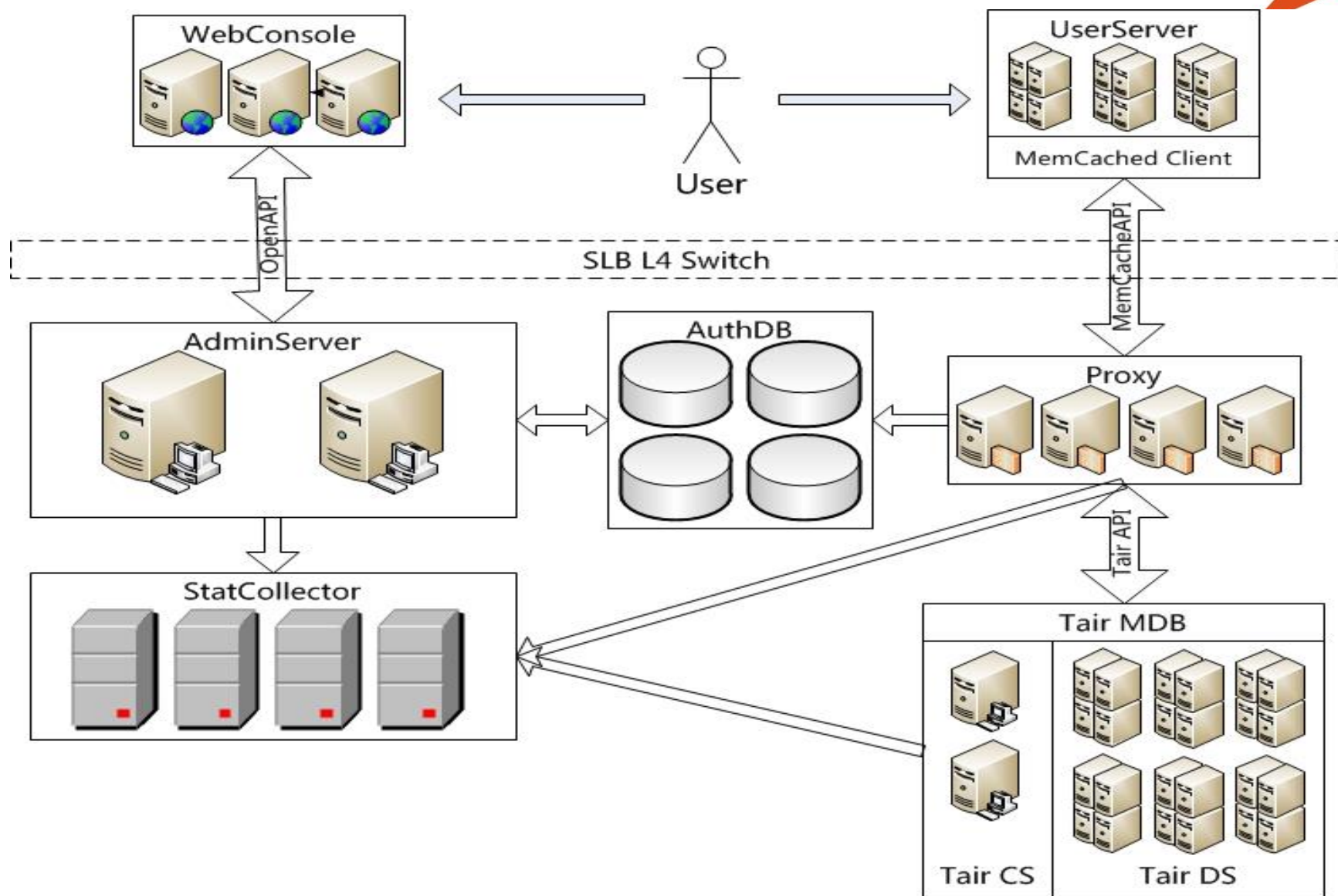


RDS的架构设计（续）



- RDS数据通道架构
 - 标准三层架构，每层都做到机房和部件冗余，无状态设计。
 - 中间层关键特性： a) 安全防护，抗攻击和SQL注入； b) 业务层面流量均衡和调度； c) 桥接功能，规避运维带来的闪断。
 - 数据操作功能，可以做分库分表、匹配不同的后端。
- RDS管理通道架构
 - 元数据库（MySQL）为中心，消息驱动，各组件异步通信。
 - 任务工作流引擎，如CreateDB步骤细化达到128步，任务错误可重试、回滚。
 - 组件无状态，可热升级。


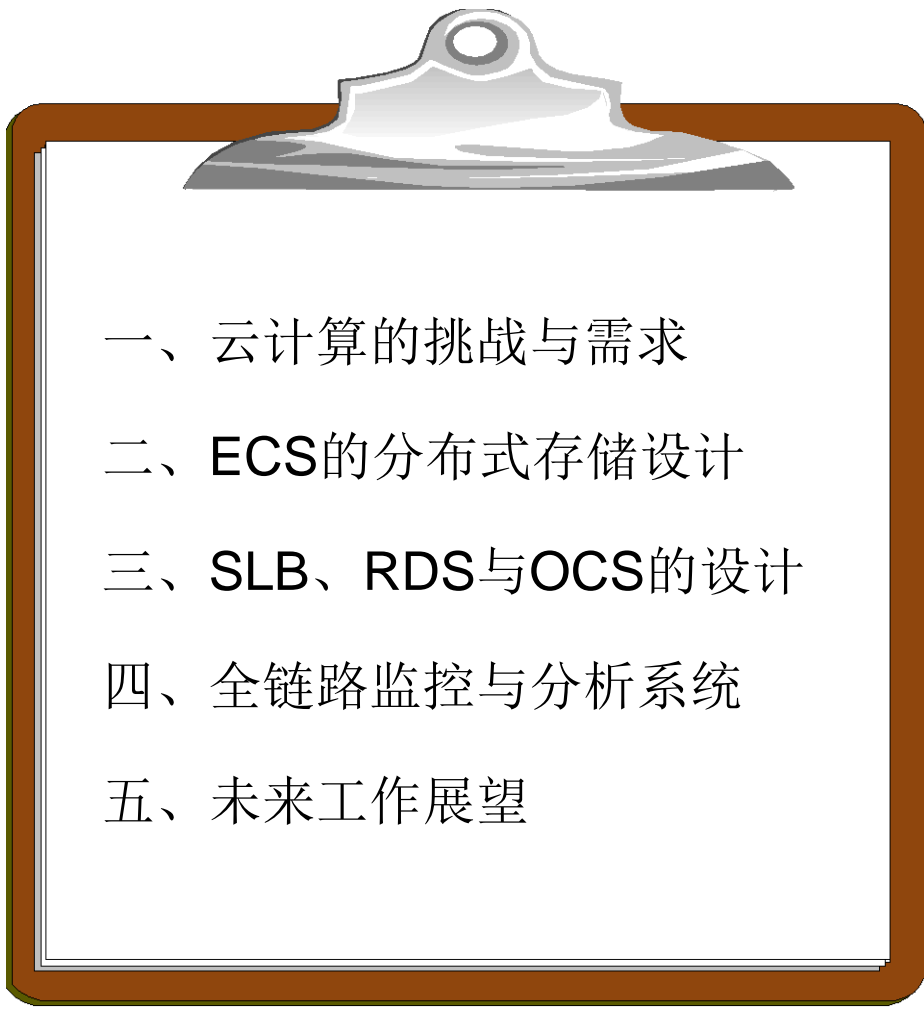
OCS的架构设计



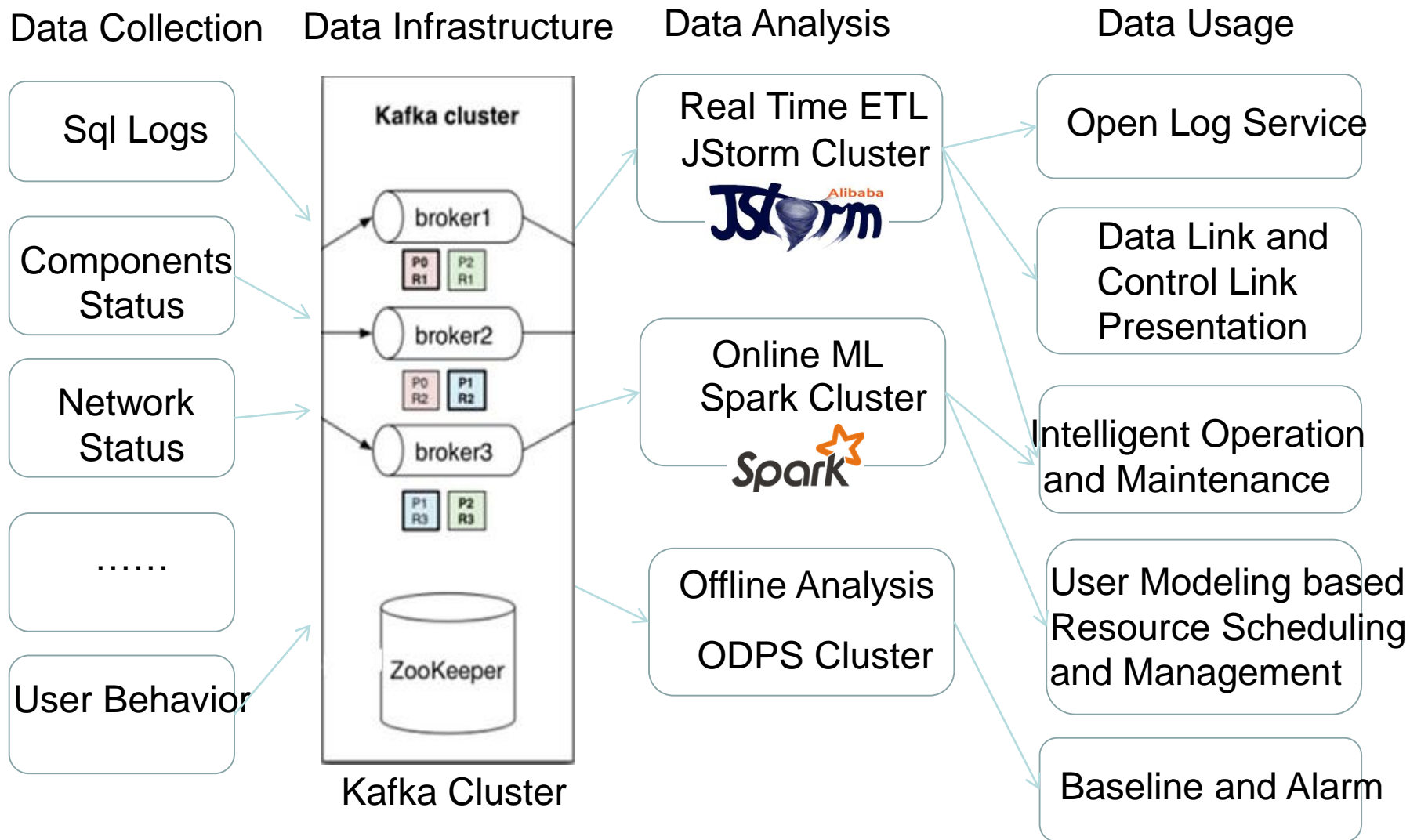
OCS的特点

- 无需自己搭建和运维
 - 扩容，缩容一键自助完成，无需等待
 - 高稳定，宕机自动处理，分布式架构
 - 丰富的监控数据与图形展示
- 高性能
 - 99% 请求在2ms以内响应
 - 并发性能稳定，百万OPS级别的处理能力
- 低成本
 - ECS上自建Memcached成本的一半
- 简单易用
 - Memcached 客户端丰富
 - API简单明了

议程

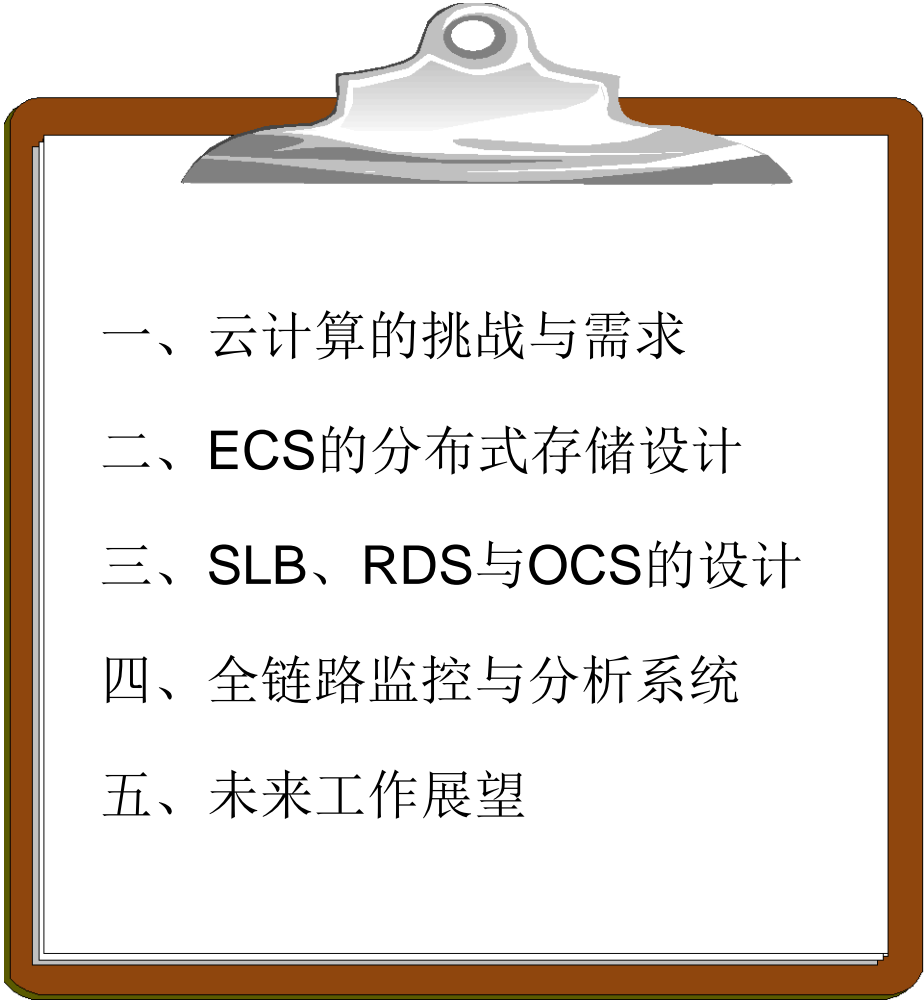

- 
- 
- 一、云计算的挑战与需求
 - 二、ECS的分布式存储设计
 - 三、SLB、RDS与OCS的设计
 - 四、全链路监控与分析系统
 - 五、未来工作展望

全链路监控与分析系统



- 优势：
 - 日志收集器收集率和代价可控，单元化部署。
 - 准实时数据分析，匹配多种流分析工具。
 - 统一的平台，容易扩展和维护，用户专注实现业务逻辑。
- 挑战：
 - 数据量，单SQL采集每天几十T。
 - 秒级实时性，先于用户发现问题。
- 应用：已在RDS中使用，SLB也有初步使用

议程

- 
- 
- 一、云计算的挑战与需求
 - 二、ECS的分布式存储设计
 - 三、SLB、RDS与OCS的设计
 - 四、全链路监控与分析系统
 - 五、未来工作展望

未来工作展望

- ECS：全路径I/O持续优化，Cache策略的优化，带SSD的读写缓存，存储与计算分离，万兆纯SSD集群，动态热点迁移技术，GPU支持，LXC/cgroups支持等。
- RDS：PostgreSQL支持，更低成本的可容忍一定切换时间RDS服务等。
- 全链路的监控与分析系统应用到全线云产品。
- 推出更多的云产品，无线网络加速、AliBench服务质量监测、OCR识别服务、深度学习的CNN/DNN计算服务等。

讨论

- Q&A
- 谢谢！