OpsWorld 运维世界大会·深圳站

# 新存储，新运维
# 分布式对象存储的运维

李明宇 @ OStorage

# 分布式对象存储的运维

运维世界 Ops World

数据增长刚需和对象存储技术

OpenStack Swift开源分布式对象存储系统

Swift跨地域部署和存储池

Swift日志分析

Swift监控报警

李明宇 @ OStorage

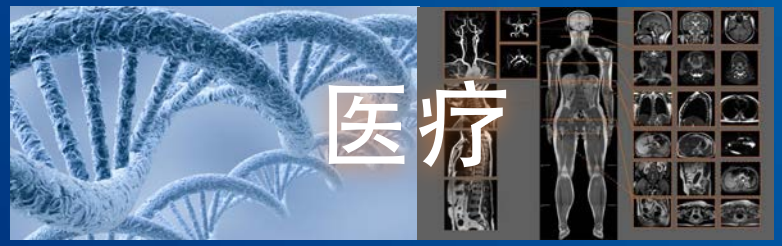
# 数据增长刚需和对象存储技术

运维世界 Ops World

1. 数据量持续快速增长
   - 90%以上是非结构化数据
   - 海量小文件与大体积文件共存


2. 数据访问模式变化
   - 虚拟化、云化、互联网访问
   - 这些数据不太冷
   - 数据共享


3. 数据管理方式变化
   - 基于文件夹的管理被
     基于元数据的管理取代



# 数据增长刚需和对象存储技术

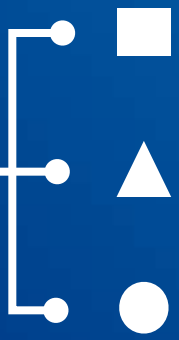
运维世界 Ops World

**Object Store** — S3-like storage

Objects are stored in buckets containers

RESTful HTTP API

GET PUT DELETE



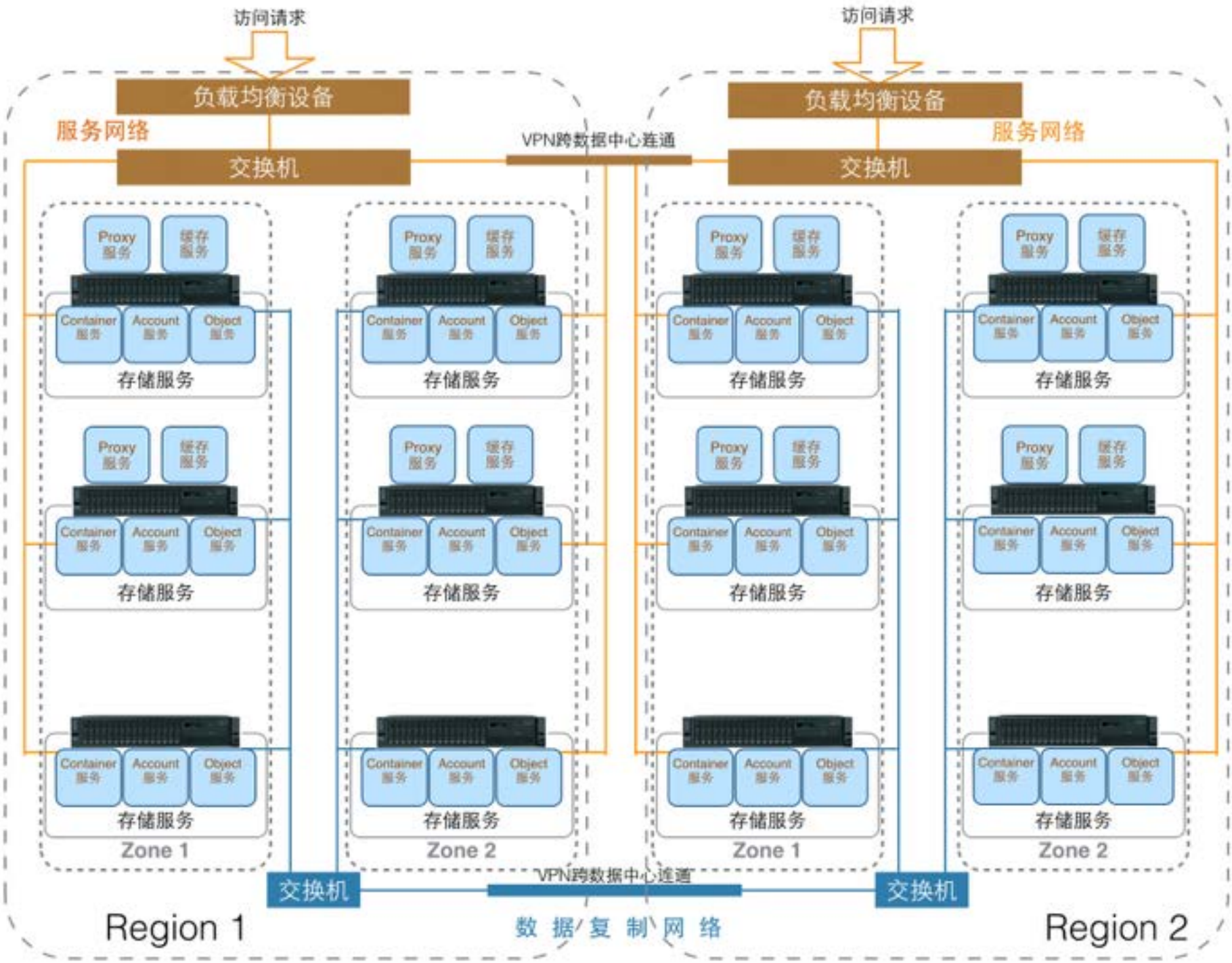
Data

简洁!
易扩展!
易共享!

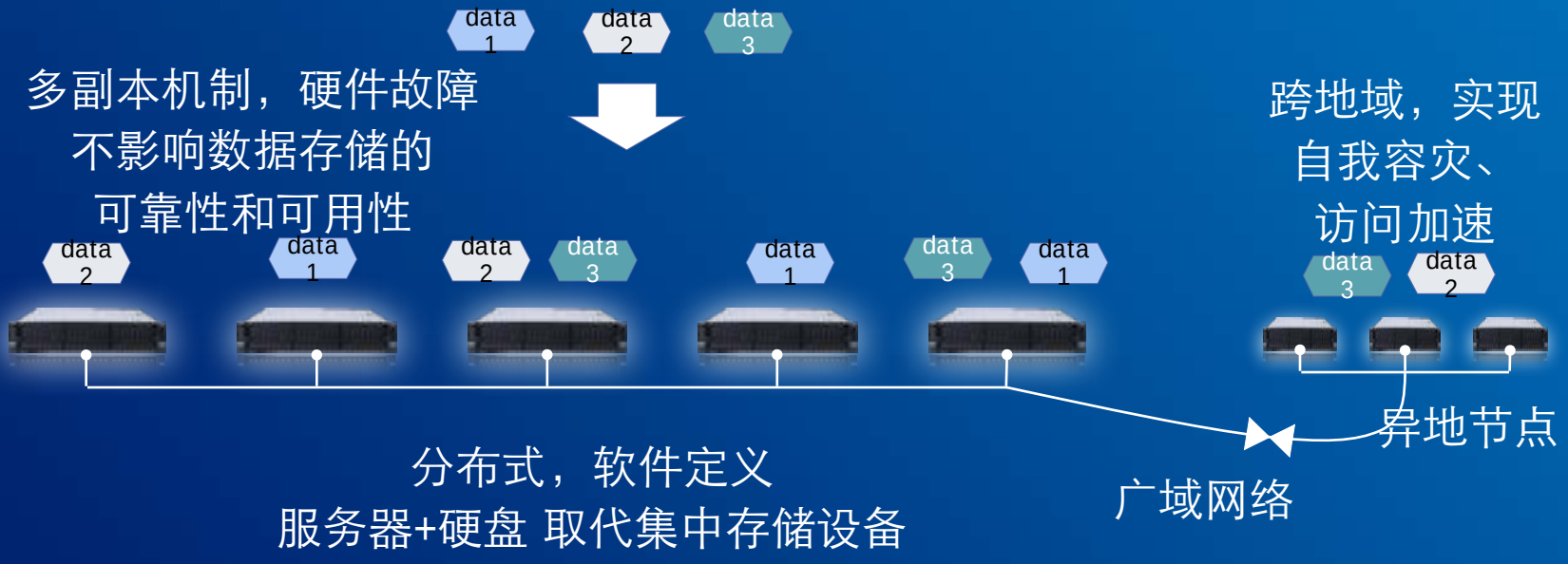
从根本上解决传统文件系统
在文件数量很多、存储规模较大时
性能衰减、可扩展性差、共享能力差的问题

# OpenStack Swift开源对象存储


- OpenStack六个核心服务之一
- 非结构化数据存储，百PB级实际案例
- 无集中节点或组件，全分布式架构，可用性、可靠性极高
- 支持跨机房、跨地域多活，实现单集群全国、全球范围分布
- 支持多租户
- 支持Hadoop
- 支持OpenStack各种备份数据存储，帮助OpenStack环境实现灾备和双活。


访问请求
访问请求
负载均衡设备
负载均衡设备
服务网络
服务网络
VPN跨数据中心连通
交换机
交换机
Proxy 服务
缓存 服务
Container 服务
Account 服务
Object 服务
存储服务
Zone 1
Zone 2
交换机
VPN跨数据中心连通
交换机
Region 1
数据复制网络
Region 2

# Swift跨地域部署和存储池



# Swift跨地域部署和存储池


Only affect object PUT

# Swift日志分析

运维世界 Ops World

- 每秒钟多少次PUT/GET/POST/DELETE请求?
- 这些访问请求来自谁?
- 完成这些请求需要消耗多少时间?
- 这些请求上传（写入）/下载（读取）多少数据?
- 跨地域多副本，数据是否完成同步?
- ……

# Swift日志分析

运维世界 Ops World

```
Jul 27 01:32:50 node2 account-replicator: 2 successes, 0 failures
Jul 27 01:32:50 node2 account-replicator: no_change:2 ts_repl:0 diff:0 rsync:0 diff_capped:0
hashmatch:0 empty:0
Jul 27 01:32:51 node2 container-server: 192.168.93.130 - - [26/Jul/2016:17:32:51 +0000] "PUT /sdb/
826/AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56" 201 - "PUT http://192.168.93.30:8080/
sdb/734/AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56"
"tx87955217f57444bfaa9e6-0057979ec3" "object-server 2079" 0.0004 "-" 5108 0
Jul 27 01:32:51 node2 object-server: 192.168.93.132 - - [26/Jul/2016:17:32:51 +0000] "PUT /sdc/
734/AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56" 201 - "PUT http://192.168.93.30:8080/
v1/AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56"  "tx87955217f57444bfaa9e6-0057979ec3"
"proxy-server 5111" 0.0163 "-" 5110 0
Jul 27 01:32:51 node2 proxy-server: 192.168.93.241 192.168.93.241 26/Jul/2016/17/32/51 PUT /v1/
AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56  HTTP/1.0  201  -  python-requests/
2.7.0%20CPython/2.7.6%20Linux/3.13.0-32-generic  AUTH_tk031c8ad99...  10240  -  -
tx87955217f57444bfaa9e6-0057979ec3 - 0.0232 - - 1469554371.622150898 1469554371.645317078 0
Jul 27 01:32:51 node2 Keepalived_vrrp[4499]: VRRP_Instance(LVS) Received higher prio advert
Jul 27 01:32:51 node2 Keepalived_vrrp[4499]: VRRP_Instance(LVS) Entering BACKUP STATE
Jul 27 01:32:53 node2 account-server: 192.168.93.130 - - [26/Jul/2016:17:32:53 +0000] "HEAD /sdb/
151/AUTH_test"  204  -  "HEAD  http://192.168.93.30:8080/v1/AUTH_test"
"txc467dba681f34a4e818fa-0057979ec5" "proxy-server 2082" 0.0014 "-" 5109 -
Jul 27 01:32:55 node2 object-auditor: Begin object audit "forever" mode (ZBF)
Jul 27 01:32:55 node2 object-auditor: Object audit (ZBF). Since Tue Jul 26 17:32:55 2016: Locally:
1 passed, 0 quarantined, 0 errors, files/sec: 702.33, bytes/sec: 0.00, Total time: 0.00, Auditing
time: 0.00, Rate: 0.00
Jul 27 01:32:56 node2 object-replicator: Starting object replication pass.
Jul 27 01:33:00 node2 object-replicator: 920/920 (100.00%) partitions replicated in 3.76s (244.37/
sec, 0s remaining)
```

# Swift日志分析

运维世界 Ops World

```
Jul 27 01:32:50 node2 account-replicator: 2 successes, 0 failures
Jul 27 01:32:50 node2 account-replicator: no_change:2 ts_repl:0 diff:0 rsync:0 diff_capped:0
hashmatch:0 empty:0
Jul 27 01:32:51 node2 container-server: 192.168.93.130 - - [26/Jul/2016:17:32:51 +0000] "PUT /sdb/
826/AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56" 201 - "PUT http://192.168.93.30:8080/
sdb/734/AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56"
"tx87955217f57444bfaa9e6-0057979ec3" "object-server 2079" 0.0004 "-" 5108 0
Jul 27 01:32:51 node2 object-server: 192.168.93.132 - - [26/Jul/2016:17:32:51 +0000] "PUT /sdc/
734/AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56" 201 - "PUT http://192.168.93.30:8080/
v1/AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56"  "tx87955217f57444bfaa9e6-0057979ec3"
"proxy-server 5111" 0.0163 "-" 5110 0
Jul 27 01:32:51 node2 proxy-server: 192.168.93.241 192.168.93.241 26/Jul/2016/17/32/51 PUT /v1/
AUTH_test/nacho/c80a81ee-7705-4866-b363-618d3e9849f8-56  HTTP/1.0  201  -  python-requests/
2.7.0%20CPython/2.7.6%20Linux/3.13.0-32-generic  AUTH_tk031c8ad99...  10240  -  -
tx87955217f57444bfaa9e6-0057979ec3 - 0.0232 - - 1469554371.622150898 1469554371.645317078 0
Jul 27 01:32:51 node2 Keepalived_vrrp[4499]: VRRP_Instance(LVS) Received higher prio advert
Jul 27 01:32:51 node2 Keepalived_vrrp[4499]: VRRP_Instance(LVS) Entering BACKUP STATE
Jul 27 01:32:53 node2 account-server: 192.168.93.130 - - [26/Jul/2016:17:32:53 +0000] "HEAD /sdb/
151/AUTH_test"  204  -  "HEAD  http://192.168.93.30:8080/v1/AUTH_test"
"txc467dba681f34a4e818fa-0057979ec5" "proxy-server 2082" 0.0014 "-" 5109 -
Jul 27 01:32:55 node2 object-auditor: Begin object audit "forever" mode (ZBF)
Jul 27 01:32:55 node2 object-auditor: Object audit (ZBF). Since Tue Jul 26 17:32:55 2016: Locally:
1 passed, 0 quarantined, 0 errors, files/sec: 702.33, bytes/sec: 0.00, Total time: 0.00, Auditing
time: 0.00, Rate: 0.00
Jul 27 01:32:56 node2 object-replicator: Starting object replication pass.
Jul 27 01:33:00 node2 object-replicator: 920/920 (100.00%) partitions replicated in 3.76s (244.37/
sec, 0s remaining)
```

# Swift日志分析

- 日志拆分
- 结构化
- 构造搜索条件

# Swift日志分析

- 日志拆分

```
if $programname == 'swift' then /var/log/swift/swift.log
if $programname == 'account-server' then /var/log/swift/account-server.log
if $programname == 'account-replicator' then /var/log/swift/account-replicator.log
if $programname == 'account-auditor' then /var/log/swift/account-auditor.log
if $programname == 'account-reaper' then /var/log/swift/account-reaper.log
if $programname == 'container-server' then /var/log/swift/container-server.log
if $programname == 'container-replicator' then /var/log/swift/container-replicator.log
if $programname == 'container-updater' then /var/log/swift/container-updater.log
if $programname == 'container-auditor' then /var/log/swift/container-auditor.log
if $programname == 'container-sync' then /var/log/swift/container-sync.log
if $programname == 'object-server' then /var/log/swift/object-server.log
if $programname == 'object-replicator' then /var/log/swift/object-replicator.log
if $programname == 'object-updater' then /var/log/swift/object-updater.log
if $programname == 'object-auditor' then /var/log/swift/object-auditor.log
```

# Swift日志分析

- 结构化

Aug 4 14:33:47 swift11 proxy-server: 172.17.204.71 172.17.204.71 04/Aug/2016/06/33/47 PUT /v1/AUTH_94cc9b871be647f2bd505aba2f71374e/test/etc/apparmor.d/abstractions/ldapclient HTTP/1.0 201 - python-swiftclient-3.0.0 16253a1d16b4433a... 686 - - tx82d631d7e2a94d2dada82-0057a2e1cb - 0.0482 - - 1470292427.701508045 1470292427.749711037 0

系统时间戳，主机名，服务名，客户端**IP**，远端节点**IP**，代理服务时间戳，**HTTP**请求方法，**HTTP**请求路径，**HTTP**版本，响应状态码，**HTTP referer**，用户代理，认证令牌，接收数据量，发送数据量，客户端**etag**，**Transaction ID**，头信息，延迟时间，**HTTP source**，日志信息，请求开始时间，请求结束时间，存储策略编号。

- 构造搜索条件

# Swift日志分析

- 结构化

Aug 4 14:33:47 swift11 proxy-server: 172.17.204.71 172.17.204.71 04/Aug/2016/06/33/47 PUT /v1/AUTH_94cc9b871be647f2bd505aba2f71374e/test/etc/apparmor.d/abstractions/ldapclient HTTP/1.0 201 - python-swiftclient-3.0.0 16253a1d16b4433a... 686 - - tx82d631d7e2a94d2dada82-0057a2e1cb - 0.0482 - - 1470292427.701508045 1470292427.749711037 0

系统时间戳，主机名，服务名，客户端**IP**，远端节点**IP**，代理服务时间戳，**HTTP**请求方法，**HTTP**请求路径，**HTTP**版本，响应状态码，**HTTP referer**，用户代理，认证令牌，接收数据量，发送数据量，客户端**etag**，**Transaction ID**，头信息，延迟时间，**HTTP source**，日志信息，请求开始时间，请求结束时间，存储策略编号。

- 构造搜索条件

# Swift日志分析

- 后台数据检查与复制进程日志

```
Jul 29 08:42:21 Main object-replicator: Starting object replication
pass.
Jul 29 08:42:33 Main object-replicator: 637/637 (100.00%) partitions
replicated in 12.14s (52.46/sec, 0s remaining)
Jul 29 08:42:33 Main object-replicator: 1274 successes, 0 failures
Jul 29 08:42:33 Main object-replicator: 999 suffixes checked - 0.00%
hashed, 0.00% synced
Jul 29 08:42:33 Main object-replicator: Partition times: max 0.0443s,
min 0.0091s, med 0.0175s
Jul 29 08:42:33 Main object-replicator: Object replication complete.
(0.20 minutes)
```

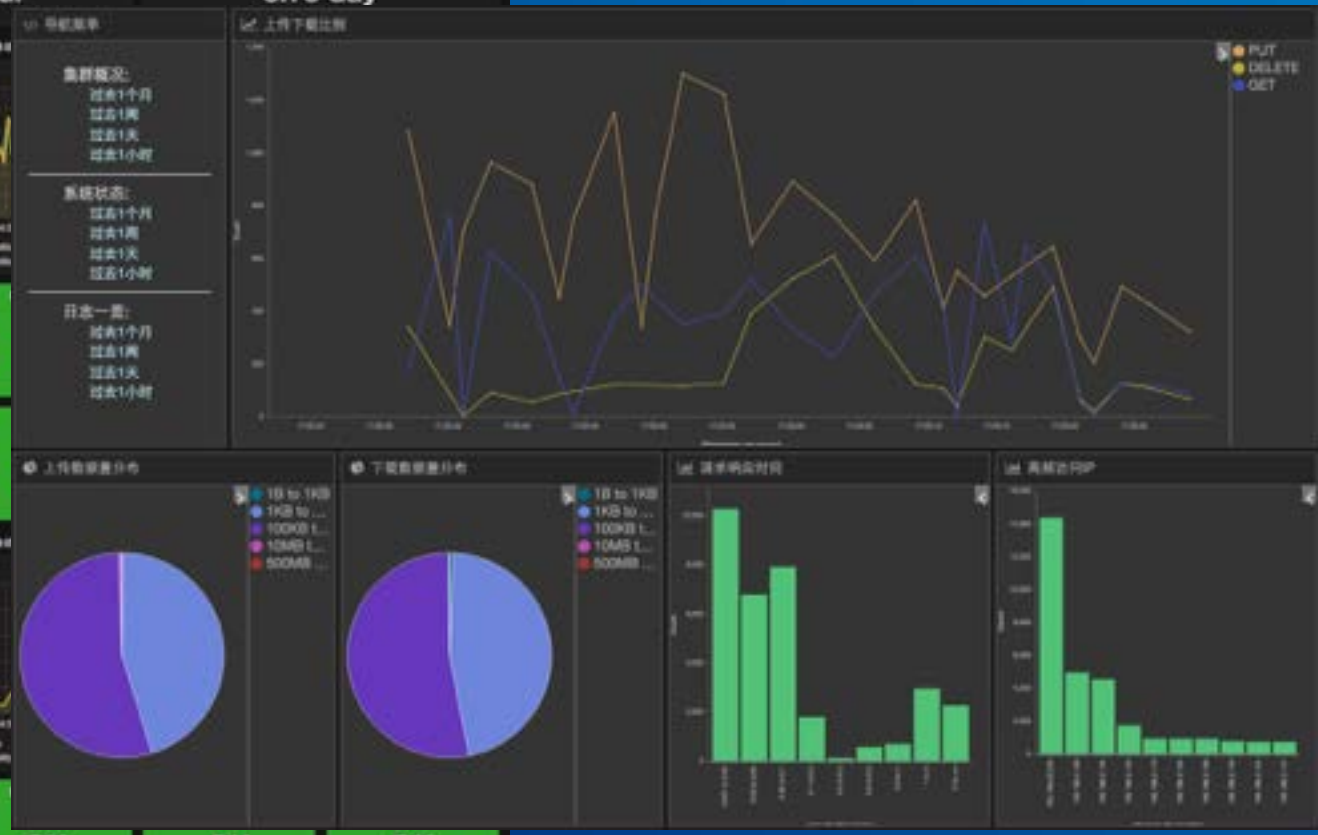
# Swift监控报警

- 存储系统常规指标
  - 盘的利用率
  - 网络利用率
  - CPU利用率
  - ……
- Swift特有的指标
  - 数据分布是否均衡?
  - 时间同步
  - ring的一致性检查
  - storage policy/存储池
  - ……

## 演示

# 运维工具开发

运维世界 Ops World


李明宇 @ OStorage

运维世界 Ops World


李明宇 @ OStorage