GOPS2016
Beijing

2016 全球运维大会 北京站

DevOps 2.0：重塑运维价值

会议时间：12月16日 - 12月17日

会议地点：北京国际会议中心

主办单位：

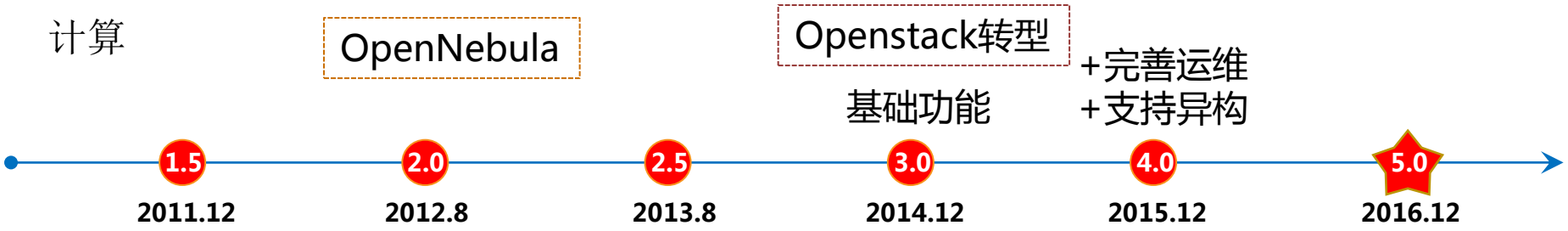
大规模分布式存储开发与实践

刘源　高级架构师

目录

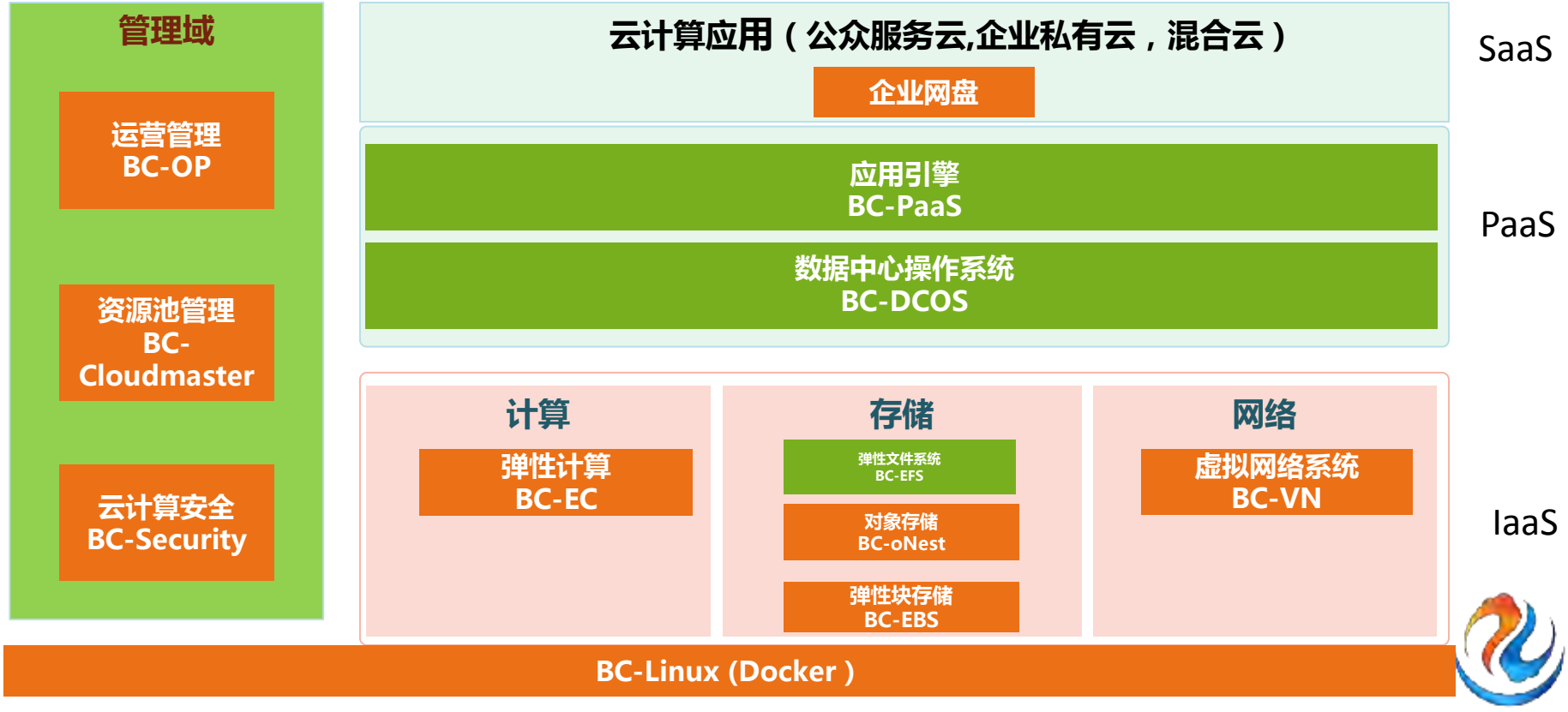
移动大云项目介绍

计算



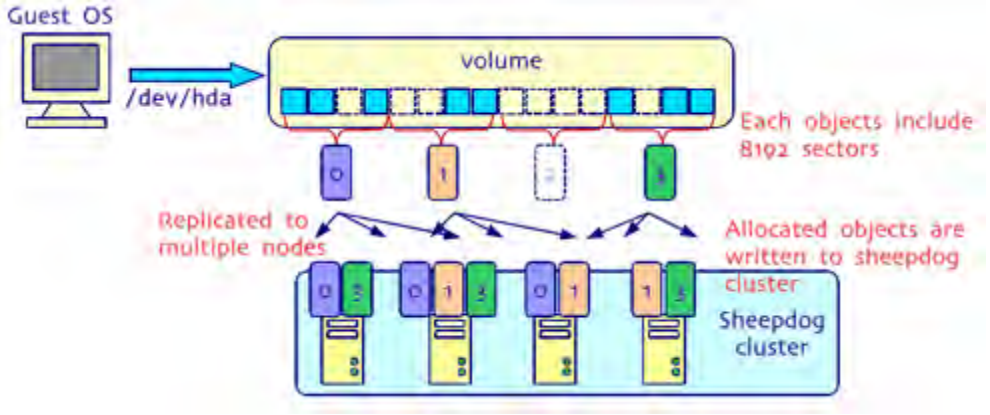
存储

移动云计算产品

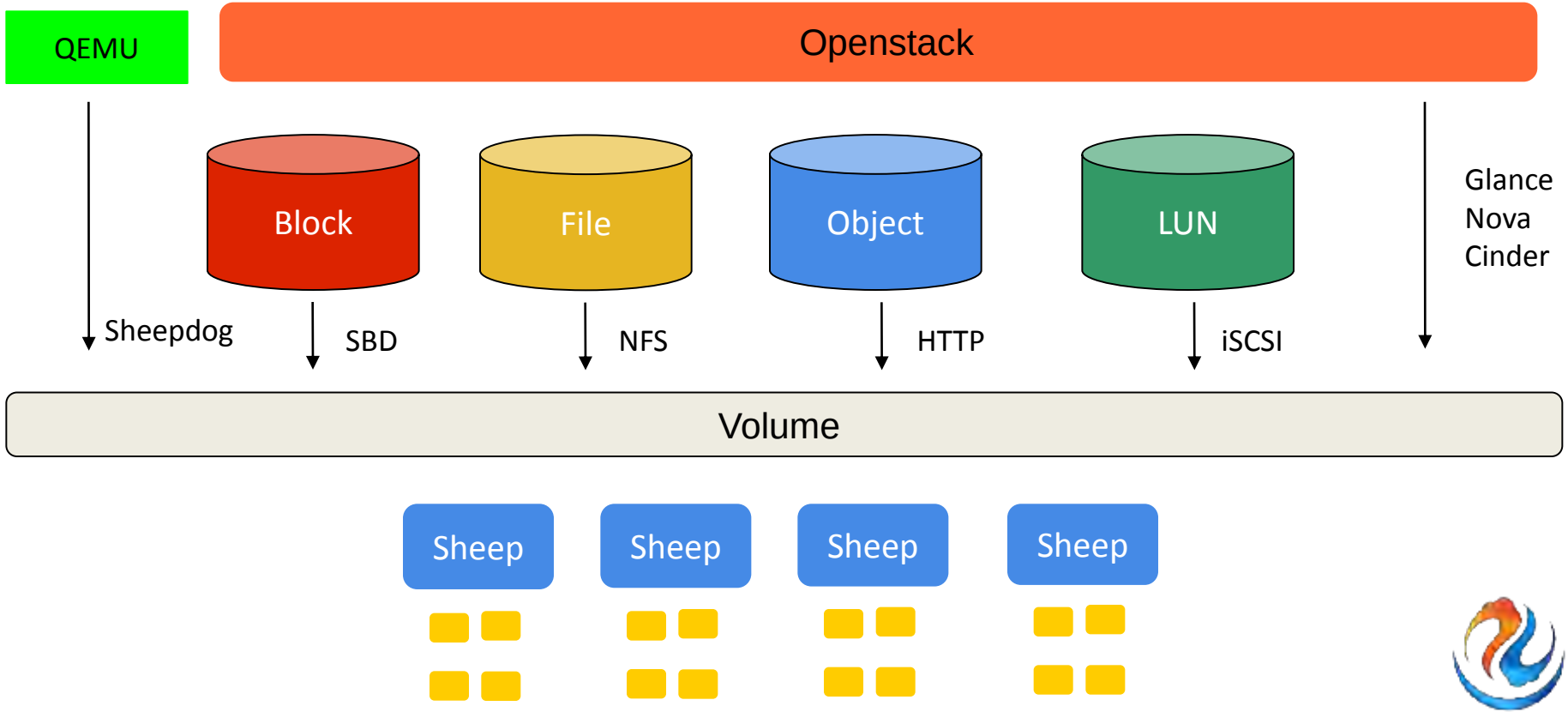

管理域
运营管理
BC-OP
资源池管理
BC-
Cloudmaster
云计算安全
BC-Security
云计算应用（公众服务云,企业私有云，混合云）
企业网盘
SaaS
应用引擎
BC-PaaS
数据中心操作系统
BC-DCOS
PaaS
计算
弹性计算
BC-EC
存储
弹性文件系统
BC-EFS
对象存储
BC-oNest
弹性块存储
BC-EBS
网络
虚拟网络系统
BC-VN
IaaS
BC-Linux (Docker)

Sheepdog介绍

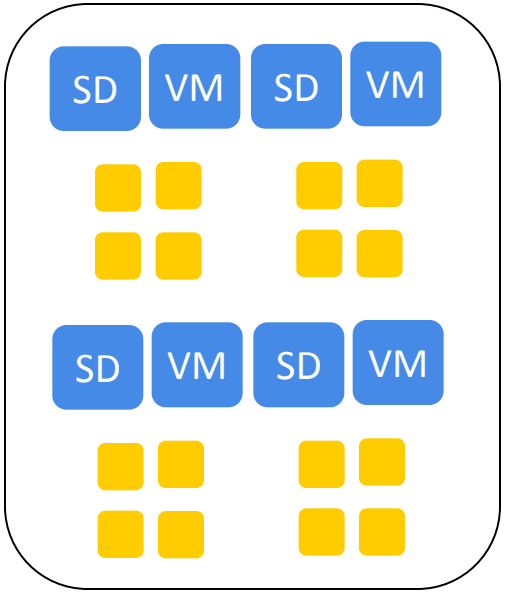
- 用户态分布式对象存储系统
  - **管理磁盘和节点**
    - 聚合容量和能力(IOPS + 吞吐量）
    - 隐藏硬件故障
    - 动态扩容或缩容
  - **保护数据**
    - 提供冗余机制(复制和纠删码)
    - 提供数据自动迁移和修复机制
  - **多样化接口（单个集群）**
    - 虚拟机虚拟卷
    - 对象存储(Swift和S3协议）
    - Openstack (Cinder, Glance, Nova) 原生支持
    - 通过NFS协议提供文件接口
    - Linux内核级块设备



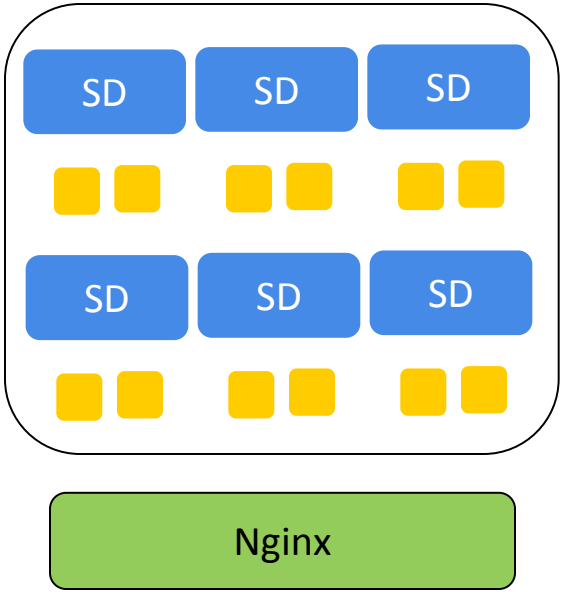
软件全栈视图


QEMU
Openstack
Block
File
Object
LUN
Glance
Nova
Cinder
Sheepdog
SBD
NFS
HTTP
iSCSI
Volume
Sheep
Sheep
Sheep
Sheep

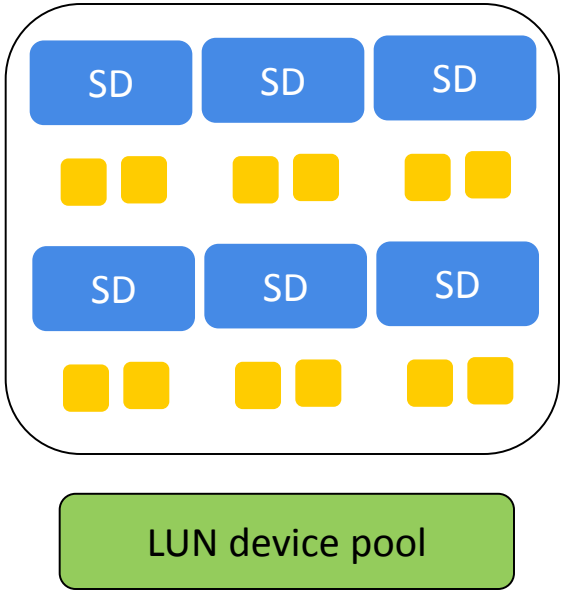
应用方案



计算存储融合



对象存储



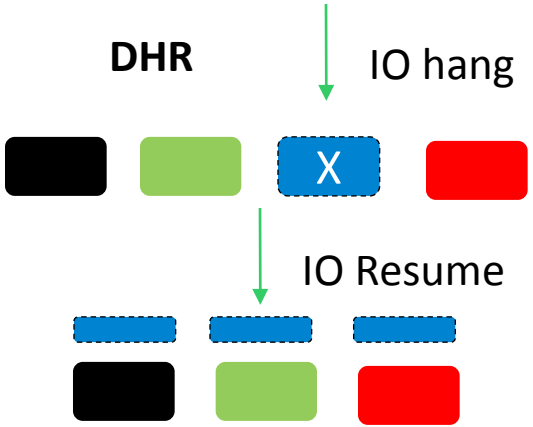
计算存储分离

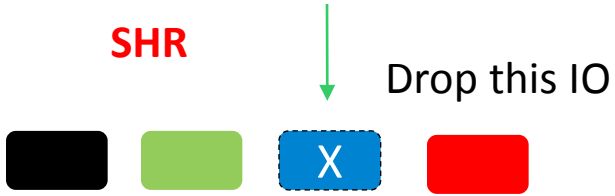
Sheepdog上线情况

- 500+节点
  - 单集群10~40个节点，单节点12块机械硬盘，计算存储分离
- 服务
  - STGT + sheepdog: 通过iSCSI LUN提供SAN服务
  - QEMU 直连：提供虚机块服务

线上问题

- 大多数都是网络故障（网卡down，交换机问题，网络抖动等），少量磁盘故障
  - 建议：Zookeeper最好定时重启
- 集群恢复时服务质量下降严重
  - 解决方案：开发静态哈希特性，避免单节点故障触发全集群恢复

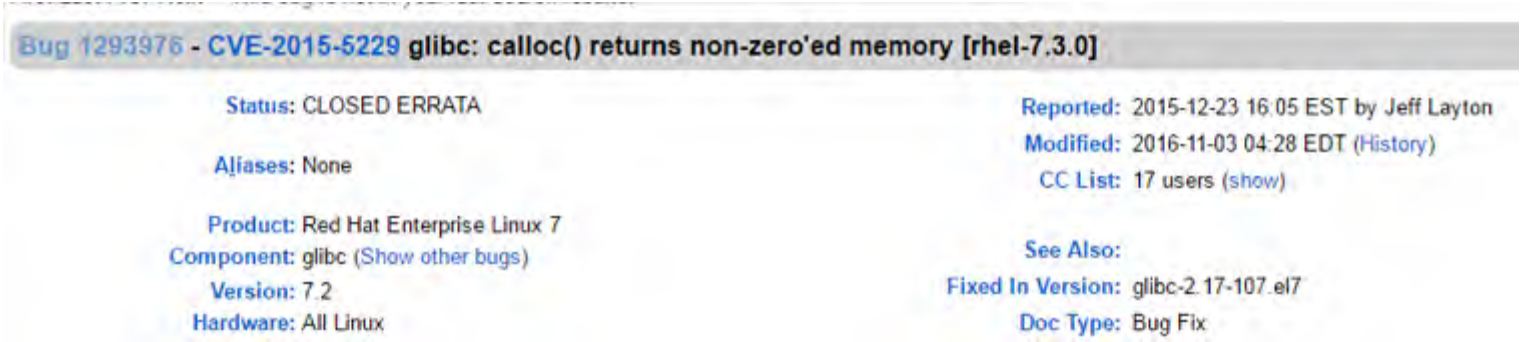




- 80%的故障都是单节点故障
- 数据有冗余，短期内不用恢复
- 单节点恢复，而不是集群恢复
- 避免集群恢复

线上问题

- 缺乏有效监控，运维效率低
- 系统升级内核以及glibc Bug导致数据异常，应用Crash
  - 建议：不要贸然升级操作系统

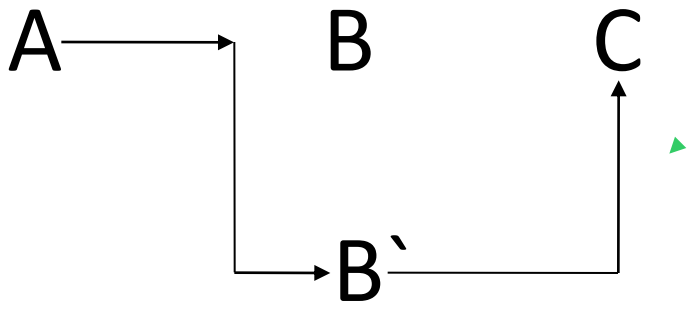


线上问题

- 不能热升级
  - 解决方案：热补丁
    - 原理：0 二进制代码提取 1 动态加载(类似.so) 2 函数替换 (类似内核Ftrace)
    - 通常对小问题修复很有效，不能修改数据结构和全局变量。

A ----> B ----> C **After Patching**

```
diff --git a/sheep/cluster/zookeeper.c b/sheep/cluster/zookeeper.c
index 487455d..c3f9f8a 100644
--- a/sheep/cluster/zookeeper.c
+++ b/sheep/cluster/zookeeper.c
@@ -1241,6 +1264,7 @@ static const int zk_max_event_handlers = ARRAY_SI
 static inline void handle_session_expire(void)
 {
        /* clean memory states */
+
        close(efd);
        zk_tree_destroy();
        INIT_RB_ROOT(&zk_node_root);
```



B` is loaded by Linux's dynamic loader on the fly

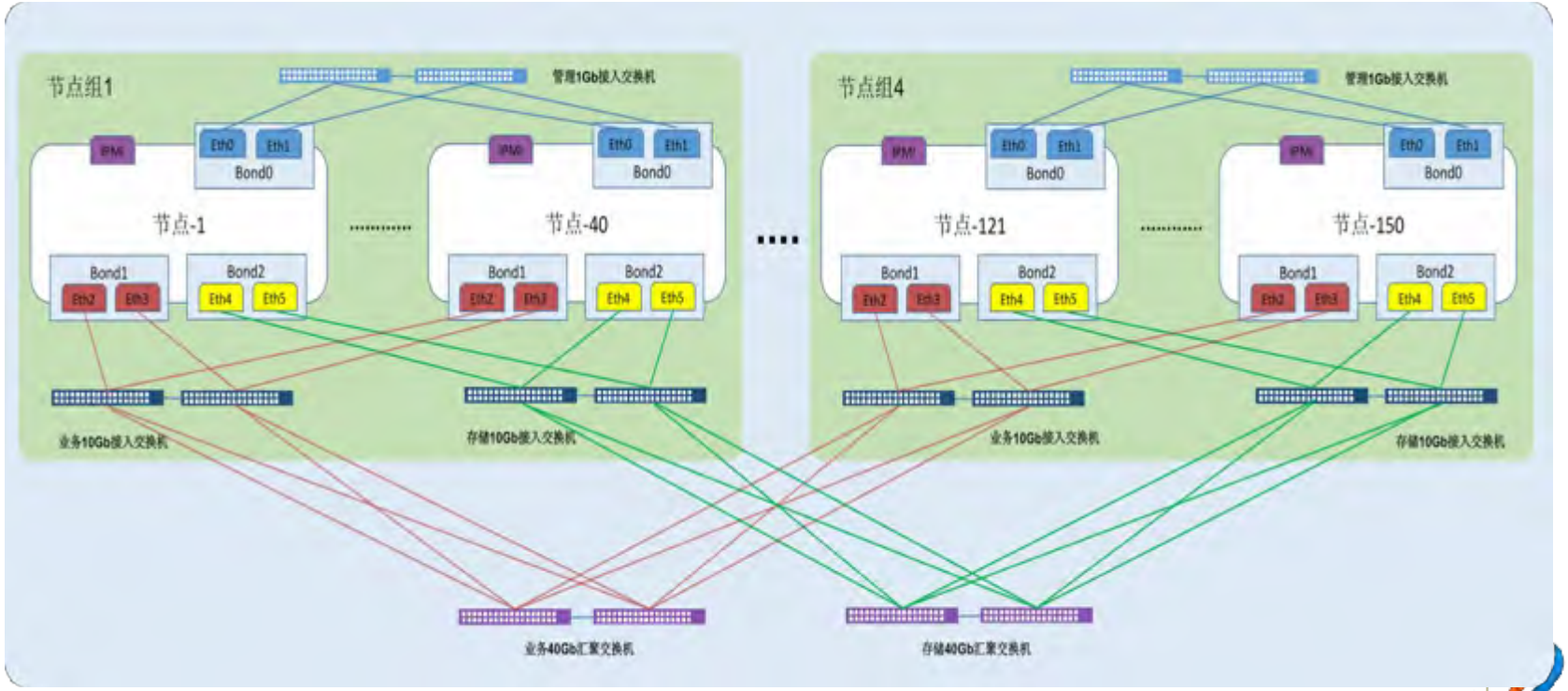
转型Ceph

- 主要原因
  - Ceph越来越稳定
  - 虽然复杂，但社区非常活跃，问题快速响应
  - Openstack支持度最好
  - 厂商支持（开发运维监控）

苏研Ceph上线情况

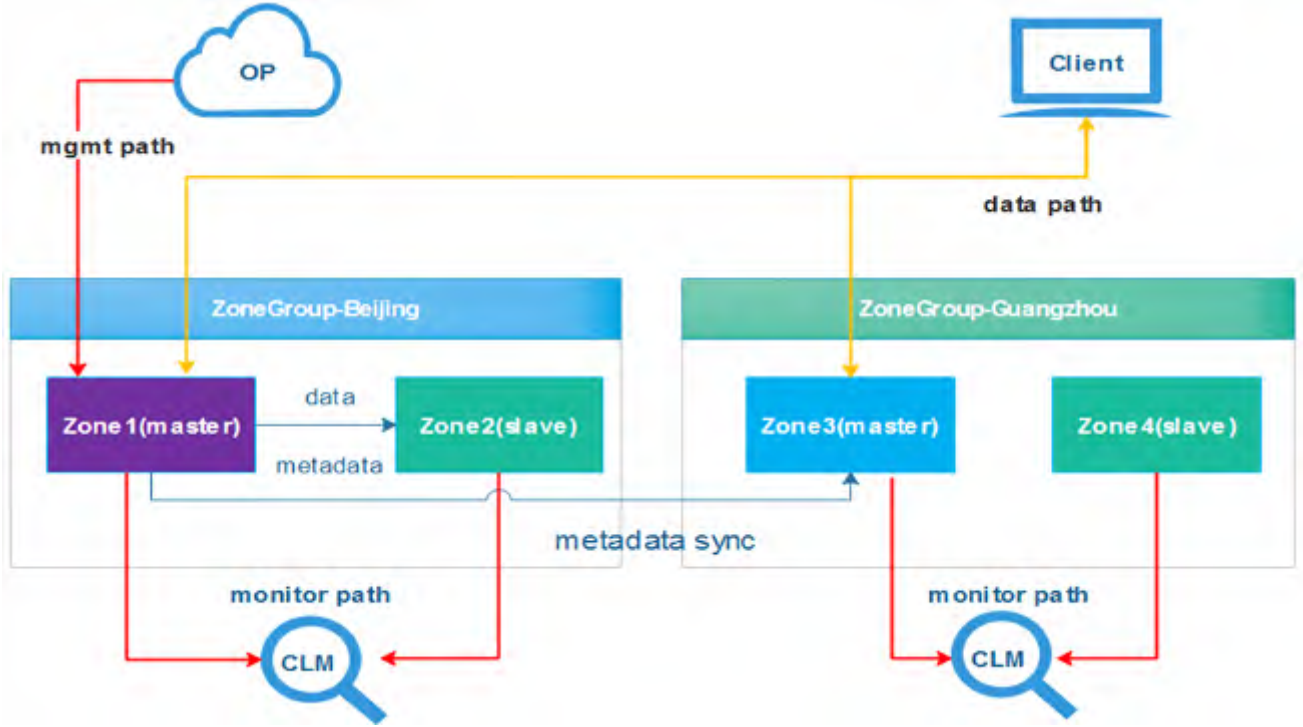
- 基于J版
- 对象存储
  - 上线500+节点
  - 单集群100+节点，单节点24个硬盘，可能是全球最大的集群(2k+ OSD)
  - 可能是全球首个异地容灾生产案列
- 块存储
  - 虚机块设备 (librbd)和iscsi LUN服务
  - NAS服务 (librbd + nfsd in VM)
  - 上线100+节点
  - 小集群部署

Ceph部署架构图


节点组1
管理1Gb接入交换机
IPMI
Eth0
Eth1
Bond0
节点-1
节点-40
Bond1
Bond2
Eth2
Eth3
Eth4
Eth5
业务10Gb接入交换机
存储10Gb接入交换机
节点组4
节点-121
节点-150
业务40Gb汇聚交换机
存储40Gb汇聚交换机

Ceph多数据中心应用

两个数据中心，异地备份



线上问题

- QEMU Centos 7系列虚机 virtio_blk性能很差
  - virtio_scsi性能正常
  - 原因：centos7.1内核的bdev文件系统未实现readpages接口
  - 建议线上部署多用virtio_scsi而不是virtio_blk，这也是红帽推荐的

苏研存储技术号

DevOpsDays 即将首次登陆中国

DevOps 之父 Patrick Debois 与您相约

DevOpsDays 北京站 2017年3月18日

门票早鸟价仅限前100名，请从速哟

http://2017-beijing.devopsdayschina.org/

想第一时间看到
高效运维社区公众号
的好文章吗？

请打开高效运维社区公众号，点击右上角小人，如右侧所示设置就好

Thanks

高效运维社区

开放运维联盟

荣誉出品