Ceph Practices in China Mobile

Zhang Shaowen
China Mobile (Suzhou) software technology Co., Ltd
2019/5
1. Application Cases In China Mobile
2. Practices In Block Storage
3. Practices In Object Storage
Cloud Solution

SaaS platform
- Cloud storage
- Backup
- IT support
- Enterprise Cloud App
- Streaming media processing
- OA
- Government affairs app

BCP Cloud Management Platform

BC-IaaS
Based on OpenStack
- BC-EC
  Server virtualization
  Based on KVM
- BC-SDS
  Cloud storage
  Ceph/GFS
- BC-VN
  Virtual network integration scheme
- BC-SEC
  Cloud Security Integration Solution

BC-PaaS
Based on Kubernetes/Docker
- Software asset library
- MQ
  Distributed cache
  ...

BC-Linux OS
Based on CentOS

BC-Cube
Hyper-Converged cloud computing / cloud storage / database and customized server

Application Layer

Operation Management Layer

Cloud Bus Layer

Resource Service Layer

Infrastructure Layer
Choose Ceph & Contributions

<table>
<thead>
<tr>
<th>#</th>
<th>Company</th>
<th>Commits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Red Hat</td>
<td>27173</td>
</tr>
<tr>
<td>2</td>
<td>Inktank</td>
<td>13097</td>
</tr>
<tr>
<td></td>
<td>*Independent</td>
<td>11362</td>
</tr>
<tr>
<td>3</td>
<td>SUSE</td>
<td>3874</td>
</tr>
<tr>
<td>4</td>
<td>DreamHost</td>
<td>2861</td>
</tr>
<tr>
<td>5</td>
<td>ZTE Corporation</td>
<td>1906</td>
</tr>
<tr>
<td>6</td>
<td>Intel</td>
<td>1562</td>
</tr>
<tr>
<td>7</td>
<td>Mirantis</td>
<td>1283</td>
</tr>
<tr>
<td>8</td>
<td>XSky</td>
<td>952</td>
</tr>
<tr>
<td>9</td>
<td>Deutsche Telekom</td>
<td>900</td>
</tr>
<tr>
<td>10</td>
<td>Cloudwatt</td>
<td>678</td>
</tr>
<tr>
<td>11</td>
<td>Digiware</td>
<td>284</td>
</tr>
<tr>
<td>12</td>
<td>Reliance</td>
<td>255</td>
</tr>
<tr>
<td>13</td>
<td>China Mobile</td>
<td><strong>215</strong></td>
</tr>
<tr>
<td>14</td>
<td>EasyStack</td>
<td>166</td>
</tr>
<tr>
<td>15</td>
<td>SanDisk</td>
<td>148</td>
</tr>
</tbody>
</table>
1. Application Cases In China Mobile
2. Practices In Block Storage
3. Practices In Object Storage
Hardware Selection

- **Capacity**
  - Higher Capacity
  - Lower Cost
    - 10 SATA HDD + 2 SATA SSD
    - 12 SATA HDD + 1 PCIe SSD

- **Balance**
  - High cost-effective
  - Data tiering
    - 12 SATA HDD + 1-3 PCIe SSD
    - 24 SATA HDD + 12 SATA SSD

- **Performance**
  - Higher Performance
  - Better Experience
    - 12 SATA SSD + 1 PCIe SSD
    - 24 SATA SSD
Effects of Cache

Cache Solution Selection

**Bcache**
- Merged into kernel since 3.10
- Excellent performance
- Optimized for SSD
- SSD pooling, thin-provision
- Rich features
- Developed by Facebook
- Device-mapper
- Good performance
- Easy to deploy and maintain
- Concise code

**Flashcache**
- Based on flashcache
- Device-mapper
- Good performance
- Fewer features

**Enhance IO**
- Merged into kernel since 3.9
- Device-mapper
- Good performance
- Fewer features

**DM-cache**
- Merged into kernel since 3.10
- Excellent performance
- Optimized for SSD
- SSD pooling, thin-provision
- Rich features

**Bcache**
- Merged into kernel since 3.10
- Excellent performance
- Optimized for SSD
- SSD pooling, thin-provision
- Rich features
Cache Solution Selection

Logical Mapping Of Flashcache

SSD

HDD

Effect of Flashcache

Effect of Bcache
Application Case of RBD & iSCSI

Typical use case of RBD

iSCSI solution for RBD

KVM-Qemu
NBD
krbd
librbd
Monitor
Ceph Cluster
Osd
Osd
Osd
Osd
initiator
initiator
iSCSI Target
iSCSI Target

SCSI Command processing
cmd handling daemon
User
Kernel
/config/target
block
file
disk_img
/backstore
/controllers
/LIO core
## Object Storage

<table>
<thead>
<tr>
<th>Interface Layer</th>
<th>S3</th>
<th>Swift</th>
<th>NFS</th>
<th>Admin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service Layer</td>
<td>Usage statistics</td>
<td>User Manager</td>
<td>Compression</td>
<td>Data tiering</td>
</tr>
<tr>
<td>Engine Layer</td>
<td>Hash</td>
<td>Strong consistency</td>
<td>Intelligent recovery</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-healing</td>
<td>Data balance</td>
<td>State management</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hardware awareness</td>
<td>Failure domain</td>
<td>Linearly scale</td>
<td></td>
</tr>
<tr>
<td>Device Layer</td>
<td>x86 server</td>
<td>Disk error checking</td>
<td>SAS/SATA/SSD</td>
<td>Cluster Network (10GE/IB)</td>
</tr>
<tr>
<td></td>
<td>RAID/HBA Management</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Public Network (10GE/IB)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Storage Manager
- Resource Monitor
- Automated deployment
- Online upgrade
- Alarm Management
- Log management
- Disk change
- User management
- Performance analysis
- Cluster Expansion
Multisite in Public Cloud

- Beijing-1
  - zg1-zone1 (master zonegroup)
- Beijing-2
  - zg2-zone1 (slave zonegroup)
- Hunan-1
  - zg3-zone1 (slave zonegroup)
- Guangzhou-1
  - zg4-zone1 (slave zonegroup)

Metadata sync:
- realm-1
Typical Application Scenarios

Backup & Archive
Content Delivery
Big Data Analytics
Website

User
bucket_1
Backup
File Share
Static Website
bucket_2
bucket_3
Thanks!