Sheepdog:
Alternative software-defined storage on your OpenStack cloud

Takashi Menjo <menjo.takashi@lab.ntt.co.jp>

NTT Software Innovation Center
Agenda

- Introduction of Sheepdog
  - What it is
  - Performance
  - Features
- Sheepdog and OpenStack
  - Integration
  - Implementation (hyper-converged)
  - Installation
  - Configuration (Cinder and Glance)
- Conclusions
What's Sheepdog?

"Alternative" open-source distributed storage software
- Fast
- Scalable
- Hyper-converged

Released in 2009 by NTT Lab, now on production use

Users:
- Alibaba (CHN), Taobao (CHN), Extensys (ITA), A.T.WORKS (JPN), etc.
- NTT DATA (JPN)
What's Sheepdog? (cont.d)

- Makes storage cluster† with commodity Linux servers
- Integrates internal disks to storage pool
- Provides virtual volumes for KVM‡ and/or iSCSI usage

[Diagram showing Sheepdog storage cluster with commodity servers, virtual volumes, and backend storage for IaaS or virtual private servers.]

- Virtual server integration environment
- Shared storage

[† with ZooKeeper or Corosync ‡ Kernel-based Virtual Machine]
Performance of Sheepdog

- Faster than Ceph
- Scales when # storage hosts increases
Performance of Sheepdog (cont.d)

- Faster than Ceph
- Scales when # storage hosts increases

---

**write throughput**

<table>
<thead>
<tr>
<th></th>
<th>3 hosts</th>
<th>6 hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceph</td>
<td>100 MB/s</td>
<td>200 MB/s</td>
</tr>
<tr>
<td>Sheepdog</td>
<td>300 MB/s</td>
<td>400 MB/s</td>
</tr>
</tbody>
</table>

**read throughput**

<table>
<thead>
<tr>
<th></th>
<th>3 hosts</th>
<th>6 hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceph</td>
<td>150 MB/s</td>
<td>300 MB/s</td>
</tr>
<tr>
<td>Sheepdog</td>
<td>400 MB/s</td>
<td>800 MB/s</td>
</tr>
</tbody>
</table>

---

(sequential 1024k workload)
Features of Sheepdog

- **Durable**
  - Data redundancy

- **Manageable**
  - Auto-recovery

- **Reliable**
  - No SPOF

- **Scalable**
  - Better performance
Integrating Sheepdog with OpenStack (1 of 3)

• Cinder and/or Glance backend

- Create
- Snapshot
- Delete
- ...
- Boot from
- Attach
- Detach
Integrating Sheepdog with OpenStack (2 of 3)

• Cinder multi-backend supported

Cinder driver

Glance driver

QEMU/KVM driver

Sheepdog cluster

Sheepdog cluster

• Create
• Snapshot
• Delete
• ...

• Boot from
• Attach
• Detach

Nova
Integrating Sheepdog with OpenStack (3 of 3)

• Sharing a cluster improves image-to-volume performance

- Create
- Snapshot
- Delete
- …

Sheepdog cluster

Cinder driver

Glance driver

QEMU/KVM driver

Nova

• Boot from
• Attach
• Detach
Scalability (small start)

• Three nodes required at least

controller

cinder-volume

glance-api
...  
ZooKeeper†

Compute [1]
nova-compute
QEMU/KVM
ZooKeeper†

Compute [2]
nova-compute
QEMU/KVM
ZooKeeper†

sheep  HDD

sheep  HDD

sheep  HDD

Large-capacity

† It is best to run on three or more odd number of machines.
Scalability (scale out)

- Get more performance and capacity by adding nodes

Controller
- cinder-volume
- glance-api ...
- ZooKeeper

sheep  HDD

Compute [1]
- nova-compute
- QEMU/KVM
- ZooKeeper

sheep  HDD

Compute [2]
- nova-compute
- QEMU/KVM
- ZooKeeper

sheep  HDD

Compute [3]
- nova-compute
- QEMU/KVM

sheep  HDD

Large-capacity
Hyper-converged

• Scale both compute and storage resources
Multi-backend

• Manage multiple clusters for your own purposes

Controller
- cinder-volume
- glance-api...
- ZooKeeper

Compute [1]
- nova-compute
- QEMU/KVM
- ZooKeeper

Compute [2]
- nova-compute
- QEMU/KVM
- ZooKeeper

Compute [3]
- nova-compute
- QEMU/KVM

Large-capacity
High-performance
Installing Sheepdog to OpenStack nodes

• Install dependencies
  • OS: Ubuntu or CentOS
    • If CentOS, EPEL† and Apache Bigtop‡ are required
  • QEMU: Use 2.5.0 or later
    • Ubuntu 16.04: Package is available
    • CentOS 7.2: Please build manually

• Deploy Sheepdog
  • Build and install
    • In short, ./configure, make, make install
    • For detail, see right
  • Run sheep on each node
    • sheep --cluster zookeeper:host:port[,...] /ID_A --port 7000 /mnt/hdd
    • sheep --cluster zookeeper:host:port[,...] /ID_B --port 7001 /mnt/ssd

$ git clone https://github.com/sheepdog/sheepdog.git
$ cd sheepdog
$ git fetch origin stable-1.0
$ git fetch origin --tags
$ git checkout -b 1.0.1 refs/tags/v1.0.1
$ ./autogen.sh
$ ./configure --prefix=/usr
   --enable-zookeeper --disable-corosync
$ make && sudo make install

† for yasm and userspace-rcu‡ for ZooKeeper
# cinder.conf
[DEFAULT]
enabled_backends=sheep0,sheep1
[hddsheep]
volume_driver=cinder.volume.drivers.sheepdog.SheepdogDriver
sheepdog_store_address=127.0.0.1
sheepdog_store_port=7000
volume_backend_name=bigsheep
[ssdsheep]
volume_driver=cinder.volume.drivers.sheepdog.SheepdogDriver
sheepdog_store_address=127.0.0.1
sheepdog_store_port=7001
volume_backend_name=rapidsheep
Configuring Glance backend

```plaintext
# glance-api.conf
[DEFAULT]
default_store=sheepdog
stores=sheepdog
sheepdog_store_address=127.0.0.1
sheepdog_store_port=7000
sheepdog_store_chunk_size=64
```
Conclusions

• Sheepdog is "alternative" OSS distributed storage
  • Fast, scalable and hyper-converged

• You can integrate Sheepdog with Cinder and Glance

• Sheepdog stable-1.0 branch released
  • Ready for production use!
  • New features (volume over 4-TiB, recovery speed throttling, etc.)
  • Latest version is v1.0.1 (released in this October)

• Repository:
  • https://github.com/sheepdog/sheepdog

• Mailing list:
  • Developers: sheepdog@lists.wpkg.org
  • Users: sheepdog-users@lists.wpkg.org
Appendix: Performance evaluation

• Comparing Sheepdog with Ceph under even conditions
  • Same hardware, same replication configuration

![Diagram comparing Sheepdog and Ceph configurations](image)

Hardware:
- CPU: Intel(R) Xeon(R) CPU E5-2630L v3 @ 1.80GHz (8C) × 1P
- Memory: 64 GiB
- SSD: Samsung PM863 MZ-7LM480

Software:
- Ceph: 10.2.1 Jewel
- Sheepdog: v1.0
- OS: CentOS 7.2 (Both host and guest)
- QEMU: 2.6.0
- fio: 2.12

Configuration:
- Filesystem: XFS
- # replicas: 2
- Volume: 20 GiB, thick-provisioned
- Guest: 1 vCPU and 4-GiB memory

Procedure:
1. Run 4 guests per host
2. Attach volume provided by Ceph or Sheepdog to each guest
3. Run fio on all guests simultaneously to issue I/O requests to volume for 1 minute
4. Sum up all stats (IOPS and throughput) on guests as a performance result
5. Repeat the steps above 5 times then take the average of performance results excluding best and worst ones as the final result

• Evaluating scalability of Sheepdog and Ceph

Pattern (1): 3 hosts

Pattern (2): 6 hosts

10Gb Ethernet (Both public and private networks)

† run on only 3 hosts