中国移动
China Mobile

Bring SCSI support into QEMU block layer

白耀伟
baiyaowei@cmss.chinamobile.com

www.10086.cn

Agenda

- Backgroud
- Solution
- Howto
- Status
- Future work

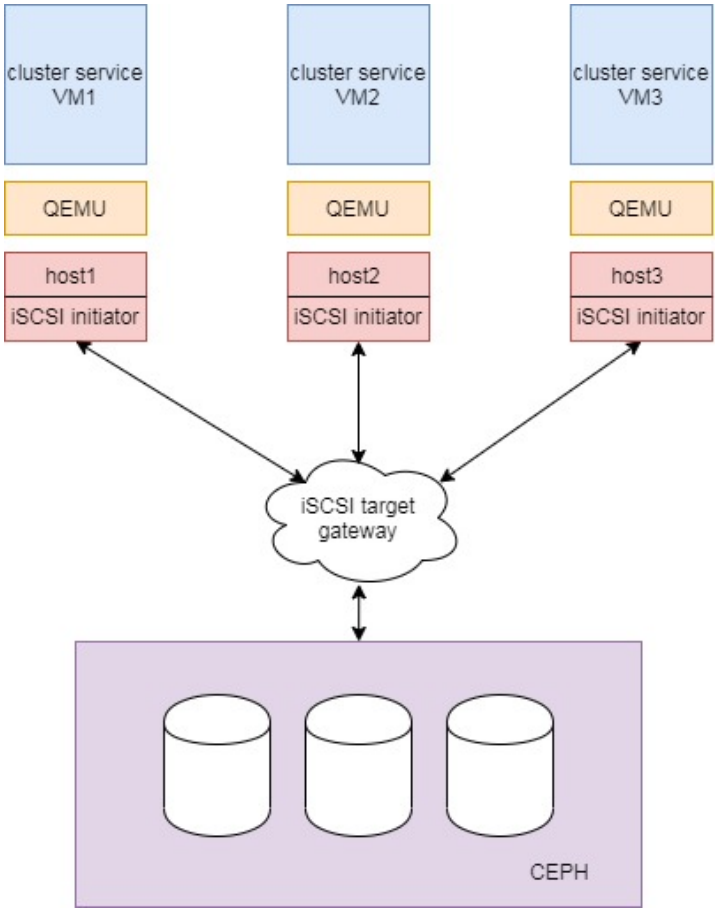
中国移动
China Mobile

Background

Background

- Clustering services, like OCFS, MSCS, share disks with concurrency control mechanisms
- Concurrency control mechanisms are usually implemented via block-level protocols, like iSCSI
- Shared disks are supplied by distributed block storage, like Ceph

Background


cluster service VM1
cluster service VM2
cluster service VM3
QEMU
QEMU
QEMU
host1
iSCSI initiator
host2
iSCSI initiator
host3
iSCSI initiator
iSCSI target gateway
CEPH

Background

中国移动 China Mobile

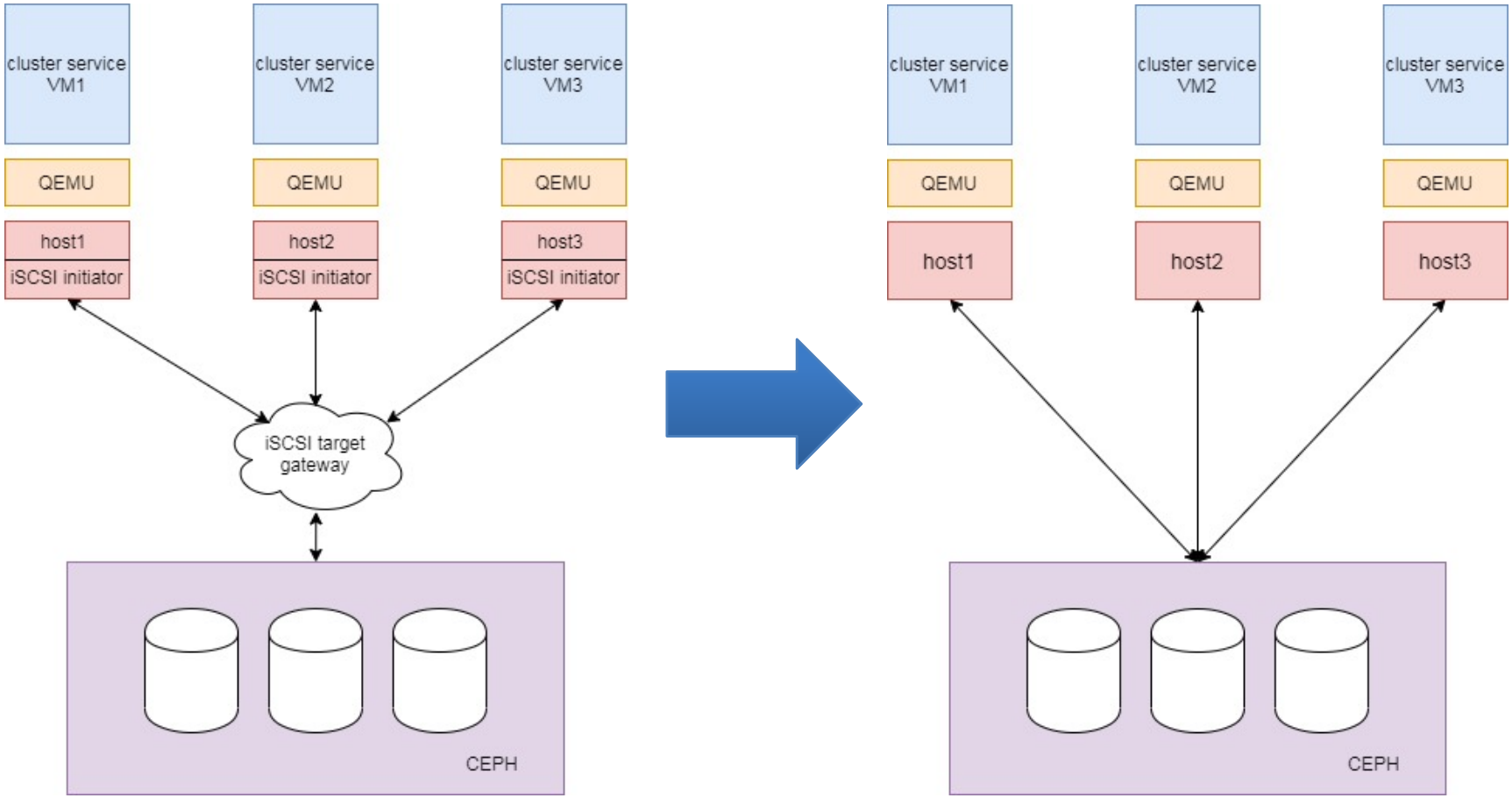
- Problems:
  - Long IO path
  - Immature components
  - Hard to maintain

中国移动
China Mobile

Solution

www.10086.cn

Solution


cluster service VM1
cluster service VM2
cluster service VM3
QEMU
QEMU
QEMU
host1
iSCSI initiator
host2
iSCSI initiator
host3
iSCSI initiator
iSCSI target gateway
CEPH
cluster service VM1
cluster service VM2
cluster service VM3
QEMU
QEMU
QEMU
host1
host2
host3
CEPH

Solution

中国移动 China Mobile

- Work to do
  - SCSI support in Ceph(done)
    - COMPARE AND WRITE(upstream)
    - PERSISTENT RESERVATION(private)
  - SCSI support in QEMU(miss)
    - COMPARE AND WRITE
    - PERSISTENT RESERVATION

中国移动
China Mobile

Howto

www.10086.cn

- SCSI support in QEMU
  - SCSI device emulation
  - Block layer interface
  - Block IO path interface
  - Block driver interface

- SCSI support in QEMU
  - SCSI device emulation
    - hw/scsi/scsi-disk.c
    - COMPARE AND WRITE, PERSISTENT RESERVATION emulation
  - Block layer interface
  - Block IO path interface
  - Block driver interface

- SCSI support in QEMU
  - SCSI device emulation
  - Block layer interface
    - block/block-backend.c
    - reuse blk_aio_pwritev for COMPARE AND WRITE
    - new blk_persistent_reserve_{in,out,check}
  - Block IO path interface
  - Block driver interface

- SCSI support in QEMU
  - SCSI device emulation
  - Block layer interface
  - Block IO path interface
    - block/io.c
    - reuse bdrv_driver_pwritev for COMPARE AND WRITE
    - new bdrv_persistent_reserve_{in,out,check}
  - Block driver interface

- SCSI support in QEMU
  - SCSI device emulation
  - Block layer interface
  - Block IO path interface
  - Block driver interface
    - block/rbd.c
    - new bdrv_aio_compare_and_write and bdrv_co_persistent_reserve_{in,out,check}

中国移动
China Mobile

Status

Status

- Online in our IaaS cloud service
- Patchset(COMPARE AND WRITE part)
  :
  - https://patchwork.ozlabs.org/project/qemu-devel/list/?series=141595

中国移动
China Mobile

Future work

Future work

中国移动 China Mobile

- Upstream
  - CEPH
    - PERSISTENT RESERVATION
  - QEMU
    - WRITE SAME
    - PERSISTENT RESERVATION

中国移动
China Mobile

Q&A

www.10086.cn