Bring SCSI support into QEMU block layer

Yaowei Bai <baiyaowei@cmss.chinamobile.com>
Agenda

• Background
• Solution
• Howto
• Status
• Future work
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
Background

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

• 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
Howto
Howto

• SCSI support in QEMU
  – SCSI device emulation
  – Block layer interface
  – Block IO path interface
  – Block driver interface
Howto

• 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
Howto

- **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
Howto

• 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
Howto

• 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}
Status
Status

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

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