Distributed Filesystem

曹隽诚　李晋

2022 年 11 月 9 日

Distributed Filesystem

NFS

NFS

```
mount -t nfs master:/mnt /mnt
```

Distributed or just Remote

NFS is a remote filesystem, not necessarily a distributed one

> *NFS does not support aggregating the storage resources on multiple systems into a virtual storage pool, instead it only allows accessing remote sotrage resources in a transparent way.*

Distributed Filesystem

GlusterFS

GlusterFS

```
gluster volume create gv replica 2 node0:/data node1:/data
```

Just like on NFS, we see an identical filesystem tree on all participating nodes, but unlike on NFS where all files are located on master, they are distributed or replicated among all nodes.

Distributed Filesystem

Communication

Requirement

A reliable transport for supporting a distributed filesystem

Choices

- stream
- message
- RMA (Remote *Memory* Access)

Distributed Filesystem

Communication

Do not communicate by sharing memory; instead, share memory by communicating.

stream

Explicit connections and states

message

Asynchronous and full of callbacks

RMA

RDMA with an optional D

Distributed Filesystem

RMA

RMA

Effective to implement on modern infrastructures with RDMA
Gracefully fallbacks to stream or message without code change
Naturally maps to NVME-like storage technologies

Addressing

```
0           63            127
| node id | block id |
```

Operation

- read
- write
- discard

RMA

Effectively, a distributed virtual block device
Simply, run an existing filesytem atop

Locality

Normal filesystems have no idea of the distributed natual of the underlying block device, resulting in poor locality and poor performance

Bottomline

Preferably allocate blocks where they are created

A Tale of Two Traits

rcore_fs::vfs::FileSystem

Allows a single filesystem implementation to be used in rCore, zCore and fuse

rcore_fs_dfs::transport::Transport

Allows the distributed filesystem to function reguardless of the underlying network and storage implementation

FUSE

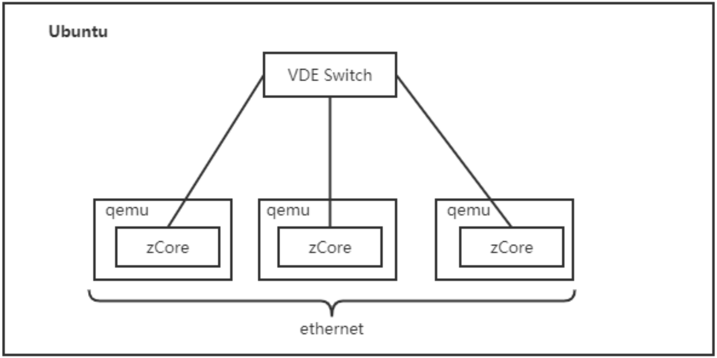
rcore_fs_dfs::transport::loopback::LoopbackTransport
point-to-point RPC-style TCP transport

zCore

linux_object::net::DistriTran
Broadcast TCP transport with rendezvous point

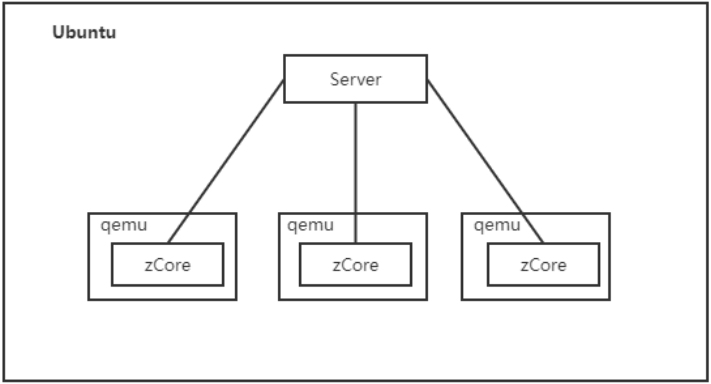
Distributed Filesystem

Topology



Distributed Filesystem

Topology



Distributed Filesystem

Future works

- address consistency issues with locking primitives
- implement more fileystem operations
- automatic rebalance and migration of blocks

Distributed Filesystem

曹隽诚　李晋

2022 年 11 月 9 日