TGT 服务器的优化

块设备协议

- NBD
  - Linux专有块设备协议
- iSCSI
  - 广泛支持的外部设备协议（块，磁带等）

Curve云原生存储支持块设备

- 通过NBD，只支持Linux
- 通过SDK API，目前只支持Linux
  - PFS
- 扩大使用范围
  - 通过iSCSI支持更多系统，例如Windows, 类UNIX系统等，使用两项基础技术
    - TCP/IP
    - SCSI
  - 替代SAN
    - 可靠性、稳定性方面有自己的的特色，使用raft副本一致性和copyset概念可以自动修复损坏的副本，并且可扩容。无论在可靠性、稳定性还是性价比方面都很有优势，使用廉价硬件搭建。

iSCSI软件

- Client端: iscsi initiator，系统自带
  - Linux open-iscsi
  - Windows iSCSI 发起者
- 服务器端
  - 必须是CurveBS原生支持的平台，因为需要curve原生接口，目前是Linux

iSCSI target服务器

- LINUX LILO
  - 一般用于输出内核本地块设备
  - TCMU
    - 作为LILO支持用户态的接口
- 如何评价LILO
  - 输出内核块设备I/O效率高
  - 不利于把复杂的存储协议代码搬进内核，例如(curve, brpc, c++, protobuf等)
  - TCMU多了一层转接，配置过程复杂，业界踩的坑不够多。
  - TCMU的用户态代码会受到框架约束，不够灵活。

iSCSI target 服务器

- TGT(STGT)
  - 比较久的历史，原来叫STGT，后来改成TGT
  - 纯用户态，不与内核绑定
  - 支持复杂的存储系统，例如ceph rbd, sheepdog, glfs
  - 纯C代码，外加一些脚本
  - 完整的源代码和维护工具、手册
  - 编写IO驱动比较容易，容易扩展支持新的存储系统
  - 代码独立，容易编译、调试、修改，适应性强

让TGT支持curve

- 编写curve驱动，底层异步提交I/O，pipeline
  - 利用NEBD PART 1接口，需要与nebdserver运行在同一台机器
- 支持共享打开，两台TGT服务器可以同时打开一个curve卷
  - 让Initiator可以使用multipath
- 支持卷resize
  - 添加新的命令
    - tgtadm --mode logicalunit --op update --tid 1 --lun 1 --params disksize=auto
  - Initiator 重新发送SCSI READ CAPACITY命令
    - Windows 磁盘管理器refresh
    - Linux open-iscsi, iscsiadm --mode node -R

DPO & FUA

- DPO是disable page out的缩写,FUA是force unit access的缩写
- FUA可以让某些文件系统在做写操作时，不需要提交一个SCSI FLUSH COMMAND，提高性能
- 已经修改TGT，让驱动可以声明自己是否支持DPO & FUA
- 由于增加的Curve 驱动没有本地cache，所以DPO & FUA可以turn on.
- sd 0:0:0:0: [sda] Write cache: enabled, read cache: enabled, doesn't support DPO or FUA
  - 这个对于curve驱动，Linux Initiator的dmesg不会显示这个信息

TGT的性能问题

- 性能问题主要体现在不能有效使用多CPU
  - 对多个socket connection，在单线程里做event loop多路复用。
  - 多个target时，如果挂的设备多，一旦客户端请求量大，就会忙不过来。
- 开源界有尝试修改
  - 例如sheepdog的开发者提交过一个patch，但是测试效果不理想，分析原因，event loop依然是瓶颈

对TGT的性能优化

- IO是使用多个epoll 线程，充分发挥多CPU能力
- 当前策略是每个target一个epoll线程，负责Initiator发过来的I/O
  - 好处是各target上的CPU使用由OS负责分配，CPU分配粒度更细
  - 也可以多个卷的lun都分配到一个target上，这样多个卷共享一个target，限制使用一个CPU。
- 管理平面不变。主线程里的事件循环及问题：

  管理面是主线程，登录，增、删、改target,lun,session,connection,params都在主线程，而target epoll 线程也要使用这些数据，多线程冲突，数据一致性问题就来了

对TGT的性能优化（续)

- 为每一个target增加一把锁
  - Target event loop (TEL)线程和管理面线程使用这把锁互斥
  - TEL在运行时锁住这把锁，管理面只能等待，等TEL线程进入epoll wait状态，会释放这把锁，管理面可以增删改target信息。
- 不需要target list lock
  - 因为TEL线程只存取自己负责的target，不存取别的target，所以TEL线程不需要target list lock。
  - 管理面是单线程，只有它遍历target list，没有需要互斥的情况。

```
void tgt_event_loop(struct tgt_evloop *evloop)
{
        int nevent, i, sched_remains, timeout;
        struct epoll_event events[1024];
        struct event_data *tev;

retry:
        sched_remains = tgt_exec_scheduled(evloop);
        timeout = sched_remains ? 0 : -1;

        if (evloop->release)
                evloop->release(evloop);
        nevent = epoll_wait(evloop->ep_fd, events, ARRAY_SIZE(events), timeout);
        if (evloop->acquire)
                evloop->acquire(evloop);
        if (evloop->event_need_refresh) {
                evloop->event_need_refresh = 0;
                goto next;
        }

        if (nevent < 0) {
                if (errno != EINTR) {
                        eprintf("%m\n");
                        exit(1);
                }
        } else if (nevent) {
                for (i = 0; i < nevent; i++) {
                        tev = (struct event_data *) events[i].data.ptr;
                        if (tev != &evloop->async_event)
                                tev->handler(evloop, tev->fd, events[i].events, tev->data);
                        else {
```

FIO性能测试（配置）

- [global]
- rw=randread
- direct=1
- iodepth=128
- ioengine=aio
- bsrange=16k-16k
- runtime=60
- group_reporting
- [disk01]
- filename=/dev/sdx
- [disk02]
- filename=/dev/sdy
- size=10G
- [disk03]
- filename=/dev/sdz
- size=10G

Tgt优化前的FIO性能

```
disk01: (g=0): rw=randread, bs=(R) 16.0KiB-16.0KiB, (W) 16.0KiB-16.0KiB, (T) 16.0KiB-16.0KiB, ioengine=libaio, iodepth=128
disk02: (g=0): rw=randread, bs=(R) 16.0KiB-16.0KiB, (W) 16.0KiB-16.0KiB, (T) 16.0KiB-16.0KiB, ioengine=libaio, iodepth=128
disk03: (g=0): rw=randread, bs=(R) 16.0KiB-16.0KiB, (W) 16.0KiB-16.0KiB, (T) 16.0KiB-16.0KiB, ioengine=libaio, iodepth=128
fio-3.28-95-gfe91-dirty
Starting 3 processes
Jobs: 1 (f=1): [r(1),_(2)][42.3%][r=254MiB/s][r=16.2k IOPS][eta 01m:22s]
disk01: (groupid=0, jobs=3): err= 0: pid=48316: Thu Jan 20 11:28:55 2022
  read: IOPS=38.8k, BW=607MiB/s (636MB/s)(35.5GiB/60009msec)
    slat (nsec): min=1593, max=2052.0k, avg=6594.47, stdev=4672.01
    clat (usec): min=371, max=40825, avg=7867.61, stdev=3619.08
     lat (usec): min=376, max=40834, avg=7874.45, stdev=3619.18
    clat percentiles (usec):
     |  1.00th=[ 1975],  5.00th=[ 2540], 10.00th=[ 3130], 20.00th=[ 4293],
     | 30.00th=[ 5407], 40.00th=[ 6587], 50.00th=[ 7701], 60.00th=[ 8848],
     | 70.00th=[10028], 80.00th=[11207], 90.00th=[12649], 95.00th=[13829],
     | 99.00th=[16188], 99.50th=[17171], 99.90th=[20317], 99.95th=[22938],
     | 99.99th=[30278]
   bw (  KiB/s): min=572736, max=943290, per=100.00%, avg=774234.77, stdev=31664.14, samples=285
   iops        : min=35796, max=58955, avg=48389.65, stdev=1978.98, samples=285
  lat (usec)   : 500=0.01%, 750=0.01%, 1000=0.01%
  lat (msec)   : 2=1.13%, 4=16.45%, 10=52.51%, 20=29.80%, 50=0.11%
  cpu          : usr=10.89%, sys=18.37%, ctx=2291214, majf=0, minf=1572
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.1%
     issued rwts: total=2329408,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=128

Run status group 0 (all jobs):
   READ: bw=607MiB/s (636MB/s), 607MiB/s-607MiB/s (636MB/s-636MB/s), io=35.5GiB (38.2GB), run=60009-60009msec

Disk stats (read/write):
  sdx: ios=1016310/1, merge=5/0, ticks=7647328/0, in_queue=7647768, util=99.86%
  sdy: ios=651362/0, merge=3917/0, ticks=5293560/0, in_queue=5294556, util=99.81%
  sdz: ios=648853/1, merge=3025/0, ticks=5298900/0, in_queue=5299060, util=99.83%
```

Tgt优化后的FIO性能

```
disk01: (g=0): rw=randread, bs=(R) 16.0KiB-16.0KiB, (W) 16.0KiB-16.0KiB, (T) 16.0KiB-16.0KiB, ioengine=libaio, iodepth=128
disk02: (g=0): rw=randread, bs=(R) 16.0KiB-16.0KiB, (W) 16.0KiB-16.0KiB, (T) 16.0KiB-16.0KiB, ioengine=libaio, iodepth=128
disk03: (g=0): rw=randread, bs=(R) 16.0KiB-16.0KiB, (W) 16.0KiB-16.0KiB, (T) 16.0KiB-16.0KiB, ioengine=libaio, iodepth=128
fio-3.28-93-ga87e
Starting 3 processes
Jobs: 1 (f=1): [r(1),_(2)][54.5%][r=649MiB/s][r=41.5k IOPS][eta 00m:50s]
disk01: (groupid=0, jobs=3): err= 0: pid=44616: Thu Jan 20 11:20:57 2022
  read: IOPS=60.9k, BW=951MiB/s (997MB/s)(55.7GiB/60003msec)
    slat (nsec): min=1517, max=2065.0k, avg=5989.93, stdev=5149.54
    clat (usec): min=256, max=33703, avg=3874.12, stdev=2219.19
     lat (usec): min=321, max=33710, avg=3880.21, stdev=2219.64
    clat percentiles (usec):
     |  1.00th=[  816],  5.00th=[ 1123], 10.00th=[ 1418], 20.00th=[ 1958],
     | 30.00th=[ 2507], 40.00th=[ 3032], 50.00th=[ 3556], 60.00th=[ 4080],
     | 70.00th=[ 4621], 80.00th=[ 5276], 90.00th=[ 6587], 95.00th=[ 8160],
     | 99.00th=[11338], 99.50th=[12387], 99.90th=[14222], 99.95th=[15270],
     | 99.99th=[22938]
   bw (  MiB/s): min= 1106, max= 2048, per=100.00%, avg=1441.07, stdev=80.62, samples=220
   iops        : min=70844, max=131128, avg=92228.42, stdev=5159.79, samples=220
  lat (usec)   : 500=0.02%, 750=0.58%, 1000=2.50%
  lat (msec)   : 2=17.53%, 4=37.67%, 10=39.51%, 20=2.17%, 50=0.02%
  cpu          : usr=7.44%, sys=29.25%, ctx=3336855, majf=0, minf=1573
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.1%
     issued rwts: total=3651260,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=128

Run status group 0 (all jobs):
   READ: bw=951MiB/s (997MB/s), 951MiB/s-951MiB/s (997MB/s-997MB/s), io=55.7GiB (59.8GB), run=60003-60003msec

Disk stats (read/write):
  sdx: ios=2334692/1, merge=9/0, ticks=7639992/0, in_queue=7643980, util=99.87%
  sdy: ios=649187/0, merge=2971/0, ticks=3777188/0, in_queue=3778636, util=99.71%
  sdz: ios=646323/1, merge=2394/0, ticks=2655080/0, in_queue=2655904, util=99.58%
```