NVMe Better File System

结题报告

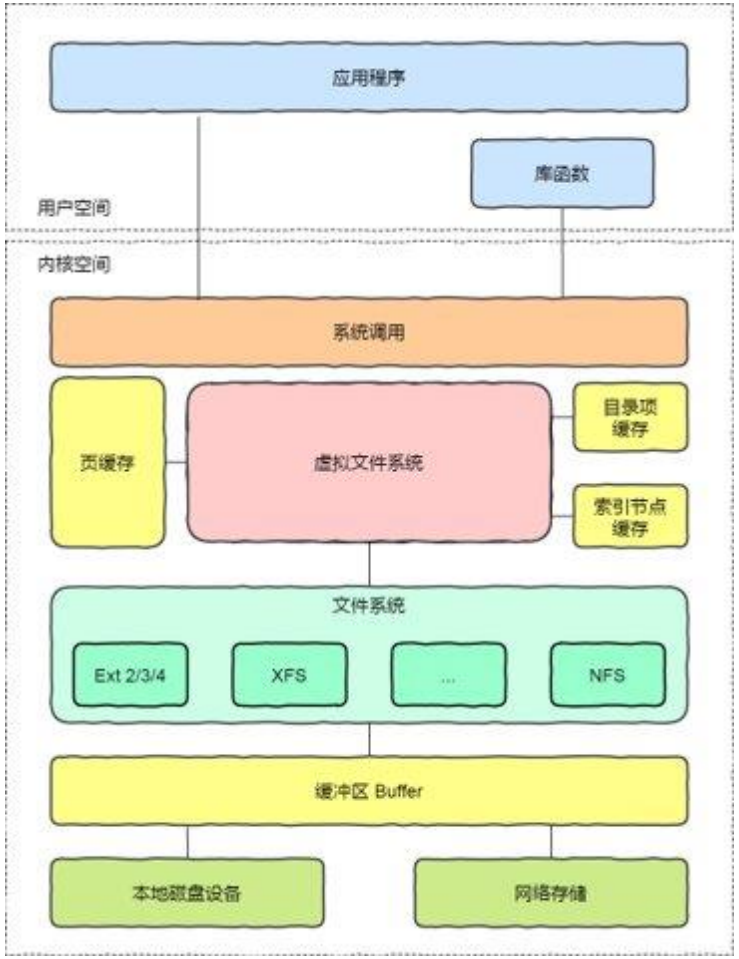
中国科学技术大学
University of Science and Technology of China

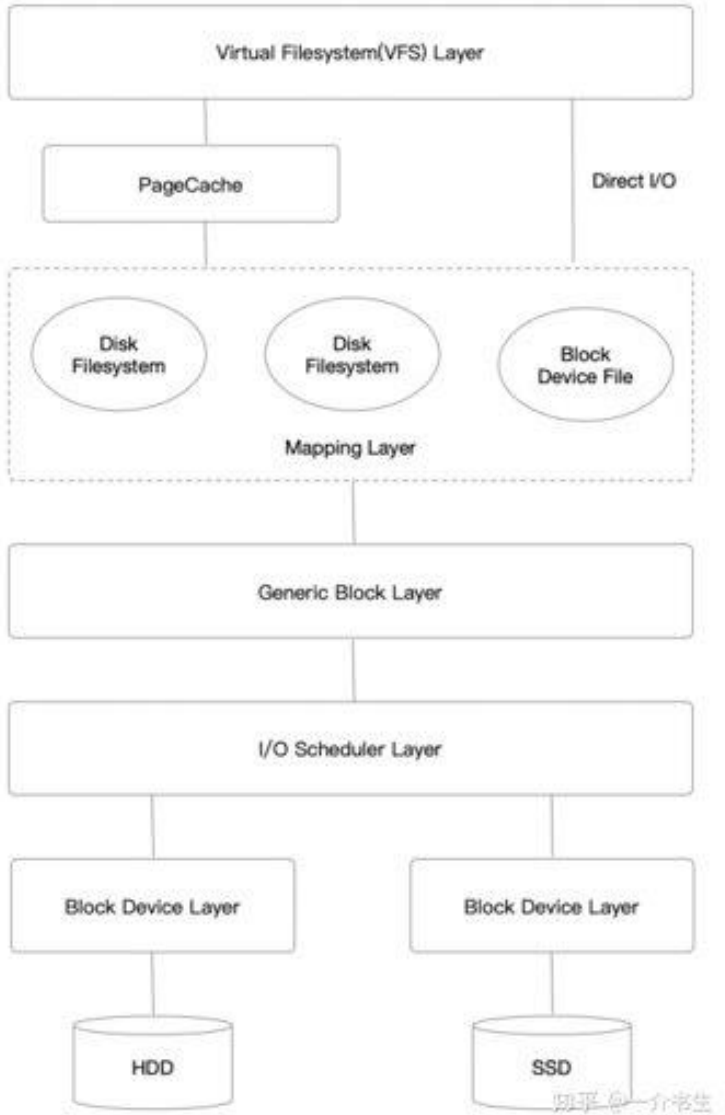
目录

- 背景介绍
- 文件系统架构
- 挂载与测试

背景介绍

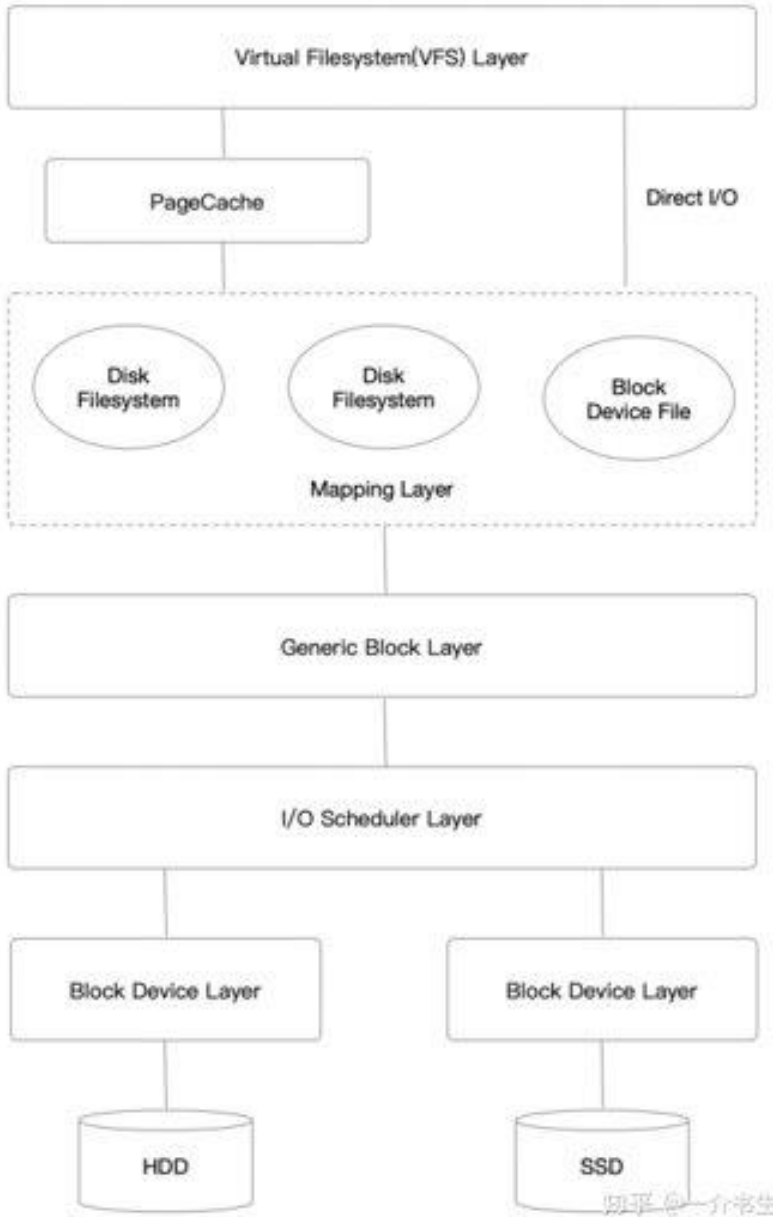
文件系统I/O队列机制


应用程序
库函数
用户空间
内核空间
系统调用
页缓存
虚拟文件系统
目录项缓存
索引节点缓存
文件系统
Ext 2/3/4
XFS
...
NFS
缓冲区 Buffer
本地磁盘设备
网络存储


Virtual Filesystem(VFS) Layer
PageCache
Direct I/O
Disk Filesystem
Disk Filesystem
Block Device File
Mapping Layer
Generic Block Layer
I/O Scheduler Layer
Block Device Layer
Block Device Layer
HDD
SSD

当前文件系统不足

1. 中断
2. 长I/O
3. 对NVMe适配不好



使用SPDK搭建用户态文件系统

- 用户态
- 中断->轮询

背景介绍

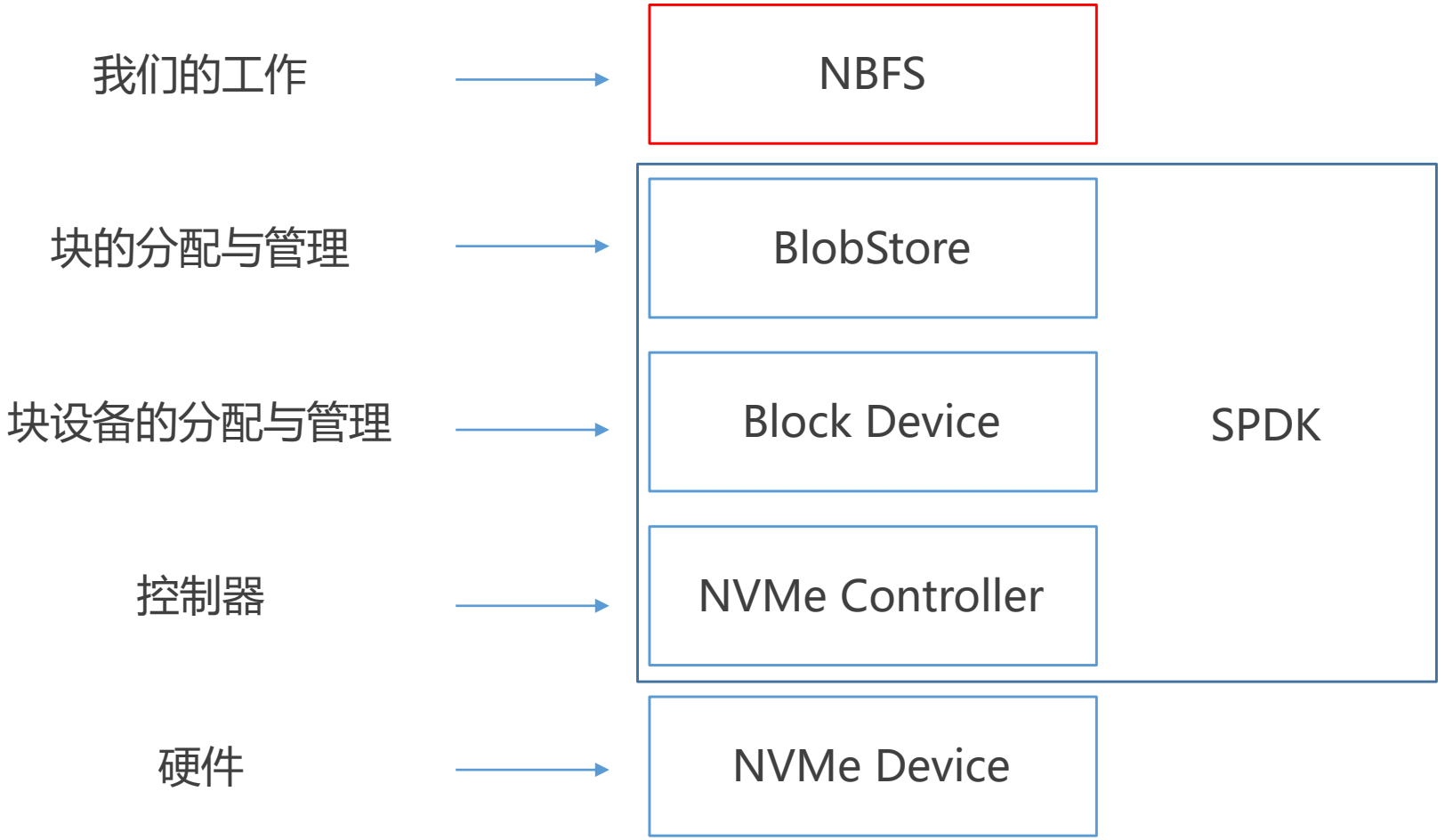
当前基于SPDK文件系统的不足

- BLOBFS是基于SPDK实现的用户态文件系统, 功能较为简单, 没有提供齐全的POSIX接口, 不提供对目录的支持, 查找文件复杂度为O(n)， 不支持原地址更新, 只能被一个进程独享使用。
- BlueStore是一个针对SSD特性优化的存储引擎, 其中包含一个轻量级的文件系统BLUEFS, 并含有用户态块设备抽象层NVMEDevice, 调用SPDK的用户态块设备驱动。 与BLOBFS相似,BLUEFS只提供少量文件接口, 只支持顺序写和两层目录, 只能被一个进程独享使用。
- DashFS是一种利用进程间通信机制的用户态文件系统, 同样是基于SPDK开发, 仅提供简单的文件操作, 不支持目录, 不考虑崩溃一致性, 研究重点在于通过微内核参与的方式实现进程的信任和安全认证, 但缺乏页缓存机制

文件系统架构

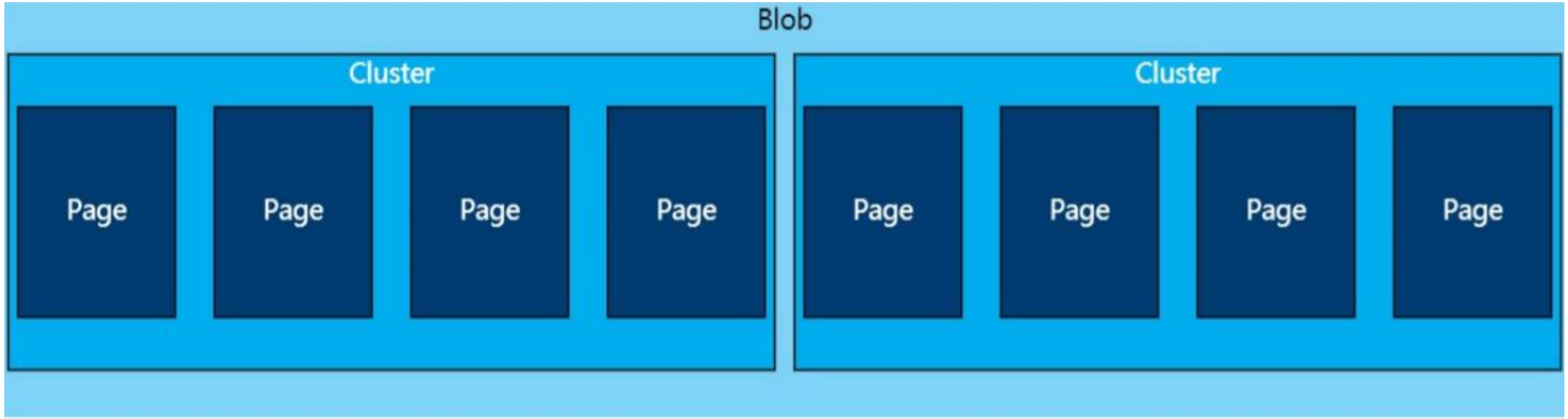
文件系统架构

基本结构


我们的工作
NBFS
块的分配与管理
BlobStore
块设备的分配与管理
Block Device
SPDK
控制器
NVMe Controller
硬件
NVMe Device

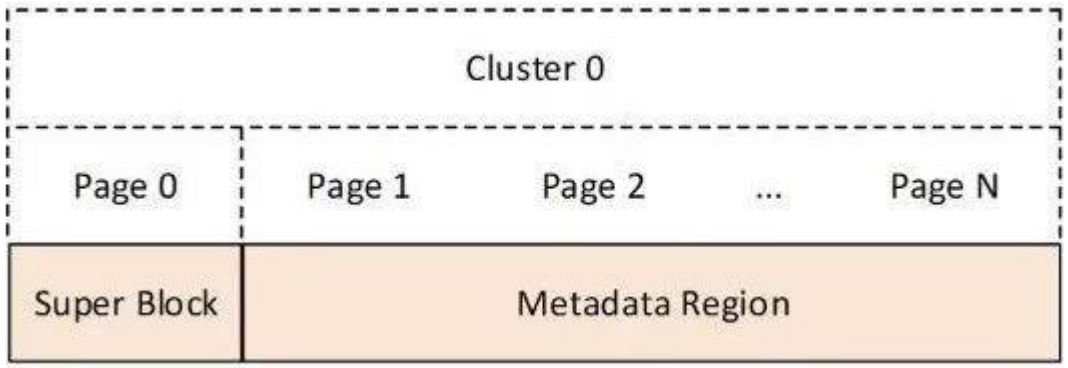
文件系统架构

BlobStore



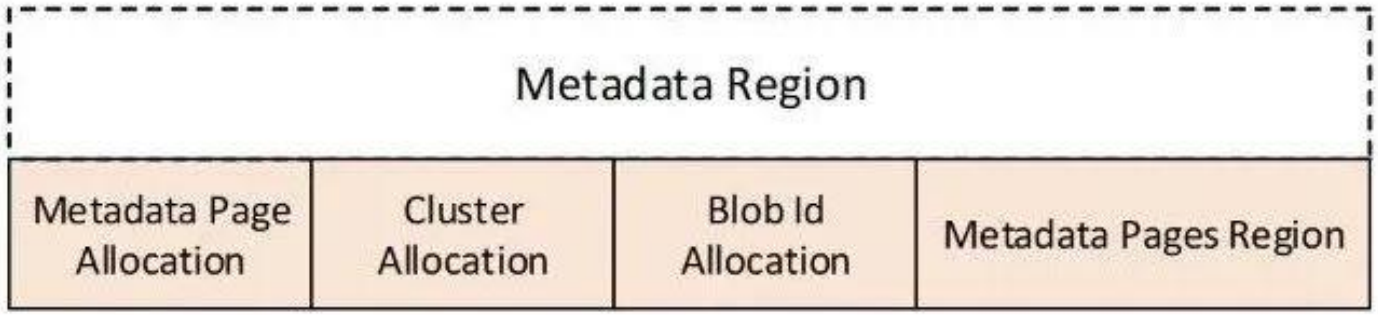
BlobStore

- cluster 0用于存放Blobstore的所有信息以及元数据，对每个blob数据块的查找、分配都是依赖cluster 0中所记录的元数据所进行的。
- Blobstore初始化后的一些基本信息都存放在super block中，例如cluster的大小、已使用page的个数、已使用cluster的个数、Blobstore的大小等信息。



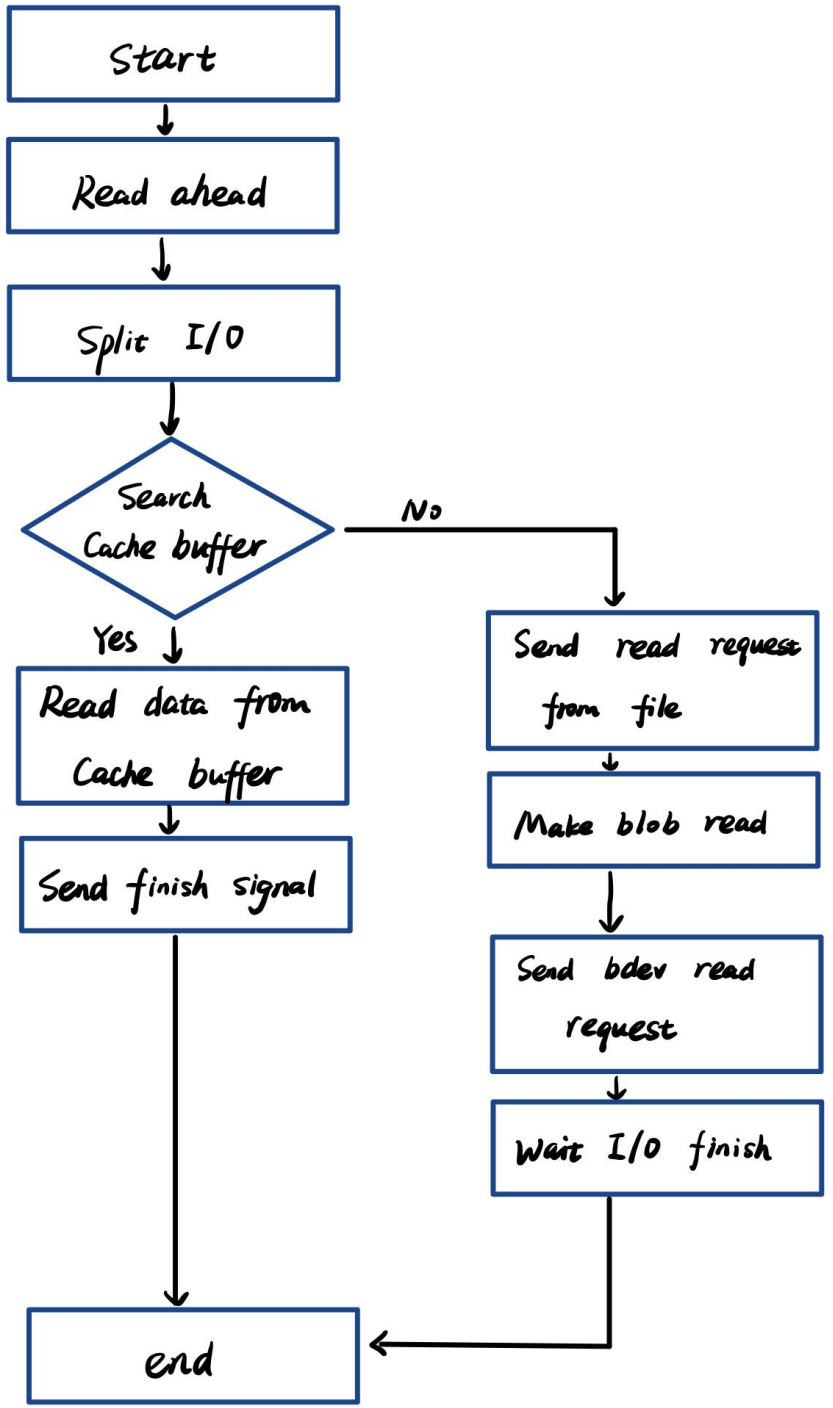
BlobStore

- Metadata Page Allocation：用于记录所有元数据页的分配情况
- Cluster Allocation：用于记录所有cluster的分配情况
- Blob Id Allocation：用于记录blob id的分配情况
- Metadata Pages Region：元数据页区域中存放着每个blob的元数据页。



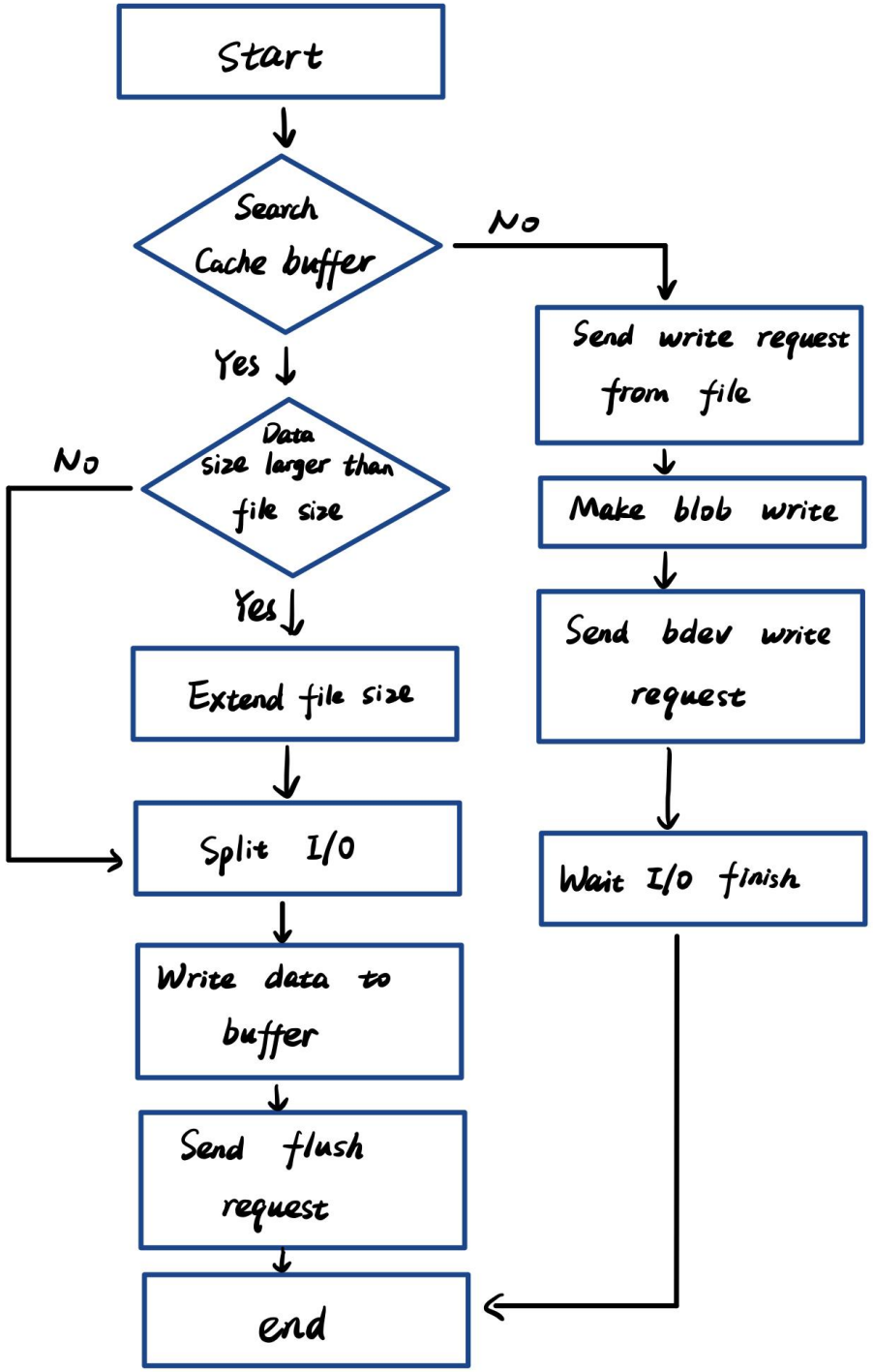
文件系统架构

NBFS读操作



文件系统架构

NBFS写操作



NBFS的优点

- 异步无锁并发
- 采用轮询方式而不是中断
- 直接与NVMe设备通信，规避多次内核态与用户态切换造成的性能损失

挂载与测试

专用接口

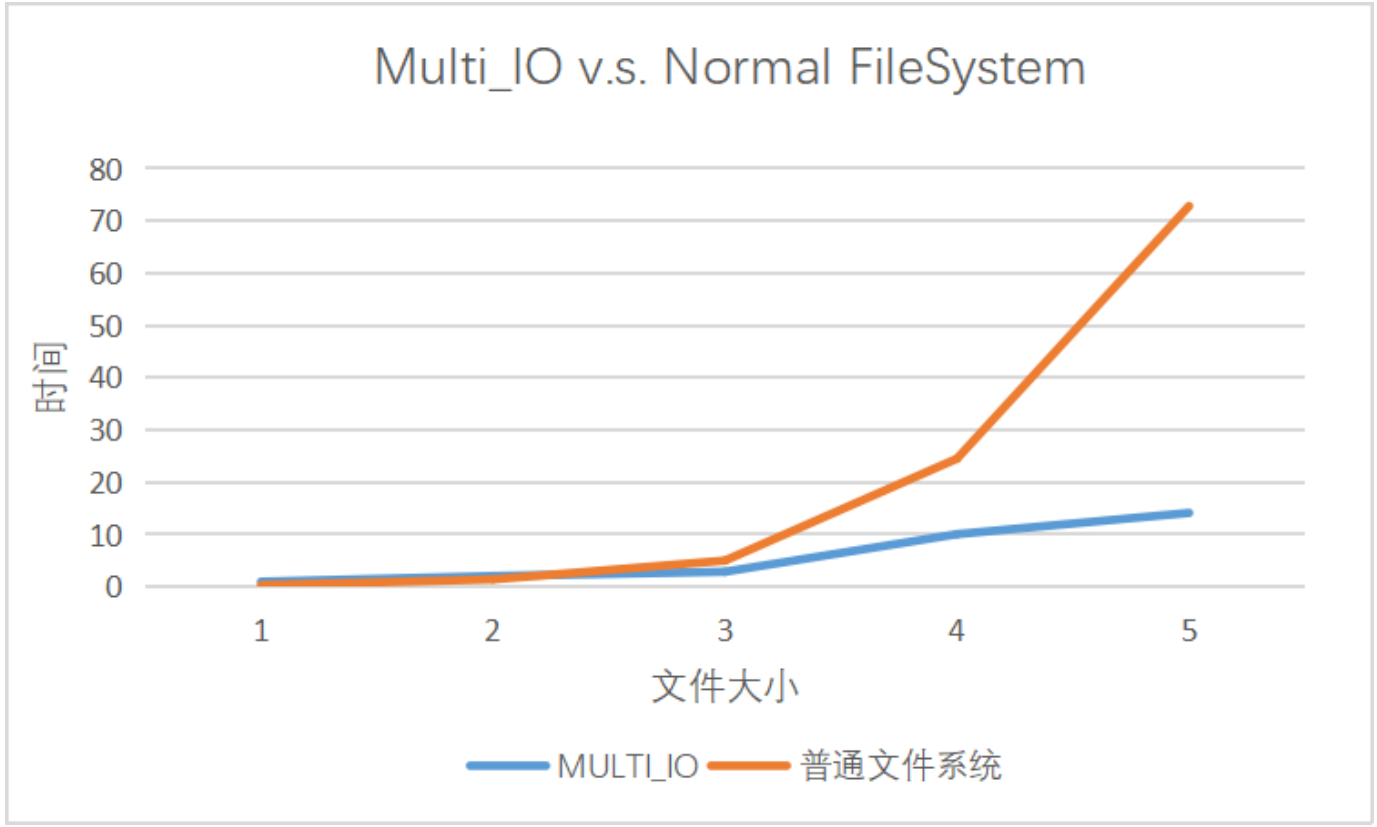
- 绕过多层封装，减小上下文切换开销
- 轻量化，压缩体积
- 使用哈希表的结构保存文件信息

挂载与测试

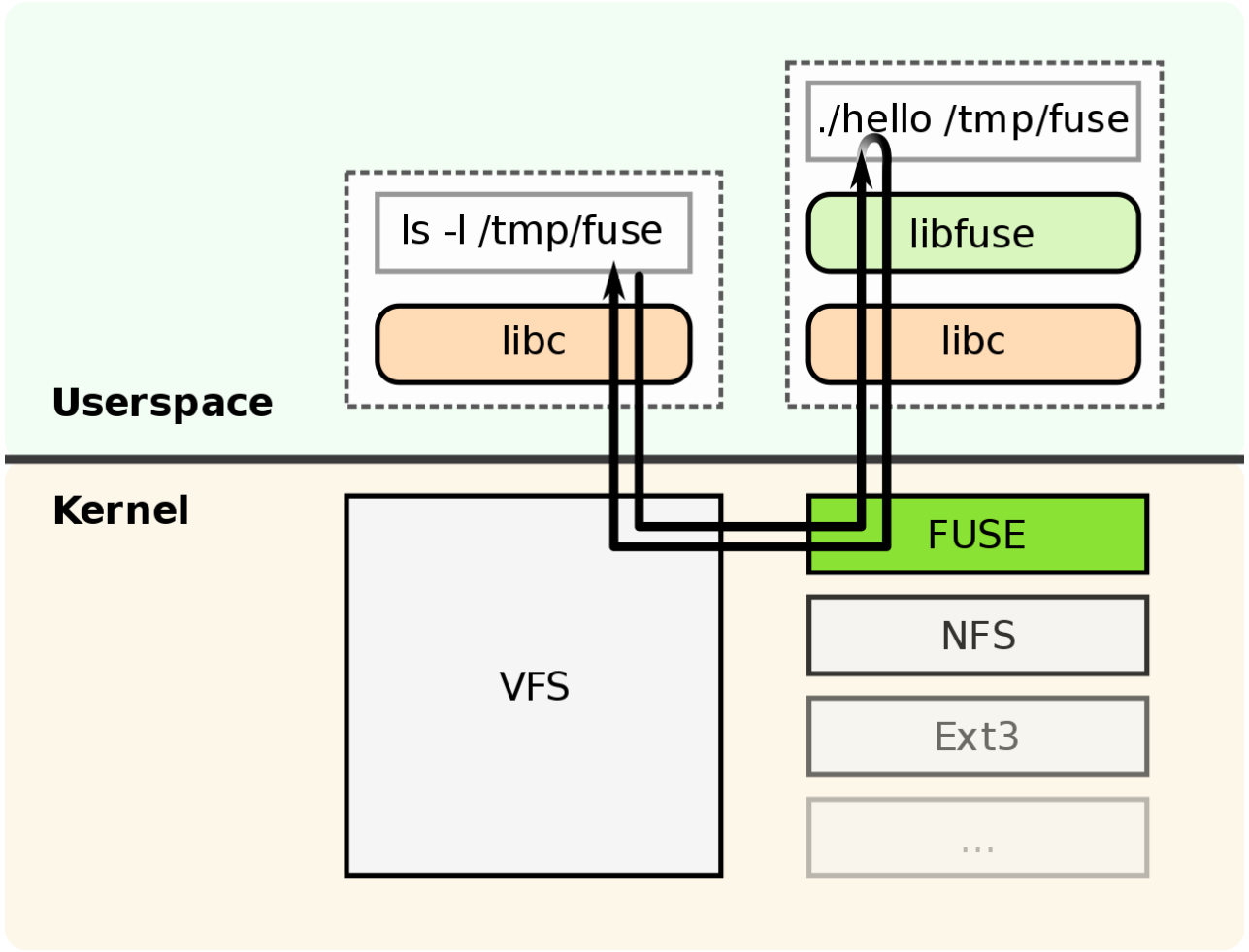
测试

```
[2021-07-08 22:28:50.440401] test_multi_IO.c: 380:main: *NOTICE*: entry
[2021-07-08 22:28:50.440494] Starting SPDK v21.07-pre git sha1 fb68d4e9a / DPDK 20.11.0 initialization...
[2021-07-08 22:28:50.440508] [ DPDK EAL parameters: [2021-07-08 22:28:50.440529] test_multi_io [2021-07-08 22:28:50.440554] --no-shconf [2021-07-08 22:28:50.440561] -c 0x1 [2021-07-08 22:28:50.440567] --log-leve
l=lib.eal:6 [2021-07-08 22:28:50.440574] --log-level=lib.cryptodev:5 [2021-07-08 22:28:50.440580] --log-level=user1:6 [2021-07-08 22:28:50.440607] --iova-mode=pa [2021-07-08 22:28:50.440613] --base-virtaddr=0x20
0000000000 [2021-07-08 22:28:50.440618] --match-allocations [2021-07-08 22:28:50.440623] --file-prefix=spdk_pid31810 [2021-07-08 22:28:50.440628] ]
EAL: No available hugepages reported in hugepages-1048576kB
EAL: No legacy callbacks, legacy socket not created
[2021-07-08 22:28:50.449206] app.c: 535:spdk_app_start: *NOTICE*: Total cores available: 1
[2021-07-08 22:28:50.526316] reactor.c: 929:reactor_run: *NOTICE*: Reactor started on core 0
[2021-07-08 22:28:50.526805] accel_engine.c: 853:spdk_accel_engine_initialize: *NOTICE*: Accel engine initialized to use software engine.
[2021-07-08 22:28:50.629099] test_multi_IO.c: 357:test_start: *NOTICE*: entry
[2021-07-08 22:28:51.484519] test_multi_IO.c: 303:bs_init_complete: *NOTICE*: entry
[2021-07-08 22:28:51.484539] test_multi_IO.c: 312:bs_init_complete: *NOTICE*: blobstore: 0x56064b07eb40
[2021-07-08 22:28:51.484547] test_multi_IO.c: 293:create_blob: *NOTICE*: entry
[2021-07-08 22:28:51.484556] test_multi_IO.c: 293:create_blob: *NOTICE*: entry
[2021-07-08 22:28:51.484583] test_multi_IO.c: 275:blob_create_complete: *NOTICE*: entry
[2021-07-08 22:28:51.484589] test_multi_IO.c: 284:blob_create_complete: *NOTICE*: new blob id 4294967296
[2021-07-08 22:28:51.484621] test_multi_IO.c: 275:blob_create_complete: *NOTICE*: entry
[2021-07-08 22:28:51.484628] test_multi_IO.c: 284:blob_create_complete: *NOTICE*: new blob id 4294967297
[2021-07-08 22:28:51.484653] test_multi_IO.c: 257:open_complete: *NOTICE*: entry
[2021-07-08 22:28:51.484917] test_multi_IO.c: 257:open_complete: *NOTICE*: entry
[2021-07-08 22:28:51.484950] test_multi_IO.c: 246:resize_complete: *NOTICE*: resized blob now has USED clusters of 8192
[2021-07-08 22:28:51.485324] test_multi_IO.c: 246:resize_complete: *NOTICE*: resized blob now has USED clusters of 8192
[2021-07-08 22:28:51.485779] test_multi_IO.c: 222:sync_complete: *NOTICE*: entry
[2021-07-08 22:28:51.485789] test_multi_IO.c: 190:blob_write: *NOTICE*: entry
[2021-07-08 22:28:51.486147] test_multi_IO.c: 222:sync_complete: *NOTICE*: entry
[2021-07-08 22:28:51.486159] test_multi_IO.c: 190:blob_write: *NOTICE*: entry
[2021-07-08 22:28:51.489697] test_multi_IO.c: 176:write_complete: *NOTICE*: entry
[2021-07-08 22:28:51.489711] test_multi_IO.c: 139:read_blob: *NOTICE*: entry
[2021-07-08 22:28:51.490155] test_multi_IO.c: 176:write_complete: *NOTICE*: entry
[2021-07-08 22:28:51.490167] test_multi_IO.c: 166:cal_time: *NOTICE*: multi io write using:0.005620175
[2021-07-08 22:28:51.490173] test_multi_IO.c: 139:read_blob: *NOTICE*: entry
[2021-07-08 22:28:51.492804] test_multi_IO.c: 114:read_complete: *NOTICE*: entry
[2021-07-08 22:28:51.492816] test_multi_IO.c: 130:read_complete: *NOTICE*: read SUCCESS and data matches!
[2021-07-08 22:28:51.492824] test_multi_IO.c:  96:delete_blob: *NOTICE*: entry
[2021-07-08 22:28:51.493100] test_multi_IO.c: 114:read_complete: *NOTICE*: entry
[2021-07-08 22:28:51.493108] test_multi_IO.c: 130:read_complete: *NOTICE*: read SUCCESS and data matches!
[2021-07-08 22:28:51.493114] test_multi_IO.c:  96:delete_blob: *NOTICE*: entry
[2021-07-08 22:28:51.495258] test_multi_IO.c:  80:delete_complete: *NOTICE*: entry
[2021-07-08 22:28:51.497144] test_multi_IO.c:  80:delete_complete: *NOTICE*: entry
[2021-07-08 22:28:51.497337] test_multi_IO.c:  37:unload_complete: *NOTICE*: entry
[2021-07-08 22:28:51.541789] test_multi_IO.c: 398:main: *NOTICE*: SUCCESS!
```

测试


Multi_IO v.s. Normal FileSystem
80
70
60
50
40
30
20
10
0
时间
1
2
3
4
5
文件大小
MULTI_IO
普通文件系统

FUSE (Filesystem in User Space)



FUSE接口

- 作为FUSE(File System in User Space)挂载，利用VFS作为普通文件系统使用
- 兼容性好，程序可以直接使用
- 使用B+树结构保存文件信息

挂载与测试

挂载FUSE

```
root@cyq-Mi-Gaming-Laptop-15-6:/home/cyq/spdk# HUGEMEM=5120 scripts/setup.sh
0000:07:00.0 (1179 011a): Active mountpoints on nvme1n1:nvme1n1p1,nvme1n1:nvme1n1p6, so not binding PCI dev
0000:06:00.0 (144d a809): nvme -> uio_pci_generic
root@cyq-Mi-Gaming-Laptop-15-6:/home/cyq/spdk# scripts/gen_nvme.sh --json-with-subsystems > rocksdb.json
root@cyq-Mi-Gaming-Laptop-15-6:/home/cyq/spdk# mkfs/mkfs rocksdb.json Nvme0n1
[2021-07-09 00:15:56.109332] Starting SPDK v21.07-pre git sha1 fb68d4e9a / DPDK 20.11.0 initialization...
[2021-07-09 00:15:56.109513] [ DPDK EAL parameters: [2021-07-09 00:15:56.109549] spdk_mkfs [2021-07-09 00:15:56.109576]
 --no-shconf [2021-07-09 00:15:56.109600] -c 0x3 [2021-07-09 00:15:56.109625] --log-level=lib.eal:6 [2021-07-09 00:15:5
6.109650] --log-level=lib.cryptodev:5 [2021-07-09 00:15:56.109675] --log-level=user1:6 [2021-07-09 00:15:56.109702] --i
ova-mode=pa [2021-07-09 00:15:56.109727] --base-virtaddr=0x200000000000 [2021-07-09 00:15:56.109752] --match-allocation
s [2021-07-09 00:15:56.109776] --file-prefix=spdk_pid2342 [2021-07-09 00:15:56.109801] ]
EAL: No available hugepages reported in hugepages-1048576kB
EAL: No legacy callbacks, legacy socket not created
[2021-07-09 00:15:56.121803] app.c: 535:spdk_app_start: *NOTICE*: Total cores available: 2
[2021-07-09 00:15:56.180508] reactor.c: 929:reactor_run: *NOTICE*: Reactor started on core 1
[2021-07-09 00:15:56.180551] reactor.c: 929:reactor_run: *NOTICE*: Reactor started on core 0
[2021-07-09 00:15:56.181065] accel_engine.c: 853:spdk_accel_engine_initialize: *NOTICE*: Accel engine initialized to us
e software engine.
Initializing filesystem on bdev Nvme0n1...
Welcome to use Nvme Better File Systerm
Refreshing the filesysterm ...
done.
```

挂载与测试

挂载FUSE

```
root@cyq-Mi-Gaming-Laptop-15-6:/home/cyq/spdk# fuse/fuse rocksdb.json Nvme0n1 /mnt/fuse
[2021-07-09 00:17:15.644400] Starting SPDK v21.07-pre git sha1 fb68d4e9a / DPDK 20.11.0 initialization...
[2021-07-09 00:17:15.644568] [ DPDK EAL parameters: [2021-07-09 00:17:15.644618] spdk_fuse [2021-07-09 00:17:15.644642]
 --no-shconf [2021-07-09 00:17:15.644668] -c 0x3 [2021-07-09 00:17:15.644693] --log-level=lib.eal:6 [2021-07-09 00:17:1
5.644724] --log-level=lib.cryptodev:5 [2021-07-09 00:17:15.644764] --log-level=user1:6 [2021-07-09 00:17:15.644800] --i
ova-mode=pa [2021-07-09 00:17:15.644840] --base-virtaddr=0x200000000000 [2021-07-09 00:17:15.644878] --match-allocation
s [2021-07-09 00:17:15.644916] --file-prefix=spdk_pid2349 [2021-07-09 00:17:15.644945] ]
EAL: No available hugepages reported in hugepages-1048576kB
EAL: No legacy callbacks, legacy socket not created
[2021-07-09 00:17:15.677299] app.c: 535:spdk_app_start: *NOTICE*: Total cores available: 2
[2021-07-09 00:17:15.735866] reactor.c: 929:reactor_run: *NOTICE*: Reactor started on core 1
[2021-07-09 00:17:15.735904] reactor.c: 929:reactor_run: *NOTICE*: Reactor started on core 0
[2021-07-09 00:17:15.736388] accel_engine.c: 853:spdk_accel_engine_initialize: *NOTICE*: Accel engine initialized to us
e software engine.
Mounting filesystem on bdev Nvme0n1 to path /mnt/fuse...

Welcome to use Nvme Better File Systerm
done.
[2021-07-09 00:17:15.905079] blobfs_fuse.c: 267:fuse_loop_new_thread: *NOTICE*: Start to loop blobfs on bdev Nvme0n1 mo
unted at /mnt/fuse
```

挂载与测试

挂载FUSE

```
root@cyq-Mi-Gaming-Laptop-15-6:/# cd mnt/fuse
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# echo HelloWorld>1.txt
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# echo HelloWorld>2.txt
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# echo HelloWorld>3.txt
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# echo HelloWorld>aaa
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# echo HelloWorld>bbb
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# echo HelloWorld>ddd
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# echo HelloWorld>ccc
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# ls
1.txt  2.txt  3.txt  aaa  bbb  ccc  ddd
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# rm ccc
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# mv 2.txt eee
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# ls
1.txt  3.txt  aaa  bbb  ddd  eee
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# rm 1.txt
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# rm 3.txt
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# rm eee
root@cyq-Mi-Gaming-Laptop-15-6:/mnt/fuse# ls
aaa  bbb  ddd
```

谢谢！