溢思得瑞科技创新集团
ISTUARY INNOVATION GROUP

基于全闪存的块设备解决方案

Pan Liu

2016.12.11

www.istuary.com

内容

- 公司、个人简介
- 硬件简单拓扑结构
- SPDK ISCSI
- NBD flow
- BlueStore

公司简介

- **Istuary Innovation Group** cultivates innovation into businesses for launch in fast-growing markets globally.
- Istuary partners with innovators, entrepreneurs, technical talent and experts in North America and China to build technology companies for fast-growing markets, such as China, and to capitalize on quickly expanding and unmet demand.
- **Hope(得瑞厚朴)** is a wholly owned subsidiary of Istuary. Our business focuses on full stack flash storage solution:

➢ Storage software
➢ PCIE Flash Chip
➢ PCIE Flash Card
➢ Storage Server

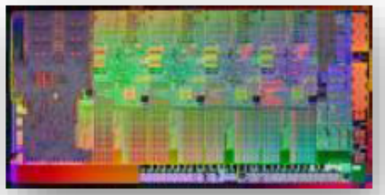
个人简介

- My name is Pan Liu.
- R&D director of storage software.
- Based in Beijing.
- pan.liu@istuary.com

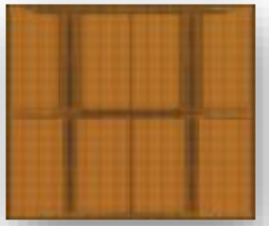
3D XPoint™ 技术
弥补内存和存储间的鸿沟

STORAGE

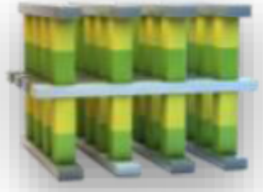
SRAM
时延: 1X
数据容量: 1X



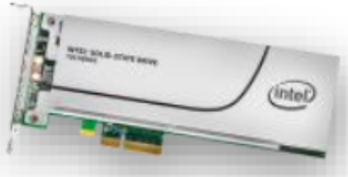
DRAM
时延: ~10X
数据容量: ~100X



3D XPoint™ 存储介质
时延: ~100X
数据容量: ~1,000X



NAND 固态盘
时延: ~100,000X
数据容量: ~1,000X



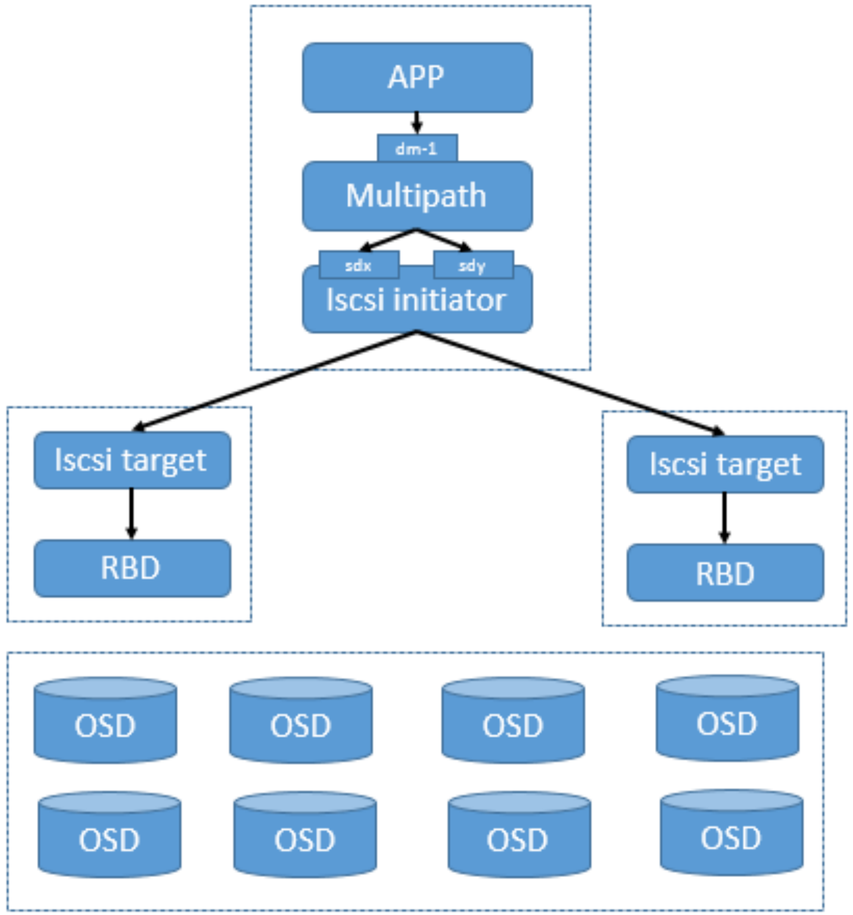
硬盘
时延: ~10 MillionX
数据容量: ~10,000X

MEMORY

Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications.

IDF16 英特尔信息技术峰会

ISCSI->RBD Gateway



- Virtual machine providers and cloud software of many types can speak iSCSI.
- If Ceph could export storage as an iSCSI device, it would be easy to glue all those providers to a Ceph cluster.

ISCSI->RBD Gateway

IO 模式 / 测试模式	4k-randw-256iodepth(iops)	4k-randr-256iodepth(iops)
SPDK iscsi + kernel nvme	117K	140K
SPDK iscsi + userspace nvme	138K	173K
Targetcli iscsi + kernel nvme	100K	145K

- CPU: Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz
- Only enable Core 0
- So we choose to consider SPDK iscsi

ISCSI->RBD Gateway

Ceph server

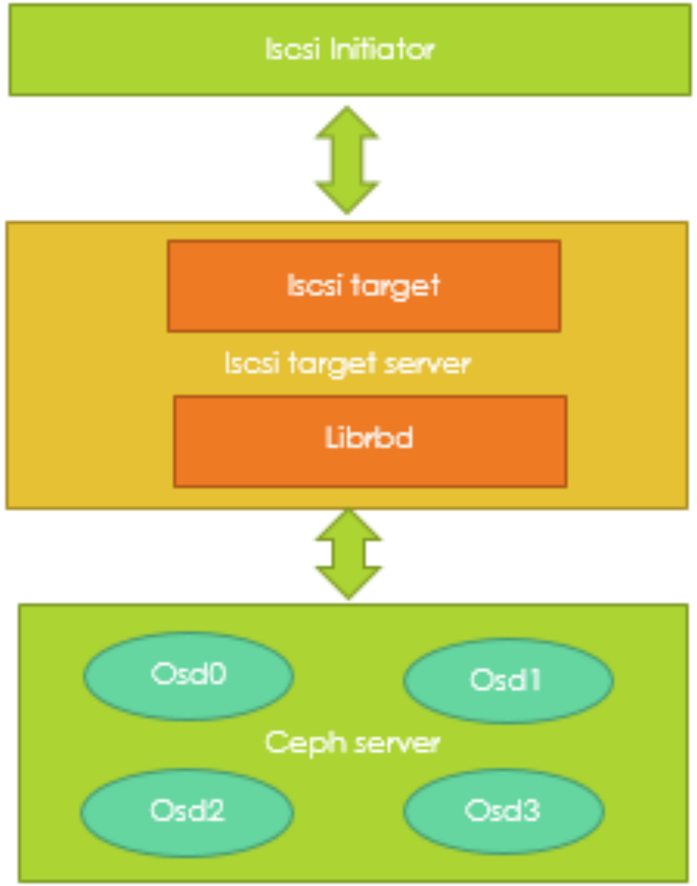
- CPU:Intel(R) Xeon(R) CPU E5-2660 v4 @2.00GHz
- Four intel P3700 SSDs
- One OSD on each SSD, total 4 osds
- 4 pools PG number 512, one 10G image in one pool

Iscsi target server (Librbd+SPDK / Librbd+tgt)

- CPU:Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz
- Only one core enable

Iscsi initiator

- CPU:Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz



ISCSI->RBD Gateway

◆ Two Core

Fio+img / iscsi type+op	1 fio + 1 img (iops)	2 fio + 2 img (iops)	3 fio + 3 img (iops)	4 fio + 4 img (iops)	Spdk/tgt ratio
Tgt +read	12K	24K	26K	26K	181%
Spdk iscsi + read	37K	47K	47K	47K	
Tgt +write	9.5K	13.5K	19K	22K	123%
Spdk iscsi + write	16K	24K	25K	27K	

代码贡献

- #53 Make a wrapper that spdk can call a function without thread affinity, and call this wrapper to open rbd image.
- #48 iscsi: fix comment issue. If not specified reactor mask , we only use core 0
- #41 Fix some cppcheck errors on lib/bdev & lib/iscsi & lib/scsi.
- Add/remove ISCSI LUN from rbd image dynamically.

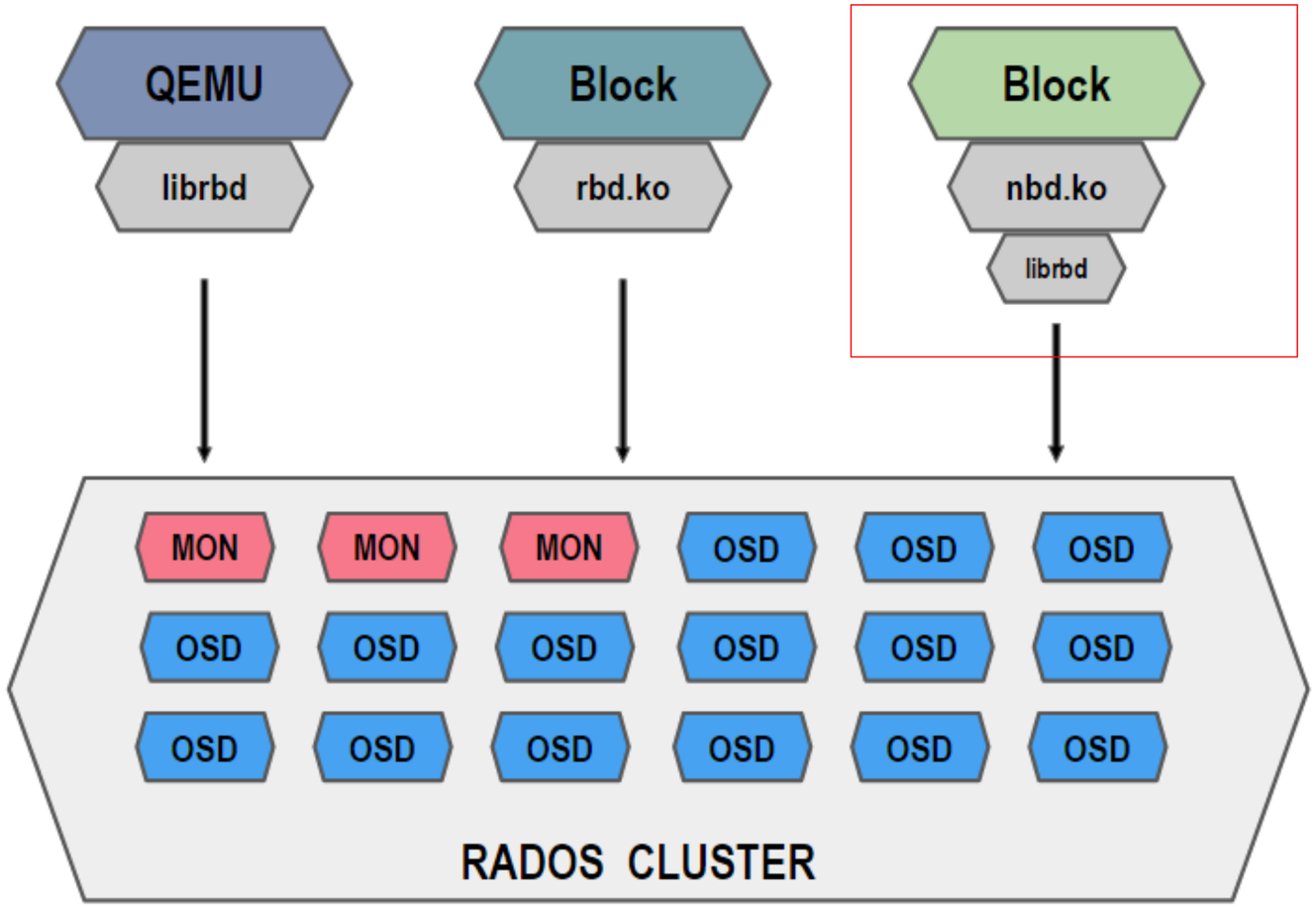
SPDK ISCSI期待的改进

- If master core down?
- Hot Plugin
- While(1) or polling
- Dynamically add/remote iscsi LUN

ISCSI gateway方案的缺陷

- 需要额外的ISCSI gateway
- 如果把ISCSI gateway放在存储服务器内部，则CPU占用过多。
- 如果把ISCSI gateway放在存储服务器外部，则需要用户额外的成本，包括机房空间。
- ISCSI gateway，会有额外的性能损耗。

Network Block Device



用户态librbd的优点

- Kernel rbd非常不稳定，甚至ceph的官网，也说了kernel rbd在一些linux kernel版本下，有dead lock的问题。
- Kernel rbd涉及到内核版本兼容问题
- librbd的功能，是多于Kernel rbd的。

Issues - NBD

- Cannot be parted by default.
  - https://github.com/ceph/ceph/pull/12379
  - https://github.com/ceph/ceph/pull/12259
- Fix log and sock issue.
  - https://github.com/ceph/ceph/pull/12433
- Support auto map after reboot.
- Support list-mapped.

性能比较

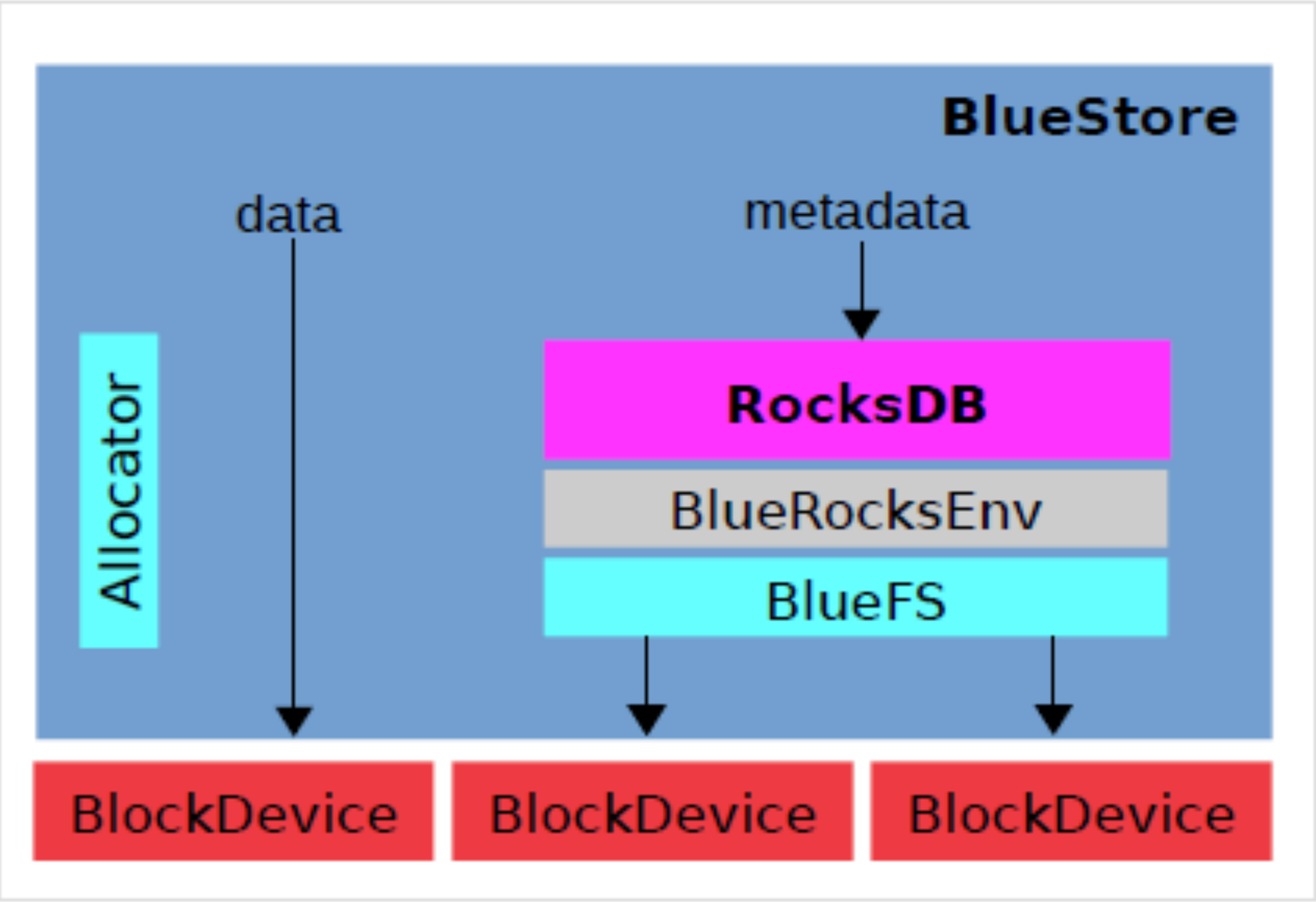
跟iscsi方案比较：

Summary		
op \ data	Single stream FIO	Ratio(ISCSI/NBD)
SPDK ISCSI 4K randwrite	41.8K	196.10%
NBD 4K randwrite	82.0K	
SPDK ISCSI 4K randread	84.0K	138.30%
NBD 4K randread	116.2K	

跟fio+librbd比较

- 读性能： 82%
- 写性能： 90%

BlueStore


BlueStore
data
metadata
Allocator
RocksDB
BlueRocksEnv
BlueFS
BlockDevice
BlockDevice
BlockDevice

BlueStore

```
[osd.0]
        host = ubu-machine
        osd data = /var/lib/ceph/osd/osd-device-0-data
        bluestore block wal path = /dev/disk/by-partlabel/osd-device-0-wal
        bluestore block db path = /dev/disk/by-partlabel/osd-device-0-db
        bluestore block path = /dev/disk/by-partlabel/osd-device-0-block
```

```
nvme0n1                 259:5    0 372.6G  0 disk
├─nvme0n1p1             259:1    0   9.5M  0 part
├─nvme0n1p2             259:2    0   1.9G  0 part
├─nvme0n1p3             259:3    0   7.6G  0 part
└─nvme0n1p4             259:4    0 336.3G  0 part
nvme1n1                 259:0    0 372.6G  0 disk
├─nvme1n1p1             259:6    0   9.5M  0 part
├─nvme1n1p2             259:7    0   1.9G  0 part
├─nvme1n1p3             259:8    0   7.6G  0 part
└─nvme1n1p4             259:9    0 336.3G  0 part
```

Q && A