When Ceph Meets DDPK

Company: XSKY
Title: Technical Director
Name: Haomai Wang

About

- I'm Haomai Wang
- Work at XSKY
- Active Ceph Developer
- Maintain AsyncMessenger and NVMEDevice module in Ceph
- haomaiwang@gmail.com

Outline

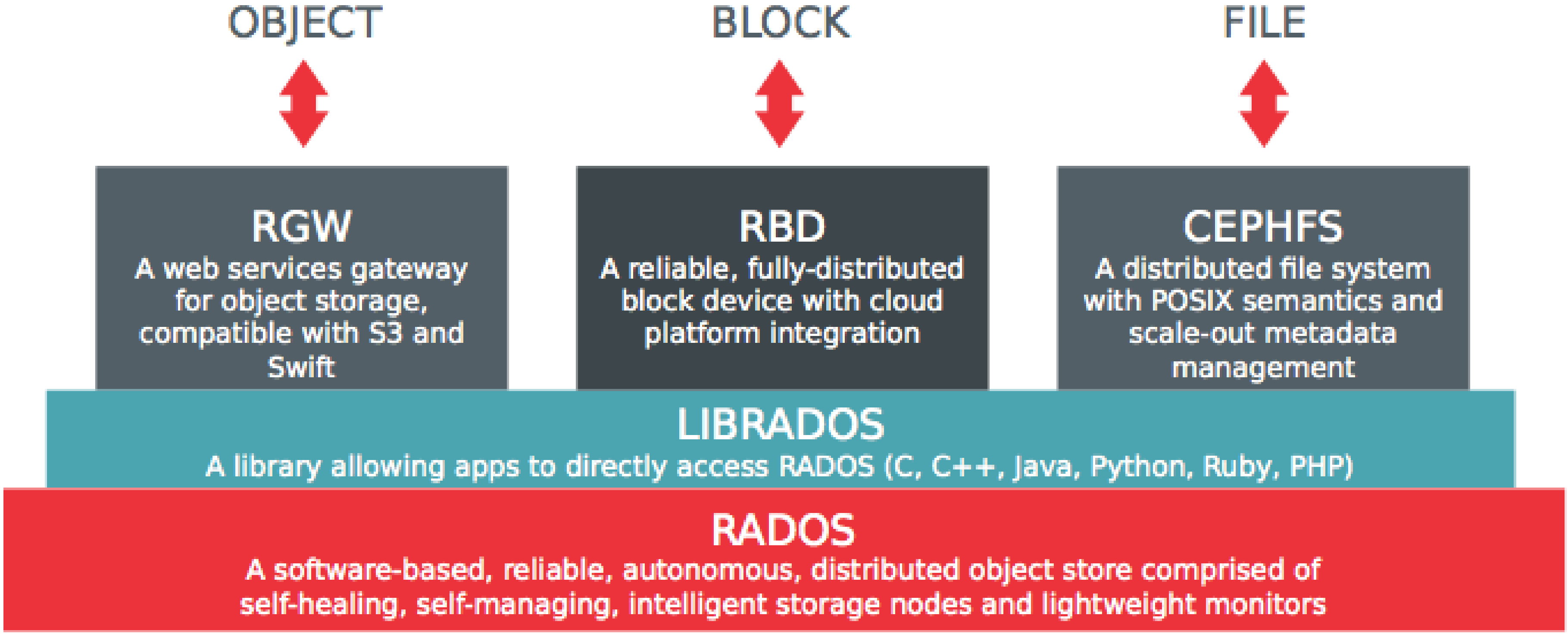
- What is Ceph?
- High performance gap
- Ceph + DPDK
- Future work

What is Ceph?

- Object, block, and file storage in a single cluster
- All components scale horizontally
- No single point of failure
- Hardware agnostic, commodity hardware
- Self-manage whenever possible
- Open source (LGPL)

- “A Scalable, High-Performance Distributed File System”
  “performance, reliability, and scalability”

Ceph Components



OBJECT

BLOCK

FILE

RGW
A web services gateway for object storage, compatible with S3 and Swift

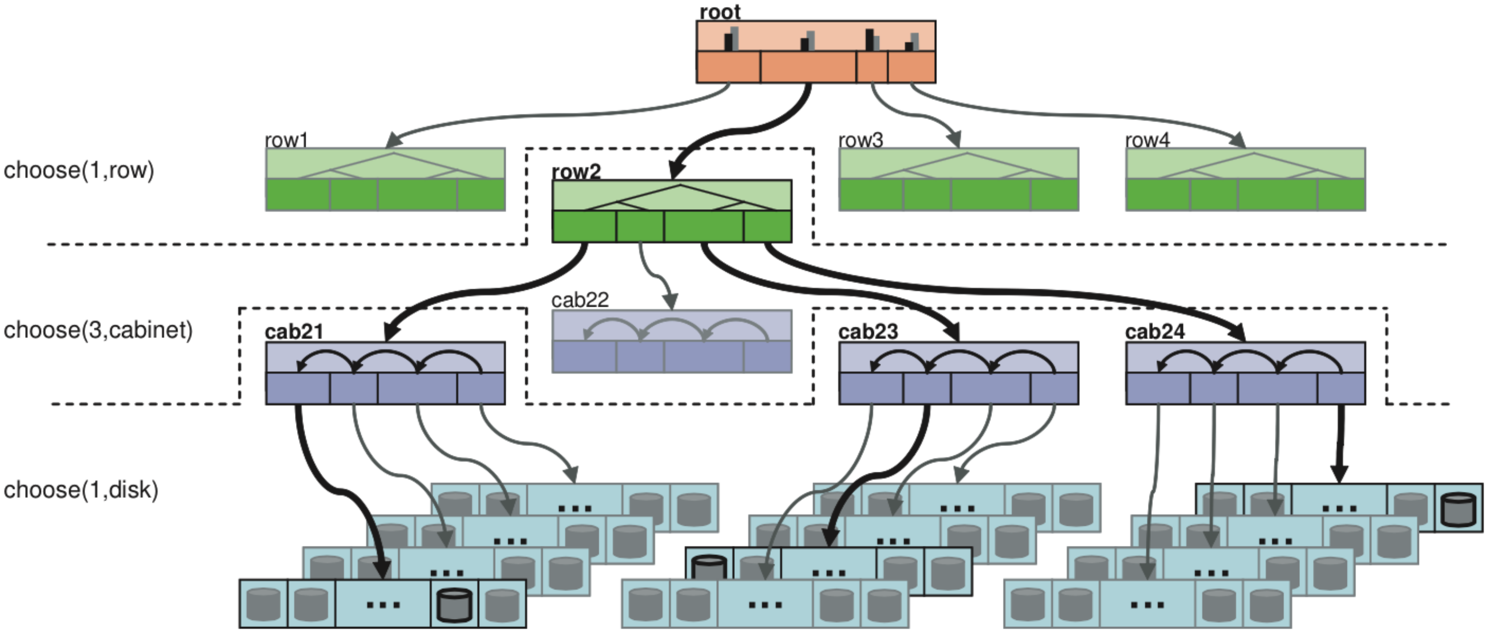
RBD
A reliable, fully-distributed block device with cloud platform integration

CEPHFS
A distributed file system with POSIX semantics and scale-out metadata management

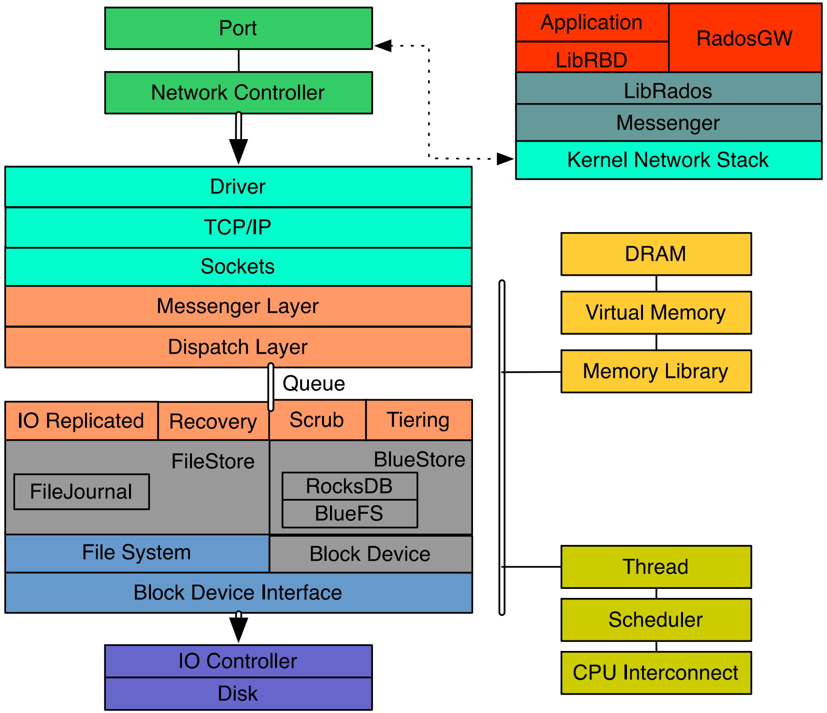
LIBRADOS
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)

RADOS
A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

Crush—Data Placement Algorithm

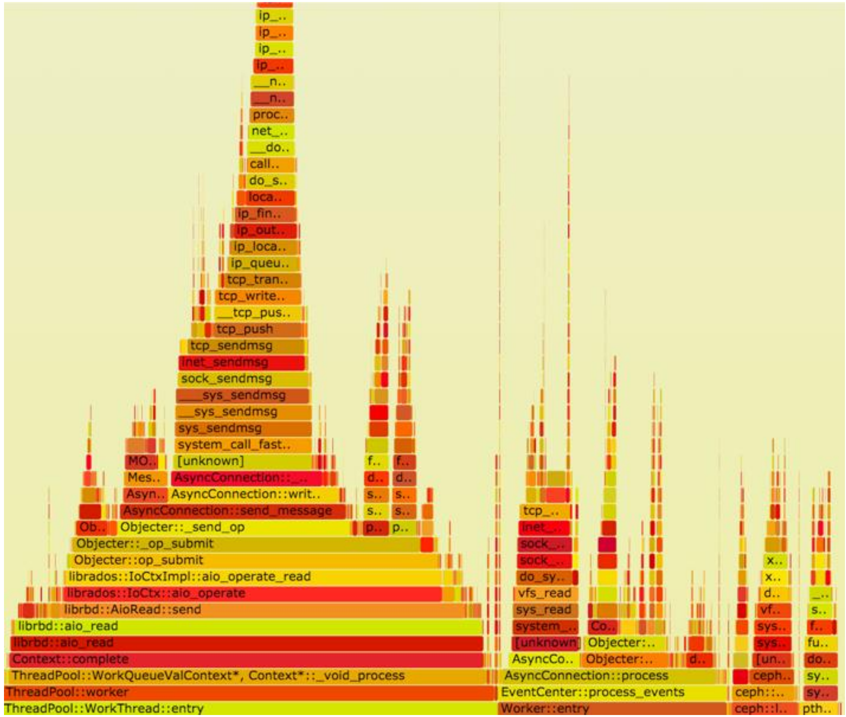


Internal Overview



HIGH PERFORMANCE GAP

Performance Bottleneck



Kernel Bottleneck

- Non Local Connections
  - NIC RX and application call in different core
- Global TCP Control Block Management
- Socket API Overhead

TCP

- TCP protocol optimized for:
  - Throughput, not latency
  - Long-haul networks (high latency)
  - Congestion throughout
  - Modest connections/server

Hardware Revolution

Component	Delay	Round Trip(2009)	Round Trip(2016)
Switch	10-30us	100-300us	5us
OS	15us	60us	2us
NIC	2.5-32us	2-128us	3us
Propagation Delay	0.5us	1.0us	1us
Total	25-70us	200-400us	11us

Alternative Solutions

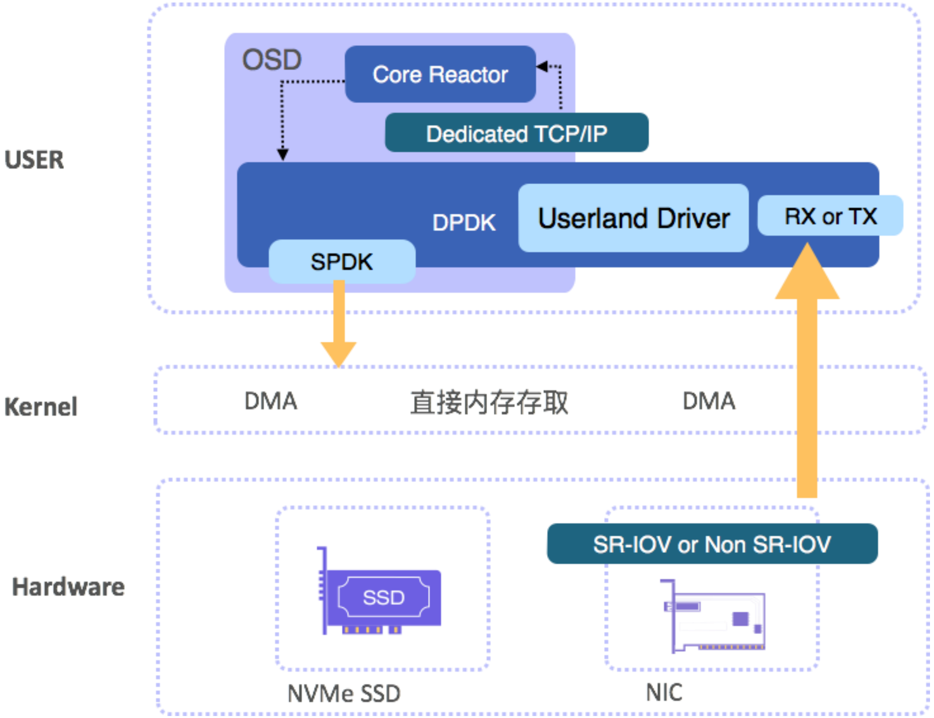
- Hardware Assistance
  - SolarFlare(TCP Offload)
  - RDMA(Infiniband/RoCE)
  - GAMMA(Genoa Active Messange Machine)
- Data Plane
  - DPDK + Userspace TCP/IP Stack
- Linux Kernel Improvement

TCP or Another?

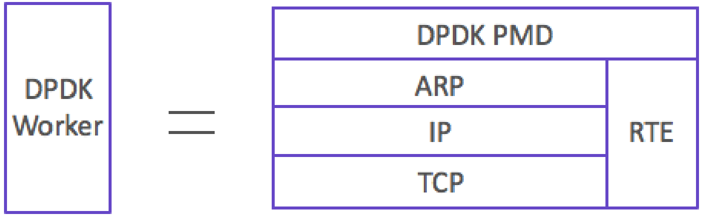
- Pros:
  - Compatible
  - Proved
- Cons:
  - Complexity
- Notes:
  - Try lower latency and scalability but no need to do extremely

CEPH MEETS DPDK

Overview


OSD
Core Reactor
Dedicated TCP/IP
USER
DPDK
Userland Driver
RX or TX
SPDK
Kernel
DMA
直接内存存取
DMA
Hardware
SSD
NVMe SSD
SR-IOV or Non SR-IOV
NIC

DPDK-Messenger Plugin



DPDK Worker

=

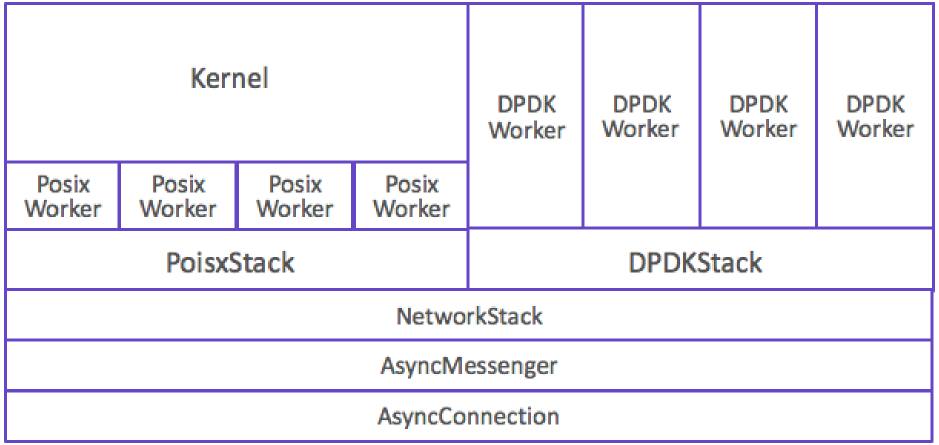
DPDK PMD

ARP

IP

TCP

RTE



Kernel

Posix Worker

Posix Worker

Posix Worker

Posix Worker

DPDK Worker

DPDK Worker

DPDK Worker

DPDK Worker

PoisxStack

DPDKStack

NetworkStack

AsyncMessenger

AsyncConnection

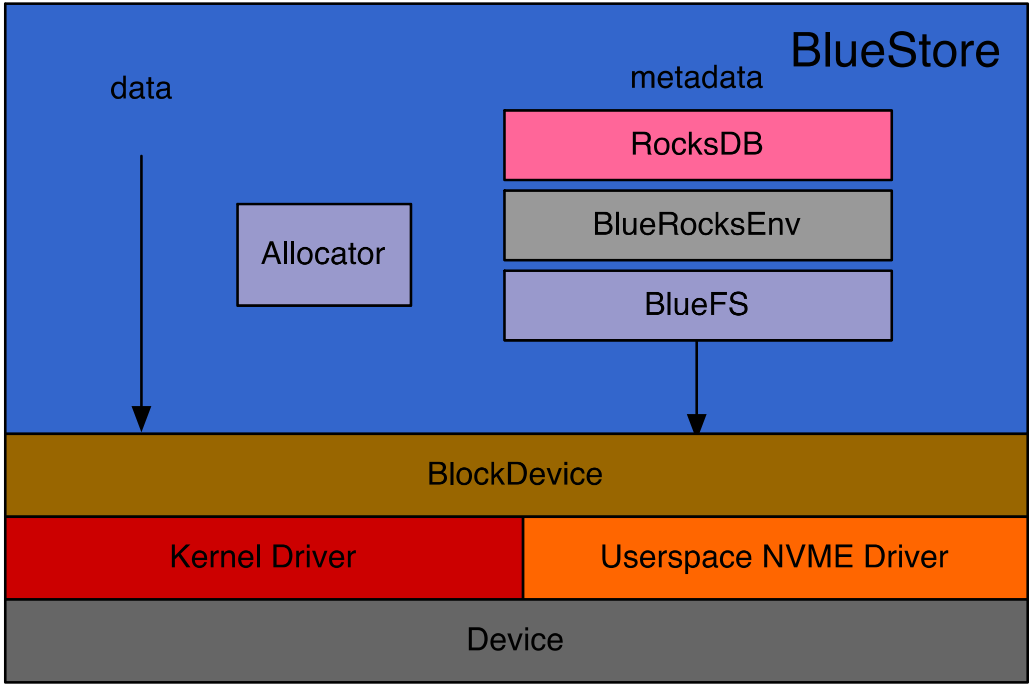
Design

- TCP, IP, ARP, DPDKDevice:
  - hardware features offloads
  - port from seastar tcp/ip stack
  - integrated with ceph’s libraries
- Event-drive:
  - Userspace Event Center(like epoll)
- NetworkStack API:
  - Basic Network Interface With Zero-copy or Non Zero-copy
  - Ensure PosixStack <-> DPDKStack Compatible
- AsyncMessenger:
  - A collection of Connections
  - Network Error Policy

Shared Nothing TCP/IP

- Local Listen Table
- Local Connection Process
- TCP 5 Tuples -> RX/TX Cores(RSS)
- Mbuf go through the whole IO Stack

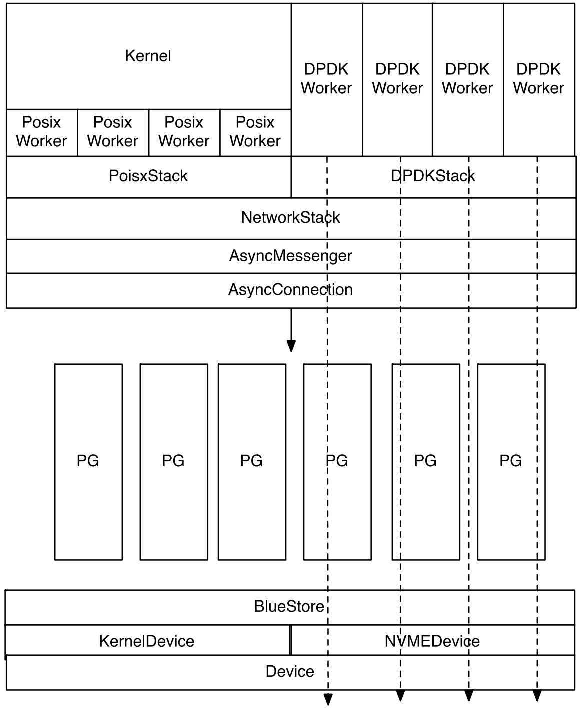
BlueStore


BlueStore
data
metadata
RocksDB
BlueRocksEnv
BlueFS
Allocator
BlockDevice
Kernel Driver
Userspace NVME Driver
Device

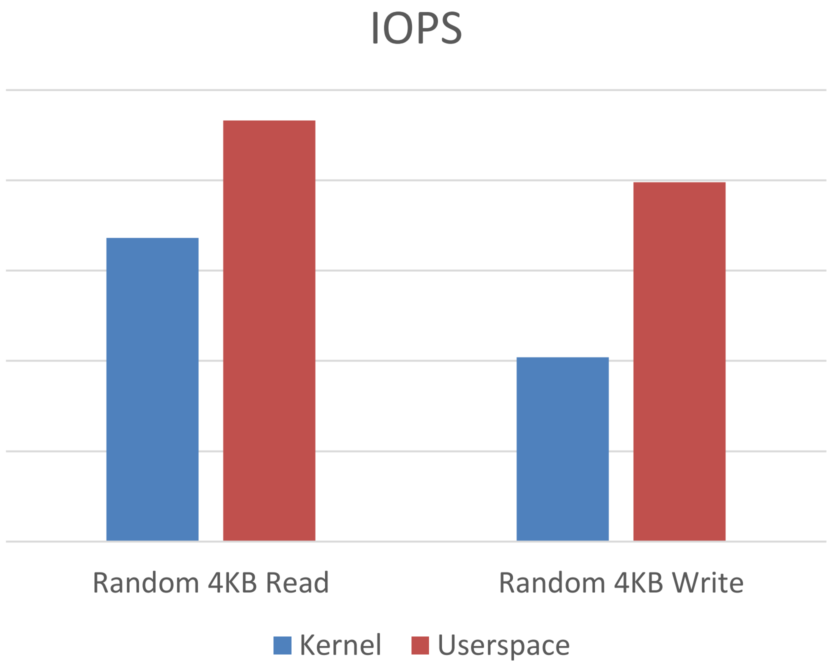
NVMEDevice

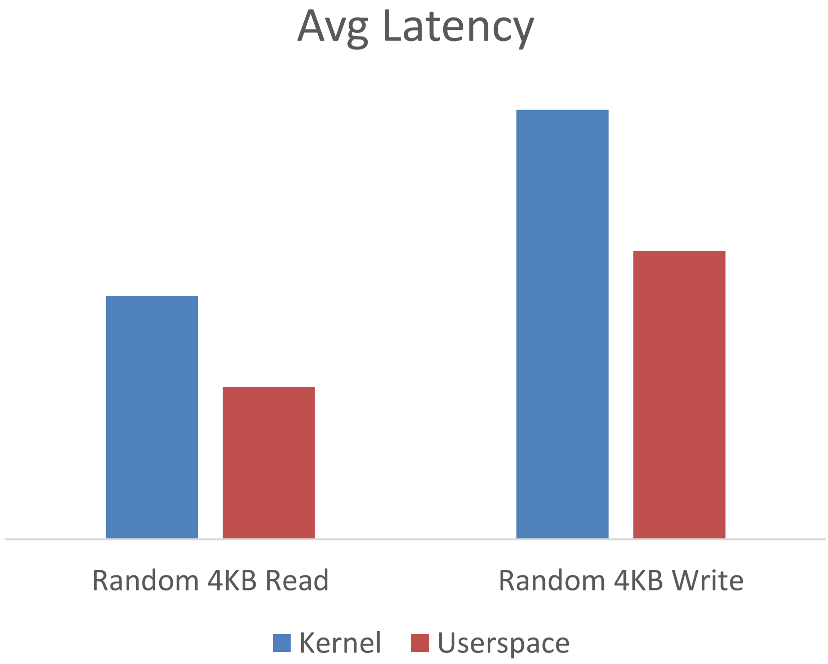
- Status
  - Userspace NVME Library(SPDK)
  - Already in Ceph master branch
  - DPDK integrated
  - IO Data From NIC(DPDK mbuf) To Device
- Lack
  - Userspace Cache

Details


Kernel
Posix Worker
Posix Worker
Posix Worker
Posix Worker
DPDK Worker
DPDK Worker
DPDK Worker
DPDK Worker
PoisxStack
DPDKStack
NetworkStack
AsyncMessenger
AsyncConnection
PG
PG
PG
PG
PG
PG
BlueStore
KernelDevice
NVMEDevice
Device

Improvements





Roadmap

- Core Logics
  - no signal/wait
  - future/promise
  - full async
- Memory Allocation
  - rte_malloc isn’t effective enough
  - mbuf livecycle control

Summary

- Storage device is fast
- Storage system need to refactor to catch up hardware
- Ceph is changing to share-less implementation
- DPDK library is expected to be integrated to office Ceph repo(K release)
- Lots of details need to work(coming soon)

DPDK

DATA PLANE DEVELOPMENT KIT