SPDK在字节跳动存储业务中的应用

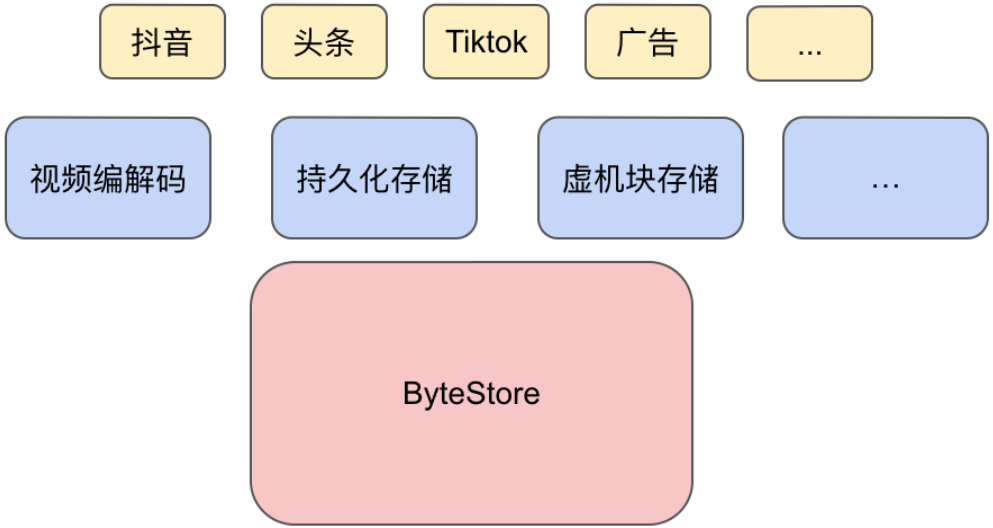
SPDK in ByteDance Storage

苗宇 miaoyu.01@bytedance.com

ByteDance 字节跳动

截至2019年第三季度，字节跳动旗下产品全球总DAU超过7亿，总MAU（月活跃用户）超过15亿，其中仅抖音DAU超过3.2亿。

自研的统一存储平台：ByteStore



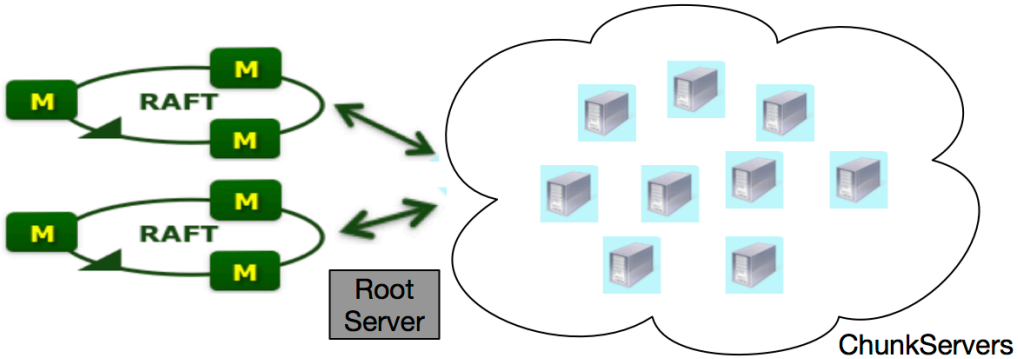
- 服务器万级规模，存储规模EB级，单集群百PB级，日增新数据PB级
- 高可靠，扩展能力强，能够跟上业务快速发展的步伐
- 高可用，上层业务无状态，快速迁移，容忍故障且问题定位迅速
- 低成本，高性能，享受生态和硬件红利

ByteStore：功能模块

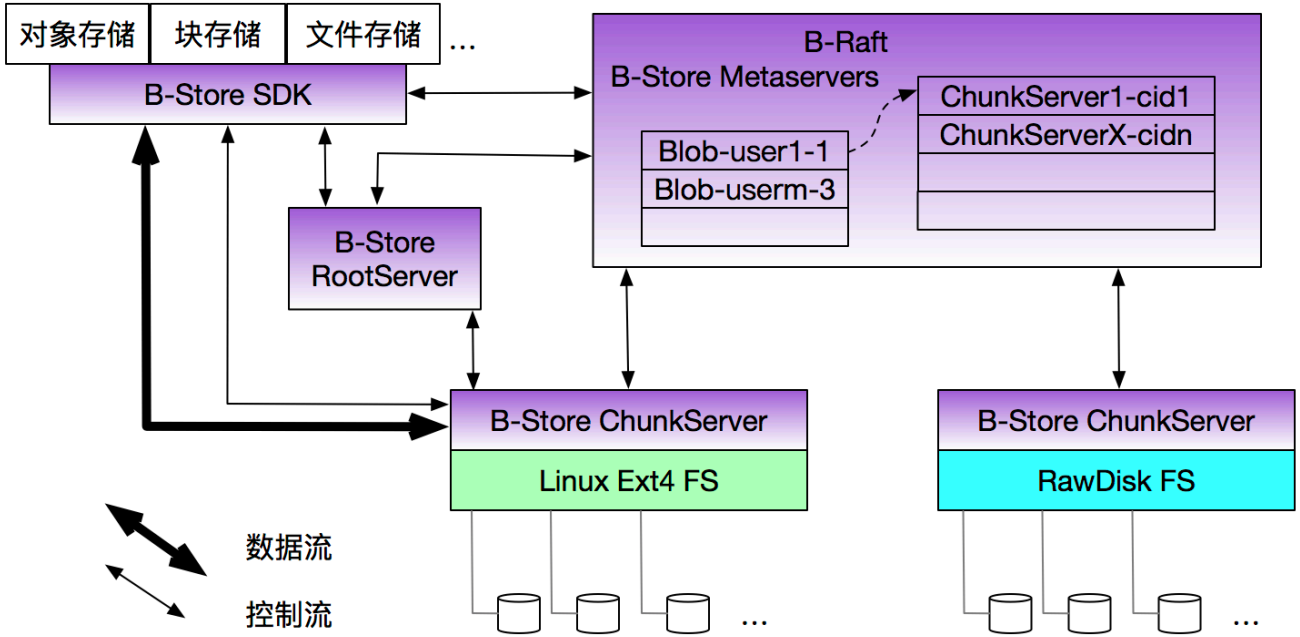
RootServer（一个）：集群中心节点, 负责集群管理，管理Pool到ByteMaster的路由关系

MetaServer（多组）：元数据节点，存储Blob信息，管理Blob到chunk之间的映射

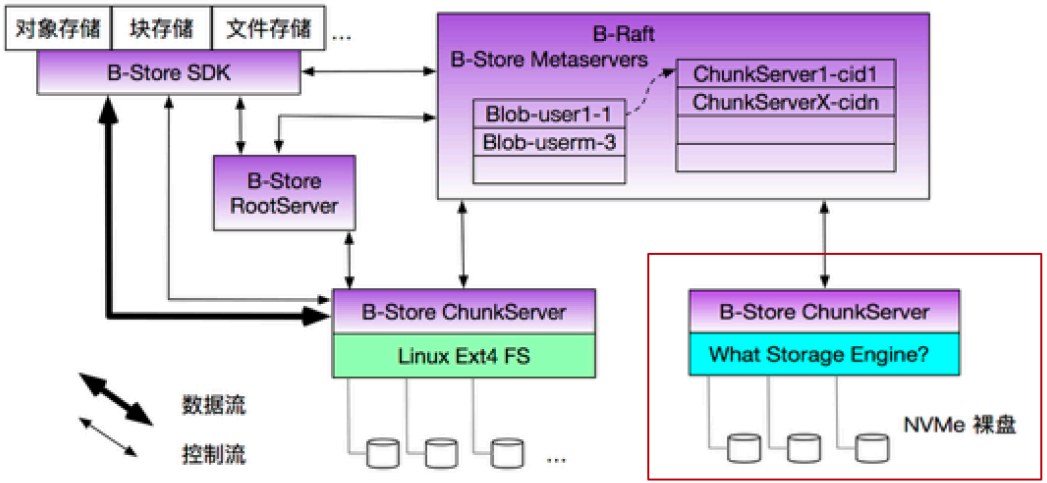
ChunkServer（多个）：数据节点，存储实际分块（chunk）数据



ByteStore：整体架构



ByteStore：面临的问题



ChunkServer对单机引擎的需求：

- 建立在裸盘基础上的非日志型本地“文件系统”
- 面向高性能存储介质设计，减少写放大
- 低延迟I/O软件栈，支持Run-To-Completion执行模型提升CPU I/O利用率
- 易于管控和维护
- 匹配高性能网络表现

ByteStore：我们的方案

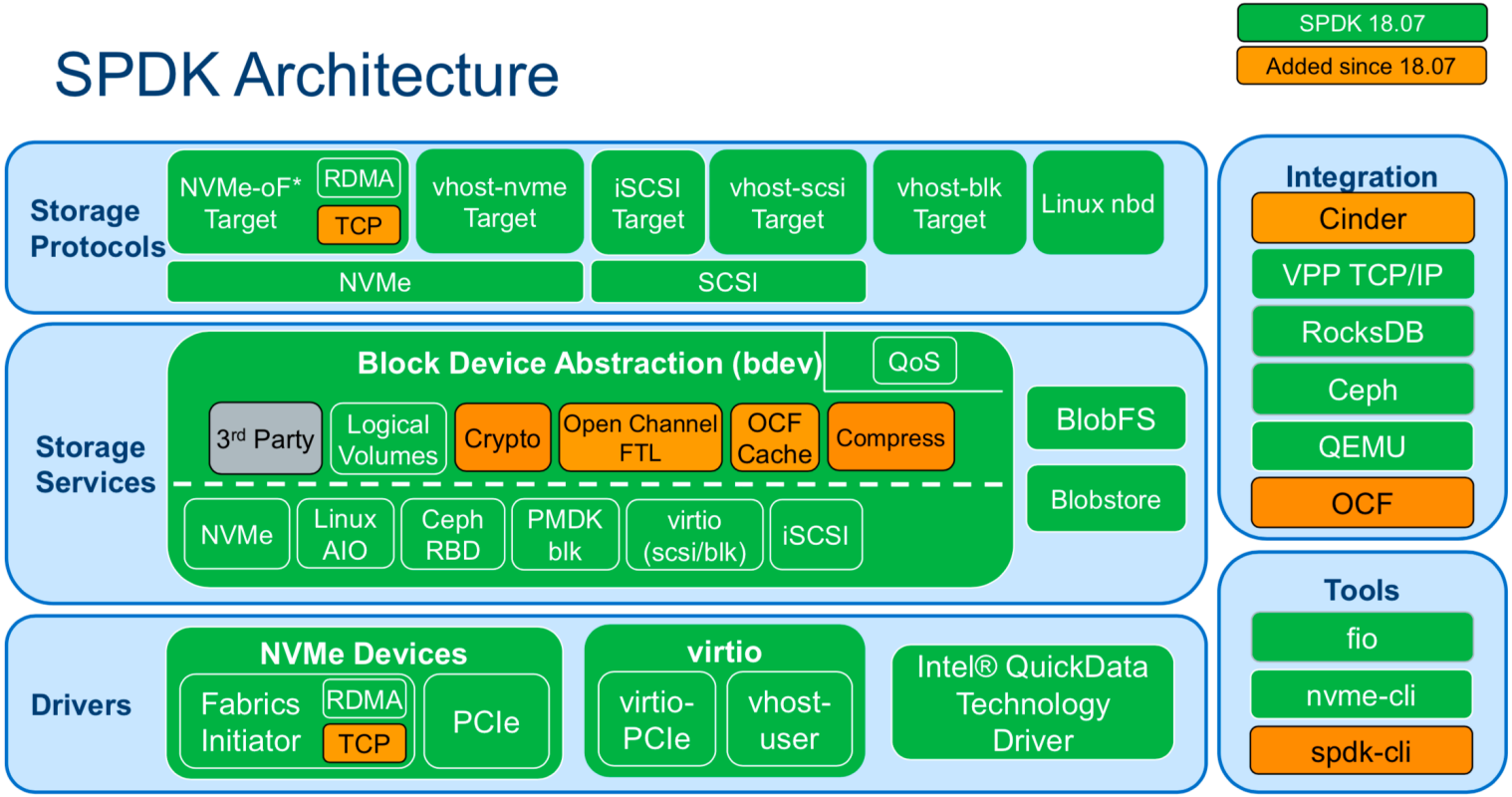
我们选择了SPDK！

SPDK是什么

Storage Performance Development Kit　存储性能开发套件

- Intel开源的用户态存储框架，能够为用户提供不同层面高性能存储服务；根据支持的存储功能来讲，可以分为驱动层，块设备层和存储协议层；相邻层之间的模块共享Polling Mode操作的资源；
- 旨在帮助客户优化存储系统软件栈的性能；主要应用场景是高性能的NVMe本地盘、Fabric Target盘及虚拟化盘；有针对最新一代的CPU和存储介质的优化；
- 拥有较为完善的测试流程和工具；社区开放，很欢迎用户积极参与提交建议和补丁；

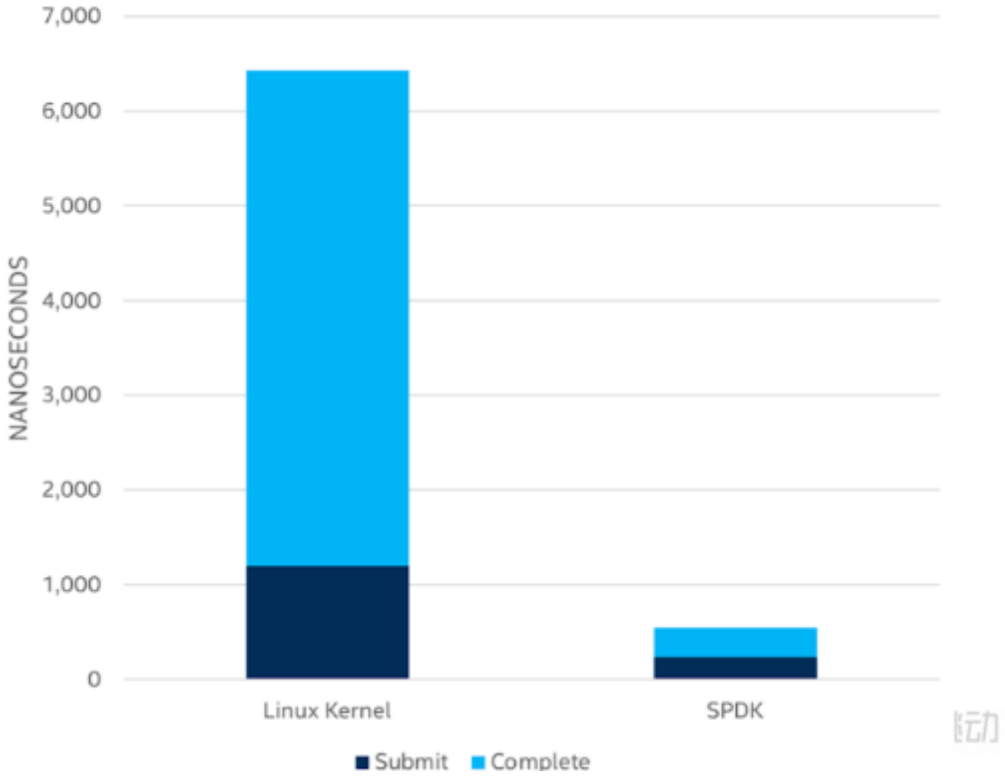
SPDK Architecture



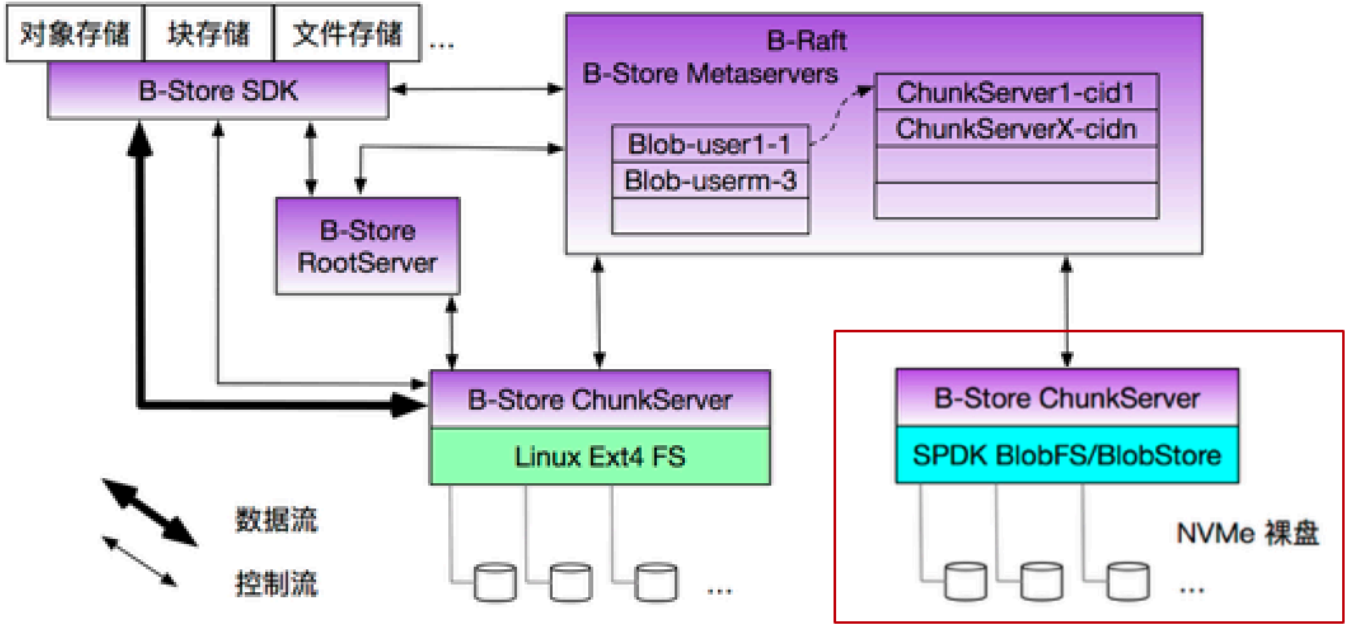
Invite SPDK to dance with ByteStore as our FIRST step

SPDK框架提供：

- 异步、无锁、零拷贝和轮询；
- Bypass 内核，全程用户态；
- 使用DPDK提供的环境抽象层，可支持链接时优化LTO；

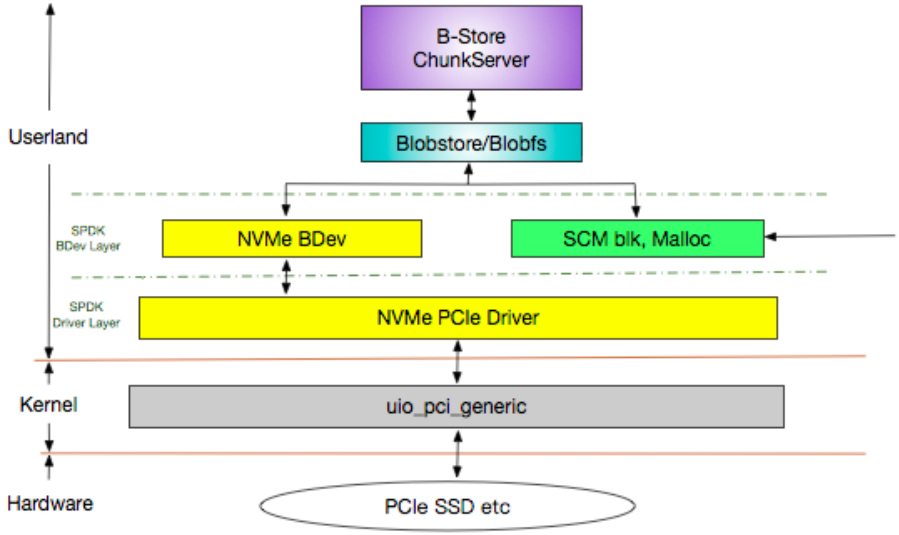


新的SPDK存储引擎



ByteStore单机引擎的变化

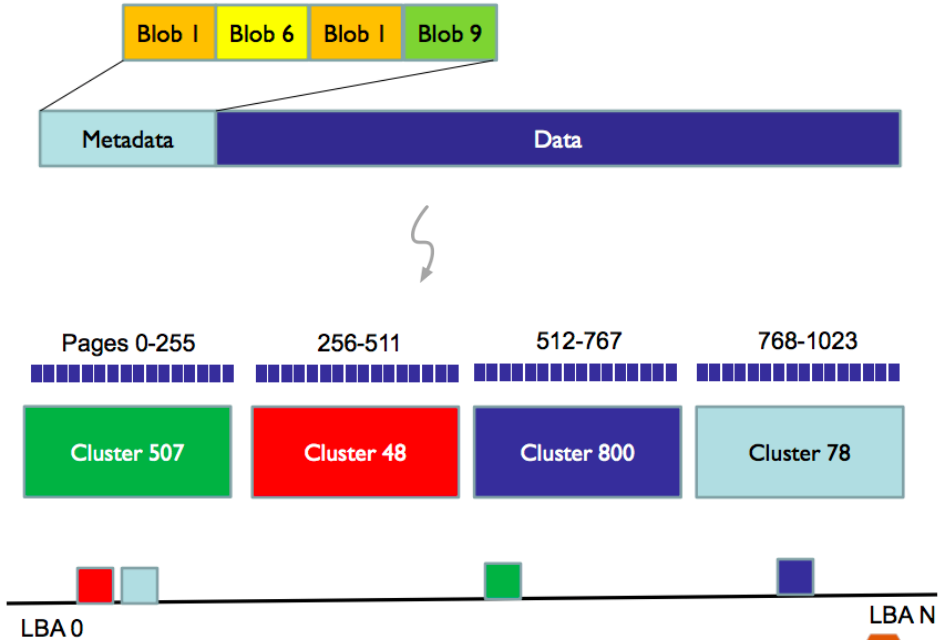
- 从PCIe 驱动开始接管NVMe SSD，使用高度优化的无锁驱动，负责SSD的资源功能初始化和管理工作；
- 用户态抽象块层连接SSD驱动和上层的存储服务，提供灵活的块设备操作API；
- BlobStore 提供了对SSD存储单元的逻辑抽象管理，借助轻量级的BlobFS实现了高精简的类Posix文件式语义API；



SPDK BlobStore/BlobFS

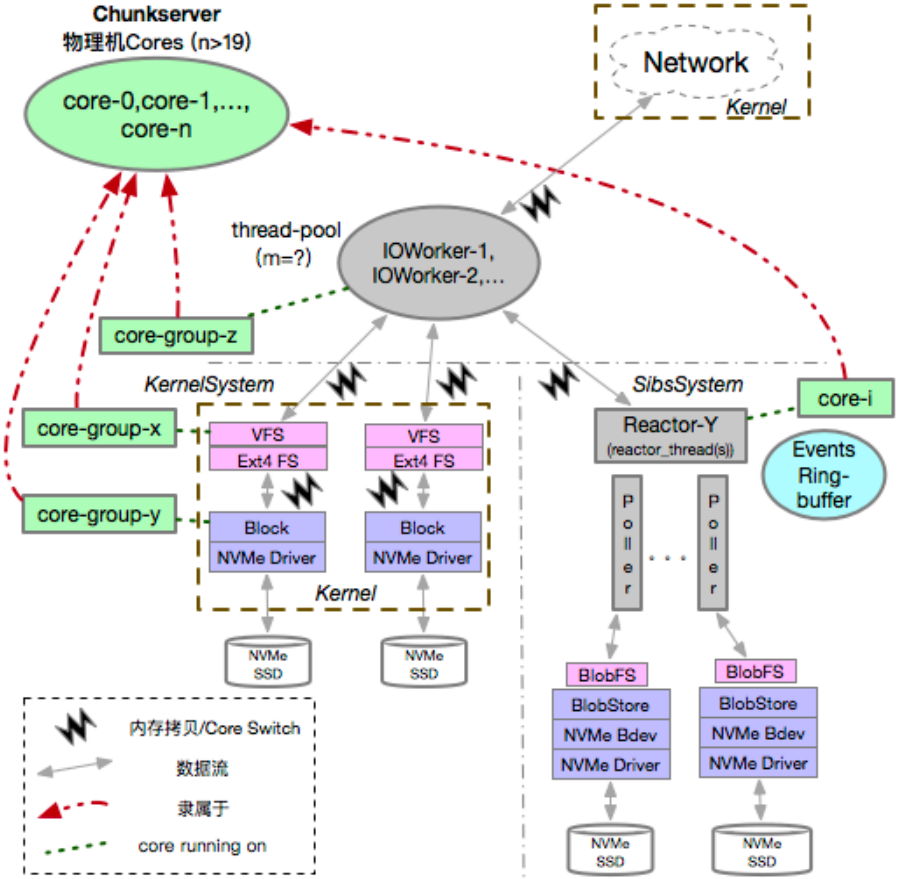
- 存储I/O单元抽象为blob，cluster，page，block；用户操作对象为blob
- 全局元数据（in pages）存储在盘的预留区域，与数据blob分开存放
- 整体实现异步无锁
- 提供类Posix接口
- 数据direct I/O，元数据cached

On-disk Layout：



ChunkServer引入SPDK前后对比

- 充分匹配和发挥高性能存储的硬件能力，减少了CPU切换和内存拷贝
- 单core可以酌情伺服多块存储设备，提升CPU使用效率
- SPDK Reactor的轮询机制和BlobFS的读缓存帮助降低I/O栈延时



SPDK 给 ByteStore带来的整体架构级优化

- 使存储节点能力匹配高性能网络技术的应用和带宽的提高，能够在ChunkServer中大幅提高SSD存储密度；
- 充分利用SPDK绑核轮询机制，实现（引擎内）I/O路径无锁，显著提高了CPU的单核利用率；
- 全栈用户态，且天然兼容Intel QAT驱动等；
- 提供了池化分层存储和混合存储的构建基础；
- 当前版本ChunkServer延时降低了～60%， 仍有很大优化空间；

ByteStore SPDK相关工作计划

性能方面：

- 与ByteStore的网络模块同步优化：存储与网络的I/O内存零拷贝；轮询线程的共享；
- 引擎之外的ChunkServer相关架构优化：与SPDK一致的绑核支持和相应线程资源调度；

功能方面：

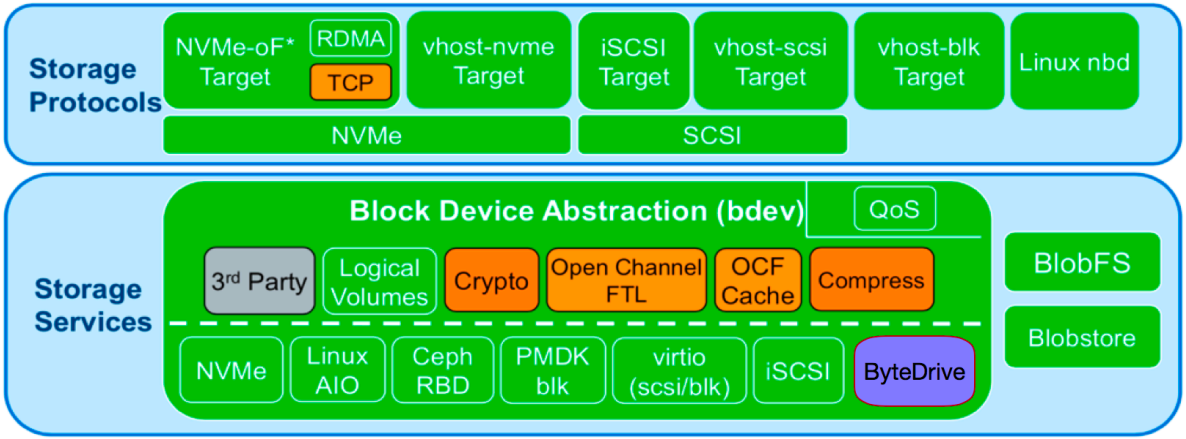
- 支持NVMe SSD在BlobStore层的分区工作，实现多ChunkServer共享单盘。

SPDK的未来

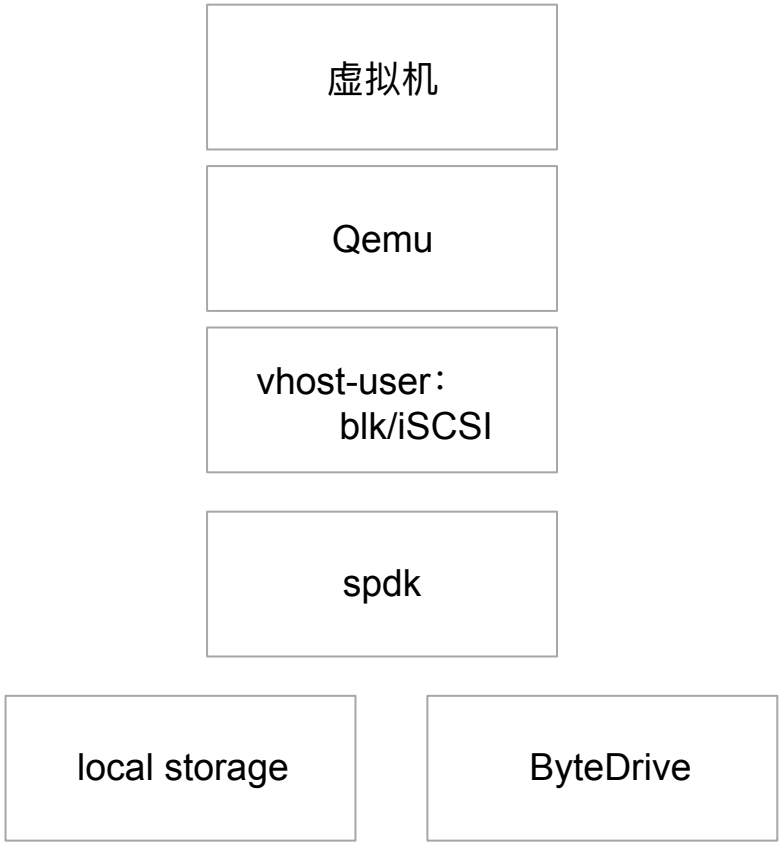
基于ByteStore和SPDK的存储延展

块存储

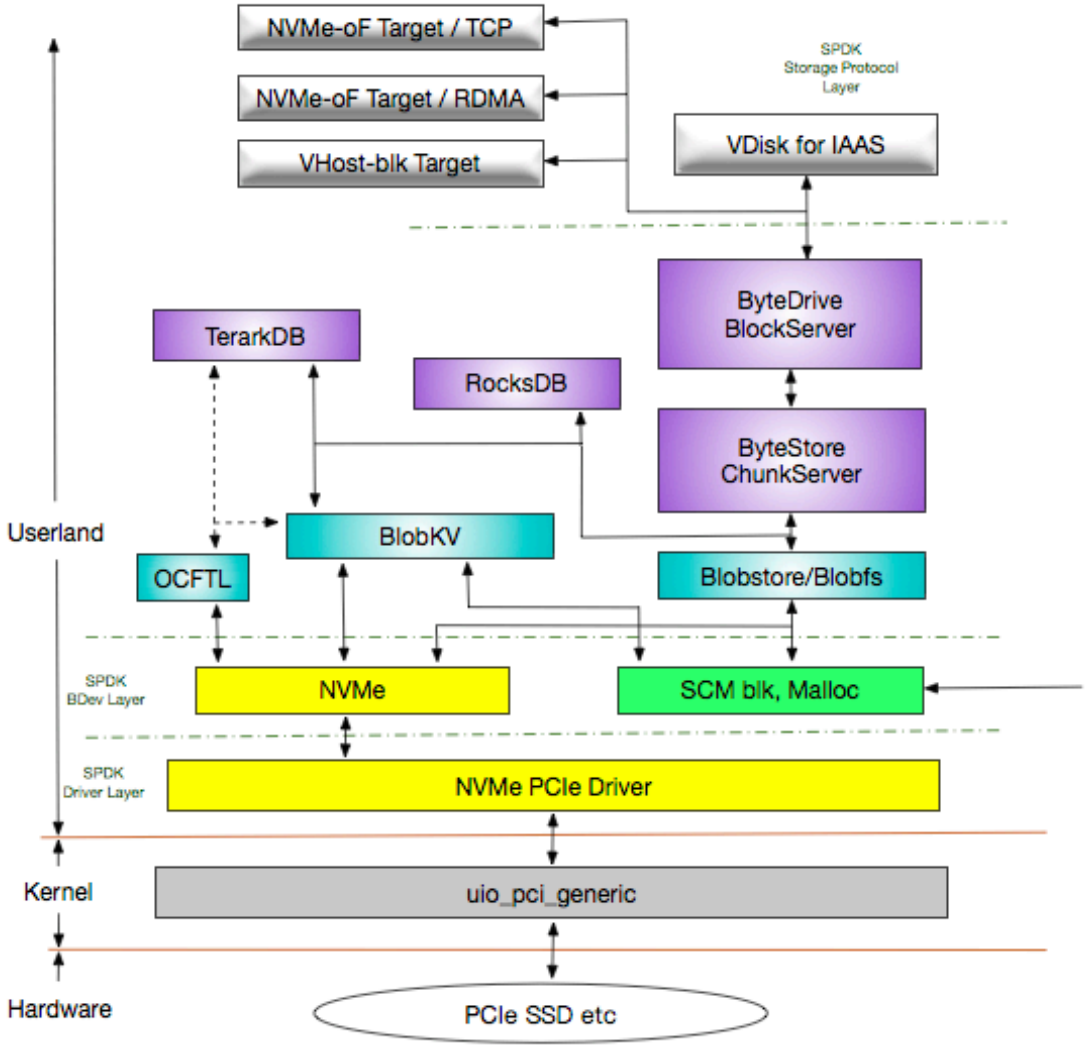
- 构建于ByteStore之上的自研块存储产品**ByteDrive**，替代公司内部使用的开源Ceph云磁盘
- 利用SPDK良好的抽象分层架构，编写ByteDrive对接模块嵌入SPDK框架，实现与存储协议iSCSI、 NVMe-oF和Vhost Target无缝对接；在Bytestore 和 ByteDrive整体I/O路径高性能表现的基础上，存储协议Target用户可以享受接近本地SSD的块设备服务；



SPDK+虚拟机



整体概览



最终目标：实现基于SPDK的一套功能完整的存储服务框架！

THANKS！

Welcome to ByteDance

We are hiring ~~

Let’s dance together!