sodacon
— Global 2021 — July 13-14

# H3C与Intel在全闪架构下Ceph的优化实践

**H3C** 存储架构师 毛宏华 mao.honghua@h3c.com

**Intel** 云与企业应用方案存储架构师 冯添 tian.feng@intel.com

soda foundation

#sodacon2021

# 内容概要

- 高性能全闪存储诉求
- 优化要点
  - 重删压缩场景技术要点与英特尔® QuickAssist 技术助力
  - I/O路径写缓存优化与英特尔® 傲腾持久内存应用
  - 英特尔® 810 网卡RDMA优化
  - 全栈优化要点总结
- 下一代Crimson OSD最新进展

# 全闪架构: 市场与技术趋势

**软件定义存储+全闪存+AI赋能**，承载新型应用海量数据


**核心应用闪存化**：闪存加速来达成高性能

**关键能力智能化**：智能加速、智能监测、智能运维

**业务全面分布式化**：计算与容量线性拓展

# 全闪架构: 内部诉求

**既有架构缺陷**

- 目前的存储系统，基于机械盘或者混闪设计的，无法充分发挥出闪存盘的性能，随机小IO的写带宽低
- 没有重删能力，压缩对系统够开销大且压缩率低，TCO竞争力不足
- 读写IO路径长，读写时延高，整体性能低
- EC长尾效应问题导致EC性能低下，商用前景堪忧
- 网络通信时延在整个IO时延占比居高不下
- 内存使用没有统一监管，内存浪费严重，内存不足导致的性能、功能问题风险高

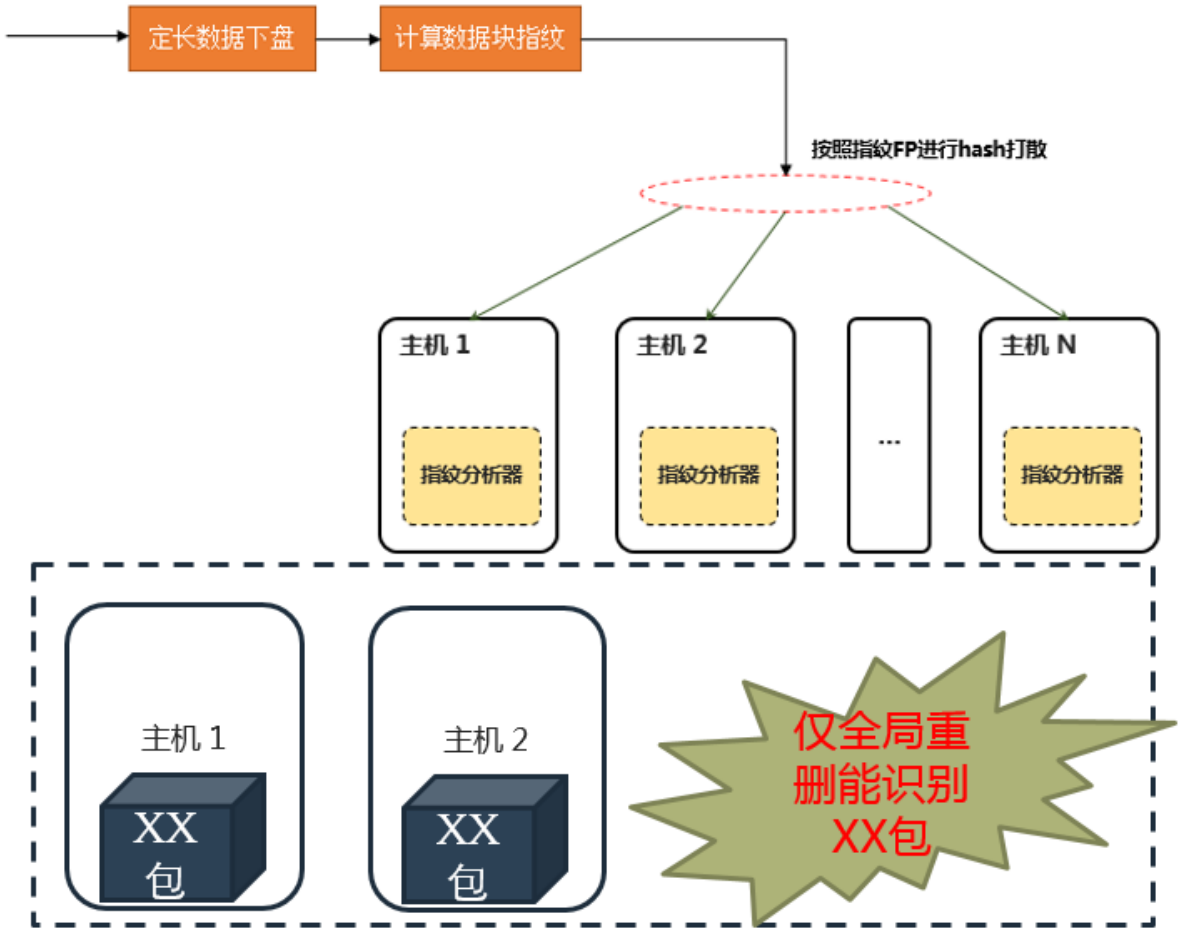
**Gartner八项能力，三项关键特性能力不足**

1. **容量**：数百TB甚至PB级别的容量扩展潜力
2. **空间效率**：支持压缩、重删、Thin Provision、自动分层的能力
3. **性能**：总IOPS、带宽吞吐量、时延指标，以及在真实应用场景下的性能发挥程度
4. **平台适配性**：与第三方软件平台的适配能力
5. **可管理性**：支持自动化管理、监控以及提供分析报告工具的能力
6. **自愈能力** – 提供高可用性和数据保护的能力
7. **多租户及安全性**：包括细粒度访问控制、用户控制的加密、防恶意软件等
8. **价格**：提供优异性价比以及定价模型的能力

**不破不立，破而后立**

需要一套具备**适应存储介质闪存化趋势+高性能+高TCO竞争力+高可靠性的**新架构，才能适应硬件及应用的变化，立足未来市场

# 关键技术-全局重删




**价值**

- 提高重删率，特别是云主机场景，效果明显。

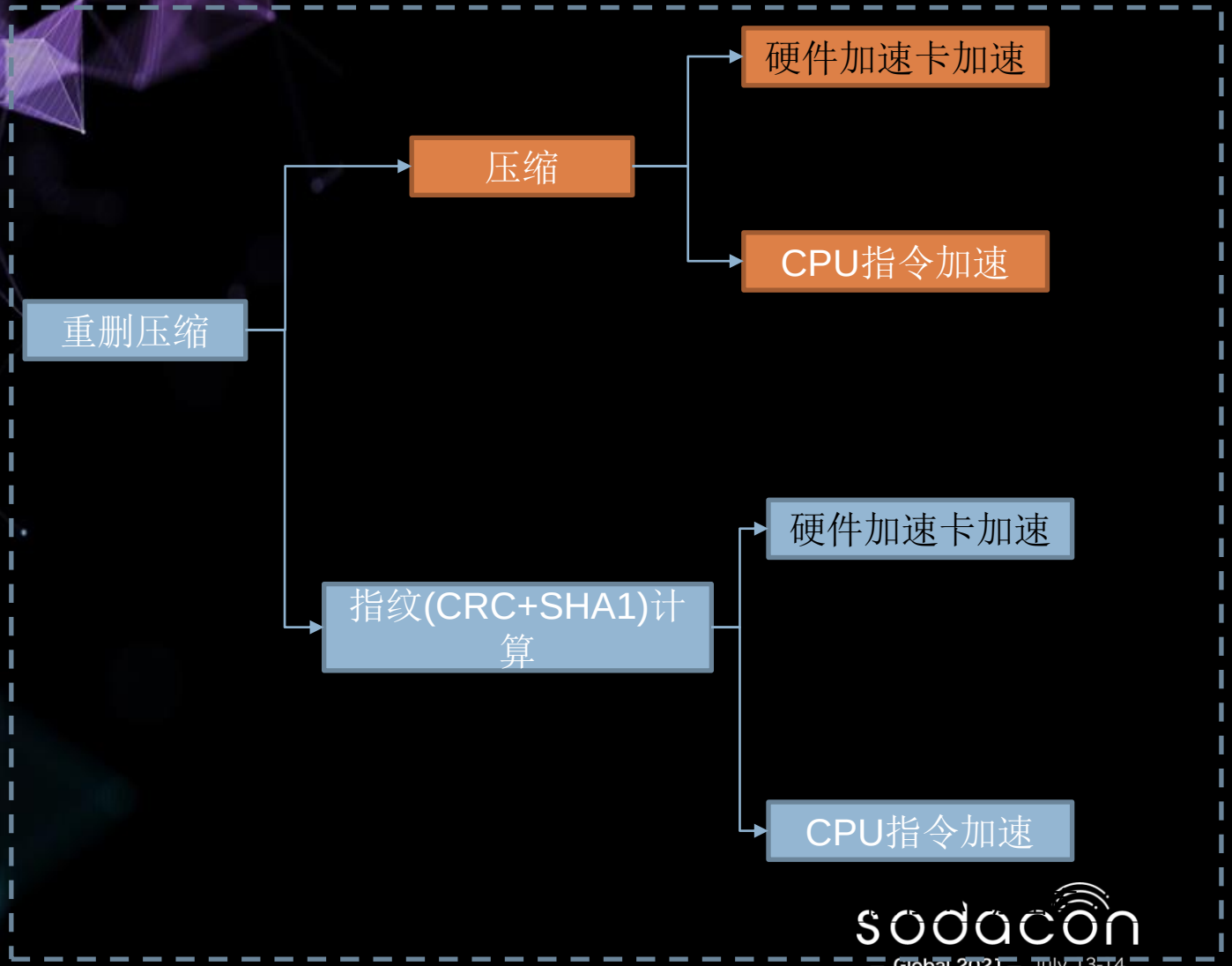
**难点：**

- 需要对整个存储池，进行重复数据识别，查询范围广，消耗网络和CPU资源大。
- 需要在整个存储池内，存储重删数据，需要考虑数据均衡问题。

**创新点：**

- 按照指纹HASH将重删元数据分散存储到不同主机上，由于指纹采用SHA1算法，能保证元数据的均衡性。
- 特定指纹元数据，一定存储于特定主机，数据下盘过程中，每一块数据，只需要在指定主机上发送重删请求。

# 英特尔® CPU指令加速与QuickAssist硬件加速提升计算效率




**价值**

- 指令加速，提升CPU计算效率
- 硬件加速，减轻CPU负载，保证IO性能
- 压缩和指纹同步计算，提升40%性能

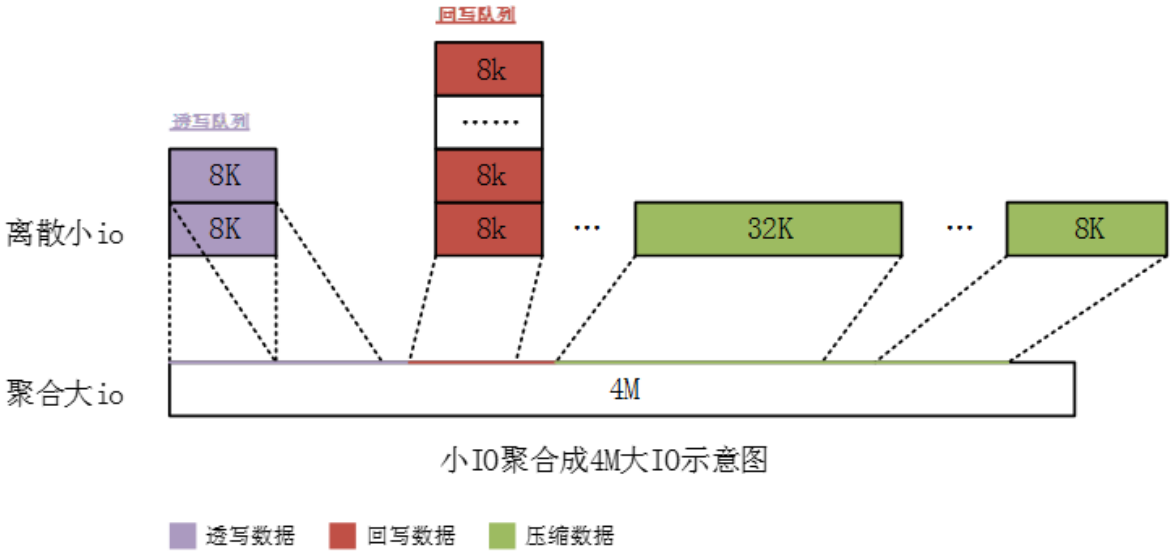
**难点：**

- QAT加速卡适配异步编程模型。
- 如何适配QAT卡，减少内存到设备的数据拷贝。

**创新点：**

- 实现了QATzip异步接口，采用单线程轮训模式，大大减少了在使用QAT卡场景下，对CPU的占用。
- 计划使用自主研发的统一内存管理，实现内存到QAT硬件的免拷贝。
- 计划在QAT硬件加速卡内，同时计算压缩和指纹，提升计算性能。

# 关键技术-IO聚合




小IO聚合成4M大IO示意图

由于所有的新写IO，不管逻辑空间需要写到什么位置，在实际写到盘上都会被追加写到盘上，只要下发IO量足够大，就能聚合足够大的连续IO。

**IO聚合目的：**

聚合大IO，一方面，能保证所有磁盘写入都为顺序大IO写入，提升写入性能；另一方面，能很容易凑齐EC满条带，减少非满条带写入读惩罚问题。

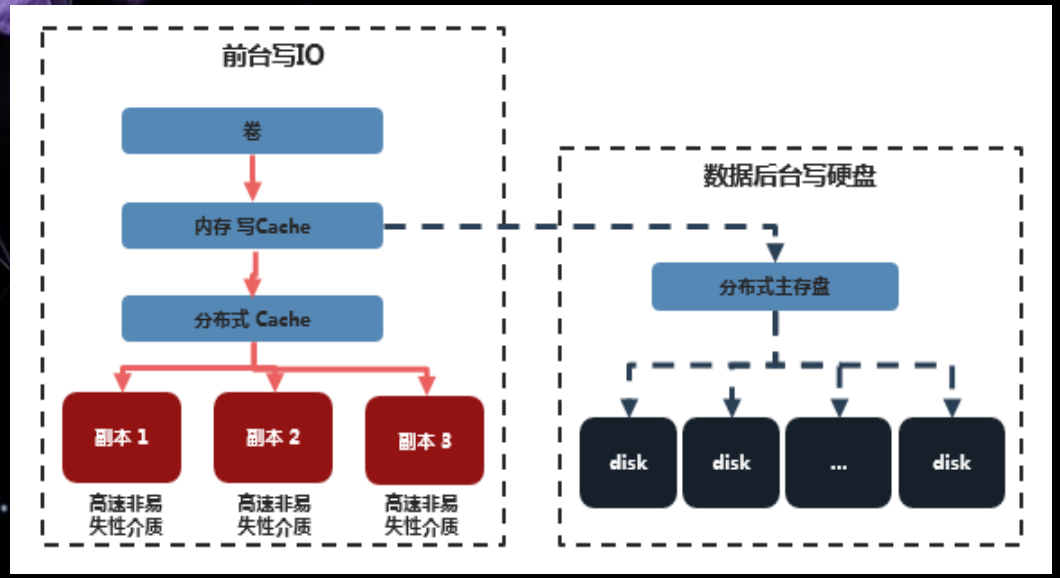
**技术难点：** 压缩数据介于8k到32k之间，以32k为主，非压缩数据都是8k大小。对于聚合模块来讲，需要聚合的小io大小是不均匀的，如何保证良好的聚合效果呢？

**技术亮点：** 优先聚合压缩数据，因为压缩数据已经在压缩模块消耗了一定的时延，优先处理可以降低整体IO时延，保证客户端IO响应迅速。聚合算法需要采取优先聚合大io，不足4M的部分用小io来补的策略以达到更有的聚合效果。

**技术难点：** 考虑聚合模块之存在cache层，那么聚合模块就需要对透写io和回写io进行聚合，两种IO对时延的要求不同，因此将两种IO分开管理。

**技术亮点：** 透写IO是从客户端直接下发的IO，聚合算法要保证透写IO在队列中等待的时间足够短，优先聚合透写IO以保证客户端IO的低时延；回写IO尽可能大量聚合。

# 英特尔® 傲腾持久内存加速方案




## PMEM

类似于 ceph，对于缓存数据池的每一个副本或分片，都有一个 store 模块，以提供缓存数据的写入、读取和删除的功能。缓存数据在 store 中以对象的形式存在。

基于 PMEM 实现的 store，其数据布局如下图：


其中："Header"中保存一些配置信息；"obj meta"中保存对象的元数据，包括对象名，数据存储地址，数据长度等等；"obj data"中保存对象的数据。

数据写入流程大致如下：

1、 在"obj data"中分配一块数据空间（如果需要的话），并写入数据；

2、 在"obj meta"中分配一块元数据空间（如果需要的话），并写入元数据；

价值：

1.写IO充分发挥AEP新介质的高性能与持久性能力，缩减IO路径，简化可靠性处理

2.少量的DRAM空间资源使用

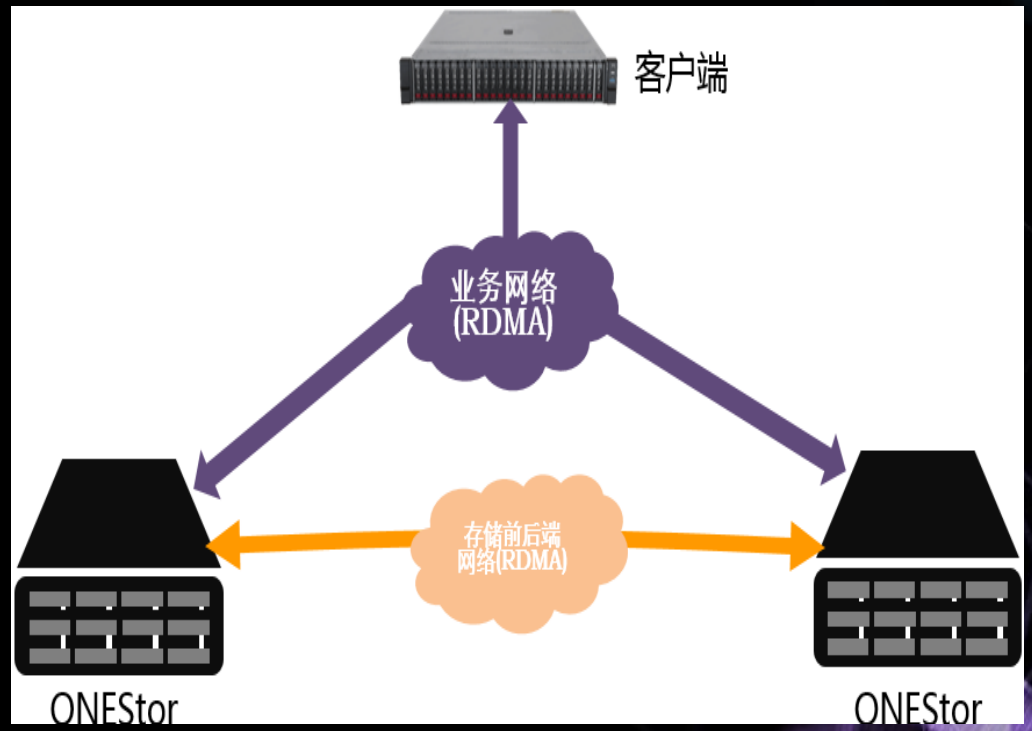
# 英特尔® 810 网卡加速方案

◆ **更好地支持HPC、金融等高性能低时延场景:**
RDMA的零拷贝、内核旁路、直接访问远端内存等诸多特点使得ONEStor使用RDMA网络时，整体拥有更低的IO时延，同时具有更高的系统资源利用率(CPU使用率大大降低)

◆ **已实现存储集群统一的RDMA通信机制:**
- 存储前后端网都采用RDMA时，IO请求从(存储前端网)接收到(从存储后端网)发送的整流程已实现完全的内存免拷贝
- 当前小IO场景性能提升10%，大IO场景性能提升30%

◆ **性能提升空间仍然较大:**

- 使用RDMA最大的挑战在于，RDMA设计与应用需要匹配，没有通用万能的方法，适配到最好与最坏选择，性能差距可以达到70倍
- 当前小IO场景下性能提升有限，需要持续研究RDMA技术细节，反复进行修改验证，最终实现小IO性能优化效果与大IO基本持平

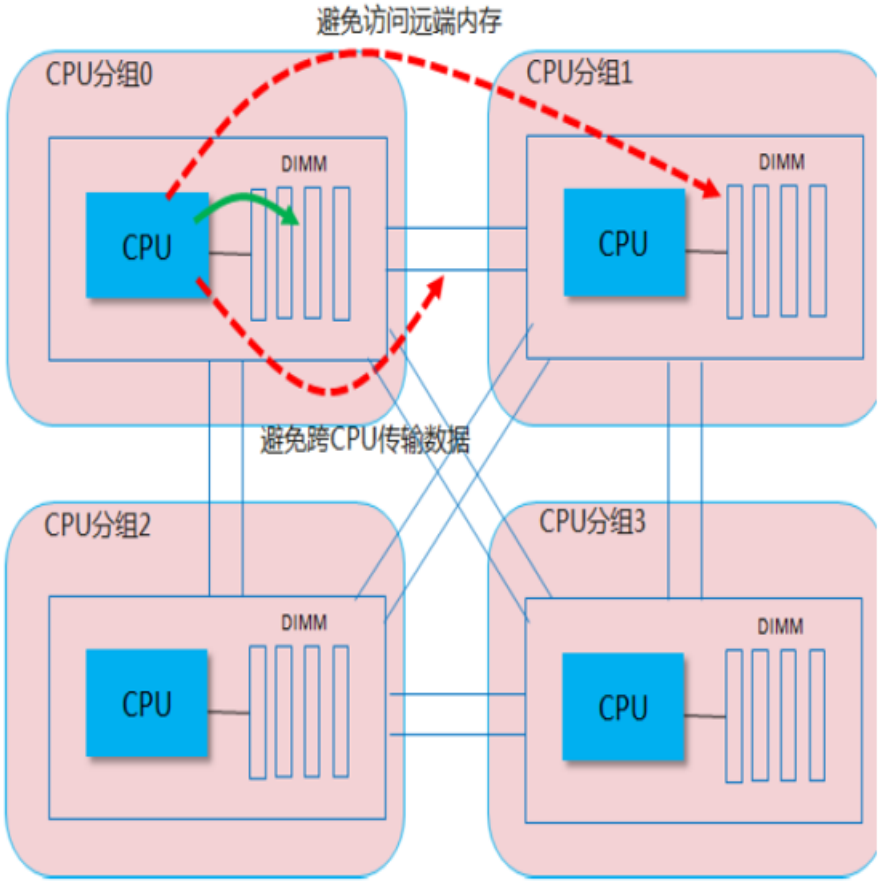


# 关键技术-CPU分组

原理：在多CPU的架构中（NUMA架构），每个CPU有本地内存（该CPU内存控制器上连接的内存），也可以访问远端内存（连接在其他CPU内存控制器上的内存），访问本地内存比访问远端内存时延更低，开销更小，对内存的访问效率会直接影响到存储系统的性能。CPU分组技术，控制器内每颗CPU加上本地的内存，形成一个CPU分组。

**价值**：每个CPU分组接收到的主机读写请求，就尽量在本CPU分组内完成端到端的读写流程处理，避免访问远端内存，以及垮CPU传输的开销。通过CPU分组技术，每个CPU分组处理不同的主机读写请求，实现了存储系统性能随CPU数量增加而线性增长。右图4颗CPU，分成4个CPU分组，独立的处理不同的主机读写请求，避免跨CPU传递数据及访问远端内存。

难点：不同服务器上CPU个数不同、每个CPU上的核数不同，如果做到业务负载均衡成为技术难点



# 性能提升技术点

1. 前、后台IO在不同的线程处理，避免后台IO影响前台IO
2. IO在关键路径上保持在同一线程上处理，减少切换开销
3. IO根据优先级来分配资源，保证高优先IO及时处理
4. 对象在IO路径的打散规则一致，做到免锁
5. 数据在IO路径上零拷贝
6. 内存不跨NUMA访问
7. IO路径避免频繁的申请与释放内存
8. 提升ROW对象、OSD对象的亲和性，减少网络开销
9. 数据均匀打算到不同的OSD对象

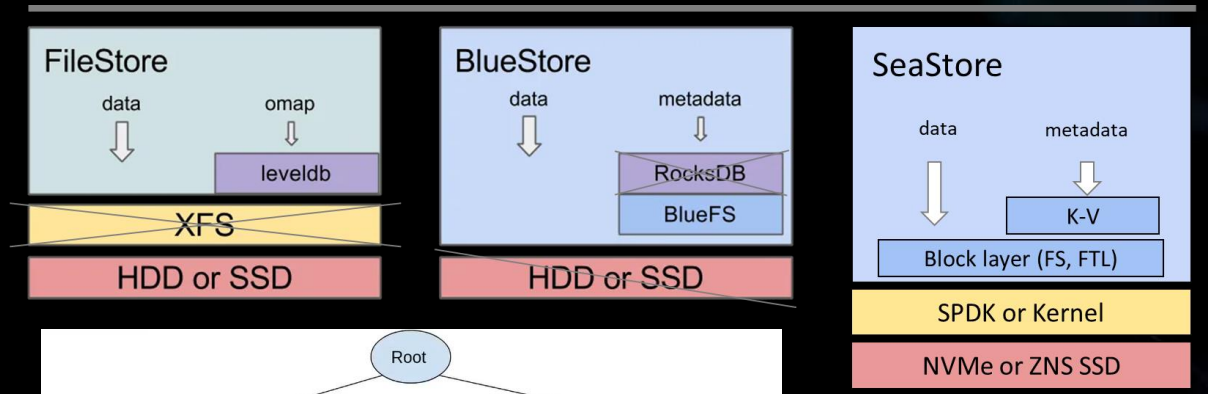
   Client上objecter数量与数据处理实例数量保持一一对应
10. 将一块PMEM部署成多个OSD，提高并发
11. 进程合一，减少进程间通信

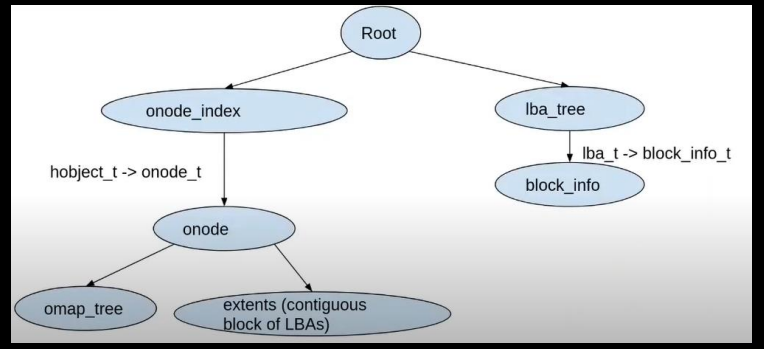
# Crimson面向未来，精工细作

基本思想：

Seastore:Log-structured File System

Seastar framework



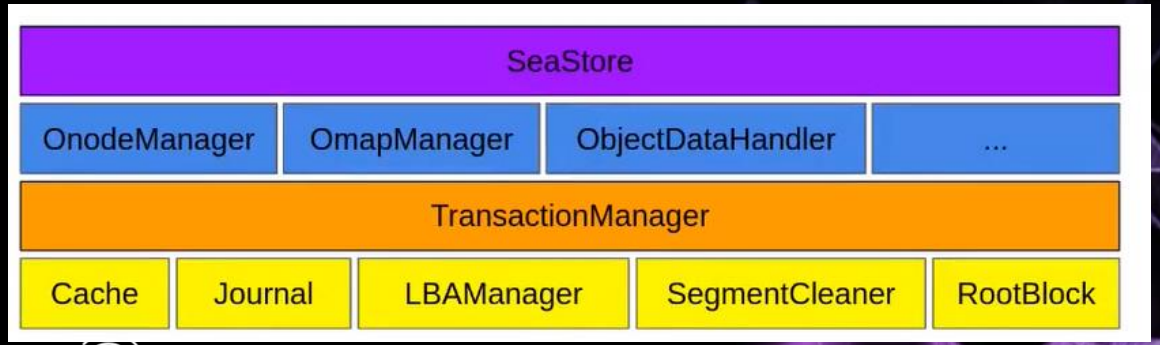




Ceph Quincy发布

- Transaction layer
- LBA layer
- Garbage Collection
- Onode Index
- Omap
- crimson-store-nbd

目前进展

- Optane SSD 和Optane Pmem支持（Non-Trim）
- 稳定性和性能优化重构
- Transaction 机制面向Pmem优化




https://www.youtube.com/watch?v=vzJPOA7aJMk

sodacon
— Global 2021 — July 13-14

Respect

ceph

sodacon
— Global 2021 — July 13-14

# Backup

# DPDK+VPP用户态协议栈

存储后端1M IO性能对比：

- 当前测试模型6系CPU不是瓶颈，实测带宽数据距离网卡理论带宽还有很大差距；
- OSD的业务在vdbench 256并发的时候，性能最优；
- VPP协议栈带宽比内核协议栈低（约低4%）；VPP时延低并发优于内核；
- 随着并发数量增加，VPP的拷贝占比增加，两种协议栈时延基本一致，但性能没有提升，对比iperf测试结果，可以推断出当前业务模型下，上层业务给协议栈及网卡提交的数据流量不够；

存储后端4K小IO性能对比：

- 4K IO CPU idle 有富余，CPU不是瓶颈
- 4K IO 256并发下，VPP时延比内核最大优越 14%； 最小优越3%
- 4K IO 256并发下，VPP带宽比内核优越10%

# 英特尔® QuickAssist加速方案

## 价值：

1.提升重删压缩性能，满足全闪存存储对于在线重删压缩的实时性要求

2.相对于原有同步多线程模型，异步模式能够显著降低 CPU 资源占用率

3. 显著节省了系统资源耗费所带来的成本，提升了全闪存存储系统的投资回报

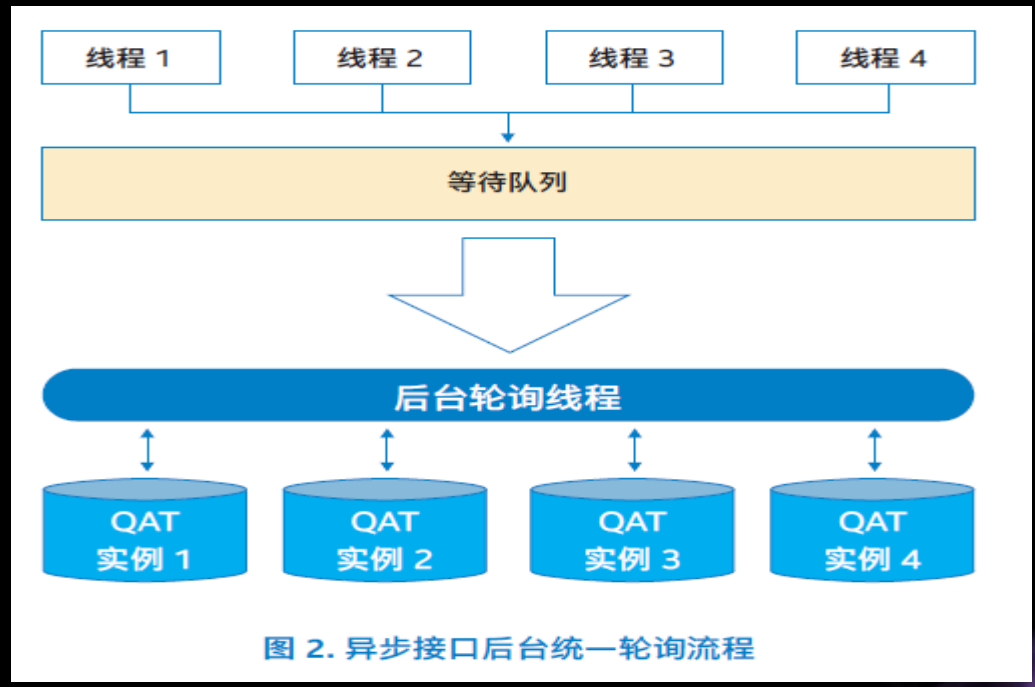


图 2. 异步接口后台统一轮询流程