# 基于 RDMA 的分布式内存存储数据传输优化

(申请中山大学工学学士学位论文答辩报告)

学　生: 兰　靖

计算机学院 计算机科学与技术

二〇二二年五月

# 目录

# 分布式计算框架

## 早期计算框架 (Hadoop, Spark, Horovod)

- 单一的任务类型
- 固定的并行模式
- 有限的表达能力

## 计算任务的变化

- 异构任务
- 复杂的组织结构
- 实时计算、在线处理

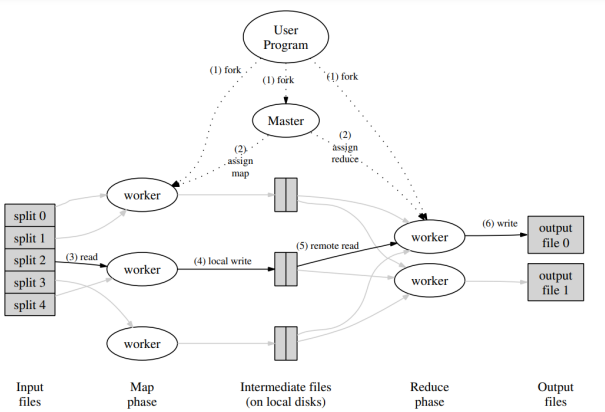


Figure 1: Mapreduce

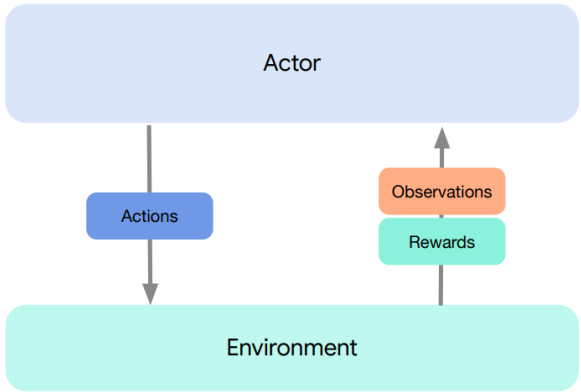


Figure 2: 强化学习

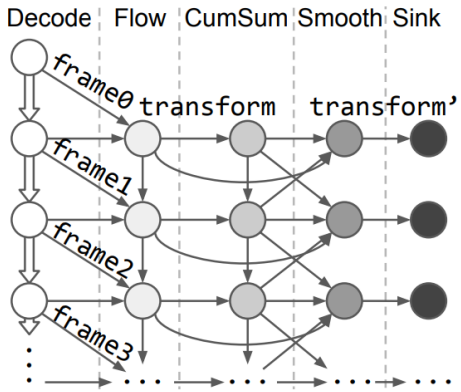


Figure 3: 视频流处理

# 新型计算框架 Ray

## 通用 & 实时

- 细粒度任务调度
  函数为单位
- **分布式内存管理**
  数据随调度移动

## 分布式内存管理

- 依赖解析机制
- 分布式对象存储
  Plasma

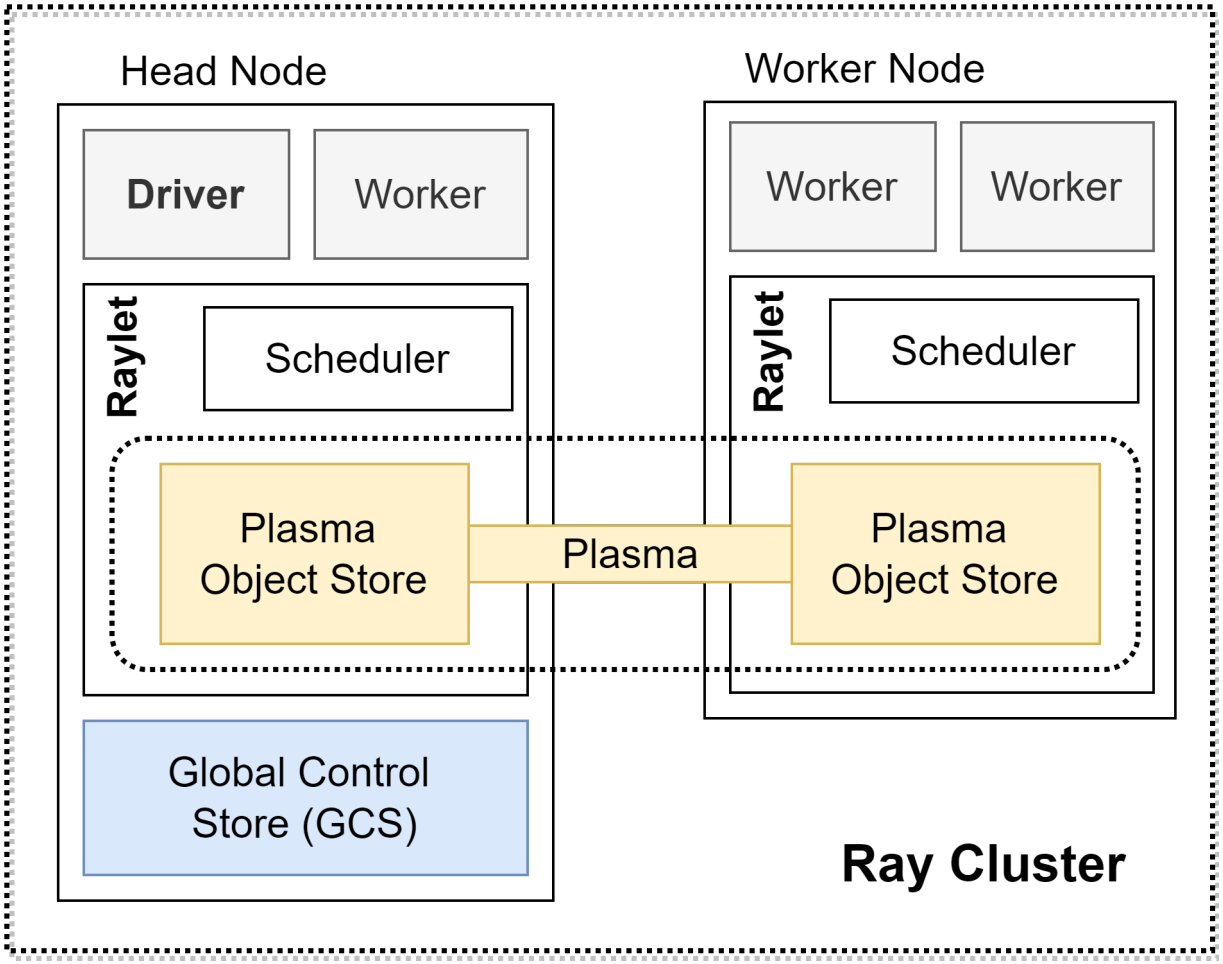


Figure 4: Ray 集群架构

# 高性能集群互联

## 分布式内存存储 Plasma

- 内存对象的网络传输
- 通用性：对象大小差异
- **是否充分利用网络？**

## Infiniband 高速网络

- 低延迟：$\sim 1\mu s$
- 高带宽：200Gb/s
- 链路层容错
- **RDMA**

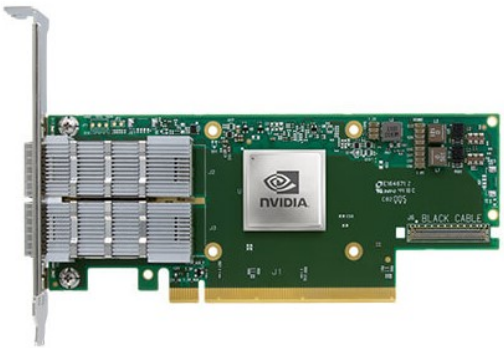


Figure 5: HDR IB 网卡

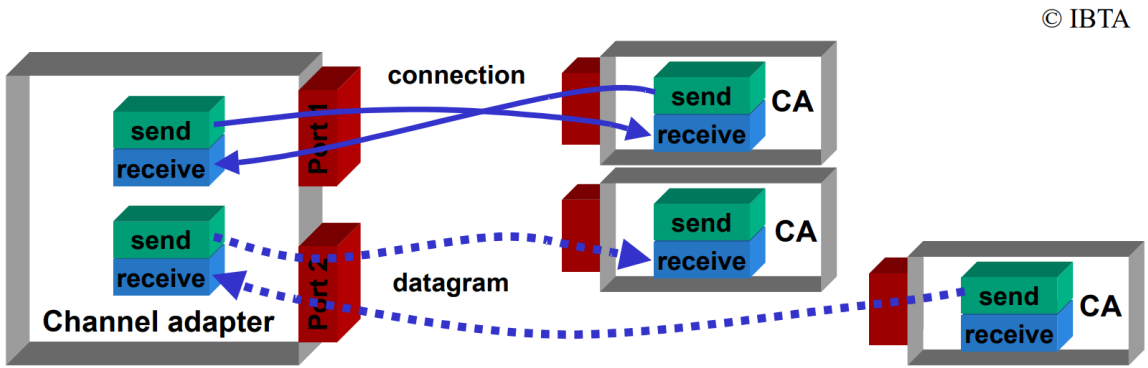


Figure 6: IB 传输层模型

# 远程直接内存访问（Remote Direct Memory Access，RDMA）

## 用户态网络栈

- 内核旁路
- 零拷贝
- CPU 卸载

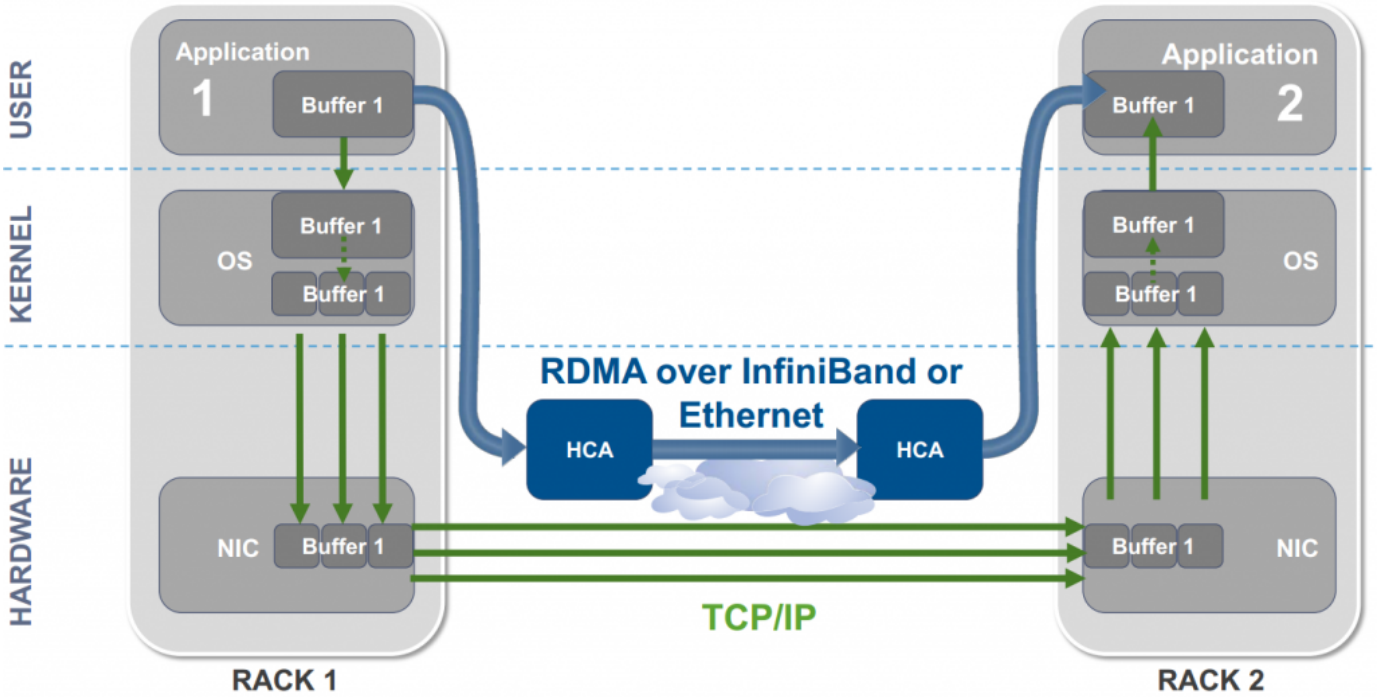


Figure 7: RDMA 示意图

# 内存对象存储

## 节点内

- ~~进程间通信~~
- 共享内存

## 节点间

- 现代智能网卡
- 功能：RDMA 原子操作
- 性能：并发能力

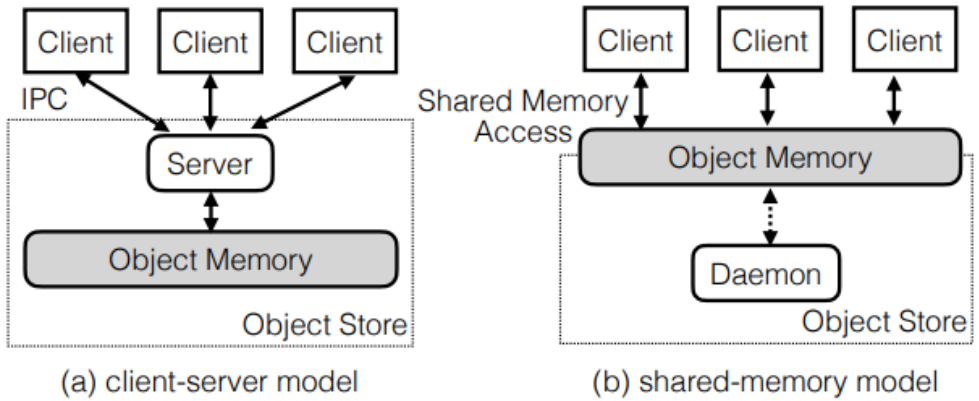


Figure 8: 服务和共享内存架构[1]

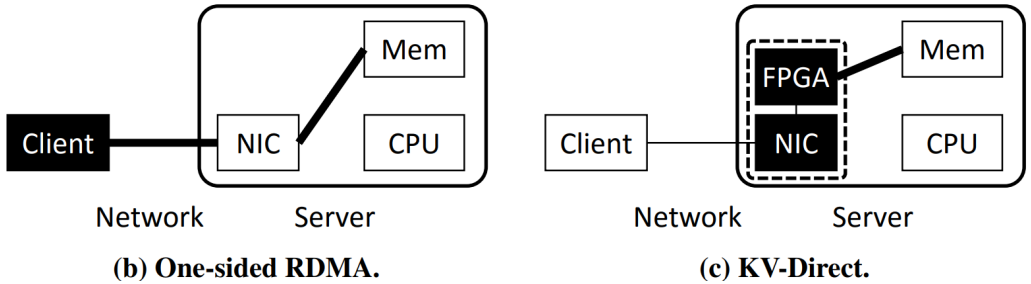


Figure 9: KV-Direct[2]

[1] Rearchitecting In-Memory Object Stores for Low Latency, PVLDB'22

[2] KV-Direct: High-Performance In-Memory Key-Value Store with Programmable NIC, SOSP'17

## 如何为分布式对象存储 (Plasma) 提供原生 RDMA 支持？

我们的实现兼容 以太网 和 Infiniband

## 大小不一的内存对象，如何性能最佳？

基于对象大小的混合传输协议

## 如何测试传输性能？

基于 MPI 的多节点测试；确定了最优决策参数

# 目录

# Plasma 集群架构

## 节点内

- 基于 mmap 的共享内存
- 无网络开销
- 常数延迟读取

## 节点间

- 对象分布信息： Redis
- Manager 拉取对象
- **套接字通信**

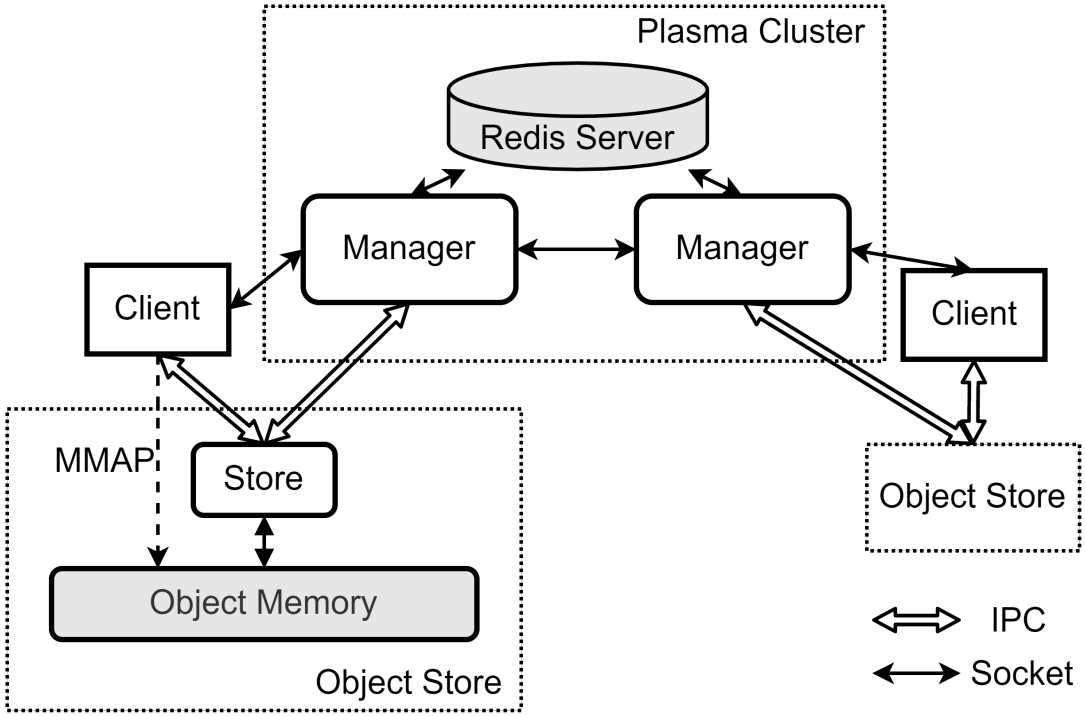


Figure 10: Plasma 集群架构

# 传输性能测试

## Plasma vs. Redis

- 小对象延迟相似
- 大对象吞吐低

## 分析

- **套接字通信**
- 元数据访问
- 本地内存分配

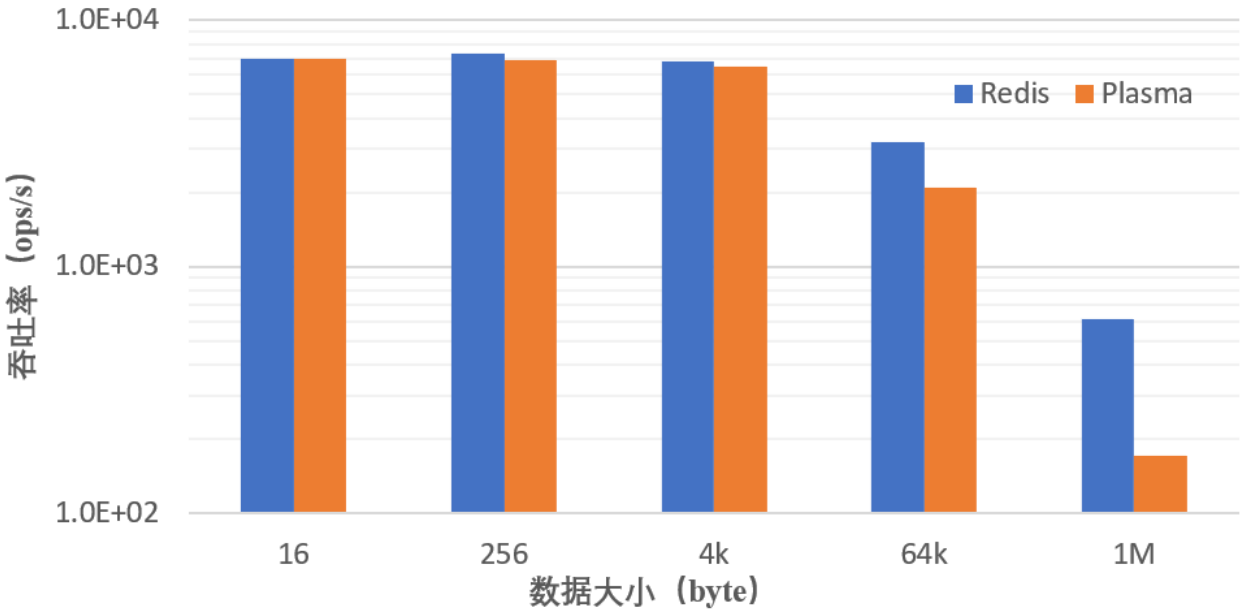


Figure 11: 传输性能测试

# 目录

# 基于 RDMA 的对象传输架构分析

## 双边通信协议

- 预注册的发送/接收缓冲区
- 避免内存注册

## 单边通信协议

- 原地注册的缓冲区
- 避免内存拷贝

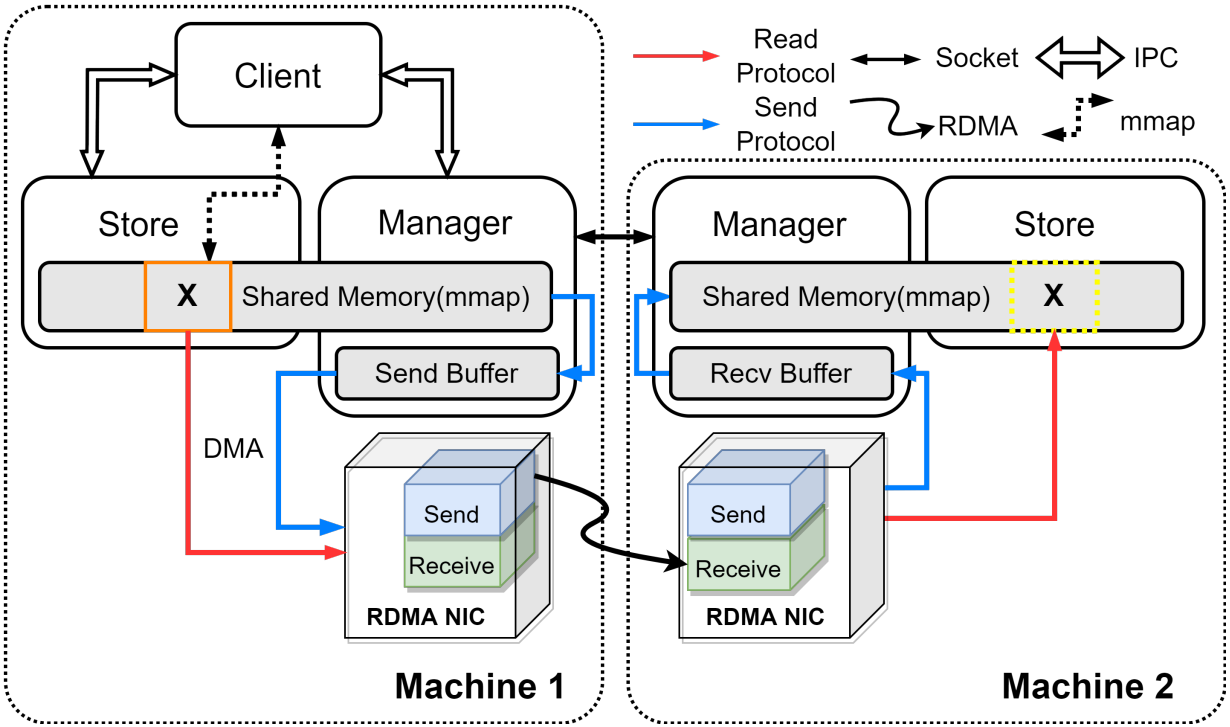


Figure 12: 基于 RDMA 的对象传输架构

# 双边传输协议 (基于 RDMA Send)

## 消息格式

- PLASMA_TRANSFER: 发起对象 X 的传输
- PLASMA_DATA: 返回 X 的元数据
- data: 对象数据

## 分析

- 清空缓冲后同步 (ACK)
- 无内存注册
- 适合传输小对象

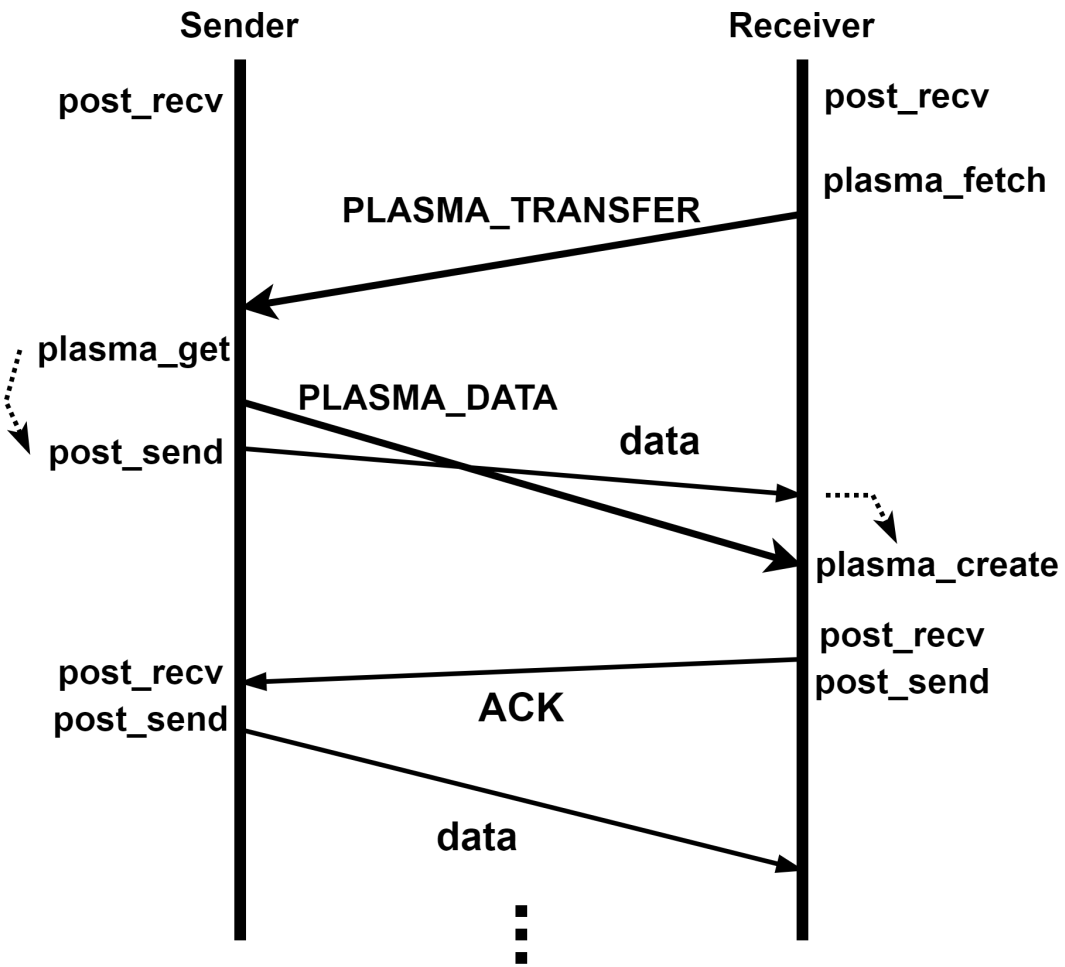


Figure 13: 双边传输协议

# 单边传输协议 (基于 RDMA Read)

## 消息格式

- PLASMA_DATA: 加入远端地址信息
- Read: 单边读

## 分析

- 即时注册/释放对象地址
- 零内存拷贝
- 适合传输大对象

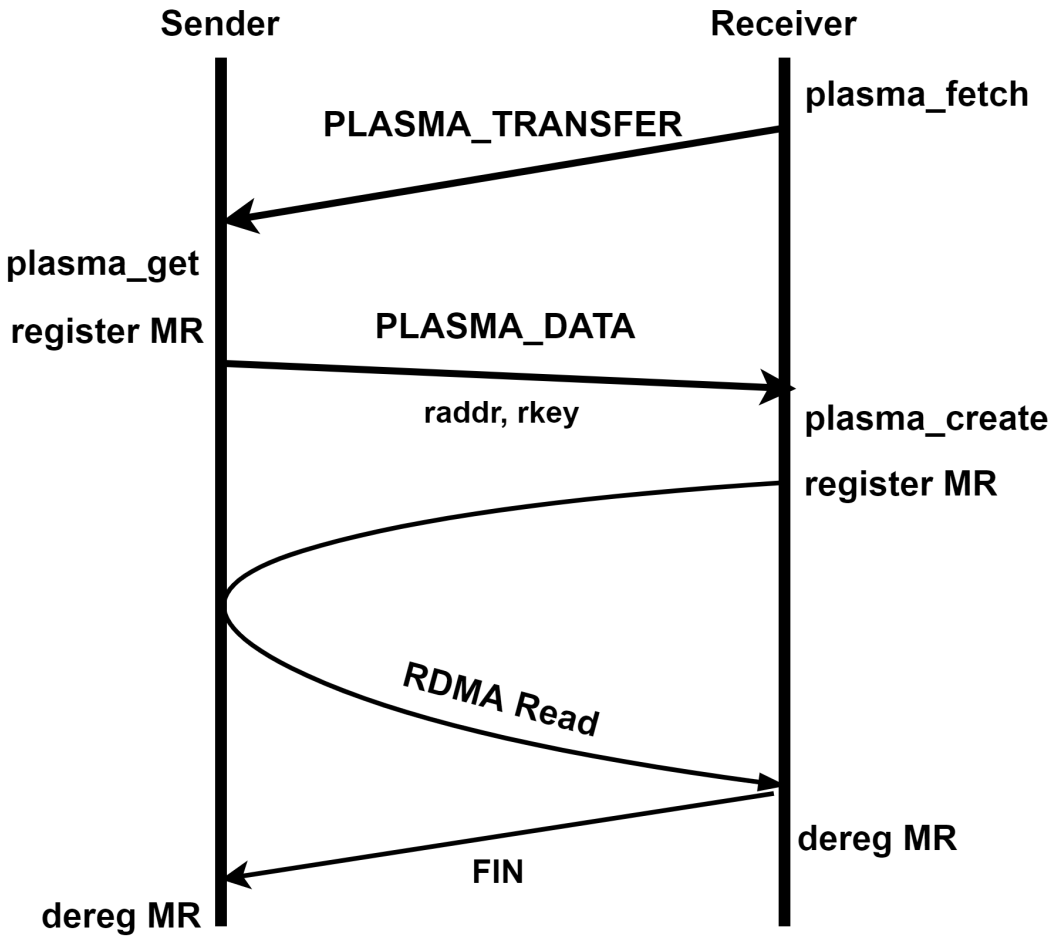


Figure 14: 单边传输协议

# 目录

# 优化实现的传输性能

## 总结

- 32KB 时切换协议：内存注册 vs. 内存拷贝
- 小对象：RDMA Send > RDMA Read ≈ Socket，1KB 15% 提升
- 大对象：RDMA Read > RDMA Send > Socket，4MB 7 倍吞吐

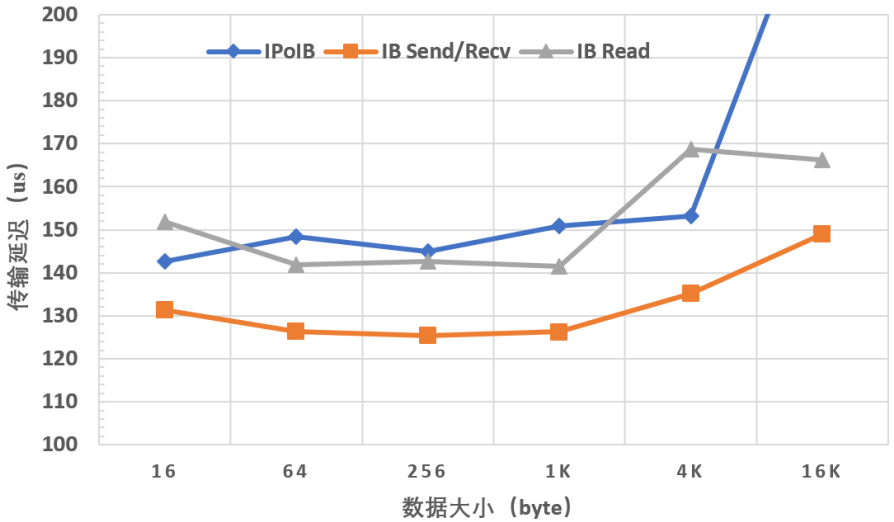


Figure 15: 小对象传输性能

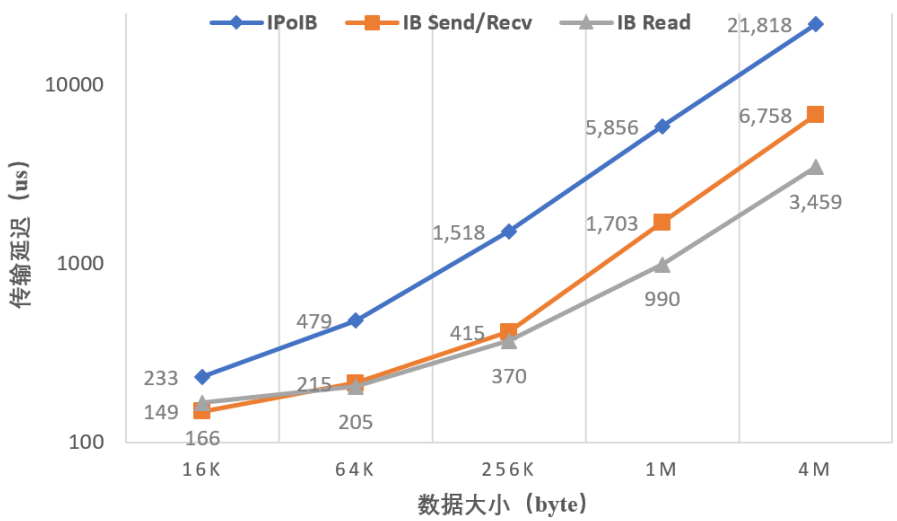


Figure 16: 大对象传输性能

# Plasma(优化后) vs. Redis

## 总结

- 功能上：分布式存储 (多副本)、分布式访问
- 性能上：常见大小对象的单机吞吐 **仍然更优**

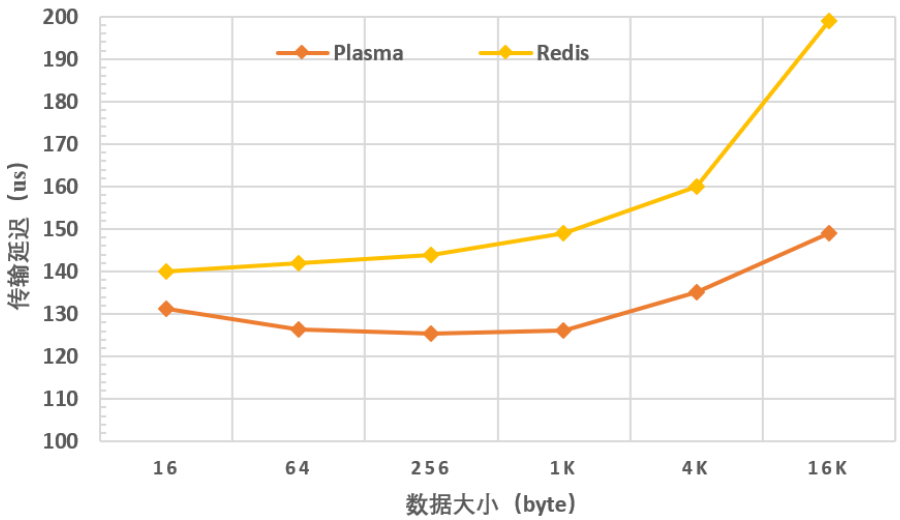


Figure 17: 小对象传输性能

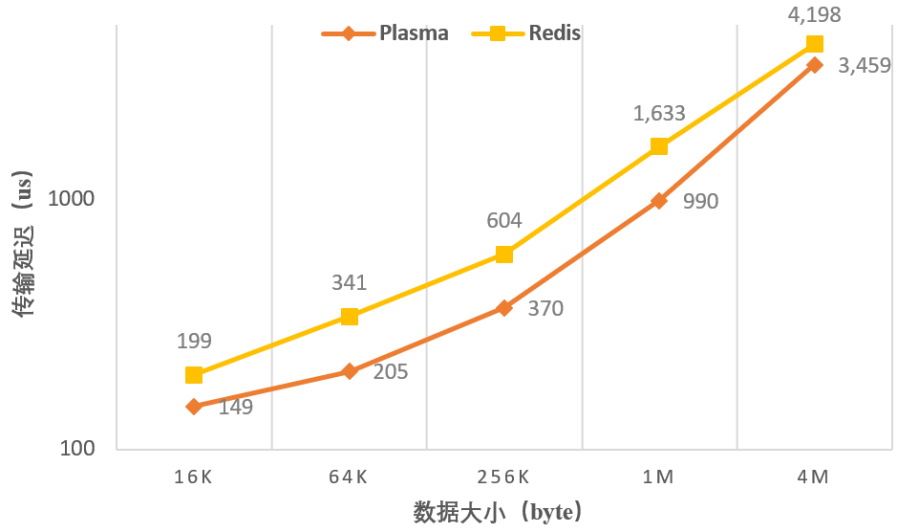


Figure 18: 大对象传输性能

# 目录

## 工作总结

- 分布式内存存储 Plasma 的架构和性能分析，挖掘出网络性能瓶颈
- 基于 RDMA 通信设计了一种全面优于原实现的对象传输机制
- 基于 MPI 的多节点传输性能测试验证了明显的性能优势

## 展望

- 基于 RDMA 的 RPC 调用框架：进一步重构调用机制
- 对象管理机制的协同设计：（基于局部性的）启发式调度、集合通信
- 基于网卡对显存 DMA 支持，实现主存、GPU 显存的异构内存存储

# Q & A

## Questions?

Thank you!