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

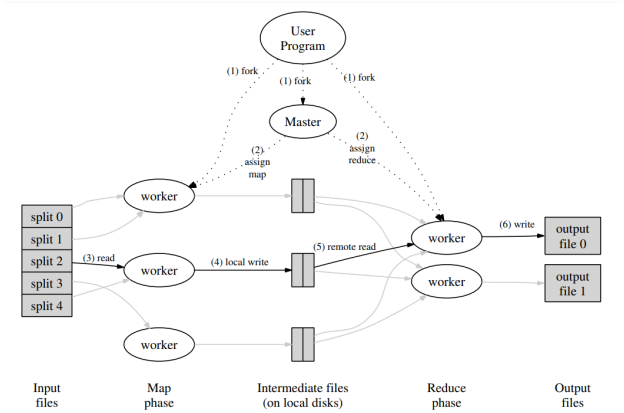

Figure 1: Mapreduce

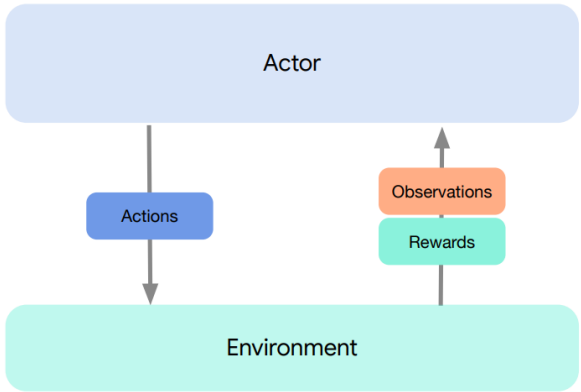

Figure 2: 强化学习

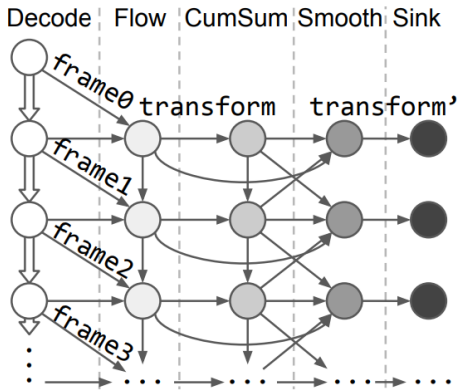

Figure 3: 视频流处理

新型计算框架 Ray

通用 & 实时

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

集群内存管理

- 依赖解析机制
- 分布式内存存储
  Plasma

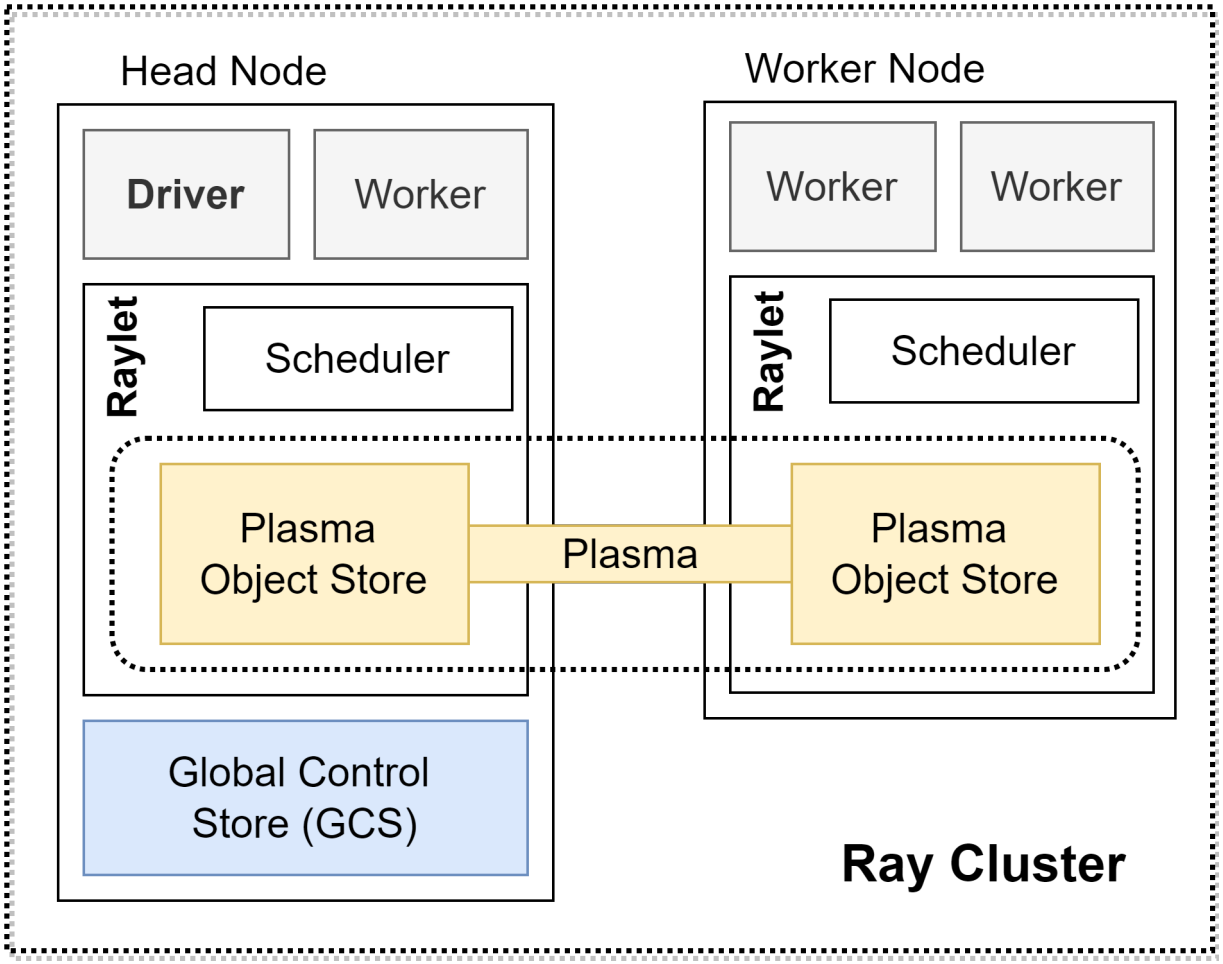

Figure 4: Ray 集群架构

硬件的变化

分布式内存存储 Plasma

- 内存对象的网络传输
- 大小不一的对象
- **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**

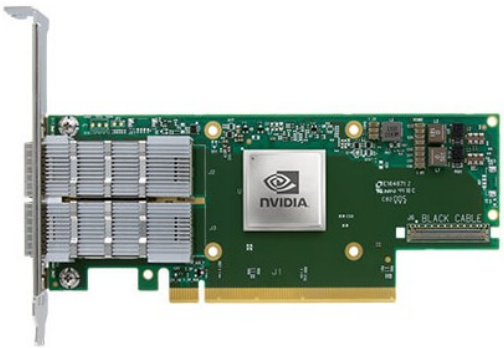

Figure 5: HDR IB 网卡

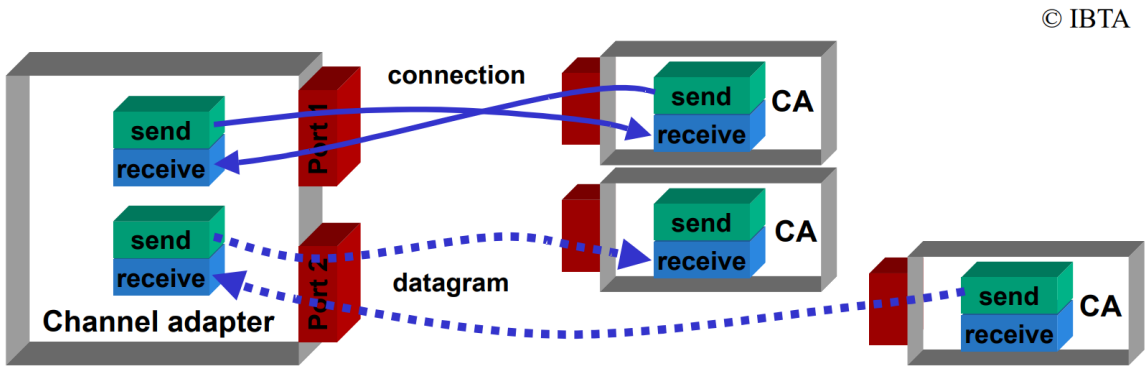

Figure 6: IB 传输层模型

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

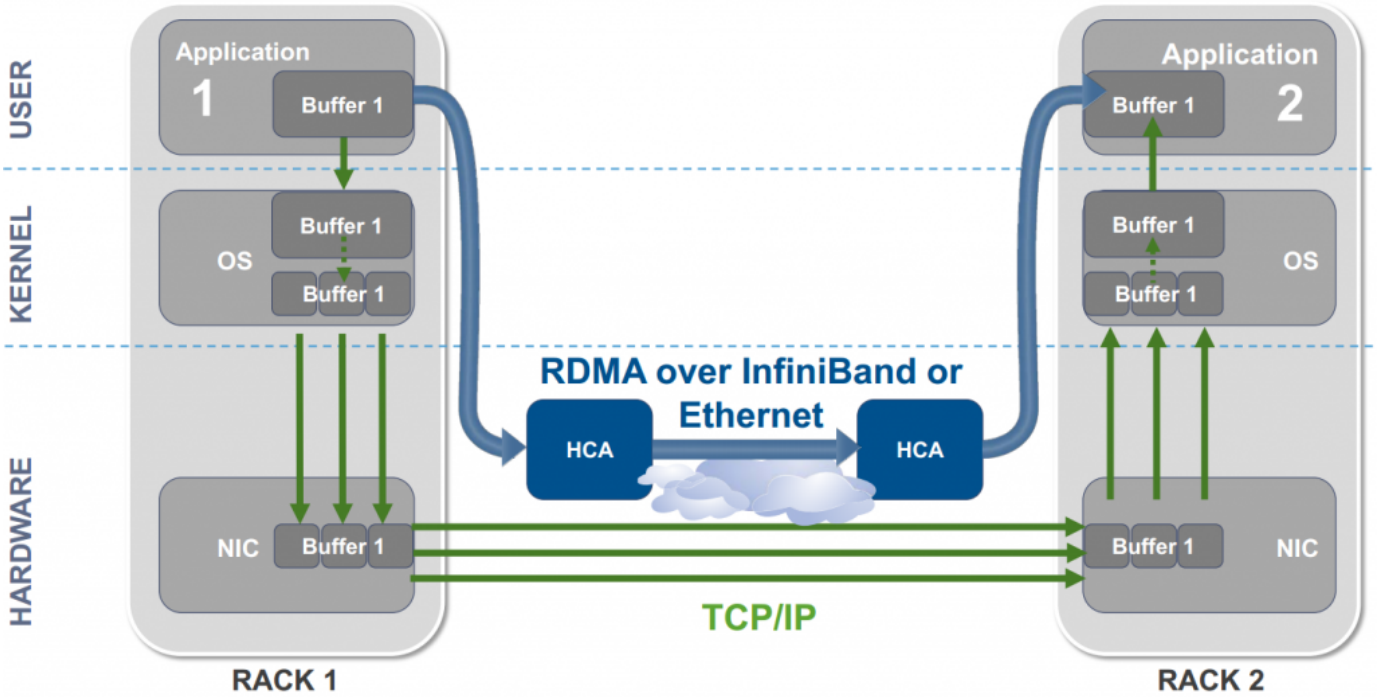

Figure 7: RDMA 示意图

基于 RDMA 的内存系统

如何为 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

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

- 套接字通信
- 对象分布：Redis Server
- Manager 拉取对象

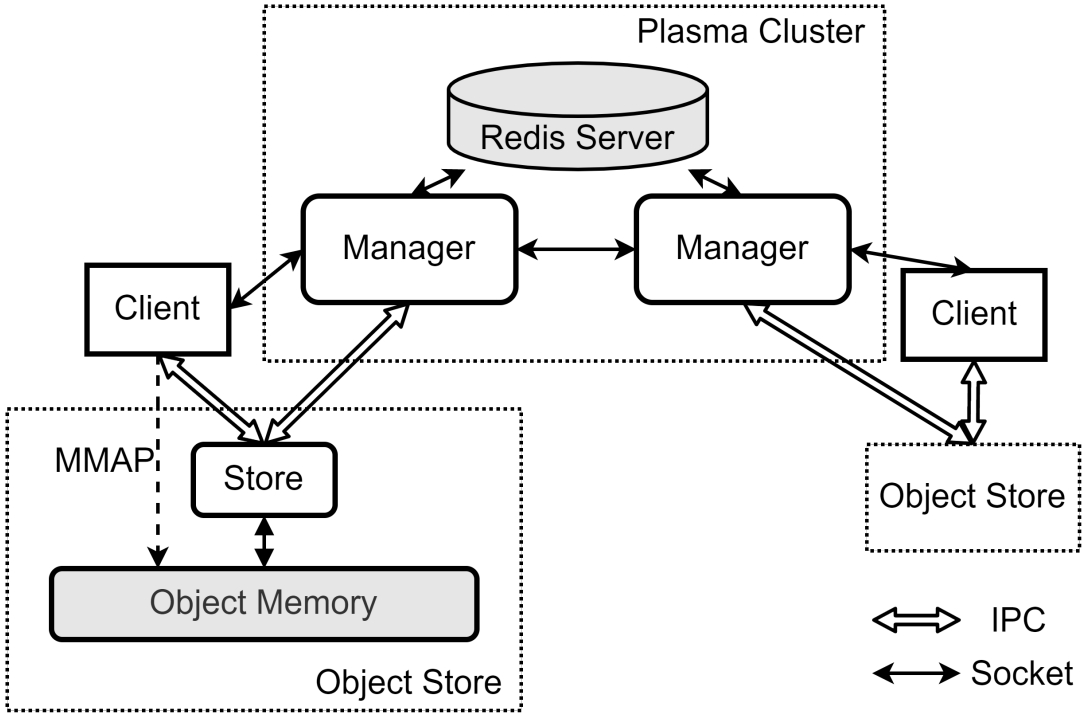

Figure 8: Plasma 集群架构

传输性能测试

Plasma vs. Redis

- 小对象延迟相似
- 大对象带宽低

分析

- **套接字通信**
- 元数据访问
- 本地内存分配

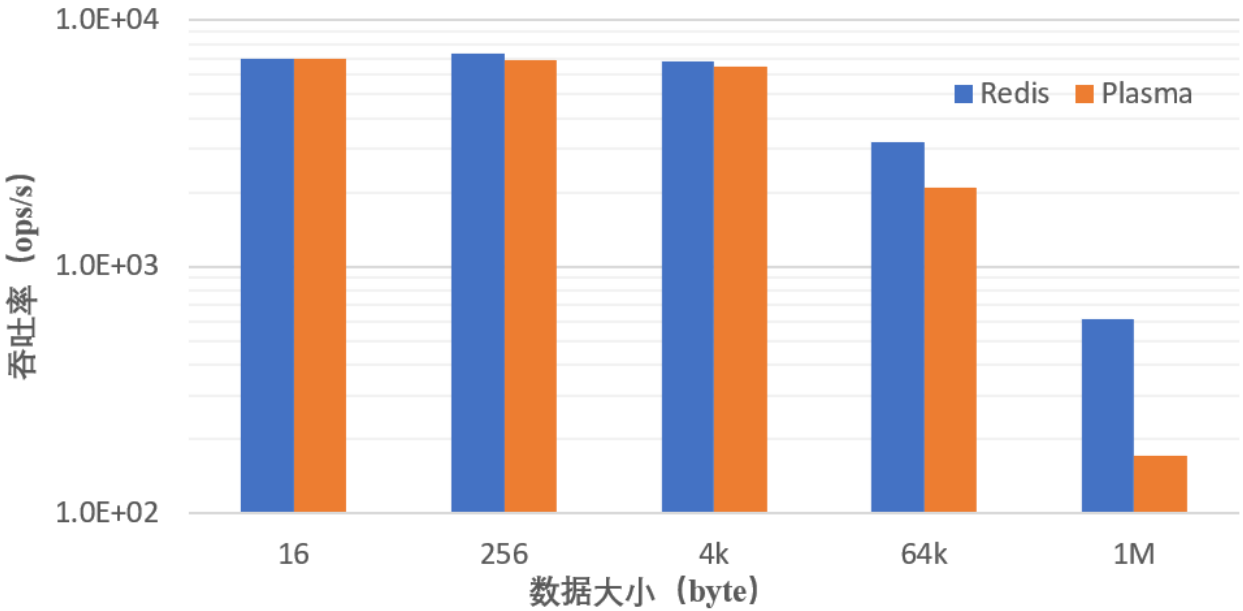

Figure 9: 传输性能测试

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
- 无需重复注册

单边通信协议

- 原地注册的缓冲区
- 零拷贝

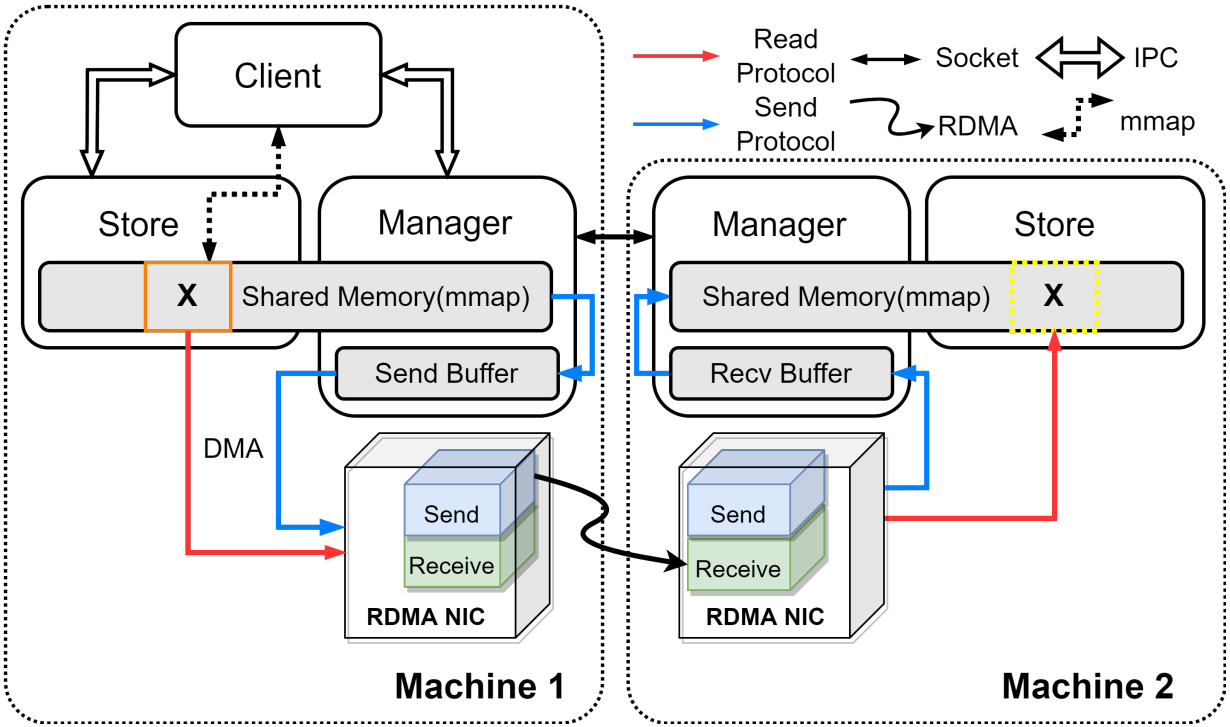

Figure 10: 基于 RDMA 的对象传输架构

双边传输协议

消息

- 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

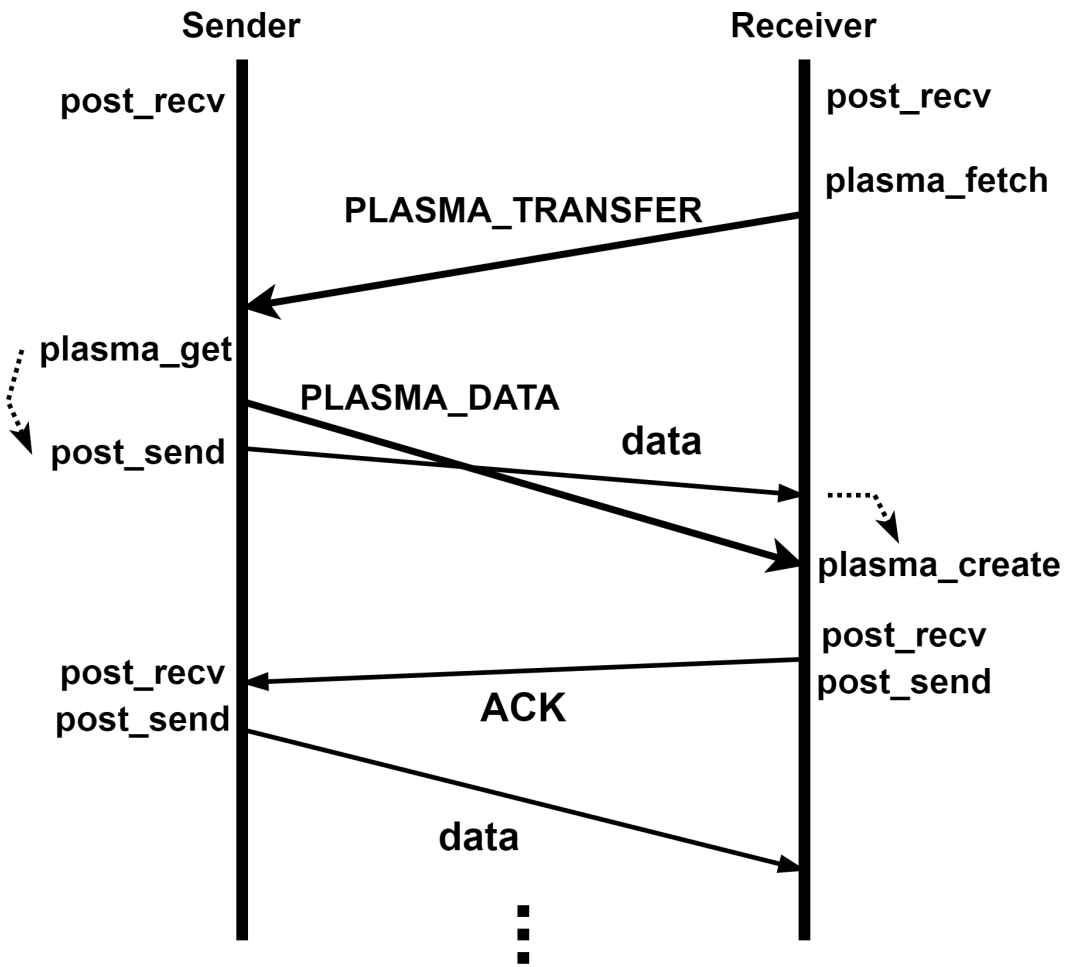

Figure 11: 双边传输协议

单边传输协议

消息

- 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

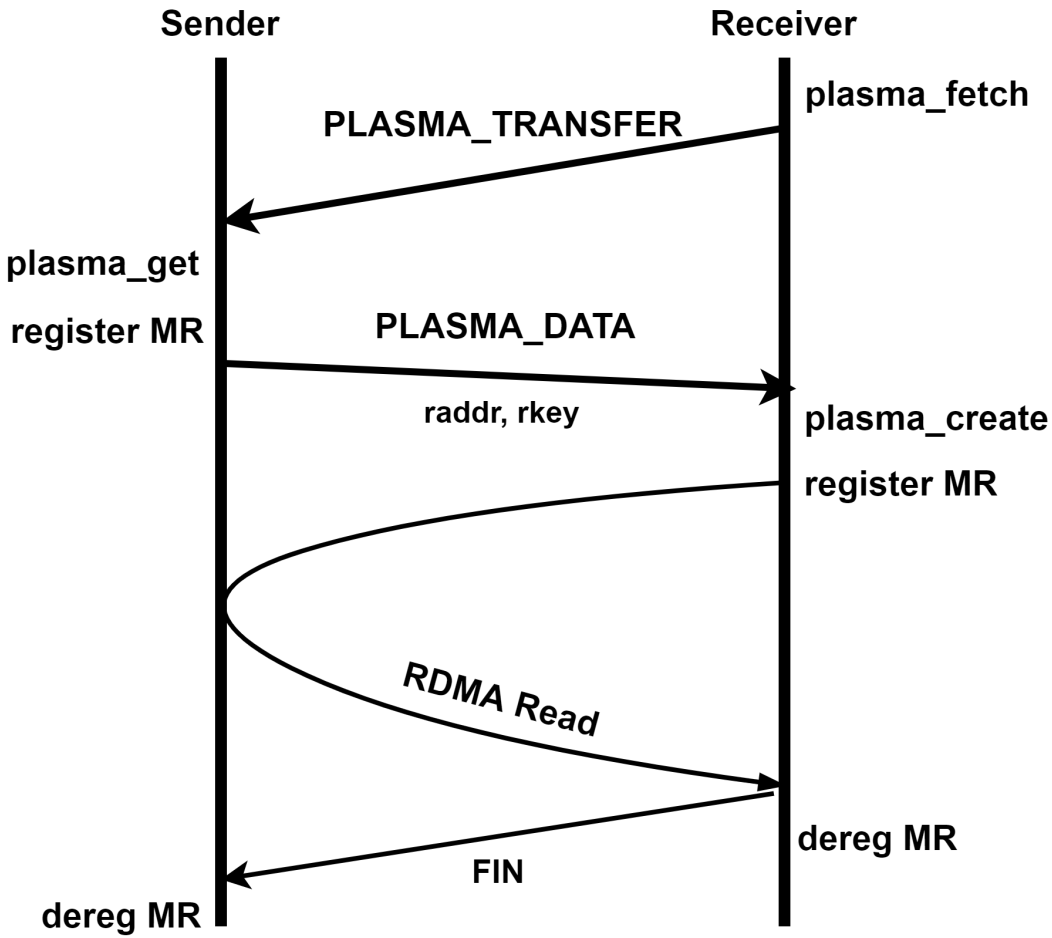

Figure 12: 单边传输协议

工作总结

- 本文的模型结合了卷积网络的特征学习能力与长短记忆网络对整体局部建模的能力，相比于全卷积网络，大幅度地提高了了模型性能
- 大量的对比实验与结果分析证明了模型的有效性

展望

- 模型性能：提高网络的深度来学习更高层次的特征，提高模型效果 (He et al. ResNet, CVPR 2016)
- 模型大小：通过裁剪网络冗余部分 (Han et al. Deep Compression, ICLR 2016 Best Paper) 或使用二值网络减少模型参数 (Courbariaux et al. Binaryconnect, NIPS 2015)
- 训练数据：使用无监督或弱监督的方式训练网络 (Papandreou et al. Weakly-and semi-supervised learning, ICCV 2015)

致谢

感谢每一个帮助过我的人

- 首先要感谢的是我的指导老师的悉心指导
- 感谢师兄师姐、同学的帮助
- 感谢家人的支持
- 感谢答辩委员会的聆听和指导

Q & A

Questions?

Thank you!