

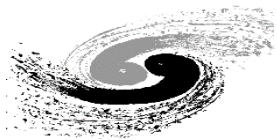


中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

八月份考核报告

张 航 畅
导师：顾旻皓
计算机应用技术

2024. 08



截止到2024.04

➤ 基于分布式内存缓存的在线计算方案

◆ 基于 MINIO 的分布式内存缓存池上线

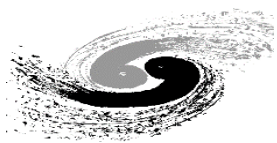
- 读写速率低
- CPU消耗大

◆ 开发缓存池读写接口

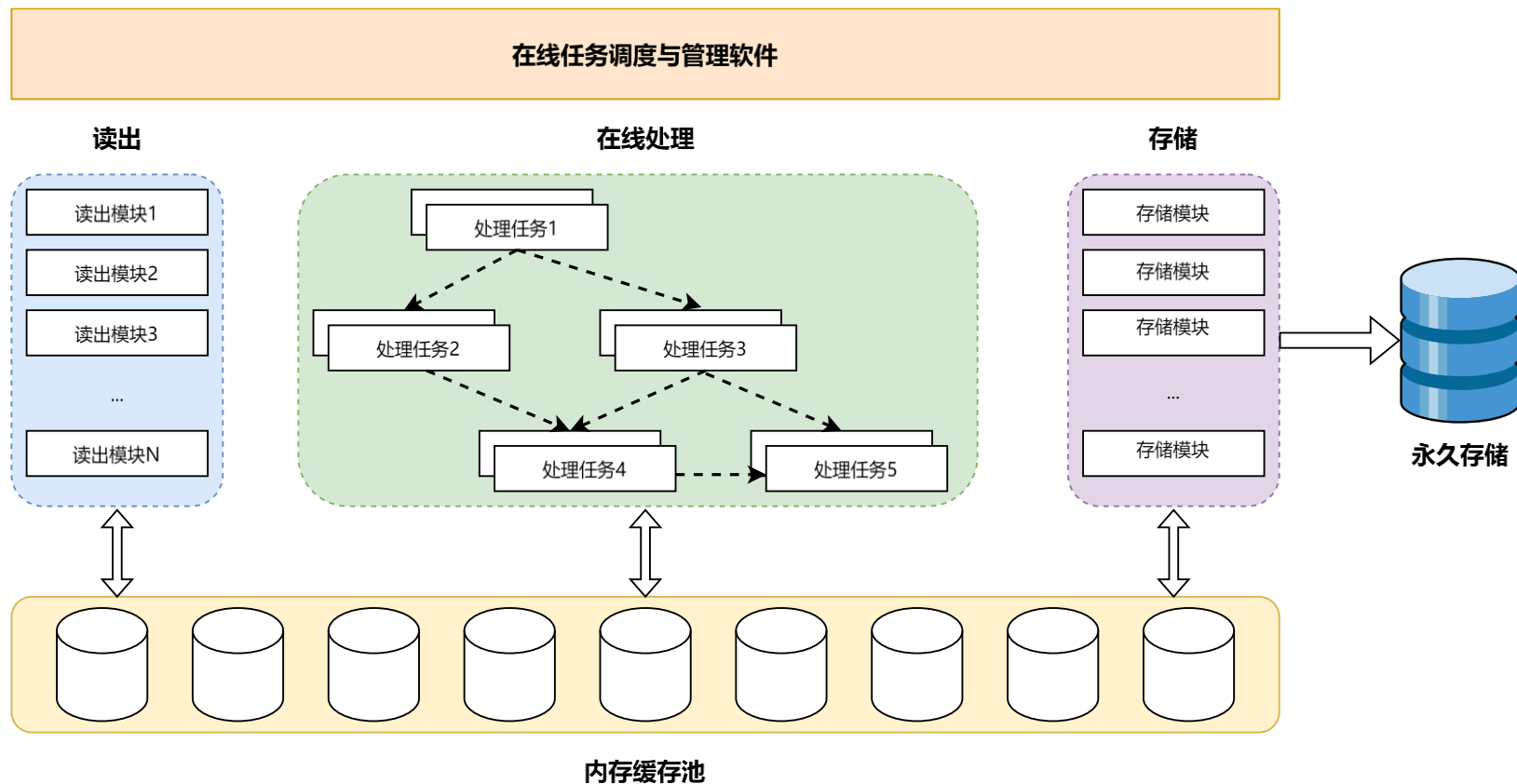
截止到2024.08

➤ 基于分布式内存缓存的在线计算方案

- ◆ 内存缓存池关键技术调研
- ◆ 编写技术方案和开发计划
- ◆ 开发基于 RDMA 的通讯层软件



基于分布式内存缓存的在线计算方案



5月完成博士开题报告，题目为：《基于内存缓存的高吞吐在线计算软件框架及其关键技术研究》

要求：

- 高性能
- 高效率
- 高可靠

关键技术：

- RDMA通讯层
- 分布式内存管理
- 分布式任务调度



基于分布式内存缓存的在线计算方案

6月-8月：详细调研，编写EDR

➤ 阅读文献，调研技术方案

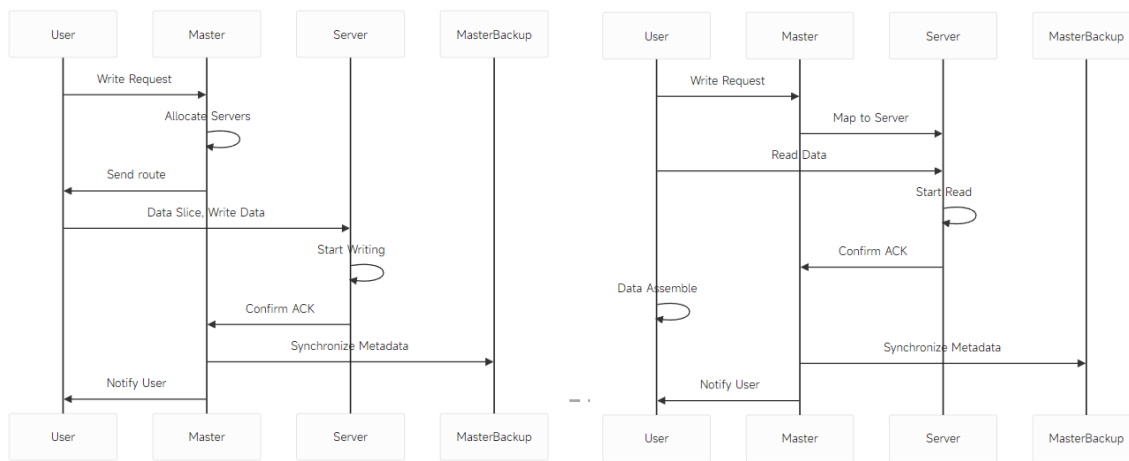
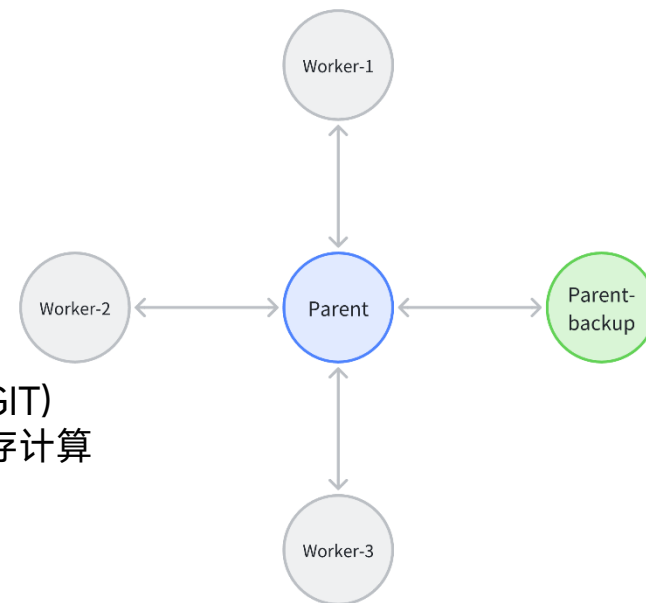
➤ 确定方案：

- ◆ 主从结构
- ◆ 多主备份
- ◆ 统一故障处理管道

➤ 制定业务逻辑

调研内容：

1. 实验室开源产品的源代码
Gengar(HUST), Octopus(THU), FaRM(MS), RDMABox (GIT)
2. 已发表论文：内存缓存池、RDMA、内存解耦理论、内存计算
3. 成熟非内存产品的设计与原理
HDFS, MinIO, Gluster, Crail



开发难点：

1. 主节点事务压力
2. 从节点负载均衡
3. 碎片化问题
4. 故障处理与高可靠

对内存缓存池方案进行全面调研和设计，在开发前尽量考虑充分



基于分布式内存缓存的在线计算方案

开发基于 RDMA 的通讯层软件 HYLIA

- 基于 librdmacm 库开发
- 对 TCP/epoll, RDMA/RoCE 进行 C++ 包装。
- 统一接口, 简单易用
- 性能测试:

```
HyliaComm tcpClient;  
tcpClient.init("tcp", "0.0.0.0", 12345, "client")  
tcpClient.send(data, size);
```

```
HyliaComm rdmaServer;  
rdmaServer.init("rdma", "0.0.0.0", 12345, "server")  
rdmaServer.listen();  
rdmaServer.poll();
```

测试环境: 1GB文件

测试环境: 128KB数据包

	Hylia / TCP	Hylia / RDMA	Hylia / TCP	Hylia / RDMA	QPERF / TCP	QPERF / RDMA
速率	2.2 GB/s	5.2 GB/s	2.2 GB/s	5.3 GB/s	2.5 GB/s	9 GB/s
CPU利用率	100%	80%	100%	60%	100%	-
延迟	512us	167 us	512us	167 us	320us	160 us

对RDMA尤其是RoCE的应用, 还需要进一步了解和优化



其他

成员管理系统

- 为 JUNO 定制功能开发
- 推送 6 个版本更新，12 项功能新增，20 余项问题修复和优化
- 9 月份开始 JUNO 公开测试

论文

- IEEE-TNS 论文投稿（5月20日），小修（7月18日）



总结与下一步计划

总结

➤ 内存缓存池:

- ◆ 博士开题 《基于内存缓存的高吞吐在线计算软件框架及其关键技术研究》
- ◆ 调研, 编写项目实施方案, 制定开发计划
- ◆ RDMA 通讯层软件开发

➤ 成员管理系统:

- ◆ 为 JUNO 开发新版本, 9月份公开测试

➤ 其他:

- ◆ IEEE-TNS 投稿, 小修

下一步计划

➤ 内存缓存池完成 Prototype 开发

- ◆ Parent 节点: 可以管理从节点, 处理事务
- ◆ Worker 节点: 可以与用户交换原始数据, 文件分片管理