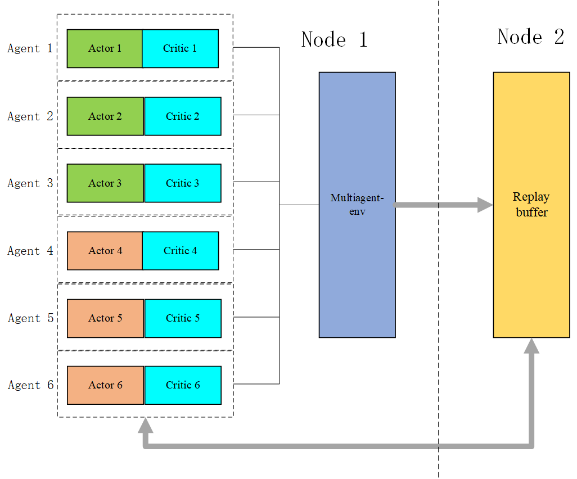
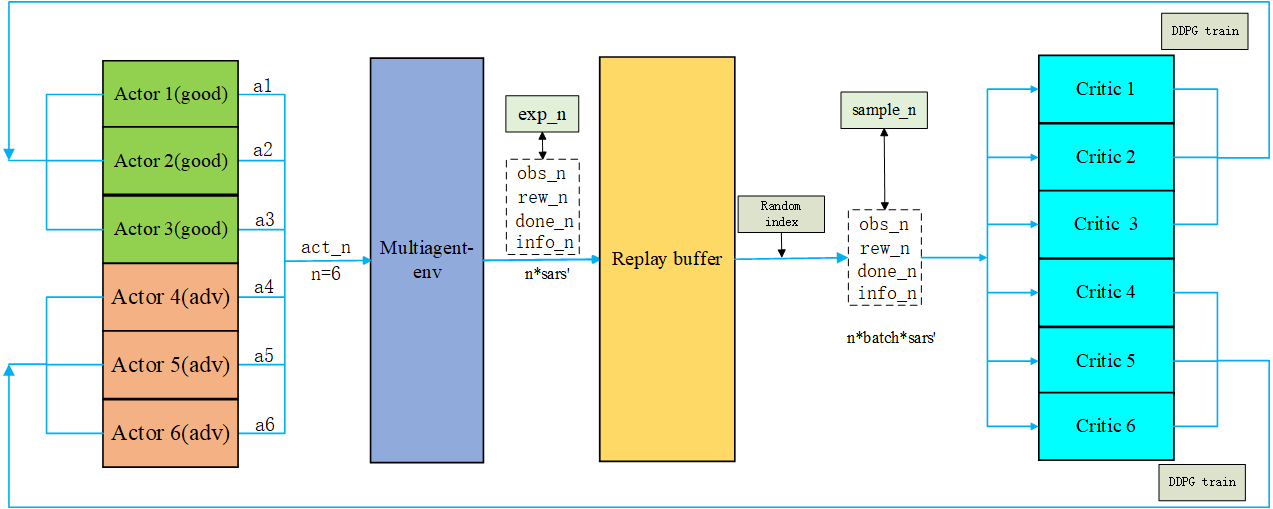
1. 上周计划

4月4日-8日

绘制基准代码的框图，具体分析其和系统框图之间的差异；使用MPI进行基本编程，通信测试等。

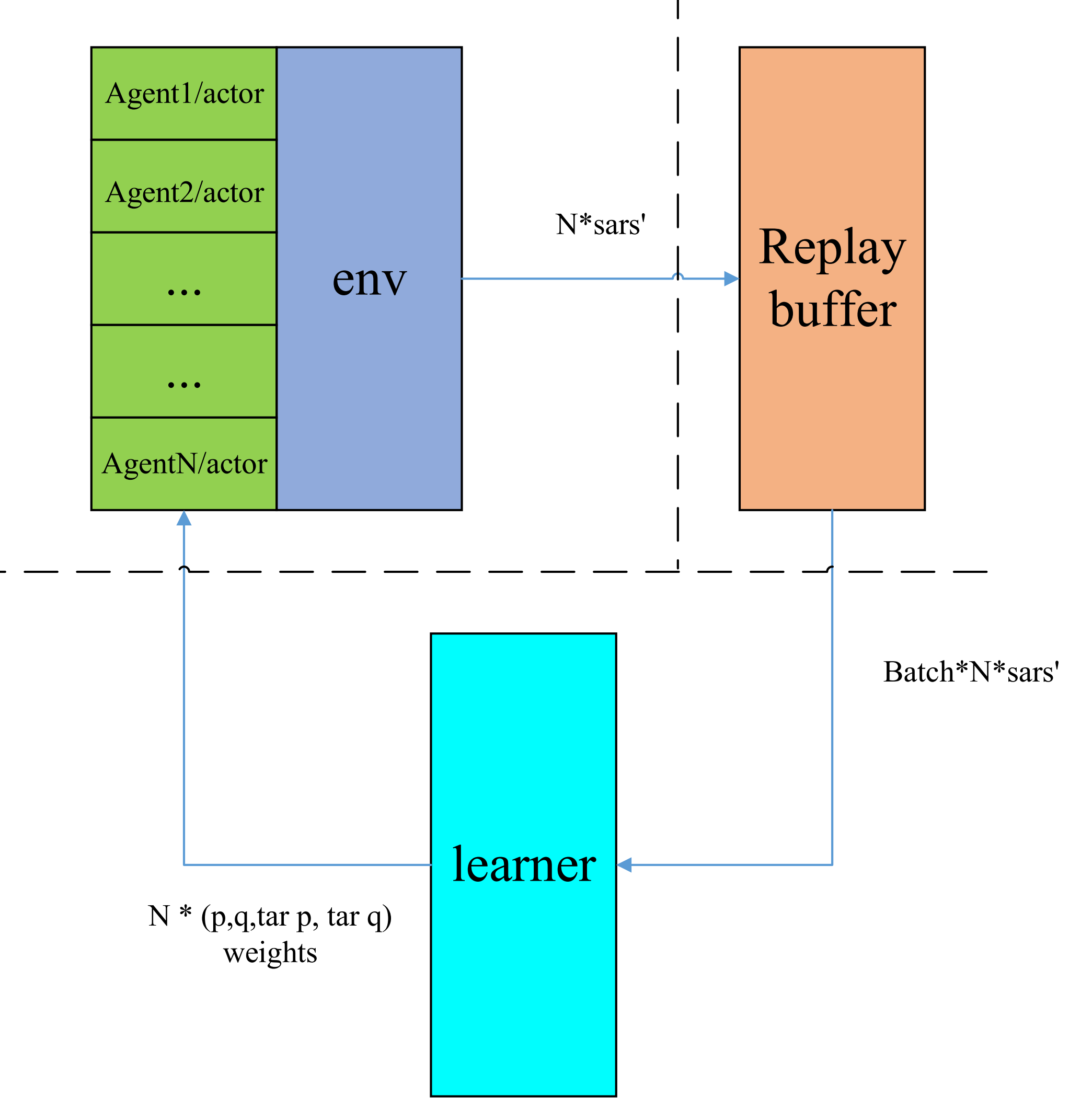
分离结构框架图



分离结构数据流向图

4月8日-11日

根据系统框图，等价拆分基准代码，将replay拆分出来；结合系统框图，整理出几个部分之间需要通信的量和时机，用MPI实现。



1. 完成情况

4月4日，进行任务规划，分工。林：明确基准，张冠宇版本的代码，以上述框图为框架，细化张代码中不一致的部分和其实现的细节。

4月5日，李，以上述框图为框架，用MPI实现3个部分之间的通信，首先熟悉MPI编程。

4月7日，了解进度，明确林和李各自的侧重点，实现内部高聚合、外部松耦合，过程中讨论做好接口设计，将问题尽量拆分成算法问题、分布式问题、综合问题，便于区分性解决。

4月8日，组会上林和李报告进度。

4月9日，和林讨论其报告中关于共享env的动机。经过讨论，该实现的本质是actor乱序发给env，env乱序反馈给actor。这个想法破坏了强化学习中的t和t+1时刻之间的因果关系，和replay buffer的随机采样的动机混淆了。

因此，需要明确一下，首先是Markov性质，在t时刻决策时，不需要历史信息，即t-1时刻及之前的信息，仅和当前状态St有关，根据策略pi做出动作At，环境根据St和At，返回奖励Rt+1和下一个状态St+1。其次，环境本身，除了明确说存在随机因素的外，St和At到Rt+1和St+1的转移是确定的。对于不同的智能体，其策略pi的不同，可能出现不同的轨迹，注意这个特异性是智能体引入的，而不是环境。再次，replay buffer使用的动机，回到Markov性质和强化学习的目的，根据当前状态St，学习状态函数S（St）、状态-动作函数Q（St，At）、策略pi。智能体和环境交互产生的样本之前前后相关性强，很容易产生过拟合。通过大量收集样本，然后随机采样，可以有效地减弱样本之间的相关性，减弱过拟合对性能的影响。

接下来重点是，按照框图将张的代码拆分出来，验证正确性。

4月10日，和李讨论方案二中，learner计算三次梯度才更新给actor的动机。align到串行版本。actor产生3个样本，learner计算3次梯度，更新3次。方案二将更新改到了1次。

首先回顾了一下前面强化学习课中几种学习方法中更新机制。MC是使用当前策略完成一局，然后使用中间产生的样本更新策略。TD(0)是每一步后，用产生的当前一个样本对策略进行更新，然后使用新的策略进行下一步，如此迭代直到该局结束，直到策略收敛。到了深度强化学习，在TD(0)的基础上，出现了每隔若干步，策略和价值函数异步更新的方式。因此，从提高actor和learner的计算资源利用率的角度看，方案二的方式显然不是最优的，可以通过调整更新的频次，将方案一变成方案三。

实现非阻塞通信的方案二和三。

4月11日，组会上林和李报告进度。

林拆分了代码，进行了训练，但是无法验证是否正确。随机初始化训练分数一直为负。使用张提供的模型，可以得到正分。

李实现了阻塞通信的方案一和二。原始串行版本，200ms左右，MPI串行版本（方案一），500ms左右，MPI优化版本（方案二），300ms左右。非阻塞通信，没有达到预期目的。

会后经过讨论，看过代码后，是MPI编程问题，需要进一步联系改进，理解非阻塞通信的本质，Isend/Irecv和wait配对，在中间插入计算，实现通信与计算重叠隐藏。对于如何实现发了不管、收了就更新，可能现在的MPI原语支持不了，需要考虑类似zeroMQ等的消息队列来实现。

1. 下周计划

4月11日-15日

林，确定一个正确性确认方式，确认拆分的正确性。

李，进一步熟悉MPI编程，理解并正确实现需要的通信方式。

4月15日-4月18日

整合基准重组代码和MPI通信，得到里程碑代码1。

4月18日-4月25日

里程碑代码1进行单节点（上图两个模式之一，从左到右顺序实现）训练，和基准代码训练效果进行比较，控制超参数设置相同，随机数相同；分析训练过程中通信的瓶颈，使用MPI高级编程或者其他机制来优化通信性能。