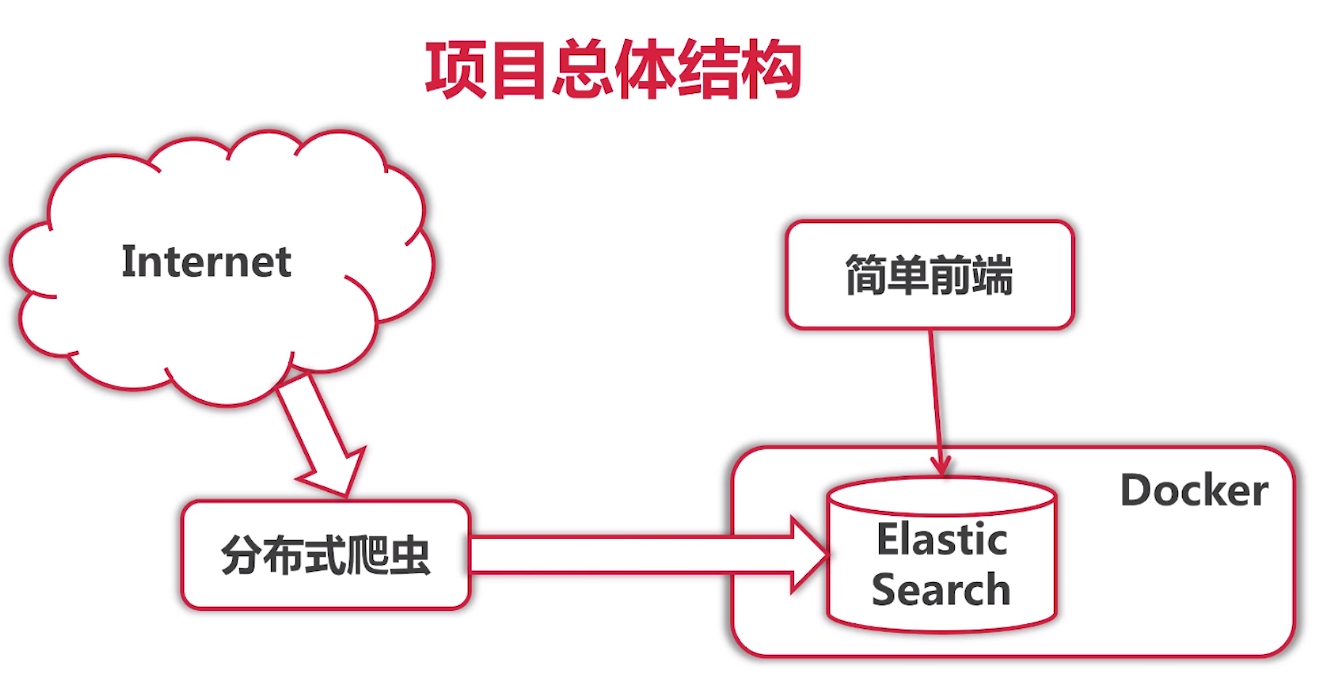
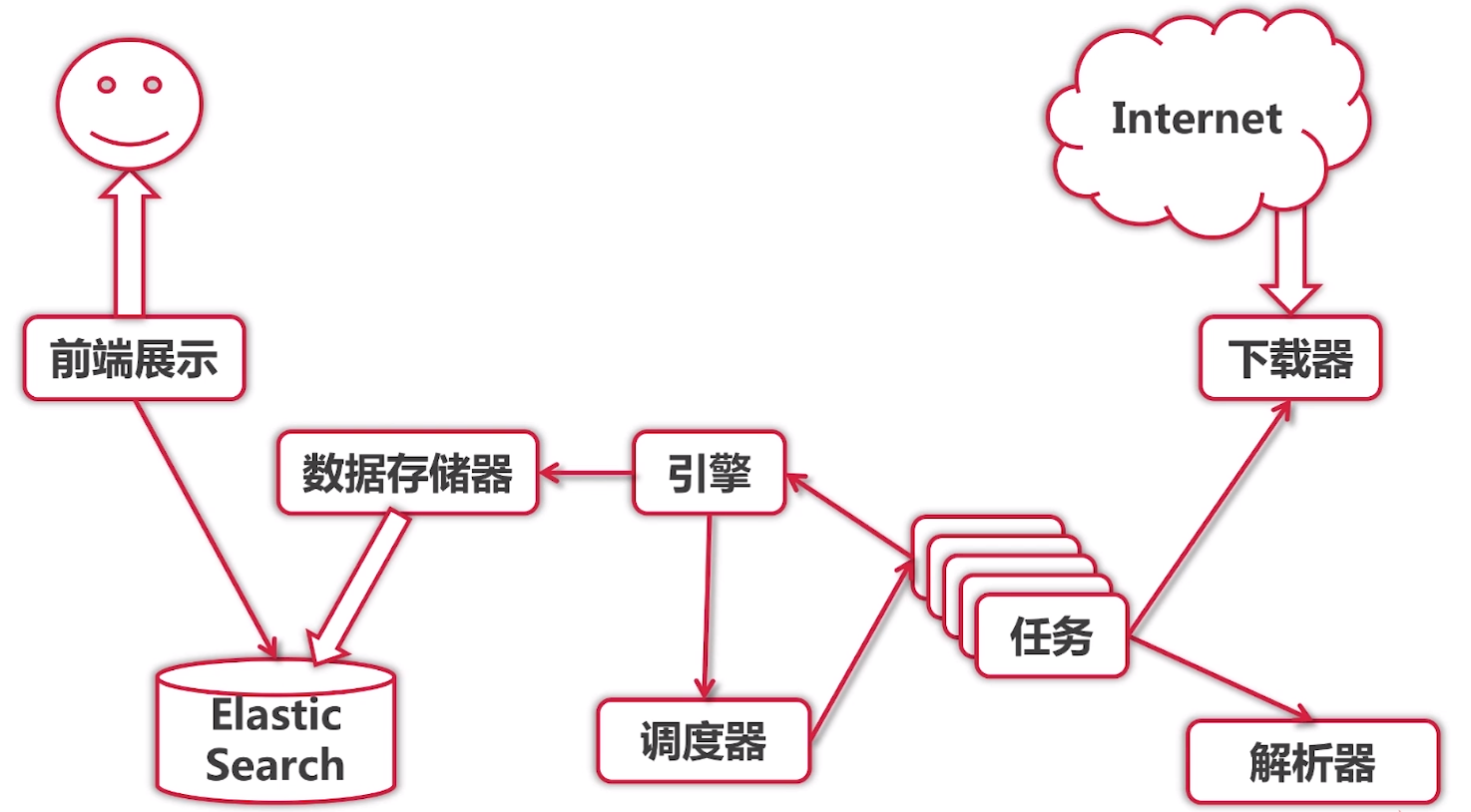
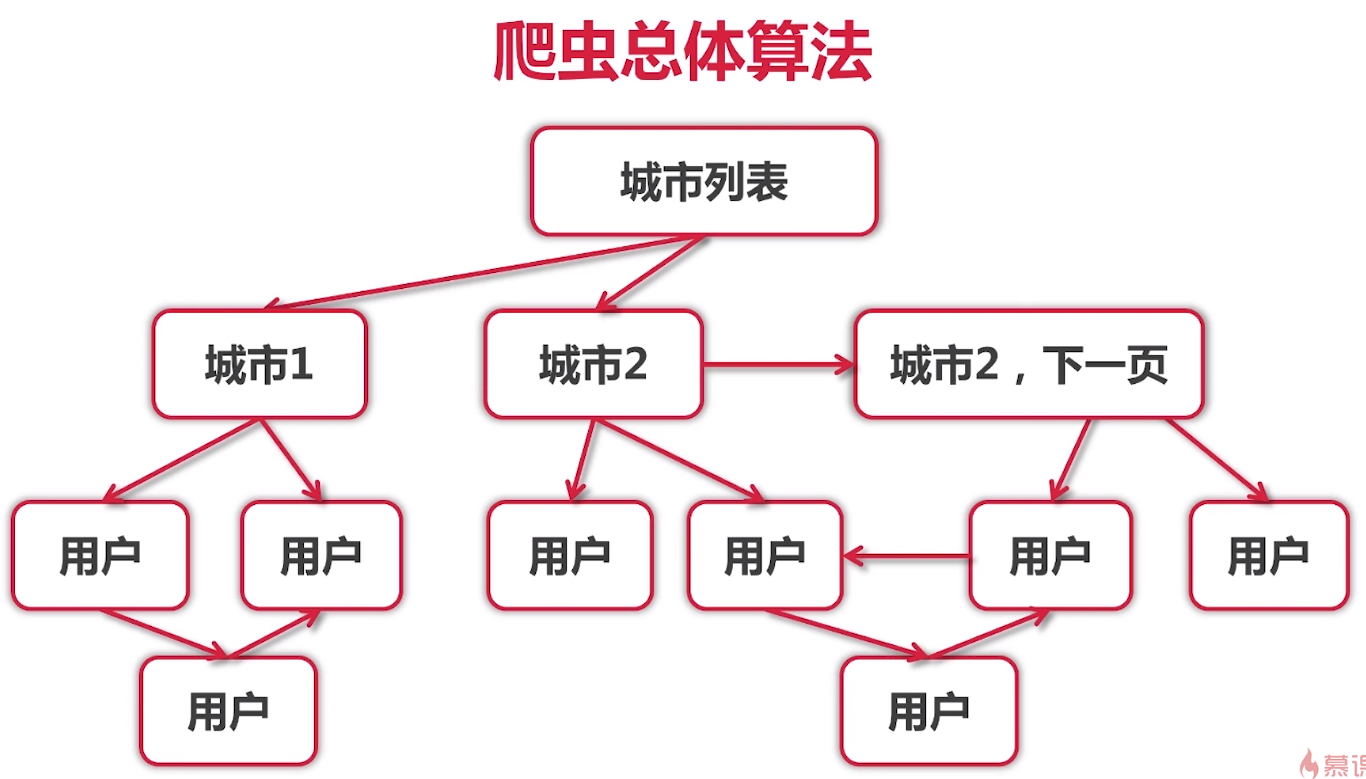
# 项目总体结构

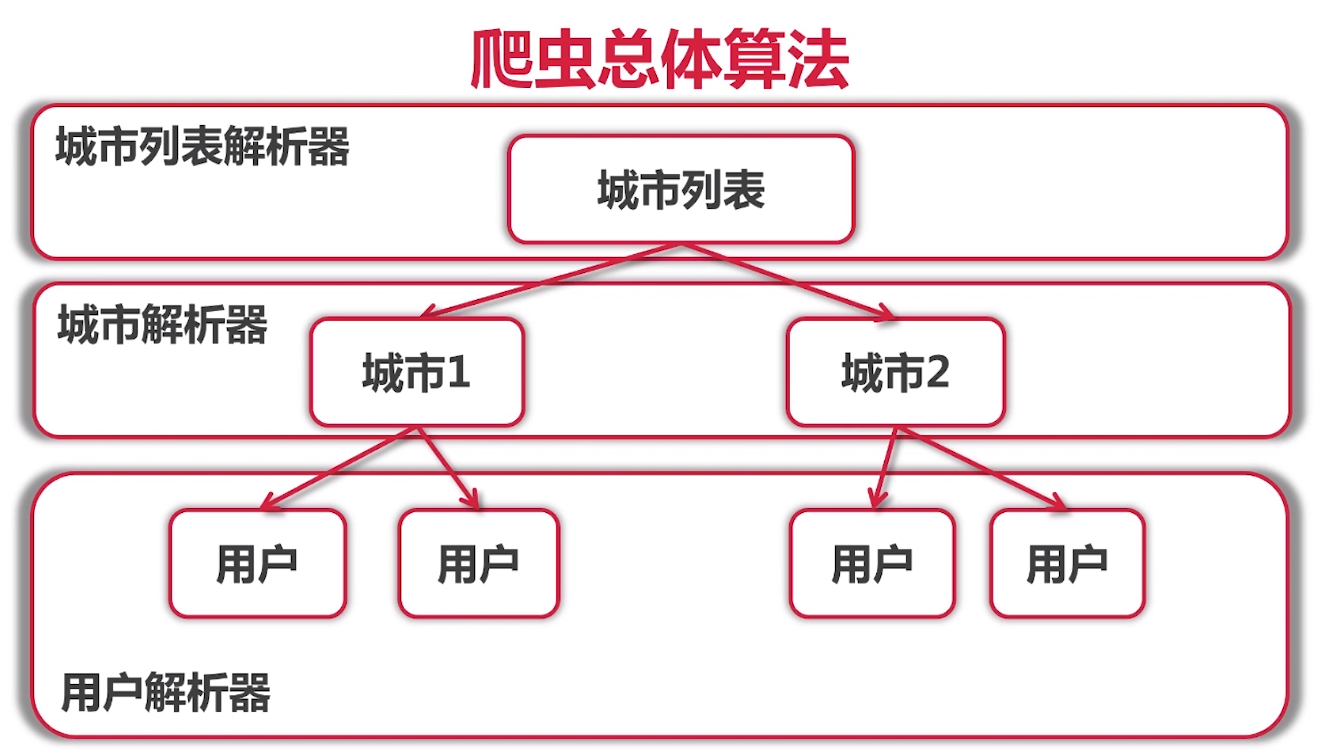




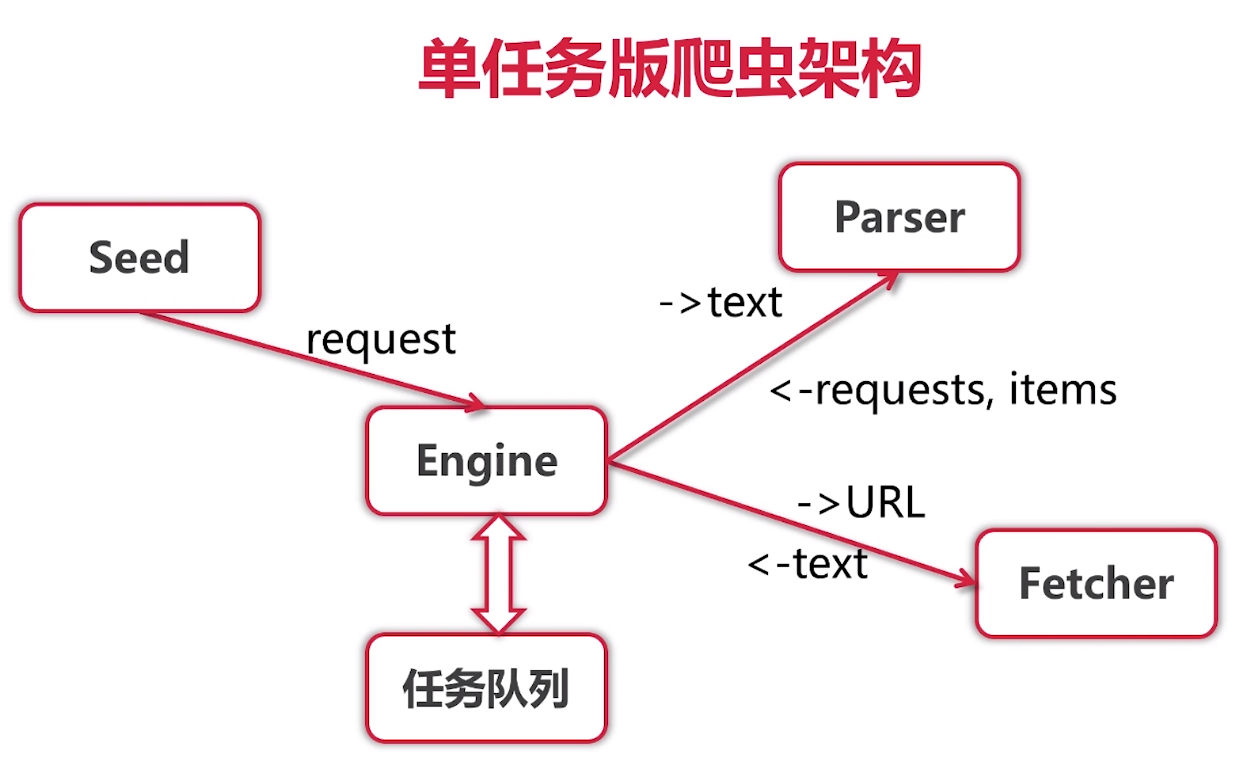
# 爬虫总体算法



每个页面的内容不一样，所以每个页面创建一个解析器



# 单任务版爬虫架构

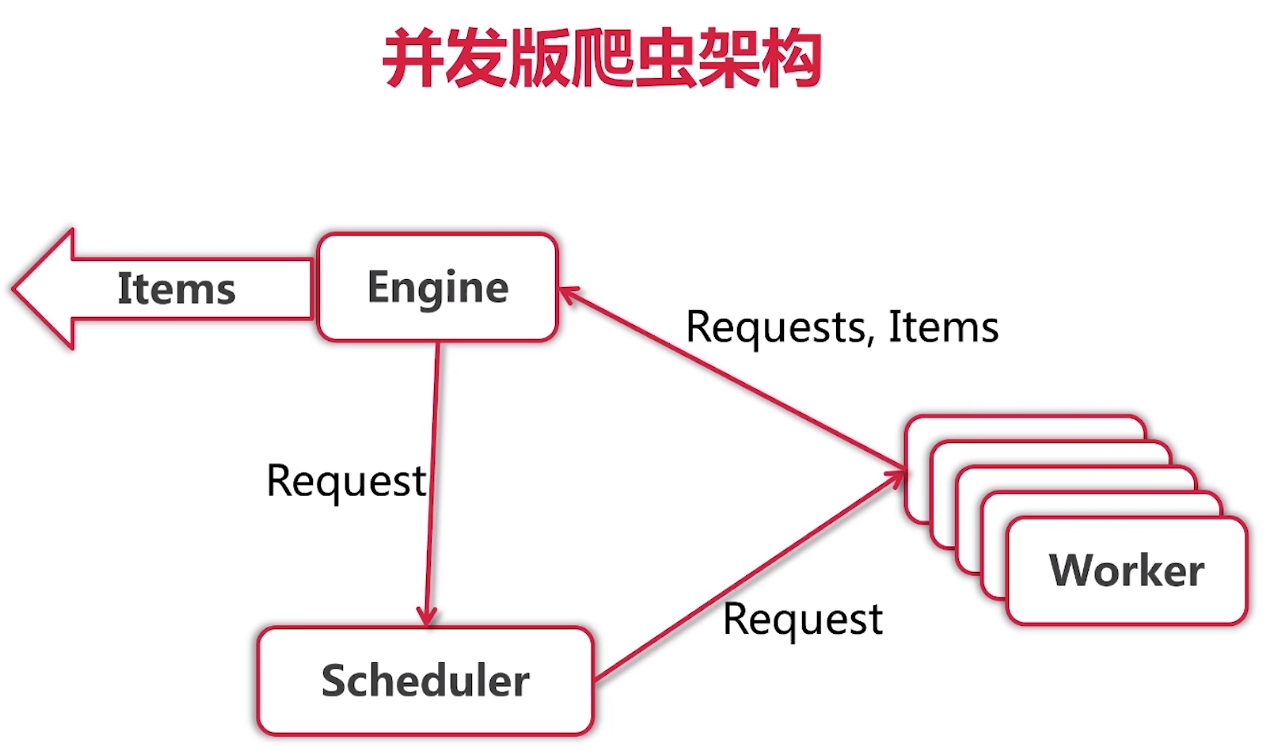


# 并发版爬虫架构

把parser和fetcher合并成worker

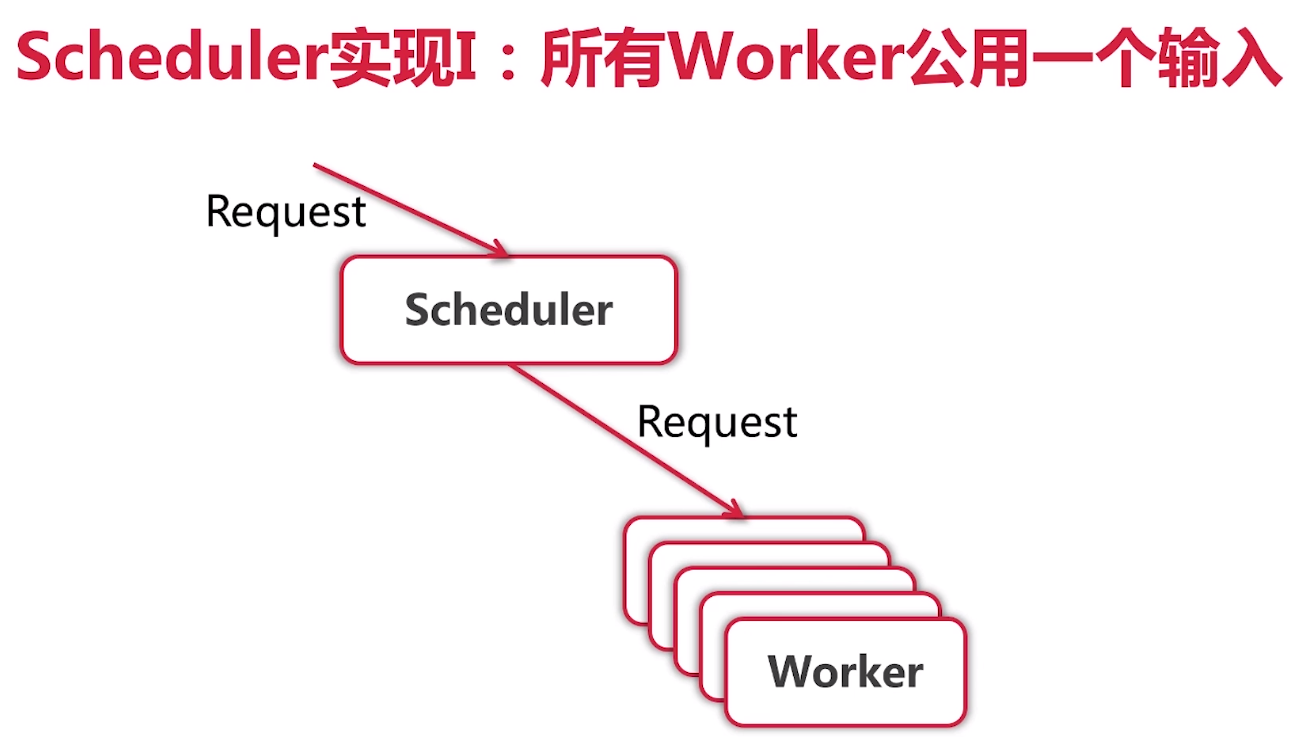
把worker并发

Worker并发的关键是实现了scheduler，通过request队列和worker chan队列把指定的request送给指定的worker



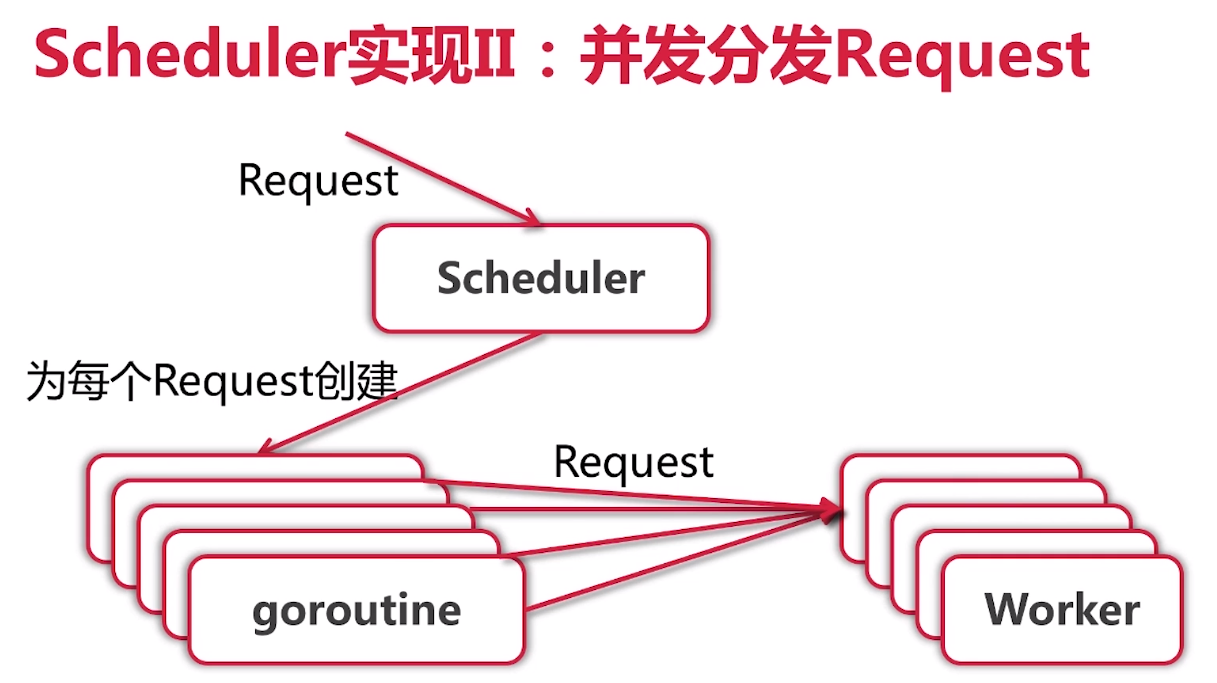
## Scheduler实现一

很简单，而且还不能工作，因为chan死锁了

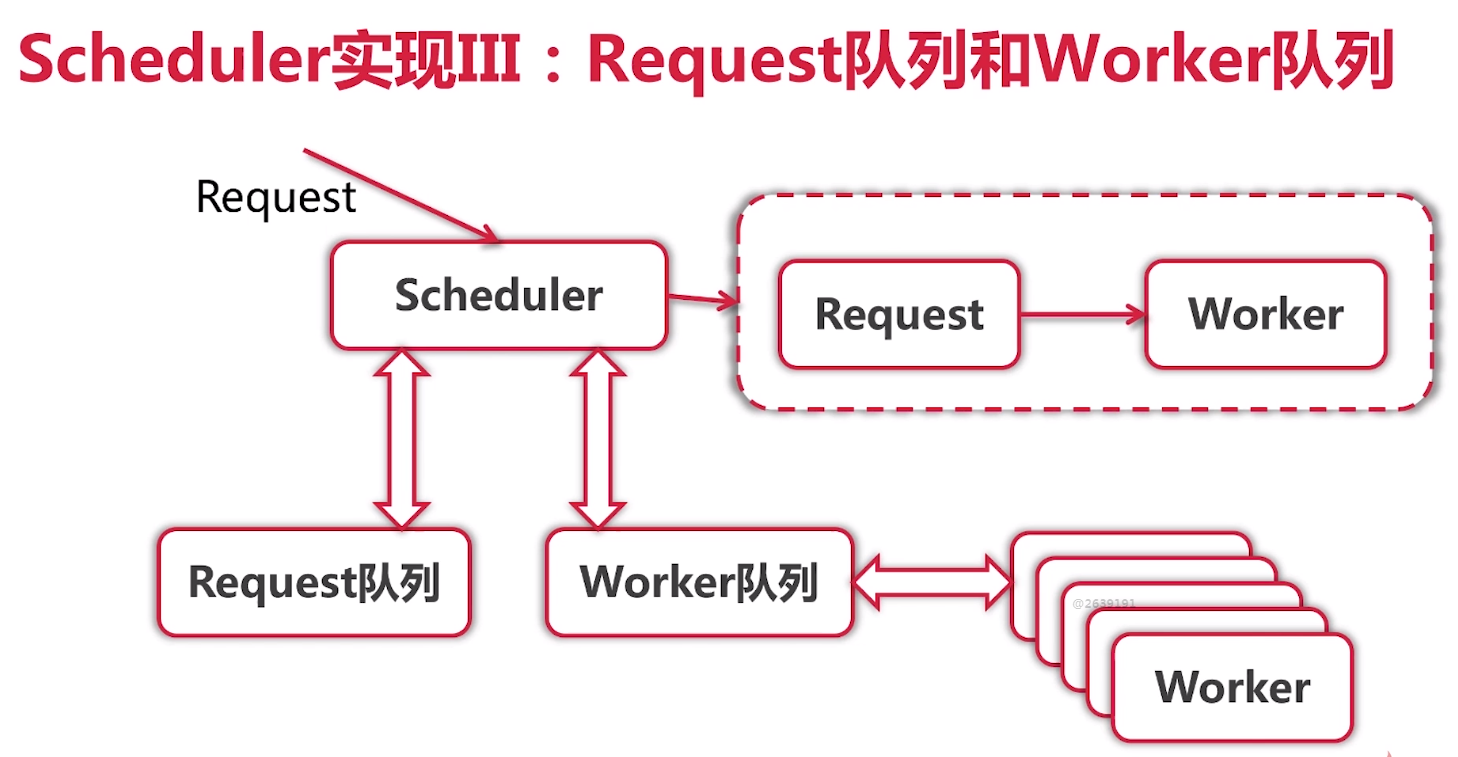


## Scheduler实现二

太浪费资源了



## Scheduler实现三



# 分布式爬虫架构

分布式系统最关键的地方是消息传递

本项目用的是RPC，复杂的地方是把Parser由匿名方法改成了接口并序列化

