**基于策略梯度的卫星频谱资源动态分配算法**

# 一、问题描述及建模

卫星通信系统通过星载多波束发射机在地面上形成个波束，并由表示。系统中可用的信道用集合表示，其中每个信道资源正交且互不重叠（对应于LTE体制中的RBG（物理资源块组）），且其大小为，其中表示系统总共可用的信道资源。该多波束卫星系统地面覆盖区域内的卫星终端分布在不同波束内，各个卫星终端用户采取相应的准则选择合适的波束进行接入。每个卫星终端与接入波束一一对应，且有唯一的标识ID（对应于LTE体制中用户唯一标识RNTI），因此可以用二元组集来唯一标识卫星终端用户，其中代表的是某用户的RNTI，代表该用户接入的波束。进一步的，波束的信道分配状态可以用矢量标识，记为，其中每一项表示波束下的用户终端对子信道的占用情况，1表示占用，0表示空闲，则卫星系统中所有波束的信道分配状态向量构成了卫星系统总的信道分配矩阵，表示为。在LTE通信体制中，每个TTI，用户（UE）都会向基站（eNB）报告信道质量CQI反映下行信道的质量，并且。用户终端可能发起的通信业务类型不同，在LTE体系中，用QCI（Qos等级标识）来区分不同用户的业务承载类型，标准QCI属性见下表，在每个TTI开始时，每个正在通信的用户终端待传送的数据量记为。

标准QCI属性

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| QCI | 资源类型 | 优先级 | 时延预算 | 丢包率 | 服务示例 |
| 1 | GBR  （保证比特速率） | 2 | 100 ms | 10-2 | 语音会话 |
| 2 | 4 | 150 ms | 10-3 | 视频会话（实时传送数据流） |
| 3 | 3 | 50 ms | 10-3 | 实时游戏 |
| 4 | 5 | 300 ms | 10-6 | 非实时视频会话（缓冲流） |
| 5 | Non-GBR  （不保证比特速率） | 1 | 100 ms | 10-6 | IMS信号 |
| 6 | 6 | 300 ms | 10-6 | 视频（缓冲流）  基于tcp的(例如，www，电子邮件，聊天，ftp，p2p文件共享，渐进视频等等) |
| 7 | 7 | 100 ms | 10-3 | 语音  视频（实时传送数据流） 互动游戏 |
| 8 | 8 | 300 ms | 10-6 | 视频（缓冲流）  基于tcp的(例如，www，电子邮件，聊天，ftp，p2p文件共享，渐进视频等等) |
| 9 | 9 |

在每个TTI周期内，LTE资源调度器将信道集合分配给接入各个波束的用户终端UE，我们要解决的问题是要给哪些用户分配资源，以及每个用户分配多少资源、哪些资源的问题，如果使用固定的频率服用规则，每个波束固定的平均使用总带宽的一部分频段，这样势必会造成用户业务接入较少的波束的频谱资源利用率低，用户业务接入较多的波束频谱资源不够用，造成系统总体性能的恶化，希望能将强化学习的思想引入到多波束卫星频谱资源分配的问题中，动态地给每一个接入的用户配置频率资源，尽量提高系统的总体性能。

# 二、强化学习模型概述

在强化学习的模型中，需要有智能体（Agent）和环境（Environment）这两个基本部分，智能体通过感触环境得到环境的观测值State，记为，根据State智能体做出一系列的行为Action，记为，Action作用于环境后会改变环境，与此同时环境会对Action做出反馈Reward，记为，智能体根据Reward不断地调整Action以使得到的Reward更好，这就是强化学习的过程，通过不断的试错，来提高智能体Agent的决策能力。

## 2.1 状态State

在我们的频谱资源分配问题中，将接入卫星的用户业务及其相关信息的集合建模为状态State，包括每个接入用户的标识（BeamId,RNTI），用户终端的CQI和QCI信息，以及当前TTI该用户待传输的数据量，另外定义一个该用户在此次调度中是否分到资源的标识，即资源编号集合，这样用户的状态可以表示为



第个TTI对应的环境状态表示为



其中，为该TTI下接入的总请求数。

假设想要满足用户的业务请求需要给该请求分配个资源块，需要把状态进行拆分操作，拆分操作如下所述。

根据该用户上报的信道质量表示，查表确定传输数据量所需的资源块数量，然后把重新表示为如下所示



## 2.2 动作Action

将卫星对接入用户业务请求的资源分配策略建模为强化学习中的动作Action，分配给用户资源的动作可表示为



其中，为分配给用户的个资源的编号集合，



第个TTI对应的action表示为



对应状态拆分后的形式为



其中。

## 2.3 奖励Reward

智能体产生调度策略后，卫星通信系统执行进行实际的通信仿真，可以从系统中提取一系列指标对调度策略做出整体评价Reward，记为，目前系统中主要以吞吐率作为性能评价标准，一个TTI内系统的吞吐率定义为



其中表示第在该TTI内，用户正确传输的数据量，表示第在该TTI内，用户想要传输的数据总量。

则可表示为



## 三、算法概述

在基于策略梯度的强化学习算法中，我们直接对策略进行参数化得到策略函数，通过控制参数的权重，直接选择比较好的行为，建立目标函数



通过优化参数，达到优化策略的目的。

基于策略梯度的动态资源分配算法在一个调度周期TTI内的流程如下图所示



一个调度周期的算法流程

基于策略梯度的动态资源分配算法流程如下表所示：

基于策略梯度的动态资源分配算法流程

|  |
| --- |
| 初始化，  For each TTI（episode）:  初始化，，，， |
| 获得请求序列，将中的请求序列依据待传输量、QCI等信息进行优先级排序 |
| For each step:  选取中优先级最高的未参与调度的业务请求 |
| 产生随机数，根据策略： |
| if ： |
| 根据策略依概率从可用资源集合中选取分配给该业务的资源编号 |
| else: |
| 随机从可用资源集合中选取分配给该业务的资源编号 |
| 将保存为动作并存入，更新、和 |
| if 是最后一个业务请求 or 已满： |
| 将送入环境中执行，计算奖励 |
| 将存入，将存入 |
| else: |
|  |
| 将存入 |
| 使用这个episode存储的序列训练策略网络 |
| 参数的更新方向是梯度下降最快的方向，优化策略函数 |
| 更新 |

# 三、在NS-3中搭建场景

利用NS-3搭建基于LTE的卫星通信场景

1、使用eNB模拟卫星的波束。

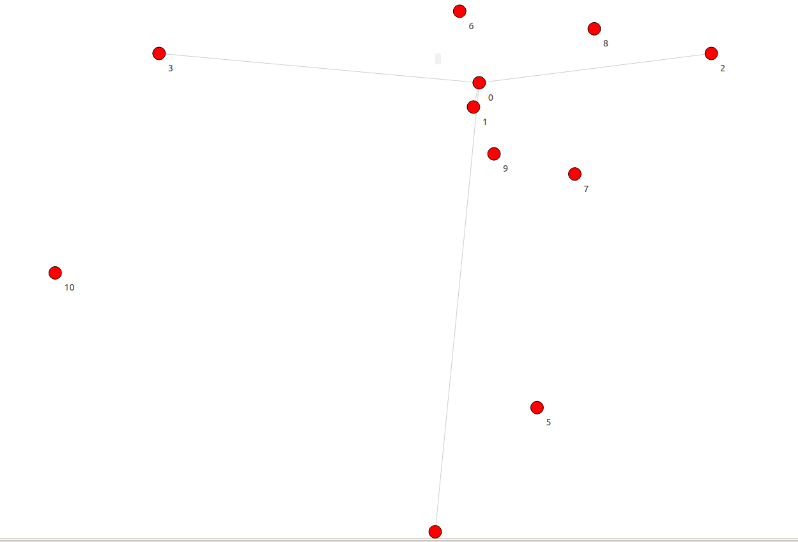
2、每个波束（eNB）下连接若干随机分布的UE作为用户终端（UE）。

3、通过设置UE的业务承载类型（体现为QCI），区分终端业务类型。

4、通过给UE安装Application设置终端的业务的开始时间、持续时间以及数据发送情况，仿真中使用OnOff模型。

5、通过设置UE节点的移动性mobility来模拟终端的移动，如RandomWalk2dMobilityModel。

通过NS-3可视化工具NetAnim可以实现通信节点可视化如下图所示

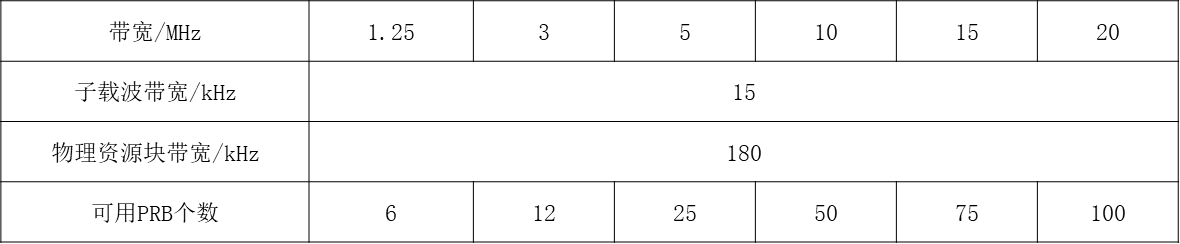


通信场景拓扑图

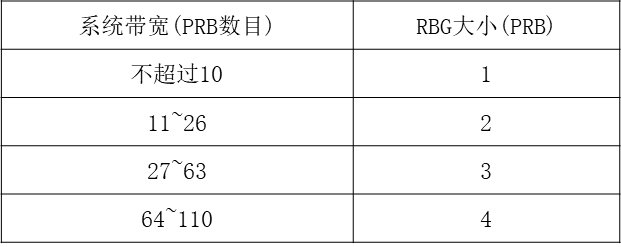
## 3.1 LTE中频谱资源划分情况

LTE系统资源调度的周期为一个TTI(通常为1ms)，资源调度的基本单位是RBG，1个RBG由若干个连续的PRB组成，PRB的数量和RBG的大小与系统带宽有关，如下两表所示。

系统带宽与可用PRB数目的关系



系统带宽与可用RBG大小的关系



## 3.2 改写NS-3中的频率资源调度器

NS-3中原本实现了LTE系统的频率资源调度器scheduler， scheduler有收集环境信息和分配资源的功能，根据我们的需要，将scheduler进行改写，保留收集环境信息的功能，资源分配模块改为从外部(来自Gym)传入资源调度策略(action)， scheduler现在只负责执行该调度策略。



改写后的LTE资源调度器示意图

## 3.3 业务到达情景构建

通信业务请求到达情况服从参数为的泊松分布，则在时间上先后到达的业务之间的时间差服从参数为的负指数分布，在NS-3中实现了产生服从负指数分布的随机数的生成函数，我们利用这个函数生成业务之间到达的时间间隔序列interval，将这个序列元素逐次累加后得到的序列则可得到业务的到达时间序列arrive。

决定通信系统业务负载的强度的另一个因素是业务持续时间，在仿真中暂时设定为业务持续时间服从处于范围[a,b]的均匀分布。

在NS3中构造通信场景，带宽为100个RB，RBG大小为4，RBG数量25，总用户数500，通信业务请求到达情况服从参数为的泊松分布，业务持续时间服从[1,2]的聂云分布，分别在为30,50,70,90,的情况下进行了仿真，并计算系统的平均吞吐量和平均吞吐率。

时刻的平均吞吐量计算为系统在一段时间内()吞吐量的平均值,即



时刻的平均吞吐率计算为系统在一段时间内()正确传输的总数据量比上该段时间内总传输数据量





吞吐量与时间关系图

由于用户总数确定有限，所有用户完成通信后仿真即停止，所以取值不同的情况下对应的仿真时长不同。

吞吐量与时间的关系图主要为了验证业务到达服从泊松分布，例如时，大约在10000ms左右吞吐量开始明显下降，也就是说这个时候只有用户离开，不再有业务到达，总用户数为500，则到达率500/10=50/s，即每秒到达用户数约为50，符合我们的要求。

# 四、算法仿真分析与比较

## 4.1 已有算法简介

最大载干比MAX C/I调度算法是一种充分利用无线信道增益来提高系统传输速率的调度算法，它的基本思想是，用户通过测量下行链路的瞬时信干噪比SINR，然后反馈给基站，基站通过对这些用户的信干噪比进行从大到小的排序，在每个调度周期内，那些信干噪比最大的用户将会优先获得调度机会，MAX C/I调度算法优先考虑系统的吞吐量性能，NS-3中的CqaFfMacScheduler调度器实现了该算法。

## 4.2 DAC算法仿真与比较

在第一节的环境中，使分别为30和70，在不施加频率复用规则限定的情况下，分别使用MAX C/I调度算法和基于强化学习的资源分配算法DAC，对比两者的平均吞吐量和平均吞吐率，如下图所示









观察上图可知，DAC算法相对最大载干比MAX C/I调度算法在吞吐率和吞吐量方面有更好的性能。