

state

input state : $s_t = ((\vec{p}_{1t}, \vec{q}_{1t}, u_{1t}, r_{1t}), (\vec{p}_{2t}, \vec{q}_{2t}, u_{2t}, r_{2t}))$

\vec{p}_{it} : 在 t 时刻, 子流 i 过去 k 个时间片内, 发送数据包的数目。

\vec{q}_{it} : 在 t 时刻, 子流 i 过去 k 个时间片内的 RTT 值。

u_{it} : 在 t 时刻, 子流 i 上未收到 ack 的包的数目。

r_{it} : 在 t 时刻, 子流 i 过去 k 个时间片内, 累计的重传次数。

action

output action : $a_t = (n_{1t}, n_{2t})$

n_{it} : 一个整数, 用于计算下一个时间片内, 子流 i 发送数据包的比例 s_{it} , 计算公式为:

$$s_{it} = \frac{n_{it}}{\sum_{j=1}^N n_{jt}}$$

当 $n_{it} = 0$ 时, 表示下个时间片, 不使用子流 i 发送数据包。

reward

调度目标, 提高吞吐量, 在较快的时间内完成传输, 降级接收方的缓冲区大小, 减少重传。reward 定义如下:

$$r = \sum_{j=1}^N (\lambda \sum_{j=1}^k \vec{p}_j - \mu \sum_{j=1}^k \vec{q}_j - \alpha u_j - \beta r_j)$$