基于多 Agent 强化学习的 Ad hoc 网络跨层拥塞控制策略

邵 飞[①] 伍 春[①] 汪李峰[②]
[①](西安电子科技大学通信工程学院 西安 710071)
[②](中国电子系统工程公司研究所 北京 100141)

摘 要：该文首先证明基于 MAC 层竞争造成的网络拥塞模型中存在纳什均衡点。其次，基于 WOLF-PHC 学习策略提出了一种跨层拥塞控制(WCS)机制。它在路由层中选择一对去耦合节点作为转发节点，同时在 MAC 层对源节点的发送数据进行分流，从而提高链路的空间重用性。仿真结果表明：在不需要交互任何信息的情况下，通过节点之间的相互博弈以后，采用 WOLF-PHC 算法能够找到每个节点的最佳分流概率进而使整体网络吞吐量达到最大值；同时当外界环境发生改变时，该算法能够较快地找到新的最佳分流概率从而实现对环境的自适应能力。

关键词：Ad hoc；拥塞控制；跨层设计；博弈论；WOLF-PHC
中图分类号：TN915 **文献标识码**： A **文章编号**：1009-5896(2010)06-1520-05
DOI: 10.3724/SP.J.1146.2009.01092

Research on Cross-layer Congestion Control Strategy Based on Multi-agent Reinforcement Learning in Ad hoc Network

Shao Fei[①] Wu Chun[①] Wang Li-feng[②]
[①](*The School of Telecommunications Engineering*, *Xidian University*, *Xi'an* 710071, *China*)
[②](*Institute of China Electronic System Engineering Corporation*, *Beijing* 100141, *China*)

Abstract: In the paper, the existence of an Nash equilibrium in the network congestion mode induced by MAC layer competition is proved firstly; Secondly, a cross-layer congestion-control mechanism named WCS is proposed based on WOLF-PHC learning strategy. WCS selects a couple of decoupled node as next-hop nodes at routing layer; Meanwhile, source's traffic is spitted and forwarded at MAC layer, which improves the space reusing efficiency of link. Simulation result shows that: without any exchanging information, optimum split-flow point of source node will be sought by WOLF-PHC in order to maximize the network throughput; Furthermore, WOLF-PHC will discover new optimum split-flow point in order to adapt to new network environment.

Key words: Ad hoc; Congestion control; Cross-layer design; Game theory; Win-Or-Lose-Fast Policy Hill Climbing(WOLF-PHC)

1 引言

本文主要研究在带宽受限的 Ad hoc 网络中，节点之间如何通过博弈学习而实现最佳的数据分流。通过对节点自私性和网络拥塞的关系进行建模，在随机博弈的框架下，将强化学习[1]与对策论相结合，使节点(也称为 agent)通过对各种对策形势的学习自动掌握如何在不削弱整体网络性能的同时增加自己的性能。WOLF-PHC(Win-Or-Lose-Fast Policy Hill Climbing)算法[2]将“赢否则就要学得更快”策略扩展到 PHC 的学习策略中，从而既满足单个个体的理性又满足整体的收敛性[2]。本文基于最佳响应多 agent 强化学习方法中的 WOLF-PHC 算法使节点通过非协同不完全信息下的博弈找到最佳分流点。

2 网络拥塞模型以及 Nash 均衡点的存在性

2.1 基于 MAC 层竞争造成的网络拥塞模型

如图 1 所示，节点 A，B 为两个源节点，分别要给目的节点 C，D 发送数据。节点 E，F 为 $A\rightarrow C$ 的两个可用中继节点；节点 G，H 为 $B\rightarrow D$ 的两个可用中继节点。节点 F，G 处于轻度干扰区域(图 1 中白色椭圆区域)，节点 E，H 处于重度干扰区域(图 1 中灰色椭圆区域)。假设，4 条路径 $A\rightarrow E\rightarrow C$，$A\rightarrow F\rightarrow C$，$B\rightarrow G\rightarrow D$，$B\rightarrow H\rightarrow D$ 的链路质量(以丢包率为指标)分别为：$A\%$，$B\%$，$C\%$，$D\%$，且 $A<B$，$D<C$。节点 F 和节点 G 分别与节点 E，H 是 MAC 层“去耦”的 1)[3]。传统的路由协议在选择路径时主要

2009-08-17 收到，2009-12-29 改回
国家 973 计划项目(2009CB320403)和国家自然科学基金(60832008，60832006)资助课题
通信作者：邵飞 shaofei715@163.com

1)这里的“去耦”是指采用位置信息，发送功率控制或定向天线等技术使两个无线节点不能够互相干扰。

根据路径的连通特性，文献[4]指出：在有干扰的环境下，采用路径连通特性反而会增加网络的业务开销，并给出了一种基于链路统计概率的路由协议 SAMPLE。本文中路由的选择主要是根据路由表中路径的链路质量(假设路由表中反映链路质量的链路统计概率已知)。因此，源节点 A 和 B 分别会选择路径 $A\rightarrow F\rightarrow C$ 和 $B\rightarrow G\rightarrow D$ 作为前向传输路径(图 1(a)中虚线所示)。从而使节点 F,G 很快变为“繁忙”节点，直至拥塞；而节点 E 和节点 H 却一直会成为“空闲”节点，直到链路质量发生重大变化。针对上述拥塞现象，本文对源节点 A，B 的数据流采用分流机制，如图 1(b)所示。t 时刻，源节点 i 的分流概率为 α_i^t，数据包的到达概率为 λ_i^t。假定每个源节点都只能有两条前向传输路径，即节点 i 存在竞争节点路径(图 1(b)中粗虚线所示)和非竞争节点路径(图 1(b)中细虚线所示)，且竞争路径的路径丢包率为 $\rho_i^{t'}$，非竞争路径的丢包率为 $\rho_i^{t''}$。节点 i 走竞争路径时数据成功发送的时间比例为 $s_i^{t'}$：$s_i^{t'}=P_{s_i}(\alpha_i^t\lambda_i^t)L_i/E_s$，$P_{s_i},L_i,E_s$ 的表达式见文献[3]；节点 i 走非竞争路径时发送成功的时间比例为 $s_i^{t''}$：$s_i^{t''}=(1-\alpha_i^t)\lambda_i^t\times\sigma$，其中 σ 表示数据平均发送速率。t 时刻，节点 i 的吞吐量为 $s_i^t=s_i^{t'}(\alpha_i^t\lambda_i^t)\times\rho_i^{t'}+s_i^{t''}((1-\alpha_i^t)\lambda_i^t)\times\rho_i^{t''}$，所以整个网络的吞吐量为

$$s^t=\sum_{i=1}^{N}s_i^t \tag{1}$$

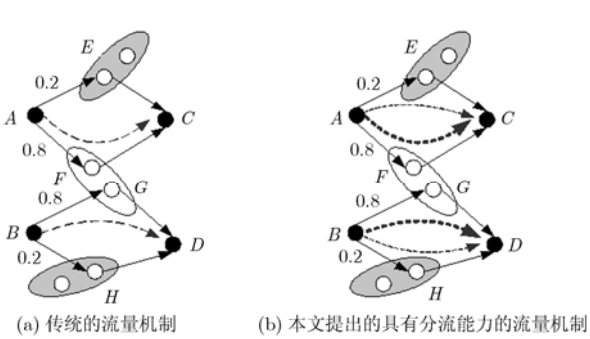
从而，整个网络的平均吞吐量：$s=\sum_{t=1}^{\infty}s^t$。各个节点分流的目标就是找到最佳的分流点，即

$$(\alpha_1,\alpha_2,\cdots,\alpha_n)=\arg\max_{\alpha_i\in[0,1]}\sum_{t=1}^{\infty}\sum_{i=1}^{N}[s_i^{t'}(\alpha_i^t\lambda_i^t)\times\rho_i^{t'}+s_i^{t''}((1-\alpha_i^t)\lambda_i^t)\times\rho_i^{t''}] \tag{2}$$

2.2 Nash 均衡及其存在性证明

下面，首先给出拥塞博弈中纳什均衡点(Nash Equilibrium Point，NEP)的定义：

定义 1　$s^t(\alpha_i,\alpha_{-i})$ 为节点 i 的吞吐量，那么 $(\alpha_1^*,\cdots,\alpha_{-i}^*,\cdots,\alpha_N^*)$ 是一个纳什均衡点，当且仅当 $\forall i\in N$；$0\le\alpha_i\le1$，都有 $s^t(\alpha_i^*,\alpha_{-i}^*)>s^t(\alpha_i,\alpha_{-i}^*)$。



(a) 传统的流量机制　　(b) 本文提出的具有分流能力的流量机制

图 1

由定义 1 可知，在由 N 个节点构成的博弈系统中，纳什均衡点是一个稳态，当系统达到纳什均衡点后，任何节点的分流概率如果偏离该均衡点(即节点 i 不选择分流概率 α 进行数据的分流)将不能够获得比在均衡点时更大的收益。

命题 1　在上述 N 个节点竞争的博弈环境中，采用分流概率的节点吞吐量如式(1)所示，那么该博弈的纳什均衡点存在且唯一。

证明　由定义 1 可知，Nash 均衡点在每个节点的吞吐量达到最大时达到，由式(1)可知，$s_i^t(\bullet)$ 是连续函数，对 α_i 二阶连续可微，且 $dS_i^2\big/d\alpha_i^2<0$，所以 $s_i^t(\bullet)$ 是单峰凸函数，存在极大值 α_i^*；由式(1)可知，对于 $\forall\alpha_i\in[0,1]$，$s_i^t(\bullet)>0$，$s_i^t(\bullet)$ 的最大值在极大值点 α_i^* 取得，即对于 $\forall\alpha_i\in[0,1]$，都有 $s_i^t(\alpha_i^*,\alpha_{-i}^*)>s_i^t(\alpha_i,\alpha_{-i}^*)$。另外，$s^t=\sum_{i=1}^{N}s_i^t$ 在其有界变量区域上无界，则根据文献[4]中有限维优化的标准结果，可知 s^t 在有界变量区域中存在唯一的最大值。所以 Nash 均衡点存在且唯一，每个效用函数的最大的分流概率序列 $(\alpha_1^*,\cdots,\alpha_{-i}^*,\cdots,\alpha_N^*)$ 构成了该纳什均衡点。

3　随机重复博弈

本文采取的分流机制实际上是对节点之间“自私性”的有效折衷，各个节点在选择分流概率的时候既要考虑“个人利益”，同时也要兼顾“整体利益”。同时受无线网络中带宽资源的限制，agent 之间不可能通过协商的方法来得到一种折衷策略。本文选择随机重复博弈框架[5]来对多 agent 在上述非协同不完全信息下折衷特性进行建模。

本文中分流机制对应的随机博弈元素为

***S* 状态空间**　节点 i 的状态包括两部分 $s_i=\{s1,s2\}$。其中 $s1$ 表示节点 i 当前的分流概率 $\alpha_i\in[0:1]$，$s2$ 表示节点 i 吞吐量的变化情况，$s2=\{\Delta_{s_i}>\delta,\Delta_{s_i}<-\delta,\left|\Delta_{s_i}\right|\le\delta\}$，其中 δ 表示吞吐量变化门限值，可以事先设定；

***A* 动作空间**　A^i 表示第 i 个 player 的动作，则 $A^i=\{\alpha_i+\Delta,\alpha_i-\Delta,\alpha_i+2\Delta,\alpha_i-2\Delta\}$，其中 Δ 表示分流概率增加的步进制；

***R* 回报函数**　r^i 表示第 i 个 player 的回报函数：r^i 直接用节点 i 当前的吞吐量 $[s_i^{t'}(\alpha_i^t\lambda_i^t)\times\rho_i^{t'}+s_i^{t''}((1-\alpha_i^t)\lambda_i^t)\times\rho_i^{t''}]$ 来计算。

4　基于 WOLF-PHC 的跨层分流拥塞控制机制 — WCS

本文提出一种基于 WOLF-PHC 的跨层分流拥

塞控制机制(Wolf-PHC based Cross-layer Split-flow congestion control，WCS)。WCS 主要机理是：由于节点的自私性，从而造成非干扰区域网络的拥塞；当选择一个非干扰区域内的“去耦”分流节点，节点之间采取合作的方式能够大大缓解非干扰区域的网络拥塞。在路由层中选择一对去耦合节点作为转发节点，同时在 MAC 层对源节点的发送数据进行分流，从而提高链路的空间重用率。WCS 算法步骤后面详细叙述，其中，拥塞的提前预测，非干扰区域“去耦”节点的选择以及节点之间最佳分流概率的实现是 WCS 的关键。前两个方面即拥塞提前预测和非干扰区域“去耦”节点的选择不是本文研究的重点，这里不再详细阐述，有兴趣的读者可以参考文献[5]。本文主要研究节点是如何通过博弈学习来找到最佳的分流概率点。

分流机制所对应随机博弈中的元素在第 3 节已经有比较详细的描述。下面主要描述如何利用 WOLF-PHC 实现策略的学习。和 Q 学习算法类似，在初始状态 s_i 源节点通过 ε 贪婪策略随机地选择动作 A^i，并根据回报函数 $r^i(A^i|S^i)$ 来更新 Q 值：$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha(r+\gamma\max_{a'}Q(s',a'))$。与 Q 学习采用单一的策略即最大化 $Q^{a'}$ 值不同，WOLF-PHC 算法定义了两种策略即当前策略 $\pi(s,a)$ 和平均策略 $\overline{\pi}(s,a')$。当前策略实际上是一种概率分布函数，初始值为 $\pi(s,a)\leftarrow 1/|A_i|$。这个概率分布函数当 agent 选择动作 a 时进行更新，更新方法是：对于 Q 学习来说最好的动作即 $a=\arg\max_a Q(s,a')$，则增加其所占的比重。实现方法是：对最大化 Q 值的动作增加一个相同的量 δ，即 $\pi(s,a')\leftarrow\pi(s,a')+\delta$ if $a=\arg\max_a Q(s,a')$；否则从所有的动作中减去相同的量 $\pi(s,a')\leftarrow\pi(s,a')-\delta(|A_i|-1)$ if $a\neq\arg\max_a Q(s,a')$。WOLF-PHC 会不断地更新平均策略 $\overline{\pi}(s,a')$，并将当前的策略 $\pi(s,a)$ 与平均策略进行比较。比较时利用当前的 Q 值来分别计算：(1)采用当前策略的平均奖励 $\sum_a\pi(s,a)Q(s,a)$；(2)采用平均策略的平均奖励 $\sum_a\overline{\pi}(s,a)Q(s,a)$。如果计算得到的当前策略平均奖励值大于平均策略的奖励值即 $\sum_a\pi(s,a)Q(s,a)>\sum_a\overline{\pi}(s,a)Q(s,a)$，认为 agent 是“赢”的，此时平均策略 $\overline{\pi}(s,a')$ 将采用赢时的学习速率 δ_w 来慢慢地更新策略；否则，认为当前 agent 是“输”的，此时平均策略 $\overline{\pi}(s,a')$ 将采用“输”时的学习速率 δ_l 来更快地自适应学习。

基于 WOLF-PHC 的跨层分流拥塞控制机制——WCS 算法步骤：

步骤1 拥塞提前预测：When $m<$Threshold_free_memory(eg. $m<0.25$)，其中，m 表示当前节点内存中空闲空间的比例；

步骤2 去耦节点选择：(I, J)=decoupling_node_searching(I, Z)，其中，Z 表示处在节点 I 通信范围之内的所有节点；

步骤3 算吞吐量：T =through_average_calculating(I, J)to each node pair (I, J)，其中，T 表示当前节点吞吐量的变化。

步骤4 WOLF-PHC 分流机制：

(1) let α，α_{in}，α_{of}，δ_{in} ,δ_{of}，C，$\delta_l>\delta_w$ be learning rates. Initialize,

$$Q(s,a)\leftarrow 0,\ \pi(s,a)\leftarrow\frac{1}{|A_i|},\ C(s)\leftarrow 0$$

(2)Repeat,

(a)From state s select action a with probability $\pi(s,a)$ with ε greedy strategy.

(b)Observing reward r and next state s',

$$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha(r+\gamma\max_{a'}Q(s',a'))$$

(c) update estimate of average policy, $\overline{\pi}$,

$$C(s)\leftarrow C(s)+1$$

$$\forall a'\in A_i,\ \overline{\pi}(s,a')\leftarrow\overline{\pi}(s,a')+\frac{1}{C(s)}(\pi(s,a')-\overline{\pi}(s,a'))$$

(d)update $\pi(s,a')$ and constraint it to a legal probability distribution,

$$\pi(s,a')\leftarrow\pi(s,a')+\begin{cases}\delta, & \text{if } a=\arg\max_a Q(s,a')\\ \dfrac{-\delta}{|A_i|-1}, & \text{otherwize}\end{cases}$$

$$\text{Where, }\delta=\begin{cases}\delta_w, \text{ if } \sum_a\pi(s,a)Q(s,a)>\sum_a\overline{\pi}(s,a)Q(s,a)\\ \delta_l, \text{ otherwize}\end{cases}$$

(e)decay learning rates and sigma rates slowly, $\alpha=\alpha_{in}\times\alpha_{of}/(\alpha_{of}+i)$，$\delta_l=\delta_{in}\times\delta_{of}/(\delta_{of}+i)$，$\delta_w=C\times\delta_l$

(3) When $\forall i\in N$，$|\Delta_{s_i}|\le\delta$ 时，算法收敛。

5 仿真结果及分析

仿真部分主要研究 WCS 机制对整个网络吞吐量提高的效果，侧重的是如何利用 WOLF-PHC 算法实现节点之间的博弈学习。对于拥塞的提前预测、路由层“去耦”节点的选择等问题仿真中假设已经完全实现。同时，直接利用路径的丢包率来表示网络中干扰区域和非干扰区域的链路质量。由文中第 2 节可知：当各个节点的发送概率已知时，整个网络的吞吐量就可以直接利用公式计算得到。基于上述考虑，整个仿真采用 MATLAB 来实现。仿真中 WOLF-PHC 算法的典型配置如下：α_{in}=0.001，

$\alpha_{of}=10^6$，$\delta_{in}=0.001$，$\delta_{of}=10^6$，$C=10$。仿真结果的分析，主要从 3 个方面：(1)从单个节点来看 WOLF-PHC 算法的收敛性能；(2)随着外界环境的变化，WOLF-PHC 的自适应能力；(3)随节点数的不同，整个网络吞吐量的性能变化。

5.1 从单个节点来看 WOLF-PHC 算法的收敛性能

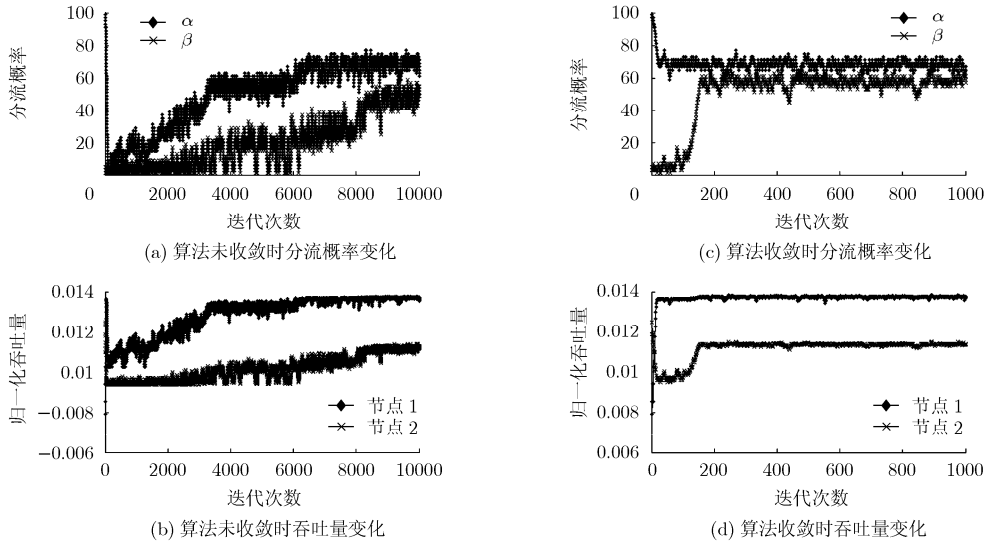
当网络的外界环境为(A%，B%，C%，D%)=(0.5, 0.7, 0.9, 0.5)时，假设网络中只有 5 个节点，且源节点的发送概率为 $\lambda_1=0.55$，$\lambda_2=0.55$ 时，从单个节点来看吞吐量以及分流概率的收敛性能。当网络的外部环境和 A 部分相同时且 $\varepsilon=0.4$，由式(1)，式(2)可知，最佳的分流均衡点为(75，60)。图 2 表示当源节点的初始分流概率为(0.99，0.01)时，分流概率和吞吐量随迭代次数的变化情况。图 2(a)，2(b)表示学习算法未收敛时节点 1 和节点 2 的分流概率和吞吐量分别随仿真迭代次数的变化情况；图 2(c)，2(d)表示学习算法收敛以后节点 1 和节点 2 的分流概率和吞吐量分别随迭代次数的变化情况。比较图 2(a)，2(c)可以发现：学习算法未收敛以前，节点 1 和节点 2 都在做大量的探索工作直到在第 8000 次迭代以后才逐渐找到最佳的分流均衡点；而当学习算法收敛以后，节点 1 和节点 2 会在第 120 次迭代时就找到了最佳的分流均衡点为(75，60)。综合图 2 说明：初始阶段，节点 1 仅仅把总数据流的 1%通过竞争节点发送，而节点 2 却把总数据流的 99%通过竞争节点发送；有趣的是，经过一段时间节点之间的强化学习以后，两个 agent 博弈的结果却是节点 1 要把通过竞争节点发送的数据流提高到 25%，节点 2 却要把通过竞争节点发送的数据流降低到 40%。可以得出这样的规律：当节点 2 牺牲自己的一些利益后(节点 2 均衡时的吞吐量与初始时吞吐量相比大大减小)，整个网络的吞吐量会得到提高。这也是纳什均衡作用的效果。从图 2 可以发现：博弈的结果是使 agent 尽可能的克服自己的自私性。

5.2 随着外界环境的变化，WOLF-PHC 的自适应能力

当外界环境变化时，agent 需要重新学习来实现对外界环境的自适应能力。为了说明本文提出的 WCS 算法的效果，设计了如下两个场景，如表 1 所示。仿真中设置如下，仿真共有 4 幕，在前两幕中环境参数为场景 1(仿真次数为 $1-4\times10^4$)，在后两幕中环境参数为场景 2(仿真次数为 4×10^4–8×10^4)。从图 3 可以发现：在前两幕中，因为环境参数未发生变化，所以学习算法收敛时的吞吐量是相同的；而在第 3 幕加载新的环境参数以后，归一化吞吐量却发生了很大的变化，如图 3(a)所示。节点 1 归一化的吞吐量从 0.014 迅速下降到 0.012。图 3(b)是把第 2 幕和第 3 幕中间变化的部分进行了时间轴上的放大以后，观察到的节点 1 的归一化吞吐量的变化情况，可以发现：当新的环境参数加载后，节点的吞吐量发生了较大的变化，同时经过很短暂的时间，节点 1 的吞吐量又回到了一个新的稳定状态。这说明本文提出的 WCS 算法能够实现对外界环境的适应能力。同时，也说明：有一定的学习经验以后，节点从加载新的环境参数到吞吐量回到新的稳定状态所经历的时间大大减小，即学习算法收敛的速率明显加快。

5.3 不同的节点数对吞吐量提高的影响

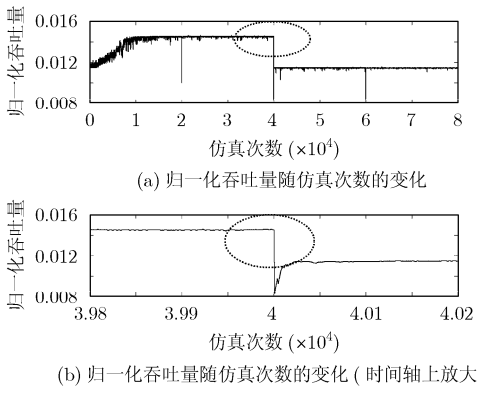
图 4 可以发现：在网络内存在 2 个发送节点时，



(a) 算法未收敛时分流概率变化　(c) 算法收敛时分流概率变化

(b) 算法未收敛时吞吐量变化　(d) 算法收敛时吞吐量变化

图 2 源节点的初始分流概率为 (α, β)= (0.99，0.01)，分流概率和吞吐量随迭代次数的变化情况



(a) 归一化吞吐量随仿真次数的变化

(b) 归一化吞吐量随仿真次数的变化 (时间轴上放大)

图 3 随着网络外界环境的变化，吞吐量的变化性能

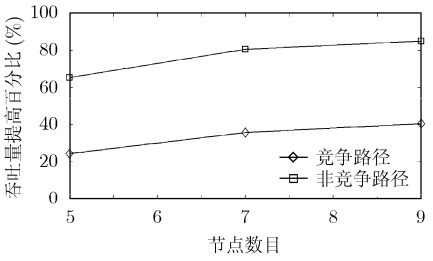


图 4 不同的节点数对吞吐量提高的影响

表 1 外界环境的变化

场景	A%	B%	C%	D%	α	β
1	0.5	0.7	0.9	0.5	0.74	0.65
2	0.5	0.5	0.9	0.5	0.84	0.6

采用分流机制的网络吞吐量比完全竞争情况下增加 24%，比完全没有竞争情况下增加 65%；在 3 个以及 4 个发送节点时，采用分流机制的网络吞吐量相对于完全竞争下则分别增加 35%和 40%。这也说明了当合作的节点数增加时，节点之间采取合作的方式能够提高整个网络的吞吐量。这与更加充分的利用了非干扰区域的网络容量是相吻合的。同时，随着节点数的增加，网络吞吐量增加的速度在逐渐变缓。这是因为整个网络容量是有限的。另外，需要指出的是因为本文提出的分流方案中不需要额外的带宽开销，因此对网络吞吐量的提高就是实际的有效数据传送效率的提高。

6 结论

本文采用随机博弈框架对节点自私性和网络拥塞的关系进行建模，并证明了存在纳什均衡点。提出了一种基于 WOLF-PHC 的拥塞跨层分流控制机制(WCS)，主要解决节点之间如何通过博弈学习来找到最佳的分流概率点。仿真结果说明：WCS 在不需要任何额外带宽开销的情况下，能够使各个节点通过博弈学习达到最佳分流点。

参 考 文 献

[1] 高阳. 强化学习研究进展. http://cs.nju.edu.cn/gaoy/documents/Agent/RL.doc, 2009-07-25.

[2] Bowling M and Veloso M. Rational and convergent learning in stochastic games[C]. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, Washington, 2001: 1021–1026.

[3] Malone D, Duffy K, and Leith D. Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions[J]. *IEEE/ACM Transactions on Networking*, 2007, 15(1): 159–172.

[4] 邵飞. 基于人工智能的认知无线网络关键技术研究. [博士论文], 西安电子科技大学通信工程学院, 2009.

[5] Thulasiraman P and Shen X. Decoupled optimization of interference aware routing and scheduling for throughput maximization in wireless relay mesh networks[C]. 2009 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad hoc Communications and Networks Workshops, Rome, 2009: 1–6.

邵 飞： 男，1980 年生，博士生，研究方向为认知无线电、认知无线网络、人工智能.

伍 春： 男，1978 年生，博士生，研究方向为认知无线电、认知无线网络、人工智能.

汪李峰： 男，1975 年生，高级工程师，研究方向为认知无线电、Ad hoc 网络.