doi:10.11959/j.issn.1000-436x.2015149

基于强化学习的 IEEE 802.15.4 网络区分服务策略

钱亮，钱志鸿，李天平，全薇
（吉林大学 通信工程学院，吉林 长春 130012）

摘 要：为了弥补 IEEE 802.15.4 协议原有区分服务机制的不足，提出了一种基于 BCS(backoff counter scheme)与强化学习的区分服务策略。从终端节点出发，在原优先级区分服务策略的基础上增加 BCS 退避策略以解决流量较大场合业务区分问题；针对协调器节点，提出了基于强化学习的占空比调整策略，该策略能根据不同应用需求和环境变化自适应调整占空比。仿真结果表明，提出算法能针对不同环境满足高优先级业务性能需求，并能根据流量变化进行占空比调整，具有极强环境适应性。

关键词：IEEE 802.15.4/LR-WPAN；区分服务；退避机制；强化学习；占空比
中图分类号：TN92 **文献标识码**：A

IEEE 802.15.4 differentiated service strategy based on reinforcement-learning

QIAN Liang, QIAN Zhi-hong, LI Tian-ping, QUAN Wei
(College of Communication Engineering, Jilin University, Changchun 130012, China)

Abstract: To provide better support in differentiated service for IEEE 802.15.4, a novel differentiated service mechanism was proposed based on BCS(back off counter scheme) and reinforcement learning. In terms of end-device, BCS backoff strategy was added to original priority-based differentiated strategy to solve the service differentiation problem under higher traffic condition. While for the coordinator, a reinforcement learning based duty-cycle adjustment algorithm was proposed to "self-learning" an optimal duty-cycle according to different application requirements and environmental changes. Simulation shows that the proposed algorithm can meet the performance requirements of high-priority service under different environments and adjust the duty-cycle when traffic is changed, which showed a strong environmental adaptability.

Key words: IEEE 802.15.4/LR-WPAN; differentiated-service; BCS; reinforcement-learning; duty-cycle

1 引言

低功耗芯片技术和无线通信技术的不断发展，为在典型的家居自动化和工业控制中使用低功耗、低成本的多功能无线传感器提供了可能，并使低速率无线个域网（LR-WPAN, low rate wireless person area networks）[1]逐渐得到了关注。低速率无线个域网是一种新的短距离无线通信技术，能提供低速率、低成本、低复杂度和低功耗的无线数据传输服务。LR-WPAN 网络在许多场合得到了应用，包括家居自动化、医疗监测和军事应用等[2,3]。为了满足工程应用的需要及保持不同设备间的兼容性，IEEE 标准委员会成立了 TG4 工作组，针对 LR-WPAN 网络制定了 IEEE 802.15.4 标准[4]，该标准定义出了相应的物理层（PHY）和媒体访问控制接入层（MAC）。在 WSN 网络中，事件触发性的数据与周期性采集的数据对传输延迟的要求具有显著的不同。在定位跟踪[5]中，当检测到相应目标出现时，必须及时将跟踪信息进行反馈，否则跟踪精度将受到极大影响；而对于周期性采集的数据，如环境指标的监测等，相应的传输延迟则影响较小。IEEE 802.15.4 作为 WSN 底层采用的可选标准

收稿日期：2014-07-11；修回日期：2014-10-18
基金项目：国家自然科学基金资助项目(61071073, 61371092)
Foundation Item: The National Natural Science Foundation of China(61071073, 61371092)

之一，虽然其在信标工作模式下提供了用于实时通信场合的 GTS（guaranteed time slot）机制，但是仍然有一些因素限制了 GTS 机制的应用。首先，设备如果需要 GTS 机制，必须先发送 GTS 请求；其次，对于一些时间敏感的事件（如 GTS 请求、报警、PAN 管理命令等），在传输的过程中并未进行区分；最后，由于在每个超帧中最多容纳的 GTS 数量为 7，大大限制了其在大规模网络中的使用。基于此，为 IEEE 802.15.4 协议提供相应的区分服务及 QoS 支持成为必须解决的问题，并逐渐成为研究热点。

为了给 IEEE 802.15.4 标准提供更强的区分服务机制支持，现有的研究者采用的方法主要集中在 3 个方面：①改进 GTS 分配机制，如多个设备共享相同的 GTS，从而提高 GTS 的利用率；②从 CSMA/CA 算法出发，为不同的业务设置不同的优先级，并根据优先级的大小赋予不同的 CSMA/CA 算法执行参数，从而使具有较高优先级的设备实现优先传输；③在基于 IEEE 802.15.4 协议提供服务的基础上添加上层区分服务支持。

HO C 等提出了一种 MFDGAS 算法（multi-factor dynamic GTS allocation scheme）[6]来提高 GTS 的利用率和网络性能，算法综合考虑了数据分组大小、延迟及 GTS 的利用率来决定 GTS 请求的优先级，根据优先级的高低，分配相应的 GTS 时隙。该算法在 GTS 的利用率上大大增强，但请求的优先级必须建立在广泛收集信息的基础上，信息收集的准确与否将极大影响算法的有效性。KOUBÂA A 在文献[7]提出了一种区分服务的机制，macMinBE 和 macMaxBE 为其涉及到的主要参数，在该机制中，作者认为命令帧的优先级较高，数据帧的优先级较低。这一机制实现了为高优先级业务提供相应的服务，但在网络内流量、节点数较大的场合，由于 CSMA/CA 算法碰撞次数的增加，即使是高优先级业务也需要等待较长时间才能进行数据的传输，从而降低了原有算法的有效性。相比于直接针对 IEEE 802.15.4 协议进行修改，KIPINS D 在文献[8]中基于 IEEE 802.15.4 协议之上实现了 AEL 层（ANGEL IEEE 802.15.4 enhancement layer），该层基于 IEEE 802.15.4 的底层服务，能够为上层具有不同优先级的服务提供区分服务支持。

从 GTS 角度出发，虽然上述提出的几种算法都能在一定程度上满足需求，但是在进行动态决策之前，都需要首先收集相应的信息，从而加大了开销；同时在进行算法设计中，上述的算法复杂度一般较高且没有考虑能耗的要求，如何实现在降低开销的同时又不损失算法的有效性是需要继续研究的问题。现有基于 IEEE 802.15.4 的区分服务策略，在满足区分服务基础上，希望尽量降低网络能耗需求，而均未对占空比进行合理的设置。由此分析，在满足不同业务性能的基础上，提出一种能够最大程度降低算法复杂度、减少网络能耗，并动态适应不同环境的区分服务机制，已经成为亟待解决的问题。本文分别从终端设备和协调器出发，提出了一种基于 BCS 退避策略与强化学习的区分服务机制：①针对终端设备，在基于优先级区分服务策略的基础上增加了 BCS 退避策略，能较好地解决网络内流量较大场合下的业务区分问题，同时，满足不同应用下高优先级业务的指标要求，并最大程度地节省网络能量；②针对协调器设备，首先仿真分析了不同占空比对于网络内节点延时、分组投递率及网络剩余能量的影响，并基于此分析结果提出了一种基于强化学习的占空比调整策略，该策略能根据不同应用环境下高优先级业务的性能需求进行“自学习”，得到一个较优的占空比，并能实时跟踪环境变化对占空比进行调整。

2 终端 BCS 退避策略

针对 IEEE 802.15.4 网络，设备在接入信道过程中，不同的参数设置对于 CSMA/CA 算法会有不同的影响[9]，而通过合理的参数配置，可以使高优先级业务具有更好的延迟指标。但是在数据量相对较大的场合，信道接入竞争冲突加大，此时设备的退避次数会大大增加，从而影响区分服务机制的有效性。为了适应数据量较大的场合，必须在原有为不同业务分配不同参数的基础上提出一种更加有效的策略来满足区分服务的要求。

2.1 时隙 CSMA/CA 算法的不同业务参数配置

时隙 CSMA/CA 算法主要应用在基于信标使能的网络中，在超帧结构的竞争接入时隙（CAP, contention access period）内，每个设备利用 CSMA/CA 接入信道进行数据的传输。执行该算法对应的一个基本时间单元称为退避周期（BP, backoff period），其值为 aUnitBackoffPeriod= 80 bit (0.32 ms)，时隙 CSMA/CA 的每个操作（CCA，退

避计数器等）都必须在 BP 的边界进行，而 BP 的边界必须与超帧中的基本时隙进行对齐。时隙 CSMA/CA 算法的执行主要基于 3 个关键的变量[4]。

1) 退避指数（BE, backoff exponent）：用于计算在执行 CCA 之前的退避延时，大小分布在 0 和 $2^{BE}-1$ 之间。

2) 竞争窗口（CW, contention window）：代表在接入信道之前，检测信道必须为空闲状态提供的退避周期（BP）数量，标准中定义的默认值为 $CW=2$（对应于执行 2 次 CCA 检测），在每个退避周期内，信道感知在 BP 的前 8 个 symbol 时间内完成。

3) 退避次数（NB, number of backoff）：代表算法在接入信道之前必须退避的次数，标准中默认初始值 $NB=0$。

为了对不同的业务进行区分，可以为不同的业务分配不同的 *CW* 和 *BE* 值。假设 2 种不同业务的优先级分别为 P_1、P_2，且 $P_1<P_2$，则为了满足区分服务的要求，应 $CW[P_1]>CW[P_2]$且 $BE[P_1]>BE[P_2]$。同时，借鉴文献[7]中优先级队列的思想，为不同的业务分别分配了不同的等待队列，并且设备在发送数据分组时优先从高优先级队列中选择数据分组进行传输。

2.2　终端 BCS 退避策略实现及分析

在数据量相对较大的场合，随着碰撞次数的增加，设备的退避次数也会相应增加，此时虽然不同的业务具有不同的退避参数，但是由于退避次数的增加，高优先级业务也需要等待较长时间才能进行数据传输，大大降低了基于优先级策略的有效性。因此，如果在发生碰撞后能够拉大高优先级业务和低优先级业务的退避距离，防止低优先级业务与高优先级业务再次发生碰撞，将会大大加强高优先级业务传输数据分组成功的概率。BCS 策略的思想即来源于此。

为了传输某一个业务的数据分组，在退避开始阶段 0，设备在$[0,W_0-1]$、$W_0=2^{BE}$ 内，随机选择一个值作为退避计数器。当检测到信道忙碌时，退避阶段和 *BE* 值均加 1 且设备将会在$[W_0,W_1-1]$、$W_1=2\ W_0$ 内选择一个退避计数器，而非按照原协议中从$[0, W_1-1]$中选择，具体算法实现流程如图 1 所示。

具体证明过程如下：假设 2 个变量 *x*, *y* 均匀分布于$[0,X-1]$和$[0,Y-1]$区间内，则当 $X<Y$ 时，变量 *x*、*y* 的期望关系为 $E_x=\frac{X}{2}<E_y=\frac{Y}{2}$，因此具有更小 *BE* 值的高优先级业务，其平均选择的退避计数器更小，将具有更高的传输概率。而如果在第一个退避阶段信道处于忙碌状态，根据 BCS 退避策略，此时 *BE* 值将会加 1，并且退避时间选择区间将会分别变为$[X,2X-1]$和$[Y,2Y-1]$，选择退避时间的期望关系分别为 $E_x=X+\frac{X}{2}<E_y=Y+\frac{Y}{2}$，此时高优先级选择退避时间的期望与低优先级选择退避时间的期望差值明显大于之前的第一个阶段，由于在发生碰撞后拉大了高优先级业务和低优先级业务的退避距离，因而能够优先保证高优先级业务的数据传输。

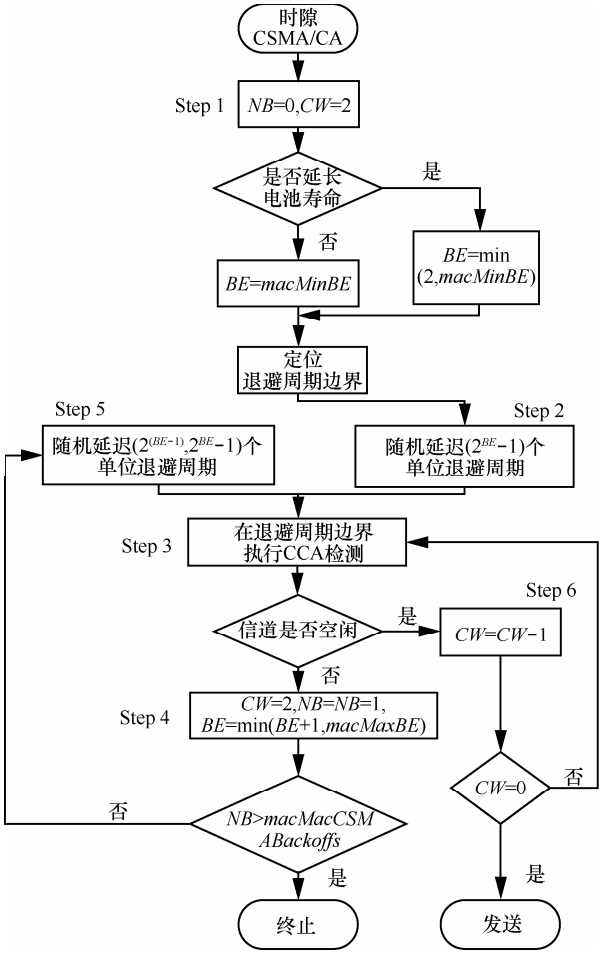


图 1　BCS 改进退避策略

3　协调器端基于强化学习的占空比调整算法

前人基于优先级的区分服务策略，虽然在某些特定环境下能够满足所提要求，但是所有解决方案

均只从终端节点角度出发，而未考虑协调器端占空比的影响。在信标网络中，协调器会周期性地广播信标帧，信标帧中包含了超帧的活跃周期与非活跃周期，终端节点只会在活跃周期内进行数据的交互，而在非活跃周期内会处于休眠状态，因此，一个超帧内活跃周期与非活跃周期的分配情况会极大地影响终端节点的性能指标，如延时、分组到达率及能耗等。因而，简单从终端角度利用基于优先级的信道接入算法来对不同的业务进行区分是不够的，必须在此基础上，再从协调器角度出发，根据应用对区分服务的需求，动态的调整占空比，从而不仅满足高优先级性能指标，同时尽量降低网络能耗。强化学习[10](reinforcement learning)是一种重要的机器学习方法，其实现简单，不需要对环境的精确建模，且具有自学习能力，能实时跟踪环境变化。本节将首先分析超帧占空比对节点性能的影响，并在 BCS 算法基础上提出了一种基于强化学习的占空比调整区分服务策略。

3.1 IEEE 802.15.4 超帧占空比影响分析

在信标使能网络中，协调器周期性的发送信标，节点在超帧的 CAP（contention access period）内利用 CSMA/CA 算法接入相应的信道进行数据传输，而在 CFP（contention-free period）内节点处于睡眠状态。一般采用 DC（duty-cycle）表示 $\frac{SD}{BI}$ 的比值大小，即 $DC=\frac{SD}{BI}=2^{SO-BO}$（$SD$ 为超帧宽度，BI 为信标间隔，SO 为超帧指数，BO 为信标指数）。为了验证不同占空比下，超帧结构的不同对于终端节点性能的影响，利用 NS2 仿真软件，在不同占空比下对于节点传输延时、分组投递率及节点能耗的影响进行了仿真，仿真结果如图 2~图 4 所示。

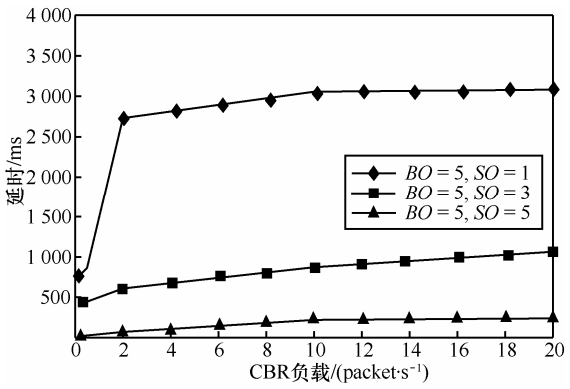


图 2 不同占空比对节点传输延时的影响

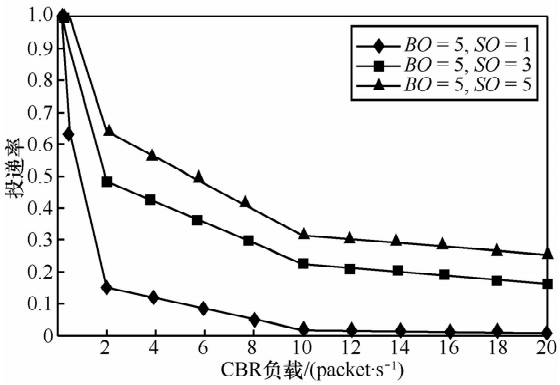


图 3 不同占空比对节点分组投递率的影响

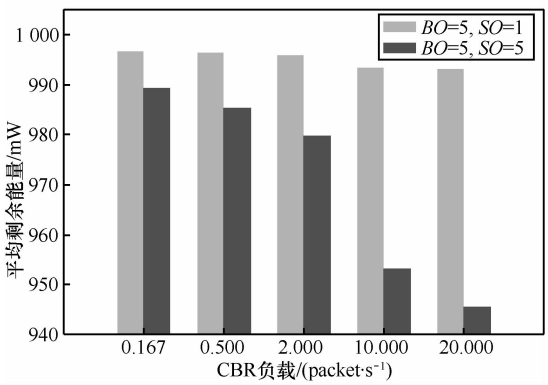


图 4 不同占空比对网络剩余能量的影响

根据仿真结果，可以得出如下结论：无论网络内流量情况如何，随着占空比 DC 的增加，节点发送数据分组产生的延时越小、分组投递率越高，但是网络整体剩余能量越小，大流量情况下尤其明显。可见，节点的延时、分组投递率指标与网络整体剩余能量指标呈现负相关关系，而基于前文所述的 BCS 算法，对于高优先级业务的延时、分组投递率及网络整体能耗，尤其会随占空比的设置不同而变化。因此，必须找到一种能动态调整占空比的方法，使之在网络流量状态变化时，对占空比能够进行实时调整，从而满足高优先级业务的延时、分组投递率需求，并且最大程度地减少网络能量消耗。

3.2 强化学习方法

人类通常从与外界环境的交互中学习。所谓强化学习是指从环境状态到行为映射的学习方式，以使系统行为从环境中获得的累积奖励值最大，是一种以目标为导向的学习方式[10]。将强化学习中的学习者和决策者称之为代理（agent），将与之交互的外界称之为环境（environment）。通过不断的与环境进行交互，Agent 选择不同的动作执行，而环境

则对该动作进行反馈并呈现新的状态。环境会根据动作给予 Agent 相应奖励（reward），Agent 的目标即通过采取相应的动作使所产生的奖励之和达到最大。强化学习模型[11]如图 5 所示。

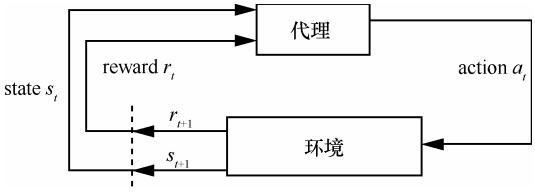


图 5　强化学习基本模型

在每个时间点上，Agent 都将会完成该状态下，对可选动作概率的映射。这种映射关系称之为 Agent 的策略（policy），用 π_t 表示，π_t（s, a）表示在状态 s_t=s 下，选择动作 a_t=a 的概率。

N-Armed Bandit 问题[12] 只考虑默认一种状态下评估不同动作的选择，是强化学习领域最简单的一类问题。在解决该问题的 Action-Value 方法中，每一个动作都对应有一个奖励的期望值，可表示为 $E(r_t \mid a_t)=Q^*(a_t)$，但实际情况往往不知道 $Q^*(a_t)$ 的实际值，只能通过估计得到相应的估计值 $Q(a_t)$。为了估计在 t 时刻，不同动作的 $Q(a_t)$ 值，最简单的方法就是求得选择动作 a 时所得到奖励的期望值大小。

$$Q_t(a)=\frac{r_1+r_2+\cdots r_{k_a}}{k_a} \tag{1}$$

当 $k_a \to \infty$ 时，有 $Q(a_t)$的极限为 $Q^*(a_t)$。而在实际情况下，为了节省内存，一般采用增量更新的方式[10]

$$\begin{aligned} Q_{k+1} &= \frac{1}{k+1}\sum_{i=1}^{k+1} r_i = \frac{1}{k+1}\left(r_{k+1}+\sum_{i=1}^{n} r_i\right) \\ &= \frac{1}{k+1}\left(r_{k+1}+kQ_k+Q_k-Q_k\right) \\ &= \frac{1}{k+1}\left(r_{k+1}+(k+1)Q_k-Q_k\right) \\ &= Q_k+\frac{1}{k+1}\left[r_{k+1}-Q_k\right] \end{aligned} \tag{2}$$

一般形式为 $Q_{k+1}(a)=Q_k(a)+\alpha(r-Q_k(a))$，在平稳状态下，$\alpha=\frac{1}{k+1}$。

如果已经得到了每个不同动作下的 $Q(a_t)$值（一般用查表实现），那么在任何一个时刻均会做出动作的情况下，$Q(a_t)$值所对应的动作是最优的，称这个动作是一个贪婪（greedy）动作。如果选择动作时，选择的是一个贪婪动作，那么也即说明 Agent 正在“采用（exploiting）”对动作的经验来进行动作的选择；而相反如果选择一个非贪婪动作，那么说明 Agent 正在进行“探索（exploring）”操作。从短期来看，采用贪婪动作可能使回报最大化，但是从长远角度来看，通过“探索”也有可能得到最大化的回报。为了解决这个矛盾，必须使用相应的策略进行平衡。其中 ε-greedy 策略[13]非常简单：

$$a_t=\begin{cases} a_t^*, \text{概率为} 1-\varepsilon \\ \text{随机动作, 概率为 } \varepsilon \end{cases} \tag{3}$$

即每个阶段在进行动作选择时，以概率 $1-\varepsilon$ 选择贪婪动作，而以概率 ε 选择其他动作。

3.3　基于强化学习的占空比调整区分服务策略实现

在实际情况下，很难建立起上述指标与占空比的精确数学关系，且难以实时适应环境变化。而利用强化学习“自学习”的特点，在特定环境下完全可以找到一个相对较优的占空比，不仅能满足不同应用下高优先级业务的指标需求，同时能动态适应环境变化。

强化学习是分阶段进行学习的，在每个阶段进行动作的选择及动作 Q 值的更新。协调器进行学习的具体示意如图 6 所示。

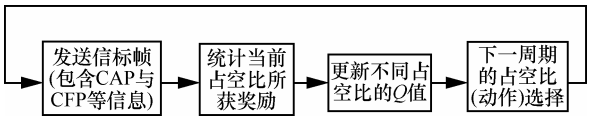


图 6　协调器强化学习示意

本系统中协调器周期性的发送信标帧，在每个信标帧中包含了超帧结构的相关信息（CAP、CFP 长度等），如 3.1 中所述，超帧中 CAP、CFP 信息与当前周期的 SD、BI 值相关，且占空比 $DC=\frac{SD}{BI}$，即确定的超帧结构决定了当前周期的占空比信息，根据上节所述动作（此处指占空比）与奖励的对应关系，统计当前占空比所获奖励，再采用增量更新的方式，根据当前周期占空比情况，更新 Q 值，在下一周期，同样按照累计奖励之和期望最大化的原则，进行占空比的选择。在此，由于高优先级业务超帧中，占空比所对应的 Q 值更大，因而选择高优先级业务的占空比作为强化学习的动作。如此反复，协调器作为一个 Agent 不断进行学习，随着时间的增长，协调器对于每个占空比的设置有了更多“经验”，从而能找到一个相对较优的占空比设置。

假设外界环境默认只有一个状态，将该学习问

题转换为一个 N-Armed Bandit 问题进行求解。下面将根据 IEEE 802.15.4 网络的特点针对强化学习中的各个元素进行建模。

1) 动作集合

IEEE 802.15.4 网络中的超帧结构如图 7 所示。

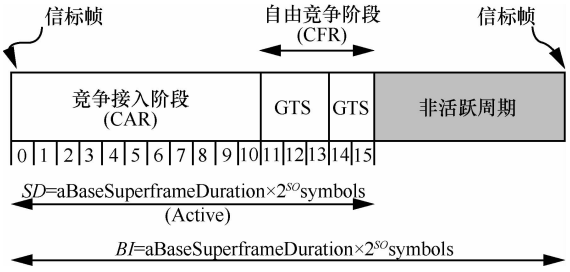


图 7 超帧结构

其中，占空比表示为 $DC=\frac{SD}{BI}=2^{SO-BO}$。占空比的改变有 3 种方式：①固定 BO，改变 SO；②固定 SO，改变 BO；③BO、SO 同时改变。BO 频繁改变容易使终端节点与协调器节点失去同步，故将 BO 设置成固定值，只通过改变 SO 的大小来调节超帧的占空比大小。那么，动作集合即为 SO 的取值集合

$$a = SO, SO_{\min} \leqslant SO \leqslant SO_{\max} \tag{4}$$

其中，$SO_{\min}=1$，而 $SO_{\max}=BO$。

2) 奖励

奖励为采取某个动作后的直接收益，在本系统中反映的是采取某一个 SO 值后，高优先级的性能指标以及网络能量消耗情况。具体有 3 个衡量因子：高优先级数据分组在当前超帧周期内经历的平均延时，终端节点高优先级队列在当前周期内的平均占用率及当前超帧周期的空闲侦听率。终端节点高优先级队列占用情况反应的是高优先级数据分组的分组丢失情况，占用率越高，说明分组丢失趋势会越明显；而超帧周期的空闲侦听率反映的是网络能量的消耗情况，超帧周期的空闲侦听率越高，说明更多的设备处于激活状态，却没有进行实际数据分组的收发，对网络能量浪费严重。在本文中决定奖励的首要因素在于高优先级业务的平均延时情况，其次在于高优先级数据分组的队列占用率（直接反映在分组丢失率上），最后才考虑超帧空闲侦听率的大小，即能量消耗的影响。奖励的具体表示如下

$$r=\begin{cases}-2\ , D>\delta \\ -1\ , D\leqslant\delta\ , O\geqslant o \\ -IL,\ O<o\ , D<\delta\end{cases} \tag{5}$$

D 表示本周期内接收到的高优先级数据分组的平均延时，δ 表示高优先级业务数据分组的容忍延时限。O 为当前周期内高优先级队列的占用率，o 为门限值。根据不同的应用需求，该容限可以设置成不同的值。在每个超帧周期内协调器利用增量更新的方式计算出高优先级数据分组的延时情况

$$D^{i}_{\text{avg}}=(1-\eta)D^{i}+\eta D^{i-1}_{\text{avg}} \tag{6}$$

为了取得这个队列的占用率，终端节点可以利用 MAC 层帧结构中的 7~9 Reserved 字段（如图 8 所示）。

利用该字段传输当前高优先级队列的占用情况，然后由协调器进行统计平均。由于 Reserved 只有 3 bit，可以进行分级表示，如表 1 所示。在超帧的末尾，协调器利用式(7)计算出在本超帧周期内的高优先级队列的占用率大小。

表 1 队列占用率分级表示

二进制表示	队列占用率/%
000	0~12.5
001	12.5~25
010	25~37.5
011	37.5~50
100	50~62.5
101	62.5~75
110	75~87.5
111	87.5~100

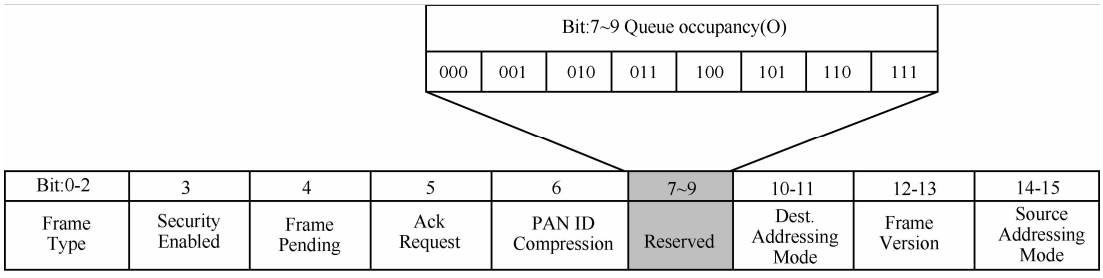


图 8 利用 Reserved 字段获取队列利用率

$$O=\begin{cases} 1, & \max(O_i)\equiv 111 \\ \dfrac{\sum_{i=1}^{n} O_i}{n}, & \text{其他} \end{cases} \tag{7}$$

IL 为一个超帧内空闲时隙所占的比率，表示如下

$$IL=1-SF_u \tag{8}$$

其中，“1”表示理论上超帧的最大利用率，即超帧的所有时隙均用来进行实际的数据传输，而没有设备处于空闲活跃状态。实际的超帧利用率为

$$SF_u=\frac{T_r}{SD-T_b}=\frac{T_s r}{SD-T_b} \quad \leqslant 1 \tag{9}$$

T_b 表示信标帧所占时隙的大小，r 表示在超帧内接收到的数据分组的个数，T_s 表示发送接收每个数据分组所占的时隙数

$$T_s=T_{CCA}+T_{DATA}+T_{IFS}+T_{ACK} \tag{10}$$

其中，T_{CCA} 表示执行 CCA 检测所用的时间，T_{DATA} 表示发送接收数据分组所占用的时间，T_{IFS} 和 T_{ACK} 分别表示帧间隔和发送应答帧所占用的时间。通过在当前超帧周期内统计接收数据分组的个数从而计算出 SF_u，进而计算出超帧空闲侦听率的大小。

3）*Q* 值更新

衡量不同动作长期收益的 *Q* 值，用 $Q(a)$ 表示，每个动作都对应一个 $Q(a)$ 值。本系统中协调器维护有一个查找表，由于 *SO* 值连续且均为特定整数值，因此可以利用数组实现，查找复杂度为 $O(1)$。其中每个 *SO* 值都对应有一个 *Q* 值，协调器在选择一个动作 a（即某个 *SO* 值）后，在周期快结束时根据统计结果及式(5)计算出相应 r，利用增量式(11)对该动作的 $Q(a)$值进行更新

$$Q_t(a)=Q_{t-1}(a)+\alpha[r_t-Q_{t-1}(a)] \tag{11}$$

其中，α 为学习速率，根据学习的稳定程度，算法会有针对性地对 α 进行更新，以及时反映环境的变化。

4) 学习方法

由于 IEEE 802.15.4 网络中流量一般比较稳定，因此采用最简单的 ε-greedy 学习方法，主要包含 2 种策略，分别表示为

$$\pi=\begin{cases} \pi_1(a)=\text{random}(a), & \varepsilon \\ \pi_2(a)=\arg\max Q(a), & 1-\varepsilon \end{cases} \tag{12}$$

其中，$SO_{\min}<\alpha<BO$，ε 表示探索速率因子，ε 越大表示在学习时，探索的概率越大。π_1 表示随机策略，协调器在每个超帧结束后以概率 ε 选择该策略并从所有动作集合（即所有 *SO* 值集合）中随机选择一个 *SO* 值作为下一个超帧周期占空比计算依据；而 π_2 表示贪婪策略，即协调器在每个超帧结束后以概率 $1-\varepsilon$ 选择该策略，并从查找表中选择对应 *Q* 值最大的 *SO* 值作为下一个超帧周期占空比的计算依据。通常 ε 较小（一般取值为 0.1 左右）。

综合上述可得到协调器端执行强化学习算法的 3 个阶段。

1) *Q* 值初始化阶段：为了对每个动作有一个“初始印象”，协调器必须首先对动作集合（即 *SO* 值集合）中的每个动作进行一个 *Q* 值初始化，协调器根据集合中动作个数的多少依次设置不同的占空比，并统计相应的奖励作为该动作的初始 *Q* 值大小。

2) 策略选择阶段即信标发送阶段：在每次发送信标前，协调器会进行相应的策略选择，得到下一个超帧周期的 *SO* 值，然后协调器发送信标帧，信标帧中包含了 *SO* 值和 *BO* 值，终端节点在接收到信标帧后根据 *SO* 值和 *BO* 值计算出相应的 *CAP* 和 *CFP*，并依次进入数据收发阶段和休眠状态。

3) 计算奖励及 *Q* 值更新阶段：在超帧周期内，协调器统计高优先级数据分组的延时和各个终端节点高优先级队列占用率情况，并在超帧末尾计算出延时、队列占用率及超帧空闲侦听率，再依据式(5)计算出当前动作（*SO* 值）的奖励大小，利用式(11)对当前动作（*SO* 值）的 *Q* 值进行更新；更新完成后，协调器进入下一个迭代阶段。

4　仿真及分析

为了验证强化学习算法的有效性，利用 NS2 仿真软件对本文提出的算法进行了仿真，仿真主要验证了强化学习针对不同应用的需求（即不同的高优先级业务对延时的需求）能否找到一个合适的 SO 值，从而保证高优先级业务的延时指标在高优先级业务延时界（HDB, high-priority delay bound）内，并且验证了在网络内流量出现变化时，强化学习的动态适应性，验证所提出的强化学习算法能否根据当前网络内流量情况及业务要求重新学习找到一个新的 *SO* 值。

4.1　仿真环境及参数设置

ZigBee 网络是一种以 IEEE 802.15.4 作为底

层通信协议的网络，应用本策略，该网络中的控制命令和数据采集可作为应用业务类型，因而，设置了如下的仿真场景。网络为信标使能的星形网络，网络中心为协调器节点，另外为网络配置 6 个终端节点，且要求均在协调器节点的通信半径内，并同时存在高优先级与低优先级业务数据流，终端节点实现了 BCS 算法，协调器端实现了强化学习算法。

仿真平台为 NS2，物理层、MAC 层、路由层分别采用 WirelessPhy/802_15_4、Mac/802_15_4、zbr 协议，业务模型为 cbr。表 2 为根据上述应用场景所设置的具体网络仿真参数。

表 2　验证强化学习仿真实验参数设置

参数		值
业务量大小/(packet·s^{-1})		0.2
数据分组长度/byte		70
节点个数		7
$BO(SO_{max})$		7
仿真时间/s		8 000
能量消耗/mW	initialEnergy	1 000
	rxPower	35.28×10^{-3}
	txPower	31.32×10^{-3}
	idlePower	712×10^{-6}
	sleepPower	144×10^{-9}
强化学习	Learning-Rate	0.1
	Exploring-Rate	0.1

4.2　仿真结果及分析

图 9 为 *HDB*=100 ms 时，*SO* 值的变化情况。

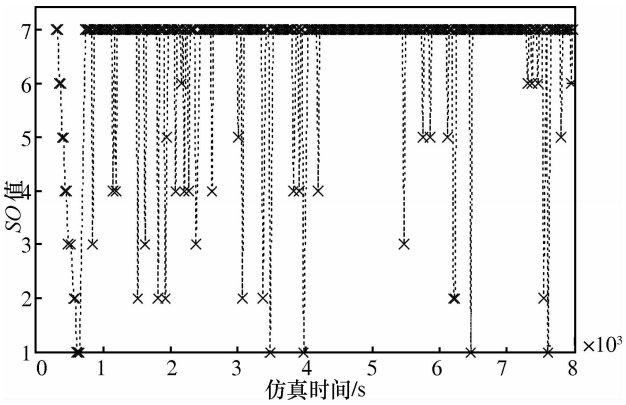
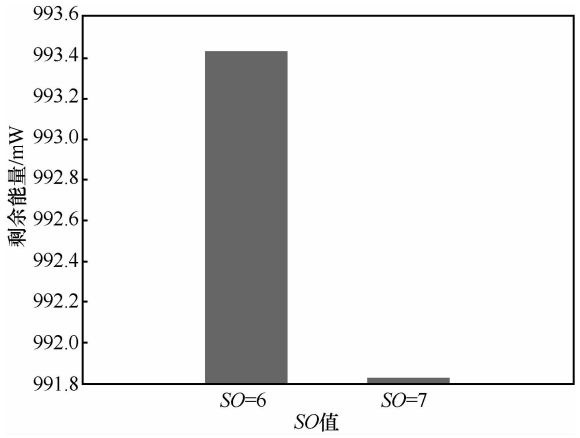
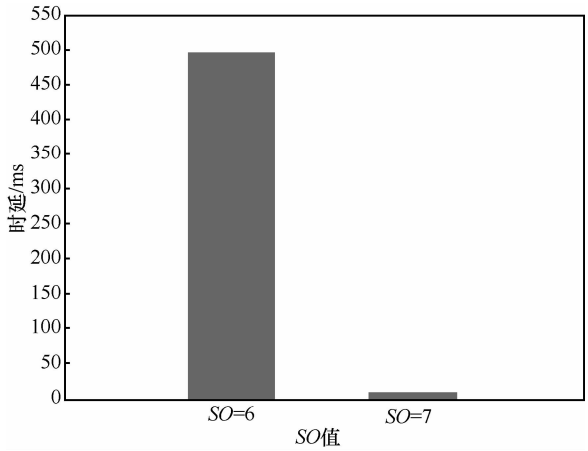


图 9　*HDB*=100 ms 时 *SO* 值变化

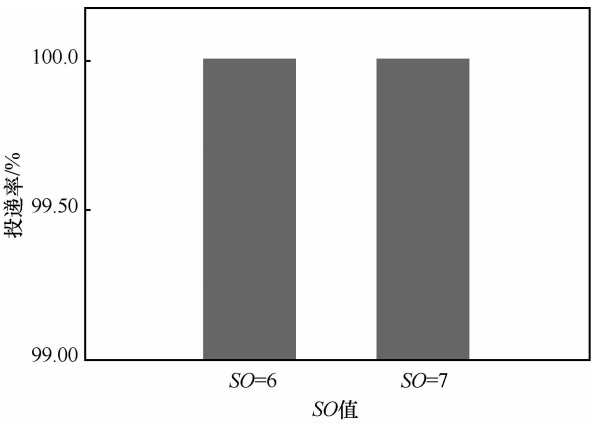
可以看出，为了满足 100 ms 的延时界需求，在 *SO*=7 被选择的次数占据绝大多数，说明 *SO*=7 即为“贪婪动作”。为了满足 100 ms 延时界的需求，此时必须要使节点的休眠时间尽量小，并且时隙数尽量多，避免相应的数据分组排队，而造成必须等到下一个超帧周期才能传输。为了证明 *SO*=7 在当前动作内是最优的，需要对 *SO*=6 时的延时、分组投递率及能量消耗情况进行对比，仿真结果如图 10 所示。



(a) *HDB*=100 ms 时相邻 *SO* 值与最优 *SO* 值剩余能量对比



(b) *HDB*=100 ms 时相邻 *SO* 值与最优 *SO* 值延时对比



(c) *HDB*=100 ms 时相邻 *SO* 值与最优 *SO* 值分组投递率对比

图 10　*HDB*=100 ms 时相邻 *SO* 值与最优 *SO* 值指标对比

针对仿真结果，从能量角度看，虽然 *SO*=6 时网络剩余能量较多，但是 *SO*=6 时高优先级业务的平均延时为 500 ms 左右，远远满足不了高优先级业务的延时需求，因而导致该动作对应的奖励较少；从分组投递率的角度来看，*SO*=7 和 *SO*=6 时分组投递率非常接近。因此，从满足高优先级业务性能指标上看，强化学习将 *SO*=7 作为“贪婪动作”是合理的。

图 11 为当高优先级业务的延时界（*HDB*）等于 1 000 ms 时，*SO* 值的变化情况。从图中可以看出，强化学习最终确定的“贪婪”动作为 *SO*=6。从图 12 中可以看出，当 *SO*=7 时，高优先级业务的延时在 1 000 ms 之下并且分组投递率和 *SO*=6 时非常接近，根据奖励计算公式可知，此时奖励的决定因素在于超帧空闲侦听率的大小，最终反映在网络整体剩余能量上。而如图 12（a）的仿真结果显示，当 *SO*=6 时，网络整体剩余能量要高于 *SO*=7 时的状态，因而，强化学习将更倾向于选择 *SO*=6 作为“贪婪动作”。而对应的 *SO*=5 不为“贪婪”动作的主要原因在于，当 *SO*=5 时，高优先级的平均延时大于 1 000 ms（如图 12(b) 所示），因而不能满足高优先级业务的延时指标需求。综上，在此情况下，选择 *SO*=6 作为“贪婪动作”也是合理的。

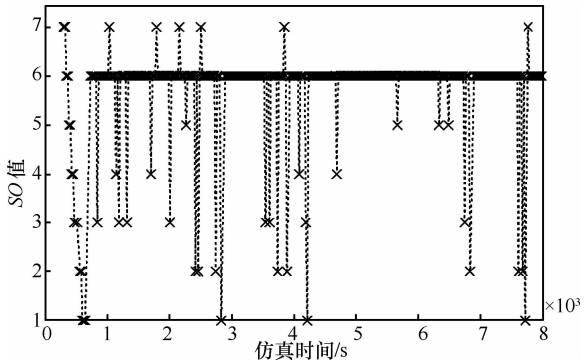
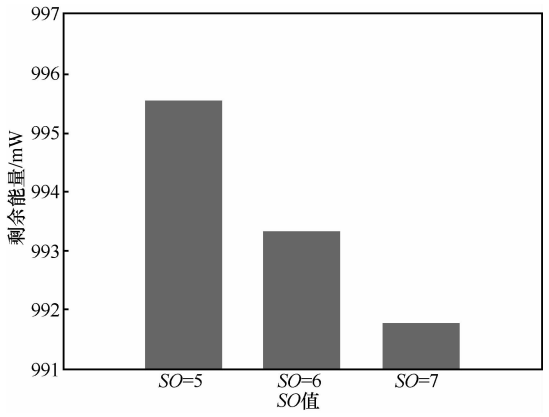
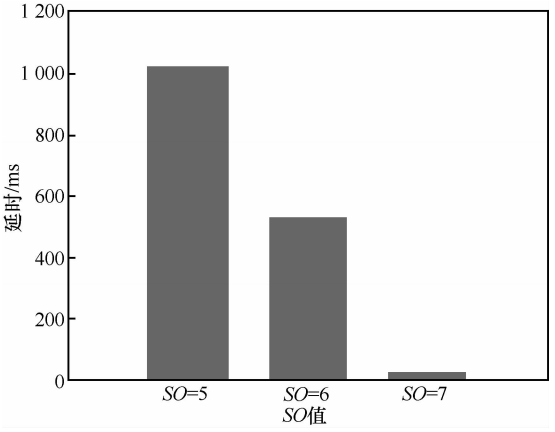


图 11　*SO* 变化情况

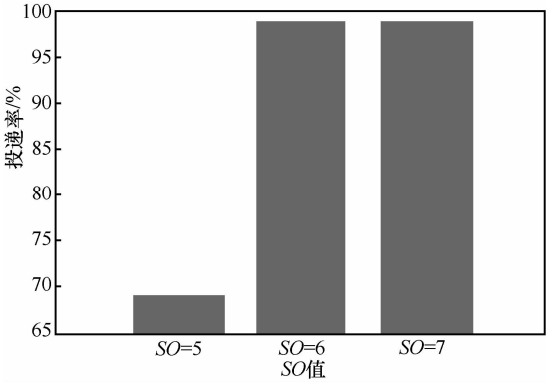
图 13 为当高优先级业务的延时界（*HDB*）为 3 500 ms 时，*SO* 值的变化情况。从图中可以看出，最终“贪婪”动作为 *SO*=4。值得注意的是，在选择 *SO*=4 前，*SO*=3 是“贪婪”动作，发生这一变化的主要原因是：在初始化后，直至 2 000 s 才再次“探索” *SO*=4 动作，并对其 *Q* 值进行了更新，因而最终决定 *SO*=4 才是“贪婪”动作。



(a) *HDB*=1 000 ms 时相邻 *SO* 值与最优 *SO* 值剩余能量对比



(b) *HDB*=1 000 ms 时相邻 *SO* 值与最优 *SO* 值延时对比



(c) *HDB*=1 000 ms 时相邻 *SO* 值与最优 *SO* 值分组投递率对比

图 12　*HDB*=1 000 ms 时相邻 *SO* 值与最优 *SO* 值指标对比

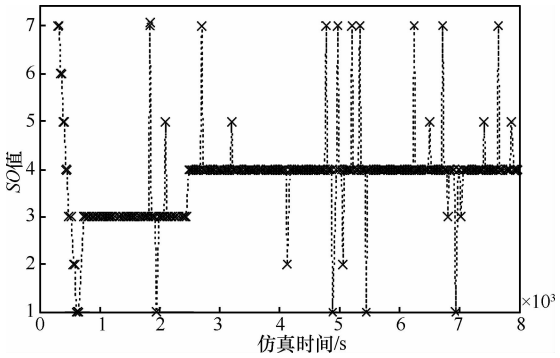
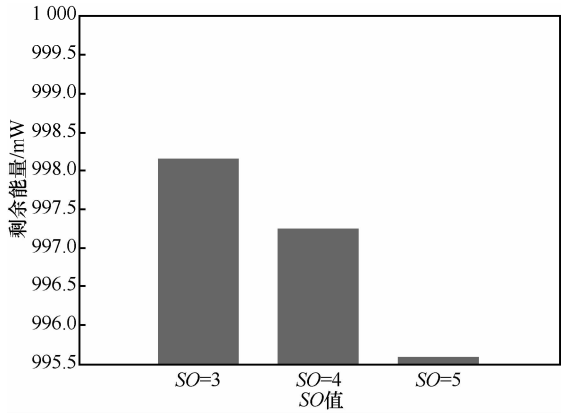
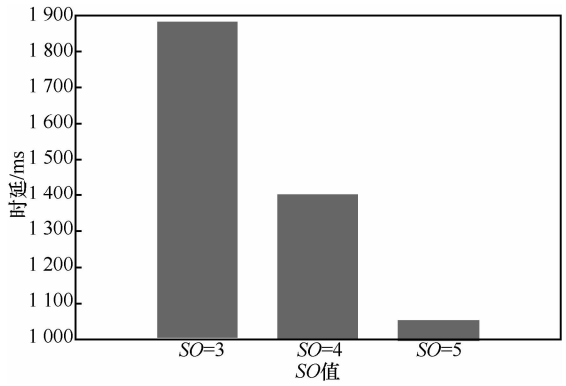


图 13　*SO* 变化情况

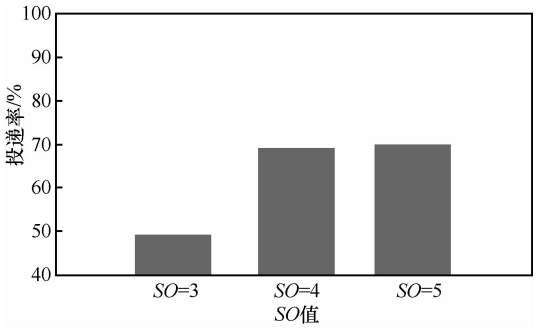
从图 14（b）的高优先级业务延时情况来看，所有数据分组的延时均远低于高优先级业务的延时界，此时奖励大小的决定性因素应在于高优先级业务队列占用率及超帧空闲侦听率。无论 *SO*=3、4 或 5 时，延时均在 3 500 ms 之下，而从分组投递率来看，*SO*=3 时，会因为分组投递率太低而被“淘汰”，而 *SO*=4 和 *SO*=5 时，由于分组投递率非常接近，因而决定它们奖励大小的因素转换为超帧空闲侦听率上面。由于 *SO*=4 时超帧空闲侦听率更低，网络内剩余能量更多，因以 *SO*=4 为“贪婪”动作。



(a) *HDB*=3 500 ms 时相邻 *SO* 值与最优 *SO* 值剩余能量对比



(b) *HDB*=3 500 ms 时相邻 *SO* 值与最优 *SO* 值延时对比



(c) *HDB*=3 500 ms 时相邻 *SO* 值与最优 *SO* 值分组投递率对比

图 14 *HDB*=3 500 ms 时相邻 *SO* 值与最优 *SO* 值指标对比

图 15 给出在 4 000 s 时，在 6 个终端节点中任选 3 个节点流量突然增大（变为 5 packet/s），*SO* 的变化情况，相对应的图 16 为高优先级业务和低优先级业务的延时情况。

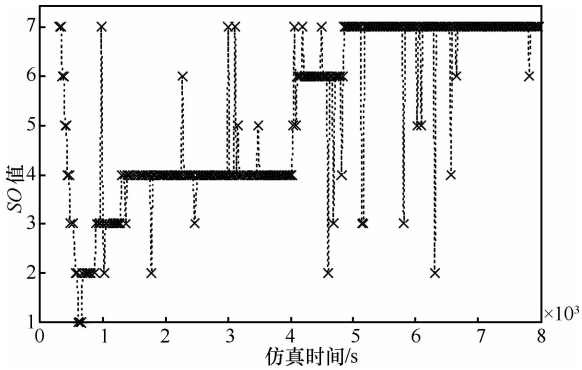


图 15 流量变化时 *SO* 值变化

从图中可以看出，在第 4 000 s 流量变化时，Agent 在对 *SO*=6 这个动作进行“探索”后发现该动作的奖励更好，同时更新该动作的 *Q* 值并将其作为新的“贪婪”动作。而通过不断的“探索”，最终将“贪婪”动作锁定在 *SO*=7。从图 16 延时的变化情况可以看出，在流量变大后，无论是高优先级业务还是低优先级业务，其延时急剧增加，这主要是由于 CAP 内时隙数量的限制，引起竞争冲突的加大。为了降低这种冲突，必须要加大 *SO* 的值，增加 CAP 内时隙的数量。协调器 Agent 在学习过程中，发现更大的 *SO* 值能满足相应的要求，并最终将“贪婪”动作锁定为 *SO*=7。但值得注意的是，当 *SO*=7 时，无论是高优先级数据分组的延时还是低优先级数据分组的延时都已经远远小于 3 500 ms，此时不选择 *SO*=6 作为“贪婪”动作的原因主要在于此时网络内流量较大，在 *SO*=7 的情况下高优先级业务的分组投递率相对较高。

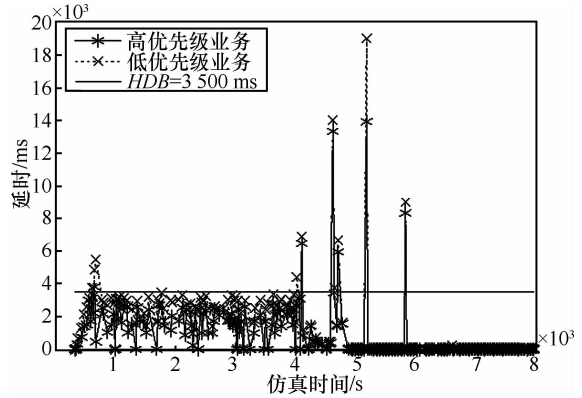


图 16 变化时高优先级业务和低优先级业务延时

综上，本文提出的 BCS 及强化学习方法能根据不同应用场景下高优先级业务的性能需求找到一个较优的占空比，并且能最大程度地节省网络流量。同时，提出的算法能在网络流量变化时实时调整占空比设置，真正体现了强化学习的自学习特性，具有非常好的应用价值。

5　结束语

本文重点研究了 IEEE 802.15.4 信标星型网络中的区分服务机制支持问题。为了弥补该协议下对区分服务机制支持的不足，本文在基于优先级区分策略的基础上分别从终端设备和协调器出发，创新性地提出了一种基于 BCS 退避策略与强化学习的区分服务机制，并基于 NS2 平台进行了仿真实验及结果分析。仿真结果表明，提出的算法能根据不同应用需求找出最佳占空比，在满足高优先级业务性能需求的同时最大程度节省网络能量；同时由于所提出强化学习算法的“自学习”特性，协调器端能实时追踪网络内流量变化并找到最优占空比设置，具有较好的应用价值。

参考文献：

[1] CALLAWAY E, GORDAY P, HESTER L. Home networking with IEEE 802.15.4: a developing standard for low-rate wireless personal area networks[J].IEEE Communications Magazine, 2002, 40(8):70-77.

[2] OTTO C, MILENKOVIĆ A, SANDERS C. System architecture of a wireless body area sensor network for ubiquitous health monitoring[J]. Journal of Mobile Multimedia, 2006, 1(4): 307-326.

[3] CAO X, CHEN J, ZHANG Y. Development of an integrated wireless sensor network micro-environmental monitoring system[J]. ISA Transactions, 2008, 47(3): 247-255.

[4] IEEE 802.15.4 Standard (2003) Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)[S]. IEEE-SA Standards Board, 2003.

[5] TSENG P, FENG K, LIN Y. Wireless location tracking algorithms for environments with insufficient signal sources[J]. IEEE Transactions on Mobile Computing, 2009, 8(12): 1676-1689.

[6] HO C, LIN C, HWANG W. Dynamic GTS allocation scheme in IEEE 802.15.4 by multi-factor[A].2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)[C]. Piraeus, 2012.457-460.

[7] KOUBÂA A, ALVES M, NEFZI B. Improving the IEEE 802.15.4 slotted CSMA/CA MAC for time-critical events in wireless sensor networks[A].Proceedings of the Workshop on Real Time Networks (RTN ’06)[C]. Dresden, 2006.270-277.

[8] KIPINS D, WILLIG A, HAUER J. The angel IEEE 802.15.4 enhancement layer: coupling priority queueing and service differentiation[A]. Wireless Conference, EW 2008. 14th European[C]. Prague, 2008.1-7.

[9] ROHM D, GOYAL M, HOSSEINI S. A simulation based analysis of the impact of IEEE 802.15.4 MAC parameters on the performance under different traffic loads[J].Mobile Information Systems, 2009, 5(1): 81-99.

[10] SUTTON R, BARTO A. Reinforcement Learning: An Introduction[M]. Massachusetts: MIT Press, 1998.

[11] SUTTON R, BARTO A. Introduction to Reinforcement Learning[M]. Massachusetts: MIT Press, 1998.

[12] Multi-armed bandit[EB/OL]. http://en.wikipedia. org/wiki/ Multi- armed_bandit, 2014.

[13] HERNANDEZ-GARDIOL N, MAHADEVAN S. Hierarchical memory-based reinforcement learning[A]. 14th Annual Neural Information Processing Systems Conference[C]. 2001.1047-1053.

作者简介：

钱亮（1987-），男，满族，吉林四平人，吉林大学硕士生，主要研究方向为短距离无线通信技术、IEEE 802.15.4 网络、射频识别技术。

钱志鸿（1957-），男，吉林长春人，吉林大学教授、博士生导师，主要研究方向为近程无线网络通信技术、无线传感器网络技术、RFID（射频识别）技术、UWB（超宽带）通信技术和物联网等。

李天平（1989-），男，湖北荆州人，吉林大学硕士生，主要研究方向为低速率无线个域网。

全薇（1964-），女，辽宁沈阳人，吉林大学教授、硕士生导师，主要研究方向为光学信息处理。