超蜂窝网络能效延时理论关系及最优基站休眠控制

郭雪莹^{1,2}, 冷冰洁^{1,2*}, 吴健^{1,2}, 牛志升^{1,2}

1. 清华大学电子工程系, 北京 100084

2. 清华大学信息科学与技术国家实验室(筹), 北京 100084

* 通信作者. E-mail: lengbj14@mails.tsinghua.edu.cn

收稿日期: 2017-02-15; 接受日期: 2017-03-21; 网络出版日期: 2017-06-09

国家重点基础研究发展计划(973)(批准号: 2012CB316000)和国家自然科学基金(批准号: 61461136004, 61571265, 91638204)资助项目

摘要 超蜂窝架构通过控制覆盖与业务覆盖的适度分离,使得业务基站休眠成为可能,实现按需调节业务覆盖,提升网络能效.但业务基站休眠会对业务延时带来潜在影响.为此,定量分析基站休眠机制对系统能效、业务延时的影响,并权衡二者给出最优休眠机制成为重要的研究课题.本文总结了笔者课题组针对此问题的一系列研究成果:针对业务基站休眠场景,借助休假排队模型给出了系统平均功率与业务延时的闭式表达式,并总结二者存在折中的条件;进一步在给定延时约束下给出能效最优的基站休眠策略,由此得到最优能效延时折中关系,同时分析了探测能耗对能效延时折中关系的影响.同时,在突发业务到达模型下,给出基站休眠机制设计,并将其与功率匹配机制相结合;基于部分可观测的 Markov 决策过程 (partially observable Markov decision process, POMDP),给出突发业务到达下的最优基站休眠策略.

关键词 超蜂窝网络, 休假排队模型, 业务基站休眠, 能效延时折中, 突发业务

1 引言

移动通信业务的快速发展带来了巨大的网络能耗,使绿色通信得到了学术界和工业界的广泛关注,在此情况下超蜂窝架构应运而生^[1].超蜂窝是资源与能效全面优化的未来移动通信网络架构,其中控制基站与业务基站并存,控制基站提供持续的控制覆盖,业务基站提供动态变化的业务覆盖.蜂窝网络传统上基于峰值流量进行部署,而业务流量时空分布的不均匀性导致大量基站长期处于低载或空载状态,考虑到基站的基础功率占到总功率的 60% 以上^[2],这一问题带来巨大的能量浪费,从而使基站休眠得到广泛关注.但传统蜂窝网络中,引入基站休眠将会引起控制信令覆盖空洞,且该问题由于网络

引用格式: 郭雪莹, 冷冰洁, 吴健, 等. 超蜂窝网络能效延时理论关系及最优基站休眠控制. 中国科学: 信息科学, 2017, 47: 771–788, doi: 10.1360/N112017-00033
Guo X Y, Leng B J, Wu J, et al. Energy-delay tradeoff and optimal base station sleeping control in hyper-cellular networks (in Chinese). Sci Sin Inform, 2017, 47: 771–788, doi: 10.1360/N112017-00033

结构限制不易解决. 新兴的超蜂窝架构通过控制覆盖与业务覆盖的适度分离, 可根据实际需求开启或关闭业务基站, 而不引起控制信令覆盖空洞的问题, 从架构上方便网络提供灵活、动态的服务, 提升网络能效.

然而, 这一新的网络架构也会引入新的问题. 超蜂窝中, 业务基站休眠在降低网络能耗的同时, 也对业务延时带来潜在的影响, 由此需要定量衡量业务基站休眠机制对系统平均功率与用户延时的影响, 并权衡两者设计最优基站休眠机制. 同时, 需要回答一个基本问题: 牺牲一定的业务延时, 能换取多大的网络能效提升? 这一能效延时折中问题, 是网络能效理论的重要组成部分, 也是超蜂窝网络中业务基站休眠机制设计的重要依据.

基站休眠对服务质量的影响一直是学术研究的关注点. 文献 [3] 在保证用户维持可接受的吞吐量的同时, 探索节能的基站休眠操作, 基于 Markov 模型得到了用户达到目标吞吐量的概率. 文献 [4, 5] 在阻塞概率约束下, 研究休眠操作. 但仅有部分文献考虑了基站休眠对业务延时的影响. 文献 [6] 以同时优化流级别延时和基站平均功率开销为目标, 将基站的动态休眠与基站用户关联相结合. 文献 [7] 在基于正交频分多址 (OFDMA) 的蜂窝系统中提出了“离散型传输机制”, 将延时容忍业务聚集起来, 以提升网络能效. 综上, 已有文献在研究基站休眠问题时, 往往忽略延时或仅考虑传输延时; 同时在休眠机制设计上, 往往忽略了基站状态转换开销, 如启动时间、探测能耗等, 将基站休眠等同于基站-用户关联问题, 但基站休眠机制的设计中应同时考虑基站硬件情况及具体网络架构, 深入衡量休眠机制 [8]. 另外, 已有文献中的基站休眠策略设计主要基于启发式算法, 对于最优基站休眠算法的理论分析不足.

能效延时折中作为无线通信中的基本问题得到了长期关注. 如文献 [9, 10] 在衰落信道中考虑能效延时折中, 并在渐近分析下指出, 能效延时之间存在着根方形式的折中关系. 文献 [11, 12] 将这一根方律扩展到多用户场景当中. 文献 [13] 指出, 在固定信道中, 通过增大数据包的传输时间, 发送固定个数包的总能量将大大降低, 即在信道与到达情况已知时, 增大延时是提升能效的有效手段. 文献 [14] 在码分多址 (CDMA) 系统中, 文献 [15] 在工业无线网络中, 对该问题进行衡量. 然而超蜂窝中的能效延时折中问题有其特殊性. 首先, 为考虑全网总能量, 必须加入对基站基础能耗的考量. 由文献 [2] 可知, 即使在空载的情况下, 宏基站 (macro BS) 功率开销仍占到其满负载时功率开销的 60% 左右; 而微基站 (micro BS) 中该比例达到 73%. 在微微基站 (pico BS) 和毫微微基站 (femto BS) 中, 这一比例更是高达 90% 以上. 其次, 需要考虑业务总延时, 既包含由多个用户共享基站资源带来的业务在系统中的等待、处理时间, 又包含采用业务基站休眠时带来额外的延时增长.

在业务基站休眠设计当中, 也应考虑业务突发性的影响. 数据业务在蜂窝网络中的占比愈来愈高, 其区别于传统的语音业务具有较大的突发性 [16]. 因此, 利用突发动态特性, 可根据实时业务到达率按需匹配资源 [17, 18]. 目前已有文献利用突发业务模型对网络性能进行了分析. 文献 [19] 中假设业务到达服从间歇 Poisson 过程, 在局域网中研究了具备额外活跃区间的交换机休眠机制, 文献 [20] 中分别以 Poisson 过程以及间歇 Poisson 过程对业务建模, 对光交换网络进行了性能分析. 而以降低总功率为目的, 针对基站休眠机制研究突发业务到达影响的已有文献不足.

由此, 本文希望全面考虑网络总功率与业务总延时, 定量衡量业务基站休眠机制的影响. 同时, 根据能效延时折中关系考量业务基站休眠中的非理想因素, 设计最优基站休眠策略, 并进一步衡量业务突发性对基站休眠机制的影响. 工作内容总结如下:

(1) 基于业务基站休眠的能效延时折中关系分析. 为定量分析业务基站休眠对系统总功率与业务延时的影响, 需要对一般化的模型开展理论研究. 其中需要衡量基础能耗、排队延时, 并考虑基站休眠中非理想因素的影响. 为此, 本研究利用休假排队模型, 考虑了迟滞休假机制及单重休假、多重休假和累积任务等典型唤醒机制, 给出了系统长时平均功率与平均延时的闭式分析. 在此基础上, 分别对各唤

醒机制分析了基站启动时间、休假时间、累积任务数和迟滞时间等系统参数对功率及延时的影响,并得到了多组参数变化下能效延时关系的闭式表达^[21].一方面,这一分析为业务基站设备制造提供指导,如指出减小启动时间均值或离散系数均可同时提升能效及延时性能^[22].另一方面,给出了能效及延时间存在折中的条件:在休眠增益条件成立时,能效及延时随迟滞时间变化呈线性折中关系;在休假时间或累积任务数变化时,能效延时折中条件更为复杂,且与启动时间紧密相关^[23].同时,累积任务机制的非折中区间中,给出了能效延时双赢的参数选择方法^[22, 23].

(2) 最优业务基站休眠策略设计及最优能效延时折中.对基站休眠控制参数进行联合优化,通过分析发现,最优参数组在 3 种典型唤醒机制下均具有两段式的特点:当延时约束小于一定阈值时,只需调整迟滞时间;当延时约束大于该阈值且启动时间的离散系数较小时,只需调整休假时间或累积任务数^[23],且根据此特点得到最优控制参数组.进一步闭式分析了最优系统平均功率与延时约束之间的关系,指出当延时约束小于一定阈值时,最优长时平均功率与延时约束呈线性关系,即一定量的延时容忍换取的能效提升保持不变.而当延时约束大于该阈值时,能效延时关系呈单调减的凸函数,即一定量的延时容忍换取的能效提升随延时增长逐渐下降^[24].另外,通过数值分析指出,尽管累积任务机制性能优于多重休假机制,多重休假机制又优于单重休假机制,前两种机制的优势都将在探测能耗增加后迅速减小直至消失^[24].

(3) 突发业务到达模型下的基站休眠机制设计.考虑单业务基站下行场景,首先将业务到达建模为间歇 Poisson 过程 (interrupted Poisson process, IPP),采用累积任务唤醒机制并结合功率匹配,给出了系统长时平均功率 (亦代表系统总功耗) 以及平均延时的闭式表达式,并发现随着业务突发性的增大,系统总功耗下降,延时性能在特定情况下变好^[25].进一步地,考虑业务的相关性,将业务到达建模为 Markov 调制 Poisson 过程 (Markov modulated Poisson process, MMPP),发现业务的相关性对总功耗影响不明显,且仅在系统利用率较高时对延时性能产生影响.进一步在延时约束下优化休眠门限以及传输功率,使系统总功耗最小,由此得到功耗与延时之间的最优折中曲线,发现此时总功耗的下界仅与平均到达率有关,而与业务突发性无关^[26].另一方面,利用部分可观测的 Markov 决策过程 (partially observable Markov decision process, POMDP) 对系统进行建模,综合考虑功率与延时,发现最优基站休眠策略存在双门限结构,且随着业务突发性的增加,系统性能将提升^[27].

本文组织如下:第 2 节给出业务基站休眠机制的设计方式,并对系统性能进行定量分析;第 3 节给出各休眠参数影响下能效延时关系及折中区间分析;第 4 节给出最优业务基站休眠机制设计并给出最优能效延时关系;第 5 节给出突发到达下的基站休眠机制,并将其与功率匹配相结合;第 6 节给出突发到达下,基于部分可观测的 Markov 决策过程的最优业务基站休眠机制.

2 业务基站休眠机制下系统性能定量分析

首先建立超蜂窝网络的排队模型,具体地,将业务基站整体建模为服务器,将流级别 (flow level) 的业务抽象为有待处理的业务.考虑任务到达过程服从 Poisson 分布的情况,到达率为 λ .假设基站同一时刻只能为某一任务提供服务¹⁾,其服务时间用随机变量 B 表示,用 h_B 表示其均值, c_B^2 表示其方差系数, $\tilde{B}(s)$ 表示其 Laplace-Stieltjes 变换 (LST 变换).系统利用率表示为 $\rho = \lambda h_B$.

在基站能耗模型上,采用线性模型^[28],即总功率与传输功率呈线性关系,该模型受到学术界的广

1) 以下结果也适用于基站按照服务者共享规则提供服务的情况^[24].

泛认可. 基站总功率 P 满足

$$P = \begin{cases} P_{\text{ID}} + P_{\text{TR}}/\Delta, & \text{工作模式;} \\ P_{\text{SL}}, & \text{休眠模式,} \end{cases}$$

其中, P_{ID} 是基站空载时的功率, P_{TR} 是基站工作状态下的射频传输功率, Δ 是负载相关功率开销的效率, P_{SL} 是基站休眠模式下的功率开销. 另外, 假设基站在启动 (setup) 阶段的功率为 P_{ST} .

业务基站休眠机制 (何时进入休眠) 及唤醒机制 (何时及如何唤醒) 需根据网络运行中的实际因素制定. 本文根据超蜂窝网络情况, 考虑了迟滞 (hysteresis) 休眠机制和 3 种典型的唤醒机制.

迟滞休眠机制 (hysteresis sleep). 当基站负载变空后, 系统启动一个定时器, 如果在定时器规定的等待时间内有新业务到达, 则基站立即进行服务; 如果直至定时结束仍无新业务到达, 则使基站进入休眠模式. 这一迟滞休眠机制能够有效地减少频繁的基站模式切换, 从而降低模式切换开销. 其中迟滞时间用随机变量 D 表示, 分别用 h_D , c_D^2 和 $\tilde{D}(s)$ 代表其均值、方差系数及 LST 变换.

单重休假唤醒机制 (single sleep scheme, 即 SS scheme). 基站休眠一段特定的时间后即自动醒来. 这一机制的突出特点在于其易于实施, 即并不需要为唤醒基站增加新的本地或网络设备, 从而具有较广的应用范围. 具体来讲, 基站仅需在休眠前做好定时, 而并不需要在休眠期间对业务、网络状态有所感知. 因此, 应用单重休假机制时, 休眠模式下的功率相对较低. 休眠时长的设计对于单休假机制下的系统性能至关重要. 休假时间用随机变量 V 表示, 分别用 h_V , c_V^2 和 $\tilde{V}(s)$ 代表其均值、方差系数及 LST 变换.

多重休假唤醒机制 (multiple sleep scheme, 即 MS scheme). 基站休眠特定时长后醒来并探测是否已有新业务到达, 如已有新业务到达则进行唤醒, 否则再次进入休眠状态. 当基站能够部分地唤醒并探测是否有新业务正在等待时, 这一机制极为实用. 相对于单休假机制, 多休假机制有更多的节能潜力, 因为其利用部分唤醒探测的机制使总体休眠时间得到延长. 然而, 部分唤醒探测也会带来额外的能量开销. 其中, 假定各段休假时间独立同分布, 用随机变量 V 表示, 分别用 h_V , c_V^2 和 $\tilde{V}(s)$ 代表其均值、方差系数及 LST 变换. 另外, 探测能耗为 E_{DT} .

累积任务唤醒机制 (N scheme). 在有 N 个新服务请求到达后唤醒基站. 为实施这一机制, 已休眠基站需要有部分设备 (可能具有较低功率) 保持工作以计算目前等待的业务数, 或有网络机制计算不断到达的服务请求并能够唤醒已休眠基站. 由于累积任务机制下, 休眠基站需对网络状态有所了解, 因此该机制有带来更好系统性能的潜力. 然而, 在休眠期间统计到达业务数的需求也可能使系统在休眠期间的功率有所增加. 假设休眠期间的探测功率为 P_{DT} .

图 1 以单重休眠唤醒机制为例给出了业务基站操作的示意图. 可以看到, 基站操作具有阶段性, 且功率水平随之变化. 在以下分析中, 假设 $P_{\text{ST}} > P_{\text{ID}} > P_{\text{SL}} + P_{\text{DT}}$, 这一假设不失一般性: 其中前一个不等式刻画了基站的状态转换开销, 后一个不等式意味着基站休眠模式的功率小于工作模式的空载功率.

通过更新过程及休假排队论相关理论, 可得到系统长时平均功率和业务平均延时的闭式分析. 推导过程参见文献 [21, 23]. 考虑到已有文献 [29, 30] 中指出了延时界对视频及语音类业务的重要性, 工作也给出了延时分布拉式变换的闭式分析, 并利用数值结果分析了基站休眠机制下业务延时的统计特性, 详见文献 [24]. 但是, 由于针对延时界不易给出闭式分析, 且平均延时改善时延时界性能也会有所提升, 本文主要考虑平均延时给出超蜂窝网络能效延时理论分析. 下面给出具体结果.

图1 单重休假唤醒机制下, 业务基站操作阶段示意图

Figure 1 BS operation transition diagram for the single sleep (SS) scheme

3 种唤醒机制下的系统长时平均功率分别为

$$E[P^{(SS)}] = (1 - \rho) \left\{ p_v h_V P_{SL} + p_v h_S P_{ST} + \left[\frac{1}{\lambda} (1 - p_v) + \frac{1}{\lambda} p_v \tilde{V}(\lambda) \tilde{S}(\lambda) \right] P_{ID} \right\} \\ \left/ \left[\frac{1}{\lambda} (1 - p_v) + p_v \left(h_V + h_S + \frac{1}{\lambda} \tilde{V}(\lambda) \tilde{S}(\lambda) \right) \right] + \rho (P_{ID} + P_{TR}/\Delta) \right\}, \quad (1)$$

$$E[P^{(MS)}] = (1 - \rho) \left[(1 - p_v) P_{ID}/\lambda + \frac{p_v h_V}{1 - \tilde{V}(\lambda)} P_{SL} + p_v h_S P_{ST} + \frac{p_v}{1 - \tilde{V}(\lambda)} E_{DT} \right] \\ \left/ \left[(1 - p_v)/\lambda + p_v h_V \left(1 - \tilde{V}(\lambda) \right)^{-1} + p_v h_S \right] + \rho (P_{ID} + P_{TR}/\Delta) \right\}, \quad (2)$$

$$E[P^{(N)}] = \rho (P_{ID} + P_{TR}/\Delta) + (1 - \rho) \left[(1 - p_v) P_{ID}/\lambda + p_v \frac{N}{\lambda} (P_{SL} + P_{DT}) + p_v h_S P_{ST} \right] \\ \left/ \left[\frac{1}{\lambda} (1 - p_v) + p_v \frac{N}{\lambda} + p_v h_S \right] \right\}, \quad (3)$$

其中, p_v 为休眠概率, 表示经过迟滞阶段的基站最终进入休眠模式的概率, 即

$$p_v \triangleq \tilde{D}(\lambda) = E[e^{-\lambda D}]. \quad (4)$$

3 种唤醒机制下的平均延时分别为

$$E[T^{(SS)}] = \bar{T}_0 + \frac{\lambda p_v [h_V^2 (1 + c_V^2) + h_S^2 (1 + c_S^2) + 2h_V h_S]}{2 + 2p_v [\lambda (h_S + h_V) + \tilde{V}(\lambda) \tilde{S}(\lambda) - 1]}; \quad (5)$$

$$E[T^{(MS)}] = \bar{T}_0 + \lambda p_v \left/ \left[2(1 - \tilde{V}(\lambda)) \right] \frac{h_V^2 (1 + c_V^2) + h_S^2 (1 + c_S^2) (1 - \tilde{V}(\lambda)) + 2h_V h_S}{p_v [\lambda h_V / (1 - \tilde{V}(\lambda)) + \lambda h_S - 1] + 1} \right. \right]; \quad (6)$$

$$E[T^{(N)}] = \bar{T}_0 + \frac{p_v [N(N-1) + 2N\lambda h_S + \lambda^2 h_S^2 (1 + c_S^2)]}{2\lambda [p_v (N + \lambda h_S) + 1 - p_v]}, \quad (7)$$

其中 c_V^2 , c_S^2 和 c_B^2 分别代表休假时间、启动时间和服务时间的方差系数, 且

$$\bar{T}_0 \triangleq h_B + \frac{\rho^2 (1 + c_B^2)}{2\lambda(1 - \rho)}$$

为业务基站不休眠场景下业务的平均延时.

3 能效延时关系和折中区间分析

本节具体分析各休眠参数对系统性能的影响. 首先, 为使引入基站休眠带来能效增益, 基站休眠控制参数需满足以下基本不等式, 称为休眠增益条件:

$$\left. \begin{aligned} h_V(P_{ID} - P_{SL}) &> h_S(P_{ST} - P_{ID}), && \text{单重休假机制;} \\ \frac{h_V(P_{ID} - P_{SL}) - E_{DT}}{1 - \tilde{V}(\lambda)} &> h_S(P_{ST} - P_{ID}), && \text{多重休假机制;} \\ N(P_{ID} - P_{SL} - P_{DT}) &> \lambda h_S(P_{ST} - P_{ID}), && \text{累积任务机制.} \end{aligned} \right\} \quad (8)$$

以下在上述休眠增益条件下, 分析各参数影响.

(1) 启动时间的影响. 首先, 对 3 种唤醒机制下的分析都表明, 越小的启动时间方差系数 c_S^2 将改善系统性能, 即同时降低功率和减小延时. 方差系数 c_S^2 衡量了启动时间的随机性. 由于启动时间与业务基站的具体设计直接相关, 这一结论的实际指导意义在于, 各设备商在制造业务基站时, 除了降低启动时间均值外, 还应尽量降低启动时间的方差系数. 同时, 在大多数情况下, 平均启动时间 h_S 的增长同时带来功率与延时同时增加, 为二者的非折中区间, 存在合理选择参数实现双赢的可能. 仅在多重休假唤醒机制下且休眠控制参数随机性过大时存在例外情况. 图 2 给出了完全非折中情况和存在非折中区间的示例.

(2) 休假时间或累积任务数的影响. 在以下不等式成立时, 较大的平均休假时间 h_V (单重/多重休假机制) 或较大的累积任务数 N (累积任务机制), 将降低功率却增加延时, 即能效延时为折中关系: (i) 单重休假机制下, $c_S^2 \leq 1$ (充分条件); (ii) 多重休假机制下, $\lambda \leq 1/h_S$ 且 $c_S^2 \leq 1$ (充分条件); (iii) 累积任务机制下, $c_S^2 \leq 1$ (充分条件) 或

$$\lambda^2 h_S^2 (1 - c_S^2) p_v + 2\lambda h_S + 2 \geq 0 \text{ (充要条件)}. \quad (9)$$

当以上不等式不成立时, 能效与延时不再是简单的折中关系. 以累积任务机制为例, 式 (9) 不成立时, 累积任务数 N 存在拐点

$$N' = \sqrt{\lambda^2 c_S^2 h_S^2 + \lambda h_S + \left(\frac{1-p_v}{p_v}\right)^2} - \lambda h_S - \frac{1-p_v}{p_v}.$$

当 N 值小于该拐点 N' 时, 平均延时和功率会随着 N 的增大而同时减小, 为共赢区间; 大于该拐点时, N 增长时会导致平均延时增加却可以减小功率, 仍为折中区间. 即累积任务机制下, 累积任务数变化时, 平均延时与能效的共赢条件为不等式 (9) 不成立且 $N < N'$. 图 3 给出了式 (9) 成立与否的两种情况下, 延时随 N 变化的示例. 对共赢区间存在情况的直观解释如下: 当启动时间方差系数过大时, 即启动时间随机性过大时, 应使累积任务数 N 足够大, 以应对业务基站状态转换时可能出现较大启动时间的情况.

图 2 多重休眠唤醒机制下, 平均启动时间变化时归一化长时平均功率 (依无休眠时的平均功率归一化) 与平均延时的关系. $h_D = 0$ s, $h_V = 10$ s

Figure 2 For the multiple sleep (MS) scheme, the normalized average power consumption vs. average delay when the setup time changes. $h_D = 0$ s, $h_V = 10$ s

图 3 累积任务唤醒机制下, 平均延时随累积任务数变化的情况. $h_S = 3$ s, 迟滞时间具有指数分布且 $h_D = 1$ s

Figure 3 For the N -limited scheme, average delay vs. N . $h_S = 3$ s, exponentially distributed hysteresis time with $h_D = 1$ s

更进一步, N 变化时, 平均延时与功率间的闭式关系为

$$E[T^{(N)}] = \bar{T}_0 + \frac{1}{2\lambda} \left\{ 1 - \frac{2}{p_v} + \frac{1}{p_v F_2^{(N)}} + F_2^{(N)} \left[\lambda h_S p_v + \lambda^2 p_v c_S^2 h_S^2 + \frac{1 - p_v}{p_v} \right] \right\},$$

其中, $F_2^{(N)} = \{[E[P^{(MS)}] - \rho(P_{ID} + \frac{P_{TR}}{\Delta})]/(1 - \rho) - P_{SL} - P_{DT}\} / [\lambda h_S p_v (P_{ST} - P_{SL} - P_{DT}) + (1 - p_v)(P_{ID} - P_{SL} - P_{DT})]$.

(3) 迟滞时间的影响. 首先迟滞时间对系统性能的影响仅通过休眠概率 p_v 体现, 其定义见式 (4),

图 4 (网络版彩图) 单重休假唤醒机制下, 迟滞时间变化时归一化长时平均功率 (依无休眠时的平均功率归一化) 与平均延时的关系. 启动时间具有指数分布且 $h_S = 5$ s

Figure 4 (Color online) For the single sleep scheme, the normalized average power consumption vs. average delay when hysteresis time changes. Exponentially distributed setup time with $h_S = 5$ s

物理意义为迟滞时间内无新任务到达的概率, 若考虑特定迟滞时间分布, 增大休眠概率意味着减小迟滞时间均值. 进一步地, 迟滞时间变化时, 功率延时呈线性关系,

$$\begin{aligned} E[P^{(SS)}] &= P_{ID} + \rho P_{TR}/\Delta + 2(1-\rho) \frac{-h_V(P_{ID} - P_{SL}) + h_S(P_{ST} - P_{ID})}{h_V^2(1+c_V^2) + h_S^2(1+c_S^2) + 2h_Vh_S} (E[T^{(SS)}] - \bar{T}_0), \\ E[P^{(MS)}] &= P_{ID} + \rho P_{TR}/\Delta \\ &\quad + 2(1-\rho) (E[T^{(MS)}] - \bar{T}_0) \frac{-h_V(P_{ID} - P_{SL}) + E_{DT} + h_S(P_{ST} - P_{ID})(1 - \tilde{V}(\lambda))}{h_V^2(1+c_V^2) + (1 - \tilde{V}(\lambda))h_S^2(1+c_S^2) + 2h_Vh_S}, \\ E[P^{(N)}] &= P_{ID} + \rho P_{TR}/\Delta + 2\lambda(1-\rho) (E[T^{(N)}] - \bar{T}_0) \frac{-N(P_{ID} - P_{SL} - P_{DT}) + \lambda h_S(P_{ST} - P_{ID})}{N^2 - N(1 - 2\lambda h_S) + \lambda^2 h_S^2(1+c_S^2)}. \end{aligned}$$

图 4 以单重休假机制为例展示了这一关系.

4 保障延时约束的功率最优基站休眠参数设计及给定休眠机制下的最优能效延时折中

考虑优化问题

$$\min E[P] \quad \text{s.t.} \quad E[T] \leq \tau, \quad (10)$$

即在平均延时约束 τ 下, 设计功率最优基站休眠策略. 求解该优化问题, 得到最优休眠控制参数.

定理1 在单重休假、多重休假及累积任务机制下, 延时约束均存在一个阈值 τ_{th} .

(1) 在 $\tau \leq \tau_{th}$ 时, 最优休假时间 V (单重、多重休假机制) 或累积任务数 N (累积任务机制) 对不同的 τ 保持不变, 而最优的休眠概率 p_v 随 τ 单调增, 即此时只需调整迟滞时间. (2) 在 $\tau \geq \tau_{th}$ 时, 最

优休眠概率 p_v 恒为 1 (即最优迟滞时间恒为 0), 而最优休假时间 h_v 或最优累积任务数 N 则随 τ 单调增, 即此时迟滞时间恒为 0, 只需调节休假时间或累积任务数.

证明 参见文献 [24].

以累积任务机制为例给出最优控制具体结果.

(1) 在 $\tau \leq \tau_{th}$ 时, 最优休眠概率和最优累积任务数为

$$p_v^* = 2(\tau - \bar{T}_0) / [\lambda h_S^2(1 + c_S^2) + 2N^*h_S + N^*(N^* - 1)/\lambda - 2(\tau - \bar{T}_0)(\lambda h_S + N^* - 1)], \quad (11)$$

其中,

$$N^* = \begin{cases} 1, & \text{若 } [1 - \lambda^2 h_S^2(1 + c_S^2)](P_{ID} - P_{SL} - P_{DT}) - \lambda h_S(P_{ST} - P_{ID})(1 + 2\lambda h_S) \geq 0; \\ \frac{\lambda h_S(P_{ST} - P_{ID})}{P_{ID} - P_{SL} - P_{DT}} \left\{ 1 + \sqrt{1 + \left(2 - \frac{1}{\lambda h_S}\right) \frac{P_{ID} - P_{SL} - P_{DT}}{P_{ST} - P_{ID}} + (1 + c_S^2) \left[\frac{P_{ID} - P_{SL} - P_{DT}}{P_{ST} - P_{ID}}\right]^2} \right\}, & \text{其他情况.} \end{cases} \quad (12)$$

阈值 $\tau_{th}^{(N)}$ 为

$$\tau_{th}^{(N)} = \bar{T}_0 + \frac{\lambda h_S^2(1 + c_S^2) + 2N^*h_S + N^*(N^* - 1)/\lambda}{2\lambda h_S + 2N^*}.$$

(2) 在 $\tau \geq \tau_{th}$ 时, 最优 $p_v = 1$ (即 $D = 0$). 此时最优累积任务数为

$$N^* = \max \left\{ 1, \left\lfloor \lambda(\tau - \bar{T}_0 - h_S) + \frac{1}{2} + \lambda \sqrt{\left(\tau - \bar{T}_0 + \frac{1}{2\lambda}\right)^2 - h_S^2 c_S^2 - \frac{h_S}{\lambda}} \right\rfloor \right\}.$$

更进一步地, 得到最优功率与延时约束之间的闭式关系, 即最优能效延时折中.

定理2 在累积任务机制下, 使用最优休眠控制参数时, 最优长时平均功率 $\bar{P}_{opt}^{(N)}$ 与平均延时约束 τ 之间的闭式关系为:

(1) 在 $\tau \leq \tau_{th}$ 时,

$$\bar{P}_{opt}^{(N)} = P_{ID} + \rho P_{TR}/\Delta + 2\lambda(1 - \rho)(\tau - \bar{T}_0) \frac{\lambda h_S(P_{ST} - P_{ID}) - N^*(P_{ID} - P_{SL} - P_{DT})}{(N^*)^2 - N^*(1 - 2\lambda h_S) + \lambda^2 h_S^2(1 + c_S^2)};$$

(2) 在 $\tau \geq \tau_{th}$ 时,

$$\bar{P}_{opt}^{(N)} = \rho(P_{ID} + P_{TR}/\Delta) + (1 - \rho)(P_{SL} + P_{DT}) + \frac{(1 - \rho)(P_{ST} - P_{SL} - P_{DT})h_S}{(\tau - \bar{T}_0) + \frac{1}{2\lambda} + \sqrt{\left(\tau - \bar{T}_0 + \frac{1}{2\lambda}\right)^2 - h_S^2 c_S^2 - \frac{h_S}{\lambda}}}.$$

即延时约束小于阈值时, 最优关系为线性; 大于阈值时, 为单调减的凸函数.

证明 参见文献 [24].

图 5 给出了不同唤醒机制下, 最优平均功率与延时约束之间的关系, 即最优能效延时折中. 从中也可以看到探测能耗的影响: 无探测能耗时, 累积任务机制性能优于多重休假机制, 而多重休假机制又优于单重休假机制, 然而前两者的优势随探测能耗的增加迅速减小.

5 业务突发性对基站休眠与功率匹配机制的影响

考虑单业务基站下行场景, 将首先介绍用于业务建模的典型突发性随机过程. 假定超蜂窝网络中业务基站休眠采用累积任务唤醒机制, 即系统空闲时基站进入休眠状态, 直至累积 N 个用户再次进入

图 5 (网络版彩图) 最优控制下归一化最优平均功率 (依无休眠时的平均功率归一化) 与平均延时约束之间的关系. $c_S^2 = 0$, $h_S = 0.25$ s

Figure 5 (Color online) Optimal power consumption (normalized by the case without sleep mode) vs. average delay constraint for different schemes. $c_S^2 = 0$, $h_S = 0.25$ s

活跃状态. 本节及之后所有章节将迟滞时间和启动时间均简化为 0, 而用 E_{SW} 刻画状态转换开销: 每次在活跃状态和休眠状态之间互相切换将带来 E_{SW} 的能耗开销. 功率匹配指活跃状态下基站传输功率 P_{TR} 将随着业务量的变化动态调整. 休眠门限 N 和基站的传输功率 P_{TR} 为目标机制的控制变量. 利用排队理论, 分析了业务突发性以及自相关性对基站休眠与功率匹配机制, 系统功率与平均延时的影响; 并求解延时受限的功率最小化问题, 得到最优的功率延时折中关系. 另外, 这里不再引入迟滞时间, 相当于迟滞时间恒为 0, 即可看作延时约束大于阈值时的情况. 另外, 此时启动时间简化为 0, 而用 E_{SW} 刻画状态转换开销: 每次在活跃状态和休眠状态之间互相切换将带来 E_{SW} 的能耗开销.

5.1 突发性业务到达模型

间歇 Poisson 过程. 如图 6 所示, 用户的到达服从参数为 (λ, r_1, r_2) 的间歇 Poisson 过程 (简称为 IPP). IPP 为一连续时间随机过程, 分为开启与关闭两个阶段, 各自的时长分别服从均值为 r_1^{-1} 和 r_2^{-1} 的指数分布. 在开启阶段, 用户的到达服从强度为 λ 的 Poisson 过程, 而关闭阶段无用户到达. 假定随着每位用户的到达, 系统将接收到一个随机大小的服务请求, 该请求服从平均长度为 l 的指数分布²⁾. 当该请求被完成后, 对应用户即离开系统.

Markov 调制 Poisson 过程. 上述 IPP 仍属于更新过程, 它无法描述业务的自相关性. 因此, 进一步考虑具有正相关的非更新过程, 以关注业务自相关性对系统的影响. Markov 调制 Poisson 过程 (MMPP) 是其中的典型代表, 主要描述业务到达的短时正相关特性, 这里以二相位 Markov 调制 Poisson 过程为例展开研究, 又称交互 Poisson 过程 (switched Poisson process, SPP). 图 7 展示了 SPP 业务模型, 该模型分为两个强度不同的阶段, 用户到达分别服从强度为 λ_1 和 λ_2 的 Poisson 过程, 停留时间分别服从均值为 r_1^{-1} 和 r_2^{-1} 的指数分布.

方差系数和自相关系数是描述随机事件的重要参数, 前者可衡量随机事件抖动 (偏离平均值程度), 后者可衡量随机事件之间关联性^[31]. 本节将主要分析二者对基站总功率以及用户延时的影响.

2) 这里主要考虑延时容忍业务, 例如平均长度为 l 的文件下载.

图 6 (网络版彩图) 间歇 Poisson 过程业务到达示意图

Figure 6 (Color online) Interrupted Poisson process (IPP)

图 7 (网络版彩图) 交互 Poisson 过程业务到达示意图

Figure 7 (Color online) Switched Poisson process (SPP)

图 8 速率可控的扩展 IPP/M/1 队列模型的状态转移图

Figure 8 State transition diagram for the extended IPP/M/1 queueing model with the N -based BS sleeping mechanism and power matching

5.2 IPP 到达下基于累积任务唤醒机制的基站休眠与功率匹配机制

根据第 2 节中给出的排队模型, 用户离开系统的速率可表示为 $\mu = \frac{1}{h_B}$, 基站的利用率表示为 $\rho = \lambda/\mu = \lambda h_B$.

IPP 业务到达模型下的队列状态转移图如图 8 所示. 状态空间被分为活跃状态与休眠状态两个子集. 系统状态由 IPP 到达过程状态以及系统中的总用户数组成. 利用排队论、生成函数等数学工具, 能够给出系统总功率与平均延时的显式表达式, 推导过程详见文献 [25, 26].

系统长时平均功率 (亦反映系统总功耗大小) $E[P]$ 为

$$E[P] = (1-p_s) \left(P_{ID} + \frac{1}{\eta} P_{TR} \right) + p_s P_{SL} + E_{SW} F_m, \quad (13)$$

其中 p_s 表示业务基站处于休眠状态的概率, 由下式给出:

$$p_s = 1 - \frac{\hat{\lambda}}{\mu} = \frac{\mu(r_1 + r_2) - \lambda r_2}{\mu(r_1 + r_2)}. \quad (14)$$

易知 p_s 结果与休眠门限 N 相互独立. 平均功率由 3 部分组成, 分别对应基站处于活跃状态时的平均功率、基站处于休眠状态时的平均功率消耗以及单位时间内基站状态切换能耗 $E_{SW} F_m$. 这里状态切换频率 F_m 表示单位时间内基站在活跃状态与休眠状态之间的切换次数.

利用 Little 定理得到系统的平均延时为

$$E[T] = \frac{\lambda + r_1 + r_2 - \mu}{\mu r_1 + \mu r_2 - \lambda r_2} + \frac{1}{\lambda r_2} \frac{N\lambda + \frac{N(N-1)}{2}(r_1 + r_2) \left(1 + \frac{r_1}{r_2} \right)}{N \left(1 + \frac{r_1}{r_2} \right) - \left[\left(\frac{\lambda}{r_2} + \frac{\lambda}{\mu} \right) z_0 - \frac{\lambda}{r_2} \right] \frac{1 - z_0^N}{1 - z_0}}, \quad (15)$$

其中,

$$z_0 = \frac{(\lambda + \mu + r_1 + r_2) - \sqrt{(\lambda + \mu + r_1 + r_2)^2 - 4\lambda(\mu + r_2)}}{2\lambda(1 + r_2/\mu)}. \quad (16)$$

图 9 (网络版彩图) 总功耗随传输功率 P_{TR} 的变化情况. $\lambda = 5$, $l = 2$ MB, $k = 1$, $N = 3$

Figure 9 (Color online) Total power consumption varying with transmit power P_{TR} . $\lambda = 5$, $l = 2$ MB, $k = 1$, $N = 3$

图 10 (网络版彩图) 不同业务突发性条件下的平均延时性能. $\lambda = 5$, $l = 2$ MB, $k = 1$, $P_{TR} = 10$ W

Figure 10 (Color online) Average delay varying with traffic burstiness. $\lambda = 5$, $l = 2$ MB, $k = 1$, $P_{TR} = 10$ W

将以上分析的基站休眠与功率匹配联合机制与仅考虑功率匹配的机制进行对比, 可分析基站休眠对系统性能的影响, 进而得出以下定理.

定理3 对于参数为 (λ, kr_2, r_2) 的 IPP 到达模型, 当传输功率给定时, 引入累积任务唤醒机制 (休眠门限 N) 能够带来节能增益的条件是

$$N + f(N, \mu, \lambda, k, r_2) > \frac{2\lambda E_{sw}}{(1+k)(P_o - P_{sleep})}, \quad (17)$$

其中 $f(N, \mu, \lambda, k, r_2) = -\frac{\lambda}{1+k}[(\frac{1}{r_2} + \frac{1}{\mu})z_0 - \frac{1}{r_2}]^{\frac{1-z_0^N}{1-z_0}}$, 并且具有以下性质:

$$(a) f(N, \mu, \lambda, k, r_2) > 0; \quad (b) \frac{\partial f(N, \mu, \lambda, k, r_2)}{\partial r_2} < 0; \quad (c) \frac{\partial f(N, \mu, \lambda, k, r_2)}{\partial N} > 0. \quad (18)$$

证明 参见文献 [26].

可发现业务的平均到达率给定时, 业务的突发性越大, 能够带来节能增益的休眠门限 N 的取值范围越大, 即可休眠的平均时间更长, 节能空间更大, 从而进一步降低能耗. 此外, 在传输功率 P_{TR} 恒定的条件下, 系统总功耗随休眠门限 N 的增加而减小, 而平均延时随 N 的增加而增大. 而在恒定休眠门限 N 时, 平均延时随传输功率 P_{TR} 的增加而减小. 总功耗随传输功率的变化情况如图 9 所示, 随着传输功率的增大, 系统总功耗先减小后增加, 存在令总功耗最小的最优传输功率点. 考虑业务突发性的影响, 从图 9 中可见, 当控制变量 (N, P_{TR}) 和业务平均到达率给定时, 业务的突发性越大, 系统总功耗越低. 对此的直观解释如下: 当业务突发性较大时, 其到达较为集中, 从而为基站提供更多的休眠机会.

图 10 给出了业务突发性对系统平均延时的影响, 发现只有 N 的取值相对较大时, 平均延时随业务突发性的提高而降低, 而 N 相对较小时, 情况相反. 这是因为当 N 较小时, 基站的唤醒门限较低, 此时, 如果业务的突发性越大, 则处于活跃状态的基站越容易累积更多的用户, 从而造成比较大的延时;

图 11 (网络版彩图) 对于 IPP 业务, 不同到达率下的最优休眠门限 N^* 以及传输功率 P_{TR}^* . $\tau = 2 \text{ s}$, $k = 1$, $l = 2 \text{ MB}$

Figure 11 (Color online) Optimal sleeping threshold N^* and transmit power P_{TR}^* under IPP traffic with different arrival rate. $\tau = 2 \text{ s}$, $k = 1$, $l = 2 \text{ MB}$

当 N 较大时, 如果业务的突发性越大, 休眠状态的基站反而更容易累积到门限 N , 以尽早唤醒基站开始服务, 故用户的延时性能会越好.

在以上分析的基础上, 求解优化问题 (10), 即平均延时约束下使得平均功率最小化的最优休眠门限和传输功率, 其中优化变量为 N 和 P_{TR} . 得到以下结果, 其证明参见文献 [26].

定理4 对于参数为 (λ, kr_2, r_2) 的 IPP 到达模型, 给定传输功率 P_{TR} , 在有延时约束的总功耗最小化问题 (10) 中, 它的最优休眠门限 N 是以下方程的唯一解:

$$N(N - E) = A(1 - z_0^N), \quad (19)$$

并具有上、下界约束

$$\max\{0, E\} < N < \frac{E + \sqrt{E^2 + 4A}}{2}, \quad (20)$$

其中参数 $A > 0$ 和 E 分别为

$$A = \frac{-2\lambda^2}{(1+k)^2(1-z_0)} \left[\left(\frac{1}{r_2} + \frac{1}{\mu} \right) z_0 - \frac{1}{r_2} \right] \left[\tau - \frac{kr_2 + r_2 + \lambda - \mu}{\mu r_2(1+k) - \lambda r_2} \right], \quad (21)$$

$$E = 1 + \frac{2\lambda}{1+k} \left[\tau - \frac{k\lambda + r_2(1+k)^2}{r_2(1+k)(k\mu + \mu - \lambda)} \right]. \quad (22)$$

图 11 给出了不同到达率下的最优休眠门限和传输功率的取值. 首先, 可以看到随着业务突发性的增加 (r_2 减小), 休眠机制的可行到达率取值范围变小, 即突发性较大时, 只有系统利用率足够低才能通过休眠机制带来节能增益. 其次, 业务到达率增加时, 最优的休眠门限取值先增加后减小, 而对于传输功率, 可以看到它在可行范围的后半部分急剧上升, 可知功率匹配主要在业务到达率相对较高时发挥作用.

图 12 (网络版彩图) SPP 到达业务模型的总功耗与平均延时 vs. 系统利用率 ρ 以及自相关系数 θ , $C^2 = 10$. $r_1 = r_2$, $l = 2$ MB, $N = 1$, $P_{TR} = 10$ W

Figure 12 (Color online) Total power consumption and average delay vs. system utilization ρ and autocorrelation coefficient θ , with SPP traffic $C^2 = 10$. $r_1 = r_2$, $l = 2$ MB, $N = 1$, $P_{TR} = 10$ W. (a) Total power consumption; (b) average delay

5.3 业务相关性对基站休眠与功率匹配机制的影响

将突发性业务模型扩展到非更新过程, 以 SPP 为例研究业务相关性对系统性能的影响. 此时, 业务基站休眠与功率匹配机制可以建模为速率可控的扩展 SPP/M/1 队列. 为了得到总功耗与平均延时性能, 仍然采用生成函数的方法, 但是 SPP 到达模型下无法给出显式表达式, 而是需要通过数值计算的方法得到结果进行分析.

通过数值结果发现, SPP 到达的业务模型下总功耗与平均延时随两个控制变量的变化情况与采用 IPP 到达的业务模型时影响相同. 平均延时随休眠门限 N 的增加或者传输功率 P_{TR} 的减小而增大. 总功耗随 N 的增加而减少; 当 N 较小时, 总功耗随传输功率 P_{TR} 单调增加, 而当 N 较大时, 总功耗随 P_{TR} 的增大先减小后增加.

同样的, 通过数值结果研究方差系数 C^2 以及自相关系数 θ 对系统性能的影响, 发现当业务的平均到达率和自相关系数给定时, 业务的突发性越大, 系统总功耗也就越低; 而只有在特定的情况下, C^2 越大, 平均延时才越小, 这与 IPP 业务到达模型下业务突发性对系统性能的影响是一致的.

当方差系数 C^2 给定时, 图 12(a) 中给出 SPP 业务模型下总功耗随系统利用率和自相关系数 θ 的变化情况. 可以发现对于不同的自相关系数 θ , 总功耗基本保持不变. 与来自方差系数的影响相比, 可以得出结论: 业务的自相关性对系统总功耗基本没有影响. 图 12(b) 展示了自相关系数 θ 对平均延时性能的影响, 发现只有系统利用率较高时, 较大的自相关系数 θ 才会导致系统延时性能变差.

6 业务突发到达下基于部分可观测模型的最优基站休眠策略

第 5 节中休眠策略采用了累积任务唤醒机制假设, 是否存在优于该机制的休眠策略成为了进一步研究的关注点. 以下考虑离散时间系统, 以间歇 Bernoulli 过程 (IBP) 为例, 利用 POMDP 对系统进行了建模, 并求解最优单基站休眠策略, 对其结构进行了分析^[27]. IBP 可视为离散化的间歇 Poisson 过程, 在 ON 状态时段业务到达服从几何分布. 此问题中到达过程状态不可观测, 只能通过经验推断.

图 13 (网络版彩图) 间歇 Bernoulli 过程下, 不同基站状态下的最优策略. $E_{\text{active}} = 25 \text{ J}$, $\omega = 10$, $\rho = 0.2$
Figure 13 (Color online) Optimal BS sleeping mechanism with varying BS current mode under IBP traffic. $E_{\text{active}} = 25 \text{ J}$, $\omega = 10$, $\rho = 0.2$. (a) $E_{\text{SW}} = 417 \text{ J}$, arrival state is OFF; (b) $E_{\text{SW}} = 417 \text{ J}$, arrival state is ON

其中, 系统状态可表示为 $i = (t, Q, \delta)$, t 表示当前时隙与上一次业务到达之间的时间间隔, 用来估计未知的到达状态 (活跃或关闭), Q 和 δ 分别表示系统中业务队长以及当前基站的开关状态. 在每一时隙, 假设至多只有一个业务到达或离去, 系统将更新当前状态, 并且基站会根据当前状态做出开关决策, 以决定下一时隙是否休眠. 根据系统的当前状态以及所采取的决策, 系统将付出相应的开销, 为系统延时与能耗的加权和, 并考虑基站的状态转换开销, 表示为

$$C(i, A_S(i)) = |\delta - A_S(i)|E_{\text{SW}} + \omega Q + A_S(i)E_{\text{active}}, \quad (23)$$

其中 $A_S(i)$ 表示在特定 S 休眠策略下, 系统处于状态 i 时作出的决策. 系统开销由 3 部分组成, $|\delta - A_S(i)|E_{\text{SW}}$ 表示基站开关状态转换的能耗开销, ωQ 表示业务等待的延时开销, 其中 ω 表示能耗与延时开销之间的权重参数. $A_S(i)E_{\text{active}}$ 则表示下一时隙的基站运行能耗开销. 最终优化目标是将长时的折现系统总开销最小化, 并利用策略迭代法求解最优策略的数值结果.

考虑到到达状态已知的特殊情况, 图 13 中给出了一组系统参数下, 不同基站状态下的最优策略. 观察可知, 最优基站休眠策略具备双门限结构, 即当到达状态一定时, 存在双门限 Q_{active} (唤醒门限) 和 Q_{sleep} (休眠门限), 当 $Q \leq Q_{\text{sleep}}$ 时, 基站将选择休眠; 当 $Q \geq Q_{\text{active}}$ 时, 基站将选择开启; 当 Q 处于 Q_{sleep} 和 Q_{active} 之间时, 基站将选择保持现有状态不变.

图 14 中给出了不同能耗延时权重参数 ω 下, 系统总开销随业务突发性的变化情况. 可以发现, 随着业务突发性的增大, 系统总开销将减小. 此结论与第 5 节中在累积任务机制假设下的研究结论一致. 此外, 随着权重参数 ω 的增大, 业务的延时开销占比加重, 系统的总开销增大.

7 结论

本文针对超蜂窝网络中的业务基站, 考虑能效与延时的理论关系设计最优休眠机制, 完成了以下工作.

量化分析基于业务基站休眠的能效延时折中问题. 考虑迟滞休眠机制和单重休假、多重休假及累积任务 3 种典型唤醒机制, 利用休假排队模型得到系统长时平均功率、系统平均延时的闭式表达. 分析了各系统参数对能效和延时的影响. 指出大部分情况下, 较小的启动时间均值和离散系数都使延时

图 14 (网络版彩图) 不同能耗延时权重参数 ω 下, 系统总开销随业务突发性的变化情况. $E_{\text{sw}} = 41.7 \text{ J}$, $E_{\text{active}} = 25 \text{ J}$, $\rho = 0.2$

Figure 14 (Color online) System cost with varying traffic burstiness under different weight ω . $E_{\text{sw}} = 41.7 \text{ J}$, $E_{\text{active}} = 25 \text{ J}$, $\rho = 0.2$

与能效性能同时提升, 即存在能效延时非折中区间, 对业务基站设备制造提供指导. 同时指出, 能效延时随迟滞时间变化呈线性的折中关系. 在累积任务机制的非折中区间中, 给出能效延时双赢的参数选择方法.

应用折中关系设计延时约束下能效最优的业务基站休眠策略, 并衡量所得的最优能效延时折中. 指出在单重休假、多重休假及累积任务唤醒机制下, 最优基站休眠策略参数组具有两段式的特点: 当延时约束小于一定阈值时, 仅最优迟滞时间随延时约束增长单调下降; 当延时约束大于该阈值时, 启动时间的离散系数较小时, 仅最优休假时间或最优累积任务数随延时约束单调增长. 根据此特点得到最优控制参数, 并对最优长时平均功率与平均延时约束进行闭式分析, 指出当延时约束小于一定阈值时, 一定量的延时容忍换取的能效提升保持不变. 而当延时约束大于该阈值时, 一定量的延时容忍换取的能效提升随延时增长逐渐下降.

考虑超蜂窝网络中突发业务到达模型下的业务基站休眠问题, 分别以间歇 Poisson 过程和 Markov 调制 Poisson 过程为例对突发业务进行建模, 考虑累积任务唤醒机制, 分析了业务突发性对系统总功耗以及延时性能的影响. 分析得到, 在给定休眠参数和传输功率时, 业务突发性越大, 系统总功耗越低, 而只有在特定情况下, 突发性越大, 延时性能才越好; 业务相关性对系统总功耗的影响并不明显, 延时性能也只在系统利用率较高时受到影响. 利用部分可观测的 Markov 决策过程对业务基站休眠机制进行建模, 发现最优基站休眠策略存在双门限结构, 且随着业务突发性的增大, 系统性能将提升. 以上结论说明了在引入基站休眠的系统中, 业务的突发性可能提升网络性能.

参考文献

- 1 Niu Z S, Zhou S, Zhou S D, et al. Energy efficiency and resource optimized hyper-cellular mobile communication system architecture and its technical challenges. *Sci Sin Inform*, 2012, 42: 1191–1203 [牛志升, 周盛, 周世东, 等. 能效与资源优化的超蜂窝移动通信系统新架构及其技术挑战. *中国科学: 信息科学*, 2012, 42: 1191–1203]
- 2 Auer G, Blume O, Giannini V, et al. How much energy is needed to run a wireless network? *IEEE Wirel Commun*, 2011, 18: 40–49
- 3 Elayoubi S E, Saker L, Chahed T. Optimal control for base station sleep mode in energy efficient radio access networks. In: *Proceedings of the IEEE International Conference on Computer Communications, Shanghai*, 2011. 106–110

- 4 Gong J, Zhou S, Niu Z. A dynamic programming approach for base station sleeping in cellular networks. *IEICE Trans Commun*, 2012, 95: 551–562
- 5 Oh E, Krishnamachari B. Energy savings through dynamic base station switching in cellular wireless access networks. In: *Proceedings of the Global Telecommunications Conference*, Miami, 2010. 1–5
- 6 Son K, Kim H, Yi Y, et al. Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. *IEEE J Sel Area Commun*, 2011, 29: 1525–1536
- 7 Gupta R, Strinati E C. Base-station duty-cycling and traffic buffering as a means to achieve green communications. In: *Proceedings of the Vehicular Technology Conference*, Quebec City, 2012. 1–6
- 8 Ashraf I, Boccardi F, Ho L. Power savings in small cell deployments via sleep mode techniques. In: *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications Workshops*, Istanbul, 2010. 307–311
- 9 Berry R A, Gallager R G. Communication over fading channels with delay constraints. *IEEE Trans Inf Theory*, 2002, 48: 1135–1149
- 10 Berry R A. Optimal power-delay tradeoffs in fading channels-small-delay asymptotics. *IEEE Trans Inf Theory*, 2013, 59: 3939–3952
- 11 Neely M J. Optimal energy and delay tradeoffs for multi-user wireless downlinks. In: *Proceedings of the IEEE International Conference on Computer Communications*, Barcelona, 2006. 1–13
- 12 Neely M J. Optimal energy and delay tradeoffs for multiuser wireless downlinks. *IEEE Trans Inf Theory*, 2007, 53: 3095–3113
- 13 Uysal-Biyikoglu E, Prabhakar B, Gamal A E. Energy-efficient packet transmission over a wireless link. *IEEE/ACM Trans Netw*, 2002, 10: 487–499
- 14 Meshkati F, Poor H V, Schwartz S C. Energy efficiency-delay tradeoffs in CDMA networks: a game-theoretic approach. *IEEE Trans Inf Theory*, 2009, 55: 3220–3228
- 15 Petreska N, Al-Zubaidy H, Gross J. Power minimization for industrial wireless networks under statistical delay constraints. In: *Proceedings of the Teletraffic Congress*, Karlskrona, 2014. 1–9
- 16 Zhang Y, Rvidsson A. Understanding the characteristics of cellular data traffic. *ACM Sigcomm Comput Commun Rev*, 2012, 42: 461–466
- 17 Niu Z. TANGO: traffic-aware network planning and green operation. *IEEE Wirel Commun*, 2011, 18: 25–29
- 18 Hasan Z, Boostanimehr H, Bhargava V K. Green cellular networks: a survey, some research issues and challenges. *IEEE Commun Surv Tut*, 2011, 13: 524–540
- 19 Tamura H, Yahiro Y, Fukuda Y, et al. Performance analysis of energy saving scheme with extra active period for LAN switches. In: *Proceedings of the Global Communications Conference*, Washington, 2007. 198–203
- 20 Venkatesh T, Murthy C S R. *An Analytical Approach to Optical Burst Switched Networks*. New York: Springer, 2010
- 21 Niu Z, Guo X, Zhou S, et al. Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control. *IEEE J Sel Area Commun*, 2015, 33: 641–650
- 22 Niu Z, Zhang J, Guo X, et al. On energy-delay tradeoff in base station sleep mode operation. In: *Proceedings of the IEEE International Conference on Communication Systems*, Singapore, 2012. 235–239
- 23 Guo X, Zhou S, Niu Z, et al. Optimal wake-up mechanism for single base station with sleep mode. In: *Proceedings of the Teletraffic Congress*, Shanghai, 2013. 1–8
- 24 Guo X, Niu Z, Zhou S, et al. Delay-constraint energy-optimal base station sleeping control. *IEEE J Sel Area Commun*, 2016, 34: 1073–1085
- 25 Wu J, Bao Y, Miao G, et al. Base station sleeping and power control for bursty traffic in cellular networks. In: *Proceedings of the IEEE International Conference on Communications Workshops*, Sydney, 2014. 837–841
- 26 Wu J, Bao Y, Miao G, et al. Base station sleeping control and power matching for energy-delay tradeoffs with bursty traffic. *IEEE Trans Veh Tech*, 2016, 65: 3657–3675
- 27 Leng B, Krishnamachari B, Guo X, et al. Optimal operation of a green server with bursty traffic. In: *Proceedings of the IEEE Global Communications Conference*, Washington, 2016. 1–6
- 28 Auer G, Blume O, Giannini V, et al. D2.3: Energy Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown. *INFOICT-247733 EARTH (Energy Aware Radio and NeTwork Technologies)*. 2010
- 29 3GPP. Further advancements for E-UTRA physical layer aspects. *TSG RAN TR 36.814 v9.0.0*, Mar. 2010.

<https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2493>

- 30 She C Y, Yang C Y. Energy efficiency and delay in wireless systems: is their relation always a tradeoff? *IEEE Trans Wirel Commun*, 2016, 15: 7215–7228
- 31 Niu Z S. *Fundamentals of Communication Networking Theory*. Beijing: Tsinghua University Press, 2006. 21–22 [牛志升. 通信网理论基础. 北京: 清华大学出版社. 2006. 21–22]

Energy-delay tradeoff and optimal base station sleeping control in hyper-cellular networks

Xueying GUO^{1,2}, Bingjie LENG^{1,2*}, Jian WU^{1,2} & Zhisheng NIU^{1,2}

1. *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China;*

2. *Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China*

* Corresponding author. E-mail: lengbj14@mails.tsinghua.edu.cn

Abstract By decoupling control coverage and traffic coverage, dynamic base station (BS) sleeping becomes possible in hyper-cellular networks. Traffic coverage can be adjusted based on real-time demands, to enhance the energy efficiency of the networks. However, sleeping mechanisms may affect the networks' delay performance. Thus, it is important to obtain analytical results on the influence of sleeping mechanisms on the system's energy efficiency and delay performance. The optimal sleeping mechanism can be derived based on the energy-delay tradeoff. This work summarizes the relevant contributions of our group: based on the vacation queueing model, closed-form expressions for average power and mean traffic delay were obtained. The conditions for the energy-delay tradeoff were analyzed. Further, the optimal BS sleeping design and the optimal energy-delay tradeoff were obtained, assuming fixed delays. The influence of the mode-changing energy cost was studied. BS sleeping designs for bursty traffic were analyzed, combined with power matching. The optimal BS sleeping mechanism for bursty traffic was obtained with a partially observable Markov decision process (POMDP).

Keywords hyper-cellular networks, vacation queue, base station sleeping, energy-delay tradeoff, bursty input traffic

Xueying GUO was born in 1988. She received her Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2017. Currently, she is a postdoctoral researcher in the Computer Science Department at the University of California, Davis, CA, USA. Her research interests include wireless networking, edge cloud system, and reinforcement learning.

Bingjie LENG was born in 1992. She received her Bachelor degree in engineering from Tsinghua University, Beijing, China, in 2014. Currently, she is a Master's student in engineering at Tsinghua University, Beijing, China. Her research interests include green wireless communications, infrastructure sharing, and mobile big data analysis.

Jian WU was born in 1988. She received her Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2015. She is currently a postdoctoral researcher at the University of California, Davis, CA, USA. Her research interests include green wireless communications, cloud computing, and vehicular networks.

Zhisheng NIU was born in 1964. He graduated from Beijing Jiaotong University, Beijing, China, in 1985, and obtained his M.E. and D.E. degrees from the Toyohashi University of Technology, Japan, in 1989 and 1992, respectively. Presently, he is a professor at the Department of Electronic Engineering, Tsinghua University, Beijing, China. He is also a guest chair professor at Shandong University, China. His major research interests include queueing theory, traffic engineering, mobile Internet, radio resource management of wireless networks, and green communication and networks.