中国科学: 信息科学 2017 年 第 47 卷 第 5 期: 607–619
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
基于超蜂窝架构的绿色通信专刊 · 论文

无线网络能效 – 服务质量的基本关系及应用

佘昌洋*, 杨晨阳*

北京航空航天大学电子信息工程学院, 北京 100191
* 通信作者. E-mail: cyshe@buaa.edu.cn, cyyang@buaa.edu.cn

收稿日期: 2017–02–15; 接受日期: 2017–03–20; 网络出版日期: 2017–05–04
国家重点基础研究发展计划 (973) (批准号: 2012CB316003) 和国家自然科学基金 (批准号: 61120106002) 资助项目

摘要 提升通信系统能效必须以满足各类业务的服务质量为前提, 因此研究能效与服务质量的关系是优化高能效无线网络的一个基本问题. 本文对过去 5 年来我们在超蜂窝网络架构下对能效与服务质量需求的基本关系及其在弹性接入机制优化方面的应用进行了总结. 由于延时界是服务质量的一个代表指标, 首先介绍在能效与延时界需求关系方面的研究进展. 研究发现, 如果基站的总功耗 (包括发射功率和电路功耗) 随着平均数据率线性增加, 那么能效 – 延时关系存在非折中区域. 而后, 概括如何针对不同类型的业务进行高能效资源分配, 包括传统实时和非实时业务, 以及超可靠低延时业务. 对于实时业务, 要达到最优的能效延时关系, 系统需要根据队列长度调整资源配置. 对于延时容忍性较强的非实时业务, 利用预测信息进行资源规划和提前推送成为可能. 结果表明, 利用移动用户的轨迹以及用户对内容的偏好都能大大提升系统能效. 对于服务质量需求极高的超可靠低时延业务, 初步探讨了在保证其服务质量的前提下使系统能效最大所需要的系统资源.

关键词 能效, 服务质量, 资源分配, 无线通信

1 引言

能量效率 (energy efficiency, EE) 已经成为第五代 (fifth generation, 5G) 移动通信系统的一个重要设计目标 [1], 在过去几年中, 关于能效的研究变得备受关注. 与另外一个关键设计目标 —— 频谱效率 (spectral efficiency, SE) 不同, 能效最优设计的核心问题不再是最大化系统能够传输的数据率, 而是根据用户的需要传输数据. 这意味着提供一个比服务质量 (quality of service, QoS) 需求更高的数据率是一种资源的浪费.

无线通信系统中的一大类业务是延时敏感的, 例如视频、音频和互动数据传输都要求较低的端到端延时. 已有文献中的延时性能指标可以分为确定延时界 [2]、平均延时 [3] 和统计 QoS 需求 [4]. 由于信道的衰落, 在无线通信系统中保证确定延时界所需要的发射功率太高, 因而在实际系统中难以满足

引用格式: 佘昌洋, 杨晨阳. 无线网络能效 – 服务质量的基本关系及应用. 中国科学: 信息科学, 2017, 47: 607–619, doi: 10.1360/N112017-00037
She C Y, Yang C Y. Energy efficiency-QoS relation and its application in wireless networks (in Chinese). Sci Sin Inform, 2017, 47: 607–619, doi: 10.1360/N112017-00037

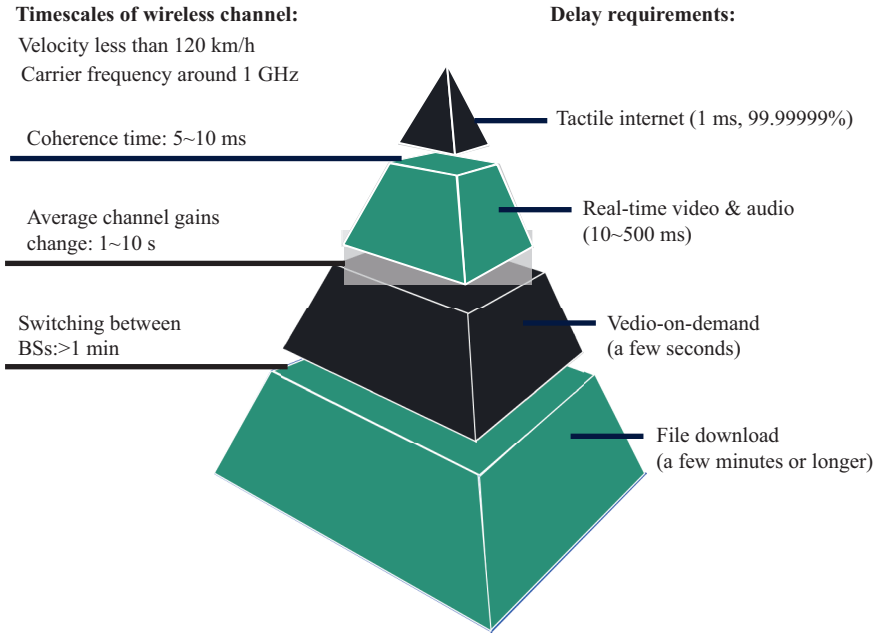

图 1 (网络版彩图) 无线信道变化的典型时间尺度与不同业务延时需求

Figure 1 (Color online) Typical timescales of wireless channel and delay requirements of different services

确定延时界需求. 对于多媒体应用, 平均延时无法确保业务所需要的延时性能, 因为如果一个数据包没有在给定延时界内传完, 则这个数据包会被丢掉. 由延时界和允许超过延时界的最大概率定义的统计 QoS 需求更适于反映无线多媒体业务的性能指标. 因此, 在第四代 (fourth generation, 4G) 移动通信系统中, 对网络语音业务 (voice over IP, VoIP) 规定了无线接入网络的延时需求为 50 ms 的延时界, 且允许超过延时界的最大概率为 2% [5].

除了延时敏感业务, 另一大类业务是内容分发等延时不太敏感 (如点播式视频传输), 甚至完全不敏感的业务 (如文件下载), 将之统称为非实时业务. 如文献 [6] 所示, 2016 年在无线网络中超过一半的业务量来自移动视频业务, 而且预计到 2021 年这个比例将达到 78%. 为了保障用户体验, 视频质量和播放中断是反映点播式视频业务用户体验的两个重要性能指标 [7]. 相对于延时敏感业务, 点播式视频业务的延时需求较为宽松, 因此可以在更长的时间内优化资源分配以提升能效.

除了传统的延时敏感与非实时业务, 在 5G 移动通信系统中将出现一类需要超短延时 (比如 1 ms 的端到端延时) 和超高可靠性 (比如 99.99999% 的包丢失概率) 的新业务. 超可靠低延时通信 (ultra-reliable and low-latency communications, URLLC) 的典型应用场景包括自动驾驶汽车、移动机器人、虚拟现实、增强现实和智能工厂等. 触感互联网能够让这一类业务成为可能 [8]. 在长期演进 (long term evolution, LTE) 系统中, 传输时间间隔被设置为 1 ms, 这意味着数据包被传输之前要在基站缓冲区中等待超过 1 ms, 因而在 LTE 系统中无法满足超短端到端延时需求. 达到如此严格的服务质量已经成为 5G 的主要设计目标之一 [9].

图 1 给出了无线信道时变的典型时间尺度和不同类型业务延时需求的关系. 信道相干时间是小尺度信道衰落变化的典型时间尺度. 对于中低速运动的用户, 在 2 GHz 频段的信道相干时间为 5 ∼ 10 ms. 平均信道增益取决于大尺度信道, 包括路径损耗和阴影衰落. 只有当用户位置明显发生变化时, 平均信道增益才会改变, 因而其随时间的变化速度远远低于小尺度信道衰落的变化速度, 一

般为秒级. 不同业务的延时需求也存在较大差异, 触感互联网业务的端到端延时一般为毫秒级, 而文件下载业务的延时则可以达到分钟甚至小时级.

未来无线通信网络不仅仅需要达到更高的能效和谱效, 还需要保证不同类型业务的 QoS 需求. QoS 需求不仅仅包括延时, 不同的业务还对可靠性有着不同的需求, 而延时和可靠性之间也存在着折中关系. 为了满足可能存在冲突的各种性能指标, 几个基本的折中关系需要进一步研究, 如能效 – 谱效 – QoS 需求之间的基本关系. 因为延时是一类典型的 QoS 需求, 能效 (或功率) 和延时的折中关系在过去十年以来受到了广泛关注.

自从 Berry 和 Gallagar 于 2002 年在 IEEE 信息论会刊上发表了影响深远的论文以来, 学术界已众所周知, 当信源或者信道随机时, 平均发射功率与平均延时需求之间是严格的折中关系. 因此, 至今已有的研究几乎都认为无论对哪种延时需求, 能效与延时需求之间也总是严格的折中关系. 实际上, 能效/功率 – 延时折中关系已被认为是无线通信的基本属性. 根据文献 [10] 中所揭示的结论, 功率 – 延时折中关系是广泛存在的, 这一结果无论延时指标是平均延时、严格延时界还是统计 QoS 需求都成立. 给定信道衰落的状态, "可靠" 传输 1 bit 数据所需要的功率与数据率是严格凸函数关系. 基于这一事实, 文献 [10] 中的研究表明, 平均发射功率和平均排队延时不能同时达到最小, 除非到达率和信道都是恒定的. 为了研究如何达到功率 – 延时的 Pareto 最优 (Pareto optimal) 关系, 该文献进一步指出了优化设计资源分配策略的准则. 对于到达过程或者服务过程不恒定的情形, 当延时趋于无穷时, Pareto 最优的功率 – 延时关系趋于最小功率极限. 如果采用一个与队列状态信息 (queue state information, QSI) 和信道状态信息 (channel state information, CSI) 相关的功率分配策略, 则当延时趋于无穷时, 平均功率趋于功率极限的速度要比采用一个只与 QSI 或 CSI 相关的功率分配策略快. 因此, 一个只依赖于 QSI 或 CSI 的资源分配策略不是 Pareto 最优的.

在针对不同类型的业务设计传输策略时, 信道模型和分析工具也各不相同. 传统的资源分配不仅难以满足各类业务的服务质量需求, 而且会导致较低的资源利用率, 因此必须采用跨层优化.

本文对过去 5 年来我们在超蜂窝网络架构下对能效与延时界需求的基本关系及其在高能效资源跨层优化方面的应用进行总结. 首先介绍在能效与延时界需求关系方面的研究进展, 而后总结针对实时、非实时业务以及超可靠低延时等不同类型业务得到高能效资源优化策略.

2 能效 – 延时界关系与跨层资源优化

2.1 能效 – 延时界需求关系与功率 – 数据率关系

(1) 统计 QoS 需求. 统计 QoS 需求的定义为 $(D_{\max}, \varepsilon_D)$, 其中 $D_{\max}$ 是延时界, ε_D 是业务允许的超过延时界的最大概率. 对于除了高可靠低延时业务以外的传统实时业务, 可以只考虑在基站的排队延时, 忽略编码和传输延时.

有效带宽和有效容量是在保证统计 QoS 需求的前提下分析和设计资源分配的有力工具. 当 Gärtner-Ellis 理论的假设满足时 (也就是下述极限存在 [11]), 随机到达过程 $\{a(t), t \geqslant 0\}$ 的有效带宽可以表示为 [11]

$$E_{\mathrm{B}}(\theta) = \lim_{t\to\infty} \frac{1}{\theta t} \ln \mathbb{E}\left[\mathrm{e}^{\theta \int_0^t a(\tau)\mathrm{d}\tau}\right], \tag{1}$$

其中 $\theta > 0$ 是 QoS 指数, 较大的 θ 意味着较严格的延时界. 随机服务过程 $\{s(t), t \geqslant 0\}$ 的有效容量可以表示为 [12]

$$E_{\mathrm{C}}(\theta)=-\lim_{t\to\infty}\frac{1}{\theta t}\ln\mathbb{E}\left[\mathrm{e}^{-\theta\int_0^t s(\tau)\mathrm{d}\tau}\right]. \tag{2}$$

根据文献 [11] 中的渐近分析, 对于平稳的到达和服务过程, 如果平均到达率 $\mathbb{E}[a(t)]$ 小于平均服务率 $\mathbb{E}[s(t)]$, 则队列长度超过一个给定门限 $Q_{\max}$ 的概率随着门限的增加指数衰减. 假设队列处于稳态, 并用 Q_∞ 表示稳态队列长度, 则 $\Pr\{Q_\infty>Q_{\max}\}\leqslant \mathrm{e}^{-\theta Q_{\max}}$ 成立的充分必要条件是 $E_{\mathrm{C}}(\theta)\geqslant E_{\mathrm{B}}(\theta)$. 满足约束 $(Q_{\max},\varepsilon_Q)$ 的 QoS 指标可以通过如下表达式得到 [12]:

$$\varepsilon_Q=\Pr\{Q_\infty>Q_{\max}\}\approx\xi\mathrm{e}^{-\theta Q_{\max}}, \tag{3}$$

其中 $\xi\triangleq\Pr\{Q_\infty>0\}$ 是队列非空概率. 根据大偏差理论, 当最大队列长度 $Q_{\max}$ 较长时, 上面的近似准确. 当 $Q_{\max}=E_{\mathrm{B}}(\theta)D_{\max}$ 且 $\varepsilon_Q=\varepsilon_D$ 时, $(D_{\max},\varepsilon_D)$ 和 $(Q_{\max},\varepsilon_Q)$ 对应于相同的 θ.

(2) 相关定义.

定义1 (能效 – 延时界关系)　能效 – 延时界关系是在满足用户 QoS 需求 θ 的前提下系统所能达到的最大能效, 用 $\eta_{\mathrm{EE}}^{\max}(\theta)$ 表示.

当只考虑发射功率时, 能效 – 延时关系退化为功率延时关系. 对于平均延时, 这一关系在已有文献中已经进行了大量研究 [3].

定义2 (能效极限)　能效极限的定义为 $\eta_{\mathrm{EE}}^{\lim}\triangleq\lim_{\theta\to0}\eta_{\mathrm{EE}}^{\max}(\theta)$, 与之相应的平均总功耗称为功耗极限, 记为 $P_{\lim}^{\mathrm{tot}}=\lim_{\theta\to0}\mathbb{E}_{Q^\infty,\boldsymbol{h}}\{P^{\mathrm{tot}}\}$. 给定允许超过延时界的最大概率 ε_D, 当 $D_{\max}\to\infty$ 时 (即 $\theta\to0$ 时), $E_{\mathrm{C}}(\theta)\geqslant E_{\mathrm{B}}(\theta)$ 退化为

$$\mathbb{E}_{\boldsymbol{h}}\{s(t)\}\geqslant\mathbb{E}\{a(t)\}, \tag{4}$$

其中 $\boldsymbol{h}$ 为信道状态向量. 根据上述定义, 能效极限是能效 – 延时界关系的上界. 当延时界趋于无穷时, 能效 – 延时界曲线趋于能效极限.

定义3 (功率 – 数据率关系)　功率数据率关系 $\mathbb{E}_{\boldsymbol{h}}\{P_{\min}^{\mathrm{tot}}\}\sim(\bar{s})$ 为支持平均服务率 $\mathbb{E}_{\boldsymbol{h}}\{s(t)\}=\bar{s}$ 所需的最小平均总功耗 (包括发射功耗和电路功耗).

为了分析能效 – 延时界的关系, 引入一个依赖于 QSI 的 “两状态策略”. 当队列长度大于零时, 设计资源分配策略满足 QoS 约束, 这个状态被称为 “ON” 状态. 当队列长度为零时, 为了避免服务空队列, 服务率为零, 这个状态被称为 “OFF” 状态.

(3) 能效 – 延时界需求关系与功率 – 数据率关系.　为了讨论能效 – 延时界需求关系的性质, 将功率 – 数据率关系分为线性和非线性区间.

(i) 线性区间.　在这个区间内, 支持平均服务率 $\bar{s}$ 所需要的最小平均总功耗与 $\bar{s}$ 之间是线性关系: $\mathbb{E}_{\boldsymbol{h}}\{P_{\min}^{\mathrm{tot}}\}=c_1\bar{s}+c_0$, 其中 c_1 和 c_0 是与具体系统有关的正常数.

为了使平均发射功率最小, 资源分配策略不能服务空队列. 这意味着要在满足延时需求的前提下最大化能效, 一个两状态策略必须满足如下条件:

$$\mathbb{E}_{Q^\infty,\boldsymbol{h}}\{s(t)\}=\mathbb{E}\{a(t)\}. \tag{5}$$

由式 (4) 和 (5), 可以得到能效极限为

$$\eta_{\mathrm{EE}}^{\lim}=\mathbb{E}\{a(t)\}/(c_1\mathbb{E}\{a(t)\}+c_0). \tag{6}$$

下面这个命题表明, 给定平均到达率 $\mathbb{E}\{a(t)\}$, 当功率数据率呈线性关系时, 能效与延时界需求之间不存在折中关系, 且当业务的延时界需求很大时达到的最高能效等于能效极限 [4].

定理1 在线性区间中, 能效与延时界需求 $D_{\max}$ 的关系不随 $D_{\max}$ 的值变化, 且 $\eta_{\mathrm{EE}}^{\max}(\theta)=\eta_{\mathrm{EE}}^{\lim}$ 可以由两状态策略达到.

下面为定理 1 的结果提供一个直观的解释. 由于两状态策略不服务空队列, 实际传输的数据率等于系统的服务能力 $s(t)$. 在线性区间内, $\mathbb{E}_{\boldsymbol{h}}\{P_{\min}^{\mathrm{tot}}\}=c_1\mathbb{E}_{\boldsymbol{h}}\{s(t)\}+c_0$. 因此, 在很短的时间 $\mathrm{d}t$ 内, 平均能耗和平均传输的数据量满足 $\mathbb{E}_{\boldsymbol{h}}\{P_{\min}^{\mathrm{tot}}\}\mathrm{d}t=c_1\mathbb{E}_{\boldsymbol{h}}\{s(t)\}\mathrm{d}t+c_0\mathrm{d}t$, 其中 $\mathbb{E}_{\boldsymbol{h}}\{s(t)\}$ 是随队列长度变化的随机变量. 此外, 在服务一个实时业务的总时间 T 内需要传输的总数据量由信源决定, 即 $\int_0^T\mathbb{E}_{\boldsymbol{h}}\{s(t)\}\mathrm{d}t=\int_0^T\{a(t)\}\mathrm{d}t$. 在线性区间内, 两状态策略所消耗的能量由需要传输的总数据量决定, 因而也由信源确定, 为

$$\int_0^T\mathbb{E}_{\boldsymbol{h}}\{P_{\min}^{\mathrm{tot}}\}\mathrm{d}t=\int_0^T c_1\mathbb{E}_{\boldsymbol{h}}\{s(t)\}\mathrm{d}t+c_0T=c_1\int_0^T\{a(t)\}\mathrm{d}t+c_0T. \tag{7}$$

尽管延时界越短, 在队列非空时所需要的平均数据率 $\mathbb{E}_{\boldsymbol{h}}\{s(t)\}$ 越高, 但这并不改变需要传输的总数据量和线性区内系统的总能耗, 因而也不改变系统的能效. 注意到式 (7) 中第 2 个等式交换了两个线性运算的顺序, 即积分与数据率到功率的线性映射. 而如果功率和数据率不是线性关系, 则不能保证交换运算顺序后等式依然成立. 例如, 如果功率是数据率的严格凸函数, 则根据 Jansen 不等式, 交换顺序后这个等号将变为大于等于号.

下面的推论表明, 如果用平均延时作为延时指标, 则在线性区间内能效也与延时需求的大小无关.

推论1 在线性区间中, 能效不依赖于平均延时需求 $\bar{D}$.

(ii) 非线性区间. 在这个区间, $\mathbb{E}_{\boldsymbol{h}}\{P_{\min}^{\mathrm{tot}}\}$ 关于 $\bar{s}$ 是严格凸函数. 对于这种情形, 现有文献已经分别在延时很大和很小的渐进场景下对平均发射功率和平均延时的关系进行了研究 [13,14], 其结论表明, 如果到达过程或者服务过程是随机的, 则存在能效和延时的折中关系; 且要达到 Pareto 最优的能效 – 延时关系, 资源分配策略需要同时考虑队列状态信息和信道状态信息.

(4) 能效 – 延时界 – 资源的折中关系. 以多天线系统为例, 文献 [4] 研究了何时会出现线性区和非线性区, 并发现了能效 – 延时 – 资源的折中关系. 为了更好地揭示上述关系, 该文献以数据包的到达过程为复合 Poisson 过程 (其包到达时间间隔和数据包的大小分别符合参数为 λ^a 和 λ^u 的指数分布) 和大规模多天线系统为例, 得到了能效 – 延时关系的折中区和非折中区边界的闭式表达. 所得出的延时界需求边界和 QoS 指标边界分别如下:

$$D_{\mathrm{th}}^{\max}=\frac{(\lambda^u-\theta^{\mathrm{th}})\ln(1/\varepsilon_D)}{\lambda^a\theta^{\mathrm{th}}},\quad \theta^{\mathrm{th}}=\lambda^u-\frac{\lambda^a}{W^{\max}\log_2\left(1+\frac{\mu N_T}{N_0}\bar{P}_{\mathrm{TW}}^*\right)}, \tag{8}$$

其中, $W^{\max}$ 是分配给一个用户的最大带宽, μ 是该用户的平均信道增益, N_T 是基站的发射天线数, N_0 是单边噪声谱密度, $\bar{P}_{\mathrm{TW}}^*$ 是经过优化后得到的的单位带宽发射功率.

(i) 资源与延时界的折中关系. 当业务的延时界需求大于 $D_{\mathrm{th}}^{\max}$ 时, 达到这一延时界所需的数据率较低. 此时, 所需的带宽小于系统最大带宽. 如图 2[15] 所示, 当最大带宽约束不起作用时, 功率 – 数据率关系处于线性区间. 同时, 图 3[15] 中的能效 – 延时界关系处于非折中区 (即饱和区). 通过增加 $W^{\max}$ 和/或 N_t, 可以在保证系统能效不降低的前提下, 支持延时界需求更短的业务, 使其统计 QoS 需求得到满足. 因此, 在能效 – 延时界关系的非折中区域, 保证 QoS 所需的资源 (即带宽与天线数) 与延时界存在折中关系, 这也意味着谱效与延时界存在折中关系.

(ii) 能效与延时界的折中关系. 当业务的延时界小于 $D_{\mathrm{th}}^{\max}$ 时, 达到这一延时界所需的数据率较高. 此时, 所需的带宽等于系统最大带宽. 如图 2 所示, 当最大带宽约束起作用时, 系统只能通过增加

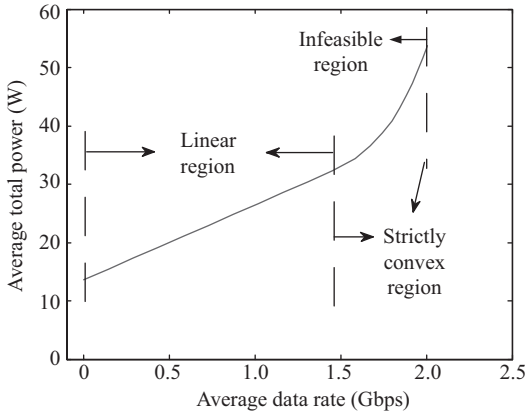

图 2　功率 – 数据率关系 [15]

Figure 2　Power-rate relation [15]

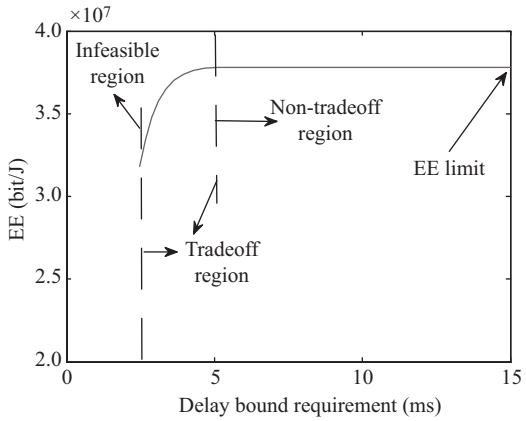

图 3　能效 – 延时界关系 [15]

Figure 3　EE-delay relation [15]

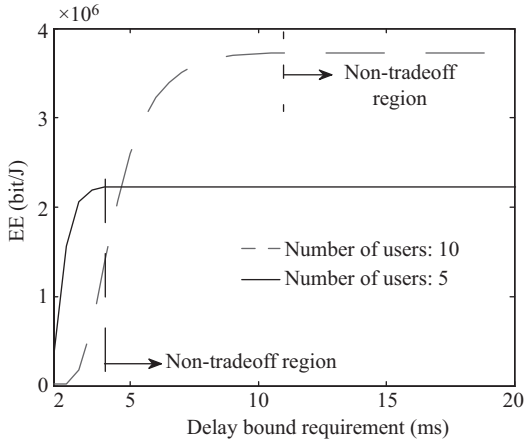

图 4　用户数对能效 – 延时界关系的影响 [4]

Figure 4　EE-delay relation with different number of users [4]

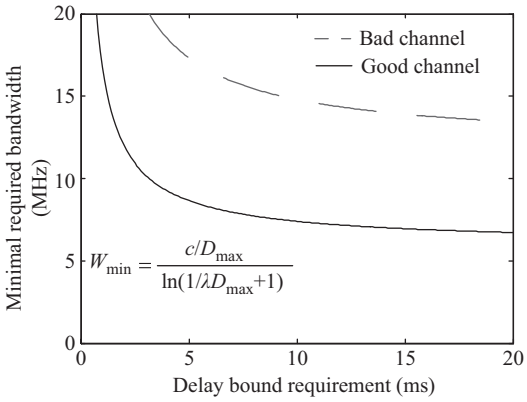

图 5　信噪比对所需带宽的影响

Figure 5　Required minimal bandwidth with different SNR

发射功率进一步提升数据率. 根据容量公式, 此时的功率 – 数据率关系处于严格凸区域. 由于功率 – 数据率关系是严格凸的, 图 3 中的能效 – 延时界关系处于折中区. 由于无法通过增加带宽来满足更短延时界需求, 因而出现了能效 – 延时界的折中关系.

(iii) 用户数与信噪比 (signal-to-noise ratio, SNR) 对能效 – 延时界 – 资源关系的影响.　当用户数增加时, 由于系统总带宽是固定的, 因而分给每个用户的最大带宽变小. 由式 (8) 可以看出, 随着 $W^{\max}$ 的减小, $D_{\text{th}}^{\max}$ 变长. 这意味着非折中区变小.

另一方面, 信道条件也会影响保障服务质量所需的资源. 图 4 给出了在不同信噪比下满足延时界且达到能效极限所需的最小带宽. 从图 5 可以看出, 满足延时界所需的最小带宽大致反比于延时界. 这意味着要保证能效不变, 当延时界减小时, 系统所需带宽迅速达到最大带宽. 由式 (8) 可以看出, 给定 $W^{\max}$, 当信道较好, 即 μ 较大时, $D_{\text{th}}^{\max}$ 较小. 因此, 信道好时, 能效 – 延时界关系的非折中区更宽.

(5) 能效 – 延时界基本关系的适用范围与实际意义.　上述能效 – 延时界需求的基本关系是在分

析传统实时业务时得到的, 在分析过程中隐含地假设用户的平均信道增益不变, 且基站没有休眠. 因此, 这个关系能够直接指导延时界需求小于平均信道发生变化时间的多媒体、语音和视频会议等传统实时业务的跨层资源分配. 研究表明, 当基站可以休眠时, 功率 – 数据率关系还包括一个既非线性也非严格凸的区域, 而能效 – 延时界关系依然包括一个折中区域和非折中 (即饱和) 区域. 不过, 此时当延时界需求很大时, 系统的最高能效小于能效极限.

对于图 1 中的非实时业务, 也可以采用有效容量和有效带宽分析在基站的排队延时. 不过, 由于其延时需求较长, 移动用户的平均信道增益会随着用户位置的改变而发生变化. 因此, 计算有效容量的方法会与实时业务有所不同. 另外, 上述分析表明, 当信道条件好时, 系统满足延时界所需的带宽较小. 对于移动用户, 由于非实时业务的平均信道增益会发生明显的变化, 这意味着利用信道分集 (即在平均信道好的时候多传) 节约带宽成为可能. 如果可以预知用户的请求 (如用户提前定制某业务) 并可以预测用户未来的平均信道增益, 则通过提前推送可以进一步增加业务对延时的容忍, 从而大大提升系统能效和节约带宽 [16].

对于图 1 中的超可靠低延时业务, 排队延时、传输延时和编码延时都不可忽略. 由于在这种业务的典型应用场景中延时界需求小于信道的相干时间, 在给定资源配置时, 服务过程是常数. 因此, 对于此类业务, 有效容量退化为恒定服务率, 依然可以利用有效带宽计算最小恒定服务率需求, 进而分析排队延时. 不过, 由于传输短包的延时极短, 不能再用 Shannon 容量公式反映有限码长编码的最大可达数据率, 能效 – 延时关系是否还存非折中区有待进一步研究.

2.2 跨层资源优化

有效带宽和有效容量是在统计 QoS 需求下进行跨层资源优化的有用工具. 有效容量这一概念来源于信息论, 它刻画了一个无线系统在保证统计 QoS 约束时能够服务的最大恒定信源到达率. 与 Shannon 容量不同, 有效容量可以让人们用一个统一的方法分析各种延时敏感业务的不同延时需求. 已有利用有效容量进行高谱效和高能效设计的文献均未考虑弹性接入, 都假设数据到达基站缓冲区的数据率恒定且始终等于有效容量. 在实际系统中, 业务的到达很难是恒定速率的. 实际上, 业务波动既包括大尺度波动 (平均到达率在不同时段变化) 也包括小尺度波动 (瞬时到达率的随机变化). 当信源到达率或者信道随机时, 资源分配策略需要同时考虑 CSI 和 QSI. 然而, 如何将 QSI 与有效容量的框架结合依然是一个难题.

与队列状态相关的资源分配. 文献 [17] 同时考虑信源和信道的随机性, 建立了一个在保证统计 QoS 前提下使能效最高的弹性接入框架. 为了设计依赖于队列状态的资源分配策略, 文献 [17] 引入了依赖于队列长度的多状态 QoS 指数. 这个想法曾在文献 [11] 中用于设计有线传输网络的动态带宽配置.

类似于文献 [11], 将 $Q_{\max}$ 分为长度为 $l=\frac{Q_{\max}}{N_q}$ 的 N_q 段, 其中 $Q_\infty\in((i-1)l,il]$ 是队列的第 i 个状态. 当队列长度处于第 i 段中时, QoS 指数是一个常数. N_q- 状态 QoS 指数 θ_i 是 $\frac{i}{N_q}$ 的函数, $i=1,2,\ldots,N_q$. 记对应于 $\{\theta_1,\ldots,\theta_{N_q}\}$ 的多状态资源分配策略为 $\{\phi_1,\ldots,\phi_{N_q}\}$. 当在每个状态中都满足 $E_{\mathrm{C}}(\theta_i,\phi_i)\geqslant E_{\mathrm{B}}(\theta_i)$ 时, 式 (3) 中的上界变为 $\Pr\{Q_\infty>Q_{\max}\}\leqslant\exp(-l\sum_{i=1}^{N_q}\theta_i)$. QoS 需求 $(Q_{\max},\varepsilon_Q)$ 可以由 $\Pr\{Q_\infty>Q_{\max}\}=\varepsilon_Q$ 保障. 这意味着多状态 QoS 指数 $\theta_1,\ldots,\theta_{N_q}$ 需要满足

$$\frac{1}{N_q}\sum_{i=1}^{N_q}\theta_i=\frac{\ln(1/\varepsilon_Q)}{Q_{\max}}\triangleq\bar{\theta}. \tag{9}$$

为了满足 $(D_{\max}, \varepsilon_D)$, 多状态 QoS 指数 $\theta_1,\ldots,\theta_{N_q}$ 需要随着队列长度增加而增加. 也就是说, 如果 $i \leqslant j$, 则 $\theta_i \leqslant \theta_j$.

根据多状态的 QoS 指数, 可以在满足 $E_{\mathrm{C}}(\theta_i, \phi_i) \geqslant E_{\mathrm{B}}(\theta_i)$ 的前提下设计高能效资源分配策略 $\{\phi_1,\ldots,\phi_{N_q}\}$. 为此, 文献 [17] 以多天线正交频分复用系统为例, 提出了如下的两步策略, 联合优化了载波数和发射功率: 第 1 步对于任意给定的 θ_i, 优化 ϕ_i 使在第 i 个状态下的总功耗最小. 通过第 1 步的优化, 可以获得最小总功耗与 θ_i 的函数关系, 记作 $P^*_{\mathrm{tot}}(\theta_i)$. 第 2 步优化不同队列状态下的 θ_i, 使平均总功耗对队列长度的数学期望最小. 为了得到最优的 $\{\theta_1,\ldots,\theta_{N_q}\}$, 一个可行的方法是将离散的队列状态连续化. 记 QoS 指标随着队列长度变化的连续函数为 $\theta(q)$, 则平均总功耗可以表示为 $\mathbb{E}\{P^*_{\mathrm{tot}}[\theta(q)]\}$. 为了最小化平均总功耗, 只需要找到 $\theta(q)$ 的最优函数形式. 这个问题是一个泛函极值问题, 可以通过求解 Euler-Lagrange 方程得到最优解.

基于以上框架, 文献 [17] 进一步以大规模多天线系统服务一个一阶自回归 (first order autoregressive, AR(1)) 信源为例, 提供了最优策略的闭式解.

3 延时容忍业务的高能效优化

近年来的研究发现, 人类行为具有可预测性, 因此利用预测的用户级情境信息 (用户移动轨迹) 和网络级情境信息 (业务流量) 来提高无线通信系统的性能和移动用户的体验开始引起学术界的关注 [16,18].

对于非实时业务, 利用所预测的未来网络级和用户级情境信息优化资源分配, 不但可以提升每个用户的服务质量, 还可以提升系统的性能. 这是因为对于流媒体业务等典型的非实时业务, 所需要传输的视频已经存储在内容服务器 (甚至在基站等无线边缘节点) 上, 并且用户对延迟也不那么敏感, 使得基站能在用户距离基站较近的时候进行传输. 早期的文献通过假设未来瞬时数据率可以准确预测来提升系统性能, 如文献 [18]. 由于瞬时数据率依赖于瞬时信道信息, 而瞬时信道仅在信道相干时间内可预测, 更合理、更现实的假设是已知未来的平均信道和平均流量, 即已知未来数据率的统计信息.

3.1 基于平均信道增益的高能效预测资源分配

未来平均信道增益可以根据预测的用户运动轨迹和射频信号地图来获得. 以文件下载为例 (在规定的时间段内传完一定的数据量), 文献 [19] 通过理论分析得出如下结论, 只知道未来平均信道增益时系统的能效与完全已知未来瞬时信道时达到的能效非常接近. 文献 [20] 假定已知未来的平均信道增益, 在保障文件传输中断概率 (即在规定的时间内未传完用户所需文件的概率) 的前提下, 优化了使系统总能耗最小的资源分配策略. 对于边看边传的点播式视频业务, 在保障视频不中断的前提下, 文献 [21] 在视频播放开始的时候, 基于未来的平均信道增益对未来一段时间内的资源分配进行了规划, 而后再根据在线估计的瞬时信道增益进一步优化了在视频播放过程中的功率控制, 得到了一个双时间尺度的高能效传输策略. 这些方法获得能效提升的本质原因, 是因为如果可以预测未来平均信道增益, 则可以更激进地利用信道分集.

当同时可以预测未来的平均信道增益和平均网络流量时, 文献 [16] 进一步得到了一个低复杂度的资源规划与分配方法, 并拓展到了多个非实时用户的场景中.

平均信道增益与功率 – 数据率的关系. 由于非实时业务的延时相对于平均信道增益发生明显改变的时间 (即平均信道的相干时间) 较长, 因此系统可以选择在平均信道增益好的时候多传数据. 平均

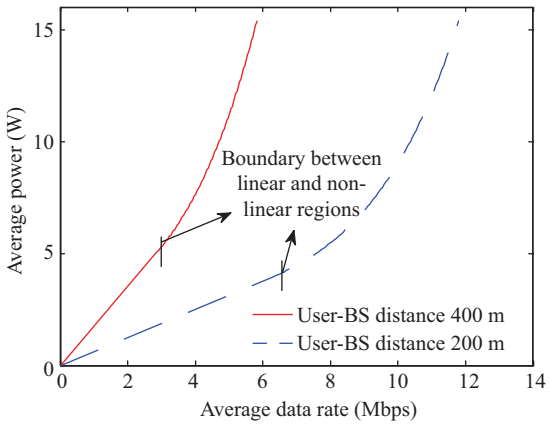

图 6　(网络版彩图) 平均信道增益对功率 – 数据率关系的影响 [21]

Figure 6　(Color online) Impact of large scale channel gain on power-rate relation [21]

信道增益对功率 – 数据率关系的影响如图 6 所示. 当用户离基站较远时, 功率 – 数据率关系曲线较陡. 这意味着要达到相同的数据率, 平均功耗随着用户与基站距离的增加而增加. 此外, 功率 – 数据率线性区域的范围随着用户离基站距离的增加而缩小.

3.2　基于用户偏好的高能效业务推送

通过预测公众或者单个用户对内容的喜好, 还可以通过主动缓存来提升网络能效. 在无线边缘节点进行主动缓存可以分为 3 类: 一是根据文件流行度, 将热门文件提前缓存在基站的存储设备中; 二是根据每个用户的个人喜好, 将内容直接缓存在用户的存储卡中, 即提前推送; 三是将热门文件提前缓存在用户的存储卡中, 当用户所请求的文件不在自己的存储里时, 可以通过终端直传从附近已经缓存所请求文件的用户获取文件. 下面考虑利用用户个人偏好给用户提前推送.

主动缓存过程包含实际请求到达前基于对内容流行度或用户喜好的预测进行内容存放的阶段, 以及用户发出请求后进行内容传输的阶段. 预测用户可能喜好的文件并提前将其缓存在用户端, 即将内容推送给用户, 一直被认为是一种可以有效提高用户体验的技术. 基站一般通过单播的方式在用户信道状态较好时进行内容的下载 [22]. 最近的研究表明, 基于用户的共同喜好通过广播的方式推送相同的内容给用户, 可以将无线的负载向专用带宽或非高峰时段转移. 这时, 由于基站仅需要在内容传输阶段服务那些缓存未命中的用户, 即那些当请求文件时该文件未在本地缓存的用户, 这种提前推送方式除了可以降低用户的平均时延, 还可以提高无线网络的吞吐量.

尽管不同用户的需求可能相关, 但不会完全一致. 因此, 从缓存命中率的角度考虑, 向每个用户推送其感兴趣的文件比向所有用户推送流行文件更加有效. 此外, 假如用户所需求的文件、需求的时间、用户的运动轨迹和网络资源的平均使用情况可预测, 那么网络中的中心处理器可以为用户制定一个传输计划, 即何时、何地以及用多少资源向用户推送什么内容, 并将传输计划发送给该用户运动轨迹经过的小区基站. 通过在较长时间 (分钟甚至小时级) 内利用网络的剩余资源并在用户的平均信道较好时进行传输, 虽然单播会在内容存放阶段比广播消耗更多的资源, 但基于用户喜好进行提前推送可以比基于广播流行文件推送的总能耗 (包括内容存放和传输阶段) 更低 [23].

然而, 对用户的需求, 如请求的内容及请求到达时间, 都不可能完全准确地预测. 对于提前推送, 预测的不确定性将不可避免地导致基站推送不需要的文件给用户, 这不但浪费网络中的剩余资源, 也

浪费基站的能量 [24], 更不用说用户是否允许基站占用其很大的存储空间来进行推送. 为了以可接受的代价获得提前推送带来的增益, 研究存在不确定性时何时推送能带来好处至关重要. 文献 [24] 考虑了单个用户请求的内容及请求到达时间的不确定性, 对提前推送所消耗的基站能量进行了初步分析. 研究结果表明, 提前推送的文件数对基站节能至关重要; 请求到达时间的预测不确定性对节能的影响小于请求内容预测不确定性的影响, 二者均小于用户偏好对节能的影响.

4 超可靠低延时业务的跨层优化

超可靠低延时业务是第五代无线通信中的一大类典型业务 [25]. 不同于传统业务, URLLC 不仅仅要求超短端到端延时与极高的可靠性, 还对网络可用性有很高的要求. 例如: 1 ms 端到端延时、10^{-7} 的丢包率以及 99.999 % 的网络可用性 [26].

(1) URLLC 的 QoS 需求. URLLC 的 QoS 需求包括端到端延时和总丢包率两个方面. 端到端延时的各个组成部分与造成丢包的因素依赖于传输模式与网络架构.

如果采用蜂窝通信, 当发送与接收用户在同一个小区内时, 端到端延时包括上下行传输延时 D^{u} 和 D^{d}, 以及基站的排队延时 D^{q} [27]. 如果发送用户和接收用户不在一个小区, 则需要通过回传网传输数据包, 因此还包括回传网延迟.

对于通信距离很长的场景, 传播延时是端到端延时的主要组成部分. 如何在长距离通信场景中实现超短延时和超高可靠性非常困难. 然而, 即便在一个局部通信的场景下 (即所有用户都接入相邻的几个基站), 如何实现超短延时和超高可靠性也依然非常有挑战.

由于传输延时很短, 信道编码的长度是有限的. 使用有限长信道编码无法做到传输差错为零. 因此, 一部分数据包会因为传输差错而丢失. 记上行和下行的传输差错概率为 ϵ^{u} 和 ϵ^{d}. 在蜂窝通信中, 由于多个上行用户的数据包是随机到达的, 要保证基站缓冲中的延时以概率 1 小于排队延时界非常困难. 如果一个数据包的排队延时超过了延时界, 这个数据包就不会再被传输. 因此, 一部分数据包会因为超出了延时界而丢失, 记为 ϵ^{q}. 对于 URLLC, 端到端延时极短, 通常小于典型场景中的信道相干时间, 在 Rayleigh 衰落信道中满足排队延时所需的最大发射功率可能无界 [27]. 为了在最大发射功率约束下满足排队延时, 文献 [27] 中提出了主动丢包机制, 并控制总的丢包概率. 为了保证排队延时, 当信道差时, 一部分数据包直接被丢弃. 记主动丢包的概率为 ϵ^{h}. 则在蜂窝通信中 URLLC 的 QoS 需求可以表示为

$$D^{\mathrm{u}} + D^{\mathrm{q}} + D^{\mathrm{d}} = D_{\max}, \tag{10}$$

$$\epsilon^{\mathrm{u}} + \epsilon^{\mathrm{q}} + \epsilon^{\mathrm{d}} + \epsilon^{\mathrm{h}} = \epsilon_{\max}, \tag{11}$$

其中 $D_{\max}$ 为端到端延时界, $\epsilon_{\max}$ 为总的丢包率约束.

如果发送与接收用户之间直接使用终端直传通信模式, 那么端到端延时只含有传输延时. 造成丢包的因素只有传输差错. 无论使用何种传输模式, 保障 URLLC 无线接入的 QoS 约束都可以表示为式 (10) 和 (11) 的形式, 根据不同的传输模式, 延时和丢包的组成部分和每个部分的表达式会有所不同.

(2) URLLC 的网络可用性. 网络可用性与覆盖概率类似, 文献 [28] 将可用性定义为系统保障网络中用户 QoS 需求的概率. 对于超可靠低延时通信的典型业务, 如自动驾驶, 只有保障极高的网络可用性, 才能满足业务需求. 如果网络可用性为 99.999 %, 则每十万个用户中有一个用户的 QoS 需求无法满足.

(3) URLLC 的跨层资源分配. 要提升 URLLC 的资源利用率, 首先要保证 QoS 需求和网络可用性. 为了刻画有限长信道编码的传输差错概率, 可以采用文献 [29] 中的 Gauss 近似:

$$R(\varepsilon_{\mathrm{t}}) \approx \frac{W}{\ln 2}\left[\ln\left(1+\frac{\mu P_{\mathrm{t}} g}{N_0 W}\right)-\sqrt{\frac{V}{D_{\mathrm{t}} W}} f_{\mathrm{Q}}^{-1}\left(\epsilon_{\mathrm{t}}\right)\right](\mathrm{bit/s}), \tag{12}$$

其中 W 为带宽, D_{t} 为传输时间, P_{t} 是发射功率, g 是小尺度信道增益, ϵ_{t} 是传输差错概率, $f_{\mathrm{Q}}^{-1}(x)$ 是 Q 函数的反函数, V 可以表示为 $V=1-\frac{1}{(1+\frac{\mu P_{\mathrm{t}} g}{N_0 W})^2}$.

与 Shannon 容量公式不同, 上述有限长信道编码的可达数据率不再是发射功率的严格凸函数. 由于功率 – 数据率关系发生了变化, 能效 – 延时关系也与利用 Shannon 容量得出的结果不同. 如何刻画能效 – 延时 – 可靠性的基本关系有待进一步研究.

为了保证排队延时, 文献 [30] 基于简化的有限长信道编码可达数据率, 以 Poisson 到达过程为例验证了有效带宽在排队延时很短的时候依然适用, 对系统带宽和基站天线数配置进行了跨层高能效优化, 并初步得到了所需系统资源与延时界和可靠性的关系. 研究结果表明, 为了保证 URLLC 的 QoS, 使系统能效最高的系统带宽和基站天线数上界正比于 $1/D_{\max}^{\mathrm{q}}$ 和 $\ln(1/\varepsilon^{\mathrm{q}})$. 利用 Gauss 近似和有效带宽, 文献 [27] 以下行传输为例, 通过优化与队列长度和信道状态相关的功率控制策略与主动丢包策略, 最小化了保障 QoS 所需的最大发射功率, 并得到了最优策略达到的 ϵ^{q}, ϵ^{d} 和 ϵ^{h}. 数值结果表明, ϵ^{q}, ϵ^{d} 和 ϵ^{h} 在同一数量级. 此外, 理论分析发现如果要控制其中任意一个丢包概率为零, 则所需的最大发射功率为无穷大. 因此, 这 3 个因素在蜂窝网下行传输中都不可忽略.

未来通信场景中用户密度可能非常大 [25], 因此系统所需的总带宽也可能会很大. 对于下行传输, 当用户数较多时, 一个可能的方法是采用广播. 然而, 在广播系统中, 接收端没有任何反馈, 如何保障超高可靠性和超短延时是一个非常有挑战的问题. 此外, 开环的广播系统中保障高可靠低时延是否比闭环的单播系统节约资源也还有待进一步研究.

5 结论

本文对过去 5 年来我们在超蜂窝网络架构下对能效与延时界需求的基本关系及其在高能效跨层资源分配方面的应用研究进行总结. 由于高能效优化需要考虑电路功耗, 可优化配置的资源与只降低发射功率不同, 导致了能效与延时需求的基本关系与传统分析得到的关系不同. 由于现有及未来基站的电路功耗不可忽略, 且具有开关部分电路的能力, 考虑与数据率相关的电路功耗, 正是这样的功耗模型使支持给定数据率所需的总功耗不再总是数据率的凸函数, 进而改变了能效随延时需求变化的基本规律. 研究表明, 功率 – 数据率的基本关系决定了能效 – 延时界需求的基本关系, 能效与延时界需求之间存在非折中 (即饱和) 区域, 在此区域内延时界需求减小不增加能效, 但需更多带宽或天线资源. 具体而言, 当所需平均功率是所支持平均数据率的线性函数时, 能效 – 延时界不是折中关系. 通过联合优化与队列长度有关的两状态功率与带宽分配策略, 发现当所需延时界较大, 使得系统的带宽约束不起作用时, 功率 – 数据率为线性关系, 且总能达到能效极限; 否则, 此策略可以达到 Parato 最优的能效 – 延时折中关系的下界. 进一步以大规模多天线系统和复合 Poisson 到达过程为例, 导出了能效 – 延时折中区与非折中区边界点的闭式解, 发现非折中区域随着带宽和发射天线数增加而增长. 基于对上述基本关系的理论研究, 提出了对延时敏感业务进行跨层资源分配的优化框架, 并进一步面向内容分发 (延时需求很长) 和触觉互联网 (延时需求很短) 业务, 研究了对延时不敏感业务的高能效预测资源分配方法和对超可靠低延时业务的高能效资源配置. 初步研究结果表明, 预测资源分配可通过

利用网络的剩余资源和用户的平均信道大大降低基站的能耗, 在保证触感互联网无线接入服务质量需求的前提下, 使系统能效最高的资源反比于所需的延时界, 正比于丢包概率倒数的对数.

参考文献

1 Zhang S, Wu Q, Xu S, et al. Fundamental green tradeoffs: progresses, challenges, and impacts on 5G networks. IEEE Commun Surv Tuts, 2017, 19: 33–56

2 Zafer M A, Modiano E. Optimal rate control for delay-constrained data transmission over a wireless channel. IEEE Trans Inf Theory, 2008, 54: 4020–4039

3 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

4 She C, Yang C. Energy efficiency and delay in wireless systems: is their relation always a tradeoff? IEEE Trans Wirel Commun, 2016, 15: 7215–7228

5 3GPP. Further advancements for E-UTRA physical layer aspects. TSG RAN TR 36.814 v9.0.0. http://www.3gpp.org/ftp/Specs/archive/36_series/36.814/36814-900.zip

6 Cisco. Cisco visual networking index: global mobile data traffic forecast update, 2016–2021. Cisco White Paper, Feb. 2017. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

7 Juluri P, Tamarapalli V, Medhi D. Measurement of quality of experience of video-on-demand services: a survey. IEEE Commun Surv Tut, 2016, 18: 401–418

8 Fettweis G P. The tactile internet: applications & challenges. IEEE Veh Tech Mag, 2014, 9: 64–70

9 Osseiran A, Boccardi F, Braun V, et al. Scenarios for 5G mobile and wireless communications: the vision of the METIS project. IEEE Commun Mag, 2014, 52: 26–35

10 Berry R A, Gallager R G. Communication over fading channels with delay constraints. IEEE Trans Inf Theory, 2002, 48: 1135–1149

11 Chang C S, Thomas J A. Effective bandwidth in high-speed digital networks. IEEE J Sel Area Commun, 1995, 13: 1091–1100

12 Wu D, Negi R. Effective capacity: a wireless link model for support of quality of service. IEEE Trans Wirel Commun, 2003, 2: 630–643

13 Chen Y, Zhang S Q, Xu S G, et al. Fundamental trade-offs on green wireless networks. IEEE Commun Mag, 2011, 49: 30–37

14 Berry R A. Optimal power-delay tradeoffs in fading channels—small-delay asymptotics. IEEE Trans Inf Theory, 2013, 59: 3939–3952

15 She C, Yang C. Optimal EE-delay relation in wireless systems. In: Proceedings of the IEEE Online Conference on Green Communications, Piscataway, 2015. 36–41

16 Yao C, Yang C, Xiong Z. Energy-saving predictive resource planning and allocation. IEEE Trans Commun, 2016, 64: 5078–5095

17 She C, Yang C, Liu L. Energy-efficient resource allocation for MIMO-OFDM systems serving random sources with statistical QoS requirement. IEEE Trans Commun, 2015, 63: 4125–4141

18 Abou-zeid H, Hassanein H. Predictive green wireless access: exploiting mobility and application information. IEEE Wirel Commun, 2013, 20: 92–99

19 Yao C, Yang C. Role of large scale channel information on predictive resource allocation. In: Proceedings of the IEEE Wireless Communications and Networking Conference, Doha, 2016. 1–6

20 Hu Y, Han S, Yang C. Context-aware energy saving with proactive power allocation. In: Proceedings of the IEEE Global Conference on Signal and Information Processing, Orlando, 2015. 53–57

21 She C, Yang C. Context aware energy efficient optimization for video on-demand service over wireless networks. In: Proceedings of the IEEE/CIC International Conference on Communications in China, Shenzhen, 2015. 1–6

22 Higgins B D, Flinn J, Giuli T J, et al. Informed mobile prefetching. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, Lake District, 2012. 155–168

23 Yao C, Chen B, Yang C, et al. Energy saving pushing based on personal interest and context information. In:

Proceedings of the IEEE 83rd Vehicular Technology Conference, Nanjing, 2016. 1–5
24 Yao C, Yang C. Impact of uncertainty in predicting the user's request on pushing. In: Proceedings of the IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, Valencia, 2016. 1–6
25 3GPP. Study on scenarios and requirements for next generation access technologies. TSG RAN TR 38.913. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2996
26 Fettweis G P. The tactile internet: applications & challenges. IEEE Vehic Tech Mag, 2014, 9: 64–70
27 She C, Yang C, Quek T Q S. Cross-layer transmission design for tactile internet. In: Proceedings of the IEEE Global Communications Conference, Washington, 2016. 1–6
28 Popovski P, Mange G, Fertl P, et al. Deliverable D6.3 intermediate system evaluation results. ICT-317669-METIS/D6.3, 2014. https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D6.3_v1.pdf
29 Yang W, Durisi G, Koch T, et al. Quasi-static multiple-antenna fading channels at finite blocklength. IEEE Trans Inf Theory, 2014, 60: 4232–4264
30 She C, Yang C. Energy efficient design for tactile internet. In: Proceedings of the IEEE/CIC International Conference on Communications in China, Chengdu, 2016. 1–6

Energy efficiency-QoS relation and its application in wireless networks

Changyang SHE* & Chenyang YANG*

School of Electronics and Information Engineering, Beihang University, Beijing 100191*, China*
* Corresponding author. E-mail: cyshe@buaa.edu.cn, cyyang@buaa.edu.cn

Abstract Since improving energy efficiency (EE) must not sacrifice quality-of-service (QoS), the EE-QoS relation is a fundamental issue for EE optimization. This paper provides an overview of our research on the EE-QoS relation and its application to elastic access optimization in hyper-cellular networks over the past five years. Because delay bound is a representative QoS requirement, we first discuss our findings on the EE-delay relation, and show that if average total power consumption is linear with average service rate, then there exists a non-tradeoff region in the EE-delay relationship. We then summarize how to design energy efficient resource allocation for different kinds of services, such as real-time and non-real-time services, and ultra-reliable and ultra-low latency service. For real-time service, in order to achieve the optimal EE-delay relation, resource allocation should depend on queue length. For delay tolerant service, it is possible to leverage predictive information for making a resource allocation plan and pushing data. The results show that the trajectories and preferences of users are very useful for improving EE in wireless systems. Finally, we summarize preliminary results for the resources required to maximize EE under the constraints of ultra-reliable and low-latency network.

Keywords energy efficiency, quality-of-service, resource allocation, wireless communications

Changyang SHE received his B.E. degree from the Honors College of Beihang University (formerly Beijing University of Aeronautics and Astronautics, BUAA), Beijing, China, in 2012. He is currently pursuing a Ph.D. degree from the School of Electronics and Information Engineering at BUAA. His research interests include tactile internet, machine type communication, big data for resource allocation in wireless networks, and energy efficient transmission in 5G systems.

Chenyang YANG received her Ph.D. degree in electrical engineering from Beihang University (formerly Beijing University of Aeronautics and Astronautics, BUAA), China, in 1997. She has been a professor with the School of Electronics and Information Engineering, BUAA since 1999. Her recent research interests include green radio, wireless big data, local caching, and tactile internet.