SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •
Special Focus on Machine-Type Communications

October 2017, Vol. 60 100302:1–100302:15
doi: 10.1007/s11432-017-9183-2

Coexistence of delay-sensitive MTC/HTC traffic in large scale networks

Jianghong SHI[1], Chen LIU[1], Xuemin HONG[1]* & Cheng-Xiang WANG[2]

[1]*Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Ministry of Education, Xiamen University, Xiamen University, Xiamen* 361005*, China;*
[2]*Institute of Sensors, Signals and Systems, School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh* EH14 4AS*, UK*

Received April 22, 2017; accepted July 3, 2017; published online September 6, 2017

Abstract The support for mission critical machine-type-communication (cMTC) services is indispensable for the 5th generation (5G) mobile communication systems. As the cMTC and (part of) the conventional human-type-communication (HTC) services are broadband and delay-sensitive services, how to ensure their coexistence is a new and challenging problem. This paper investigates the problem of service-level resource allocation, which decides how cMTC and HTC traffic share a limited amount of radio resource. Considering a large-scale network, we put forth a system model that integrates queuing models and stochastic geometric models to characterize the delay performance in self-interfering scenarios. A service-level resource allocation scheme called load division is proposed. The delay and throughput performance of cMTC and HTC are derived under different resource allocation schemes and priority scheduling policies. We show that compared with the baseline scheme of frequency division, the proposed load division scheme can significantly improve the delay performance of cMTC service, at a cost of slightly degraded MTC and HTC service capacities.

Keywords machine type communication, coexistence, resource allocation, priority queue, delay sensitive

Citation Shi J H, Liu C, Hong X M, et al. Coexistence of delay-sensitive MTC/HTC traffic in large scale networks. Sci China Inf Sci, 2017, 60(10): 100302, doi: 10.1007/s11432-017-9183-2

1 Introduction

It is envisioned that the majority of wireless connections in the near future will be originated by autonomous machines and devices instead of human-operated mobile terminals. Due to the increasing heterogeneity of end devices, the 5th generation (5G) mobile communication system [1–8] is required to support diverse applications categorized into three scenarios [9]. The first scenario addresses traditional mobile broadband services such as voice/video streaming and multimedia content delivery. The second scenario targets wide-area Internet of Things (IoT) services, which aim to connect millions of small embedded devices to the network [10–12]. The third scenario targets mission critical communication services such as real-time control of vehicles, robots, and industrial process automation [13]. While the first scenario belongs to the traditional paradigm of human-type communications (HTCs), the latter two scenarios are novel paradigms called machine-type communications (MTCs) [14, 15].

* Corresponding author (email: xuemin.hong@xmu.edu.cn)

How to best accommodate HTC and MTC services in a seamlessly integrated network is a critical engineering challenge for 5G. The difficulty comes from the disparate design priorities imposed by different scenarios [9]. In the first scenario of HTC, the design priority is to provide ultra high data rate (e.g., 1 Gbit/s) access to mobile user devices with high spectrum efficiency and energy efficiency [1, 2]. In the second scenario of massive MTC (mMTC), high priorities are placed for low device cost, low device power, ubiquitous coverage, and high access density [12]. In the third scenario of critical MTC (cMTC), the priority is to support ultra-reliable (e.g., 1e-5), low latency (e.g., 1 ms) communications with medium data rate (e.g., 50 Mbit/s). It is widely believed that specially designed air interfaces are needed to support different scenarios.

The coexistence of HTC and mMTC services has attracted significant research attention in recent years [16–28]. The focus is to minimize the impact of dense, impulsive, delay-tolerant mMTC traffic on the quality-of-service (QoS) of HTC services. Various aspects such as resource allocation [17, 21, 22, 25], energy efficiency [19, 23, 24, 26], and random access schemes [16, 20, 27, 28] have been investigated in the literature. Apart from theoretical studies, three mMTC standards have been developed by the 3GPP: EC-GSM-Io [29], LTE-eMTC [30, 31], and narrowband IoT (NB-IoT) [32–35]. EC-GSM-IoT and LTE-eMTC are backward compatible standards that aim to enhance existing Global System for Mobile Communications (GSM) [36] and Long Term Evolution (LTE) [37] networks, respectively. NB-IoT is a new 3GPP air interface designed to coexist with legacy GSM and LTE systems. Taking NB-IoT for example, it is a narrowband system with a minimum bandwidth of 180 kHz. NB-IoT has three modes of deployment: standalone, guard-band, and in-band. The first two modes use separated frequency bands to ensure coexistence. In the third mode, one dedicated GSM carrier (200 kHz) or one LTE physical resource block (180 kHz) can be allocated for NB-IoT. The narrowband nature of NB-IoT makes it easier to coexist with the broadband HTC by frequency division.

As both HTC and cMTC are broadband and delay-sensitive services, how to ensure their coexistence is a new and challenging problem. Different from mMTC, the cMTC service has a higher priority than HTC service. To date, both the theoretical research [38–45] and standardization [13] of cMTC systems are still in an early stage. Existing literature on cMTC mainly addressed different aspects of physical and medium-access-control (MAC) layer design [38–45], while focused studies on the coexistence of HTC and cMTC have so far received limited coverage. In particular, to our best knowledge, the cMTC/HTC coexistence problem with respect to large and self-interfering networks has not been studied.

The cMTC/HTC coexistence problem is particularly challenging because it requires an analytical framework that is able to describe the delay/queuing behavior in the temporal domain as well as the interfering phenomenon in the spatial domain. A framework of "timely throughput" was proposed in [46] and recently adopted for the analysis of HCN in [47, 48]. It assumes that a queuing packet will be dropped if the packet passes a critical delay. This is a useful framework suitable for loss-tolerant traffic, but falls short to characterize loss-sensitive cMTC traffic. In this paper, we put forth a new analytical framework that integrates stochastic geometry models and queuing models. Our model differs from [46] in that packets are not allowed to be dropped, hence the service is reliable. In particular, our model is shown to be useful in revealing key analytical insights with respect to the mean measures.

To our best knowledge, our paper is the first research effort focused on the coexistence performance analysis of delay-sensitive MTC/HTC traffic, under a setting of large scale networks with inter-cell interference. The main contributions of our paper are as follows: First, we put forth a new analytical framework that integrates stochastic geometry models and queuing models to jointly capture the spatial and temporal behavior of a network. Compared with existing alternatives, our framework is particularly relevant for delay-sensitive and loss-intolerant traffic. Our framework is shown to be useful and analytically tractable in revealing key insights into the mean measures of performance. Second, we propose a new service level resource allocation scheme called load division, which is a new paradigm different from traditional schemes such as frequency and time division. Our analysis show that compared with frequency division, the load division scheme yields a different tradeoff behavior in the delay domain, making it a promising candidate for mission critical and delay sensitive MTC services.

The remainder of the paper is organized as follows. Section 2 describes the system model. Section 3

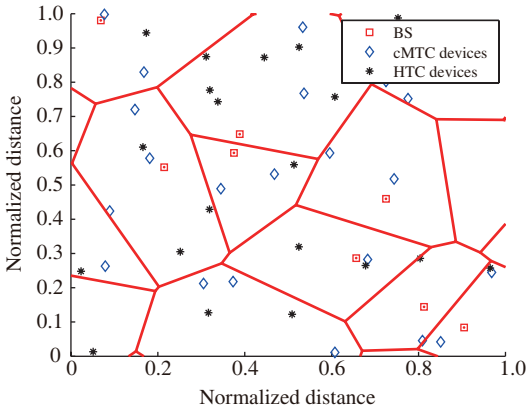

Figure 1 (Color online) Distribution of BSs, cMTC devices, and HTC devices on the plane.

gives some preliminaries. Section 4 presents the performance analysis, followed by numerical results and discussions in Section 5. Finally, Section 6 concludes the paper.

2 System model

2.1 General settings

We consider the downlink of a large scale cellular network with multiple base stations (BSs), cMTC devices and HTC devices. As illustrated in Figure 1, BSs are distributed in the two-dimensional Euclidean plane according to a stationary Poisson point process (PPP) with intensity λ_b. Similarly, the locations of the cMTC and HTC devices are assumed to follow two stationary PPPs on the plane with intensities λ_m and λ_h, respectively. A device is associated with the nearest BS. This means that the dimensioning of the cellular network is characterized by Poisson Voroini cells defined with respect to the BS point process. The available system bandwidth is W and the frequency reuse factor is assumed to be 1. The BSs are assumed to be homogeneous and active with a constant transmit power P.

Let us now consider a typical cell randomly selected from the network. The number of cMTC devices currently located in the cell is denoted by N_m. All the cMTC devices in the cell are assumed to have a homogeneous traffic of packets characterized by a Poisson arrival process with mean interval $\overline{\alpha}_m$. Each packet has a fixed size denoted by L_m. The transmission rate of a typical cMTC device is denoted by R_m, which is a random variable determined by the instantaneous signal-to-noise-and-interference ratio (SINR) γ_m. The SINR is assumed to vary on a packet-by-packet basis, which means that the SINR values remain constant during the transmission of a packet, but vary randomly and independently across different packets. This assumption implies that the average transmission time of a packet is roughly comparable to the coherent time of the channel, which is a reasonable assumption that resembles the widely-applied assumption of block fading channels. By making the block length to be random and continuous instead of fixed and discrete, this assumption enables the integration of queueing models and stochastic geometric models into a coherent analytical framework.

Let β_m be the transmission time of a cMTC packet, we have

$$\beta_m = \frac{L_m}{R_m}. \tag{1}$$

The mean of β_m is denote as $\overline{\beta}_m$. According to the properties of Poisson process, the aggregated traffic at the BS is also a Poisson process. Let ρ_m denote the aggregated cMTC traffic load at the BS, we have

$$\rho_m = \frac{N_m \overline{\beta}_m}{\overline{\alpha}_m}. \tag{2}$$

Similarly, the HTC service in the cell can be characterized by the number of HTC devices N_h, the mean packet arriving interval $\overline{\alpha}_h$, the size L_h of the HTC packet, the transmission rate R_h, the transmission

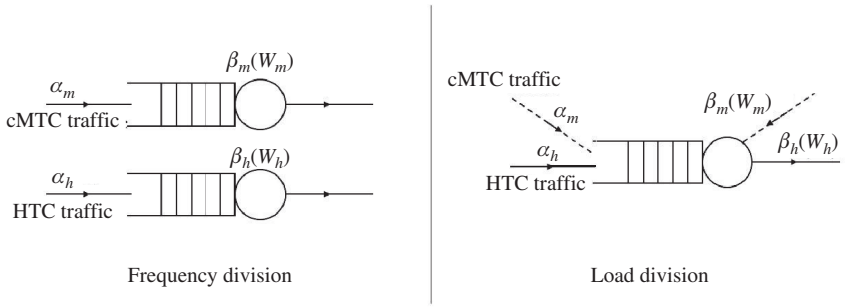

Figure 2 Queue models for the frequency division scheme and load division scheme.

time β_h, and the traffic load ρ_h. We have

$$\beta_h = \frac{L_h}{R_h}, \tag{3}$$

$$\rho_h = \frac{N_h \overline{\beta}_h}{\overline{\alpha}_h}. \tag{4}$$

The total traffic load of the cell is given by $\rho = \rho_m + \rho_h$ and we have $0 < \rho < 1$. In practice, the values of ρ_m and ρ_h should be monitored and kept to a small value to ensure a manageable delay performance.

2.2 Resource allocation for cMTC and HTC services

For the cMTC and HTC services to coexist in a cell, the radio resource should be properly allocated to each user. We consider a two-stage resource allocation procedure: in the first stage, the radio resource is allocated/partitioned between the cMTC and HTC services. In the second stage, the radio resource is equally shared among multiple users in the same service. Our paper focuses on the first stage of service-level resource allocation. Two allocation schemes are considered: frequency division and load division. The former is a baseline scheme, while the latter is a new proposal.

Scheme 1: Frequency division.

The frequency division scheme simply divides the entire system band into two orthogonal sets and allocate each set for a service. We introduce a parameter ϵ to indicate the ratio of bandwidth allocated to the cMTC service. It follows that

$$\begin{cases} W_m^F = \epsilon W, \\ W_h^F = (1-\epsilon) W, \end{cases} \tag{5}$$

where W_m^F and W_h^F are bandwidth allocated to the cMTC and HTC services, respectively. Now consider a typical cMTC user in the cell and define γ_m as the instantaneous SINR of the user, the transmission rate R_m of the user is given by

$$R_m = W_m^F \log_2(1+\gamma_m). \tag{6}$$

It follows that the transmission time of a cMTC packet is given by

$$\beta_m^F = \frac{L_m}{\epsilon W \log_2(1+\gamma_m)}. \tag{7}$$

Here, β_m^F is a random variable. Its mean and second moment are denoted as $\bar{\beta}_m^F$ and $\hat{\beta}_m$, respectively. The transmission time of an HTC packet is denoted by β_h^F, which can be calculated following a similar procedure. We note that the frequency division scheme does not necessarily mean that the resource partition is fixed. In practice, ϵ can be dynamically adjusted according to the traffic load of each type of service. As shown in Figure 2, the frequency division scheme allows the two services to enter independent queuing processes.

Scheme 2: Load division.

The load division scheme allows the cMTC and HTC packets to enter and mix in a single queue. The entire bandwidth is used to transmit each packet. Hence we have

$$W_m^L = W_h^L = W. \tag{8}$$

The transmission time of a cMTC packet is given by

$$\beta_m^L = \frac{L_m}{W\log_2(1+\gamma_m)}. \tag{9}$$

The transmission time of an HTC packet can be defined in a similar fashion.

Instead of allocating physical resource blocks to each types of service, load division means that each service has a maximum allowable traffic load. Given the maximum allowable total traffic load of the queue $\rho_{\max}$, the load division scheme gives

$$\rho_m^L = \epsilon\rho_{\max}, \tag{10}$$

$$\rho_h^L = (1-\epsilon)\rho_{\max}. \tag{11}$$

The two resource allocation schemes introduced above will yield different queuing dynamics. In the case of frequency division, we have two independent M/G/1 queues. In the case of load division, we have a two-level M/G/1 priority queue, where different priority scheduling policies can be further introduced to enhance the performance of the cMTC service. The queuing of a packet includes two phases: the queuing phase and transmission phase. In the queuing phase, we assume a time-varying priority policy, which will be introduced latter in detail. In the transmission phase, we consider two policies: non-preemptive and preemptive-resume. In the non-preemptive policy, a newly arriving cMTC packet should wait for an ongoing HTC packet transmission to be completed. In the preemptive-resume policy, a newly arriving cMTC packet can immediately interrupt an ongoing HTC transmission. Afterwards, the interrupted HTC packet can resume transmission from the point where it was interrupted.

It is easy to see that load division is closely related to the concept of statistical time division multiplexing, which is the cornerstone of packet-switch communication networks. Both schemes share the same feature that the full bandwidth is utilized for the transmission of a packet and the time resource is statistically multiplexed to be used on demand. However, we note that in our paper, load division is proposed as a service-level (in contrast to packet-level) resource partition scheme. A key feature of load division is that the total traffic load of a particular service is controlled and bounded to provide manageable delay performance. To our best knowledge, this differs from previous time-division based studies, which lack the control aspect on the traffic load.

3 Preliminaries

This section presents some useful Lemmas.

Lemma 1. Let N_m be the random number of cMTC devices in a typical cell. The probability mass function (PMF) of N_m is given by [49]

$$f_{N_m}(n) = \frac{3.5^{3.5}\Gamma(3.5+n)(\lambda_m/\lambda_b)^n}{\Gamma(3.5)n!(\lambda_m/\lambda_b+3.5)^{n+3.5}}. \tag{12}$$

Similarity, denote N_h as the random number of HTC devices in a cell. The PMF of N_h is $f_{N_h}(n)$, which can be obtained by replacing λ_m with λ_h in (12).

Lemma 2. Denote γ as the SINR of a typical device in the Poisson field of transmitting BSs. Under the assumption that the pass loss exponent equals 4 and the channel is subject to i.i.d. Rayleigh fading, the complementary cumulative distribution function (CDF) of γ is given by [50]

$$\widetilde{F}_\gamma(x) = \frac{\pi^{\frac{3}{2}}\lambda_b}{\sqrt{x/P}}\mathrm{e}^{\frac{a^2}{4x/P}}\mathrm{Q}\left(\frac{a}{\sqrt{2x/P}}\right), \tag{13}$$

where λ_b is the BS density, P is the BS transmit power, $\mathrm{Q}(\cdot)$ is the Q-function, and $a = \pi\lambda_b(1+\sqrt{x}\arctan(\sqrt{x}))$. In the interference limited case, which means P is sufficiently large so that the interference power dominates the noise power, Eq. (13) can be further simplified to

$$\widetilde{F}_\gamma^{\lim}(x) = \frac{1}{1+\sqrt{x}\arctan(\sqrt{x})}. \tag{14}$$

Lemma 3. Given the definition of β_m in (7), the CDF of β_m is given by

$$F_{\beta_m}(x) = \widetilde{F}_{\gamma}\left(2^{\frac{L_m}{W_m}\frac{1}{x}} - 1\right). \tag{15}$$

The derivation simply follows the definition of CDF. Furthermore, we can derive the probability density function (PDF) of β_m by taking the first order derivative of (15). This gives the following corollary:

Corollary 1. The PDF of β_m is given by

$$f_{\beta_m}(x) = \frac{\arctan\phi_x + \frac{\phi_x}{2^{\frac{L_m}{W_m}\frac{1}{x}}}}{\left(1+\phi_x\arctan\phi_x\right)^2}\frac{\ln 2\frac{L_m}{W_m}2^{\frac{L_m}{W_m}\frac{1}{x}}}{2\phi_x}x^{-2}, \tag{16}$$

where

$$\phi_x = \sqrt{2^{\frac{L_m}{W_m}\frac{1}{x}} - 1}. \tag{17}$$

Proof. The PDF of β_m can directly get by taking derivation of (15):

$$f_{\beta_m}(x) = \frac{\mathrm{d}F_{\beta_m}(x)}{\mathrm{d}x}, \tag{18}$$

where

$$F'_{\beta_m} = f_1(x)f_2(x), \tag{19}$$

$$f_1(x) = -\frac{\arctan\phi_x + \frac{\phi_x}{2^{\frac{L_m}{W_m}\frac{1}{x}}}}{\left(1+\phi_x\arctan\phi_x\right)^2}, \tag{20}$$

$$f_2(x) = \frac{\ln 2}{2}\frac{L_m}{W_m}\frac{2^{\frac{L_m}{W_m}\frac{1}{x}}}{\phi_x}x^{-2}. \tag{21}$$

Lemma 4. Let $\overline{D}$ be the mean delay of a packet in an M/G/1 queue without priority, we have [51]

$$\overline{D} = \overline{W} + \overline{\beta} = \frac{\hat{\beta}}{2\alpha(1-\rho)} + \overline{\beta}, \tag{22}$$

where $\overline{W}$ denotes the mean queueing delay of packets, α denotes the time interval of incoming packets, β denotes the transmission time of packets, $\hat{\beta}$ denotes the second-order moment of transmission time, ρ denotes the traffic load of devices, and $\rho = \overline{\beta}/\overline{\alpha}$.

Denote $\overline{D}_p$ as the mean delay of level p, $p \in (1, 2, \ldots, P)$ packet in a time varying priority M/G/1 queue with no-preemptive policy. We have [51]

$$\overline{D_p} = \overline{W_p} + \overline{\beta_p} = \frac{\overline{W_0} + \sum_{i=p}^{P}\rho_i W_i + \sum_{i=1}^{p-1}\rho_i W_i(\frac{\tau_i}{\tau_p})}{1 - \sum_{i=p+1}^{P}\rho_i[1 - (\frac{\tau_p}{\tau_i})]} + \overline{\beta_p}, \tag{23}$$

where $\overline{W_p}$ denotes the mean queueing delay of packets, $\overline{W_0}$ denotes the mean residual life of a packet and τ_p denotes the priority coefficient of level p packets.

$$\overline{W_0} = \sum_{i=1}^{P}\frac{\hat{\beta}_i}{2\alpha_i}. \tag{24}$$

4 Delay and throughput performance analysis

4.1 Frequency division scheme

In this subsection, we derive the delay and throughput performance for cMTC and HTC services with the frequency division scheme.

4.1.1 *Delay performance*

Proposition 1. In the case of frequency division, the mean delay of a cMTC packet and an HTC packet are given by

$$\overline{D}_m^F = \frac{\rho_m}{2(1-\rho_m)}\frac{\hat{\beta}_m^F}{\overline{\beta}_m^F} + \overline{\beta}_m^F, \tag{25}$$

$$\overline{D}_h^F = \frac{\rho_h}{2(1-\rho_h)}\frac{\hat{\beta}_h^F}{\overline{\beta}_h^F} + \overline{\beta}_h^F, \tag{26}$$

respectively, where $\hat{\beta}_m^F$ and $\hat{\beta}_h^F$ are the second-order moments of β_m^F and β_h^F, respectively.

Proof. Consider an arbitrary cell with N_m cMTC devices. According to Lemma 4, the mean delay of cMTC devices in the cell is

$$\overline{D}_{m|N_m}^F = \frac{N_m\hat{\beta}_m^F}{2\overline{\alpha}_m(1-\rho_m)} + \overline{\beta_m^F}. \tag{27}$$

Substituting (2) into (27) yields (25). Parameter $\overline{D}_h^F$ can be derived similar to (27).

We note that the delay of a device is the total time that the device spends in the queue, which consists of two parts. The first part is queueing delay time Q_1, which is the duration from the moment of arrival to the moment when the transmission first starts. The second part it the transmission delay time Q_2, which is the duration from the moment when the transmission first starts to the moment when the transmission ends and the device leaves the system. It follows that $D = Q_1 + Q_2$, where Q_1 and Q_2 are independent RVs.

4.1.2 *Throughput performance*

Proposition 2. In the case of frequency division, the mean throughput of a randomly selected cMTC device is given by

$$\overline{C}_m^F = \frac{L_m\rho_m}{\overline{\beta}_m^F}\sum_{n=1}^{\infty}\frac{1}{n}f_{N_m}(n). \tag{28}$$

Proof. Consider a randomly selected cMTC device. The throughput per device is given by

$$\overline{C}_{m|N_m}^F = \frac{L_m}{\overline{\alpha}_m}. \tag{29}$$

Substituting (2) into (29), we get

$$\overline{C}_{m|N_m}^F = \frac{L_m\rho_m}{N_m\overline{\beta}_m^F}. \tag{30}$$

Here, L_m, ρ_m and β_m are deterministic parameters, while the number of (active) users per cell N_m is a random variable. Parameter N_m indicates that the total throughput is equally shared among all users in a cell. Taking expectations over N_m yields the average per user throughput given by Proposition 2. The average throughput can be seen as either the spatial average of multiple users' throughput over a large network, or the long-term temporal average of the throughput of mobile user that moves across the network.

4.2 Load division scheme

In this section, we adopt the time varying two-class M/G/1 priority queuing model with non-preemptive and preemptive-resume policies to describe the coexistence of cMTC devices and HTC devices. In a time varying priority queue, the priority of a packet in a particular class of service is governed by a time-dependent coefficient, which is dynamically determined by the time of devices staying in the queue. Denote τ_p as the time-dependent coefficient of class p. We assume that for a tagged packet from class p and who has waited in the queue for a duration t, its priority coefficient is given by the following linear equation

$$q_p(t) = t \times \tau_p. \tag{31}$$

Figure 3 illustrates such a queuing process with time-varying priority.

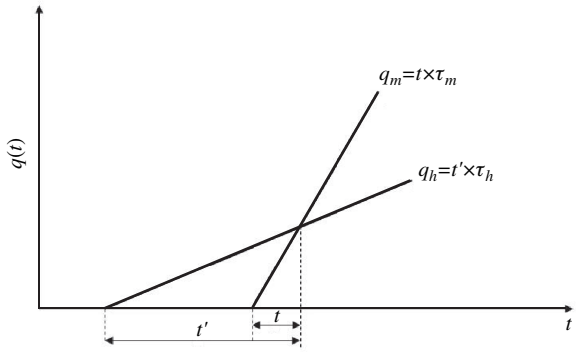

Figure 3 Illustration of time varying priority.

4.2.1 *Time varying priority with non-preemptive policy*

Under the non-preemptive policy, upon the arrival of a cMTC packet, if there is an HTC device receiving transmission from BS, the cMTC device should wait until this HTC device finish its transmission. The delay performance of cMTC devices and HTC devices are given by the following two propositions, respectively.

Proposition 3. For the non-preemptive policy with time varying priority, the mean delay of an HTC packet at a random BS is given by [51]

$$\overline{D}_m^{L|np} = \frac{(1-\rho_h-\rho_m+\rho_h\frac{\tau_h}{\tau_m}+\rho_m\frac{\tau_h}{\tau_m})(\rho_m\frac{\hat{\beta}_m^L}{\overline{\beta}_m^L}+\rho_h\frac{\hat{\beta}_h^L}{\overline{\beta}_h^L})}{2(1-\rho_h-\rho_m)(1-\rho_m(1-\frac{\tau_h}{\tau_m}))}+\overline{\beta}_m^L, \tag{32}$$

$$\overline{D}_h^{L|np} = \frac{\rho_m\frac{\hat{\beta}_m^L}{\overline{\beta}_m^L}+\rho_h\frac{\hat{\beta}_h^L}{\overline{\beta}_h^L}}{2(1-\rho_h-\rho_m)(1-\rho_m(1-\frac{\tau_h}{\tau_m}))}+\overline{\beta}_h^L. \tag{33}$$

Proof. Consider an arbitrary cell with N_m cMTC devices and HTC devices. According to Lemma 4, the mean queueing delay of cMTC and HTC packets is

$$\overline{W}_m^{L|np} = \overline{W_0}+\rho_m\overline{W}_m^{L|pr}+\rho_h\frac{\tau_h}{\tau_m}\overline{W}_h^{L|pr}, \tag{34}$$

$$\overline{W}_h^{L|np} = \frac{\overline{W_0}+\rho_h\overline{W}_h^{L|pr}+\rho_m\overline{W}_m^{L|pr}}{1-\rho_m(1-\frac{\tau_h}{\tau_m})}, \tag{35}$$

Substituting (35) into (34) yields

$$\overline{W}_m^{L|np} = \frac{(1-\rho_h-\rho_m+\rho_h\frac{\tau_h}{\tau_m}+\rho_m\frac{\tau_h}{\tau_m})\overline{W_0}}{(1-\rho_h-\rho_m)(1-\rho_m(1-\frac{\tau_h}{\tau_m}))}. \tag{36}$$

Substituting (2), (4), (24) and (36) into (23) yields (32). Similarly, $\overline{D}_h^{L|pr}$ can also be derived.

Notice that the mean delay of a packet is jointly determined by the traffic load, moments of transmission time, and coefficient ratio τ_h/τ_m. We are interested in the influence caused by the priority policy. Considering an extreme case where the coefficient ratio $\tau_h/\tau_m \to 0$, this case means that the priority of the cMTC packets is strictly higher than HTC packets. This result in the Head-of-Line (HoL) priority scheme. We can easily derive the delay performance of HoL scheme by taking $\tau_h/\tau_m = 0$.

Corollary 2. In the extreme case of HoL priority with non-preemptive policy, the mean delays of cMTC and HTC packets are given by

$$\overline{D}_m^{L|np'} = \frac{\rho_m\frac{\hat{\beta}_m^L}{\overline{\beta}_m^L}+\rho_h\frac{\hat{\beta}_h^L}{\overline{\beta}_h^L}}{2(1-\rho_m)}+\overline{\beta}_m^L, \tag{37}$$

$$\overline{D}_h^{L|np'} = \frac{\rho_m \frac{\hat{\beta}_m^L}{\overline{\beta}_m^L} + \rho_h \frac{\hat{\beta}_h^L}{\overline{\beta}_h^L}}{2(1-\rho_m)(1-\rho_m-\rho_h)} + \overline{\beta}_h^L. \tag{38}$$

Proof. Substituting $\tau_h/\tau_m = 0$ into (32) and (33) yields (37) and (38).

4.2.2 *Time varying priority with preemptive-resume policy*

Under the preemptive-resume policy, upon the arrival of a cMTC packet, if there is an HTC device receiving transmission from BS, the HTC transmission is interrupted by the cMTC packet until the queue is empty for cMTC. Then the HTC packet continues transmission from the point where it was interrupted. The mean delay performance of cMTC and HTC packets are given by the following two propositions, respectively.

Proposition 4. For time varying priority with preemptive-resume policy, the mean delays of cMTC and HTC service are given by

$$\begin{aligned}\overline{D}_m^{L|np} = &\left[\left(1-\rho_h-\rho_m\left(1-\frac{\tau_h}{\tau_m}\right)+\rho_h\frac{\tau_h}{\tau_m}\right)\rho_m\frac{\hat{\beta}_m^L}{2\overline{\beta}_m^L}+\frac{\tau_h}{\tau_m}\rho_h\frac{\hat{\beta}_h^L}{2\overline{\beta}_h^L}\right.\\ &\left.+\left(1-\rho_h-\rho_m\left(1-\frac{\tau_h}{\tau_m}\right)\right)\rho_h\frac{\tau_h}{\tau_m}\left(\frac{\overline{\beta}_h^L}{1-\rho_m(1-\tau_h/\tau_m)}\right)\right]\\ &\Bigg/\left[(1-\rho_h-\rho_m)\left(1-\rho_m\left(1-\frac{\tau_h}{\tau_m}\right)\right)\right]+\overline{\beta}_m^L,\end{aligned} \tag{39}$$

$$\overline{D}_h^{L|pr} = \frac{\rho_m\frac{\hat{\beta}_m^L}{2\overline{\beta}_m^L}+(1-\rho_m)\rho_h\frac{\hat{\beta}_h^L}{2\overline{\beta}_h^L}+\rho_m\rho_h\frac{\tau_h}{\tau_m}\left(\frac{\overline{\beta}_h^L}{1-\rho_m(1-\frac{\tau_h}{\tau_m})}\right)}{(1-\rho_h-\rho_m)(1-\rho_m(1-\frac{\tau_h}{\tau_m}))}+\frac{\overline{\beta}_h^L}{1-\rho_m(1-\frac{\tau_h}{\tau_m})}, \tag{40}$$

respectively. The proof of Proposition 4 is given in Appendix A.

Corollary 3. In the extreme case of HoL priority with preemptive-resume policy, the mean delays of cMTC and HTC service are given by

$$\overline{D}_m^{L|pu'} = \frac{\rho_m\frac{\hat{\beta}_m^L}{\overline{\beta}_m^L}}{2(1-\rho_m)}+\overline{\beta}_m^L, \tag{41}$$

$$\overline{D}_h^{L|pu'} = \frac{\rho_m\frac{\hat{\beta}_m^L}{\overline{\beta}_m^L}+(1-\rho_m)\rho_h\frac{\hat{\beta}_h^L}{\overline{\beta}_h^L}}{2(1-\rho_m)(1-\rho_m-\rho_h)}+\frac{\overline{\beta}_h^L}{1-\rho_m}. \tag{42}$$

Proof. Substituting $\tau_h/\tau_m = 0$ into (39) and (40) yields (41) and (42).

4.2.3 *Throughput performance*

Proposition 5. In the case of load division, the mean throughput of a randomly selected cMTC user is given by

$$\overline{C}_m^L = \frac{L_m\rho_m}{\overline{\beta}_m^L}\sum_{n=1}^{\infty}\frac{1}{x}f_n(n). \tag{43}$$

Proof. Consider an arbitrary cell with N_m cMTC devices. The throughput per device is given by

$$\overline{C}_{m|N_m}^L = \frac{L_m}{\overline{\alpha}_m}. \tag{44}$$

Substituting (2) into (44) yields

$$\overline{C}_{m|N_m}^L = \frac{L_m\rho_m}{N_m\overline{\beta}_m^L}. \tag{45}$$

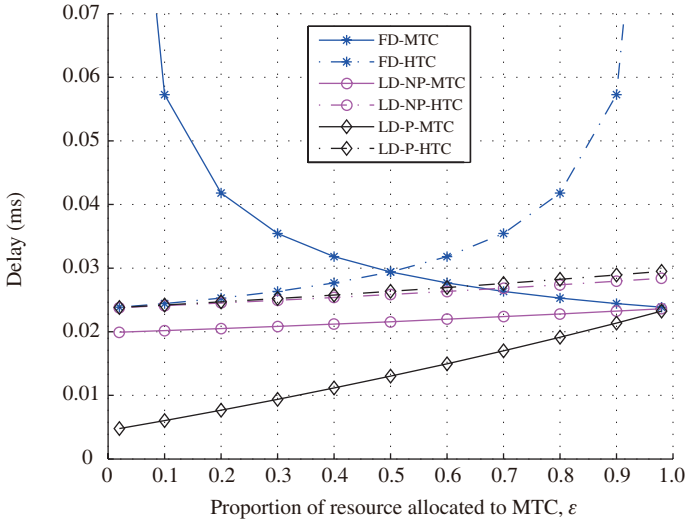

Figure 4 (Color online) Mean delay as a function of the resource partition parameter ϵ ($\rho_m + \rho_h = 0.2$).

The mean throughput per user, taking average over the random variable N_m, can then be easily derived as (44).

The throughput of HTC service can be obtained in a similar fashion. Comparing (28) with (43), we can find that the difference between $\overline{\beta}_m^F$ and $\overline{\beta}_m^L$ determines the difference in throughput. We note that the former is calculated with partial system bandwidth, while the latter is calculated with full system bandwidth.

5 Numerical results and discussions

This section presents numerical results and discusses their implications. For simplicity, we consider an interference-limited system, which means the accumulated interference is much higher than the noise to justify the use of (13). Without loss of generality, the densities of BSs, MTC users, and HTC users are set to be $\lambda_b = 10^{-6}/m^2$, $\lambda_m = 10^{-5}/m^2$, and $\lambda_h = 5 \times 10^{-6}/m^2$, respectively. The packet length $L_m = L_h = 20$ kbits, the system bandwidth $W = 20$ MHz.

According to (25), (26), (37), (38), (41) and (42), Figures 4–6 illustrates the mean packet delay as a function of the resource partition parameter ϵ with different resource allocation policies. Figures 4–6 corresponds to the cases of low traffic load ($\rho_m + \rho_h = 0.2$), medium traffic load ($\rho_m + \rho_h = 0.5$), and high traffic load ($\rho_m + \rho_h = 0.8$), respectively. The abbreviation 'FD' and 'LD' denotes frequency division and load division, respectively; 'NP' and 'P' denotes non-preemptive and preemptive, respectively. For simplicity, we consider the HoL priority, which is an extreme case of the time dependent priority with $\tau_h/\tau_m = 0$. We can see that in all three cases, the 'LD-P' policy gives the lowest MTC delay, followed by 'LD-NP' and 'FD' policies. This indicates that the LD scheme outperforms the FD scheme in providing guaranteed better performance for MTC delay. However, for HTC delay performance, the relative performance of LD and FD schemes depends on the total traffic load. It can be observed that LD outperforms FD with low traffic low. But the LD delay gradually becomes worse than that of FD when the traffic load increases. In the case of high traffic load, the LD becomes constantly worse than FD for HTC delay.

According to (25), (26), (37), (38), (41) and (42), Figure 7 shows the relationship between cMTC delay and HTC delay, under different conditions of traffic load. The curve is obtained by varying the value of ϵ from 0.02 to 0.98. It is observed that the FD scheme yields a typical tradeoff curve, such that improving the cMTC delay performance comes at a cost of degrading the performance of coexisting HTC. Moreover, the tradeoff curve has a sharp turn, indicating that the cMTC and HTC performance should be jointly optimized. For the LD schemes, the delay performance of cMTC and HTC is no longer

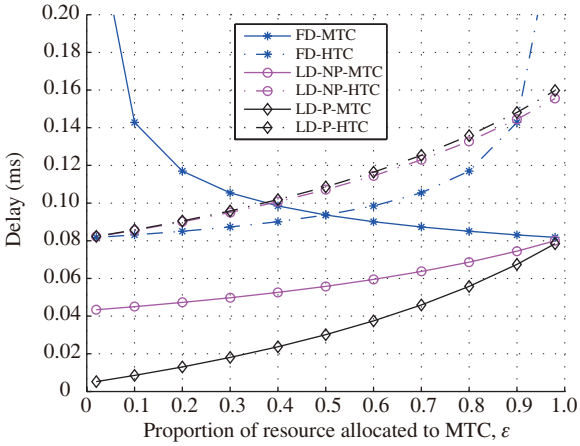

Figure 5 (Color online) Mean delay as a function of the resource partition parameter ϵ ($\rho_m + \rho_h = 0.5$).

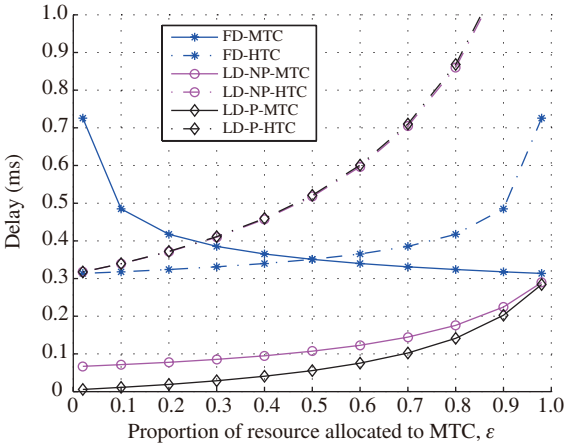

Figure 6 (Color online) Mean delay as a function of the resource partition parameter ϵ ($\rho_m + \rho_h = 0.8$).

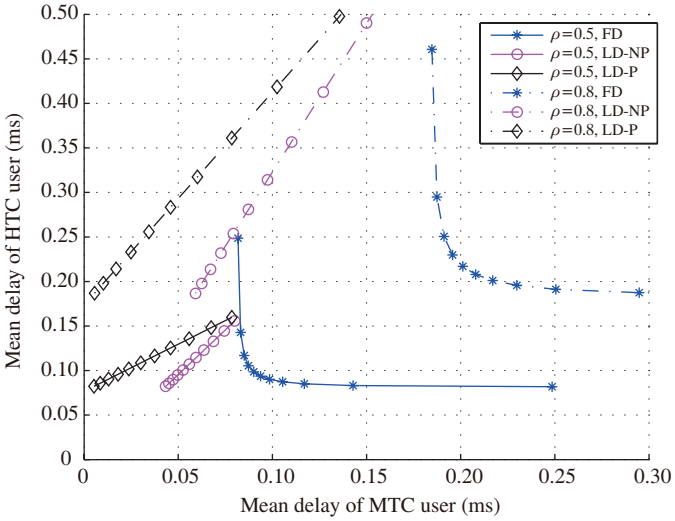

Figure 7 (Color online) Tradeoff between cMTC delay and HTC delay under different conditions of traffic load.

a tradeoff. Instead, a close-to-linear, positively correlated relationship is observed. This implies that choosing a proper resource partition parameter ϵ is important for the LD scheme.

The results we presented so far assume the case of HoL policy, which is the extreme case of time varying priority with $\tau_h/\tau_m = 0$. More generally, the ratio of time-dependent coefficients τ_h/τ_m can vary between 0 to 1. The case of $\tau_h/\tau_m = 1$ means that both classes of traffic have the same priority. According to (32) and (33), Figure 8 illustrates the impact of τ_h/τ_m on the delay performance. It can be observed that when the ratio becomes large, the delays of HTC decrease while the delays of cMTC increase gradually. Therefore, by choosing proper values for τ_h/τ_m, we can have a smooth transition between absolute priority and absolute fairness for cMTC and HTC traffic.

In practice, the MTC traffic can be dominated by short packets. It is desirable to understand the impact of small cMTC packet length on the delay performance. According to (25), (26), (32), (33), (39) and (40), Figure 9 shows the mean packet delay with $L_h = 20$ kbits and $L_m = 200$ bits, under different resource allocation policies. Compared Figure 9 with Figure 5, it can be observed that a smaller packet length of cMTC traffic helps to reduce the delay of both HTC and cMTC traffic under all resource allocation policies. However, it should be noted that smaller packet size is typically associated with larger overhead, which is an aspect not addressed in this paper.

Finally, according to (28) and (43), Figure 10 shows the relationship between the throughput capacities for cMTC and HTC. The curve is obtained by varying the value of ϵ from 0.02 to 0.98. It is observed

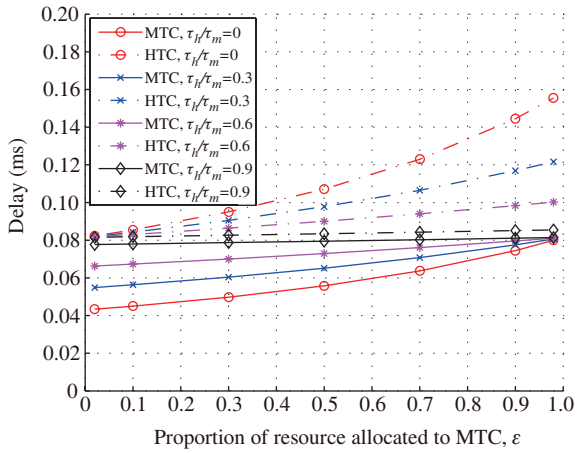

Figure 8 (Color online) Mean delay as a function of the resource partition parameter ϵ under time varying priority with different values of τ_h/τ_m ($\rho_m + \rho_h = 0.5$).

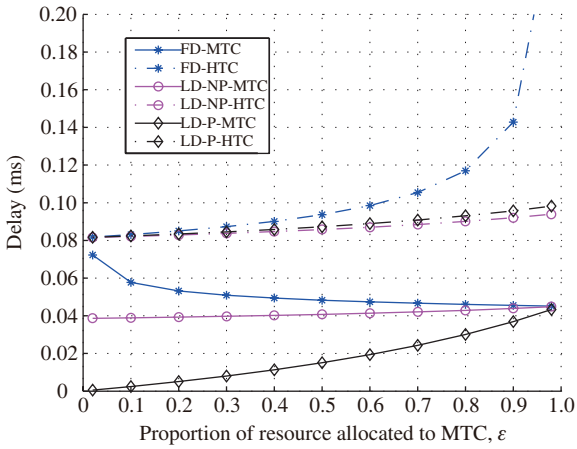

Figure 9 (Color online) Mean delay as a function of the resource partition parameter ϵ with small packet size of cMTC services ($\rho_m + \rho_h = 0.5$, $L_h = 20$ kbits, $L_m = 200$ bits).

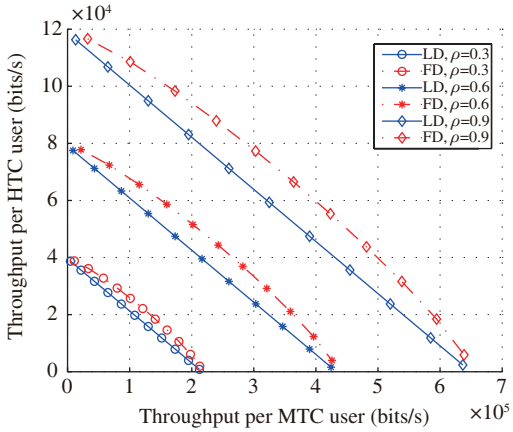

Figure 10 (Color online) Tradeoff between cMTC throughput and HTC throughput under different conditions of traffic load.

that the LD scheme gives a linear throughput tradeoff. The FD scheme yields a close-to-linear tradeoff curve, which is concave and slightly better than the throughput of LD. Figure 8 implies that the LD performance gains with respect to delay comes at a cost of slightly degraded throughput performance.

Summarizing the above results, we conclude that the proposed LD scheme can trade throughput performance for delay performance. When the total traffic load is relatively low, both the delay performances of cMTC and HTC are improved. However, when the total traffic load becomes high, the LD scheme further trades the delay of HTC traffic to make sure that the delay performance of cMTC is optimized.

6 Conclusion

This paper has investigated the coexistence problem of mixed cMTC and HTC traffic in a large scale cellular network. A service-level resource allocation scheme called load division has been proposed. The delay and throughput performance of the proposed scheme has been analyzed under a new analytical framework, which integrates queuing models and stochastic geometric models to jointly capture the spatial and temporal behavior of the network. Numerical results have shown that compared with frequency division, load division yields a guaranteed better performance for the mean cMTC service delay, a mixed performance (depending on the total traffic load) for the HTC service delay, and a slightly worse perfor-

mance for cMTC and HTC throughput. We conclude that the load division scheme is well-suited for the coexistence of delay-sensitive services.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61571378), EU H2020 RISE TESTBED Project (Grant No. 734325), EU FP7 QUICK Project (Grant No. PIRSES-GA-2013-612652), and EPSRC TOUCAN Project (Grant No. EP/L020009/1).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Agiwal M, Roy A, Saxena N. Next generation 5G wireless networks: a comprehensive survey. IEEE Commun Surv Tutor, 2016, 18: 1617–1655
2 Wang C X, Haider F, Gao X, et al. Cellular architecture and key technologies for 5G wireless communication networks. IEEE Commun Mag, 2014, 52: 122–130
3 Ge X H, Chen J Q, Wang C X, et al. 5G green cellular networks considering power allocation schemes. Sci China Inf Sci, 2016, 59: 022308
4 Patcharamaneepakorn P, Wu S, Wang C X, et al. Spectral, energy and economic efficiency of 5G multi-cell massive MIMO systems with generalized spatial modulation. IEEE Trans Veh Technol, 2016, 65: 9715–9731
5 Ge X H, Tu S, Mao G Q, et al. 5G ultra-dense cellular networks. IEEE Wirel Commun, 2016, 23: 72–79
6 Ma X, Sheng M, Li J D, et al. Concurrent transmission for energy efficiency of user equipment in 5G wireless communication networks. Sci China Inf Sci, 2016, 59: 022306
7 Ma Z, Zhang Z Q, Ding Z G, et al. Key techniques for 5G wireless communications:network architecture, physical layer, and MAC layer perspectives. Sci China Inf Sci, 2015, 58: 041301
8 Cao J Y, Zhang Y, An W, et al. VNF-FG design and VNF placement for 5G mobile networks. Sci China Inf Sci, 2017, 60: 040302
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 Ratasuk R, Mangalvedhe N, Ghosh A. Overview of LTE enhancements for cellular IoT. In: Proceedings of IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong, 2015. 2293–2297
11 Raza U, Kulkarni P, Sooriyabandara M. Low power wide area networks: an overview. IEEE Commun Surv Tutor, 2017, 19: 855–873
12 Gozalvez J. New 3GPP standard for IoT [mobile radio]. IEEE Veh Technol Mag, 2016, 11: 14–20
13 Ashraf S A, Aktas I, Eriksson E, et al. Ultra-reliable and low-latency communication for wireless factory automation: from LTE to 5G. In: Proceedings of IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), Berlin, 2016. 1–8
14 Ratasuk R, Prasad A, Li Z, et al. Recent advancements in M2M communications in 4G networks and evolution towards 5G. In: Proceedings of 18th International Conference on Intelligence in Next Generation Networks, Paris, 2015. 52–57
15 Shariatmadari H, Ratasuk R, Iraji S, et al. Machine-type communications: current status and future perspectives toward 5G systems. IEEE Commun Mag, 2015, 53: 10–17
16 Hossain M I, Azari A, Zander J. DERA: augmented random access for cellular networks with dense H2H-MTC mixed traffic. In: Proceedings of IEEE Globecom Workshops (GC Wkshps), Washington, 2016. 1–7
17 Aijaz A, Tshangini M, Nakhai M R, et al. Energy-efficient uplink resource allocation in LTE networks with M2M/H2H co-existence under statistical QoS guarantees. IEEE Trans Commun, 2014, 62: 2353–2365
18 Niyato D, Wang P, Kim D I. Performance modeling and analysis of heterogeneous machine type communications. IEEE Trans Wirel Commun, 2014, 13: 2836–2849
19 Tian H, Xu Y Y, Xu K, et al. Energy-efficient user association in heterogeneous networks with M2M/H2H coexistence under QoS guarantees. China Commun, 2015, 12: 93–103
20 Aijaz A, Aghvami A H. PRMA-based cognitive machine-to-machine communications in smart grid networks. IEEE Trans Veh Technol, 2015, 64: 3608–3623
21 Hamdoun S, Rachedi A, Ghamri-Doudane Y. Radio resource sharing for MTC in LTE-A: an interference-aware bipartite graph approach. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), San Diego, 2015. 1–7
22 Hamdoun S, Rachedi A, Ghamri-Doudane Y. A flexible M2M radio resource sharing scheme in LTE networks within an H2H/M2M coexistence scenario. In: Proceedings of IEEE International Conference on Communications (ICC), Kuala Lumpur, 2016. 1–7
23 Azari A, Miao G. Battery lifetime-aware base station sleeping control with M2M/H2H coexistence. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), Washington, 2016. 1–6
24 Tian H, Xie W, Gan X Y, et al. Hybrid user association for maximising energy efficiency in heterogeneous networks with human-to-human/machine-to-machine coexistence. IET Commun, 2016, 10: 1035–1043
25 Lien S Y, Cheng S M, Shih S Y, et al. Radio resource management for QoS guarantees in cyber-physical systems. IEEE Trans Parall Distr Syst, 2012, 23: 1752–1761

26 Tian H, Xu L L, Pei Y S, et al. Power ramping schemes for M2M and H2H co-existing scenario. China Commun, 2013, 10: 100–113
27 Zheng K, Ou S, Alonso-Zarate J. Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications. IEEE Wirel Commun, 2014, 21: 12–18
28 Ali M S, Hossain E, Kim D I. LTE/LTE-A random access for massive machine-type communications in smart cities. IEEE Commun Mag, 2017, 55: 76–83
29 3GPP TR 45.820. Cellular system support for ultra low complexity and low throughput Internet of Things. http://www.3gpp.org/ftp/Specs/archive/45series/45.820/45820-d10.zip
30 Ericsson and Nokia Networks. Further LTE physical layer enhancements for MTC. http://www.3gpp.org/ftp/tsgran/tsgran/TSGR65/Docs/RP-141660.zip
31 Lauridsen M, Kovacs I Z, Mogensen P. Coverage and capacity analysis of LTE-M and NB-IoT in a rural area. In: Proceedings of IEEE 84th Vehicular Technology Conference (VTC-Fall), Montreal, 2016. 1–5
32 Wang Y P E, Lin X, Adhikary A. A primer on 3GPP narrowband Internet of Things. IEEE Commun Mag, 2017, 55: 117–123
33 Ratasuk R, Vejlgaard B, Mangalvedhe N. NB-IoT system for M2M communication. In: Proceedings of IEEE Wireless Communications and Networking Conference, Doha, 2016. 1–5
34 Ratasuk R, Mangalvedhe N, Zhang Y. Overview of narrowband IoT in LTE Rel-13. In: Proceedings of IEEE Conference on Standards for Communications and Networking (CSCN), Berlin, 2016. 1–7
35 Yu C S, Yu L, Wu Y. Uplink scheduling and link adaptation for narrowband Internet of Things systems. IEEE Access, 2017, 5: 1724–1734
36 Stuckmann P. The GSM Evolution: Mobile Packet Data Services. Chichester: Wiley, 2002
37 Erik D, Stefan P, Johan S. 4G - LTE/LTE-advanced for mobile broadband. Burlington: Elsevie, 2011
38 Shariatmadari H, Iraji S, Jantti R. Analysis of transmission methods for ultra-reliable communications. In: Proceedings of IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong, 2015. 2303–2308
39 Udesh O, Furqan A, Olav T. Ultra-reliable link adaptation for downlink MISO transmission in 5G cellular networks. Information, 2016, 7: 14
40 Yilmaz O N C, Wang Y P E, Johansson N A. Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case. In: Proceedings of IEEE International Conference on Communication Workshop (ICCW), London, 2015. 1190–1195
41 Johansson N A, Wang N A, Eriksson E. Radio access for ultra-reliable and low-latency 5G communications. In: Proceedings of IEEE International Conference on Communication Workshop (ICCW), London, 2015. 1184–1189
42 Durisi G, Koch T, Ostman J. Short-packet communications over multiple-antenna Rayleigh-fading channels. IEEE Trans Commun, 2016, 64: 618–629
43 Durisi G, Koch T, Popovski P. Toward massive, ultrareliable, and low-latency wireless communication with short packets. Proc IEEE, 2016, 104: 1711–1726
44 Farayev B, Ergen S C. Towards ultra-reliable M2M communication: scheduling policies in fading channels. In: Proceedings of 23rd International Conference on Telecommunications (ICT), Thessaloniki, 2016. 1–6
45 Singh B, Li Z, Tirkkonen O. Ultra-reliable communication in a factory environment for 5G wireless networks: link level and deployment study. In: Proceedings of IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Valencia, 2016. 1–5
46 Hou I-H, Borkar V, Kumar P R. A theory of QoS for wireless. In: Proceedings of the 2009 IEEE INFOCOM, Rio de Janeiro, 2009. 486–494
47 Lashgari S, Avestimehr A S. Timely throughput of heterogeneous wireless networks: fundamental limits and algorithms. IEEE Trans Inf Theory, 2013, 59: 8414–8433
48 Zhang G Z, Quek T Q S, Huang A P, et al. Delay and reliability tradeoffs in heterogeneous cellular networks. IEEE Trans Wirel Commun, 2016, 15: 1101–1113
49 Yu S M, Kim S L. Downlink capacity and base station density in cellular networks. In: Proceedings of 11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), Tsukuba Science City, 2013. 119–124
50 Andrews J G, Baccelli F, Ganti R K. A tractable approach to coverage and rate in cellular networks. IEEE Trans Commun, 2011, 59: 3122–3134
51 Kleinrock L. Queueing Systems Volume II: Computer Applications. New York: Wiley, 1976

Appendix A Proof of Proposition 4

First, we consider a typical level p user in a time varying priority queue ($p \in (1, \ldots, P)$, level 1 has the highest priority). Denote j as a level has higher priority coefficienta level that has a higher priority than p ($j < p$). The mean number of level j users that come after and interrupt the p level users' transmission is given by

$$\overline{K}_{pj} = \lambda_j \overline{T}_p (1 - \frac{\tau_p}{\tau_j}). \tag{A1}$$

It follows that the probability for level j users to interrupt the transmission of a level p user is given by

$$\phi_{pj} = \frac{\overline{K}_{pj}}{\lambda_j \overline{T}_p} = 1 - \frac{\tau_p}{\tau_j}. \tag{A2}$$

Therefore, the mean transmission delay of the level p user is given by

$$\overline{T}_p = \overline{x}_p + \sum_{i=1}^{p-1} \overline{x}_p \overline{K}_{pi} = \overline{x}_p + \sum_{i=1}^{p-1} \rho_i \overline{T}_p \left(1 - \frac{\tau_p}{\tau_i}\right). \tag{A3}$$

Eq. (A3) can be further simplified into

$$\overline{T}_p = \frac{\overline{x}_p}{1 - \sum_{i=1}^{p-1} \rho_i (1 - \frac{\tau_p}{\tau_i})}. \tag{A4}$$

Then, we consider the queue delay of the level p user. As stated in [51], the queue delay consists of three parts: 1) delay caused by the user in service upon arrival; 2) delay caused by the user already in queue upon arrival; 3) delay caused by the user arrival after. The latter two parts are derived in the book, here we concentrate on the first part. Denote l as a level that has lower priority than p ($l > p$). Upon the arrival of the level p user, a level l user is receiving transmission. The delay caused by this level l includes two cases.

Case 1: The whole residual life of the level l user, which means the p level user cannot interrupt the l level user. The probability of this case is

$$P_{pl}^1 = \frac{\tau_l}{\tau_p}. \tag{A5}$$

Case 2: Partial of the residual life of the level l user, which means the p level user interrupts the l level user. The probability of this case is

$$P_{pl}^2 = 1 - \frac{\tau_i}{\tau_p}. \tag{A6}$$

The delay caused by this case is the catch up time of the p level user to the l level user, which is $\frac{\tau_l}{\tau_p}\overline{T}_l$. Therefore, the mean delay caused by the user in service upon arrival is given by

$$\overline{W}_p^r = \sum_{i=1}^{p} \rho_i \overline{R}_i + \sum_{i=p+1}^{P} \rho_i \frac{\tau_i}{\tau_p} \overline{T}_i, \tag{A7}$$

where $\overline{R}_i$ denotes the mean residual life of a level i user, $\overline{R}_i = \frac{E({x_i}^2)}{2\overline{x}_i}$. Therefore, the mean queueing delay of a level p user is given by

$$\overline{W}_p = \frac{\overline{W}_p^r + \sum_{i=1}^{p-1} \rho_i \overline{W}_i + \sum_{i=p+1}^{P} \rho_i \overline{W}_i \frac{\tau_i}{\tau_p}}{1 - \sum_{i=1}^{p} \rho_i (1 - \frac{\tau_p}{\tau_i})}. \tag{A8}$$