# Performance Approximations for Non-real-time Traffic in an Integrated Service System

Yue-Cai Huang*, King-Tim Ko*, and Moshe Zukerman*
* Department of Electronic Engineering, City University of Hong Kong, Hong Kong SAR, P. R. China
Email: yuechuang2-c@my.cityu.edu.hk; {eektko; m.zu}@cityu.edu.hk

***Abstract*—We consider an integrated service system, where real-time (RT) calls of multiple rate requirements and non-real-time (NRT) calls share the total system capacity. Typically, RT calls are given strict priority over NRT calls, therefore, NRT performance is dependent on the RT process. When RT call arrival processes are not Poisson processes, the effect of the RT traffic burstiness on the NRT performance has not been investigated. In this paper, we investigate the effect and provide a computationally efficient approximation for the NRT performance evaluation in the integrated service system with multi-rate non-Poisson RT traffic. With known first and second moments of the RT call arrival processes, we propose to consider the multi-rate non-Poisson RT traffic streams as an equivalent single-rate Poisson traffic stream. Then we evaluated the NRT performance by converting the original system to a system offered with the equivalent RT traffic and the NRT traffic. Our approximation is validated with extensive numerical examples.**

***Index Terms*—Integrated service, quality of service, dimensioning, real-time, non-real-time, MERM, non-Poisson, multi service**

## I. Introduction

Traffic generated by various end-user applications in current and emerging communication networks can be broadly categorized into real-time (RT) traffic and non-real-time (NRT) traffic according to their traffic nature and Quality-of-Service (QoS) requirements [1]. Generally, RT traffic is generated by real-time applications such as voice and video. They are delay sensitive and therefore require guaranteed data rate to meet their QoS requirements. NRT traffic represents non-real-time applications, such as web-browsing, email and downloads. They are elastic and can be served by time-varying data rates.

To efficiently support both RT and NRT traffic with QoS assurance, integrated service systems [2]–[7] were considered where different traffic characteristics and QoS requirements of RT and NRT traffic are exploited in the resource allocation. Generally, RT traffic is given higher priority over the NRT traffic. Each admitted RT call is served by a guaranteed constant data rate according to its requirement, and the admitted NRT calls share evenly the remaining capacity. The total capacity available for the RT calls and the maximum numbers of concurrently served NRT calls are limited, so that certain minimum capacity can be guaranteed to each NRT call.

Performance evaluation of such an integrated service system is important for network design and dimensioning. In previous publications, RT and NRT call arrivals are often assumed to follow Poisson processes. Under this assumption, performance of the RT traffic can be evaluated using known multi-rate loss model [8]–[13]. Performance evaluation for the NRT traffic is challenging because the scheduling introduces dependencies of the NRT process on the RT process. Besides the time-consuming exact Markov chain solution, computationally efficient approximations were proposed [2], [5]–[7], [14]–[17].

In some practical cases, the call arrival processes cannot be accurately modeled by Poisson processes. A typical example is the overflow traffic, i.e., traffic blocked by one trunk and redirected to an alternative trunk for possible service, which is known to be more bursty than Poisson traffic [18], [19]. The effect of the burstiness has been extensively studied for circuit-switching telephony systems in [18]–[24], and the Poisson assumption was found to underestimate the blocking probabilities in the alternative trunk. These studies can give the performance of RT traffic in our considered system. However, processor-sharing for the NRT traffic was not included.

In this paper, we provide performance evaluation approximations for the NRT traffic in an integrated service system with multiple classes of non-Poisson RT traffic. We use the concept of the Hayward approximation [19] and consider the multiple classes of non-Poisson RT traffic streams as an equivalent single class Poisson traffic stream. This consideration makes the NRT performance analysis feasible. Our contribution here is to demonstrate that the multi-rate Hayward approximation concept which has been widely applied for performance analysis for RT traffic can also give accurate performance for NRT traffic.

## II. The integrated service system

We consider an integrated service system with total capacity $C$ b/s supporting $K$ classes of RT calls and a single class of NRT calls. Assume that different classes of RT calls arrive at the system independently with each other. Let $\lambda_k$ be the arrival rate of the class $k$ RT calls. Assume that the means and the variances (definitions given in Section III) of the RT traffic streams are known. Define an RT channel as a portion of capacity with the data rate $c_{rt}$ b/s. Each admitted RT call of class $k$ requires and is allocated capacity of $d_k$ RT channels for its entire service time. Then $\mathbf{d} = [d_1, d_2, \ldots, d_K]^T$ represents the channel requirements for all the RT classes.

Denote by $n_k(t)$ the number of class $k$ RT calls being served in the system at time $t$ and define $\mathbf{n}(t) = [n_1(t), n_2(t), \ldots, n_K(t)]$. Then, the total number of RT channels occupied at time $t$ is given by $\mathbf{n}(t) \cdot \mathbf{d}$. RT calls are served in strict priority over NRT calls. To avoid starvation

of the NRT calls, the RT calls can only occupy up to $N_{rt}$ RT channels, where $N_{rt} \cdot c_{rt} \leq C$. For an RT call of class $k$ arriving at time $t$, it is admitted and is served at a data rate of $d_k c_{rt}$ b/s if $\mathbf{n}(t) \cdot \mathbf{d} + d_k \leq N_{rt}$; otherwise, it is blocked.

The call arrival process of the NRT traffic is assumed to be independent of the RT traffic and is assumed to follow a Poisson process. Denote by $\lambda_{nrt}$ the NRT call arrival rate. The NRT calls share evenly the remaining capacity left over by the RT calls according to the processor sharing policy. Denote by $n_{nrt}(t)$ the number of the NRT calls in the system at time $t$. The data rate of all the NRT calls and the data rate of an individual NRT call at time $t$ are thus given by,

$$C_{nrt}(t) = C - (\mathbf{n}(t) \cdot \mathbf{d})\, c_{rt}, \tag{1}$$

and

$$c_{nrt}(t) = C_{nrt}(t)/n_{nrt}(t), \tag{2}$$

respectively.

The data rate of an individual NRT call, $c_{nrt}(t)$, is updated upon an RT/NRT admitted arrival or departure. As mentioned, to satisfy the QoS of the admitted NRT traffic, the maximum number of concurrently served NRT calls is limited to $N_{nrt}$. Accordingly, a new NRT call arriving at time $t$ is admitted if $n_{nrt}(t) < N_{nrt}$ and is blocked if $n_{nrt}(t) = N_{nrt}$.

Since the RT traffic has strict priority over the NRT traffic, the performance of the RT traffic can be evaluated independently as if the NRT traffic does not exist, using methods in existing work [8]–[13], [18], [19], [23], [24].

The performance of the NRT traffic will be affected by the RT processes. When the RT call arrivals follow Poisson processes, exact Markov chain solution or approximation methods mentioned in Section I can be used for NRT performance evaluation. However, the case that the RT call arrivals do not follow Poisson processes has not been studied before, and this paper provides a new performance evaluation method to address this scenario.

## III. Equivalent traffic streams

As mentioned, our proposed approximation for NRT performance evaluation is based on considering the multi-rate non-Poisson RT traffic streams as an equivalent single-rate Poisson traffic stream. To better present our method, we first discuss the equivalence of traffic streams in this section.

A non-Poisson RT traffic stream is often characterized by the *customer distribution* if the traffic were offered to a trunk of infinite servers. Consider an G/G/$\infty$ queueing system. Denote by $\lambda$ and $\mu$, customer arrival rate and service rate, respectively. Let $N_c$ be the number of customers in the system. Denote by $\mathbf{E}[N_c]$ and $\mathbf{V}[N_c]$ the mean and the variance of $N_c$. The peakedness is defined as,

$$\mathbf{Z}[N_c] = \mathbf{V}[N_c]/\mathbf{E}[N_c]. \tag{3}$$

According to Little's formula, $\mathbf{E}[N_c] = \lambda/\mu$. For a Poisson traffic stream, $\mathbf{V}[N_c] = \mathbf{E}[N_c]$, i.e., $\mathbf{Z}[N_c] = 1$. For a non-Poisson traffic stream, typically, $\mathbf{V}[N_c] \neq \mathbf{E}[N_c]$. As the distribution of $N_c$ is often used to characterize the input traffic stream, we call $\mathbf{E}[N_c]$, $\mathbf{V}[N_c]$ and $\mathbf{Z}[N_c]$, the mean, the variance and the peakedness of the input traffic stream.

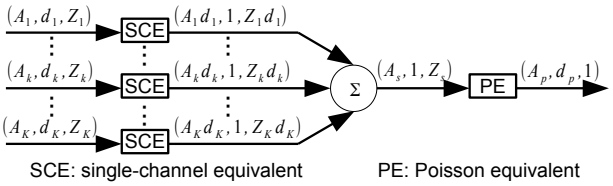



Fig. 1. Procedures to obtain the equivalent single class Poisson traffic stream of multiple classes non-Poisson traffic streams.

Assume that each customer occupies $d$ servers during its entire service time. Let $N_s$ be the number of occupied servers in the system. Then, we have $N_s = N_c d$, and we obtain,

$$\mathbf{E}[N_s] = d\mathbf{E}[N_c],\ \mathbf{V}[N_s] = d^2\mathbf{V}[N_c]. \tag{4}$$

Use $(A, d, Z)$ to represent a traffic stream, where $A$, $d$, and $Z$ denote the offered load ($A = \lambda/\mu$), the number of servers occupied by one customer during its entire service time, and the peakedness, respectively. We have the following Lemma.

**Lemma 1** *Consider traffic streams $\mathfrak{S}_1$ and $\mathfrak{S}_2$ represented by $(A, d, Z)$ and $(Ad, 1, Zd)$, respectively. Let $N_{s1}$ ($N_{s2}$) be the number of occupied servers if $\mathfrak{S}_1$ ($\mathfrak{S}_2$) were offered to a trunk of infinite number of servers. We have,*

$$\mathbf{E}[N_{s1}] = \mathbf{E}[N_{s2}],\ \textit{and}\ \mathbf{V}[N_{s1}] = \mathbf{V}[N_{s2}]. \tag{5}$$

Lemma 1 can be proved using Eq. (3) and Eq. (4). We call a traffic stream *single-channel* if each customer requires a single server during its entire service time ($d = 1$) and *non-single-channel* otherwise ($d \neq 1$). We call the traffic stream $\mathfrak{S}_2$ the *single-channel equivalent* of the traffic stream $\mathfrak{S}_1$. When $Z = 1$, we call the traffic stream $\mathfrak{S}_1$ the *Poisson equivalent* of the traffic stream $\mathfrak{S}_2$. We notice that Lemma 1 complements and enhances (19) in [24]. The equivalence of the traffic streams $\mathfrak{S}_1$ and $\mathfrak{S}_2$ is indicated by (19) in [24] while Lemma 1 proves it in terms of two moments matching.

## IV. NRT performance approximation

Lemma 1 enables us to find an equivalent single class Poisson traffic stream for the multiple classes of non-Poisson RT traffic streams. Then the NRT performance can be evaluated by converting the original system to a system offered with the equivalent RT traffic and the NRT traffic.

### A. Obtain an equivalent RT traffic stream

The equivalent single class Poisson traffic stream can be obtained in the following three steps (illustrated in Fig. 1).

**Step 1**. Find the single-channel equivalents of all the RT traffic streams. For the RT traffic stream of class $k$, define $A_k = \lambda_k/\mu_k$ and denote by $Z_k$ the peakedness of this RT traffic. Using Lemma 1, we obtain,

$$(A_k, d_k, Z_k) \sim (A_k d_k, 1, Z_k d_k). \tag{6}$$

**Step 2**. Aggregate all single-channel RT traffic streams to a single traffic stream. The mean and the peakedness (denoted

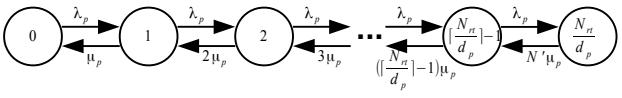



Fig. 2. State transition diagram of the modified RT processes.

by $A_s$ and by $Z_s$) of the aggregated traffic are given by,

$$A_s = \sum_{k=1}^{K} A_k d_k, \text{ and } Z_s = \left( \sum_{k=1}^{K} A_k Z_k d_k^2 \right) / A_s, \quad (7)$$

respectively.

Notice that, in Step 1, the offered load of the single-channel equivalent is $d_k$ times of the original RT traffic stream of class $k$. This can be regarded that the call arrival rate of the single-channel equivalent is $d_k$ times of its original RT traffic while the service rates of the two traffic streams are the same. In Step 2, as different RT traffic streams may have different service rates, the average service time of the aggregated traffic should be re-calculated. Assume that the service times of all RT classes follow exponential distributions. Then, the service time of the aggregated RT traffic also follows exponential distribution and its average is given by,

$$\frac{1}{\mu_s} = \sum_{k=1}^{K} \frac{\lambda_k d_k}{\sum_{k=1}^{K} \lambda_k d_k} \cdot \frac{1}{\mu_k} = \frac{A_s}{\sum_{k=1}^{K} \lambda_k d_k}. \quad (8)$$

**Step 3**. Find the Poisson equivalent represented by $(A_p, d_p, 1)$ of the aggregated traffic stream $(A_s, 1, Z_s)$. According to Lemma 1, we obtain,

$$A_p = A_s / Z_s, \text{ and } d_p = Z_s. \quad (9)$$

Till now, we have obtained the equivalent single class Poisson traffic stream for the original multiple classes of non-Poisson traffic streams. Next, we evaluate the NRT performance in the integrated service system with the RT traffic stream $(A_p, d_p, 1)$.

### *B. NRT performance evaluation*

Replace the multiple classes of RT traffic streams with traffic stream $(A_p, d_p, 1)$. Then, denote by $n'_{rt}(t)$ the number of RT calls in the system at time $t$ and $\{n'_{rt}(t),\ t \geq 0\}$ is the RT process. The RT process is a Markov process and it evolves among states in set $\Omega = \{0, 1, 2, \ldots, \lceil \frac{N_{rt}}{d_p} \rceil - 1, \frac{N_{rt}}{d_p}\}$. The associated state transition diagram is illustrated in Fig. 2, where $\mu_p = \mu_s$ and $\lambda_p = A_p \mu_p$. When $\frac{N_{rt}}{d_p} \in \mathbb{Z}^+$, $N' = \frac{N_{rt}}{d_p}$; otherwise, the value of $N'$ is chosen so that the steady-state probability of the congestion state $\frac{N_{rt}}{d_p}$ equals the RT call blocking probability, which is obtained from a linear interpolation of the Erlang loss formula or from the continuous Erlang loss formula [25].

Denote by $n'_{nrt}(t)$ the number of NRT calls in the system at time $t$. Then the data rate for all the NRT calls and the data rate of an individual NRT call at time $t$ are given by,

$$C'_{nrt}(t) = C - n'_{rt}(t) d_p c, \quad (10)$$

and

$$c'_{nrt}(t) = C'_{nrt}(t) / n'_{nrt}(t), \quad (11)$$

respectively. Then the NRT blocking probability, the NRT average queue size and the NRT average delay can be obtained from the joint RT and NRT process, $\{(n'_{rt}(t), n_{nrt}(t)),\ t \geq 0\}$, using exact Markov chain solution or existing approximations [2], [5]–[7], [14]–[17] mentioned in Section I.

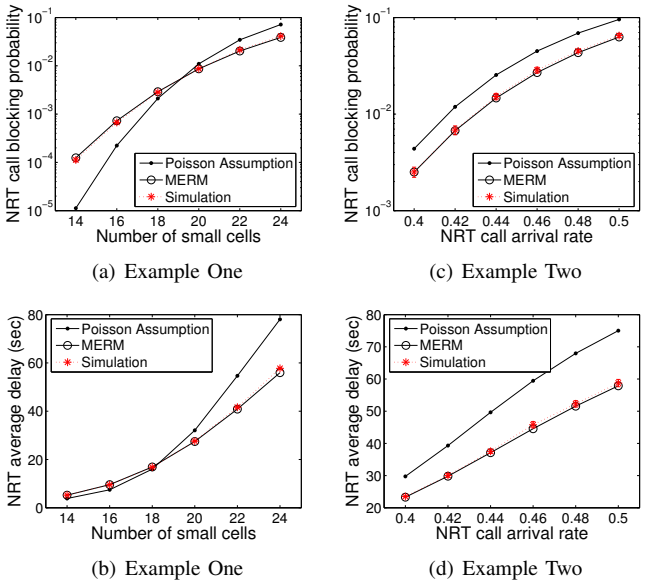
As mentioned, Lemma 1 is a detailed explanation of (19) in [24]. As the method in [24] is called Multi-Service Equivalent Random Method (MERM), we therefore call the method provided here an MERM-based approximation.

## V. Validation by numerical examples

In this section, we validate our proposed MERM-based approximation through numerical examples. We consider an hierarchical system where one macrocell overlays a number of small cells. RT calls blocked by small cells are overflowed to the macrocell. As the overflow traffic is more bursty than Poisson traffic, our examples will demonstrate the performance of our proposed method under non-Poisson inputs.

In our examples, there are in total $N_{small}$ small cells with 16 RT channels each. The small cells are evenly divided into two groups. Each small cell of group 1 (group 2) is offered with class 1 (class 2) RT traffic with arrival rate $\lambda^{(1)} = 0.08$ $\text{s}^{-1}$ ($\lambda^{(2)} = 0.05$ $\text{s}^{-1}$). The RT calls of class 1 and that of class 2 occupy 1 and 2 RT channels during their entire service time, respectively. The RT call arrival processes are independent with each other and they are all Poisson processes. Service times of class 1 and class 2 RT calls all follow exponential distributions with average service times of 180 s. When free channels left in a small cell are not sufficient to serve an incoming RT call, this RT call will be blocked and overflowed to the macrocell immediately. The macrocell is an integrated service system where the input RT traffic is the overflow traffic of small cells. The macrocell has its local Poisson NRT traffic arrivals and the NRT call sizes follow exponential distribution. In the macrocell, parameters are set as follows: $C = 6.4$ Mb/s, $c_{rt} = 64$ kb/s, $N_{rt} = 80$, $N_{nrt} = 50$, $1/\mu_1 = 1/\mu_2 = 180$ s, $d_1 = 1$, $d_2 = 2$, and $L = 500$ kB. We conduct two examples. In Example One, $\lambda_{nrt} = 0.45$ $\text{s}^{-1}$, $N_{small}$ =[14, 16, 18, 20, 22, 24]. Different number of small cells will cause different overflow call arrival rate of RT traffic to the macrocell. In Example Two, $N_{small} = 22$ and $\lambda_{nrt}$ = [0.40, 0.42, 0.44, 0.46, 0.48, 0.50].

The NRT blocking probabilities and the NRT average delay from our proposed method and those from simulations for the two examples are compared in Fig. 3. The mean and the variance of the overflow traffic can be obtained from the Erlang loss formula and Riordan's formula [18]. The 95% confidence intervals are provided for the simulation results. Results from the Poisson Approximation (PA) are also included in the comparisons, where call arrival processes of the overflow traffic are approximated as Poisson processes. It is observed that PA underestimates the results in some cases and overestimates the results in other cases, which implies that the effect of the burstiness of the RT traffic on the NRT performance is not one directional. Therefore, the burstiness of RT traffic should be considered to achieve accurate performance evaluations for




(a) Example One (c) Example Two

(b) Example One (d) Example Two

Fig. 3. Comparisons of the NRT blocking probabilities and NRT average delay obtained from MERM-based approximation, PA and simulations for Example One and Example Two.

the NRT traffic. We can see from Fig. 3 that our proposed approximation is accurate in all cases considered.

Our proposed method can also be applied for the cases when both RT and NRT call arrivals follow Poisson processes. In this case, our method will reduce the computational complexity as it replaces the multiples classes of RT traffic streams to a single type of traffic stream. It is especially useful when the number of RT classes is large. Details can be found in [26].

## VI. Conclusion

We have considered an integrated service system that serves multi-rate RT traffic streams as well as NRT traffic. We have investigated the performance of NRT traffic in such a system when the RT call arrival processes are not Poisson processes; and we have also demonstrated that the effect of the burstiness of the RT traffic on the NRT performance can be significant and is not one-directional, therefore should not be ignored. We have provided a computationally efficient approximation for performance evaluation of the NRT traffic in such cases. Our method is based on considering the multiple classes of non-Poisson RT traffic streams as an equivalent single class of Poisson traffic stream. The proposed method can also be applied to cases that the RT call arrivals follow Poisson processes. It reduces the computational complexity dramatically when comparing with the exact Markov chain solution, especially when the number of RT classes is large. The proposed method is validated by numerical results.

## Acknowledgment

The work described in this paper was supported by City University of Hong Kong (Project No. 9220040).

## References

[1] J. Roberts, "Internet traffic, QoS, and pricing," *Proceedings of the IEEE*, vol. 92, no. 9, pp. 1389–1399, Sep. 2004.

[2] A. Leon-Garcia, R. Kwong, and G. Williams, "Performance evaluation methods for an integrated voice/data link," *IEEE Trans. Commun.*, vol. 30, no. 8, pp. 1848–1858, Aug. 1982.

[3] M. Zukerman, "Bandwidth allocation for bursty isochronous traffic in a hybrid switching system," *IEEE Trans. Commun.*, vol. 37, no. 12, pp. 1367–1371, Dec. 1989.

[4] R. Litjens and R. Boucherie, "Elastic calls in an integrated services network: the greater the call size variability the better the QoS," *Perform. Eval.*, vol. 52, no. 4, pp. 193–220, 2003.

[5] F. Delcoigne, A. Proutiere, and G. Régnié, "Modeling integration of streaming and data traffic," *Perform. Eval.*, vol. 55, no. 3-4, pp. 185–209, Feb. 2004.

[6] W. Song, H. Jiang, W. Zhuang, and X. Shen, "Resource management for QoS support in cellular/WLAN interworking," *IEEE Network*, vol. 19, no. 5, pp. 12–18, Sep./Oct. 2005.

[7] O. Boxma, A. Gabor, R. Núñez-Queija, and H. Tan, "Performance analysis of admission control for integrated services with minimum rate guarantees," in *Proc. 2nd NGI*, Valencia, Spain, Apr. 2006, pp. 41–47.

[8] R. Fortet and C. Grandjean, "Congestion in a loss system when some calls want several devices simultaneously," *Electrical Communications*, vol. 39, no. 4, pp. 513–526, 1964.

[9] J. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. 29, no. 10, pp. 1474–1481, Oct. 1981.

[10] J. W. Roberts, "A service system with heterogeneous user requirements – application to multi-service telecommunications systems," in *Performance of Data Communications Systems and Their Applications*, G. Pujolle, Ed. New York: North Holland, 1981, pp. 423–431.

[11] V. B. Iversen, "The exact evaluation of multi-service loss system with access control," *Teleteknik, English Edition*, vol. 31, no. 2, pp. 56–61, Aug. 1987.

[12] D. Tsang and K. Ross, "Algorithms to determine exact blocking probabilities for multirate tree networks," *IEEE Trans. Commun.*, vol. 38, no. 8, pp. 1266–1271, Aug. 1990.

[13] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer, 1995, ch. 2, pp. 80–89.

[14] F. Hübner and P. Tran-Gia, "Quasi-stationary analysis of a finite capacity asynchronous multiplexer with modulated deterministic input," in *Proc. ITC-13*, Copenhagen, June 1991.

[15] H. Tan, R. Núñez-Queija, F. Gabor, and O. Boxma, "Admission control for differentiated services in future generation CDMA networks," *Perform. Eval.*, vol. 66, no. 9-10, pp. 488–504, Sep. 2009.

[16] Y.-C. Huang, K.-T. Ko, and M. Zukerman, "A generalized quasi-stationary approximation for analysis of an integrated service system," *IEEE Commun. Lett.*, vol. 16, no. 11, pp. 1884–1887, Nov. 2012.

[17] Y.-C. Huang, K. T. Ko, and M. Zukerman, "A generalized fluid approximation for analysis of an integrated service system," in *Proc. IEEE HPSR 2013*, Taipei, Taiwan, Jul. 2013.

[18] R. I. Wilkinson, "Theories for toll traffic engineering in the U.S.A." *Bell Syst. Tech. J.*, vol. 35, no. 2, pp. 421–514, 1956.

[19] A. A. Fredericks, "Congestion in blocking systems – a simple approximation technique," *Bell Syst. Tech. J*, vol. 59, no. 6, pp. 805–827, Jul.–Aug. 1980.

[20] S.-P. Chung and J.-C. Lee, "Performance analysis and overflowed traffic characterization in multiservice hierarchical wireless networks," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 904–918, May 2005.

[21] L.-R. Hu and S. Rappaport, "Personal communication systems using multiple hierarchical cellular overlays," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 2, pp. 406–415, Feb. 1995.

[22] M. Głąbowski, K. Kubasik, and M. Stasiak, "Modeling of systems with overflow multi-rate traffic," *Telecommun. Syst.*, vol. 37, no. 1-3, pp. 85–96, 2008.

[23] Q. Huang, K.-T. Ko, and V. B. Iversen, "Approximation of loss calculation for hierarchical networks with multiservice overflows," *IEEE Trans. Commun.*, vol. 56, no. 3, pp. 466–473, Mar. 2008.

[24] Y.-C. Huang, K.-T. Ko, Q. Huang, V. B. Iversen, and M. Zukerman, "An efficient method for performance evaluation of femto-macro overlay systems," in *Proc. ICC*, Kyoto, Japan, June 2011.

[25] R. Syski, *Introduction to Congestion Theory in Telephone Systems, Second Edition*. North Holland, 1986, ch. 9, p. 497.

[26] Y. Huang, "Performance evaluation of hierarchical multiservice wireless networks," Ph.D. dissertation, City University of Hong Kong, 2013.