System Capacity Calculation for Packet-Switched Traffic in the Next Generation Wireless Systems, Part II: Batch Arrival M/G/1 Nonpreemptive Priority Queueing Model for Transmission over a Radio Channel

Hideaki Takagi(a), Hitoshi Yoshino(b), Naoto Matoba(b), Mitsuhiro Azuma(c), and Masashige Shirakabe(b)

a) University of Tsukuba, Tsukuba-shi, 305-8573 Japan; takagi@sk.tsukuba.ac.jp
b) NTT DoCoMo, Inc., Yokosuka-shi, 239-8536 Japan; {yosino, shirakabe}@mlab.yrp.nttdocomo.co.jp matoba@nttdocomo.co.jp
c) Fujitsu Laboratories Ltd., Kawasaki-shi, 211-8588 Japan; azuma@labs.fujitsu.com

Abstract: We propose a method of determining the required system capacity for several service categories of packet-switched traffic for the future development of the third generation (3G) mobile radio systems and systems beyond 3G. Based on the batch arrival M/G/1 nonpreemptive priority queueing model, we calculate the system capacity so as to satisfy the user’s Quality of Service (QoS) requirement of the mean delay and/or the delay percentile. Numerical analysis is given for the sensitivity of the required system capacity to such parameters as the offered traffic intensity, the mean delay requirement, and the packet size distribution.

Keywords: M/G/1 nonpreemptive priority queue, mobile radio systems, packet-switched traffic, sensitivity analysis, system capacity calculation.

1 INTRODUCTION

International Telecommunication Union Radiocommunication Sector (ITU-R) is developing a new methodology for calculation of radio spectrum requirements in the year 2010 onward for the future development of the third-generation (3G) mobile radio systems and systems beyond 3G. The conventional methodology [4, 5] was developed by using the Erlang-B formula for circuit-switched traffic and using the Erlang-C formula for packet-switched traffic. In the new methodology, the traffic of a mix of packet-switched service categories should be handled appropriately due to the forecast that the majority of future traffic will arise from multi-media applications and that the communication will be based on Internet Protocol (IP). In response to the request by ITU-R, we have developed a new spectrum calculation methodology and proposed it to the ITU-R. Together with a companion paper [3], this paper presents an alternative spectrum calculation methodology that has been approved as a Japanese proposal in a recent meeting of the ITU-R Working Party 8F.

For the new methodology, a single arrival M/G/1 nonpreemptive priority queueing model was proposed by Irnich and Walke [2], and it has been elaborated further [3]. A customer in their model represents an IP packet. Considering the fact that each IP packet is segmented into several frames, which are then transmitted separately over a radio channel, we propose a batch arrival M/G/1 nonpreemptive priority queue. In our model, each arriving batch represents an IP packet and each customer constituting the batch represents a radio frame. The formula for
the mean delay of a customer is used to calculate the system capacity in order to satisfy the mean delay requirement. We also propose an approximate distribution function for the delay whose parameters are determined to match the mean and the second moment of the delay. This can be used to calculate the system capacity in order to satisfy the requirement of the delay percentile for delay-sensitive service categories.

2 BATCH ARRIVAL M/G/1 NONPREEMPTIVE PRORITY QUEUE

In a batch arrival M/G/1 nonpreemptive priority queue, there are \( N \) classes of customers indexed \( i = 1, 2, \ldots, N \), where class \( i \) has priority over class \( j \) if \( i < j \). Batches of customers of class \( i \) arrive in a Poisson process at rate \( \lambda_i \) [batches/sec], each arriving batch containing \( G_i \) customers. The factorial moments of \( G_i \) are given by

\[
\begin{align*}
    g_i := E[G_i], \quad g_i^{(2)} := E[G_i(G_i - 1)], \quad g_i^{(3)} := E[G_i(G_i - 1)(G_i - 2)], \quad \ldots \quad 1 \leq i \leq N.
\end{align*}
\]

The service time \( B_i \) [sec] of each customer of class \( i \) has moments given by

\[
\begin{align*}
    \beta_i := E[B_i], \quad \beta_i^{(2)} := E[B_i^2], \quad \beta_i^{(3)} := E[B_i^3], \quad \ldots \quad 1 \leq i \leq N.
\end{align*}
\]

This queue is analyzed by Takagi and Takahashi [8]; see also [7, Section 3.5 (p.373)].

For a batch of customers of class \( n \), we denote by \( \mathcal{W}_n \) the waiting time of the first customer in a batch, by \( T_n \) the time interval from the start of service for the first customer to the end of service for the last customer in a batch, and by \( D_n \) the delay of the last customer in a batch, i.e., the delay of the whole batch. Then we have

\[
D_n = W_n + T_n.
\]

It is known that

\[
W_n := E[\mathcal{W}_n] = \frac{\sum_{i=1}^{n} \lambda_i g_i^{(2)} \beta_i^2 + \sum_{i=1}^{N} \lambda_i g_i^{(2)} \beta_i^2}{2(1 - \rho_{\leq n})(1 - \rho_{\leq n-1})}
\]

Figure 1: Delay in a batch arrival queue.
\[ T_n := E[T_n] = \frac{(g_n - 1)\beta_n}{1 - \rho_{\leq n-1}} + \beta_n, \]  
(3)

where

\[ \rho_{\leq n} := \sum_{i=1}^{n} \lambda_i g_i \beta_i. \]

Therefore, we get

\[ D_n := E[D_n] = \frac{\sum_{i=1}^{n} \lambda_i g_i (2) \beta_i^2}{2(1 - \rho_{\leq n})(1 - \rho_{\leq n-1})^2} + \frac{(g_n - 1)\beta_n}{1 - \rho_{\leq n-1}} + \beta_n. \]  
(4)

The second moment for the delay of a batch of class \( n \) is given by

\[ D_n^{(2)} := E[D_n^{2}] = W_n^{(2)} + 2W_nT_n + T_n^{(2)}, \]  
(5)

where \( W_n \) and \( T_n \) are given in (2) and (3), respectively, and

\[ W_n^{(2)} := E[W_n^{2}] = \sum_{i=1}^{n} \lambda_i \left( g_i (3) \beta_i^3 + 3g_i (2) \beta_i \beta_i^{(2)} + g_i^{(2)} \beta_i^{(3)} \right) + \sum_{i=1}^{N} \lambda_i g_i \beta_i^{(3)} \]

\[ + \frac{3(1 - \rho_{\leq n})(1 - \rho_{\leq n-1})^2}{(1 - \rho_{\leq n})^2} \sum_{i=1}^{n-1} \lambda_i \left( g_i (2) \beta_i^2 + g_i \beta_i^{(2)} \right) + \sum_{i=1}^{n} \lambda_i \left( g_i (2) \beta_i^2 + g_i \beta_i^{(2)} \right) \]

\[ \left(1 - \rho_{\leq n-1} \right) \left(1 - \rho_{\leq n-1} \right) W_n, \]  
(6)

\[ T_n^{(2)} := E[T_n^{2}] = \frac{(g_n - 1)\beta_n}{(1 - \rho_{\leq n-1})^3} \sum_{i=1}^{n-1} \lambda_i \left( g_i (2) \beta_i^2 + g_i \beta_i^{(2)} \right) + \frac{(g_n (2) - 2g_n + 2)\beta_n^2 + (g_n - 1)\beta_n^{(2)}}{(1 - \rho_{\leq n-1})^2} + \frac{2(g_n - 1)\beta_n^2}{1 - \rho_{\leq n-1}} + \beta_n^{(2)}. \]  
(7)

In [8], the waiting time \( W_n \) of the first customer in a batch is studied as well as the waiting time of an arbitrary customer in the batch. However, the delay \( D_n \) of the last customer in the batch is not considered. Therefore, in the Appendix of this paper, we give the distribution of \( T_n \) in Laplace-Stieltjes transform in (23), from which we can derive \( T_n \) in (3) and \( T_n^{(2)} \) in (7).

### 3 Capacity Allocation Based on Mean Delay

In mobile IP networks, each IP packet is segmented into several radio frames for transmission. For the purpose of modeling the packet transmission in a mobile IP network by the batch arrival M/G/1 nonpreemptive priority queue, let us assume that each arriving batch represents an IP packet and that each customer constituting the batch represents a radio frame. Thus \( G_i \) denotes the number of radio frames generated from an IP packet.
We denote by $L_i$ the length [bits] of a radio frame of class $i$, whose moments are given by

$$l_i := E[L_i], \quad l_i^{(2)} := E[L_i^2], \quad l_i^{(3)} := E[L_i^3], \quad \ldots \quad 1 \leq i \leq N.$$  

If radio frames are transmitted over the channel of capacity $C$ [bits/sec], we have the relation

$$\beta_i = \frac{l_i}{C}, \quad \beta_i^{(2)} = \frac{l_i^{(2)}}{C^2}, \quad \beta_i^{(3)} = \frac{l_i^{(3)}}{C^3}, \quad \ldots \quad 1 \leq i \leq N.$$  

Then the mean delay [sec] of an IP packet of class $n$ is given by

$$D_n(C) = \frac{n}{2} \frac{\sum_{i=1}^{n} \lambda_i g_i^{(2)} l_i^2 + \sum_{i=1}^{N} \lambda_i g_i^{(2)} l_i^2}{C - \sum_{i=1}^{n} \lambda_i g_i l_i} + \frac{(g_n - 1) l_n}{C} + \frac{l_n}{C}. \quad (8)$$  

We propose a method for system capacity allocation that can be used when the transmission time of a radio frame is constant. Such a case occurs in several radio access techniques, for example [1]. Suppose that the transmission time $B_i$ of a radio frame of class $i$ over the channel of capacity $C$ [bits/sec] is constant $\beta_i$ [sec]. Then we have

$$l_i = \beta_i C, \quad l_i^{(2)} = (l_i)^2 = (\beta_i C)^2, \quad l_i^{(3)} = (l_i)^3 = (\beta_i C)^3. \quad (9)$$  

Furthermore, if we assume that the length $S_i$ [bits] of an IP packet of class $i$ is an integral multiple of $l_i$, we get

$$s_i = g_i l_i, \quad s_i^{(2)} = (g_i^{(2)} + g_i) l_i^2, \quad s_i^{(3)} = (g_i^{(3)} + 3g_i^{(2)} + g_i) l_i^3.$$  

Thus we have

$$g_i = \frac{s_i}{l_i}, \quad g_i^{(2)} = \frac{s_i^{(2)} - s_i l_i}{l_i^2}, \quad g_i^{(3)} = \frac{s_i^{(3)} - 3s_i^{(2)} l_i + 2s_i l_i^2}{l_i^3}. \quad (10)$$  

Substituting(9) and (10) into (8), we get

$$D_n(C) = \frac{\sum_{i=1}^{n} \lambda_i s_i^{(2)} + C \sum_{i=n+1}^{N} \lambda_i s_i \beta_i}{\left(C - \sum_{i=1}^{n} \lambda_i s_i\right) \left(C - \sum_{i=1}^{n-1} \lambda_i s_i\right)} + \frac{\sum_{i=1}^{n-1} \lambda_i s_i}{C - \sum_{i=1}^{n-1} \lambda_i s_i}, \quad (11)$$  

We use this expression to determine the channel capacity for class $n$.

Given the Quality of Service (QoS) requirement on the mean delay $D_n$ for class $n$, the required channel capacity $C_n$ is determined so as to satisfy the condition that

$$D_n(C_n) = D_n. \quad (12)$$  

This reduces to the following quadratic equation for $C_n$:

$$a_n x^2 + b_n x + c_n = 0,$$
where
\[ a_n := 2D_n, \]
\[ b_n := -2 \left[ D_n \sum_{i=1}^{n} \lambda_is_i + (D_n - \beta_n) \sum_{i=1}^{n-1} \lambda_is_i + s_n \right] - \sum_{i=n+1}^{N} \lambda_is_i\beta_i, \]
\[ c_n := 2 \left( \sum_{i=1}^{n} \lambda_is_i \right) \left[ (D_n - \beta_n) \sum_{i=1}^{n-1} \lambda_is_i + s_n \right] - \sum_{i=1}^{n} \lambda_is_i^{(2)}. \]

This equation has two real roots, and we can choose \( C_n \) that satisfies the stability condition
\[ C_n > \sum_{i=1}^{n} \lambda_is_i. \quad (13) \]

After the set \( \{C_1, C_2, \ldots, C_N\} \) has been determined, the system capacity \( C \) is obtained as
\[ C = \max(C_1, C_2, \ldots, C_N). \quad (14) \]

Then the length \( S_i \) of an IP packet of class \( i \) may not be an integral multiple of \( l_i = \beta_iC \) for all \( i = 1, 2, \ldots, N \). However, we do not adjust this situation for the sake of simplicity.

### 4 Capacity Allocation Based on Delay Percentile

Let us propose a new approximate form for the distribution function of the delay (waiting time plus service time) \( D \) with three parameters:
\[ P\{D \leq t\} = \begin{cases} 0 & 0 \leq t < \beta \\ 1 - qe^{-\gamma(t-\beta)} & t \geq \beta \end{cases}, \]
\[ (15) \]
where \( \beta \) is the minimum service time. We determine the parameters \( q \) and \( \gamma \) in (15) so that:

(a) It has the specified mean:
\[ D := E[D] = \int_{0}^{\infty} P\{D > t\} dt = \beta + \frac{q}{\gamma}. \]

(b) It has the specified second moment:
\[ D^{(2)} := E[D^2] = 2 \int_{0}^{\infty} tP\{D > t\} dt = \beta^2 + \frac{2q\beta}{\gamma} + \frac{2q}{\gamma^2}. \]

Hence we get
\[ q = \frac{2(D - \beta)^2}{D^{(2)} - 2D\beta + \beta^2} \quad \text{and} \quad \gamma = \frac{2(D - \beta)}{D^{(2)} - 2D\beta + \beta^2}. \quad (16) \]

The percentile of the delay distribution can be obtained as follows. Let \( \pi(r) \) denote the \( r \)th percentile of the distribution function of \( D \) in (15). Then, from the relation
\[ 1 - qe^{-\gamma[\pi(r) - \beta]} = \frac{r}{100} \]
we get
\[ \pi(r) = \beta + \frac{1}{\gamma} \log_e \left( \frac{100q}{100 - r} \right). \quad (17) \]
The approximate delay distribution function in (15) may be used for determining the channel capacity $C_n$ required to satisfy the QoS with respect to the delay percentile for class $n$. In doing so, $D_n(C)$ in (11) and $D_n^{(2)}(C)$ given below should be used for $D$ and $D^{(2)}$:

$$D_n^{(2)}(C) = W_n^{(2)}(C) + 2W_n(C)T_n(C) + T_n^{(2)}(C),$$

where $W_n(C)$ and $T_n(C)$ are given as the first and second terms, respectively, on the right-hand side of (11), and

$$W_n^{(2)}(C) = \frac{\sum_{i=1}^{n} \lambda_i s_i^{(3)} + C^2 \sum_{i=n+1}^{N} \lambda_i s_i \beta_i^2}{3 \left( C - \sum_{i=1}^{n} \lambda_i s_i \right) \left( C - \sum_{i=1}^{n-1} \lambda_i s_i \right)^2} + \frac{\sum_{i=1}^{n-1} \lambda_i s_i^{(2)}}{\left( C - \sum_{i=1}^{n-1} \lambda_i s_i \right)^2} + \frac{\sum_{i=1}^{n} \lambda_i s_i^{(2)}}{\left( C - \sum_{i=1}^{n} \lambda_i s_i \right) \left( C - \sum_{i=1}^{n-1} \lambda_i s_i \right)}. \quad (19)$$

$$T_n^{(2)}(C) = \frac{(s_n - \beta_n C) \sum_{i=1}^{n-1} \lambda_i s_i^{(2)}}{\left( C - \sum_{i=1}^{n-1} \lambda_i s_i \right)^3} + \frac{s_n^{(2)} - 2s_n \beta_n C + (\beta_n C)^2}{\left( C - \sum_{i=1}^{n-1} \lambda_i s_i \right)^2} + \frac{2(s_n - \beta_n C) \beta_n}{C - \sum_{i=1}^{n-1} \lambda_i s_i} + \beta_n^2. \quad (20)$$

Also, $\beta$ in (15) should be $\beta_n$, the transmission time of a radio frame of class $n$.

Then, $q_n(C)$ and $\gamma_n(C)$ are calculated by (16). Given the $r$th percentile $\pi_n(r)$ of the delay for class $n$, the condition

$$\beta_n + \frac{1}{\gamma_n(C)} \log_e \left( \frac{100q_n(C)}{100 - r} \right) \leq \pi_n(r) \quad (21)$$

is checked. The required channel capacity $C_n$ is determined by incrementing $C$ until this condition is satisfied. The system capacity $C$ is obtained by (14).

**5 SENSITIVITY ANALYSIS**

We investigate the effects of the parameter values on the required system capacity based on the mean delay formula in Section 3 for a system of three classes. The basic values of the input parameters for the sensitivity analysis are given in Table 1. The values for the offered traffic has been calculated from the data in the Report ITU-R M. 2023 [5] by assuming class 1 for High Multimedia, class 2 for Medium Multimedia, and class 3 for Simple Message. For the packet size, we have assumed $s_n^{(2)} = 2s_n^2$ as in the exponential distribution. The radio frame size is set to 2 milliseconds for all classes. The sensitivity analysis results based on the delay percentile formula in Section 4 are omitted here due to the scarcity of space.

The influence of the offered traffic intensity on the required channel capacity is shown in Fig. 2. In the left figure, the offered traffic of class 1 ($\lambda_1 s_1$) is varied while all other parameter
values are kept fixed as in Table 1. The influence of the mean delay requirement, the mean packet size, and the second moment of the packet size is similarly shown in Fig. 3, Fig. 4, and Fig. 5, respectively. The curves in these figures plot the required channel capacity for each class. Therefore, the system capacity $C$ should be obtained as the upper envelop of the curves.

In Fig. 2, the required channel capacity depends almost linearly on the offered traffic intensity. In general, when the offered traffic of class $n$, $\lambda_n s_n$, is varied, we have from (13) the inequality

$$C_m > \begin{cases} \left( \sum_{i=1}^{m} \lambda_i s_i \right) & m < n \\ \left( \sum_{i=1}^{n-1} \lambda_i s_i \right) + \lambda_n s_n & m = n \\ \left( \sum_{i=1}^{n-1} \lambda_i s_i + \sum_{i=n+1}^{m} \lambda_i s_i \right) + \lambda_n s_n & m > n \end{cases}$$

(22)

where the quantities in the parentheses are constant. The values plotted in Fig. 2 lie somewhat above the right-hand side of (22).

In Fig. 3, when $D_n = D_n(C_n) \rightarrow 0$, we have $C_n \rightarrow \infty$ while other $C_m$’s ($m \neq n$) are nearly constant. On the other hand, when $D_n = D_n(C_n) \rightarrow \infty$, we have $C_n \rightarrow \sum_{i=1}^{n} \lambda_i s_i$. These asymptotes are expected from (11).

In Fig. 4, the mean packet size $s_n$ is varied while the offered traffic $\lambda_n s_n$ is kept constant. Therefore, decreasing $s_n$ means increasing $\lambda_n$, which requires a large channel capacity $C_n$ so as to keep $D_n(C_n) = D_n$ constant. In Fig. 5, the largest channel capacity is required for the class for which the second moment of the packet size is increased.

If we compare the results in these figures with the corresponding results in the companion paper [3], we observe similar trends for the influence of offered traffic and mean delay requirement, but different trends for the influence of the mean and second moment of the packet size distribution. This is because the difference comes from the segmentation of an IP packet. Hence we need further validation work of the system capacity calculation model possibly against the simulation of radio transmissions in a realistic environment.
Figure 2: Influence of the offered traffic on the channel capacity.

Figure 3: Influence of the mean delay requirement on the channel capacity.

6 CONCLUDING REMARK

The current framework for developing the new methodology in ITU-R recommends that the system capacity be determined based only on the IP layer model as presented in [3]. The physical aspects of the radio interface such as the segmentation of each IP packet into several radio frames for transmission are supposed to be taken into consideration as multiplicative factors for the required spectrum value obtained from the IP layer model. However, in radio communication systems, the traffic performance of IP packet transmission is intrinsically dependent on the radio interface aspects. Therefore, our model in this paper can be an alternative to the one in [3].

Acknowledgement

H. Takagi’s work is supported by the Mobile IT Forum through Commissioned Research ACA16104 and ACA17044 in 2004–05.
Figure 4: Influence of the mean packet size on the channel capacity.

Figure 5: Influence of the second moment of packet size on the channel capacity.

References


Appendix: Distribution of $T_n$

In a batch arrival M/G/1 nonpreemptive priority queue described in Section 2, recall that $T_n$ denotes the time interval from the start of service for the first customer to the end of service for the last customer in a batch of class $n$. If the batch contains $k$ customers, $T_n$ consists of (i) $k-1$ delay cycles [6, Section 3.3 (p.111)], each of which consists of the initial delay for the service of a customer of class $n$ and the subsequent delay busy periods generated by the customers of class 1 through $n-1$, and (ii) the uninterrupted service time for the last customer. This consideration leads to the following Laplace-Stieltjes transform (LST) of the distribution function for $T_n$:

$$T_n(s) := E[e^{-sT_n}] = \int_0^\infty e^{-st}dP\{T_n \leq t\} = \frac{G_n\{B_n[\sigma_{\leq n-1}(s)]\}}{B_n[\sigma_{\leq n-1}(s)]}B_n(s),$$

where $B_n(s)$ is the LST of the distribution function for $B_n$, the service time of a customer of class $n$, $G_n(z)$ is the PGF of $\sigma_{\leq n-1}$, the number of customers included in a batch of class $n$, and

$$\sigma_{\leq n-1}(s) := s + \lambda_{\leq n-1} - \lambda_{\leq n-1}\Theta_{\leq n-1}(s)$$

with

$$\lambda_{\leq n-1} := \sum_{i=1}^{n-1} \lambda_i.$$

Furthermore, $\Theta_{\leq n-1}(s)$ is the LST of the distribution function for the duration $\Theta_{\leq n-1}$ of a busy period generated by the customers of classes 1 through $n-1$. It is given as the solution to the equation

$$\Theta_{\leq n-1}(s) = B_{\leq n-1}[s + \lambda_{\leq n-1} - \lambda_{\leq n-1}\Theta_{\leq n-1}(s)],$$

where

$$B_{\leq n-1}(s) := \frac{1}{\lambda_{\leq n-1}} \sum_{i=1}^{n-1} \lambda_i G_i[B_i(s)].$$

The mean $T_n$ and the second moment $T_n^{(2)}$ given in (3) and (7) can be derived from (23).