



TECHNICAL PAPER

Title: How to Dimension Wireless Networks for Packet Data Services with Guaranteed QoS (Part 1—Theoretical Issues)

Authors: S. Rasoul Safavian, Ph.D. — Bechtel Corporation

Date: August 2005

Publication/Venue: Bechtel Telecommunications Technical Journal, Vol. 3, No. 1
©2005 Bechtel Corporation. All rights reserved.

HOW TO DIMENSION WIRELESS NETWORKS FOR PACKET DATA SERVICES WITH GUARANTEED QoS (PART 1 – THEORETICAL ISSUES)

Issue Date: August 2005

Abstract—Developed primarily for 1G and 2G applications, today's commercially available tools and procedures for network design and dimensioning address only the best-effort packet data applications, where the only QoS considered is basically the nonguaranteed average data throughput. This paper, Part 1 of 2, provides a theoretical framework for dimensioning/sizing 3G+ wireless networks to provide various packet data services having different QoS requirements. (Part 2 will address the simulation results and the corresponding network dimensioning tool.) The analysis, which is based on nonpreemptive-priority M/G/s/s+r queueing systems modeling, incorporates prioritization of handover calls as well as different packet data services. The effects of channel impairments are also considered. Explicit closed-form formulas, with relative computational simplicity, are provided.

ORGANIZATION OF PAPER

This paper is organized into four sections:

- The first provides background information and highlights the main issues that are addressed.
- The second examines the subject of traffic modeling for voice and packet data applications.
- The third presents proposed solutions, including explicit delay probabilities and waiting time distributions for both nonprioritized and two-class priority cases.
- The fourth states conclusions and provides closing remarks.

INTRODUCTION

Unlike the first-generation (1G) and second-generation (2G) wireless technologies—designed primarily for voice applications—a main feature of third-generation (3G) systems and beyond (3G+, including enhanced 3G and fourth-generation [4G]) is their high-speed multimedia capability. These new technologies also enable concurrency, allowing a mobile user to simultaneously place a voice call while downloading a large file, composing or replying to e-mails, and browsing the Internet. Thus, a wireless user is no longer limited to only voice calls or very low-speed (often circuit-switched [CS]) data applications.

In fact, as voice use reaches saturation levels, packet data applications are expected to play an increasingly important role, and provide a greater source of revenue, in future-generation wireless networks. In support of this enhanced capability, approved 3G+ standards incorporate pre-defined quality of service (QoS; sometimes referred to as grade of service [GoS]) requirements for each type of application. The QoS vector includes parameters such as guaranteed user throughput, maximum allowed delay, maximum allowed jitter, bit error rate, and maximum tolerable packet loss rate.

Today's commonly used traffic models and tools designed for CS voice applications are insufficient to handle the new packet-switched (PS) multimedia applications. Increasing competition among operators and the very stringent QoS requirements make it crucial for the network designer to estimate network size as realistically as possible. An undersized network cannot provide the required QoS, leading to unsatisfied customers and high churns; an over-designed network is wasteful, particularly with network operators facing fierce competition and tough financial situations.

While there is much in the literature about QoS, call admission control (CAC), and radio resource management (RRM) [1-9], most articles either look at the architecture required to support QoS or examine various resource-sharing algorithms. More specifically, they focus on power allocation

S. Rasoul
Safavian, PhD
srsafavi@bechtel.com

ABBREVIATIONS, ACRONYMS, AND TERMS

1G	first generation
2G	second generation
3G	third generation
3G+	beyond 3G
3GPP™	Third Generation Partnership Project
4G	fourth generation
AF	activity factor
ARQ	automated repeat request
AS	application server
BTS	base transceiver station
CAC	call admission control
CDMA	code division multiple access
CN	core network
CS	circuit switched
FCFS	first come, first served
FIFO	first in, first out
FTP	file transfer protocol
GoS	grade of service (same as QoS)
GPD	generalized Pareto distribution
HOL	head of the line
HTTP	hypertext transport protocol
IMS	IP multimedia subsystem
IP	Internet Protocol
LAN	local area network
LIFO	last in, first out
MT	mobile terminal
OVSF	orthogonal variable spreading factor
PC	packet call
PS	packet switched
QoS	quality of service
RAN	radio access network
RRM	radio resource management
TCP	transmission control protocol
VoIP	voice over IP
WAP	wireless application protocol

algorithms for the base transceiver station (BTS) scheduler, who must examine the requested or required data rates for different users/applications and try to allocate the required power for the desired bit rate.

This paper takes a different approach. It determines, *a priori*, the kind of resources (time slots, frequency channels, required power, Walsh or orthogonal variable spreading factor [OVSF] codes) needed at a BTS or Node B to guarantee the required QoS. It does this by providing explicit closed form expressions for the queuing delays or waiting time distributions in BTS or Node B. Doing this helps the network designer to properly dimension the network and network resources to provide the desired services with the required QoS for the various applications.

Furthermore, the current push for voice over Internet Protocol (VoIP), such as the release of Revision A of the code division multiple access 2000 (cdma2000®) 1xEV-DO technology standard, and the recent standardization attempts for IP multimedia subsystem (IMS) seem to point the future toward all-IP, all-packet services networks. Networks would, of course, have very stringent QoS requirements for voice, video, teleconferencing, and streaming applications.

To allow these different classes of services, transmitter sides typically have different queues for different priority levels. To address this situation, this paper looks at two-class priority queuing systems. For instance, since handover calls are typically more important than new originating calls, priority schemes are used to differentiate between these classes of calls, as well as to differentiate among various packet data services. The paper shows the results for a two-class case, noting that the results can easily be extended to the m -class case. The paper also accounts for the effects of wireless channel impairments, such as fading, in the context of automated repeat request (ARQ) retransmissions.

TRAFFIC MODELING

To successfully design a network, it is important to select the appropriate source or user traffic models that reflect the behavior of the network users [10]. Since voice is currently handled as CS and data as PS, this paper looks at voice and data traffic separately.

Voice Traffic Modeling

The theory of voice traffic modeling and network dimensioning is both well understood and well established [11, 12]. Typically, the random arrival of calls is modeled as a homogenous Markov (Poisson) process with exponential inter-arrival times based on an average arrival rate of λ calls per second and with call durations modeled by exponential distribution using a mean duration or call holding time of $1/\mu$ seconds per call. By

assuming that s servers or resources (frequency channels, time slots, channel elements, etc.) can be assigned to arriving calls based on a first-come, first-served (FCFS) policy without prioritization or queuing of blocked calls, and that the number of arriving calls is infinite (or sufficiently large), then queuing theory can be used to establish a simple relation between call blocking probability P_B , λ , μ , and s . More specifically, it can be shown that:

$$P_B = \frac{(\lambda/\mu)^s / s!}{\sum_{j=0}^s [(\lambda/\mu)^j / j!]} \quad (1)$$

The significance of this relation, known as the Erlang B formula, is that, given any three of the parameters, the fourth can be computed easily. For example, given the demand traffic (basically λ and μ) and the desired QoS (here, simply the blocking probability), the required resources s can be easily determined.

We are going to focus on queuing systems and use Kendall's notation to describe the system characteristics. The notation is of the form $A/S/s/c/p/D$, where A is the arrival process; S , the service time distribution; s , the number of servers; c , the system capacity (i.e., number of servers plus queue size); p , the population size or maximum number of calls that can arrive; and D , the queuing discipline, such as first in, first out (FIFO) or last in, first out (LIFO). In this notation, Erlang B modeling corresponds to an $M/M/s/s/\infty/FIFO$ queue, where the two M 's denote that the arrival process and the service time process, respectively, are both Markovian (Poisson), and the two s 's denote that s system resources are available to handle up to s users at a time, with no queuing or buffering of blocked calls. This model is appropriate for voice because voice callers prefer that their calls be blocked during busy periods rather than being placed on hold. Also, because there is no buffer in Erlang B, the average delay for calls that go through is simply the average service time, i.e., $1/\mu$.

For certain applications where callers can be placed in an (infinite) queue when all resources are tied up, analysis is still rather simple, with the results corresponding to an $M/M/s/\infty/\infty$ queue (with P_D being the probability that a caller is delayed), better known as Erlang C:

$$P_D = \frac{\frac{(\lambda/\mu)^s}{(s-1)!(s-\lambda/\mu)}}{\sum_{j=0}^{s-1} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^s}{(s-1)!(s-\lambda/\mu)}} \quad (2)$$

In the Erlang C model, the total average delay is the sum of the average service time $1/\mu$ and the average delay in the queue W_q , which is given as:

$$W_q = \frac{(1/\mu) P_D}{s - \lambda/\mu} \quad (3)$$

Packet Data Traffic Modeling

While the characterization of voice users is fairly straightforward, the traffic generated by packet data users is highly dependent on the application and has a high degree of burstiness (i.e., the inter-arrival time between data packets, as well as the packet length, can vary widely). Also, the transmission control protocol (TCP) rate adaptation mechanism causes the amount of transmitted traffic to depend on the network load and the contacted server's processing speed. Thus, the customer's behavior is feedback-oriented and is determined not only by the application, but also by external influences that are difficult to foresee. This feedback-oriented nature of packet data applications makes the Poisson model inappropriate for most packet data traffic modeling. Furthermore, differences among hypertext transport protocol (HTTP) versions and among browsers affect the way a Web page is downloaded and how much traffic is generated.

Currently, the most frequently used wireless data applications are Web browsing, e-mail, file downloading/uploading using the file transfer protocol (FTP), and the wireless application protocol (WAP). Without loss of generality, this paper looks at HTTP and e-mail and refers interested readers to other sources—including traffic models recommended in the Third Generation Partnership Project (3GPP™) and 3GPP2 standards bodies—for FTP, WAP, etc. [13–15].

HTTP Modeling

Three basic methods have been used to measure and model Web traffic: (a) using server logs; (b) using client logs, and (c) tracing packets or measuring traffic through a local area network (LAN) and extracting the Web traffic. The models proposed in these approaches all have the same structure: The user creates a session, consisting of a random number of packet calls (PCs), where each PC has a random number of packets and durations or lengths. This HTTP model is depicted in **Figure 1**.

This model is particularly useful and is general enough to encompass many other practical PS models for both real-time and nonreal-time applications, such as e-mail and FTP. The model

Today's commonly used traffic models and tools designed for CS voice applications are insufficient to handle the new PS multimedia applications.

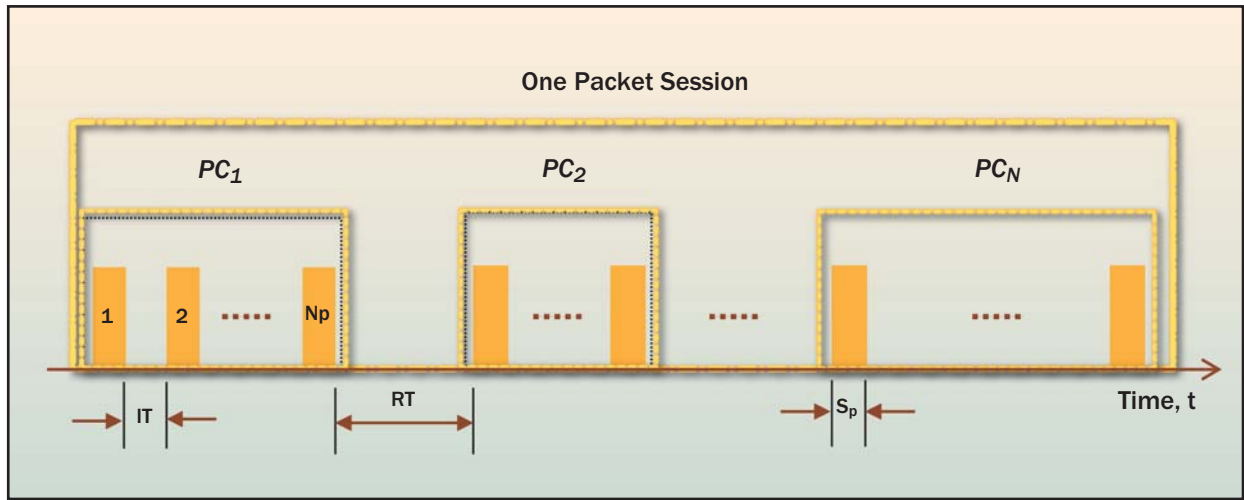


Figure 1. HTTP Traffic Model

is completely specified by the following five parameters:

- Number of PCs in a given session: N_{pc}
- Reading time between PCs: RT
- Number of packets within a PC: N_p
- Packet inter-arrival time: IT
- Packet size or length: S_p

Table 1 summarizes the characteristics of these parameters, obtained from extensive measurements.

The following general remarks apply to this HTTP model:

- Geometric distribution is a discrete version of exponential distribution.
- Exponential distribution is a special case of generalized Pareto distribution (GPD), given by:

$$GPD(x; \sigma, k) = \begin{cases} 1 - \left(1 - \frac{kx}{\alpha}\right)^{\alpha/k}, & k \neq 0 \\ 1 - e^{-x/k}, & k = 0 \end{cases} \quad (4)$$

Table 1. HTTP (Web) Traffic Model

MODEL PARAMETER	DISTRIBUTION	DISTRIBUTION COMPLETELY SPECIFIED BY	TYPICAL VALUES
N_{pc}	Geometric	Mean, $\mu_{N_{pc}}$	5
RT	Geometric	Mean, μ_{RT}	41.2 Seconds
N_p	Geometric	Mean, μ_{N_p}	25
IT	Geometric	Mean, μ_{IT}	Bit-rate Dependent
S_p	Pareto (with cutoff)	α, k	$\alpha = 1.1$ $k = 81.5$ bytes

where k is the shape parameter and $\alpha > 0$ is the scale parameter. Note that for $k = 0$, GPD reduces to exponential distribution with mean α , while for $k < 0$, it reduces to standard Pareto distribution.

- Pareto with cutoff is used in the Web model because, in real applications, packet size is always upper-bounded by a maximum value m , i.e.,

$$Packet\ Size = \min(p, m) \quad (5)$$

where p is the standard Pareto random variable and m is the maximum allowed packet size (typically, $m = 66,666$ bytes).

E-Mail Modeling

Besides HTTP, the most important wireless data application is probably e-mail, for which the literature contains various models [13–17]. During an e-mail reading session, the first information received is the header. This is followed by an off period or read time while the user reads the received information, i.e., the list of available e-mails, old and new. Then the user may decide to download the first new message, if any, with more off time to read that message. The user may then decide to download the next (new or old) message or compose a response to a message, etc. Graphically, e-mail traffic can be modeled as shown in **Figure 2**.

Other Traffic Modeling

Using similar logic, traffic models can also be constructed for FTP and other applications with a similar structure. Basically, every user/application starts a session. Session traffic can be described in terms of the arrival process for user activities and the activity-phase process. For a voice call, the session duration is just the duration of that call. For packet data, the session is

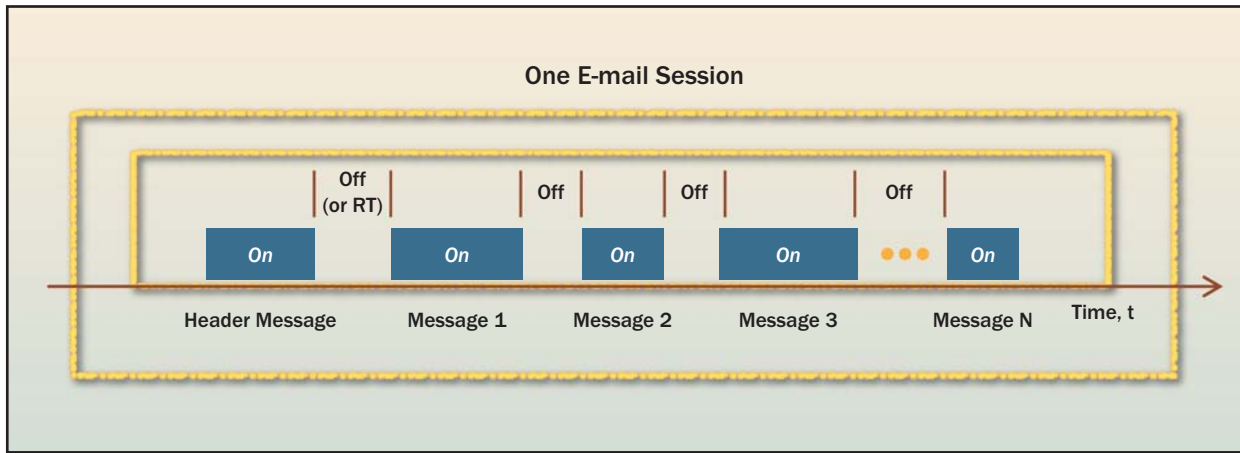


Figure 2. E-mail Traffic Model

typically composed of multiple packet data calls with periods of inactivity in between. Therefore, the service or session time distribution is generally a mixed distribution of several random quantities (e.g., the number of PCs, the duration/size of each PC). In the notation of queuing theory, CS voice can be modeled as a simple $M/M/s/s$ queue, whereas packet data has to be modeled as $M/G/s/s+r$, where G stands for a general service or session distribution time and r stands for the size of the buffer or queue. This simple departure of service time from Markov to general process has astronomical consequences in terms of the difficulty of obtaining a solution. $M/M/s/s$ queues are well understood, whereas $M/G/s/s+r$ queues do not easily lend themselves to analysis.

How Existing Network Planning Tools Use Packet Data Call Models

This section briefly examines how existing commercial network planning tools handle packet data traffic. As best as could be determined at the time this paper was written, the existing network planning tools are static, designed basically for CS traffic and incapable of tracking time stamps on individual packets. As such, they cannot provide a network design solution with any sort of time-related QoS provisioning, such as maximum tolerable delay or jitter.

Commercial tools often provide a user interface where parameter values can be entered for traffic models such as a Web model, thus creating the impression that these tools can handle packet data applications. In reality, commercial tools merely use the packet traffic model parameters to compute a mean packet activity factor (AF) and then proceed with the design as a CS design.

More specifically, mean packet AF is computed simply as:

$$\begin{aligned} \text{Mean Packet Switch Activity Factor} &= \frac{\text{Busy Time}}{\text{Total Time}} \\ &= \frac{(1 + \text{ARQ rate})(N_{pc})(N_p)(S_p)(1/R_b)}{(1 + \text{ARQ rate})(N_{pc})(N_p)(S_p)(1/R_b) + [IT(N_p - 1)](N_{pc}) + RT(N_{pc} - 1)} \end{aligned} \quad (6)$$

where the terms are as defined earlier, ARQ rate is the average overhead associated with retransmission, and R_b is the mean bit rate defined by the user.

Similar to CS traffic, once the PS activity factor is computed, a channel with the bit rate specified in the traffic model as the mean bit rate is dedicated to that PS traffic. While this approach can capture average interference contributed by a packet data call user (such as a Web user), it does not provide any insight about the appropriateness of the design to meet required QoS for various data applications.

THE PROPOSED SOLUTION

As discussed, there are various source traffic models for different packet data applications. What is clear from these and similar studies is that packet data traffic does not fit into the standard voice traffic model, i.e., the service time distribution typically is not negative exponential. Furthermore, most proposed models are based on a general form of *on/off* model where a session is generally composed of some random number of *on/off* periods described by some appropriate distribution. Each *on* period may be further specified in terms of a random number of packets, packet durations, packet inter-arrival times, etc.

An undersized network cannot provide the required QoS, leading to unsatisfied customers and high churns; an over-designed network is wasteful, particularly with network operators facing fierce competition and tough financial situations.

To successfully design a network, it is important to select the appropriate source or user traffic models that reflect the behavior of the network users.

This paper proposes, instead, to establish a framework that would allow network planning under any traffic model.

To start, it is assumed, without loss of generality, that the traffic is generated from two different priority classes. For example, class 1 could be VoIP, video stream, handover calls, etc., and class 2 could be Web, e-mail, or new originating calls. Handover calls may be considered class 1 calls, since they are typically more important than a newly originated call.

The class 1 arrival process is assumed to be Poisson (or Markovian), with an average rate of λ_1 and a retransmission rate of λ_1^R . The class 2 arrival process is also Poisson, but with a different mean rate of λ_2 and a retransmission rate of λ_2^R . These four processes are assumed to be independent. All assumptions are very realistic and have been verified in many situations. (The m -class priority case is a simple conceptual, but complex mathematical, extension of the two-class case.)

Both class 1 and class 2 service time distributions are allowed to be the same and to be any general distribution G , as long as the first three moments of this distribution are finite. The mean of G is denoted as $1/\mu$. Furthermore, arrival time and service time distributions are assumed to be independent.

Initially, the queues are given infinite capacity, i.e., the buffer size is allowed to be infinite. If

buffer size, i.e., capacity queue, becomes finite, that buffer size is denoted as r . The number of homogenous servers or resources needed to provide the desired QoS is s .

A FIFO service discipline with nonpreemptive or head-of-the-line (HOL) priority queuing is assumed; i.e., if all servers are busy and calls have to be queued, higher priority class calls (or packets) are sent in front of lower. However, no service (high or low priority class) is disrupted to allow another call.

Using Kendall's notation for queues, the queuing model considered will be $M/G/s/s+r$ for the finite capacity case and $M/G/s$ for the infinite capacity case. The graphical model for the system under study is given in Figure 3. Figure 3a corresponds to the downlink scenario and Figure 3b to the uplink scenario.

Packets are taken first from the priority class 1 queue. Only when the buffer is empty are calls or packets from priority class 2 processed. Priority class 1 packets that need retransmission are given higher priority and are directed to the priority class 1 queue. It is assumed that retransmitted packets for both classes follow the Poisson process, independent of the original processes. λ_1^R and λ_2^R denote the mean retransmission rates for handover (class 1) and new calls (class 2), respectively.

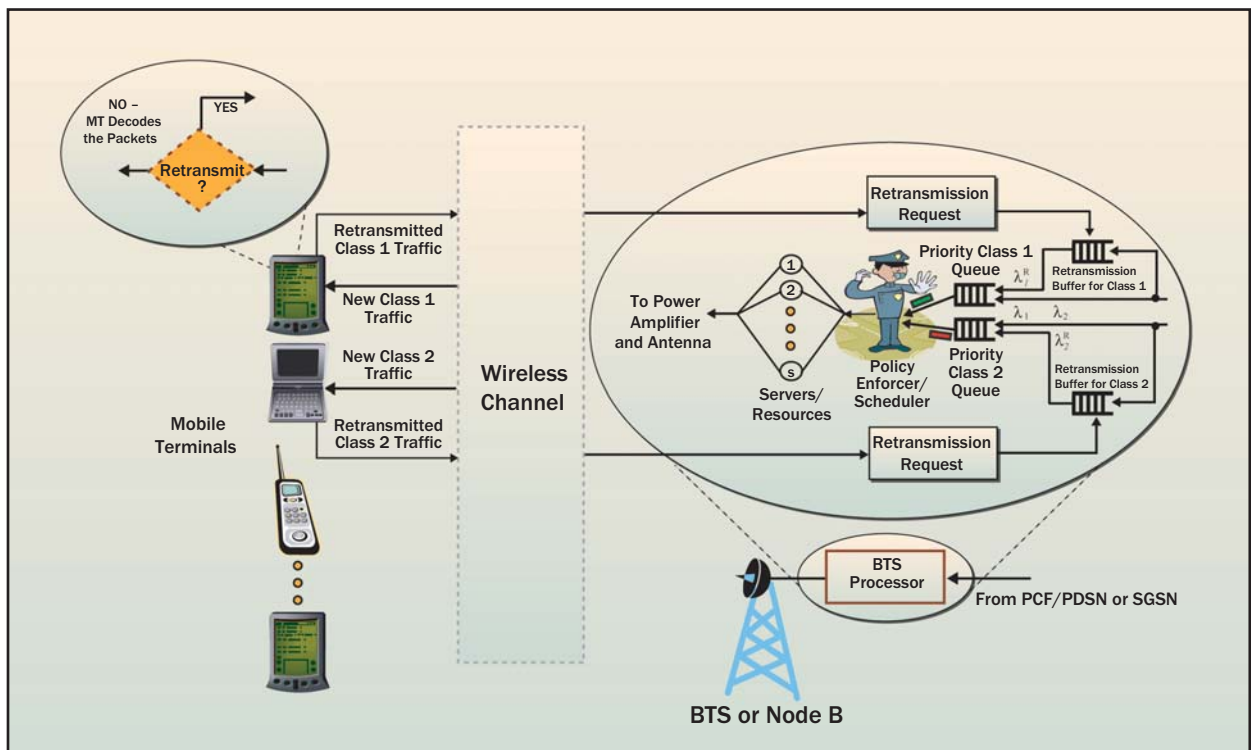


Figure 3a. Basic System Model for Downlink

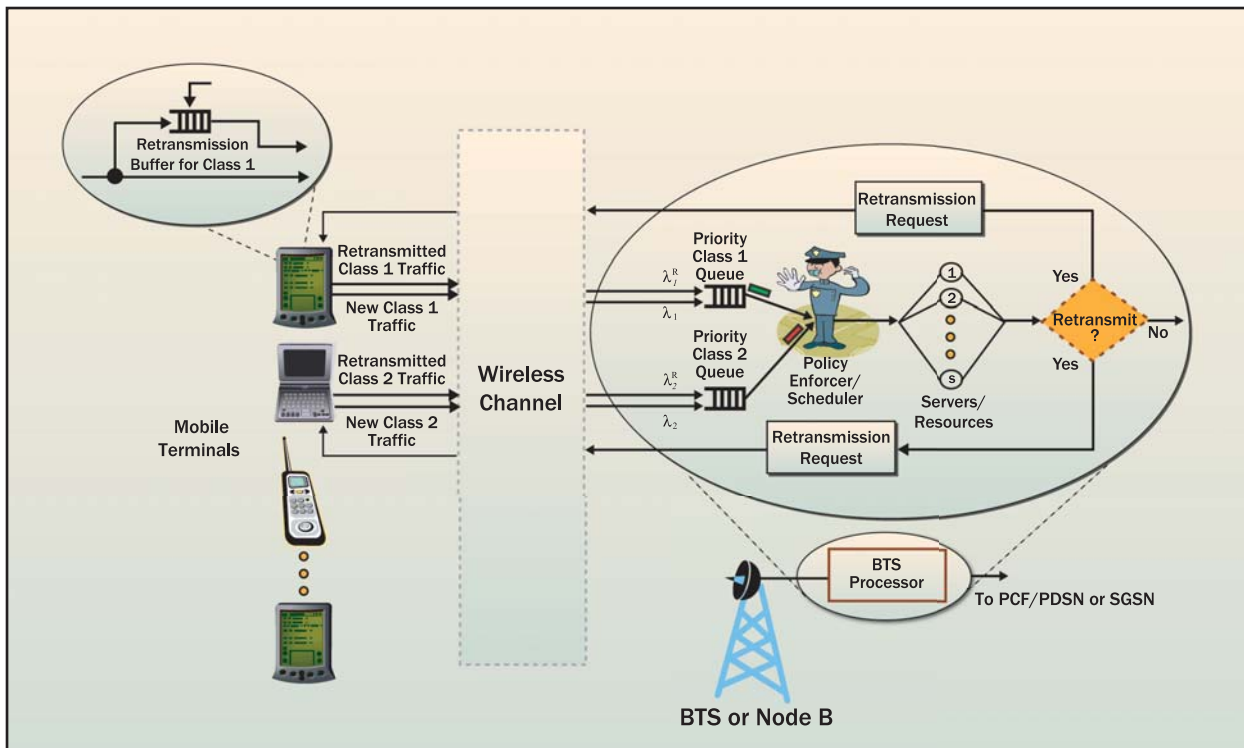


Figure 3b. Basic System Model for Uplink

Power control is assumed to be perfect. This implies that the actual locations of MTs are unimportant as far as the uplink analysis for the cell under consideration is concerned. Of course, MT locations do affect the performance of the other cells.

The QoS measures of most interest for packet data applications are maximum allowed packet delay, maximum allowed jitter, and, perhaps, average acceptable delay. Maximum packet delays are usually specified in terms of a reliability factor, e.g., a desire to have a maximum packet delay of less than τ_{max} with a certain reliability of, say, 95 percent.

Given the actual waiting time distribution (either for the queue or the combination of the queue and the servers), all three of the above QoS measures can be computed directly.

Waiting Time Distribution and Network Planning Without Priority Classes

This section derives the explicit equations describing the waiting time distribution for the no-priority case, i.e., for no differentiation among call or packet classes. The arrival process is still assumed to be Poisson, with a mean $\lambda > 0$, but the service time distribution is allowed to be any general service time distribution G with a mean of $1/\mu$, independent of the arrival process.

The nonexponential nature of the service time distribution renders analysis of the underlying embedded multidimensional Markov chain almost intractable. Several very accurate approximations have been proposed in the literature. The basic approaches can be classified into three different categories: transform-based approaches, matrix-analytic approaches, and approaches exploiting the special structure of a particular Markov chain along with the use of some heuristics.

This paper follows the approximation results suggested by Kimura [18, 19] that provide accurate, precise, and computationally feasible solutions. To proceed further, it is necessary to define certain terms. Let c^2 be the square of variation (variance divided by the square of the mean) of G . Let s denote the number of required servers to provide the desired QoS. Let $\rho = \lambda/(s\mu)$ be the traffic intensity. The system is assumed to be stable and in steady state, i.e., $\rho < 1$. The waiting time in the queue is denoted W , and the waiting time distribution is denoted $Pr(W < t)$. The delay probability (i.e., $Pr(W > 0)$) or the probability that the queue is nonempty) is denoted $\pi(G,s)$ when the service time distribution is G and $\pi(M,s)$ when the service time distribution is negative exponential.

Note that $\pi(M,s)$ is the well-known Erlang delay probability and is given as:

$$\Pi(M,s) = \frac{(sp)^s}{s!(1-\rho)} \left[\sum_{j=0}^{s-1} \frac{(sp)^j}{j!} + \frac{(sp)^s}{s!(1-s)} \right]^{-1} \quad (7)$$

And the wait time distribution is given as:

$$P(W < t) \cong 1 - \Pi(M,s) \exp \left\{ -\frac{s\mu(1-\rho)t}{R_G} \right\}, \quad t \geq 0 \quad (8)$$

Where R_G is given as:

$$R_G = \frac{R_D(1+c^2)}{(2R_D-1)J_G(s)+1} \quad (9)$$

And:

$$J_G(s) = \begin{cases} 1, & s=1 \\ \frac{s+1}{s-1} \left\{ \frac{1+c^2}{(s-1)\mu I_G(s)} - 1 \right\}, & s \geq 2 \end{cases} \quad (10)$$

$$I_G(s) = \int_0^{\infty} \{1 - G_e(t)\}^s dt, \quad s \geq 1 \quad (11)$$

$$G_e(t) = \mu \int_0^t \{1 - G(u)\} du, \quad t \geq 0 \quad (12)$$

Also:

$$R_D = \frac{1}{2} \{1 + f(s)g(\rho)h(s,\rho)\} \quad (13)$$

Where:

$$f(s) = \frac{(s-1)(\sqrt{4+5s}-2)}{16s} \quad (14)$$

$$g(\rho) = \frac{1-\rho}{\rho} \quad (15)$$

And:

$$h(s,\rho) = \xi(s, a(\rho))\eta(b(s),\rho) \quad (16)$$

With:

$$\xi(s,x) = \sqrt{1 - \exp\left(-\frac{2x}{s-1}\right)}, \quad x \geq 0 \quad (17)$$

$$\eta(y,\rho) = 1 - \exp\left(-\frac{\rho y}{1-\rho}\right), \quad y \geq 0$$

$$a(\rho) = \frac{25.6}{\{g(\rho)\eta(\beta,\rho)\}^2} \quad (18)$$

$$b(s) = \frac{s-1}{(s+1)f(s)\xi(s,\alpha)} \quad (19)$$

Where α and β are non-negative numbers satisfying the following relationship:

$$\alpha\beta^2 = 25.6 \quad (20)$$

Recommended values for α and β are:

$$\alpha = 2.2$$

$$\beta = \sqrt{25.6/2.2} = 3.41 \quad (21)$$

From Eq. 8, and with specified values for the maximum allowed delay and the desired reliability, the required number of resources or servers s can be computed. Similarly, the required number of servers to provide a desired mean delay or maximum allowed jitter, etc., can be computed.

Waiting Time Distribution and Network Planning With Priority Classes

This section examines the queue waiting distribution when there are two call or packet classes, namely, priority class 1 (e.g., handover calls) and priority class 2 (e.g., new originating calls). A FIFO discipline with nonpreemptive priority is assumed. Using the approximations of Hokstad [20] and Williams [21] yields the following results.

First, define:

λ_1 = arrival rate for priority class 1 (e.g., handover calls)

λ_2 = arrival rate for priority class 2 (e.g., new originating calls)

λ^{R_1} = retransmission rate for priority class 1 (e.g., handover calls)

λ^{R_2} = retransmission rate for priority class 2 (e.g., new originating calls)

$\lambda = (\lambda_1 + \lambda^{R_1}) + (\lambda_2 + \lambda^{R_2})$

G^* = Laplace transform of the arbitrary service time distribution G

g_i = i th moment of G

ρ_1 = traffic intensity for priority class 1 (e.g., handover calls)

ρ_2 = traffic intensity for priority class 2 (e.g., new originating calls)

ρ^{R_1} = traffic intensity for retransmitted priority class 1 (e.g., handover calls)

ρ^{R_2} = traffic intensity for retransmitted priority class 2 (e.g., new originating calls)

$\rho = (\rho_1 + \rho^{R_1}) + (\rho_2 + \rho^{R_2})$

$\pi(M,s)$ = delay probability corresponding to the negative exponential service time

The nonexponential nature of the service time distribution renders analysis of the underlying embedded multidimensional Markov chain almost intractable.

Then, for an $M/G/s$ nonpriority queue with arrival rate λ and traffic intensity ρ , Hokstad [20] provides the following approximations, which are based on a Laplace transform approach. The Laplace transform of queue waiting time distribution $W^*(S)$ is given as:

$$W^*(S) = 1 - \Pi(M, s) + \frac{\lambda[1-G^*(S/s)](1-\rho)\Pi(M, s)}{\{S-\lambda[1-G^*(S/s)]\}\rho} \quad (22)$$

And the Laplace transform of the busy period (i.e., when all servers are busy) $H^*(S)$ is the unique solution to:

$$G^*([\lambda(1-z) + S]/s) = z \quad |z| < 1 \quad (23)$$

Now, for the two-priority case, Williams [21] shows that the Laplace transform of the waiting time distribution of priority 2 calls or packets is given as:

$$W_2^*(S) = 1 - \Pi(M, s) + \frac{[1-H_1^*(S/s)](1-\rho)(\lambda_1+\lambda_2)\Pi(M, s)}{\{S-\lambda_2[1-H_1^*(S/s)]\}\rho} \quad (24)$$

Where H_1^* is the unique solution to:

$$G^*([\lambda_1(1-z) + S]/s) = z \quad |z| < 1 \quad (25)$$

The first moment of W_2 can also be found as:

$$\frac{\pi(M, s)g_2}{2(1/\mu)s(1-\rho)(1-\rho_1)} \quad (26)$$

Similarly, for the priority 1 class, the Laplace transform of the waiting time distribution W_1^* is given as:

$$W_1^*(S) = 1 - \Pi(M, s) + \frac{[1-G^*(S/s)](1-\rho_1)\lambda_1\Pi(M, s)}{\{S-\lambda_1[1-G^*(S/s)]\}\rho_1} \quad (27)$$

Note that Eqs. 7, 24, and 27 can be used to compute the number of servers or resources s needed to provide the desired maximum (or average or certain percentile) tolerable delays, time delay jitters, etc.

CONCLUSIONS

This paper has looked at the problem of network sizing for packet data applications with arbitrary general service time distributions. Using results from advanced queuing theory, it has been shown how to compute required resources (e.g., number of basic rate codes, channel elements, frequency channels, time slots)

to provide packet data services with guaranteed QoS performance. The delay components examined correspond to the radio access network (RAN) portion of a wireless network. There are also delays corresponding to the core network (CN), Internet, and application servers (ASs); these were not considered and are beyond the scope of this current work. It should be noted that CN delays are typically predictable and fixed. Performance acceleration techniques, such as proxy servers and caching, could help with AS performance issues.

The analysis considered three main QoS metrics: maximum allowable packet delays with a given reliability, maximum allowable delay jitters, and average allowable packet delays. The effects of handover were also considered using a two-class nonpreemptive priority queuing model. Note that the same process can be extended to the m -class priority case. The effects of channel impairments such as fading were taken into account by adjusting the actual mean arrival rate to the effective mean arrival rate, incorporating the effects of retransmission. Part 2 of this paper will address some simulation results and a corresponding network dimensioning tool. ■

TRADEMARKS

3GPP is a trademark of the European Telecommunications Standard Institute (ETSI) in France and other jurisdictions.

cdma2000 is a registered trademark of the Telecommunications Industry Association (TIA-USA).

REFERENCES

- [1] R. Koodli and M. Puuskari, "Supporting Packet-Data QoS in Next-Generation Cellular Networks," *IEEE Communications Magazine*, February 2001, pp. 180-188.
- [2] H.M. Chaskar and U. Madhow, "Statistical Multiplexing and QoS Provisioning for Real-Time Traffic on Wireless Downlinks," *IEEE Journal on Selected Areas in Communications*, Vol. 19, No. 2, February 2001, pp. 347-354.
- [3] S. Dixit, Y. Guo, and Z. Antoniou, "Resource Management and Quality of Service in Third-Generation Wireless Networks," *IEEE Communications Magazine*, February 2001, pp. 125-133.
- [4] O. Gurbuz and H. Owen, "Dynamic Resource Scheduling Schemes for W-CDMA Systems," *IEEE Communications Magazine*, October 2000, pp. 80-84.
- [5] N. Dimitriou and R. Tafazolli, "Quality of Service for Multimedia CDMA," *IEEE Communications Magazine*, July 2000, pp. 88-94.

The delay components examined correspond to the RAN portion of a wireless network. There are also delays corresponding to the CN, Internet, and ASs.

- [6] B. Epstein and M. Schwartz, "Predictive QoS-Based Admission Control for Multiclass Traffic in Cellular Wireless Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 3, March 2000, pp. 523–534.
- [7] T. Liu and J.A. Silvester, "Joint Admission/Congestion Control for Wireless CDMA Systems Supporting Integrated Services," *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 6, August 1998, pp. 845–857.
- [8] J. Misic and T.K. Bun, "Adaptive Admission Control in Wireless Multimedia Networks Under Nonuniform Traffic Conditions," *IEEE Journal on Selected Area in Communications*, Vol. 18, No. 11, November 2000, pp. 2429–2442.
- [9] A. Chockalingam, W. Xu, M. Zorzi, and L.B. Milstein, "Throughput-Delay Analysis of a Multichannel Wireless Access Protocol," *IEEE Transactions on Vehicular Technology*, Vol. 49, No. 2, March 2000, pp. 661–671.
- [10] V.S. Frost and B. Melamed, "Traffic Modeling for Telecommunications Networks," *IEEE Communications Magazine*, March 1994, pp. 70–81.
- [11] L. Kleinrock, *Queuing Systems, Volume I: Theory and Queuing Systems, Volume II: Computer Applications*, Wiley InterScience®, New York, 1975 and 1976 (respectively).
- [12] L. Kleinrock, *Queuing Systems: Problems and Solutions*, John Wiley & Sons, 1996.
- [13] 3GPP, A Global Initiative (<http://www.3GPP.org>).
- [14] Third Generation Partnership Project 2 (<http://www.3GPP2.org>).
- [15] ETSI TR 101 112, "Selection Procedures for the Choice of Radio Transmission Technologies of the Universal Mobile Telecommunications System UMTS (UMTS 30.03)," Version 3.2.0, April 1998.
- [16] D. Staehle et al., "Source Traffic Modeling of Wireless Applications," Report No. 261, Universität Würzburg, June 2000.
- [17] N. Antunes et al., "An Integrated Traffic Model for Multimedia Wireless Networks," *Computer Networks*, Vol. 38, 2002, pp. 25–41.
- [18] T. Kimura, "Approximations for the Delay Probability in the M/G/s Queue," *Mathematical Computer Modeling*, Vol. 22, Nos. 10–12, 1995, pp. 157–165.
- [19] T. Kimura, "A Transform-free Approximation for the Finite Capacity M/G/s Queue," *Operations Research*, Vol. 44, No. 6, 1996, pp. 984–988.
- [20] P. Hokstad, "Approximations for the M/G/m Queue," *Operations Research*, Vol. 26, 1978, pp. 510–523.
- [21] T.M. Williams, "Nonpreemptive Multi-Serve Priority Queues," *Journal of the Operational Research Society*, Vol. 31, 1980, pp. 1105–1107.

BIOGRAPHY



Rasoul Safavian brings more than 15 years of experience in the wired and wireless communications industry to his position as Bechtel Telecommunications' new executive director of technology and network planning. He is charged with establishing the overall technical vision for Bechtel's North American market as well as guiding and directing to its specific technological activities. In fulfilling this responsibility, he will be well served by his background in cellular/PCS, fixed microwave, satellite communications, wireless local loops, and fixed networks; his working experience with major 2G, 2.5G, 3G, and 4G technologies; his exposure to the leading facets of technology development as well as its financial, business, and risk factors; and his extensive academic, teaching, and research experience.

Before joining Bechtel in June 2005, Dr. Safavian oversaw advanced technology research and development activities, first as vice president of the Advanced Technology Group at Wireless Facilities, Inc., then as chief technical officer and vice president of engineering at GCB Services. Earlier, over an 8-year period at LCC International, Inc., he progressed through several positions. Initially, as principal engineer at LCC's Wireless Institute, he was in charge of CDMA-related programs and activities. Next, as lead systems engineer/senior principal engineer, he provided nationwide technical guidance for LCC's XM satellite radio project. Then, as senior technical manager/senior consultant, he assisted key clients with the design, deployment, optimization, and operation of 3G wireless networks.

Dr. Safavian is quite familiar with the electrical engineering departments of three universities: The George Washington University, where he has been an adjunct professor for several years; Purdue University, where he received his PhD in Electrical Engineering, was a graduate research assistant, and was later a member of the visiting faculty; and the University of Kansas, where he received both his BS and MS in Electrical Engineering and was a teaching and a research assistant. He is a senior member of the Institute of Electrical and Electronics Engineers and a past official reviewer of its transactions in certain categories.