A Flexible Model for Dimensioning Mixed Service 3G Wireless Networks

Keith Butterworth[†], Mansoor Shafi[††], Peter Smith[†††]

[†]Alcatel New Zealand, Auckland, New Zealand, Keith.Butterworth@telecom.co.nz
[††]Telecom New Zealand, Wellington, New Zealand, Mansoor.Shafi@telecom.co.nz
[†††]Department of Electrical and Computer Engineering, The University of Canterbury,
Christchurch, New Zealand, p.smith@elec.canterbury.ac.nz

Abstract— **In this paper we present an approximate closed form solution for dimensioning mixed service 3G wireless networks for defined service and quality constraints. Both web-browsing services alone, as well as a mixture of web and circuit voice services are considered. Dimensioning is achieved through knowledge of the distribution of the latency delay experienced by end users of the 3G network. The parameters defining the distribution are shown to be dependent on a range of network and service parameters. This allows the dimensioning of 3G wireless networks under a wide range of quality and service conditions.**

I. INTRODUCTION

For wireless networks that are capable of carrying packet switched traffic, the resources of the carrier are typically simultaneously shared between many users. Loading of the network causes delay. This delay or latency, τ, is the time the packets need to be held in the base station buffer before air interface capacity is available to transmit the packets over the air interface. The amount of latency that can be tolerated is service dependent [1]. The latency is a random variable and is influenced by the following parameters:

- Air interface throughput capacity, R (kbits/s)
- Volume of data sent in the busy hour (BH), V (Mbits)
- Traffic model parameters of the data session.

The aim of the system designer is to ensure that the end users achieve a quality of service in line with that guaranteed by the operator. This could be achieved by ensuring that for a given percentage of the time (say 95%) the value of τ, is less than a target value, subsequently referred to as $\tau(0.95)$. In this paper we address the following questions:

- What is a suitable model for the probability density function (pdf), $f_\tau(\tau)$, of the latency delay and what are its key parameters?
- How can we relate the parameters of $f_\tau(\tau)$ to other system parameters such as air interface throughput, data volume to be transmitted, the traffic model, etc.
- How can the resulting model be used to give a complete solution to dimensioning practical wireless systems?

In order to estimate $\tau_{(0.95)}$, we choose to model the pdf of τ. This approach also allows the use of the pdf to look at other issues in future, e.g. the estimation of $\tau_{(0.99)}$ and the mean of τ. In addition, the shape of $f_\tau(\tau)$ in response to the variables constituting a packet session can be studied.

Analysis of packet-switched networks is possible, but only for very simple traffic models. For example in the 1970's and 1980's there were many contributions by Wong and Lam et al [2-5]. Such approaches are not possible here as they rely on a simple Poisson process of arrivals and often focus on steady state values such as mean delay. The complexity of the web-browsing model substantially violates these assumptions and furthermore, we are interested in the delay distribution. For these reasons we resort to model fitting based on simulated data rather than an exact approach for determining $f_\tau(\tau)$. Our results show that that the latency may be well modeled by a simple mixture distribution where $\tau=0$ with probability p and with probability $1-p$ we have $\tau>0$ with pdf

$$f_\tau(\tau) = (1/\mu)\exp(-\tau/\mu)\,, \tag{1}$$

Hence, when $\tau>0$ the delay is modeled by an exponential distribution with mean, μ. In our paper we relate p and μ to air interface throughput, R, volume of the data transmitted in the BH, V, and service class (data rate) of the data sessions. This is a very simple model and is ideal for dimensioning practical 3G wireless systems. We give examples for Cdma 1X and WCDMA wireless networks. The major contribution here is the remarkable simplicity and generality of the models used. This makes the results robust and more likely to be applied usefully to other cases, e.g. 4G and beyond.

The format of the paper is as follows. Sections II and III describe the traffic and simulation models respectively. The model for the latency delay pdf $f_\tau(\tau)$ and practical dimensioning examples are given in section IV. Finally the conclusions are given in section V.

II. TRAFFIC MODEL

In this paper we consider both voice call and web browsing session models, as shown in Figs 1-2. The traffic parameters for these models have been described in [6] and subsequently measured in [7] for the web model.

0-7803-8938-7/05/$20.00 (C) 2005 IEEE

A. *Voice Traffic Model*

Voice calls are assumed to have a uniformly distributed random start time over the BH, with mean call durations of 120 seconds. Voice activity is modeled as an on-off process, with active and silent periods with durations, $D_{ativity}$, being generated according to an exponential distribution with a mean of 3 seconds (a 50% voice activity factor). Over the period of a voice call there is a uniform sequence of voice frames arriving at time intervals of, D_{frame}, which for practical systems is of the order of 20 msecs. During active periods the frames are full rate, with size, S_{va}, while during silent periods the frames are assumed to be eighth rate, with size, S_{vs}. The frame sizes are technology dependent and are set inline with the air interface throughput capacity, R, and the voice circuit limit.

The parameter, R, is the total throughput shared between all users sharing a single carrier. This throughput is determined under the assumption that users can tolerate a queuing delay that tends towards infinity but is service dependent. Consequently it is the capacity limit in terms of throughput for a carrier. This value is likely to vary for different morphologies, geographic user distributions, network deployment configurations, etc. However, a representative value can be estimated for each network deployment.

For a particular technology, the maximum number of voice circuits, Cct_{Max}, a carrier can support, is given by, $Cct_{Max} = Cct_{Lim}$ * SHO (Soft Handoff Overhead), where Cct_{Lim} is the air interface circuit limit and SHO (if applicable) is the soft handoff overhead (both of which are typically well understood). The mean data rate for each voice call is given by, $RV_{mean} = R \,/\, Cct_{Max}$. This effectively calibrates the voice loading imposed by each user to the overall throughput, R, thus, ensuring that each voice user appropriately loads the carrier when mixed with web users. It is assumed that voice calls have a 50% activity factor. Therefore, $(S_{va} + S_{vs}) / 2 = RV_{avg} * D_{frame}$. Since, S_{va} and S_{vs} are frame volumes for full and eighth rate frames respectively, $S_{vs} = S_{va} / 8$. Substituting yields, $(S_{va} + S_{va} / 8) / 2 = R * D_{frame} / (Cct_{Lim} * \text{SHO})$. Thus, $S_{va}= (2R * 8/9 * D_{frame}) / (Cct_{Lim} * \text{SHO})$ in kbits.

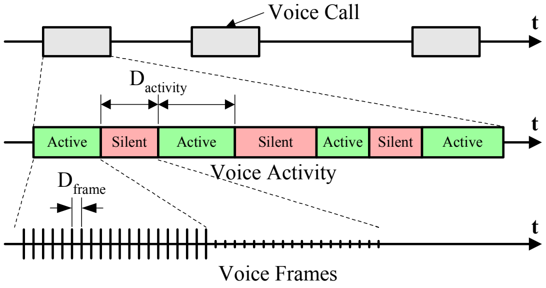

Figure 1. Voice call model.

B. *Web Browsing Traffic Model*

A commonly used web-browsing traffic model used is that given by [6]. This model assumes that web sessions consist of several packet calls interspersed by reading times. A packet call itself contains one or more packets that are transmitted according to a specific service class – the source data rate.

The following is a description of the random processes driving the traffic model. The session arrival process is modeled as a uniformly distributed random start time over the BH, with the assumption that each user has only one session. The number of packet calls per session, N_{pc}, is modeled as a geometric random variable with mean, μ_{Npc}. The reading time in seconds between packet calls, D_{pc}, is modeled as a geometric random variable with mean μ_{Dpc}. The number of packets within a packet call, N_d, is modeled as a geometric random variable with mean, μ_{Nd}. The inter arrival time, in seconds, within a packet call, D_d, is modeled as a geometric random variable with mean, μ_{Dd}. The size of a packet, S_d, is modeled as a truncated Pareto variable, i.e. S_d=min(P,m), where P is a Pareto variable and m is the maximum allowed packet size, taken as 66666 bytes. The pdf of P (for P>0) is given by:

$$f_P(P) = (\alpha.k^{\alpha})/(P^{\alpha+1}). \quad (2)$$

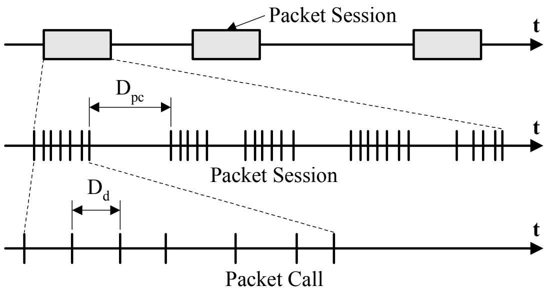

Figure 2. A WWW packet service session.

The values of all the parameters, α, k, μ_{Npc}, μ_{Dpc}, μ_{Nd}, and μ_{Dd} have been measured and published in [6-7]. These values are reproduced in Table 1 for the sake of completeness.

TABLE 1
PARAMETERS FOR PACKET TRAFFIC MODEL

Service type b(*j*) *j*=1,4	μ_{Npc}	μ_{Dpc} seconds	μ_{Nd}	μ_{Dd} seconds	Parameters for determining *P*
b(1) = 32 kbps	5	12	25	0.0125	k=81.5, α = 1.1
b(2) =64 kbps	5	12	25	0.0625	k=81.5, α = 1.1
b(3) = 144 kbps	5	12	25	0.0277	k=81.5, α = 1.1
b(4) = 384 kbps	5	12	25	0.0104	k=81.5, α = 1.1

III. SIMULATION METHODOLOGY

Simulations were undertaken for two cases, the first, for web traffic only, and the second, for a mixture of voice and web traffic.

A. *Simulation Parameters*

A range of values for, R, V and b(j), that influence the queuing delay were considered. The volume, V, in the BH can range between zero and R x 3.6 Mbits. For the purpose of comparison between simulation results with different R, the loadings are expressed as average percentage carrier utilizations, U, where:

$$U_W + U_V = U = V / (3.6R), \quad (3)$$

and U_W and U_V are the percentage carrier utilization due to web and voice traffic, respectively.

1) Web Browsing Traffic Only

The objective of web traffic only simulations was to determine the queuing delay characteristics for increasing web traffic loads. The web traffic only simulations undertaken considered the following combination of parameters. Eight air interface throughputs, R, between 30 and 2500kbps were used. For each R, the four service classes b(1-4) were used and nine average % carrier utilizations, U_W, between 10 and 90%, were implemented for each combination of R and b(j). Finally, 20 simulation runs were undertaken for each combination of U_W, R and b(j).

Table 2 summarizes the parameters, R and b(j) for some example combinations representative of Cdma 1X and UMTS air interfaces (a subset of the simulations undertaken). Consider as an example, service class b(2) for scenario I. For an air interface throughput of 250 kbps, 20 simulations were undertaken each of which considered utilizations, U_W, ranging between 10 and 90%: resulting in 20 sets of the latency delay pdf's for a range of U_W.

TABLE 2
SUMMARY OF WEB SCENARIOS CONSIDERED

Scenarios	**Air Interface Throughput, *R* (kbps)**	**Representative Technology**	**Service Class, b(*j*)**
I	250	CDMA 1X	b(1-4)
II	700	UMTS Rel. 99	b(1-4)

2) Mixed Voice and Web Browsing Services

The objective of the mixed voice and web simulations was to determine the queuing delay characteristics for increasing web traffic loads when there is an underlying voice traffic load. The mixed voice and web simulations undertaken considered a matrix of combinations of U_W and U_V. Air interface throughputs, R, between 125 and 2500kbps were considered. Web traffic was always of service class b(2).

For each value of R, the following combinations of parameters were considered. Eight values of voice utilization, U_V, were used between 0.1 and 0.8; and for each value U_V, web utilizations, U_W, between 0.1 and (0.9 - U_V) were implemented. Finally, 20 simulation runs were undertaken for each combination of R, U_V, and U_W.

As an example, consider R = 250kbps and $U_V = 0.3$. For this example, 20 simulations were undertaken for six values of U_W between 0.1 and 0.6, resulting in 20 sets of the latency delay pdf's for a range of U_W.

The overall utilization, U, for each value of U_V ranges between U_V and 0.9.

B. Generation of Traffic

For a given scenario at specified values of U_W and U_V, many voice calls and web sessions were generated in accordance with [6] for many voice and web users, N_{VUsers} and N_{WUsers}, respectively. The traffic for each voice user is comprised of many packets with 20msec inter-arrival times. Packet sizes are either S_{va} or S_{vs} depending on whether they are full or eighth rate frames. The traffic for each web user consists of:

- N_{pc} packet calls, each consisting of N_d packets with packet arrival times defined relative to the start of their respective packet call;
- the start time of the first packet call for the user;
- $N_{pc} - 1$ reading times, D_{pc}, between N_{pc} packet calls;
- the size of each packet, S_d, defined for every packet.

C. Traffic Aggregation and Queuing

The queuing delay, τ, is calculated for each second over the BH (t= 1 to 3600). For each t, the volume of data offered for each user and any data not sent (buffered) from previous time step/s is input to a weighted fair queuing (WFQ) algorithm [8]. A WFQ algorithm was selected because the majority of practical 3G wireless networks today offer little more than best effort resource allocation for packet data. For each t, there is an available volume of R kbits of data that can be carried. The WFQ algorithm allocates the available volume according to user priority weighting. Web and voice user weightings were taken as 1 and 99, respectively. This means that voice users are effectively circuit switched users, getting allocated all the bandwidth they require in preference to any contending web users. The web users are allocated any remaining bandwidth equally amongst themselves after the voice user demand is satisfied. Any traffic intended for a web user, not sent is buffered until the next time step. For each user, τ_{User} was calculated and is given by:

$$\tau_{\text{User}} = \text{Allocated User Bandwidth / Buffered User Data} \quad (4)$$

Combined queuing delays are calculated for all web users, τ_{Web}, and also for all voice users, τ_{Voice}, and are given by:

$$\tau_{Web\,or\,Voice} = \frac{\sum\limits_{Web\,or\,Voice\,Users} \tau_{User} \times Allocated\,User\,Bandwidth}{\sum\limits_{Web\,or\,Voice\,Users} Allocated\,User\,Bandwidth} \quad (5)$$

After calculating τ over the entire BH an estimate of the pdf, $f_\tau(\tau)$, can be made.

IV. RESULTS AND DISCUSSION

This section focuses on the development of a closed form model for the delay distribution experienced by web traffic with no voice traffic. The analysis and models for mixtures of voice and web traffic are very similar. Due to space limitations only key results are presented for voice-web mixtures.

A. Latency Delay PDF

Due to varying inter-arrival times between packets, non-exponential packet lengths and an arrivals process which is "batch-like" (sessions contain batches of packets), packet arrivals do not conform to a Poisson process. Consequently, a closed form solution for $f_{\tau Web}(\tau_{Web})$ is almost certainly intractable since the queuing situation is not amenable to analysis without substantial simplifying assumptions. As a result, we resort to model-fitting based on simulated data. A pdf of simulated τ_{Web} values is shown in Fig. 3. Note that the pdf is for the non-zero delays and exhibits a simple decay curve for $\tau_{Web}>0$. Consideration of the zero delay component and the shape of Fig. 3 suggests that the latency distribution may be well approximated by a simple mixture of a degenerate distribution at zero ($\tau_{Web}=0$) with an exponential for $\tau_{Web}>0$. The corresponding density function is given in (1). The excellent fit of this model is shown in Fig. 3 for a sample data set. Out of 5760 pdf's the model was accepted by a 99% confidence goodness of fit test 99.7% of the time.

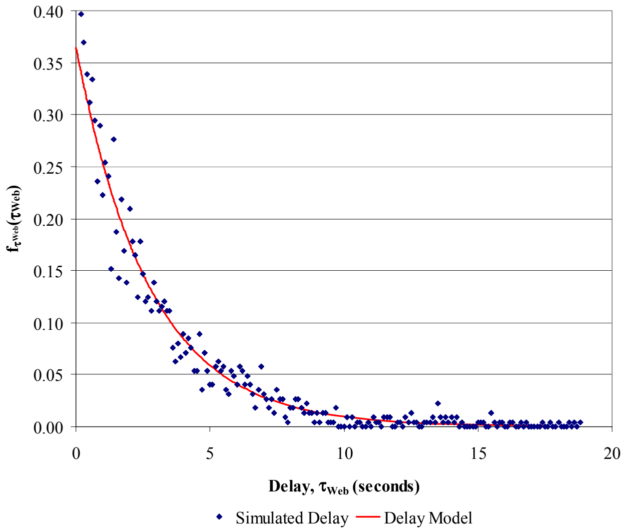

Figure 3. pdf for a sample data set (R=250, U_W=80%, U_V=0%) with the $f_\tau(\tau)$ model as defined by (1) superimposed only for τ>0.

B. *Relationship to System Parameters*

What remains is to relate the parameters p and μ in (1) to the system parameters, R, V and b(j).

Exploratory data analysis of the relationship between p and μ and V led to the following simple approximations:

$$p = 1\text{-}U^{1/b} \qquad 0 \le b \le 1 \qquad (6)$$

$$\mu = 1/(c(1\text{-}U)) \qquad 0 \le c \qquad (7)$$

Note that the parameters b and c in (6) and (7) hold for a given value of b(j) and R and are estimated by least squares. The fit of (6) and (7) was found to be excellent for all values of service class and air interface throughput considered. The next step is to relate the parameters b and c to the traffic parameters. This was performed by exploratory data analysis. The most satisfactory models found were:

$$b = g + f\ \log_{10}R \qquad (8)$$

$$c = h\,R \qquad (9)$$

f, g, and h are given in Table 3 for services classes b(1-4).

The fit of (8) and (9) to the b and c parameters are presented in Figs 4 and 5 for all values of R and b(j) considered.

TABLE 3
PARAMETERS f, g, AND h FOR DIFFERENT SERVICE CLASSES

Service type, b(*j*) *j*=1,4	***f***	***g***	***h***
b(1) = 32 kbps	0.336	1.38	0.0082
b(2) = 64 kbps	0.370	1.39	0.0078
b(3) = 144 kbps	0.388	1.48	0.0068
b(4) = 384 kbps	0.383	1.50	0.0056

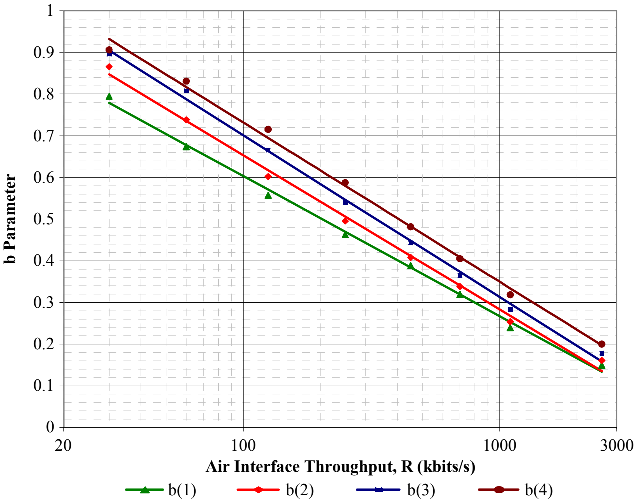

Figure 4. Fit of web only simulation results to the relationship between R and b defined in (8), for all service classes.

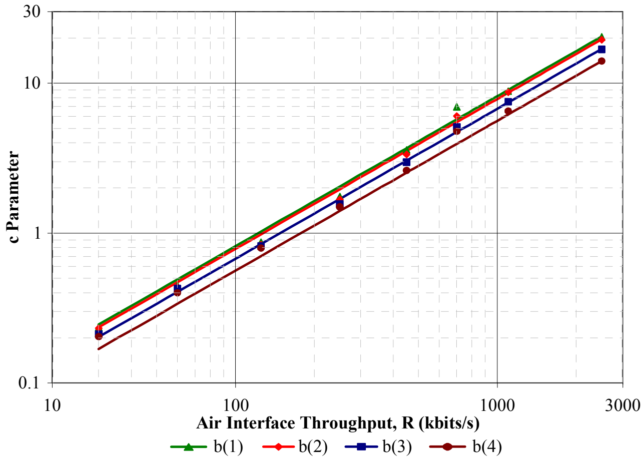

Figure 5. Fit of web simulation results to the relationship between R and c defined in (9), for all service classes.

Integrating (1) gives the distribution function for $\mu\geq0$:

$$F_{\tau_{Web}}(\tau_{Web}) = p + (1-p)(1-\exp(-\tau_{Web}/\mu))\,. \qquad (10)$$

Substituting using equations (6)-(9) gives a single model for the delay distribution for web only traffic:

$$\mathrm{F}_{\tau_{\mathrm{Web}}}(\tau_{\mathrm{Web}}) = 1\text{-}\ U^{1/(g\ +f\ \log_{10}R\ )} \times \exp(\text{-}\tau_{\mathrm{Web}}\ h\,R\ (1\text{-}U)) \quad (11)$$

Note that (11) provides a complete solution given f, g and h for the service class.

Now we must evaluate the models for $f_{\tau Web}(\tau_{Web})$ from the point of view of their overall accuracy in predicting delay probabilities based on R, b(j) and V. For all simulated data sets of service class b(2) we present in Fig. 6 a comparison of simulated $\tau_{Web}(0.95)$ and the predicted values using (11) and the *f*, *g*, and *h* parameters for b(2) from Table 3.

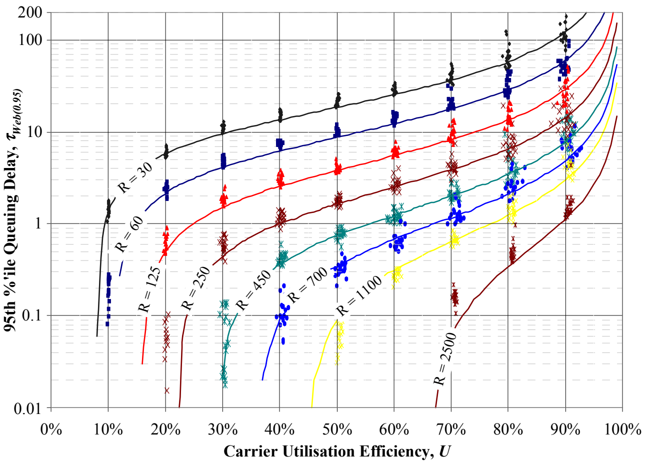

Figure 6. $\tau_{Web}(0.95)$ for web only simulations, service class b(2). Data points are from simulations. Solid lines are based on (11), *f*, *g* and *h* from Table 3.

As can be seen from Fig. 6, the model curves follow the data remarkably well for such a simple model. For lower *U*, the model tends to slightly underestimate the delay. When the model is compared with the raw data the difference is on average within 1.7% of the simulated data.

A very similar analysis can be undertaken for mixtures of circuit voice and web traffic. It can be shown that the density function given in (1) holds for web queuing delays, τ_{Web}, when mixed with voice traffic. Furthermore, the fit of (6) and (7) to *p* and μ in (1) are good for all combinations of voice and web traffic considered. Similarly, (11) and the *f*, *g*, and *h* parameters can also be derived.

For all simulated data sets of service class b(2) and U_V=0.3 we present in Fig 7 a comparison of simulated $\tau_{Web}(0.95)$ (data points) and the predicted values (solid lines) using (11) where the *f*, *g*, and *h* parameters are fitted to the web-voice simulation results (similar curves can be plotted for different voice utilizations, U_V). Also included for comparison purposes only (dashed lines), are the model curves from Fig 6 for the equivalent web only model values.

As can be seen from Fig 11, the simulated values (data points) are more variable than the data only case. However, the model curves (solid lines) still follow the simulated values remarkably well. Thus, illustrating that (11) is equally applicable to a mixture of web and circuit voice traffic as it is to web traffic only.

Comparison of the solid and dashed curves in Fig 7 illustrates the impact of mixing voice and web traffic. A consistent trend is for the same carrier utilization, is for $\tau_{Web}(0.95)$ for mixed scenarios to be less than that for web only scenarios (solid curves always lie below the dashed curves).

C. Dimensioning Mixed Service 3G Networks

Consider a hypothetical scenario where a network operator is faced with a constraint of $\tau_{Web}(0.95)$=1 sec for b(2) web traffic over either a Cdma 1X carrier, *R*=250, or a UMTS carrier, *R*=700. They want to know how much traffic can be carried while maintaining $\tau_{Web}(0.95)$=1 under two scenarios. The first with web traffic only (a dedicated carrier) and the second with a voice loading of U_V=0.3 and the remainder web traffic (a mixed carrier).

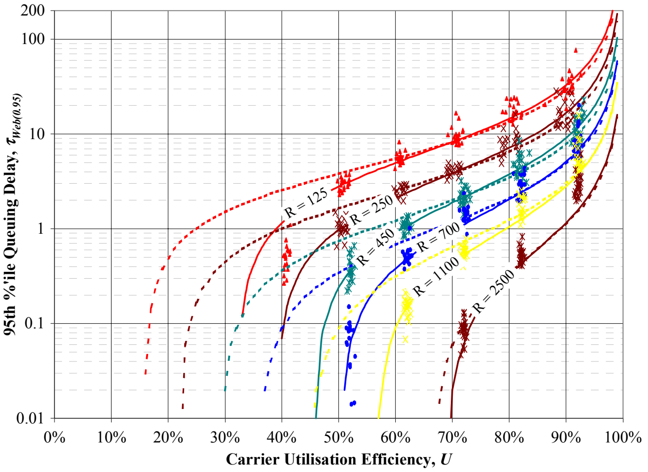

Figure 7. $\tau_{Web(0.95)}$ for mixed voice web traffic, U_V= 0.3 and service class b(2). Data points from simulations. Dashed lines are the solid lines from Fig 6, solid lines are based on (11) fitted to the web-voice simulation results.

For a dedicated carrier, referring to the solid curves in Fig 6 for $\tau_{Web}(0.95)$=1, the values of *U* for Cdma 1X (R=250) and UMTS (*R*=700) are 0.40 and 0.66, respectively. For a mixed carrier, referring to the solid curves in Fig 7 for $\tau_{Web}(0.95)$=1 the value of *U* for Cdma 1X and UMTS are 0.50 and 0.70, respectively.

For the case of Cdma 1X, this means that with a dedicated carrier the network can be operated on average at 40% of *R*, while maintaining the required $\tau_{Web}(0.95)$. For a mixed carrier this increases to 50% of *R*, meaning the carrier is being utilized more efficiently. This corresponds to a gain in overall traffic volume carried in the BH of 90Mbits.

In the case of UMTS the gains realizable with a mixed carrier are smaller (4% gain) due to the multiplexing gains associated with the higher *R*.

This is a very powerful closed form solution which can be applied to a wide variety of situations for different combinations of *R*, U_V, $\tau_{Web}(0.95)$, service classes, etc, without having to resort to very time consuming simulations to verify that the service quality constraints are achievable.

V. Conclusions

In this paper we have presented a simple closed form model for accurately estimating the queuing delay in a wireless packet network for web traffic only and mixtures of web and circuit voice traffic. The model parameters can be simply related to key parameters, namely air interface throughput, volume of voice and web traffic, service class, etc. The model avoids the need for time-consuming simulations and can be extended to cases not considered here. In addition since the delay pdf is modeled itself, any system feature which depends on the delay can be evaluated. In particular the use of equation (11) leads to a complete solution for dimensioning practical 3G wireless systems allowing the system operator, amongst other things, to understand system performance, assess trade offs between dedicated and mixed carrier deployments, deploy only enough capacity to satisfy the delay requirements for the anticipated traffic loading, etc.

A natural extension of this work would be the consideration of a mixed service model that includes other traffic types such as video traffic.

Acronyms

τ	delay or latency
R	air interface throughput capacity, (kbits/s)
V	volume of data sent in the busy hour (BH), (Mbits)
p	probability of zero queuing delay
μ	mean of exponentially distributed queuing delay
b(j)	web browsing service class
U	average percentage carrier utilization

References

[1] ITU-T Y.1540, “Network performance objectives for IP-based service”, May 2002.

[2] J. W. Wong, "Distribution of end-to-end delay in message-switched networks," Comput. Networks, vol. 2, Feb. 1978.

[3] G. Gopal and J. W. Wong, "Delay analysis of broadcast routing in packet-switched networks," IEEE Trans. on Computers, vol. 30, No. 12, pp. 915-922, Dec. 1981.

[4] S. S. Lam, "Store-and-forward buffer requirements in a packet-switching network," IEEE Trans. Commun. vol. 24, No. 4, pp. 394-403, Apr. 1976.

[5] S. S. Lam, "Delay analysis of a Time Division Multiple Access (TDMA) Channel," IEEE Trans. Commun. vol. 25, No. 12, pp. 1489-1494, Dec. 1977.

[6] “Selection procedures for the choice of radio transmission technologies of the UMTS, ETSI” TR 101 112 v 3.2.0 1998-04 (UMTS 30.03 version 3.2.0).

[7] Tapani Nieminen, “Report on WWW – traffic measurements and modeling in 1999”, Helsinki, University of Technology, 20 April 2000.

[8] Demers, A., Keshav, S., and Shenker, S., "Analysis and Simulation of a Fair-queueing Algorithm", Proc. ACM SIGCOMM '89, Sept 1989: 1-12.