Tunable Preemption Controls for a Cellular Emergency Network

Jiazhen Zhou and Cory Beard
Department of Computer Science Electrical Engineering
University of Missouri - Kansas City, Kansas City, MO 64110

Abstract— **In this paper we introduce the preemption threshold based strategy to cellular networks for supporting emergency traffic. With the new strategy, in addition to providing immediate access for emergency traffic, a certain amount of resources can be protected for public traffic by tuning the preemption threshold. Also, with the assumption that preempted users will restart the session after resumption, we find that a conservation law holds for the average duration of calls that succeed or fail to finish. Numerical results about the evaluation of this strategy are presented and show an obvious improvement over related strategies.**

Keywords: Admission control; emergency traffic; preemption; queuing; gross service time; channel occupancy; handoff; expiration; scheduling

I. INTRODUCTION

After disaster events happen, tremendous stress is placed on networks due to the rise in traffic demand, including demand from public and emergency staff. As pointed out in [1], network demand can be up to 10 times of normal. Among the traffic demands, emergency traffic should be given special priority for saving life and property. In recent years, special focus has been made to prioritize calls in wireless cellular networks [1]. And as 3G and 4G technologies emerge, methods to prioritize connection admission will be important to deploy there as well.

Nyquetek Inc. has prepared a evaluation report [1] of algorithms for the Wireless Priority Service (WPS) (provided for United States national security and emergency preparedness (NS/EP) workers) that are currently used by the main cellular network operators. A series of queueing and scheduling based policies are introduced and compared for supporting emergency traffic in the commercialized wireless cellular network. In their work, they emphasized that the priority of emergency traffic can be guaranteed by queueing emergency calls when no channels are immediately available. They show that normally the emergency traffic will not be more than 10% of the normal engineered load of a cell, so the admission of public traffic will not be affected much. Furthermore, if the emergency demand is high enough that it can take much of the available resources in a cell, queueing of both public calls and emergency calls is proposed and a round-robin like scheduling policy is provided: out of every four channels released, one will be used to serve emergency calls.

This work is supported by the United States National Science Foundation under CAREER Award ANI-0133605.

The main idea in Nyquetek's report is to guarantee that a portion of the resources will be available for public users while providing some priority for emergency traffic. But a problem with this scheme is that emergency traffic must wait significant time before being admitted, because when the network is heavily congested, emergency calls must wait 4 units of inter-departure time, which is unreasonable when there are urgent needs to save life or property. Waiting times are especially long if other emergency calls are already queued or if cell capacity is low (which results in longer inter-departure times).

The best strategy to guarantee real immediate access is preemption, which means emergency calls can break ongoing public calls and take the resources for emergency use. Yet, as mentioned in [3], a pure preemption strategy will cause immediate termination of public calls, and when emergency traffic is very high, the public calls may have little chance to finish since many might be preempted. This is harsh to public calls and that's why preemption is not adopted by the cellular network operators. So the combined preemption and queueing scheme was introduced in [3] to improve the success probability and soften the impact on public calls, by which preempted calls are queued and allowed to return when channels become available.

However, even with the best that can be achieved in [3], the channel occupancy for public traffic is not so well protected as what's proposed in [1]. So, in this paper we introduce preemption thresholds to the combined preemption and queueing scheme. This makes it feasible to adjust the channel occupancy for each class as we like, to provide the same protection for public calls as in [1], but immediate instead of delayed access for emergency calls.

The model we use incorporates preemption, queueing of preempted sessions, and expiration (i.e., impatience) of queued sessions. The applications of combined preemption and queueing schemes in wireless networks can also be seen in [7], [13]. In their works, real-time (voice) traffic can preempt resources from non-real-time (data) traffic. Each type of traffic consists of both originating and handoff traffic. However, the behavior of expiration of calls in the queues was not studied.

The main purpose of this paper is to guarantee immediate access of emergency traffic while protecting public calls when emergency traffic is high. The main contributions include: (1) The preemption threshold based strategy is introduced to provide higher flexibility of tuning the channel occupancy for public traffic. (2) With the assumption that preempted users will restart the session after resumption, we find a conservation

law holds for the average duration of calls that succeed or fail to finish. From which we can see that the average duration for those successful calls will be shorter than the case that there is no preemption.

In Section II, we introduce the preemption threshold based strategy in the wireless cellular network for the support of emergency traffic and derive important performance metrics. In Section III, an algorithm for tuning the threshold is introduced. In Section IV we show numerical results about threshold tuning and achieved channel occupancy, observe a conservation law pertaining to successful calls, and study the average service time for successful and failed calls. Finally, Section V concludes this paper.

II. Threshold-Based Preemption Control

A. Basic scheme and assumptions

As in [3], the main three types of voice calls we deal with are emergency calls, public handoff calls and public originating calls. There is no handoff for emergency calls; we assume most emergency users will be stationary within a disaster area.

The key idea of our work is to control and lessen the impact of preemption on low priority calls. To achieve this purpose, a popular way is to block further preemptions based on variant thresholds. The threshold can be based on service effort rendered for each session [9], [11], numbers of times of calls are preempted [3], or resource utilization taken by the high priority class [12]. In this paper, we take the strategy similar to [12], which sets thresholds on the number of channels occupied by emergency traffic. Also, there is no queue for emergency traffic, so we can prevent starvation of public traffic when emergency traffic is heavy. The main differences from the work in [12] are that we also consider loss due to blocking and expiration of preempted calls waiting in the queue, introduce priority queues for public handoff and originating traffic, and study the gross service time and channel occupancy.

The basic scheme we are to use is illustrated in Fig. 1. Similar to the scheme described in [3], the preempted calls will be put into a queue. To ensure public calls are well protected, we newly introduce a queue for public new calls. In the figure, Class 1 is for emergency calls, Class 2 is for public handoff calls, and Class 3 is for public calls originating from within the cell. When an incoming emergency call fails to find free channels, and if the number of occupied channels by emergency calls is less than the preemption threshold, it will preempt resources from ongoing public calls randomly (either from originating or handoff). The preempted calls are put into a queue. If the incoming call is a handoff call, it will be put into the same queue as preempted calls. And if it's an originating call, it will be put into the other queue for originating calls. When there are channels available later, one call from the queue will be served according to the FIFO policy. A priority scheduling policy will be used between the two queues, and we assume that handoff/preempted calls have higher priority over the originating calls. The queues are finite and customers can be impatient when waiting in the queue, so blocking and expiration are possible.

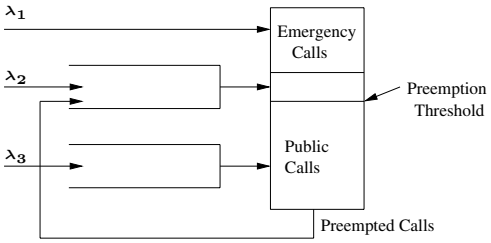

Fig. 1. Combined preemption and queueing scheme

We assume that the expiration times of both preempted calls and handoff calls waiting in the same queue are exponentially and identically distributed. The patience of customers is the same after each repeated preemption. And similar to what's used widely and assumed in [1], [13], all call durations and inter-arrival times are independently, identically, and exponentially distributed. If call durations are memoryless (i.e., exponentially distributed), this means that if at any point a call is interrupted, the remaining service time is still exponential with the same average service time as when it began. It is, therefore, reasonable to assume that the call will be restarted with re-sampling of the exponential random variable [8], also called a repeat-different approach [10].

In Fig. 2, a three dimensional example of the state diagram of this strategy is shown. In which the total number of channels $C = 2$, the queue sizes are $L_1 = L_2 = 1$. Each state is identified as (i,j,m,n), where i means the number of channels occupied by emergency calls, j represents the number of channels taken by public ones (regardless of being originating or handoff calls before being admitted by this cell), and m means the number of calls in queue 1 (handoff calls or preempted public calls), n means the number of calls in queue 2 (public originating calls). The arrival rate for emergency, handoff, and originating calls is λ_1, λ_2, λ_3, and the service rate for emergency and public calls is μ_1, μ_2 individually. In this paper we assume the average call durations for each class are the same and are denoted as $1/\mu$ in the later sections. This also means that the call duration in a single cell is exponentially distributed with mean $1/\mu$, whether the call ends in this cell or is handed off to another cell. The average expiration time for calls waiting in the queues are $1/\mu_{exp1}$ and $1/\mu_{exp2}$. State probabilities can be obtained by solving the global balance equations from this Markov chain directly.

B. Performance Evaluation

With the state probabilities solved, we can use them to calculate performance metrics like preemption probability, expiration probability, total loss probability, etc. In this paper our main concern is the system utilization and channel occupation for each class, so we just show the computation of system utilization, the channel occupancy for each class. The computation for other performance metrics like preemption probability, average number of preemption times, admission and success probability, etc. is basically the same as described in [3].

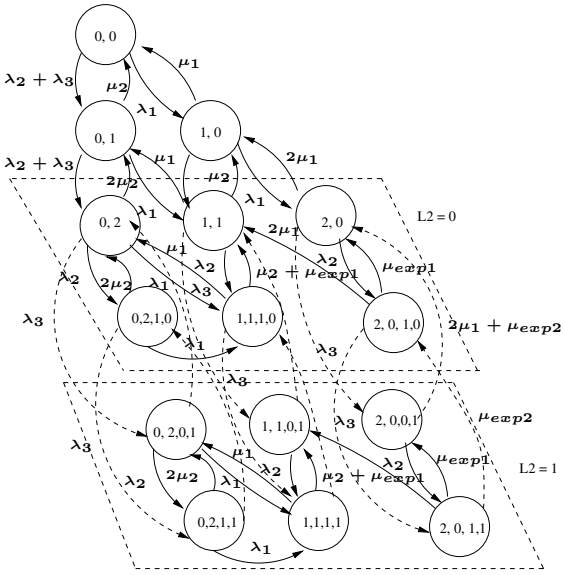

Fig. 2. State Diagram for 2 queues case

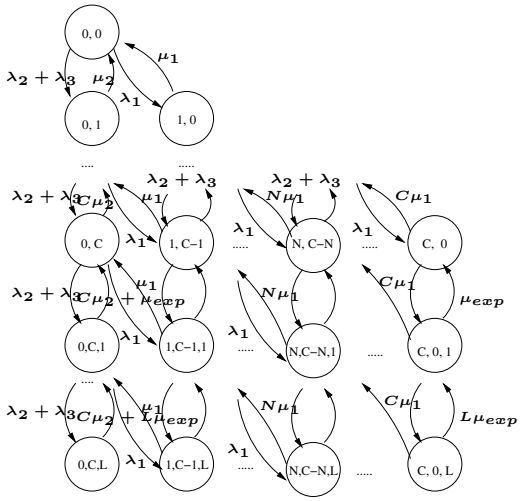

Fig. 3. State diagram with preemption threshold

(1) System utilization

A direct way to calculate the system utilization is :

$$SysUtil = 1 - \sum_{k=1}^{C-1} \sum_{i=0}^{k} \frac{(C-k)P(i,k-i,0,0)}{C} \quad (1)$$

The latter part of above formula is the average portion of channels left unused when in "not full" state.

Also, if we know the probability for calls that succeed to finish, the probability for calls that are admitted but forced to terminate, and the corresponding average call duration, we can also calculate the system utilization as:

$$\begin{aligned} SysUtil &= \frac{\lambda_1 P_{Succ,1}}{C\mu} + \frac{\lambda_2 P_{Succ,2}}{C\mu_{succ}} \\ &+ \frac{\lambda_2 P_{Fail,2}}{C\mu_{fail}} + \frac{\lambda_3 P_{Succ,3}}{C\mu_{succ}} + \frac{\lambda_3 P_{Fail,3}}{C\mu_{fail}} \quad (2) \end{aligned}$$

Where $P_{Fail,2}$, $P_{Fail,3}$ is the portion of calls admitted but lost due to waiting too long or the queue is full after being preempted for class 2 and class 3 individually; $1/\mu_{succ}$ is the average service time for the successful public calls, and $1/\mu_{fail}$ is the average service time for those that fail.

Since the failed calls will cause dissatisfaction, it is reasonable to say that only those resources used by the successful calls are effective. We just consider the portion of calls that are successful. Then we assume that there can be a system with the same number of channels that can guarantee all those calls are admitted and then finish successfully with the average call duration of $1/\mu$. The system utilization for this case (we call it *effective system utilization*) is easily computed as:

$$EffSysUtil = \frac{\lambda_1 P_{Succ,1}}{C\mu} + \frac{\lambda_2 P_{Succ,2}}{C\mu} + \frac{\lambda_3 P_{Succ,3}}{C\mu} \quad (3)$$

In the later sections we will study the system behavior by comparing the system utilization and the effective system utilization we defined above.

(2) Channel Occupancy

Channel occupancy is an important metric to measure if public traffic is protected well enough when emergency traffic is heavy. The computation is pretty straightforward. For the emergency traffic:

$$\begin{aligned} Occp_{em} &= \sum_{i=0}^{C} \sum_{m=0}^{L_1} \sum_{n=0}^{L_2} iP(i,C-i,m,n) \\ &+ \sum_{i=0}^{C-1} \sum_{j=0}^{C-1-i} iP(i,j,0,0) \quad (4) \end{aligned}$$

For the public traffic:

$$\begin{aligned} Occp_{pub} &= \sum_{i=0}^{C} \sum_{m=0}^{L_1} \sum_{n=0}^{L_2} (C-i)P(i,C-i,m,n) \\ &+ \sum_{i=0}^{C-1} \sum_{j=0}^{C-1-i} jP(i,j,0,0) \quad (5) \end{aligned}$$

III. Harshness Tuning

As described in the introduction, the pure preemption policy is harsh to public traffic. One important metric to measure harshness is the channel occupancy for public traffic. By tuning the preemption threshold, we hope to achieve the suitable channel occupancy for public traffic. In this section we are to show the algorithm to tune the preemption threshold as we desire.

When our main focus is on the channel occupancy of emergency traffic and public traffic and not the performance about handoff or originating traffic in detail, with the expiration times of both queues are identically distributed, we can use the two dimensional diagram in Fig. 3 instead of the three dimensionaln diagram shown in Fig. 1. Here we combine the two queues together into one queue, so the queue length is $L = L_1 + L_2$.

Using this two dimensional model, we can compute the system utilization and channel occupancy with much lower complexity.

A. The bound of channel occupancy

The threshold of preemption can be from 0 to C. If the threshold is 0, no preemption will be allowed and it becomes a complete sharing (CS) policy; if the threshold is C, preemption is allowed until no ongoing lower priority calls exist (we call this *full preemption*). Obviously, the larger the preemption threshold, the higher the channel occupancy for emergency traffic will be.

To compute the channel occupancy for each class, a direct way is solving the two dimensional Markov chain to get the state probabilities first, and then use the following formulas:

$$Occp_{em} = \sum_{i=0}^{C}\sum_{k=0}^{L} iP(i, C-i, k) + \sum_{i=0}^{C-1}\sum_{j=0}^{C-1-i} iP(i, j, 0) \tag{6}$$

$$Occp_{pub} = \sum_{j=0}^{C}\sum_{k=0}^{L} jP(C-j, j, k) + \sum_{i=0}^{C-1}\sum_{j=0}^{C-1-i} jP(i, j, 0) \tag{7}$$

For the full preemption case, if we treat each column as a single state, an M/M/C/C model can be formed. Note $\pi[i]$ as the steady state for i channels taken by emergency traffic, as $\pi[i] = \pi[i-1]\frac{\lambda_1}{i\mu}$, denote $\rho = \lambda_1/\mu$, the upper bound of channel occupancy for emergency traffic can be calculated directly:

$$\begin{aligned} Occp_{em} &= \sum_{i=1}^{C} \pi[i]i/C = \sum_{i=1}^{C} \pi[i-1]\lambda_1/(C\mu) \\ &= \lambda_1/(C\mu)(1-\pi[C]) = \frac{\rho\sum_{i=0}^{C-1} i!/\rho^i}{C\sum_{i=0}^{C} i!/\rho^i} \end{aligned} \tag{8}$$

When the system is overloaded, the system utilization is close to 1, so the lower bound of channel occupancy for public traffic can be estimated as:

$$Occp_{pub} = 1 - Occp_{em} = 1 - \frac{\rho\sum_{i=0}^{C-1} i!/\rho^i}{C\sum_{i=0}^{C} i!/\rho^i} \tag{9}$$

B. Algorithm

The algorithm to find the best preemption threshold is:

Step 1: Estimate the lower bound for public calls using equation (9), and compare it with the required value. If the lower bound is larger than the required value, the requirement for public calls is already satisfied and full preemption should be used, stop here. Else, go to step 2.

Step 2: Use a binary search method to search for the best threshold value: Let the threshold value be C/2, solve the two dimensional Markov chain and then use equation (7) to decide the channel occupancy of public calls. If it is larger than the required value, search the right half space [C/2, C]; otherwise search the left half space [0, C/2]. Repeat this step until the largest threshold that meets the requirement of public traffic is found.

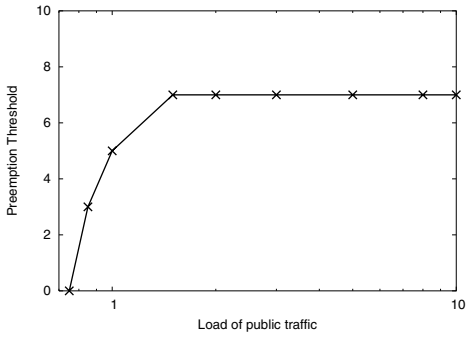

Fig. 4. The threshold for different public traffic loads

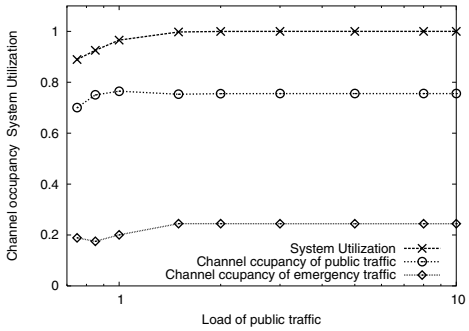

Fig. 5. The system utilization and channel occupancy with threshold applied

IV. Numerical results

A. Preemption thresholds according to different loads

As suggested in [1], the public traffic should be guaranteed 75% of the system resources. When the emergency traffic is low compared with the system's engineered load(e.g. 5%), this goal can be achieved even if full preemption is used. So no tuning of the preemption threshold is needed for this case.

We consider the case when emergency traffic is higher, say 30% of the system's capacity, and seek to keep channel occupancy of public traffic at 75% or above. Suitable thresholds will be searched according to different loads of public traffic. The parameters we use are: C = 20, service time = 100 seconds, expiration time = 50 seconds and the queue length = 10. The results are shown in Fig. 4, 5. We can see that:

(a) When the public traffic is not so high (even if it's a little higher than 75% of the engineered system load), the public traffic does not use 75%, due to expiration of calls waiting in the queue. Preemption is not allowed in this case to prevent the channel occupancy of public traffic from being even lower.

(b) When the public traffic increases, the preemption threshold needs to be increased to give emergency traffic more ability to overcome the higher volume of public traffic.

(c) Once the public traffic is beyond a certain point, with the same preemption threshold, the channel occupancy of each class almost does not change. This is because the system utilization is close to 100% and the channel occupancy of emergency traffic will not be affected by the increase of public traffic.

In Figs. 6 and 7 , we show the case when emergency traffic is very high (160% of the system's capacity). We can see that

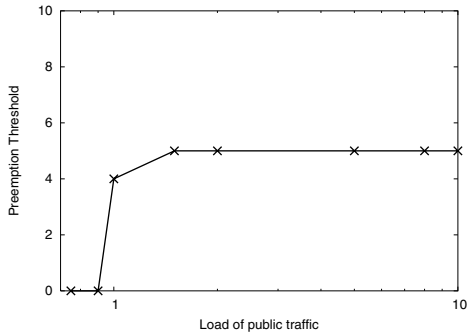

Fig. 6. The threshold for different public traffic loads

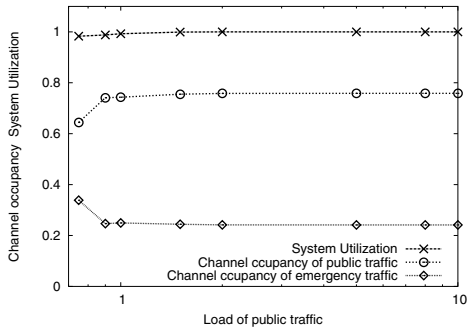

Fig. 7. The system utilization and channel occupancy with threshold applied

when the public traffic is not high enough, 75% of the capacity is not used even if we don't allow preemption. Emergency traffic can take advantage of this and its channel occupancy can be higher than 25%.

B. Comparison of achievable channel occupancy

As we stated before, preemption based strategies have the advantage of immediate access over queueing and scheduling based strategies. But we also know that the pure preemption policy can not be used to adjust the channel occupancy for each class according to different traffic demands. In Figs. 8 and 9, we compare the channel occupancy that can be adjusted for the preemption threshold strategy, the combined strategy in [3], and the pure preemption strategy at different load case.

From the graph we can see that, when the emergency traffic is about 30% of the system load, for the pure preemption

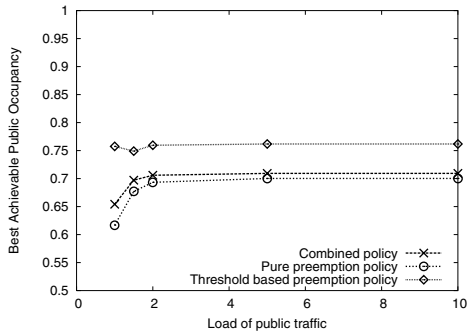

Fig. 8. Comparison of best achievable channel occupancy for public traffic - Emergency traffic = 30% system load

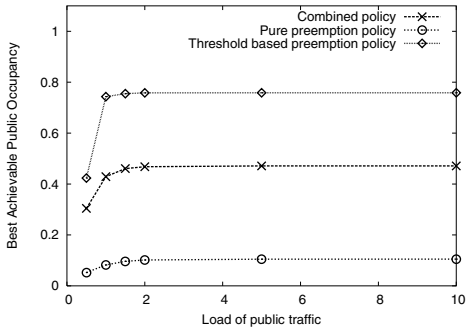

Fig. 9. Comparison of best achievable channel occupancy for public traffic - Emergency traffic = 160% system load

policy and the combined (preemption and queueing) policy, the channel occupancy of public traffic can achieve about 70%. But when emergency traffic is as high as 160% of the system load, for the pure preemption policy, the best achievable channel occupancy for public traffic is only about 10%; for the combined policy, it's much better, but still less than 50%. While for combined policy with preemption threshold, 75% can be guaranteed as desired for both load cases.

Obviously, compared with the pure preemption policy and the basic combined preemption and queueing policy, the combined policy with preemption threshold is much better in terms of guaranteeing the occupancy and thus the admission of public traffic. In fact, it can be as good as the queueing and scheduling strategies shown in [1]. The tradeoff is that when the emergency traffic is very heavy, the admission rate of emergency traffic will decrease compared to other policies, but must occur so as to protect admission of public calls. It is important to *both* protect admission of public calls and provide immediate access to emergency calls. The work in [1] meets the first objective, but not the second.

C. Relationships between system utilization and call duration

Since the public calls can be preempted and restarted, and can even be terminated, a natural question is: will the gross call duration for public calls be different than the no preemption case? In this subsection we will show the study of call durations.

When there is no preemption, each admitted call will finish by itself. So the average call duration will be $1/\mu$. When there is preemption but no loss due to preemption (which means that each preempted call will resume successfully), recall that we assume that each call will restart based on the repeat with re-sampling rule. According to Conway ([8], P.177), the gross service time of preempted calls, therefore, will still be exponential with mean $1/\mu$.

For the model we considered, loss due to preemption is unavoidable either because the queue can be full or preempted users can be impatient while waiting in the queue. To study the call durations in this case, we compare the system utilization and the effective system utilization (See equation (1), (3)) for different parameters (load, expiration time, queue length etc.); the results are shown in Table I. *Interestingly, the system*

Thr	ExpTime(Sec)	QLen	SysUtil	EffSysUtil
10	50	5	.940743	.940743
10	100	5	.957922	.957922
10	200	5	.968608	.968608
5	50	5	.937904	.937904
10	50	10	.942123	.942123

TABLE I

SYSTEM UTILIZATION AND EFFECTIVE SYSTEM UTILIZATION

PubTrffLoad	SuccTime	FailTime
1	98.306283	96.05872
1.5	95.8626359	94.021153
2	95.483025	88.515796
5	93.774451	86.078719
10	91.329356	83.40661

TABLE II

SIMULATION RESULTS FOR AVERAGE DURATION OF SUCCESSFUL AND FAILED CALLS

utilization and the effective system utilization are exactly the same.

If we compare equations (2) and (3), we can remove the part about class 1 and thus get a conservation law equation about the gross service time:

$$\frac{\lambda_2 P_{Succ,2}}{C\mu} + \frac{\lambda_3 P_{Succ,3}}{C\mu} = \frac{\lambda_2 P_{Succ,2}}{C\mu_{succ}} + \frac{\lambda_2 P_{Fail,2}}{C\mu_{fail}} + \frac{\lambda_3 P_{Succ,3}}{C\mu_{succ}} + \frac{\lambda_3 P_{Fail,3}}{C\mu_{fail}} \quad (10)$$

From the above equation, it can be shown that the service time for those successful calls must be less than $1/\mu$. This means that if a good admission strategy was taken such that all calls admitted will be guaranteed success (e.g. blocking some calls while there are still free channels or the queues are still not full, like pre-dropping packets in data network [2]), each call's average duration would be $1/\mu$. If no pre-blocking strategy is taken, the same amount of calls will eventually succeed, but their average call duration will be shorter than $1/\mu$ since some of the resources will be taken by those calls that eventually fail. This also shows that calls with shorter duration have a greater chance to succeed.

In Table II, the simulation results about average duration for successful and failed calls (those calls that have been admitted but lost after being preempted) are shown for $1/\mu = 100$ seconds. SuccTime and FailTime represent the average call duration for those successful and failed calls, and PubTrffLoad means the load of public traffic relative to capacity. We can see that as the system load increases, not only does the blocking rate increase, the call durations of both calls that succeed and fail also become shorter.

V. CONCLUSIONS

In this paper we introduce the preemption threshold based strategy for the admission control of wireless cellular network with emergency traffic supported. While the advantage of the combined preemption and queuing strategy is inherited from the work of [3], the channel occupancy for public traffic can now also be controlled. Under the preemptive repeat with re-sampling assumption, we find that the same amount of traffic will be guaranteed success compared with the pre-blocking strategy, while the average service time for those successful calls will be shorter. This work can be readily applied to the connection reservation mechanisms of 3G/4G wireless networks(e.g. EV-DO Rev.A, 802.16).

Possible future work includes the proof of the conservation law of service time, the calculation of the service time achieved by successful calls and failed calls, and possible use of different assumptions about impatience times of customers.

REFERENCES

[1] Nyquetek Inc., "Wireless Priority Service for National Security / Emergency Preparedness: Algorithms for Public Use Reservation and Network Performance," August 30, 2002. Avaiable at http://wireless.fcc.gov/releases/da051650PublicUse.pdf.

[2] M. Joshi, A. Mansata, S. Talauliker, and C. Beard, "Design and Analysis of Multi-Level Active Queue Management Mechanisms for Emergency Traffic," *Computer Communications Journal,* Volume 28, Issue 2, February 2005, pp. 162-173.

[3] J. Zhou and C. Beard, "Comparison of Combined Preemption and Queuing Schemes for Admission Control in a Cellular Emergency Network," *IEEE Wireless Communications and Networking Conference (WCNC) 2006*, Las Vegas, NV, April 3-5, 2006.

[4] C. Beard, "Preemptive and Delay-Based Mechanisms to Provide Preference to Emergency Traffic." *Computer Networks Journal*, Vol. 47:6, pp. 801-824, 22 April 2005.

[5] C. Beard and V. Frost, "Prioritized Resource Allocation for Stressed Networks," *IEEE/ACM Transactions on Networking*, Vol. 6, no. 5,October 2001, pp. 618-633.

[6] D. Hong and S. S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Tele- phone Systems with Prioritized and Non-prioritized Handoff Procediires," *IEEE Trans. on Vehicular Technology*, vol. VT-35, no. 3, pp. 77-92, Aug. 1986.

[7] J. Wang, Q.-A. Zeng, and D. P. Agrawal "Performance Analysis of Preemptive Handoff Scheme for Integrated Wireless Mobile Networks," *Proceedings of IEEE GLOBECOM 2001*, pp. 3277 - 3281.

[8] R. W. Conway, W. L. Maxwell, and L. W. Miller. "Theory of scheduling," Addison Wesley, 1967

[9] Y.Z.Cho, C.K.Un, "Analysis of the M/G/1 queue under a combined preemptive/nonpreemptive priorty discipline, " *IEEE Transaction on Communications,* V.41(1), 1993 pp.132-141.

[10] S. Drekic and D.A. Stanford, "Reducing delay in preemptive repeat priority queues," *Operations Research,* 49(2001) pp.145-156

[11] S. Drekic and D.A. Stanford, "Threshold-based interventions to optimize performance in preemptive priority queues," *Queueing Systems,* 35(2000) pp.289-315

[12] Y.H. Kim and C.K. Un "Bandwidth allocation strategy with access restriction and preemptive priority, " *Electronics Letters*, 1989, **25**(10), pp.655-656

[13] S. Tang and W. Li "An adaptive bandwidth allocation scheme with preemptive priority for integrated voice/data mobile networks, " *IEEE Transactions on Wireless Communications,* Vol.5, No.9, September 2006,