A Controlled Preemption Scheme for Emergency Applications in Cellular Networks

Jiazhen Zhou and Cory Beard, *Senior Member, IEEE*

***Abstract*—In this paper we introduce a threshold-based preemption strategy for supporting emergency traffic in cellular networks. Compared with the current commercially used policies, our scheme not only can guarantee a certain amount of resources to public customers, but also can provide immediate access for emergency users and flexibility for providers to adapt to different requirements and operating scenarios. In addition, under the combined preemption and queueing framework, interesting analytical relationships among channel occupancy, gross service time and success probability for public traffic are revealed. Based on this, guidelines for further improving satisfaction of public customers can be provided.**

Keywords: Admission control; emergency traffic; preemption; priority queue; gross service time; channel occupancy; success probability; handoff; expiration; scheduling

I. INTRODUCTION

In a society where wireless communication capabilities are pervasive, emergency personnel should be able to use both government and commercially available systems to respond to natural and man-made disasters [1]. To provide such services, a general sense of priority should be attached to emergency sessions and to how they are given access to wireless resources. Such priority, however, cannot be absolute, since the needs of the general public are also very important to address. As a result, some control policy must be applied to prevent the extreme use of resources by emergency traffic.

For several years, but especially in response to the events of September 11, 2001, the U.S. government and the wireless telecommunications industry have worked together to specify a technically and politically feasible solution to the needs of homeland security for priority access and enhanced session completion. This has resulted in definition of requirements for an end-to-end solution for national security and emergency preparedness (NS/EP) sessions called the wireless priority service full operating capability (WPS FOC) [2], [3], [4]. First-responders, NS/EP leadership, and key staff are able to use this capability by using allocated access codes.

Nyquetek Inc. has prepared an evaluation report [5] of algorithms for the Wireless Priority Service (WPS) that are currently used by some of the main cellular network operators. A series of queueing and scheduling based policies are introduced and compared for supporting emergency traffic. They emphasized that the priority of emergency traffic can be guaranteed by just queueing emergency sessions when no channels are immediately available. When the emergency traffic is less than 10% of the normal engineered load of a cell, the admission of emergency traffic can be virtually guaranteed and the admission of public traffic will not be affected much. However, if the emergency demand is high enough that it can take most of the available resources in a cell, a policy called Public Use Reservation with Queuing All Calls (*PURQ-AC*) was proposed. In PURQ-AC, two buffers are provided for emergency and public originating calls, and guard channels are reserved for public handoff traffic. When there are channels released, sessions in the two queues will be scheduled in a round robin style: the NS/EP queue is served once every 4 times a channel becomes available (giving a 1/4 allocation to the NS/EP queue). Through this scheduling policy, they hope to achieve a maximum allocation of 25% of resources for emergency use and at least 75% for public use.

A big problem with the above schemes is that emergency traffic must wait significant time before being admitted, which is unreasonable when there are urgent needs to save life or property. Another side effect of waiting in a buffer is that some emergency users may be forced to give up because they cannot wait too long. Furthermore, as shown in [5], desired resource allocation for public use cannot be guaranteed when public traffic is not extremely high.

In this paper, we propose a scheme that can address the above problems by providing faster access and better assurance of admission of emergency traffic, and guaranteed resource allocation for public use that is much less sensitive to the load of public traffic.

The best strategy to guarantee immediate access and assure the admission of emergency traffic is preemption, which means emergency sessions can break ongoing public sessions and take the resources for emergency use. However, an uncontrolled preemption strategy tends to use up all of the channel resources and will be against our goal to protect public traffic. As an effective control method to reduce the effect on public users, a threshold based controlled preemption method is introduced in this paper. The idea is that when the resources occupied by the emergency sessions surpass the threshold, preemption will be prohibited. By tuning the preemption threshold, the channel occupancy for each class can be adjusted as we like.

To support the use of preemptive controls, we have devel-

Jiazhen Zhou is with Department of Systems and Computer Science, Howard University, Washington D.C., 20059, United States (e-mail:zhoujiazhen@gmail.com)

Cory Beard is with Department of Computer Science Electrical Engineering, University of Missouri - Kansas City, Kansas City, MO 64110, United States (e-mail: BeardC@umkc.edu)

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

oped a series of theoretical models and performance metrics to measure the performance that is provided to both emergency and public users. These results are also used to tune preemption thresholds. One interesting result that was found shows that the channel occupancy of public traffic is proportional to its success probability. This means that either quantity can be measured when the cellular network is in operation, and the system control parameters can be tuned accordingly.

Related work on preemption based research, both on applications and theory, is discussed in Section II. In Section III, we introduce the main idea of the preemption threshold based strategy and derive important performance metrics. In Section IV we present the analysis about the effects of preemption on public traffic. The average gross service time, and the relationships among channel occupancy, gross service time and success probability are studied in detail. In section V, an algorithm for tuning the preemption threshold is introduced. In addition, we show numerical results in section VI, and conclude this paper in Section VII.

II. Related work in preemption based strategies

A. Control of preemption

As mentioned in Section I, preemption must be controlled to avoid the starvation of public traffic. In fact, there have been numerous methods on lessening the impact of preemption on low priority tasks. One main class of ideas is to block further preemptions based on service effort rendered or to consider alternatives to limit the frequency within which preemption occurs [7].

In Cho and Un [8], a combined preemptive/nonpreemptive priority discipline was proposed. When a discretion rule is satisfied, preemptive priority will be applied. Otherwise, nonpreemptive priority will be used. The discretion rule is based on the parameter of low priority traffic like the elapsed service time, the remaining service time, or the ratio of elapsed to total service time. In Drekic and Stanford [9], the discretion rule is based on the number of preemptions experienced by the low priority tasks. After being preempted for a certain number of times, the task can be promoted to a higher priority class, or forbidden to be preempted.

Both works mentioned above set the discretion rule based on the behavior of each session or task. Another possible approach is based on the behavior of the whole class. In Kim and Un [10], the resource utilization taken by the high priority class was considered for deciding preemption to be allowed or not.

In our work, we take the strategy similar to [10] by setting the thresholds on the number of channels occupied by emergency traffic. The novelty of our work is: (1) We consider the practical issue that preempted sessions can renege due to waiting too long in the waiting queue; (2) Compared with the performance analysis limited only to throughput and blocking in [10], our analysis provides more insights about the effects of preemption on low priority traffic, especially on the gross service time and average channel occupancy.

B. Applications of preemption based schemes

In Beard [11], different preemptive schemes for supporting emergency traffic were studied. The effects on low priority traffic and the whole system utilization were shown. But it was assumed that all preempted sessions will be dropped, which can cause high termination probability for low priority traffic and thus high dissatisfaction from the customers.

The applications of combined preemption and queueing schemes in wireless networks can be seen in Wang, Zeng and Agrawal [12], and Tang and Li [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 for expiration (reneging due to impatience, or thrown away by the system after a certain time [15], [16], [17]) of preempted sessions in the queues was not studied, and to ignore such behavior is unrealistic even for non-real-time data sessions.

In Zhou and Beard [6], a "single preemption" policy similar to [9] was introduced on the basis of the combined preemption and queueing scheme. This strategy helps protect public traffic from preemptions and is a big improvement over the "pure preemption" policy that employ no queues, but the protection is not strong enough when emergency traffic is unexpectedly high.

A preliminary version of this paper appears in [18]. The basic preemption threshold based control scheme was shown there. However, this paper shows much more detailed benefits and provides analysis on the cause of these benefits.

C. Theoretical research on multiple server based preemption strategies

The performance metrics we especially care about in this paper are channel occupancy for each class and the average gross service time of low priority classes. The channel occupancy characterizes how well the public traffic is protected, and the average gross service time shows how the communication of public users is affected. Theoretical solutions that can be readily applied in this paper have not been seen. In fact, most past theoretic studies for preemptive priority systems, like Graver [19], Takacs [20], Welch [21], Conway, Maxwell, and Miller [22], Cho and Un [8], and Drekic and Stanford [9], were for the single server case, and generally with the assumption that a preempted job will not be lost and will surely come back to finish.

Studies for the multiple server case can be seen in Segal [23], Mitrani and King [24], Buzhen and Bondi [25], and Gail, Hantler and Taylor [26]. All these papers assume no loss or expiration for the preempted tasks. With the preemptive resume or preemptive repeat assumptions they make, the gross service time is the same as with the no preemption case. So there are no studies about the gross service time of low priority tasks seen in these former works. Instead, they just study the waiting time. In the practical model we are to study in this paper, however, the loss of preempted tasks is unavoidable due to the limit on the buffer size and the impatience of preempted customers. The corresponding gross service time, and its effect on the system utilization of low priority tasks is presented. Due

to the generality of the assumptions we make, this part of work could be applied to scenarios far beyond the emergency traffic problem addressed in this paper.

III. Threshold-Based Preemption Control

A. Basic scheme and assumptions

The main types of sessions we deal with are emergency sessions, public handoff sessions and public originating sessions. We assume each session uses the same amount of wireless resources. As shown in [5], since the current WPS is provided only for leadership and key staff, it is reasonable to assume that most emergency users are stationary within a disaster area. So handoff for emergency sessions is not considered here, but our current work can be readily extended to deal with emergency handoff traffic when necessary.

The basic scheme used is illustrated in Fig. 1. When an incoming emergency session fails to find free capacity, and if the number of active emergency sessions is less than the preemption threshold, it will preempt resources from a randomly picked ongoing public session. The preempted session will be put into the handoff/preempted session queue. For an arriving public handoff session, it will also be buffered in the handoff/preempted session queue when no capacity is immediately available. Correspondingly, there is also an originating session queue, which is further helpful for preventing starvation of public traffic. If an incoming emergency session fails to find free resources to preempt, it will be simply dropped.

We suggest not to have a buffer for emergency users for two reasons: (1) Make sure there is no access delay for emergency sessions; (2) Guarantee the public traffic has enough system resources when emergency traffic is very heavy. If emergency traffic is queued in this case, public traffic could not be well protected as the FCC requires even if preemption is not allowed. The reason that we use the same buffer for handoff and preempted sessions is that both of these two types of sessions are broken sessions, so they have the same urgency to be resumed. More precise configuration like using two different buffers is possible, but will not be obviously beneficial. In fact, it will make the implementation and analysis more time consuming, because it will have a much larger Markov chain state space.

When capacity becomes available later, one session from the queues is served. A priority queue based scheduling policy will be used, and it is reasonable to assume that handoff/preempted sessions have higher priority over the originating sessions. The queues are finite and customers can be impatient when waiting in the queue, so blocking and expiration are possible.

Since customers have different patience, it is reasonable to assume their impatience behavior to be random rather than deterministic like assumed in Nyquetek's study. We assume that the expiration times of traffic in the same queue are exponentially and identically distributed, and the patience of a customer is the same after each preemption. Strictly speaking, the session duration is probably not exponentially distributed. As shown by Jedrzycki and Leung [27], the channel holding times in cellular networks can be modeled much more accurately using the lognormal distribution. However, in reality

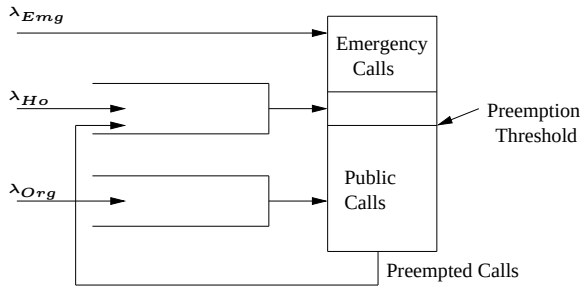

Fig. 1. Combined preemption and queueing scheme

the exponential distribution assumption for sessions is still mostly used, both in analysis-based study like Tang and Li [13], and simulation-based study like Nyquetek's report [5]. In this paper, we also assume that all session durations and inter-arrival times are independently, identically, and exponentially distributed.

If session durations are memoryless (i.e., exponentially distributed), this means that if at any point a session is interrupted, the remaining service time is still exponential with the same average service time as when it began. It is, therefore, reasonable to model a restarted session as a renewal process. In other words, the preempted session will be restarted with resampling of the exponential random variable [22], also called a repeat-different approach [9].

B. Modeling and computing complexity

Let us denote the total number of channels as C, the length of handoff/preempted queue is L_1 and the length of originating queue is L_2, and the *preemption threshold* is R. Each state is identified as (i, j, m, n), where i, j is the number of channels occupied by emergency and public sessions respectively , m, n represents the number of sessions in the handoff/preempted session queue and the public originating session queue individually. The arrival rates for emergency, handoff, and originating sessions are λ_{Emg}, λ_{Ho}, λ_{Org} respectively. The mean expiration rates for sessions waiting in the handoff/preempted queue and originating queue are denoted as $\mu_{exp}^{ho/prm}$ and μ_{exp}^{org}. To facilitate analysis, the average service rate for each class is assumed to be the same and denoted as μ. This also means that the session duration in a single cell is exponentially distributed with mean $1/\mu$, whether the session ends in this cell or is handed off to another cell. A Markov chain can be formed, and the state probabilities can be obtained by solving the following balance equations:

(1) When the channels are not full, the typical state transition is shown in Fig. 2. Since the queues are empty in this case, in the notation we replace $P(i, j, 0, 0)$ with $P(i, j)$ for simplicity. The corresponding balance equation is:

$$\begin{aligned} &P(i,j)(\lambda_{Emg}+\lambda_{Ho}+\lambda_{Org}+(i+j)\mu) \\ &= P(i-1,j)\lambda_{Emg}+P(i,j-1)(\lambda_{Ho}+\lambda_{Org}) \\ &+P(i,j+1)(j+1)\mu+P(i+1,j)(i+1)\mu. \end{aligned} \quad (1)$$

For the states on the edge, some terms of this equation will disappear.

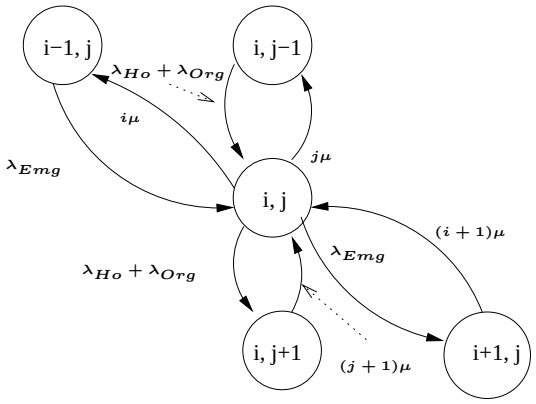

Fig. 2. The typical state change when channels are non-full

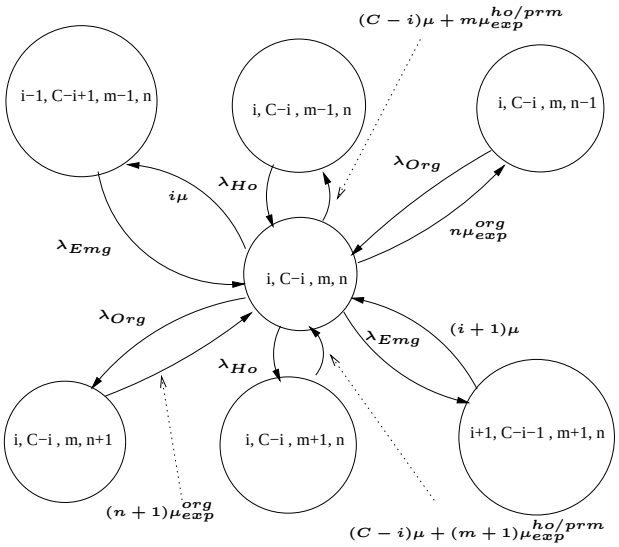

Fig. 3. The typical state change when channels are full and $i < R$

(2) When the channels are full, queueing is involved, the typical state transition is shown in Fig. 3. The corresponding balance equation is:

$$\begin{aligned}
&P(i,C-i,m,n)(\lambda_{Emg}+\lambda_{Ho}+\lambda_{Org}+C\mu\\
&+m\mu_{exp}^{ho/prm}+n\mu_{exp}^{org})\\
&=P(i-1,C-i+1,m-1,n)\lambda_{Emg}\\
&+P(i,C-i,m-1,n)\lambda_{Ho}\\
&+P(i,C-i,m,n-1)\lambda_{Org}\\
&+P(i+1,C-i-1,m+1,n)(i+1)\mu\\
&+P(i,C-i,m+1,n)((C-i)\mu+(m+1)\mu_{exp}^{ho/prm})\\
&+P(i,C-i,m,n+1)(n+1)\mu_{exp}^{org} \qquad (2)
\end{aligned}$$

Note that when $i \geq R$, no preemption will be allowed, which will make the elements involving λ_{Emg} disappear.

With the practical consideration of expiration and preemption threshold, a product form solution for the equilibrium equations has not been found. Since we have limited the system to one buffer for handoff and preempted sessions, the computation requires operations on a matrix with size CL_1L_2, which means it depends on the number of channels and the size of the two buffers. Note that as pointed out in [5], the buffer size need not to be long (=5) because the effect will not be obvious after a certain point. Due to this fact, the computation is feasible.

C. Extension of our model to 3G/4G systems

The admission control policies discussed in this paper are assumed to be load based. This means that admission is based on whether the new session will make the load surpass the capacity of the system. The load is usually measured by the number of users in a 2G system. With multiple access schemes like CDMA, WCDMA, OFDMA applied, one main difference is that interference rather than the number of users is the main factor to be considered for the admission control problem in a 3G/4G system [28].

With a CDMA based access scheme, admission can be done indirectly by setting an interference-based criteria, for example a limit on CDMA Rise over Thermal (RoT), then determining ahead of time the load where a new session would cause the system to exceed the interference limit. In fact, as pointed out in [29], load based admission control is still suitable. In their analysis for number-based CAC, the interference threshold is transferred into the maximum acceptable number of users. Then the blocking rate (measured grade of service) and the outage probability of communication quality (measured quality of service) are evaluated. The numerical results show that the number-based CAC and the interference-based CAC agree well with each other. They concluded that load-based admission is preferred because of its simplicity and the ease of implementation, although interference based admission has the advantage that the threshold value has less sensitivity on other system parameters like the propagation model, traffic distribution, or the transmission rate.

As opposed to balancing blocking rate and outage probability of communication quality like in [29], we are mainly considering the fairness in resource use between emergency users and public users. When an emergency happens, there is much more demand than the system can handle. No matter how we try to balance capacity and quality of service, there is still blocking. So the capacity of the system, in terms of the maximum number of admitted users, can be determined according to the requirements on quality of service (QoS) only. With the capacity of the system known, the preemption threshold can be tuned to achieve ideal channel occupancies for both emergency and public traffic. Note here that we assume the capacity is static for a period of time, but it can be recomputed if the SIR threshold needs to be changed, for instance, due to increased interference from neighboring cells or due to cell breathing to shift users to neighboring cells.

Another important difference is that data applications are much more common in a 3G/4G network. How would load based admission control be accomplished with both voice sessions and data sessions in the same cell? Admission of voice sessions can easily be controlled based on whether a new session would go beyond the voice loading limit. Data sessions, however, can be handled in two distinctly different ways. On the one hand, if data sessions need some level of guaranteed QoS, they can be admitted similarly to voice

sessions, by equating a data session to a certain number of voice sessions or a certain amount of needed bandwidth. On the other hand, service providers may treat data sessions differently by expecting them to use whatever is left over after the voice sessions are satisfied. For example, in the 3G EV-DO, Rev. A standard, "HiCap" data sessions are given different power levels and Hybrid ARQ termination targets as compared to "LoLat" voice traffic. HiCap data traffic is expected to be able to tolerate longer packet delays and to probably use TCP to adapt to the network congestion.

If a data session is equated to a certain number (say K) voice sessions, minor modifications to the state transition are needed for our Markov model. For example, compared with the diagram in Fig. 2, the state (i, j) will transfer to state $(i, j+K)$ rather than $(i, j+1)$ when a new public data session arrives. The Markov chain can still be solved in a similar way, and the tuning of preemption threshold can thus be performed.

If the data traffic is treated as elastic, it can use all of the bandwidth if no other sessions are present. When new voice sessions arrive, some of the bandwidth is taken and the elastic sources adjust. This can be viewed as a form of soft preemption [11] of the data sources, where voice sessions preempt some, but not all, of the bandwidth of the data sessions. A 3G/4G system with data applications supported can, therefore, also be described using diagrams shown in Fig. 2 as total admitted users increase. If emergency sessions still want to interrupt the last portion of the guaranteed bandwidth needed by data sessions, say equal to a voice session, then it means hard preemption [11] happens and the preempted data session can be queued to resume later. This is same as what has been described in Fig. 3 as total admitted users are fixed and it is the issue of replacing public users with emergency users.

In conclusion, interference-based admission control can be converted into a load based admission control problem. Furthermore, the elastic property of data sessions makes it possible for us to use the same model as that of a 2G scenario. This is why we can conclude that the work in this paper is suitable for all 2G, 3G, and 4G systems.

D. Performance Evaluation

With the state probabilities solved, performance metrics, including average channel occupancy and the success probability, i.e., probability of finishing normally without expiring or dropping for each class can be obtained and will be shown in this subsection. Computation of related parameters, like admission probability, blocking probability of each class, the expiration probability of sessions in each queue, and preemption probability for a low priority session given that it is admitted, has been provided by [6].

(1) System utilization and channel occupancy

The system is not fully used when there are still free channels available. When there are n channels occupied, that means $C - n$ channels are not used, and the total proportion of unused channels is thus $\frac{C-n}{C}$. The system utilization can be computed by considering those portion of unused channels at all possible states:

$$SysUtil = 1 - \sum_{n=1}^{C-1} \sum_{i=0}^{n} \frac{(C-n)P(i, n-i, 0, 0)}{C} \quad (3)$$

"Channel occupancy" is defined as the portion of channels occupied by each class of traffic. It is an important metric to measure whether the public traffic is well protected when emergency traffic is heavy. The channel occupancy for emergency traffic and public traffic can also be computed based on steady states:

$$ChOcp^{Emg} = \sum_{n=1}^{C} \sum_{i=1}^{n} \sum_{k=0}^{L_1} \sum_{l=0}^{L_2} \frac{iP(i,n-i,k,l)}{C} \quad (4)$$

$$ChOcp^{Pub} = \sum_{n=1}^{C} \sum_{j=1}^{n} \sum_{k=0}^{L_1} \sum_{l=0}^{L_2} \frac{jP(n-j,j,k,l)}{C} \quad (5)$$

(2) Probability flow of low priority sessions

In Fig. 4, the probability flow of low priority sessions is shown. In the frame, "F" means failed, "S" means successful.

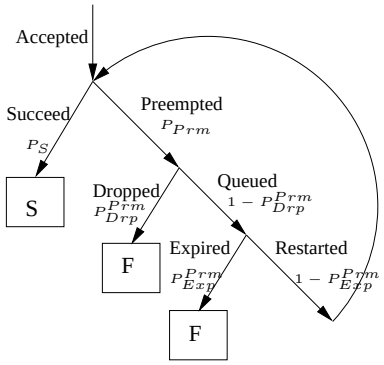

Fig. 4. Probability flow for low priority sessions

A session can be preempted multiple times, and with the renewal process assumption on resumed sessions, the number of preemption times will not affect the preemption probability of a session. Thus the preemption times is geometrically distributed with:

$$Pr(\text{Preempted } n \text{ times}) = P_{Prm}(1-A)A^{n-1}, n = 1, 2, ... \quad (6)$$

Here $A = P_{Prm}(1 - P_{Drp}^{Prm})(1 - P_{Exp}^{Prm})$ is the probability for a session to stay active; $(1 - A)$ is the probability that the session ends (succeeds, expires or be blocked after being preempted). P_{Drp}^{Prm} is the probability for a preempted session to be dropped (due to full queue) after being preempted, and P_{Exp}^{Prm} is the expiration probability for sessions waiting in the preempted session queue.

Thus the expected value of preempted times is $\frac{P_{Prm}}{1-A}$, or expressed in the form of preemption and expiration probability:

$$\overline{PrmTimes} = \frac{P_{Prm}}{1 - P_{Prm}(1 - P_{Exp}^{Prm})P_{Drp}^{Prm}} \quad (7)$$

(3) Success probability

For emergency sessions, all of the admitted sessions will be successfully finished, thus providing high dependability. This kind of dependability can not be assured for low priority sessions.

According to Fig. 4 we can compute the success probability given a session is admitted, which is denoted as P_{SGA}: for an

admitted session, it will succeed only if it does not expire or being blocked after being preempted. Note that $P_S = 1 - P_{Prm}$, we have:

$$P_{SGA} = P_S \sum_{i=0}^{\infty} (P_{Prm}(1 - P_{Drp}^{Prm})(1 - P_{Exp}^{Prm}))^i = \frac{(1 - P_{Prm})}{1 - P_{Prm}(1 - P_{Drp}^{Prm})(1 - P_{Exp}^{Prm})} \quad (8)$$

The successfully finished probabilities are decided by P_{SGA} and corresponding admission probabilities :

$$P_{Succ}^{Ho} = P_{Adm}^{Ho} P_{SGA} \quad (9)$$

$$P_{Succ}^{Org} = P_{Adm}^{Org} P_{SGA} \quad (10)$$

IV. Effects of preemption on low priority traffic

Now we move to showing interesting analytical relationships among channel occupancy, gross service time and success probability for public traffic. To facilitate analysis, first we consider a simple case where there are only two classes of traffic, and in which the higher priority class can preempt resources from low priority ongoing sessions. The generalization to the multi-class case (including the three class scenario in this paper) is shown later in this section.

When there is no preemption, each admitted session will finish by itself directly. So the average session duration is $1/\mu$. If there is preemption but no loss due to preemption (which means that each preempted session will be resumed eventually), recall that we assume that the resumed session is a renewal process, the gross service time of low priority sessions, is still exponential with mean $1/\mu$ according to Conway et al. ([22], P.177). Although the conclusion was made for the single server case, it is also true for the multiple server case because the memoryless property still holds.

For the model we consider in this paper, loss due to preemption is unavoidable either because the queue for the preempted sessions might be full, or because preempted users can become impatient while waiting in the queue. Some important questions then arise. Will the average gross duration for low priority sessions still be equal to $1/\mu$? Is there any difference between successful (finish by itself) and failed sessions? To our knowledge there has been no work dealing with this problem so far. And our research result is shown as follows.

A. Gross service time and channel occupancy

Denote $\overline{T_l} = 1/\mu_l$ as the average gross session duration for the low priority traffic , and P_{SGA}^l as the success probability given that a low priority session is admitted. The arrival rate of admitted low priority traffic is $\lambda_l P_{Adm}^l$, and the channel occupancy of low priority traffic can be computed as the ratio of admitted traffic rate compared with the service rate, which is $\frac{\lambda_l P_{Adm}^l}{C\mu_l}$.

For the computation of average gross service time, we have:

Lemma 1: For a preemptive system, assuming a renewal process for preempted sessions that are resumed, the average gross service time of the low priority traffic can be computed as:

$$\overline{T_l} = P_{SGA}^l \frac{1}{\mu} \quad (11)$$

Proof:

According to Little's law, $\overline{T_l} = \overline{N_l}/(\lambda_l P_{Adm}^l)$. Here $\overline{N_l}$ is the average number of low priority sessions in service, and thus can be expressed in terms of Markov chain steady states as: $\overline{N_l} = \sum_{i=1}^{C} \sum_{X=0}^{C-i} iP(X,i)$, where $P(X,i)$ is the steady state probability that represents X high priority customers and i low priority customers in service. If $X+i = C$, then $P(X,i)$ includes all states where low priority sessions can be buffered. So:

$$\overline{T_l} = \frac{\sum_{i=1}^{C} \sum_{X=0}^{C-i} iP(X,i)}{\lambda_l P_{Adm}^l} = \frac{\sum_{i=1}^{C} \sum_{X=0}^{C-i} i\mu P(X,i)}{\lambda_l P_{Adm}^l \mu} \quad (12)$$

Note that under the renewal process assumption, the service rate of each low priority session will be independently and exponentially distributed with average rate μ, irrespective of having been preempted or not. As $i\mu$ is the rate of low priority sessions to *finish service* given that there are i sessions in service (the state is $P(X,i)$), $\sum_{i=1}^{C} \sum_{X=0}^{C-i} i\mu P(X,i)/\lambda_l$ is the ratio of low priority sessions that finish service (become successful) compared with the arrivals, which is the definition of success probability for low priority sessions (P_{Succ}^l). Thus from equation (12) we get: $\overline{T_l} = \frac{P_{Succ}^l}{P_{Adm}^l \mu} = P_{SGA}^l \frac{1}{\mu}$, so equation (11) is proved.

Remarks 1:

(1) The average gross session duration is directly decided by P_{SGA}^l. When more sessions are lost due to short queues or high expiration (reneging) rates, the success chance is worse and P_{SGA}^l will decrease, so the average session duration will be shorter. Conversely, the average session duration will be longer if the success chance is better.

(2) Since low priority sessions may restart several times, one might think they would last longer than the no preemption case. But in reality the average duration is shorter since $P_{SGA}^l \leq 1$.

(3) As noted in [22], P.177, if no preempted sessions are lost, their gross service time will still be exponential with mean $1/\mu$. This is just a special case of the general formulae we have provided in equation (11) by letting P_{SGA}^l = 1.

(4) In the proof of equation (11) we can see that *the key assumption is that the service time for each low priority session is exponentially distributed with average value* $1/\mu$, regardless of whether it has been preempted or not. The behavior of the queues, including various reneging behaviors and different scheduling policies, will not change equation (11).

With the "preemption and resume" or "preemption and restart same" rule [22], the remaining service time after resuming for preempted sessions will not be independently and exponentially distributed with mean time being $1/\mu$, so equation (11) would not apply .

(5) Since obtaining the number of admitted and dropped

(after being admitted) sessions in a certain period (so that we can estimate P^l_{SGA}) is much easier than keeping record of service time for each session, which can consist of several broken periods, computing average gross service time using equation (11) is a more economical approach.

Theorem 1: In a preemption based system, the success probability of low priority traffic is decided by its channel occupancy. This can be expressed in the following equation:

$$ChOcp^l = \frac{\lambda_l P^l_{Succ}}{C\mu} \tag{13}$$

Proof: As mentioned earlier in this subsection, $ChOcp^l = \frac{\lambda_l P^l_{Adm}}{C\mu_l}$. On the other hand, by multiplying P^l_{Adm} on both sides of equation (11), and recall that $\overline{T_l} = 1/\mu_l$, we have:

$$\frac{P^l_{Succ}}{\mu} = \frac{P^l_{Adm}}{\mu_l} \tag{14}$$

Multiply λ_l/C to both sides of equation (14), then we have: $\frac{\lambda_l P^l_{Adm}}{C\mu_l} = \frac{\lambda_l P^l_{Succ}}{C\mu}$. So, $ChOcp^l = \frac{\lambda_l P^l_{Succ}}{C\mu}$.

Remarks 2:
(1) The channel occupancy of low priority traffic has the same mathematical expression as if only $\lambda_l P^l_{Succ}$ part of traffic were admitted and all succeed to finish by themselves. Of course the reality is that more public users ($\lambda_l P^l_{Adm}$) are admitted, and the average gross session duration is shorter.
(2) If the channel occupancy of low priority traffic with a preemption based strategy can be tuned to be the same as a queuing and scheduling based strategy, the same success probability can be achieved. This means that similar satisfaction with the service can be achieved.
(3) By improving the channel occupancy of low priority traffic, e.g. by lowering the preemption threshold, we can also increase the success probability and thus the satisfaction of low priority users.

Corollary 1: Theorem 1 can be extended to multiple classes by viewing the low priority traffic as different subclasses with a non-preemptive or preemptive scheduling rule employed.
Proof: First consider the *non-preemptive* case. As the low priority sessions are treated the same after being admitted, the distribution and average value of service times and P_{SGA} for different classes are the same. Take the 3 classes case studied in this paper as an example; the public handoff traffic has non-preemptive priority compared with public originating traffic. We have $P^l_{SGA} = P^l_{Succ}/P^l_{Adm} = P^{Ho}_{Succ}/P^{Ho}_{Adm} = P^{Org}_{Succ}/P^{Org}_{Adm}$. Through equation (11) we get $P^l_{SGA} = \mu/\mu_l$, thus $\frac{P^{Ho}_{Succ}}{\mu} = \frac{P^{Ho}_{Adm}}{\mu_l}$ and $\frac{P^{Org}_{Succ}}{\mu} = \frac{P^{Org}_{Adm}}{\mu_l}$. With the definition on channel occupancy we can get $ChOcp^{Ho} = \frac{\lambda_{Ho} P^{Ho}_{Adm}}{C\mu_l} = \frac{\lambda_{Ho} P^{Ho}_{Succ}}{C\mu}$, and $ChOcp^{Org} = \frac{\lambda_{Org} P^{Org}_{Adm}}{C\mu_l} = \frac{\lambda_{Org} P^{Org}_{Succ}}{C\mu}$. Similar derivation can be obtained for more than 3 classes.

For n classes of traffic with a *preemptive priority discipline* applied, the first $n-1$ classes can be viewed as one class, and the lowest priority class is another class, then the problem turns into two classes traffic with preemptive priority which has been studied. So the equation (13) holds for the lowest priority class. Continue this process for the first $n-1$ classes and we can prove that equation (13) holds for each class one by one.

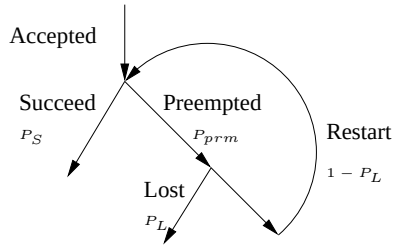

Fig. 5. Simplified Probability flow for low priority sessions

Theorem 2: When the low priority traffic is heavy loaded, P^l_{SGA} can be adjusted without affecting its success probability and channel occupancy.
Proof: In the proof of Lemma 1, we get $P^l_{Succ} = \sum_{i=1}^{C}\sum_{X=0}^{C-i} i\mu P(X,i)/\lambda_l = N_l\mu/\lambda_l$. So P^l_{Succ} is decided by the average number of low priority sessions in service (N_l). For the heavy load case, there are always low priority customers waiting in the queue whenever channels are available. Thus N_l will not be affected by more or less traffic out of the preempted session queue, although that will decide the value of P^l_{SGA}: P^l_{SGA} can be improved by allowing more preempted sessions to recover.
Remarks 3:
(1) Improvement of P^l_{SGA} can be achieved either through taking some methods to keep preempted customers more patient, or by using better scheduling with them.
(2) Since P^l_{Succ} is fixed, P^l_{Adm} will be inversely proportional to P^l_{SGA}. Thus improving P^l_{SGA} will lead less low priority traffic to be admitted.

B. Conditional gross service time

As we have shown, the average gross service time can be computed based on P_{SGA}, and is shorter than the average service time when no preemption happens. One might wonder about the conditional gross service time - the gross service time given it succeeds (denoted as T_{Succ}) or fails (T_{Fail}). Will they be longer or shorter than the normal service time? This will be discussed in this subsection.

To facilitate analysis, the flow graph in Fig. 4 is simplified into the graph shown in Fig. 5. Here we hide the detailed information about expiration or blocking after being preempted, and denote P_L as the total loss probability which includes both expiration and blocking for preempted sessions.

Now we need to know the distribution of the number of preemption times given the session succeeds or fails at last. From Fig. 5 we can see that, for a successful session, the probability mass function (PMF) of the number of preemption

times is:

$$Prob(succeed\ \&Prm\ Times = n) = P_S[(1-P_S)(1-P_L)]^n, \quad n = 0, 1, 2... \quad (15)$$

The total success probability (given that it is admitted) is:

$$Prob(succeed) = \sum_{n=0}^{\infty} P_S[(1-P_S)(1-P_L)]^n = P_S/(P_S + P_L - P_S P_L) \quad (16)$$

Define $p = P_S + P_L - P_S P_L$. The conditional PMF of preemption times will be

$$Prob(Prm\ Times = n|succeed) = \frac{Prob(succeed\&Prm\ Times = n)}{Prob(succeed)} = p(1-p)^n, n = 0, 1, 2... \quad (17)$$

Similarly we can get the conditional PMF for the case that a session fails at last:

$$Prob(Prm\ Times = n|fail) = p(1-p)^{n-1}, n = 1, 2... \quad (18)$$

Note that here $n \geq 1$ since a session must be preempted at least once before it fails. The conditional expectation of preemption times will be $1/p$ if a session fails, and is $1/p-1$ if it succeeds, which means a failed session is preempted once more than a successful session on average. On the other hand, the additional time period before a session is lastly preempted has the same distribution as the time period before a session succeeds since they can both be viewed as the dwell time before leaving the same state.

Based on above facts, it can be concluded that *the distributions of gross service time for successful sessions and failed sessions are the same*, thus they have the same average value which is $P_{SGA}\frac{1}{\mu}$. It is worthy to mention that, a neat form of the distribution of conditional gross service time can be derived for single server case, but it is is extremely difficult for the multiple-server case, especially for the practical model we are studying.

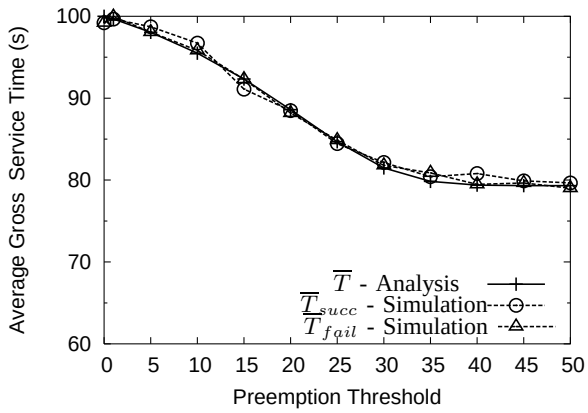

Fig. 6. Comparison of analytical and simulation results for service time

As a verification, in Fig. 6 we compare our analysis with the simulation results (using a CSIM [30] simulation environment) under different preemption thresholds in a 50-channel cell. The average duration for each session is 100 seconds. The average arrival rates for emergency traffic, public handoff traffic and public originating traffic are 15 sessions/minute, 6 sessions/minute, and 60 sessions/minute respectively. The average expiration time for handoff/preempted traffic is set as 10 seconds, and for originating traffic it is 25 seconds. The buffer sizes for both queues are 5. We can see that the simulation results for service time of successful sessions and terminated sessions are both equal to the analytical results for average service time.

V. Tuning of Preemption Threshold

For the emergency scenario we are dealing with in this paper, the preemption threshold is chosen as the control parameter. The discretion rule is: if the number of channels taken by emergency traffic is less than the threshold, the preemption is allowed. Otherwise it will be forbidden. The value of preemption threshold is determined according to the channel occupancy requirements and traffic rates, and the corresponding algorithm is shown in this section.

A. Maximum and minimum average channel occupancy

In a system with C channels, the threshold of preemption can be any value from 0 to C. If the threshold is 0, preemption will never be allowed and it becomes a complete sharing (CS) policy with queueing; if the threshold is C, preemption is allowed until no ongoing lower priority sessions exist (we call this *full preemption*). Obviously, the larger the preemption threshold, the higher the channel occupancy for emergency traffic.

For the full preemption case, if we group the states with the same number of ongoing emergency sessions into a single state, an M/M/C/C model can be formed to find maximum average channel occupancy for emergency traffic since emergency sessions only experience blocking due to other emergency sessions. Denote $P[i]$ as the steady state probability for i channels taken by the emergency traffic, we have $P[i] = P[i-1]\frac{\lambda_{Emg}}{i\mu}$. Let $\rho = \lambda_{Emg}/\mu$, the maximum average channel occupancy for emergency traffic can be calculated directly:

$$ChOcp_{Max}^{Emg} = \sum_{i=1}^{C} P[i]i/C = \sum_{i=1}^{C} P[i-1]\lambda_{Emg}/(C\mu) = \lambda_{Emg}\sum_{i=0}^{C-1} P[i]/(C\mu) = \frac{\rho\sum_{i=0}^{C-1}\rho^i/i!}{C\sum_{i=0}^{C}\rho^i/i!} \quad (19)$$

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

$$ChOcp_{Min}^{Pub} = 1 - ChOcp_{Max}^{Emg} = 1 - \frac{\rho\sum_{i=0}^{C-1}\rho^i/i!}{C\sum_{i=0}^{C}\rho^i/i!} \quad (20)$$

B. Algorithm

With minimum average channel occupancy for public traffic decided, an algorithm can be used to find the best preemption

threshold.

Algorithm 1: Preemption Threshold Tuning

Step 1: Estimate the minimum average channel occupancy for public sessions using equation (20), and compare it with the required value. If the minimum value is larger than the required value, the requirement for public sessions 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 Markov chain and then use equation (5) to decide the channel occupancy of public sessions. 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 preemption threshold that meets the requirement for public traffic is found.

VI. Numerical results

In this section, the preemption threshold based strategy is evaluated. To show the superiority of this strategy, comparison with other possible admission control strategies is provided. The main performance metrics involved are achievable channel occupancy, success probability and waiting time. The performance results are mostly obtained through the analytical approach, but simulation is also shown to verify its accuracy. The simulation tool used is CSIM, the distribution and parameters of session duration, interarrival time, expiration time etc. are all the same (all exponentially distributed) as used in the analytical result.

A. Study of the preemption threshold based strategy

1) Effects of preemption threshold: The preemption threshold affects the amount of resources that can be used by emergency traffic. In Fig 7, the corresponding channel occupancy with the change of preemption threshold is shown. Here the number of channels in a cell is set as 50. The average duration for each session is 100 seconds, so the maximum load that the system can process, called *system capacity* or *engineered system load*, is $C\mu = 0.5$ session/second $= 30$ sessions/minute. The arrival rate for emergency traffic is 15 sessions/minute, accounting for 50% of the engineered system load. The load of public handoff traffic is 6 sessions/minute (20% of engineered system load), and for public originating traffic it is 60 sessions/minute (200% of engineered system load) . In addition, the average impatience time for handoff/preempted traffic is set as 10 seconds, and for originating traffic it is 25 seconds, while the buffer sizes for both queues are 5.

As shown in Fig. 7, the simulation results match perfectly with the analytical results, which verifies the correctness of our analysis.

With the increase of preemption threshold, channel occupancy of emergency traffic increases, but that of public traffic goes down. To obtain 75% channel occupancy for the public traffic, through Fig. 7 it can be seen that the preemption threshold should be set at 13.

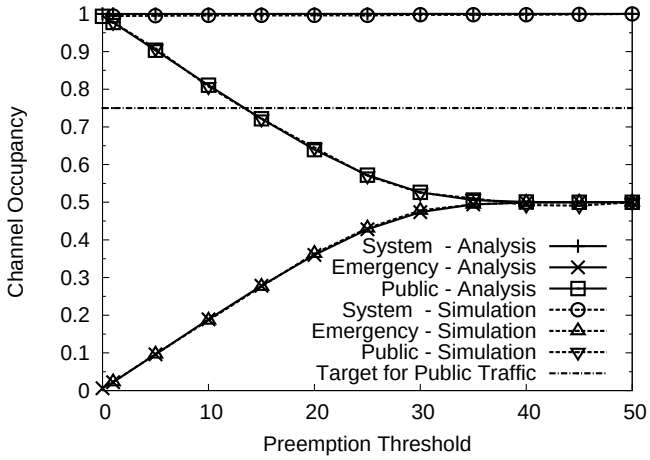

Fig. 7. The change of channel occupancy according to threshold

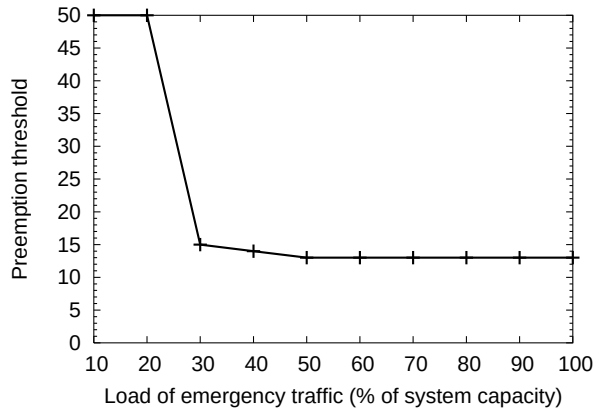

Fig. 8. The threshold for different public traffic loads

2) Preemption thresholds according to different loads: With the required channel occupancy for public traffic given, the preemption threshold can be determined using the algorithm presented in section V. An example of applying this algorithm to find the best preemption threshold according to the change of emergency traffic is shown in Fig.8.

When the emergency traffic is less than 20% of engineered load, there is no difficulty in assuring 75% of channel resources for public traffic. Thus it is not necessary to limit preemptions and the preemption threshold is simply 50. However, as emergency traffic keeps increasing, for example, to 30% of engineered load, the preemption threshold must be set at 15 to ensure enough resources used for public traffic. An interesting phenomenon observed from this figure is that the increase of emergency traffic thereafter does not require much change in preemption threshold (13 will be a value that is suitable for most load cases).

The fact that the preemption threshold is relatively *insensitive* to the change of traffic allows us to conclude that, even though the measurements of traffic might not be that timely so that preemption threshold is not adjusted soon enough, a certain preemption threshold might still work pretty well. As an example, the achieved channel occupancy with a fixed preemption threshold (=13) is shown in Fig. 9. We can see that

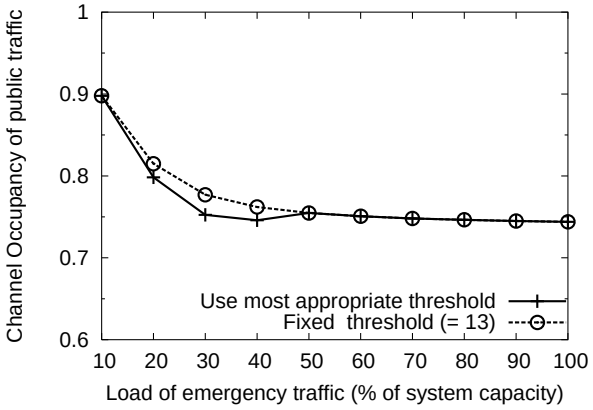

Fig. 9. Channel occupancy vs. adjustable or same threshold

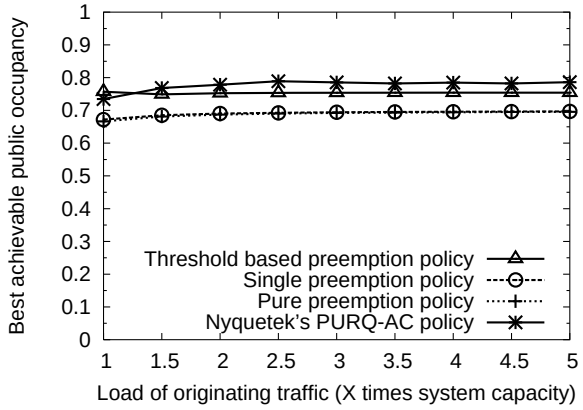

Fig. 10. Comparison of achievable channel occupancy for public traffic - Emergency traffic = 30% system capacity

by applying this single threshold, the system performs well except that public traffic might take a little bit more resources than it should when the emergency traffic load is moderate.

B. Comparison with other candidate strategies

Several other candidate policies could be considered for supporting emergency applications. They include the PURQ-AC policy [5], the "single preemption" policy [6], and the "pure preemption" policy.

In the following we will first evaluate which policies are feasible candidates by showing whether they can protect public traffic as we desire. Then more detailed performance metrics, like success probability and waiting time of each class, will be compared among feasible strategies. In Nyquetek's report, the patience times used for waiting emergency and public users in PURQ-AC are fixed at 28 seconds and 5 seconds respectively. The patience times of handoff/preempted queue and originating queue in our preemption threshold-based strategy are exponentially distributed with mean value being 5 seconds. The buffer sizes for both queues are 5.

1) Comparison of achievable channel occupancy : One main goal in this paper is to guarantee at least a certain amount (75%) of channel resources for public use. To evaluate which policies can achieve this goal, two different loads of emergency traffic are studied.

When the load of emergency traffic is at 30% of the system capacity as shown in Fig. 10, only up to 70% channel occupancy of public traffic can be achieved for the pure preemption and single preemption policies. This is because the emergency traffic uses another 30% of system capacity without being effectively constrained. For the preemption threshold based strategy and PURQ-AC, at least 75% can be guaranteed.

When emergency traffic is as high as 160% of the system load as shown in Fig. 11, it becomes much worse for pure preemption and single preemption. For the pure preemption policy, the best achievable channel occupancy for public traffic is only about 2%; the single preemption policy is obviously better (about 35%), but is still much lower than our desired value of 75%. In contrast, both the preemption threshold based strategy and PURQ-AC can still guarantee 75% of channel

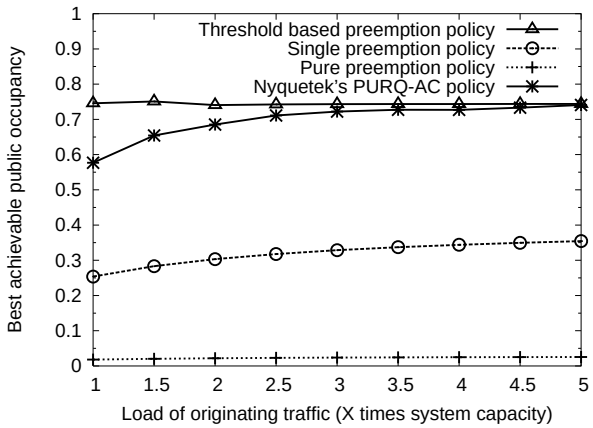

Fig. 11. Comparison of achievable channel occupancy for public traffic - Emergency traffic = 160% system capacity

resources for public use when the public traffic is heavy enough.

Another interesting phenomenon shown in Fig. 11 is that, when the public traffic is not heavy enough, the PURQ-AC policy can not guarantee 75% of channel occupancy for public traffic (also shown in Nyquetek's report, Fig. 3-7). For example, when the load of public originating traffic is at 100% of the system load, only 58% of channel resources are used for public traffic. In contrast, the preemption threshold based strategy can achieve 75% for all load cases. This means that *the preemption threshold based method is better than PURQ-AC in protecting public traffic*.

The main reason behind the above difference is that, with 1/4 scheduling, 75% of channel occupancy for public traffic is hard to be guaranteed due to some factors not considered in [5], among them are the load of public traffic and different impatience times of each class of customers. For example, very short impatience time (5 seconds) in the originating traffic queue will cause a lot of customers to drop their sessions before a channel is available, thus leading to much smaller amount of effective originating traffic to compete with emergency traffic. So strictly speaking, the statement in Nyquetek's report about using 1/4 scheduling to achieve 25% allocation for emergency and 75% allocation for public

does not consider all load cases. In contrast, our preemption threshold based method will take all factors into account and decide the best preemption threshold to achieve the desired channel occupancies.

In conclusion, among the four policies, the one proposed in this paper is the only one that can satisfy the requirement for protecting public traffic for all load cases. The PURQ-AC policy is also acceptable especially when public traffic is very heavy, which is most common when the disaster happens. On the other hand, the pure preemption and single preemption policy are not suitable since they can not protect public traffic effectively. Thus we will only compare the preemption threshold based method and PURQ-AC for other aspects of performance hereafter.

2) Success probability of each class: As another main goal, admission of emergency traffic should be guaranteed when its volume is not unexpectedly high. To compare the effectiveness of the preemption threshold strategy and PURQ-AC in this aspect, the achieved success probability of emergency traffic and handoff traffic is shown in Fig. 12. Here public handoff traffic is 6 sessions/minute, and public originating traffic is assumed to be 60 sessions/minute. It can be seen that by using the preemption threshold based method, almost all emergency requests are admitted when the load of emergency traffic is less than 20% of the system capacity. In contrast, PURQ-AC can only guarantee about 90% admission probability for emergency traffic even though its load is just 10% of the system capacity. The essential cause of this is the queueing mechanism and the impatience behavior. With PURQ-AC, emergency users need to wait some time in the buffer before being admitted, and they will drop the session requests when they become impatient. In contrast with preemption threshold strategy, emergency customers will be admitted immediately when the traffic volume is not higher than expected.

Although emergency traffic's admission is much more reasonably guaranteed in our strategy, the success probability of public handoff traffic is not as good compared with PURQ-AC since it reserves one channel for handoff traffic. However, if we also reserve one guard channel for handoff traffic, similar success probability for handoff traffic can also be guaranteed, and the behavior of emergency traffic is almost not affected. Furthermore, it is found out that the system utilization for PURQ-AC and the preemption threshold method are almost the same with the same amount of resources reserved. This is not surprising as queueing strategies are employed in both methods.

3) Waiting time: When the preemption threshold based method is applied, emergency traffic need not wait before being admitted, but public handoff and originating traffic may need to wait some time. However, the waiting time is pretty short as shown in Fig. 13. The average waiting time for handoff traffic is less than 2 seconds, while for originating traffic it is about 3.5 seconds; both are acceptable.

In contrast, emergency users in the PURQ-AC strategy have to wait, and the waiting time can be as long as 15 seconds. At the same time, the waiting for originating traffic is also longer than the preemption based strategy and approaches 5 seconds. In total we can conclude that PURQ-AC method will cause longer access time, especially for emergency users.

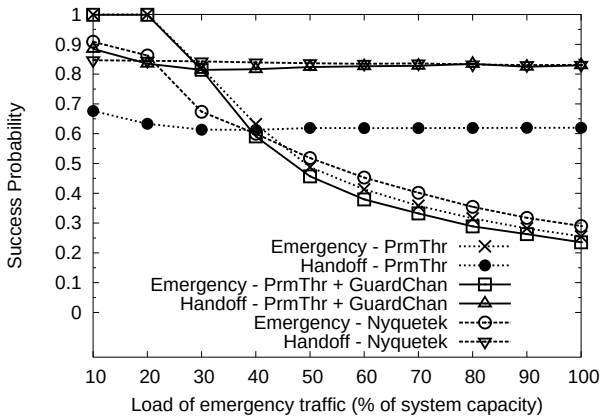

Fig. 12. Comparison of success probability

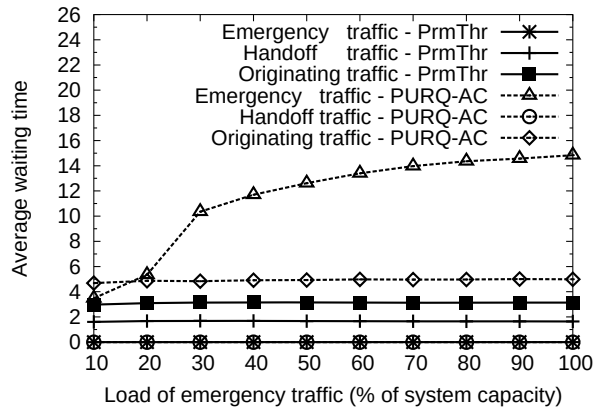

Fig. 13. Comparison of two methods: waiting time

VII. CONCLUSIONS

In this paper we introduce a preemption strategy that uses preemption thresholds to tune admission control of a wireless cellular network when emergency service is supported. Compared with the scheduling strategies in [5], our method can provide more guaranteed admissions for emergency traffic and protect public traffic more effectively when its load is not extremely heavy. Furthermore, immediate access of emergency traffic is guaranteed and the waiting times for originating traffic are shorter.

From the aspect of fundamental research, this paper also provides analytical solutions and insights for the gross service time in a multiple server system with loss possible. The result then is used for further study and we show the close connection between the system resource occupancy of low priority tasks and the number of tasks that avoid termination. Such work has not been performed to our knowledge and it can be applied to any system that uses a preemption based scheme.

Possible future work includes the use of different assumptions about impatience times of customers and applications of our theoretical results to different environments. In the future, emergency users will need to be supported in non-cellular technologies like wireless LANs, ad hoc and mesh

networks. Our preemption threshold based method could prove very useful for those applications.

REFERENCES

[1] Ken Carlberg, Ian Brown, Cory Beard, "Framework for Supporting Emergency Telecommunications Service (ETS) in IP Telephony", *Internet Engineering Task Force, Request for Comments 4190*, November 2005.

[2] Federal Communications Commission Report and Order, "The Development of Operational, Technical and Spectrum Requirements for Meeting Federal, State and Local Public Safety Agency Communication Requirements Through the Year 2010", FCC 00-242, rel. July 13, 2000.

[3] National Communications System, "Wireless Priority Service (WPS) Industry Requirements for the Full Operating Capability (FOC) for CDMA-Based Systems", prepared by Telcordia, Issue 1.0, Mar. 2003.

[4] National Communications System, "Wireless Priority Service (WPS) Industry Requirements for the Full Operating Capability (FOC) for GSM-Based Systems", prepared by Telcordia, Issue 1.0, Sept. 2002.

[5] 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.

[6] 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.

[7] S. Drekic, "A preemptive resume queue with an expiry time for retained service", *Performance Evaluation*, 54 (2003) pp.59-74.

[8] 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.

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

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

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

[12] 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.

[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.

[14] 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.

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

[16] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping", *IEEE/ACM Transactions on Networking*. vol. 2, no. 2, pp. 166-175, Apr. 1994.

[17] J. Zhou and C. Beard, "Weighted Earliest Deadline Scheduling and Its Analytical Solution for Admission Control in a Wireless Emergency Network", *International Teletraffic Conference(ITC19)*, August 29-September 2, 2005, Beijing.

[18] J. Zhou and C. Beard, "Tunable Preemption Controls for a Cellular Emergency Network", *IEEE Wireless Communications and Networking Conference (WCNC) 2007*, HongKong, March 11-15, 2007.

[19] D. P. Graver, "A waiting line with interrupted service, including priorities", *Journal of Royal Statistical Society B* 24 (1962), pp.73-90.

[20] L.Takacs, "Priority Queues", *Operations Research* 12 (1964), pp.63-74.

[21] P. D. Welch, "On preemptive resume priority queues", *Annal Mathematical Statistics* 35 (1964), pp.600-612.

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

[23] M. Segal "A Multiserver System with Preemptive Priorities", *Operations Research*, Vol. 18, No. 2, 1970, pp. 316-323.

[24] I. Mitrani and P.J.B. King. "Multiprocessor systems with preemptive priorities", *Performance Evaluation*, 1(2) pp.118-125, 1981.

[25] J.P. Buzhen and A.B. Bondi " The Response Times of Priority Classes under Preemptive Resume in M/M/m Queues", *Operations Research*, Vol. 31, No. 3, 1983, pp.456-465.

[26] H.R. Gail, S.L.Hantler and B.A. Taylor "On a preemptive Markovian queue with Multiper Servers and Two Priority Classes", *Mathematics of Operations Research*, Vol. 17, No. 2, 1992, pp.365-391.

[27] C. Jedrzycki and V. Leung "Probability distribution of channel holding time in cellular telephony systems", *Proc. IEEE Veh. Technol. Conf.*, GA, May 1996, pp.247-251.

[28] M. H. Ahmed "Call admission control in wireless networks: a comprehensive survey", *IEEE Communications Surveys & Tutorials*, Vol. 7, No. 1 2005 pp.49-68.

[29] Y. Ishikawa and N. Umeda "Capacity Design and Performance of Call Admission Control in Cellular CDMA Systems", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 8, 1997, pp.1627-1635.

[30] Mesquite Software, Inc. "CSIM19 User's Guide", Austin, Texas, 2001.

PLACE PHOTO HERE

Jiazhen Zhou received the B.S. degree in mathematics from Shandong University, China in 1995, the M.S. degree in Automatic Control from Shenyang Institute of Automation, Chinese Academy of Sciences in 19998, and the Ph.D degree in Telecommunication Networking from University of Missouri - Kansas City, United States in 2008.

He is currently a Postdoctoral researcher with the Department of Systems and Computer Science at Howard University, United States. His current research interests include performance evaluation and optimization of wireless networks, queueing theory, wireless mesh networks, emergency communications, and delay tolerant networks.

PLACE PHOTO HERE

Cory C. Beard (S'88 - M'99 - SM'08) received the B.S. and M.S. degrees from the University of Missouri, Columbia, in 1990 and 1992, respectively, and the Ph.D. degree from the University of Kansas, Lawrence, in 1999. He is currently an Associate Professor in the Department of Computer Science and Electrical Engineering in the School of Computing and Engineering at the University of Missouri - Kansas City. His current research interest is in the areas of prioritized and differentiated quality of service provisioning, network traffic engineering, queueing theory, the application of computer networks to disaster recovery operations, and the use of teams of robots for search and rescue after disasters. Dr. Beard was a recipient of the National Science Foundation CAREER Award for his project entitled "Priority Users and Applications on the Internet."