DELAY BASED APPROACH TO SUPPORT LOW PRIORITY USERS IN PREEMPTIVE WIRELESS NETWORKS

A THESIS IN
Electrical Engineering

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

MASTER OF SCIENCE

by
AASHISH CHANDRA

B.E., University of Madras, India, 1999

Kansas City, Missouri
2011

DELAY BASED APPROACH TO SUPPORT LOW PRIORITY USERS IN PREEMPTIVE WIRELESS NETWORKS

Aashish Chandra, Candidate for the Master of Science Degree

University of Missouri-Kansas City, 2011

ABSTRACT

At times of serious disasters (natural or man-made), wireless networks are quickly congested due to the sheer volume and stress on network resources, and, preferential treatment is necessary for National Security/Emergency Preparedness (NS/EP) users to combat the disaster by responding effectively and potentially save many lives. Under such circumstances, with scarce resources, the new request for sessions are denied and worse even, active sessions are dropped for general public whilst they have come to rely on these resources and depend on them especially during distressed times. Prior research has been conducted to examine upper limit (UL) and preemptive approaches to support emergency users but the traditional approach of blocking the capacity for emergency users is, from one perspective, restrictive to the general public.

In this thesis, we propose the delay-based soft preemptive approach to support the low priority users and provide an alternative to several preemptive policies by further examining them. We provide a queuing algorithm in the network that warns the low priority users with an active session of scarce resources thereby giving them an opportunity to complete their session prior to reducing the quality of service (QoS) of their session and moving their bandwidth to emergency users, if blocked. The emergency users in turn wait for the resources to become available and are on hold until resources become available. By creating a queuing modeling system for this algorithm, we present simulation model in C with results of our delay-based soft preemptive approach and examine other preemptive approaches to provide a comparative analysis. The results demonstrate that increasing the warning time also increases the number of sessions blocked for emergency users as well as general public due to further constraining the resources, however, this reduces the inconvenience of preemption caused to the low priority users.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Computing and Engineering have examined a thesis titled “Delay Based Approach to Support low priority users in preemptive Wireless Networks,” presented by Aashish Chandra, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Cory C. Beard, Ph.D., Committee Chair
Department of Computer Science and Electrical Engineering

Ghulam M. Chaudhry, Ph.D.
Department of Computer Science and Electrical Engineering

Vijay Kumar, Ph.D.
Department of Computer Science and Electrical Engineering

CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The wireless world has become a crucial component of the telecommunications industry and is growing rapidly. The question is no longer that how to get into the wireless market rather it is more like what else to leverage from it. Today, consumers and business people alike want fast, easy access to a vast variety of resources - from news and entertainment to corporate intranets and desktop capabilities. Moreover, providing access to these resources no longer depends on location, medium or device.

From the world of proprietary implementations, wireless technology has emerged to become an open solution for providing mobility as well as essential network services where wire-line installations had proven impractical [6].

To appreciate the growth of the wireless sector, it should be noted that in 1990 there were only 10 million cell phone subscribers worldwide, mostly using analog FM (first-generation) technology [10]. Today there are approximately 5.3 billion subscribers and projected to maintain the growth trajectory in the coming years (18.5% more mobile devices were sold in 2010 compared to 2009) [7]. The global view of mobile's statistics is detailed in Table 1. From the consumer retail standpoint, mobile commerce (m-Commerce) continues to show significant increase in year-after-year sales with $6.7 Billion in sales this year (increase of 91.4% from last year), and is estimated $11.6 Billion in 2012. By 2015, mobile commerce could top $31 Billion [11].

Inside of an office coupe or home, WPANs (IEEE 802.15 technologies – example, Bluetooth, HomeRF, and IrDA) allow connecting various personal portable devices without using cables. In the enterprise or campus environment, WLANs enable a company's employees to move about freely while remaining connected to their data networks. Outside of the office – in WWANs based on GPRS, UMTS or CDMA2000, consumers and business travelers are using an increasing number of Wireless Internet devices such as laptops, PDAs and smartphones (mobile phones with the Internet and e-mail access).

WLAN technology (for example, IEEE 802.11) enables a mobile lightweight device within a specific location operating in unlicensed spectrum to deliver higher data throughput capacity in concentrated areas. On the other hand, 2.5G/3G/3.xG capabilities offer extensive mobility features, significantly faster data rates and cost-effective wide area coverage [9]. WLANs operate in licensed-exempt band (example, the ISM band at 2.45 GHz) so it would be required by an operator to share the available spectrum with another operator. Spectrum sharing can potentially affect issues such as quality of service (QoS) and security.

Table 1: Key Global Telecom Indicators for the World Telecommunication Service Sector in 2010 (all figures are estimates) [7]

	Global	Developed nations	Developing nations	Africa	Arab States	Asia & Pacific	Europe	The Americas
Cellular subscriptions (mil)	5,282	1,436	3,846	333	282	2,649	741	880
Per 100 people	76.2%	116.1%	67.6%	41.4%	79.4%	67.8%	120.0%	94.1%
Fixed telephone lines (mil)	1,197	506	691	13	33	549	249	262
Per 100 people	17.3%	40.9%	12.1%	1.6%	9.4%	14.0%	40.3%	28.1%
Mobile broadband subscriptions (mil)	940	631	309	29	34	278	286	226
Per 100 people	13.6%	51.1%	5.4%	3.6%	9.7%	7.1%	46.3%	24.2%
Fixed broadband subscriptions (mil)	555	304	251	1	8	223	148	145
per 100 people	8.0%	24.6%	4.4%	0.2%	2.3%	5.7%	23.9%	15.5%

1.1 Wireless History

The 1990s were a period of tremendous growth for the wireless sector, evolving from a niche business to one of the dominant areas for growth in the 21st century. Few could have predicted the arrival of existing technological gadgets and applications possible with these hi-tech devices. Likewise, there were some amazing and startling failures in the wireless sector, despite the brilliant engineering and technological efforts that went into their formations.

One of the most successful wireless communications technologies of the previous decades was Code Division Multiple Access (CDMA), pioneered by Qualcomm, Inc. Qualcomm introduced its CDMA concept for mobile radio in 1990, at a time when U.S. cellular industry was selecting its first digital mobile telephone standard [6].

Just prior to Qualcomm's introduction of its wideband digital CDMA mobile radio standard in 1990, now known as IS-95, the U.S. cellular industry was assured to select TDMA (which became IS-136) as the digital successor to the analog AMPS standard. The European community had already adopted GSM for its own pan-European digital cellular standard a couple of years earlier, and Japan's second generation digital TDMA standard, PDC (Pacific Digital Cellular), introduced shortly after IS-136's acceptance in the U.S. As cellular telephone service caught on with customers, governments across the world auctioned additional spectrum (the Personal Communications Services, or PCS spectrum) to allow new competitors to support even more cellular telephone subscribers. The PCS spectrum auctions of the mid-1990's created a vast increase in frequencies for cellular telephone providers across the globe, thereby providing the proving ground for the second generation of cellular telephony (2G, the first generation of digital modulation technologies) [10].

While the pioneering design of the GSM, which included international billing, short messaging features, and network-level interoperability, now enjoys the lead in today's global wireless market, it is also evident that wireless CDMA was a breakthrough technology, offering increased wireless capacity by increasing channel bandwidth and moving complexity in the handset to low-cost baseband signal processing circuits [10].

Third-generation (3G) is mobile multimedia, personal services, convergence of digitalization, mobility, the Internet, new technologies based on global standards, all of the above. The end user is able to access the mobile Internet at the bandwidth (on demand) from hundreds of kilobits per second to about 2 Mbps. The 3G standard accelerated the expansion of mobile communications market post-2G and propelled the development of smartphones leading to phenomenal demand for mobile internet connectivity. Internet browsing enabled by mobile broadband in 3G propelled the m-Commerce market for retailers and businesses, securing further momentum in the development of standards.

The 3rd Generation Partnership Project (3GPP) created GSM specifications for 3G within International Mobile Telecommunications 2000 (IMT-2000) guidelines, including standards for reliability and speed. The Universal Mobile Telecommunications Systems (UMTS), created and revised by 3GPP, is derived from GSM in terms of encoding methods and hardware. The CDMA2000 system, standardized by 3GPP2 evolved from original IS-95 CDMA system is used in North America, China, India, Japan, South Korea, Southeast Asia, Europe and Africa [4].

1.2 Wireless Path to Future

With over 5.3 billion mobile phone users currently, and packet-based multimedia services accounting for a respectable part of all wireless traffic, it is natural to provide more capacity in the mobile network, and higher bandwidth in the radio link, radio access network, and core network [8]. The proliferation of m-Commerce market driven by smartphones (for example, the iPhone) has further driven the momentum in the industry to evolve the current infrastructure, network services, and end-user applications towards an end-to-end IP solution capable of supporting QoS to meet the needs of the dominant data traffic. The view of wireless is continuously evolving though it is at the beginning of the significant change of the wireless systems and services. Mobile wireless technologies beyond the 3G already exist while 3.xG standards are currently in place as many services provide broadband access of several Mb/sec to smartphones and mobile modems in laptop computers. Figure 1 shows the speeds that 3G cellular provides in different environments.

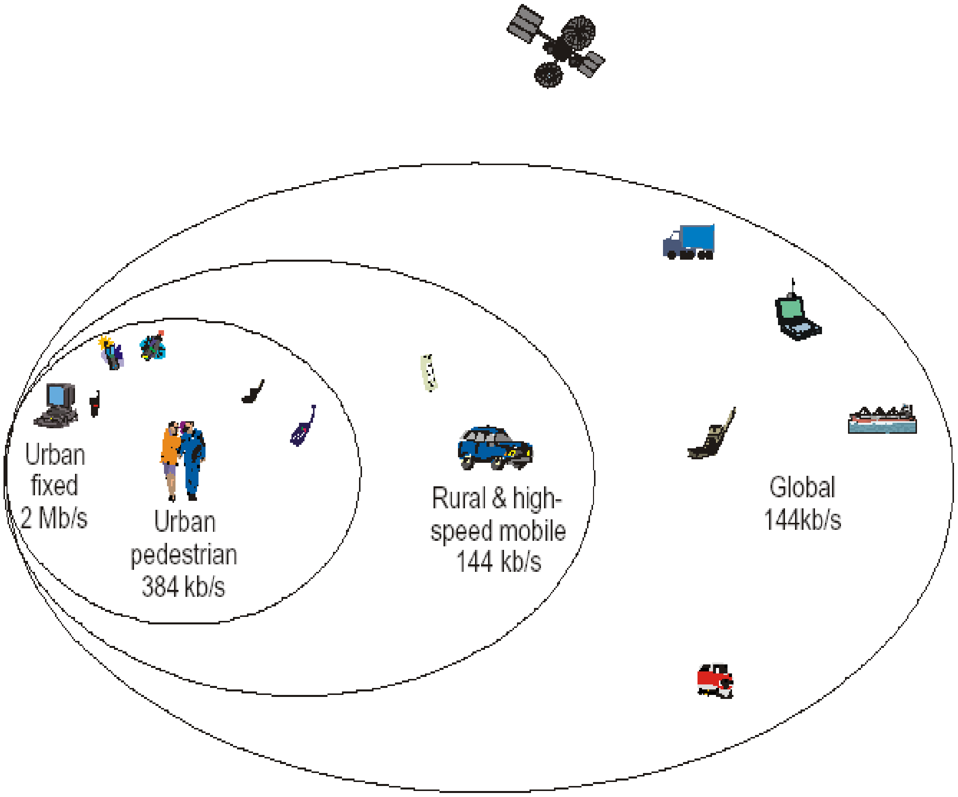

Figure 1: 3G Supplies Service from 144 Kb/s to 2 Mb/s [8]

Fourth-generation (4G) wireless is a major move towards ubiquitous wireless communications systems and seamless high-quality wireless services. The next-generation wireless beyond 3G/3.xG is an effort towards a new wireless world that is a converged broadband wireless system (wireless mobile and wireless access). This converged system will be extremely important in developing countries to greatly improve the wireless infrastructure and provide the solutions of low cost, secured applications, and integrated services to huge volumes of mobile subscribers.

Fourth-generation (4G) wireless systems are under development around the globe with the objectives of having performance of 1Gbits/sec for stationary and 100Mbits/sec for mobile operations. From a business perspective it is the business opportunity of the 21st century. Regardless of which multiple access standards are widely deployed, the challenges are significant in the area of hardware and particularly in the software architecture to realize the goals of 4G.

1.3 Congestion in Wireless Networks

There is an exponential increase in number of wireless users which has resulted in greater airwave congestion/contention and an over-subscription of available bandwidth. Many of the world's cellular telephone systems do not have sufficient capacity to support demand in urban areas. In Japan, the PDC technology has been strained on capacity in some cities, and the same is true of GSM in some European metropolitan areas. In developing countries like India and China, millions of cell phone subscribers are being added each month. But, as subscriber growth continues to increase in US, Europe and rest of the world, carriers and infrastructure providers are facing a huge challenge in addressing bandwidth problems associated with this exploding capacity [8].

QoS has always been considered the key to providing the service to Wireless users and subscribers. Admission Control policies can be used to meet the applications' requirement for the QoS by improving the connection performance and reducing the blocking probability for higher priority traffic that has more importance. Connection

Admission Control (CAC) is an Admission Control Polity that can be applied in a multi-service environment where there are different services competing for resources in the network. Several CAC policies can be defined depending upon the service and parameter requirements [1].

1. Complete Sharing
2. Complete Portioning
3. Trunk Reservation
4. Partitioning
5. Upper Limit Policy
6. Guaranteed Minimum
7. Preemption

This research examines the preemption approach, listed above and specifically details the delay-based preemption approach as a consideration. Chapter 2 provides the background research already done on the loss networks and preemption policy.

1.4 Need for Prioritization

With the advancement of technologies and the wireless world, through the continuous effort to meet the QoS demands of users and bandwidth requirements, it apparently seems impractical to meet the capacity requirements all of the time everywhere. Furthermore, several recent natural disasters in the U.S. and around the globe have shown the applicability and usefulness of cellular telephony in providing emergency telecommunications for the Local, State and Federal officials who are on a disaster site or on the move in stressed environments. However, due to the heavy cellular traffic demand placed upon the surviving cellular systems in the aftermath of disasters, severe cellular network congestion has been experienced resulting in high call blocking to the critical disaster relief officials when communications are needed the most. The tragedy at the World Trade Center in New York, on September 11th 2001 supplied anecdotal evidence of the need for a system that would enable the stress networking and disaster recovery when there is network congestion. To this end, the IST has initiated a significant effort to implement the wireless priority service (WPS) capability for the nation's critical disaster response personnel.

WPS provides a means to queue incoming National Security/ Emergency Preparedness (NS/EP) call attempts for the next available channel, thereby enhancing the NS/EP user’s ability to complete calls during a disaster over the congested cellular network. In a situation where no voice channel is available, an authorized user invokes the WPS capability by dialing an assigned feature code plus the destination number. The cell-site passes the WPS call invocation to a Mobile Switching Center (MSC) where verification occurs. Once validated, the MSC queues the WPS call attempts until a cellular channel becomes available. When a channel is available, the MSC alerts the NS/EP user who can then complete the emergency call.

In order to effectively respond to serious disasters (natural or man-made), it is necessary to empower the NS/EP personnel with higher priority since many lives could be at stake. Prioritization allows for NS/EP users to respond to the disastrous situations and also provide disaster recovery.

1.5 The Idea of Preemption

With the advent of wireless multimedia and wireless Internet business, deregulation ensures fair game for everyone to start with the minimum acceptable level of service. In today's wireless Internet, the ability to control QoS – to deliver the services that meet the end users' expectations, e.g. voice quality, viewable video, fast web site access, and safe commerce transaction, is the most essential quality required.

During the stressed wireless network situations resulting from natural or man-made disasters, while the NS/EP users need priority driven access for themselves to effectively respond to disasters, the general public has also come to rely on mobile broadband and depend on their wireless network during such disasters (example, 911 calls or other emergencies). The US Government Emergency Telephone Service (GETS) has a policy that preemption should not be used to support emergency users. However, there are several approaches to preemption that could be explored to support both the NS/EP as well as low priority users (general public). Figure 2 depicts the set-priority model wherein priority is invoked to provide preferential treatment to emergency personnel.

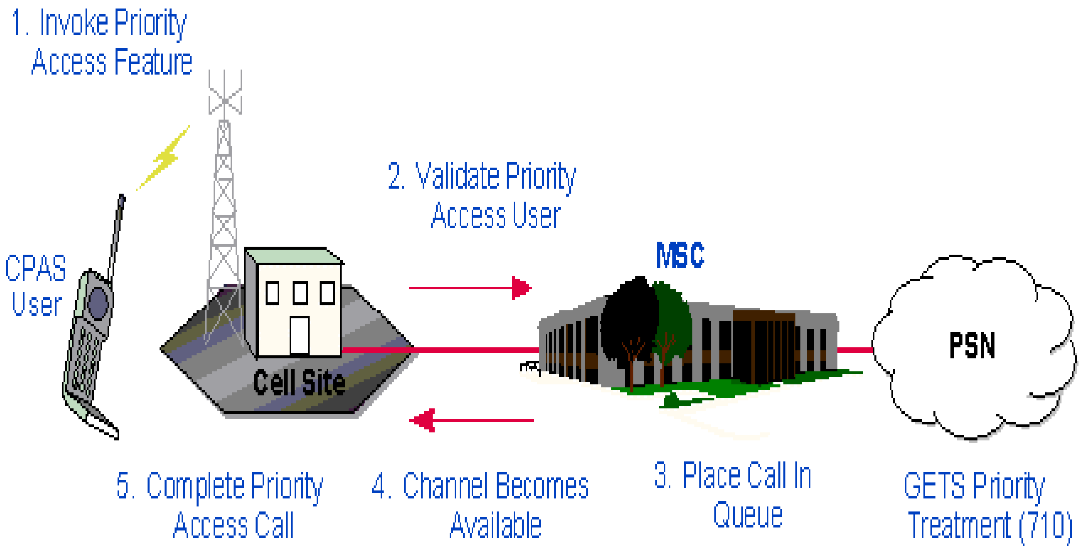

Figure 2: The Set-Priority Model

QoS has become one of the key differentiators used not only to attract the market share of new customers, but also to gain customers' loyalty and retention. Customers' perception of getting their value for the money has become more evident during the great recession in the last few years. With the economy still in turmoil, consumer's behavior has changed concerning goods and services. Consumers now, more than ever, might accept the compromise of lower QoS in serious disastrous situations and if they have an opportunity to complete their active session rather than having their active sessions dropped or connections denied (blocked).

1.6 Outline of Thesis Work

This thesis work is based on making the best use of available network resources to deliver reduced QoS to low priority users after giving them an opportunity to complete their active session while the NS/EP (high priority) user waits for the connection to become available, at times of severe network congestions during disastrous situations. This delay-based preemption approach (NS/EP user is put on hold) utilizes the soft preemptive policy wherein QoS is reduced to accommodate the connection request from high priority users if their connection is blocked due to the congestion. In this research, a hard preemptive policy was not examined due to the nuisance of “cold” session drop for low-priority users where active sessions are dropped to free up capacity for high priority users.

This thesis work would be considering connection-oriented traffic over the wireless networks. This thesis work is a companion research to [2] where it has been demonstrated how soft preemption could be a solution to giving priority to high importance users during the congestion in the stressed networks.

Chapter 2 provides the related work background that has been a motivation for work in this thesis. Loss networks and connection management is highlighted while several preemption policies have been discussed. The preemption approaches in [2] are addressed while the concept of a delay-based soft preemption approach is further detailed.

Analysis and the simulation approach for the delay-based mechanisms are primarily focused in Chapter 3. The algorithm for delay-based soft preemption is detailed to support the low priority users within wireless communication networks as well.

Chapter 4 shares the statistical results of the simulation model developed for this thesis work and Chapter 5 provides the summary of results and the conclusion of the thesis work. It closes with some thoughts on future work in this domain as well.

1.7 Summary of Results

Utilizing the queuing modeling system for our algorithm, simulation results of our delay-based soft preemptive approach demonstrate that increasing the warning time (hold time for high priority users) also increases the number of sessions denied (blocked) for emergency users as well as the general public due to further constraining the resources. This however, reduces the inconvenience of preemption caused to the low priority users.

Furthermore, one of the simulation results show that with 1000 new connection requests coming from each of the emergency and low priority users, within the stressed and congested network, and no delay mechanisms, 134 calls from low priority users were not affected (either blocked or preempted) but by introducing the hold time of 1 minute for emergency users, the number of calls not affected for low priority users went up to 196. Therefore, we found that within the soft preemptive model, the delay-based mechanism increases the number of calls not affected for low priority users.

Although it is proven that delay-based mechanisms within preemptive policies favor the low priority users with fewer affected calls, in support of [2], the decision to adopt preemptive approaches to support emergency and low priority users is more value driven rather than qualitative. The inconvenience of having sessions preempted by emergency users is a major consideration in making that determination. This thesis work builds upon [2] to provide yet another solution to support emergency traffic and low priority users in stressed network situations.

CHAPTER 2

RELATED WORK

2.1 Traffic Engineering

Considering a scenario where we get 3,200 calls in an eight-hour day and assuming that each call lasts three minutes, each person can handle 20 calls an hour. Logically, only 20 incoming lines are needed. But, following a typical distribution pattern, 550 or 600 of calls will arrive during the busiest hour of the day. And then they will bunch-up during that hour, leading to times when the network is going crazy and times when the network resources are free.

A simple Arithmetic approach in dealing with configuration decisions about telecommunications networks could result in calls bunched-up, i.e., too few trunks, too few subscribers and too many unhappy callers. The discipline that uses mathematical formulas to making decisions concerning network resources and network traffic is called “Traffic Engineering”. The basic concepts of traffic engineering including Loss Networks, Erlang B, and Blocking Probability are discussed to have an insight about why the idea of Preemption is optimal for priority traffic.

2.2 Loss Networks

A loss network has multiple nodes connected by trunks (or links), each of which contains a number of circuits, as depicted in Figure 3. In Figure 3, there are 5 nodes and 5 trunks with trunk *j* having K_j circuits. The nodes actually play no role in the model. The loss network carries multiple classes or types of calls, which are distinguished by the set of trunks (or route) they require, by the number of circuits required on each trunk (which need not be the same on all trunks) and by the average call holding time. Each call holds all its circuits for the duration of the call [5].

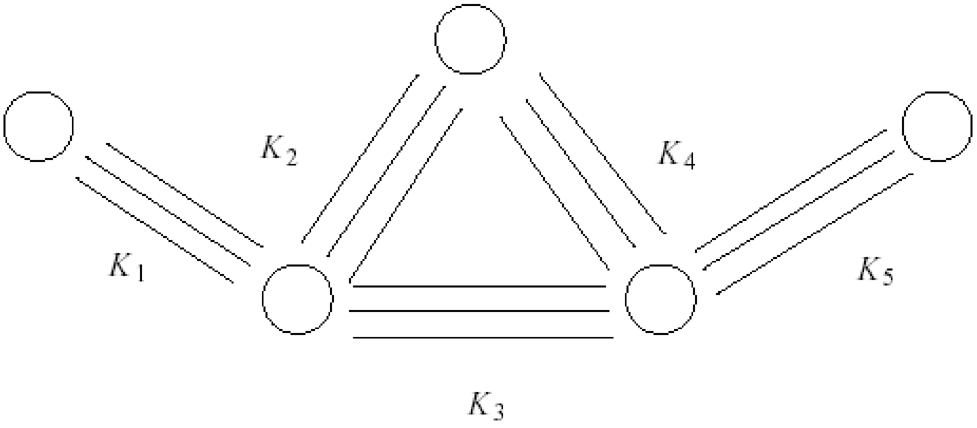

Figure 3: An example of a loss network [5]

For the example in Figure 3, the set of routes could be

5 = { {1}, {2}, {1,2}, {3,5}, {4,5}, {1,3,5} }.

In this example there could be 12 call types; two corresponding to each route (subset) with the route indicating the trunks needed for each call. The number of circuits needed by each call type on each trunk must also be specified. There might be more call types than routes, because different call types with the same route may have different circuit requirements. The steady-state distribution depends on only one more parameter for each call type: the offered load, which is the call arrival rate multiplied by the mean call holding time [5].

In the basic loss network model, calls of each type arrive according to a Poisson process. Each call is accepted only if all the trunks on its route have enough circuits available to support the call; otherwise, the call is blocked and lost (without generating retrials or otherwise affecting future arrivals). Loss networks have many applications; e.g., a loss network may represent a database, a wireless communication network or a circuit-switched computer network as well as a circuit-switched telephone network [5]. The connection oriented (C-O) protocol establishes an end-to-end logical or physical connection before any data may be sent, however since acknowledgement of receipt isn't incorporated, it is not very reliable. QoS leverages and exploits this inflexibility in resource reservation to grant enough resources to services depending on the urgency or need for emergency traffic.

The traditional loss network model assumes Complete Sharing (CS) of the circuits on a trunk among all competing traffic classes. However, considering other sharing policies could provide different grade-of-service and protect one traffic class from another.

2.3 Erlang B Formula for Blocking Probability

In 1917 the Danish mathematician A.K. Erlang published his famous formula

$$E(\lambda, C) = \frac{\lambda^c}{C!}\left[\sum_{n=0}^{C}\frac{\lambda^n}{n!}\right]^{-1}$$

for the loss probability of a telephone system [3]. The problem considered by Erlang can be phrased as follows. Calls arrive at a link as a Poisson process of rate λ. The link comprises C circuits, and a call is blocked and lost if all C circuits are occupied. Otherwise the call is accepted and occupies a single circuit for the holding period of the call. Call holding periods are independent of each other and of arrival times, and are identically distributed with unit mean. Then Erlang's formula gives the proportion of calls that are lost.

The probability that all servers will be busy and the call will be blocked or lost when a call attempt is made is called the Blocking Probability. By determining the blocking probability for a particular network, its grade-of-service (or, GoS) could be determined. GoS has always been considered the key to providing the service to wireless users and subscribers. And, one way of attaining the GoS is by reducing the blocking probability.

The Erlang B formula assumes infinite sources which jointly offer traffic to servers (example, link in trunk group). The rate of arrival of new calls is constant λ, not depending on the number of active sources (assumed to be infinite). The rate of call departure is the service time constant μ. The Erlang loss formula calculates blocking probability in a loss system, where if a connection is denied due to congestion, the call is dropped. The formula also assumes that blocked traffic is immediately cleared.

Identifying Erlang's concept of statistical equilibrium with the stationary measure of a Markov process delivers interesting results. Thus if call holding periods are exponentially distributed then the number of lines occupied is a finite Markov chain as shown in figure 4, and Erlang's formula gives the stationary probability that all C circuits are busy. If call holding periods are arbitrarily distributed then the stochastic process describing the number of circuits occupied is more complex. Nevertheless Erlang's formula still gives the stationary probability that all C circuits are busy.

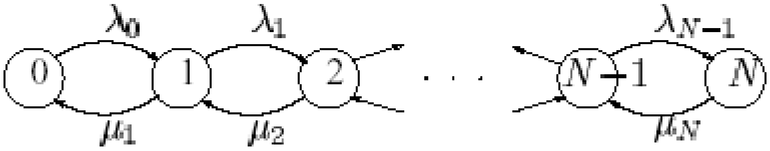

Figure 4: Markov Model

In addition to the Erlang's loss formula, a Q Matrix could be employed to determine the blocking probability of a loss network. The probability of preemption could also derived by using the Q Matrix by computing state probabilities for the Markov chain over a set of states [2].

2.4 Preemption Policies

Preemption in Wireless Communications Networks refers to a policy that allows the high priority traffic to run-over the lower priority traffic. Thereby, disconnecting low-priority calls that are already in progress over the bandwidth channels. The objective is to free the channels occupied during the network congestion by the low-priority users for the high-priority users' availability. These high–priority users could be the NS/EP or other emergency workers.

In this section all the options that are possible for managing the traffic are listed. Starting from the very basic methodology of using no preemption to using the delay-based soft-preemption for the data traffic, we make an effort to list all the various zones possible. We have also listed the options we are not considering for the analysis and simulations. Following are the different approaches:

1. Complete Sharing (CS) and Upper Limit (UL) Policies: Preemption is not allowed.
2. Hard Preemption: Preemption is allowed anytime and the preempted connection has no opportunity to resume or finish the active session.
3. Soft Preemption: Preemption is allowed but instead of dropping the active session for low priority users as in hard preemption (and thereby causing nuisance), QoS of the session is reduced for low priority users to free up resources and the freed channels (or, Bandwidth) are allocated to high-priority traffic.
4. Partial Soft Preemption: Soft preemption is allowed but with a control to optimize the throughput. A specific threshold is set and soft preemption is only allowed up to that threshold.

5. Delay-Based Hard Preemption: Preemption is allowed but only after a delay (example 1 minute) to caution low-priority users of stressed resources. After the wait time during which the emergency user is put on hold, the preemption occurs and there is no opportunity for preempted connection to resume.
6. Delay-Based Soft Preemption: Preemption is allowed but only after a delay (example 1 minute) to caution low-priority users of stressed resources. After the wait time during which the emergency user is put on hold, the soft preemption occurs and the QoS of a low-priority call is reduced to accommodate the emergency user.
7. Preempted Revival: A low-priority call is preempted but it revives after the high-priority call is finished and the channel is free. Several options could be considered in this scenario with variability to ensure a low priority user is put on hold for only once and whether or not to link the call with a preempting emergency session (might result in longer hold times).
8. The high priority call is put on hold up to 2 min to see if a channel becomes available. Then it will preempt another call with no warning to the other call.

Not allowing preemption could be restrictive to both the emergency users as well as the low priority users since optimization of available network resources and manipulation of occupied resources could benefit the NS/EP users during times of severe disasters. Prior research work has been done to examine the opportunities presented by hard preemption, soft preemption and partial soft preemption. Various scenarios with comparative analytics are examined in [2] for these policies.

As an illustration, a multi-dimensional Markov chain for a soft preemption policy is modeled at the session level in figure 5. Class 2 has preemptive priority over class 1 with Capacity (C=6), Arrival rate (L=λ=2) for both class 1 and class 2 traffic, and Service Time (Mu=μ=1) for all class 1, class 2 and class 1` sessions. Class 1` is the class for soft preempted users.

A delay-based soft preemption approach analyzed and simulated in this thesis as inspired by [2]. Although the inconvenience of wait time exists for emergency users in this approach, the low priority users get the opportunity to complete their call before the QoS is reduced for their active session.

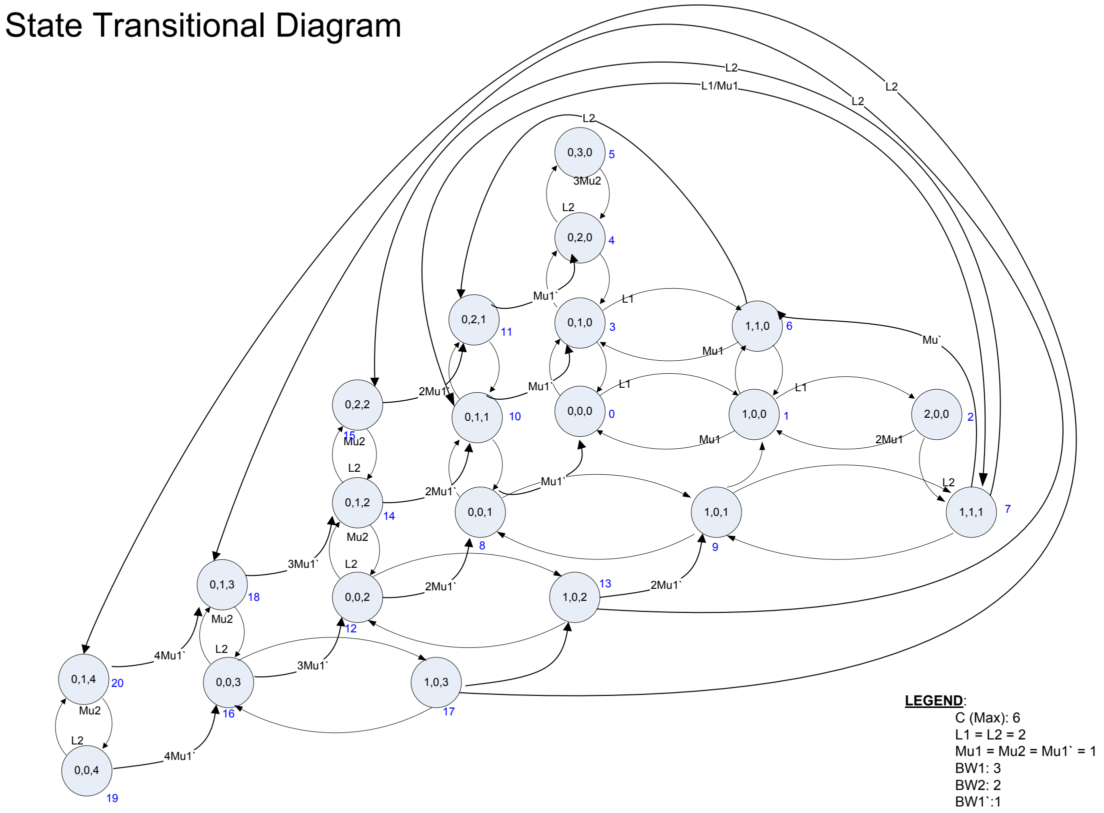

Figure 5: Multi dimensional Markov chain for soft preemption

Policies including preemptive revival are not of interest in this research since a low-priority user cannot be expected to hold for as long as the emergency traffic is utilizing the channel bandwidth. In such event, a low-priority user would likely drop the connection and try again for channel availability. In the event of stressed networks, they will not get the connection established due to a constraint on resources.

CHAPTER 3

ALGORITHM AND SIMULATIONS

3.1 Assumptions

In this paper, we have made the following assumptions to perform the analytics and simulations –

1. Only two classes of traffic are assumed. So, by implication we are only considering two priority levels – NS/EP (emergency) users and the low-priority users.
2. The arrival rate of calls and the rate of service for both the classes of traffic is the same.
3. A low priority can only be blocked at its initial request level. It cannot be admitted at lower QoS.
4. In case the network has no remaining low-priority calls (class 1 calls) since all were preempted, it is assumed that the higher priority call will be blocked and will not hard-preempt the end states (i.e., drop the reduced QoS call).
5. A generator process generates customer process arrivals and the connections each receive service in first-come, first-served basis at a set of facilities.

These assumptions aided us to concentrate on the effects of delay based mechanisms relative to the hard, soft and partial soft preemptive systems.

3.2 Algorithm for Delay-Based Mechanisms

The algorithm for the delay based soft preemptive approach utilizes the assumptions mentioned in the prior section –

Step 1: New connection request arrives

Step 2: If capacity is available, the call is admitted

Else

Step 3: Is it a high priority connection request?

Step 4: If not, deny the connection request for a low-priority call (block)

Else

Step 5: Are there enough low priority calls available (in session) to provide enough bandwidth to a new high priority call?

Step 6: If no, deny the connection request to the emergency user (block)

Else

Step 7: Request the emergency user to be on hold for at most the configured wait time and, warn the low priority session(s) of reduced QoS to give them the heads-up

Step 8: Wait for a configurable hold time (delay), and admit the high-priority call by reducing QoS for the low priority call(s) (soft preempt) or admit the high priority immediately if low priority call(s) end before the hold time.

3.3 Simulations

The details of simulations setups are provided in this section. To secure the best results, we opted for CSIM as the simulating tool. OPNET was also considered but due to some complexities, the circuit-switched model in OPENT was not available to the research team. To simulate the preemption-based model, coding was done in CSIM on Linux platform.

Class 1 is defined as the call request from a low-priority user while class 2 is an emergency user request. Class 1` is the preempted state of class 1 after class 2 preempts it. The parameters of the model, defined in the simulations are -

1. Number of arrival requests – numarvs
2. The mean inter-arrival rate for class 1 and class 2 – iarate
3. The mean service time for class 1 and class 2 calls – srvtm
4. The capacity available in the network (number of servers) – capacity
5. Bandwidth for class 1 – BW1
6. Soft preempted bandwidth for class 1` - BW1`
7. Bandwidth for class 2 – BW2
8. Delay wait time for class 2 call prior to soft-preempting the class 1 call

The simulator function starts the simulation for the arrivals as defined in the parameters while the generator function processes the generation for class 1 and class 2 calls. With calls arriving for the low-priority users, a function looks for available capacity and denies if no more capacity is available. In the meanwhile, calls arriving for emergency users look for available bandwidth and preempt any available class 1 calls, if any, after waiting for the delay time. If no class 1 calls are available to preempt, class 2 calls are denied connection as well.

If delay wait time for class 2 calls is defined as zero, the model represents a soft preemption simulating model. Furthermore, if soft preempted bandwidth of class 1` is also assigned as zero during the definition of parameters, the model becomes a hard preemption simulation model. This allows us to run the simulations and provide comparative analysis of all the three major approaches – hard preemption, soft preemption and delay based preemptions. The blocking probability of class 1 and class 2 calls along with the preemption probability of class 1 calls could be reviewed to conduct the analysis.

CHAPTER 4

ANALYSIS AND RESULTS

4.1 Hard and Soft Preemptions

Soft Preemption, from one perspective is more appealing and promising because the low priority calls are not hard preempted and hence sacrificed, instead a trade off is done with the QoS. By developing the state models and thereby developing the Q-Matrix, the probability of preemption or reduced QoS and blocking could be derived. We started with minimum values of Capacity to derive optimal capacity for simulation purposes.

Utilizing the sample Q-Matrix formula shown, with λ1 and λ2 as rate of arrival for class 1 and 2 respectively, μ1 and μ2 as rate of service time for the traffic calls, the probability of preemption for class 1 calls could be derived using the formula

$$Q = \begin{bmatrix} -(\lambda 1 + \lambda 2) & \lambda 1 & 0 & 1 \\ \mu 1 & -(\mu 1 + \lambda 1 + \lambda 2) & \lambda 1 & 1 \\ 0 & 2\mu 1 & -2\mu 1 & 1 \\ \mu 2 & 0 & 0 & 1 \end{bmatrix}$$

$$\Pr\ \{\text{preemption}\} = [0 \quad 0 \quad 0 \quad 1] * Q^{-1}$$

The probability of hard preemption, Pr {preemption} and the probability of soft preemption, Pr {Reduced-QoS} derived for $C_{max} = 3$, 4 and 5 are listed in Table 2. It should be noted that for minimum capacity values, the probabilities of preemption for hard versus soft preemption not vary. Running through the simulation tool, these results were verified to be accurate. For the purposes of this research, $C_{max} = 18$ was acceptable after calculating the probability of preemption for varied C_{max} values including $C_{max} = 30$.

Table 2: Probability of Preemption

	Pr {preemption}	Pr {Reduced-QoS}
$C_{max} = 3$	0.666	0.666
$C_{max} = 4$	0.666	0.666
$C_{max} = 5$	0.545	0.546
$C_{max} = 30$	0.758	0.799

Utilizing the $C_{max} = 18$ with NARS (number of arrivals) = 10000, the following parameters were used to derive the comparison in blocking probabilities for class 1 and class 2 calls and the preemption of probability for class 1 calls. Class 1 calls are considered as low priority traffic while class 2 is emergency traffic: λ1=λ2=3, µ1=µ2=0.2, BW1=3, BW2=2 and BW1`=0 or 1, depending on hard or soft preemption simulations. In this scenario, the blocking probability for class 1 calls, B1 reduces from 87% to 84% when switching from hard preemption to soft preemption while, the probability of being preempted for class 1 calls, P1, does not reflect much change between hard versus soft preemptive approaches. The probability of blocking for class 2 calls, B2, however, shows an increase from 46% to 52% when switching from hard to soft preemption.

4.2 Soft and Partially-Soft Preemptions

A partial preemption policy is a system that allows preemption from only some of the states. Thereby, the reduced QoS could be controlled in favor of the low priority users so that only the required capacity is freed up to accommodate the emergency traffic. Soft preemption in contrast, preempts all possible states. The soft preemption, as noted in previous section, does not provide a reasonable contribution in reducing the B2 and neither does it substantially reduce B1 or P1. Therefore, soft preemption in itself may not be the most optimal approach to benefit the emergency or low priority users.

Utilizing the $C_{max} = 18$, $\lambda1=\lambda2=4$, $\mu1=\mu2=1$, BW1=3, BW2=2 and BW1`=1, the simulation results are compared for soft and partially soft preemption approaches. In this model, end states in the state transitional chain are not allowed to be preempted by the high priority call. The states blocked from preemption are (3,0,0), (4,3,0) and (2,6,0). The results demonstrate that the blocking probability for class 1 calls, B1, reduces from 37.1% to 36.1% when using only the partial soft preemption approach while the probability of preemption for class 1 calls, P1, is also slashed from 20.2% to

9.54% only. The blocking probability for higher priority class 2 calls, B2, increases from 3.5% to 12.9% in this scenario.

Although, the numbers of calls blocked for high priority users are increased from 3.5% to 12.9%, partial soft preemption provides considerable contribution in reducing the impact on the low priority class by substantially reducing the probability of preemption.

Furthermore, we found that the total number of class 1 calls affected, F1 slipped from 49.8% to 42.2% when switching from soft preemption to partial soft preemption. Total calls affected for low priority users are found from the following formula.

$$\text{Total\_Calls\_1\_Affected} = F_1 = B_1 + (1-B_1) * P_1$$

4.3 Delay-based Soft Preemptions

A delay based preemption policy allows the emergency users to wait for capacity to become available. During the time they are put on hold, let's say 20 seconds, a low priority caller is issued a warning for that duration so that they have the opportunity to complete their call before it gets preempted by the emergency caller. This approach of inducing delay in accepting the connection request from an emergency user could be implemented with or without further preempting the low priority call. In the event no preemption is permitted, the emergency user will simply be put on hold until enough sessions are completed and capacity is freed up for them to establish the connection. As mentioned earlier, this non-preemptive approach is not of interest in this research since it could be frustrating and impractical to put the emergency user on hold for period of time when they have to respond to the disastrous situations with urgency.

Examining the delay based preemptive approach by utilizing $C_{max} = 18$, NARS= 10000, $\lambda1=\lambda2=3$, $\mu1=\mu2=0.2$, BW1=3, BW2=2 and BW1`=1, we generated the simulation results for B1, B2 and P1, as shown in table 3 below. We learnt that the B1 and B2 increase with delays while the probability of preemption for class 1, P1, decreases as hold time increases and eventually becomes zero. This provides an opportunity to identify the optimum delay (hold time) to create a reasonable and sustainable balance between the blocking probabilities and the probability of preemption so that there is not a substantial adverse affect on emergency users.

Table 3: Delay based preemption

Delay (min)	B1	B2	P1
0	0.72	0.527	0.671
0.333	0.741	0.567	0.502
1	0.767	0.603	0.278
2	0.775	0.621	0.099
100	0.787	0.631	0

To plot the graph for this simulation, the delay time (in minutes) was normalized with the call service time. For example, since µ1=0.2, the SVTM would be 5, and a normalized value of 1.0 would mean the wait time would also be 5. Figure 6 depicts the delay based preemption mechanism.

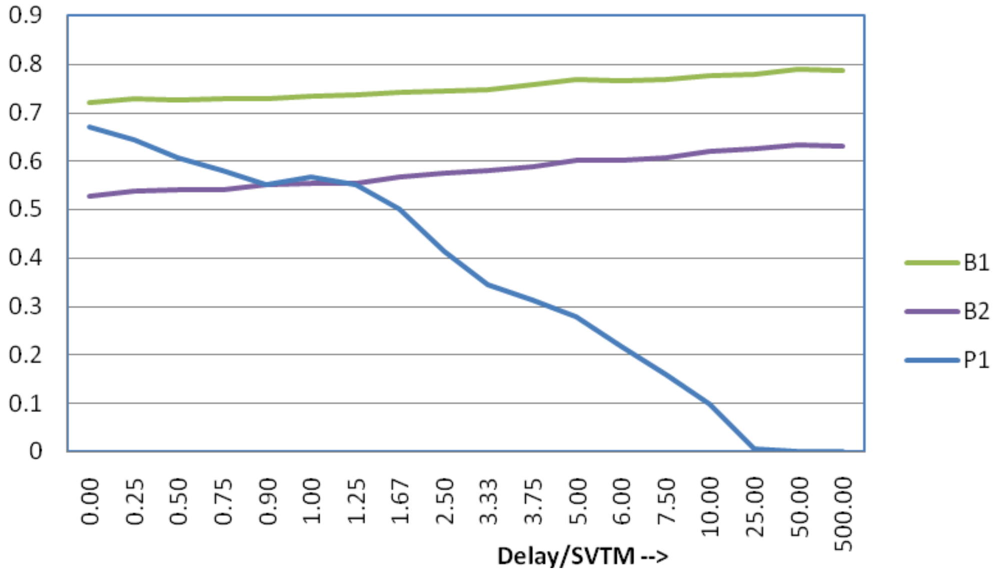

Figure 6: Delay based preemption mechanism

The total number of calls affected relative to blocking probabilities for class 1 and class 2 calls and the probability of preemption are depicted in Figure 7. The total class 1 calls affected, as demonstrated in this figure, are reduced as the delay hold time increases in value.

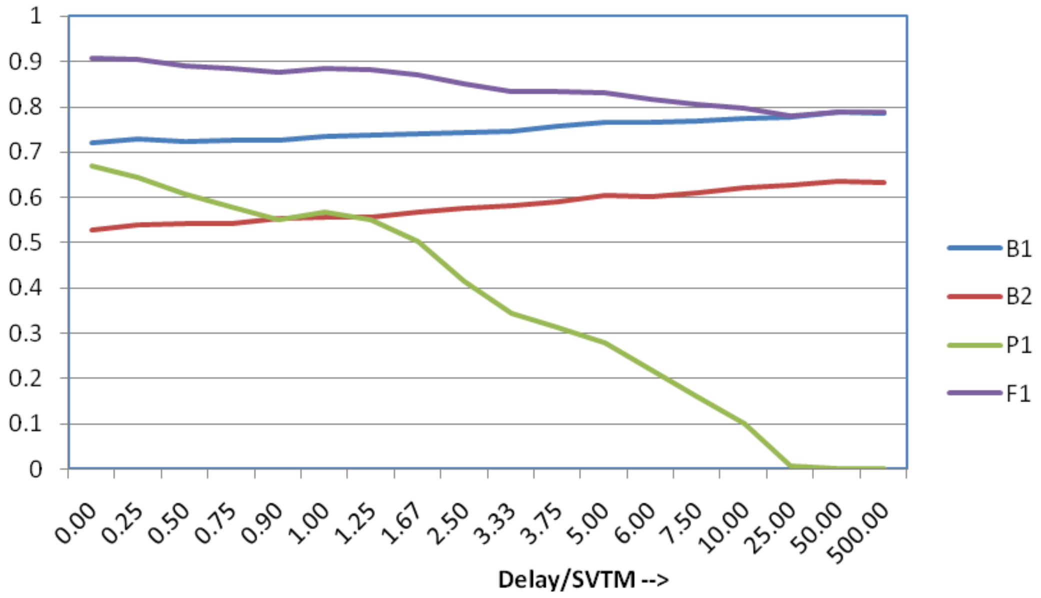

Figure 7: Affected class 1 calls in delay based preemption

We also simulated the results by varying the NARS (number of arrivals) while keeping the parameters the same to demonstrate the impact of arrival on the model, as shown in Figure 8. The delay based mechanism shows slight but inconsequential impact on B2 with higher arrivals as the delay hold time increases in value.

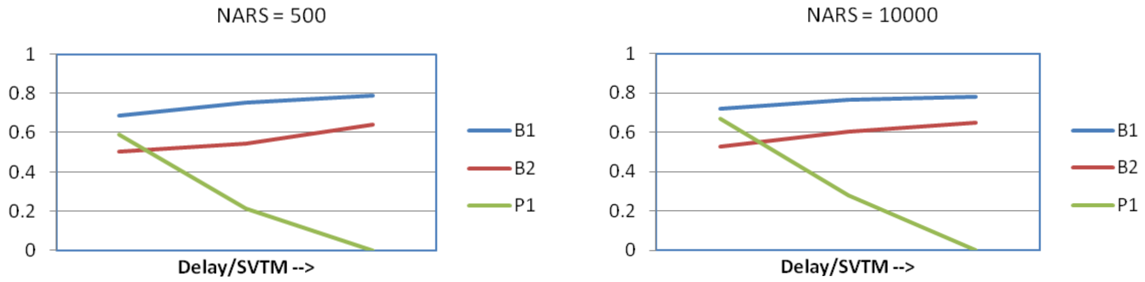

Figure 8: Delay based mechanism with NARS variance

CHAPTER 5

CONCLUSION

The motivation of this work comes from the need to support the general public and the NS/EP users at times of serious disasters when networks are severely stressed and congested. This work serves as an extension to [2] to provide a view into delay based mechanisms under the preemptive policies. A variety of preemption approaches were reviewed in this thesis leading up to the examination of the delay based soft preemption approach. We presented the algorithm for our delay based preemption model that was developed on the foundation of hard and soft preemption models.

We created a queuing modeling simulated system for hard, soft, partially-soft and delay-based preemption systems and provided comparative analysis of these varieties of preemption policies. It was demonstrated that soft preemption, relative to hard preemption, adversely affected the emergency users by blocking more high priority calls without any reasonable contribution to the preemption probability of class 1 calls. Partial soft preemption was also simulated and the results verified decent reduction in probability of preemption for class 1 calls but at some expense of the emergency users.

The simulation results of our delay-based soft preemptive approach demonstrated that increasing the delay time also increases the number of sessions blocked for emergency users as well as the general public; however, it reduces the probability of preemption and the number of class 1 calls affected.

It is essential though, to keep the impact on emergency users in perspective since preemption would not be allowed if NS/EP users incur inconvenience or increased challenges in responding to the disastrous situations. In order to benefit from the delay based mechanism, a balance would need to be determined about an acceptable delay value. In addition to the increased blocked calls, emergency users are also put on hold for the duration of time that may be a significant inconvenience at the time of disaster response or recovery. For instance, a 20 second hold time could result in a significant reduction in P1 from 67% to 50% while increasing the blocking probability for emergency users by merely 4 percent (52.7% to 56.7%).

This thesis work provides another preemptive approach to support emergency traffic and low priority users, during the times when networks are severely congested. The decision to adopt any of the preemption policies must be based on the urgency of disaster response and recovery but should also consider all the available options to ensure full utilization of the available resources.

REFERENCES

1. Beard, C.C. Dynamic agent-based prioritized resource allocation for stressed networks. Ph.D. Dissertation, University of Kansas, Lawrence, 1999.

2. Beard, C.C. Preemptive and delay-based mechanisms to provide preference to emergency traffic. *Computer Networks Journal* 47, 6 (April 2005), 801-824.

3. Brockmeyer, E., Halstrom, H.L., and Jensen, A. *The Life and Works of A.K. Erlang*. Academy of Technical Sciences, Copenhagen, 1948.

4. Available from http://en.wikipedia.org/wiki/3g.

5. Choudhury, G.L., Leung, K.K., and Whitt, W. An algorithm to compute blocking probabilities in multi-rate multi-class multi-resource loss models. *Advances in Applied Probability 27* 4 (Dec.1995), 1104-1143.

6. Geier, J. *Wireless LANs: Implementing High Performance IEEE 802.11 Networks* (2nd Edition), SAMS Publishing, 2001.

7. Available from http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats#subscribers.

8. Micrologic Research. *Overview of Cellular Market,* 2002. Available from http://www.mlr.biz/Cell_Exec.pdf.

9. Rappaport, T.S. *Wireless Communications, Principles and Practice* (2nd edition). Prentice Hall, Upper Saddle River, NJ, 2002.

10. Rappaport, T.S., Annamalai, A., Buehrer, R.M., and Tranter, W.H. Wireless Communications: Past Events and a Future Perspective. *IEEE Communications Magazine* 40, 5 (May 2002), 148-161.

11. Wheeler, R. Mobile commerce continues to grow. Available from http://www.qas.com/company/data-quality-news/mobile_commerce_continues_to_grow_8188.htm.

VITA

Aashish Chandra was born on June 16th 1978 in Roorkee, India. He completed his high school at St. Gabriel's Academy in 1995. With his keen interest in networking and telecommunications he pursued Bachelor's degree in Electronics and Communications Engineering from University of Madras, India in 1999.

After working for a year as an R&D Engineer in Telecommunications and Networking department of DSQ Software Limited he began his master's program at University of Missouri-Kansas City. While pursuing his master's degree he worked as a Graduate Research Assistant at University of Missouri, assisting with microprocessor architectures and disaster communications. Upon completion of his degree requirements, Mr. Chandra plans to continue his career as a business technology executive.