Design and Analysis of Multi-Level Active Queue Management Mechanisms for Emergency Traffic

Manali Joshi, Ajay Mansata, Salil Talauliker, Cory Beard

School of Computing and Engineering
University of Missouri-Kansas City
Kansas City, MO 64110 USA

***Abstract*—Multiple Average-Multiple Threshold (MAMT) Active Queue Management (AQM) is proposed as a solution for providing available and dependable service to traffic from emergency users after disasters. MAMT is a simple but effective approach that can be applied at strategic network locations where heavy congestion is anticipated. It can provide low loss to emergency packets while dropping non-emergency packets only as much as necessary. Fluid flow analysis and simulation is conducted to provide guidelines for proper MAMT design, especially regarding the queue size and averaging parameters that are most important. This work considers non-responsive traffic exclusively, since non-responsive traffic types are currently getting the most attention from emergency management organizations. Plus, very little work has been performed regarding AQM and non-responsive traffic. It demonstrates queue oscillation problems that previously may have been attributed to the interactions between TCP and AQM, but which are actually inherent to AQM and can be greatly reduced with proper parameter settings. MAMT is shown to perform well over a range of loads and can effectively protect emergency traffic from surges in non-emergency traffic.**

***Keywords*—System design, simulations, active queue management, emergency services, quality of service.**

I. Introduction

Packet networks, the most notable example being the Internet, have shown themselves to be tremendously useful to society. Some of the most useful applications have been the World Wide Web, electronic mail, and electronic commerce. These networks are even being extended to take over the functions of traditional public switched telephone networks to carry voice and video.

By and large, however, the main uses of the Internet have to be qualified as non-mission critical types of activities. Users are expected to understand that the reliability of the services that are provided to them is somewhat susceptible to congestion and other network configuration anomalies. While the fundamental design of the Internet is amazingly robust to work around and adapt to these types of problems, experiences of virtually every user include times where they have been unable to conduct activities they wished to perform. Part of the problem is that sometimes users do not appreciate the impact of network failures, high levels of congestion, or even misconfiguration of their own end devices. Regardless of this fact, however, most users believe they cannot trust Internet-based services the way they can trust the telephone system.

A. Emergency Users

The goal of this work is to improve the robustness of the Internet to be able to support activities that are of great value to society and are of a highly critical nature. The user type of particular interest in this paper is the National Security/Emergency Preparedness (NS/EP) user. NS/EP users conduct operations to save lives, restore the community infrastructure, and return the population to normal living conditions after serious disasters, which include floods, earthquakes, hurricanes, and terrorist attacks.

Recent terrorist attacks in the United States on September 11, 2001, have given special urgency to the development of network services for these users in the Internet. While many services and applications are already widely used, the special requirement of emergency users is that they be provided with more availability and dependability than is currently provided today, especially during the difficult operating conditions caused by disasters.

Recent events have shown what is common with disaster response – that tremendous stress is placed on networks in the aftermath of a disaster. In recent history, most stress was on wireless networks and the public switched telephone network (PSTN). The stress has come both from damaged facilities and network demand up to 400% of normal [1]. Aside from isolated problems with web sites of news organizations, the Internet has performed admirably after events such as September 11 [2]. But it is anticipated that as use of voice and video services increases over the Internet the same problems experienced by telephone and wireless networks will increasingly be seen on the public Internet. Standards bodies are working on requirements documents and solutions to address this issue [3] [4] [5].

Emergency response organizations are currently focusing on improving reliability for voice and multimedia applications over the public Internet [4]. Even though dedicated or special government telecommunications resources can be used in disaster response, they do not generally have the immediate wide

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

accessibility to be available in the critical first stages of disaster events [4].

B. Application of AQM to Emergency Services

This paper provides preferential treatment to emergency users by providing lower packet loss using strategically placed packet marking and Active Queue Management (AQM) functions. Placement of packet marking and AQM is recommended at particular places where congestion can occur. Most notably those places could be at bottleneck links between ISP's (where traffic contracts may be in place to limit traffic), in regional networks (e.g., lower tier ISP's, corporate networks, or non-profit regional educational networks) or in some types of access networks (e.g., DSL, cable modem or wireless).

The implementation of AQM for emergency purposes seeks to drop the packets of non-emergency traffic as much as needed to allow as much emergency traffic as possible to proceed through the router. The goal is to provide lower packet loss for emergency traffic so that emergency communications can proceed more dependably. It should be noted, however, that emergency traffic in general does not need to have better delay or jitter performance than normal traffic. Emergency traffic is no different than non-emergency traffic in that sense. The use of AQM, therefore, is meant to create lower packet loss for emergency traffic by acting as a filtering mechanism to perform prioritized packet discarding. The filtering also seeks, however, to provide the best service possible to non-emergency traffic by not unnecessarily favoring emergency traffic.

C. Scope of Work

As described through the paper, Multiple Average-Multiple Threshold (MAMT) Active Queue Management is proposed, justified, and then evaluated for appropriate parameter settings to ensure proper dropping priorities then good delay and jitter performance. A fourth drop precedence is used as part of the Diffserv Assured Forwarding Per Hop Behavior. Queues that receive exclusively UDP traffic are studied, from either constant or variable rate (exponential ON-OFF) sources. MAMT is implemented using coupled queues, non-overlapping curves, and linear dropping functions. Fluid-flow analytical and ns-2 simulation models are both used to evaluate performance and provide design guidelines.

D. Contributions of this Work

Four contributions are provided by this work. The first contribution is to apply AQM as a solution to the special requirements of emergency traffic. These requirements are discussed in detail in Section II. Secondly, a Multiple Average-Multiple Threshold (MAMT) approach is proposed here to meet those requirements. MAMT is a form of AQM that sets up dropping functions independently for each class of traffic as a function of the average number of packets in the queue for that class and those of higher priority (called a "coupled" approach to MAMT) [6]. It is an extension of RIO (RED with in/out bit) described in [7] to more than 2 classes. It protects emergency traffic by isolating its dropping probabilities from non-emergency traffic.

The third contribution of this work is an in-depth study of AQM as applied to non-adaptive (UDP-type) traffic. Non-adaptive traffic is the current focus of emergency management organizations, due to the synchronous nature of most emergency communications immediately after a disaster. It is most important to emergency organizations that voice traffic be supported over the Internet [3] [4], even though, certainly, adaptive traffic (i.e., TCP) should also be studied and is being studied as a follow-on to this work.

In addition, virtually all of the attention of previous AQM work has been to consider the dynamics of AQM with TCP traffic, and in some cases the interaction when both adaptive (i.e., TCP) and non-adaptive traffic mix together [8]. But it is proposed here that packet markings be implemented such that non-adaptive traffic does not mix with TCP traffic (discussed fully in Section III). Therefore, one must understand how AQM behaves with exclusively non-responsive traffic, which has received only minimal attention [9]. And unexpected results have been obtained that show that strong queue fill oscillations can occur in the absence of TCP traffic (even with constant rate traffic) if queue fill averaging parameters and dropping functions are not chosen properly.

The fourth contribution of this work is to show both through fluid flow analysis and simulation that it is possible to establish dropping functions so that AQM can perform well over a range of loads. We consider averaging functions, dropping functions, traffic burstiness, and minimum queue size. Tradeoffs must be made between average queue fill (and corresponding average packet delay through the queue) versus packet dropping probabilities.

The next sections provide a definition of emergency traffic requirements and the proposed approach to meeting those requirements. Then a fluid flow analytical investigation is provided using Poisson Counter Driven Stochastic Differential Equations, followed by further evaluation of the MAMT AQM scheme using simulations. The result is a set of design guidelines for an effective implementation of AQM for emergency traffic.

II. Emergency Service Requirements

There are three-high level requirements for emergency communications to be supported at the network layer in the public Internet-based infrastructure.

- Availability – Authorized emergency-related users must have a high likelihood of network resources being available to them when they need them.
- Dependability – Emergency users must not only be able to initiate communication sessions; they must also be able to successfully complete their intended activities.
- Security Protection – Emergency traffic needs to be protected against intrusion, spoofing, and specifically, denial of service (DoS). This is beyond the scope of this work but is discussed in detail in [10]. It is assumed that work conducted elsewhere will solve the critical security issues related to emergency traffic. User identification codes are used successfully in PSTN-based emergency services today [11], but public keys, private session keys, certificate authorities, and digital signatures may be employed for Internet-based services. The primary requirement is to protect network devices that are allowed to mark packets with emergency markings. The vulnerability of packet marking devices (to DoS attacks especially) is important to critically assess when using any packet-marking quality of service (QoS) mechanism, and is especially important here.

Being able to meet the above availability and dependability requirements is vital for emergency organizations. Otherwise, they will not be able to adapt and apply new disaster response procedures. Such operations are to some degree dynamic and adaptive to the current situation, but more importantly they are highly coordinated, structured, and hierarchically controlled, if for no other reason than to provide safety to the workers. A communications capability, if it is part of normal operating procedures, must be dependable; otherwise, time and energy gets wasted in a response effort due to frustrations or inability to communicate. The inability of the Internet to address this dependence on reliability has limited its use for emergency operations [12].

It is also illustrative to mention requirements that are *not* applicable to emergency services. Emergency services do not have delay and jitter expectations that are any different from those of normal services. The focus is on reliability, not so much on better packet-level delay performance. Voice applications for emergency workers do not need better sound quality than those for other users. This is important because a service deployment should avoid providing unnecessarily good service to emergency traffic. Thus the design objectives are to first to provide acceptable quality to emergency services with reliability that is reasonably immune to network conditions, then to provide the best service possible to non-emergency services.

This work implements preferential treatment on a packet-by-packet basis, instead of using flow level reservation and preemption mechanisms that have been extensively investigated for emergency traffic [13] [14]. A packet-based QoS mechanism uses AQM to provide dependable QoS by dropping non-emergency packets to provide preference for emergency packets. It is simpler to implement than a flow-based approach, only needs to be applied at a few routers along a path (in contrast to a flow-based approach like Intserv which requires all routers on the path to participate), and does not require per-flow maintenance of state information.

III. AQM for Emergency Traffic

To provide preferential treatment using AQM, mechanisms are needed for marking packets and for dropping packets.

A. *AQM with Assured Forwarding*

Marking of emergency packets can be done with a variety of mechanisms at both network and application layers. Different approaches have different merits, and a comprehensive study of how to mark emergency packets is beyond the scope of this work. But difficult problems must be solved related to who marks packets (what end users, what organizations), what devices mark packets, who polices, and how service providers might be monetarily compensated, especially given the decentralized nature of the Internet.

Currently, the most fully developed packet marking approaches are those related to IP Differentiated Services Per Hop Behaviors (PHB's). So this work assumes the use of Assured Forwarding (AF) codepoints [15] to mark emergency and non-emergency packets. Expedited Forwarding (EF) is not considered because it does not provide a preferential treatment capability, other than preferential treatment that could be applied to limit the number of packets marked with EF. EF also would likely provide unnecessarily good service to emergency packets by over-allocating resources to EF traffic at the expense of other traffic.

It is most reasonable to mark AF classes in such a way that the traffic type is homogeneous within a particular AF class. For example, one might allocate AF classes as follows.

- AF1 – Interactive voice.
- AF2 – Interactive video.
- AF3 – Bulk data transfers.
- AF4 – Transactions (instant messaging, database

access, interactive applications).

After that, the priorities of the different traffic flows (and packets within the flows) can be marked using the three AF drop precedences within each class.

One might consider whether emergency services would need one or more of their own per-hop behaviors. The main reason why one might propose a new PHB, however, would be to provide better delay or jitter performance, since in general each PHB receives a separate queue and bandwidth allocation at the scheduler. But as stated already, emergency services do not need any different delay or jitter performance than normal traffic, so no new PHB would be needed. Then statistical multiplexing within a PHB can be exploited for the benefit of both the emergency and non-emergency traffic. Emergency traffic can mix in a queue that has a large bandwidth allocation (instead of using a separate emergency queue that would likely have a smaller bandwidth allocation), and non-emergency traffic can use the capacity most of the time when emergency traffic is not present. By this reasoning, a fifth AF class is not needed.

One might also consider, however, having more drop precedences within an AF class, since it is beneficial to provide lower dropping probabilities to emergency packets. One of the three existing AF drop precedences could be used exclusively for emergency traffic, but then non-emergency traffic would only be left with the remaining two. But if one did not want to lose the three-level capability for non-emergency traffic, a fourth drop precedence could be implemented. Such a decision could be made specific to a particular service provider's discretion.

The focus of this work, then, is on AF. A scheduling mechanism such as Weighted Fair Queueing is used to set a minimum bound on the amount of bandwidth used by each AF class. Then AQM is used within the class to provide preferential packet dropping.

The AF specification only requires that the dropping probability of one drop precedence be less than or equal to that of a higher drop precedence [15]. But for our purposes for non-adaptive emergency traffic, a stricter requirement is used that says that the dropping probability of one priority level should be nearly equal to 100% before the dropping probability of the higher priority level goes above a small dropping level (above, say, 0.1% or 1%). This is because it is common that voice and other multimedia flows (from both emergency and non-emergency users) are marked by users to have certain packets within a flow that are marked with lower priority than others, such as packets that provide enhanced video resolution. There is little justification in the multimedia context for keeping lower priority packets at the expense of dropping some higher priority ones. Also, if flows as a whole are at different priority levels, packets can be marked appropriately so that the most important packets of lower importance flows are treated properly relative to lower priority packets of more important flows.

This means that the most important packets of emergency traffic should be dropped at very low levels (e.g., at a fixed rate of 0.1%) until traffic from all of the other classes is dropped. For the purpose of the discussions here, four drop precedences will be used and the convention will be followed that the lowest priority packets are given the color red, and then successively high priorities are given colors yellow and green. All emergency traffic is given the color blue.

This approach meets emergency service requirements as follows. For the availability requirement, emergency traffic is allowed to mix with non-emergency traffic within an AF queue. To meet the dependability requirement, once emergency traffic mixes with non-emergency traffic AQM is used to drop non-emergency traffic as much as needed.

B. AQM Design Requirements

An AQM design must provide low dropping to emergency packets and then as low of an average delay and delay variation as possible. A particular design must work effectively over the range of loads which could be expected. This range of loads can be quite large when considering what might happen after a disaster.

AQM performance is affected by the following four considerations.

- Average queue fill – Since all packets within an AF class share the same queue, all packets have the same average delay. This is directly proportional to the average queue fill. To have low average delay, average queue fill must be kept low.
- Tail drops – When a queue is full, all incoming packets are dropped, which is known as tail dropping. Tail drops must be avoided because emergency packets are then dropped with no preference over non-emergency packets. One way to address this is to have large queues; another is to use AQM to keep the queue from becoming full.
- Queue fill variations and oscillations – Variations in queue fill can cause tail drops if severe enough. Variations can also cause wide delay jitter. An important goal is to limit queue fill variations, but also allow some variation if the input traffic is bursty.
- Averaging – An averaged queue fill, instead of an instantaneous measurement, should be used as an input parameter to functions to decide whether to drop packets. The standard computation that is used to find the average is a weighted sum of the current

average and the current instantaneous queue fill. The weighting parameter is shown later to have a significant impact on MAMT performance.

C. Relationship with Related Work

When active queue management was first developed, an important early proposal was for Random Early Detection (RED) [16]. RED was designed more for congestion avoidance and fairness for TCP traffic than as a filtering mechanism for non-adaptive traffic, but is widely used and is useful to be applied here. Newly arriving packets (and only newly arriving packets) are dropped if the queue is too full. Figure 1 illustrates that the RED dropping function is a linear function of the average queue fill. Below min_{th}, no packets are dropped; above max_{th} all packets are dropped, and dropping is a linear function in between the thresholds based on a probability of max_p just before max_{th}. A gentler version of RED has also been proposed that more gradually increases to 1.0 above max_{th} [17]. The average queue fill is computed as an exponentially weighted moving average based on a parameter w_q as follows.

$$q_{avg}(n) = (1 - w_q)\, q_{avg}(n-1) + w_q q(n) \qquad (1)$$

Since RED was proposed, a large body of work has been performed to understand RED and its interaction with TCP. Limitations of RED in this context have been well documented, even when non-TCP traffic is also present [18].

Many variations to RED have been proposed. One type of variation uses per-flow information so that the history of dropping for the flow can be taken into account [19]. Another type of variation uses reward and penalty functions and control theory to stabilize the queue fill at a target value [20] [21]. And another type uses adaptation mechanisms to continually adjust the RED parameters shown in Figure 1, especially max_p [22].

To extend RED capabilities to multiple classes of traffic, RIO (RED with in/out) was proposed in [7]. RIO has two classes of traffic (call them green and red) and maintains an average of the number of packets in the queue for each class. A separate dropping function is used for each class, and each function has the same form as in Figure 1. For the red traffic, the dropping function is with respect to the average number of red packets plus the average number of green packets. This is called a coupled approach and has been abbreviated RIO-C. But for the green dropping function, only the average number of green packets is used. This means that green packets will only be dropped if there are a large number of green packets; the number of red packets is not taken into account.

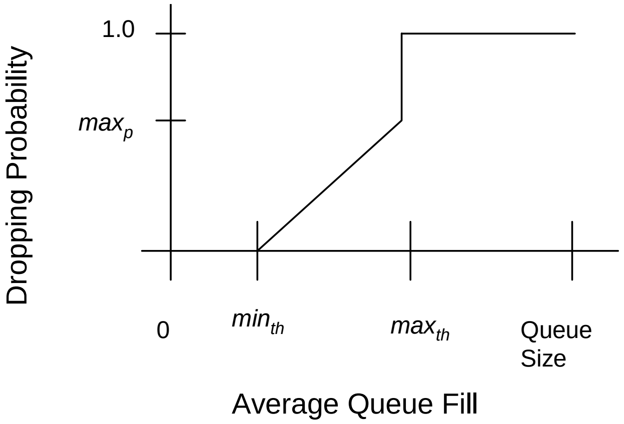

Figure 1 – RED Dropping Function

This approach can be extended to more classes in the same manner. This approach is called Multiple Average-Multiple Threshold (MAMT) [6] with coupled queues, since multiple queue averages are used and multiple dropping curves are used. A variation of this approach uses only one queue average, the overall queue fill average (a Single Average-Multiple Threshold approach), and is called Weighted RED (WRED); this is implemented in Cisco routers. WRED cannot effectively protect emergency packets as required, however, because a large burst of low priority packets could cause emergency packets to be dropped.

The approach here is to use MAMT with linear dropping curves like those used with RIO. Flow-based, control theoretic, or adaptive approaches are not considered because these seem more complex than necessary. One reason why emergency services have not been deployed is because the mechanisms that have been proposed are unduly complex; an MAMT approach only involves computing average queue fills and dropping according to a simple function. It is simple and effective.

This is especially appropriate because only non-responsive UDP types of traffic are being considered. Some work has already been performed for TCP traffic with two-color marking [23], but not for UDP traffic. More importantly, however, the goal here is to understand multi-level AQM applied to emergency traffic without TCP response issues influencing the analysis. And effectively using AQM for UDP traffic is a non-trivial exercise. Only one other paper has exclusively considered multi-level AQM with non-responsive traffic, but it uses a more complicated approach than the one here and considers more parameters than just average queue occupancy [9].

D. Complexity

Although MAMT RED here is a simple approach, there are two complexity considerations that must be addressed – where AQM should be placed and when it should be active. AQM will add processing load and processing delay at a router to classify packets, compute dropping probabilities, and drop packets, so it is best to only implement AQM where needed. As stated in Section I.B, AQM is proposed to be deployed only where congestion is to be expected. Since many service providers over-provision their core networks, AQM would not be required in core routers, which is advantageous because AQM would cause the most impact at those core routers. Instead, AQM should be implemented in some lower-volume regional networks, and some types of access networks. The traffic volume at those locations should not cause significant processing delays.

In addition, AQM need not be continuously active. Routers need not classify packets until congestion levels warrant. With coupled MAMT, the lowest priority packets are dropped according to the average number of packets for their class and all other classes, which is the total of all packets. An average queue fill can be computed based on total packets, and then once this value exceeds the min_{th} for the lowest priority class, only then would AQM be activated. So the only function that would be continuously active would be a computation of average queue fill without regard to packet markings.

IV. Analytical Model

The goal for the remainder of this paper is to understand how MAMT AQM can be implemented to provide low packet loss to emergency traffic over a range of network loads. Then as a secondary goal, it is desirable to provide low delay and low delay variation.

As a first step, analytical expressions are derived for AQM performance for two simplified RED models. They provide insight into some of the key characteristics and parameters that affect AQM behavior for non-responsive traffic. After that, a comprehensive AQM simulation model is used to fully study the MAMT approach just discussed.

A. Constant Rate Source

The first simplified AQM model to be considered is single-class RED with a constant rate source as its input. There is one class of traffic and one dropping function with min_{th}=0, and max_{th}=queuesize. The parameter $L_{red}=max_p/max_{th}$ is defined as the slope of the dropping curve. Then the probability of dropping for AQM will be $p_{drop}(t)=L_{red}\, q_{avg}(t)$ because min_{th}=0.

From there a set of two differential equations can be defined, based on the stochastic differential equation fluid-flow models in [8] [24] [25], as follows.

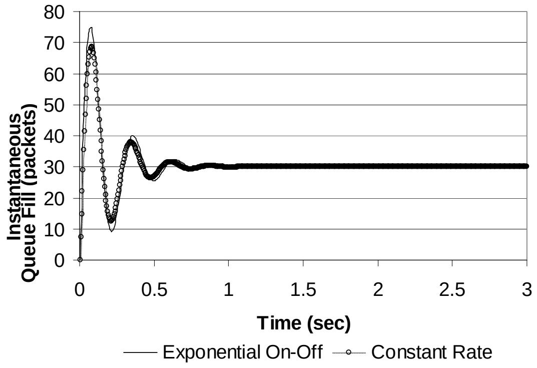

Figure 2 – Analytical Results for Constant and Variable Rate Traffic for w_q=0.003

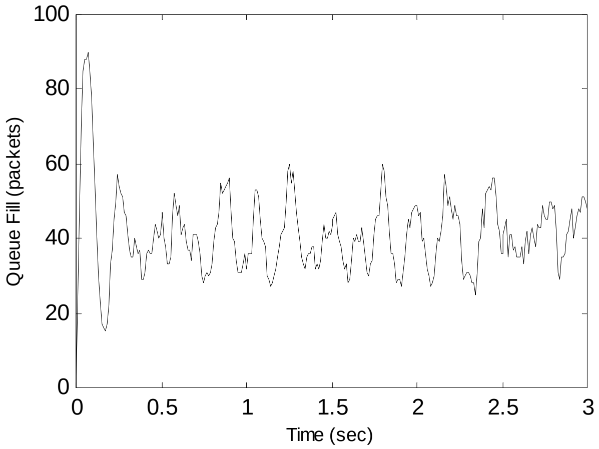

Figure 3 – Actual Simulation Results for Constant Rate Traffic and for w_q=0.003

$$\begin{aligned} d(v(t)) &= -c\mathbf{1}_{(v(t)>0)}\,dt \\ &\quad + x_0(1-L_{red}\,q_{avg}(t))\mathbf{1}_{(v(t)<v_{\max})}\,dt \\ d(q_{avg}(t)) &= -K_{red}\,q_{avg}(t)dt + K_{red}\,v(t)dt \end{aligned} \tag{2}$$

The indicator function is denoted as $\mathbf{1}_{\{v(t)>0\}}$ which equals 1 when the condition is met. The instantaneous queue fill is $v(t)$, v_{max} is the queue size, x_0 is the input traffic rate, and c is the fixed output rate of the queue. Queue fill, $v(t)$, drains at a rate c and fills at the fixed rate x_0 thinned by $p_{drop}(t)=L_{red}\, q_{avg}(t)$. K_{red} is used to convert the discrete time difference equation in (1) to a continuous time differential equation by the following [24].

$$\begin{aligned} K_{red} &= -\ln(1-w_q)/\delta \\ \delta &= \text{average time between updates of } q_{avg} \end{aligned} \tag{3}$$

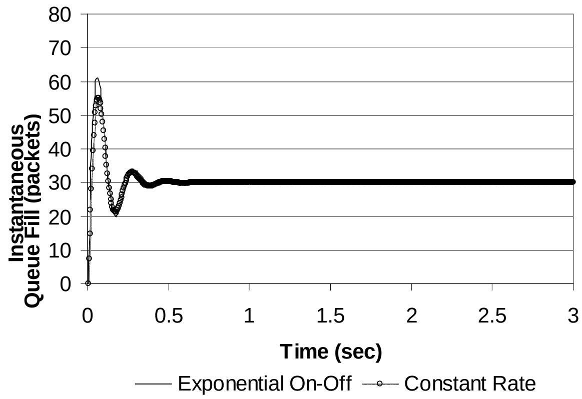

Figure 4 – Analytical Results for Constant and Variable Rate Traffic for w_q=0.005

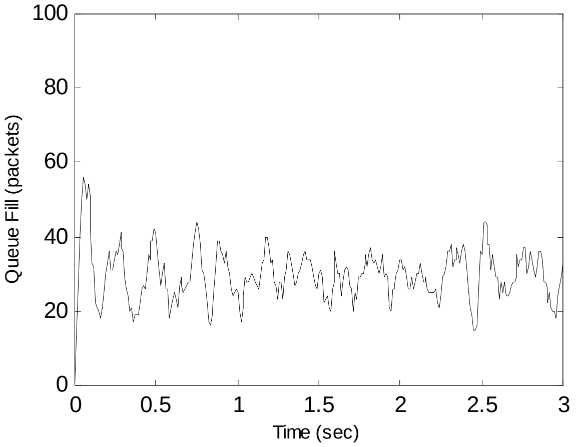

Figure 5 – Actual Simulation Results for Constant Rate Traffic and for w_q=0.005

If it is assumed that $0 < v(t) < v_{max}$ is always true, then a solution for $v(t)$ can readily be formed from

$$V(s) = \frac{(x_0 - c)(s + K_{red})}{s^2 + sK_{red} + x_0 L_{red} K_{red}}\ , \tag{4}$$

so that poles are located at

$$s_{1,2} = -\frac{K_{red}}{2} \pm \frac{1}{2}\sqrt{K_{red}^2 - 4K_{red} L_{red} x_0}\ . \tag{5}$$

Therefore, RED response to a constant rate input in many cases is a damped sinusoid, depending on K_{red}. It stabilizes at the value of q_{avg} (with the corresponding p_{drop}) that thins out just the right amount of traffic. The results for RED with fixed rate traffic are illustrated in Figure 2 by the dotted curve. Parameters are w_q = 0.003, x_0 = 20 Mbps, c = 10 Mbps, L_{red} = 1/80, and packet size = 4000 bits (fixed).

Figure 3 shows actual RED performance through simulation results. The initial transient matches that of the analytical result. However, noise in the system continues to drive the RED performance by a substantial amount after that transient period. This is not modeled well from these differential equations, which is why simulation models are used in the next section to study AQM in its fully functional form. Also, [26] shows that much of the noise in an AQM system can be eliminated by managing packet dropping over groups of packets instead of using the packet-by-packet approach that is commonly used with AQM.

But the result from (5) still has some usefulness. RED performance is affected as follows by different system parameters.

- K_{red} defines the damping of the sinusoid. If K_{red} is small, this results from a small value of w_q and results in less damping. Figure 4 shows the same results from Figure 2, but now with w_q = 0.005 (a larger value of K_{red}) and shows more damping. Changes to K_{red} may also increase or decrease the frequency of the sinusoid. As seen in Figure 3, the noise in the system drives an instantaneous queue fill variation that has a power spectral density with strong components around the same frequency as the oscillation frequency from the transient response from (3). K_{red} affects that frequency. Figure 5 shows the same as Figure 3 but with w_q = 0.005, and it can be seen that more damping and higher frequency of variation results from the increase in K_{red}.
- An increase in L_{red} causes an increase in the frequency of the sinusoid. An increase in the steepness of the curve (as seen in Figure 1) causes this increase in L_{red}.
- An increase in the rate of the input traffic (i.e., x_0) also causes an increase in the frequency of the sinusoid.

B. *Variable Rate Source*

The RED analytical model can also be extended to one that has an exponential on-off source as its input. ON and OFF times are exponentially distributed and the source transmits traffic at rate h when ON. Similar to [8], this source can be modeled using a Poisson Counter Driven Stochastic Differential Equation. The source, $x(t)$ is defined as $x(t) \in \{-1,+1\}$, which signifies OFF and ON states, and the Poisson counters N_1 and N_2 cause transitions between these two states at rates λ_1 and λ_2. No other changes are made to the previously analyzed system and the differential equations for this system are modified from (2) to become

$$\begin{aligned} d(x(t)) &= -(x(t)-1)dN_1 - (x(t)+1)dN_2 \\ d(v(t)) &= -c\mathbf{1}_{(v(t)>0)}\,dt \\ &\quad + \frac{h}{2}\big(x(t)+1\big)(1-L_{red}\,q_{avg}(t))\,\mathbf{1}_{(v(t)<v_{\max})}\,dt \\ d(q_{avg}(t)) &= -K_{red}\,q_{avg}(t)dt + K_{red}\,v(t)dt \end{aligned} \tag{6}$$

To solve the equation for the input source, first take expectations of both sides which gives

$$\frac{d}{dt}E[x(t)] = -(E[x(t)]-1)\lambda_1 - (E[x(t)]+1)\lambda_2 \tag{7}$$

Once again if $0 < v(t) < v_{max}$, the following are derived.

$$\begin{aligned} \frac{d}{dt}E[x(t)] &= -(E[x(t)]-1)\lambda_1 - (E[x(t)]+1)\lambda_2 \\ \frac{d}{dt}E[v(t)] &= -c + \frac{h}{2} + \frac{h}{2}E[x(t)] \\ &\quad - \frac{h}{2}L_{red}\,E[x(t)q_{avg}(t)] \\ &\quad - \frac{h}{2}L_{red}\,E[q_{avg}(t)] \\ \frac{d}{dt}E[q_{avg}(t)] &= -K_{red}\,E[q_{avg}(t)] - K_{red}\,E[v(t)] \\ \frac{d}{dt}E[x(t)q_{avg}(t)] &= E[x(t)q_{avg}(t)](-\lambda_1 - \lambda_2 - K_{red}) \\ &\quad + E[q_{avg}(t)](\lambda_1 - \lambda_2) \\ &\quad + E[x(t)v(t)]K_{red} \\ \frac{d}{dt}E[x(t)v(t)] &= E[x(t)v(t)](-\lambda_1 - \lambda_2) \\ &\quad + E[v(t)](\lambda_1 - \lambda_2) \\ &\quad + E[x(t)]\left(\frac{h}{2} - c\right) \\ &\quad + \frac{h}{2} - \frac{h}{2}E[q_{avg}(t)]L_{red} \\ &\quad - \frac{h}{2}E[x(t)q_{avg}(t)]L_{red} \end{aligned} \tag{8}$$

Figure 2 also shows the curve for an exponential ON/OFF source for the same RED system. The exponential ON/OFF source has the same average rate as the constant rate source. The curve for the exponential ON/OFF source is created from input parameters of λ_1 = 50 (rate from OFF to ON) and λ_2 = 12.5 transitions per second. The general observation from Figure 2 is that the burstiness of the ON/OFF source creates stronger oscillations than those of the constant rate source. But in terms of expected behavior, as modeled by these equations, the ON-OFF source produces close to the same time response, although with a little more overshoot.

To further study the performance of AQM beyond these two simple models toward understanding the MAMT model and its usefulness for emergency traffic, the remainder of this paper considers simulation results. To extend the equations in (6) to involve 4 classes of traffic, 4 averaging functions, and 4 dropping functions would be intractable. Plus, they would not capture the effects of noise as has been discussed.

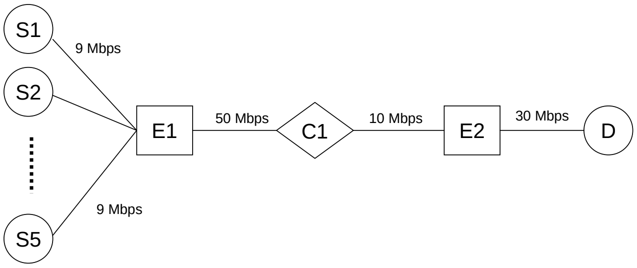

Figure 6 - Simulation Topology

Results will show, however, the same dependence on K_{red}, L_{red}, and input traffic rate that has been shown analytically. And it will be shown that the four-class MAMT system shows even more pronounced queue fill variations than for the single class case, because of interaction between the curves.

V. Simulation Results

This section presents the simulation results of the MAMT scheme implemented in the ns-2 simulator [27]. Figure 6 shows the simulation model consisting of five UDP sources S1 through S5 and a destination node D. The DiffServ domain consists of two edge devices E1, E2 and a core C1. C1 can be considered to be a node in a regional network that might experience congestion. Each source and edge device is connected by a 9 Mbps link. The link E1-C1 is 50 Mbps and the bottleneck link C1-E2 is 10 Mbps. Unless otherwise specified, the following default parameters are used.

- Fixed packet size of 500 bytes.
- Simulation time of 100 sec to observe steady state behavior.
- Size of each AF physical queue = 100 packets.
- Exponential on/off sources with average ON times of 4 ms and average OFF times of 1 ms.
- Queue averaging parameter w_q = 0.002 for all drop precedences.
- Average total load of 12 Mbps.

A Time Sliding Window (two rate, three color) meter/policer is used to mark packets for AF drop precedence levels. The following sub-sections present the findings.

A. *Base Results*

The first results given in Figure 7 show that MAMT is indeed effective at accomplishing its intended goals. As load increases, the lowest priority traffic (red, yellow, then green) is dropped up to 100% levels before the next class is dropped substantially. For this case, the balance at each load is 20% blue (emergency), 40% green, 20% yellow, and 20% red.

This shows that emergency traffic has zero dropping at all levels, and green traffic is only dropped at the higher loads. Also, it should be noted that there is some small overlapping at 12.5 Mbps, where yellow is dropped at 8.8% while red is still only at 98.6%. This small overlap also occurs at 17.5 Mbps. This is caused by AQM queue fill variations which are influenced by different factors as discussed below. Queue fill variations cause the average queue fill point to move back and forth between two curves.

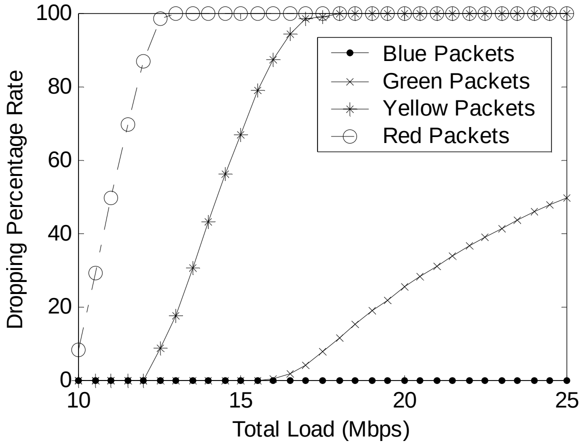

Figure 7 – Dropping Rates per Traffic Class for Various Loads

B. Effects of Queue Averaging Parameter w_q

In this section, the effect of the queue averaging parameter w_q on the queue fill is considered. In the MAMT scheme, for each arriving packet a decision is made whether to enqueue or drop the packet. This decision is directly based on the current average queue fill for the particular drop precedence (DP) and the threshold pair (min_{th}, max_{th}) for that DP. The queue weighting parameter w_q controls the level of averaging performed. If w_q is large, the AQM scheme becomes very responsive to each change in the instantaneous queue fill. On the other hand, if w_q is small then the AQM scheme has more consistent response to variations in traffic and behaves more strongly as a low pass filter. Thus w_q controls the tradeoff between robustness and responsiveness [8].

RED gateways are designed to keep the average queue size below a certain threshold [16]. This suggests that the w_q value should be higher so as to respond quickly to incoming traffic. At the same time, the queue should be capable of handling some amount of burstiness. Hence the w_q value should not be too high. In [16], w_q = 0.002 was suggested as an optimal value for the averaging parameter to achieve both the above-mentioned characteristics for TCP traffic. However, these simulation results for non-responsive traffic show that w_q could be chosen from a range of values. The choice of this parameter would be somewhat dependent on the traffic characteristics expected.

Table 1 shows how changes to w_q affect the mean queue fill and variance of instantaneous queue fill. It is seen that the mean queue fill as well as the variance decreases with increasing values of w_q. Below the suggested optimal value of 0.002, variance in the instantaneous queue fill is very high, indicating that the averaging function is responding very slowly to the changes in the actual queue size and substantial queue fluctuations are occurring before AQM can respond. High variance is undesirable because it causes large delay variation and may also cause tail drops. It is also seen that above $w_q = 0.01$, the variances of the instantaneous queue fill are very similar. So, for this set of traffic characteristics, a w_q value in the range (0.002, 0.1) should be chosen. The same approach that was used here (analyzing queue fill mean and variance) could be used for a different traffic characteristic, and it is anticipated that a w_q value in the range (0.002, 0.1) would also be effective there.

Note that these results were obtained from variable rate traffic, but it was also observed that even with constant rate traffic high variances can occur in the queue fill if w_q is not set properly.

Table 1 – Queue averaging versus queue fill

w_q	Mean Queue Fill (packets)	Variance of Queue Fill
0.00005	14	510
0.0002	12	235
0.0005	12	150
0.002	12	89
0.007	13	67
0.01	13	64
0.06	14	54
0.2	14	53
0.6	13	51
1.0	14	53

C. Effects of Queue Size

In this section the effect of the queue limit on the throughput and the oscillations in the queue fill is considered. Two cases are studied; in the first the queue size is 50 packets and in the second the queue size is 100 packets. For each queue size, (min_{th}, max_{th}) values are scaled according to queue size.

Simulations for each case were conducted where the average load was increased from 1 Mbps to 20 Mbps in equal steps. Figure 8 shows the line drops (i.e., tail drops) observed in each case.

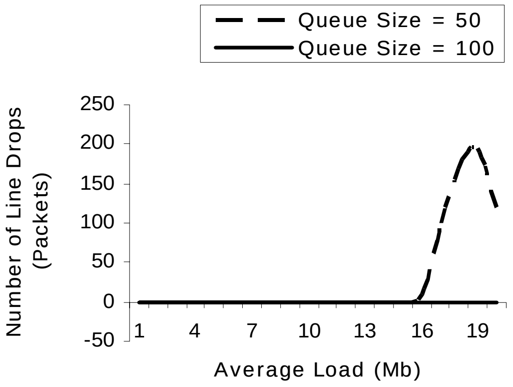

Figure 8 – Line Drops versus Queue Size

As the load increases, a large number of line drops occur for a queue size of 50 as compared to a queue size of 100. The presence of line drops is unacceptable because it causes emergency packets to be dropped indiscriminately from non-emergency packets. Note the size of 100, which contrasts with the default queue size of 40 packets in Cisco line interface cards [18].

When link capacity and traffic characteristics remain constant (only load increases), there is a direct relationship between the throughput and the required queue size. Moreover, if the queue size is not above a minimum value, then AQM cannot prevent line drops from occurring.

But if the queue size is sufficient, a wide range of loads can be supported. This is in contrast to some findings on AQM that have said that AQM performance is very sensitive to appropriate parameter settings [18]. The way to reconcile these two perspectives may be to say that AQM is not overly sensitive when applied to non-responsive traffic as seen here, but is more sensitive for TCP traffic.

Another important observation is that the average delay experienced by a packet roughly doubles as the queue size doubles. This is because the scaled (min_{th}, max_{th}) parameters cause larger average queue fill. Thus there is an important tradeoff between throughput and delay for a given queue size for a load with a particular burstiness. And it is best to have a queue size as low as possible.

D. Effects of the Characteristics of Input Traffic

This section studies the effects of the burstiness of the traffic on the queue fill. Traffic that is specified with a greater average ON time is considered to be more bursty as compared to traffic with a smaller average ON time. For all of the simulations, the ratio of ON time to OFF time is kept constant at 4:1. Figures 9 and 10 show the queue fill plots for ON times of 4 ms and 64 ms respectively. It is observed that the increase in the burstiness causes stronger oscillations in the queue fill. This is consistent with what was seen in the analysis results shown in Figure 2 and is even more pronounced here.

For the next simulations, the number of sources was fixed and the queue fill was observed as the ON time is increased from 1 ms to 64 ms.

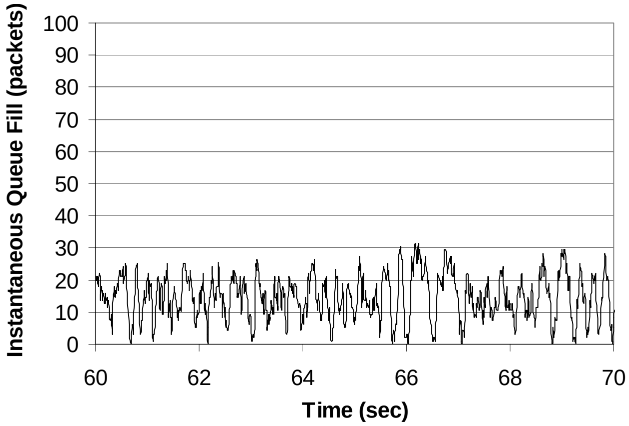

Figure 9 – Queue fill with ON time = 4 ms

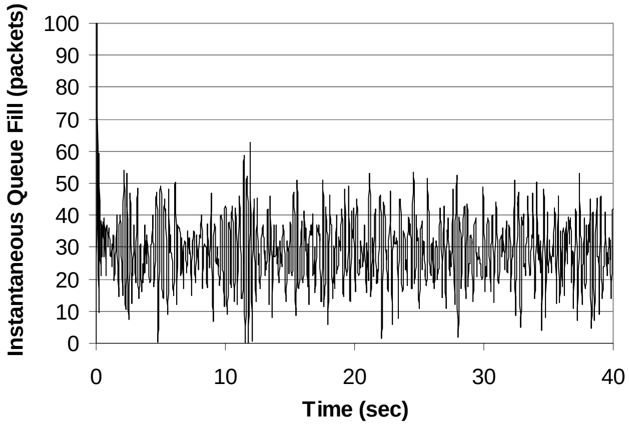

Figure 10 –Queue fill with ON time = 64 ms

Figure 11 shows that as the burstiness increases the variance in the queue fill increases. After a certain point any further increase in burstiness results in line drops as shown in Table 2.

This increase in variance can be explained as follows. As the burstiness increases, it is more likely that sources will all be ON at the same time, and can fill the queue more quickly up to its limit. The same is true during OFF times, which causes the queue to drain rapidly. This process of rapid filling and draining leads to large variance in the queue fill.

Thus for a given queue size and traffic load, increasing levels of burstiness cause increased oscillations in the queue fill which eventually leads to line drops.

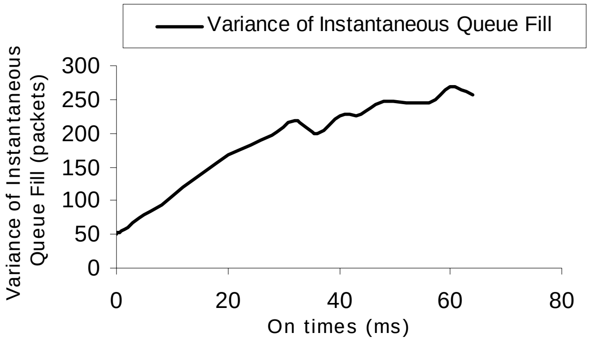

Figure 11 – Burstiness versus Variance in Queue Fill

Table 2 – Line Drops v/s Burstiness

Constant load of 12 Mbps			
No.	ON Time (ms)	OFF Time (ms)	Line Drops
1	2	0.5	0
2	4	1	0
3	8	2	0
4	16	4	0
5	32	8	0
6	64	16	273

E. Effects of Spacing between Dropping Curves

In this section, the effect on the queue fill from the spacing between the (min_{th}, max_{th}) pairs for individual traffic classes is studied. The spacing between the min_{th} and max_{th} of MAMT curves of particular traffic classes along with the maximum dropping probability max_p for that class define the dropping curve for that class. The effect of max_p is considered in the next section. Here we are interested in the effect of spacing between the dropping curves for each class on the overall throughput of the higher priority class and on the oscillations in the queue fill.

Two scenarios were investigated, the first having staggered curves with small inter-curve spacing (5 packets) and the second with wider inter-curve spacing (15 packets). Figure 12 shows the queue fill plot for the first case and Figure 13 shows the plot for the second case.

In the first case, the oscillations in the queue fill are limited to a small range. In the second case the oscillations are spread over a larger range. However, statistical results indicate that the first case causes excessive drops of higher priority packets as compared to the latter case. The reason for excessive drops is that when the curves are closely spaced, a considerable number of higher priority packets are dropped before dropping all of the lower priority packets. Some switching occurs between the active dropping functions for each of those two classes, causing higher priority packets to be dropped where their dropping function is active. In the second scenario, none of the higher priority packets are dropped before dropping all of the lower priority packets.

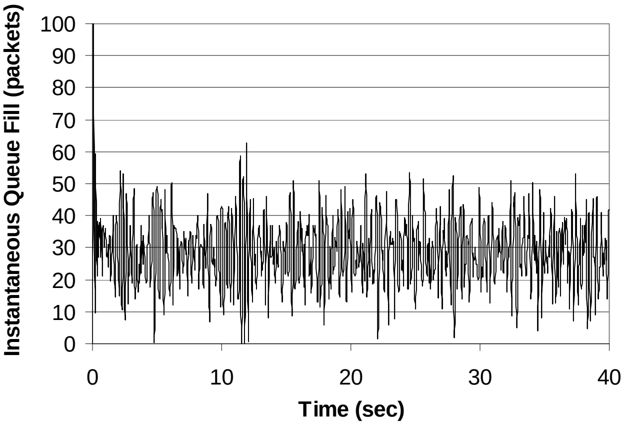

Figure 12 – Effect of closely spaced dropping curves

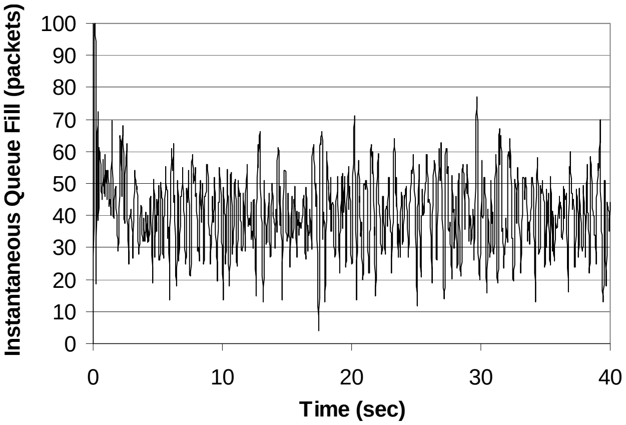

Figure 13 – Effect of widely spaced dropping curves

The first case leads to a lower overall throughput for the higher priority class. In the second case, the packet dropping is much more controlled, providing a better overall throughput for the higher priority class, because the curves are sufficiently spaced to not interact. Also, the closely spaced setting provides slightly lower average delay as compared to the other case since average queue fill is lower. Thus the spacing between the dropping curves presents a tradeoff between delay and throughput for higher priority traffic.

F. Effects of max_p

This section examines the effect of the maximum dropping probability, max_p, on the oscillations in the queue fill. A set of simulations were first performed for constant rate traffic and max_p was varied from 0 through 1 in equal steps to find its effect on the oscillations in the instantaneous queue fill values. Each simulation produced a result very similar to the other and demonstrated that for constant rate traffic, max_p has no noticeable effect on the queue fill.

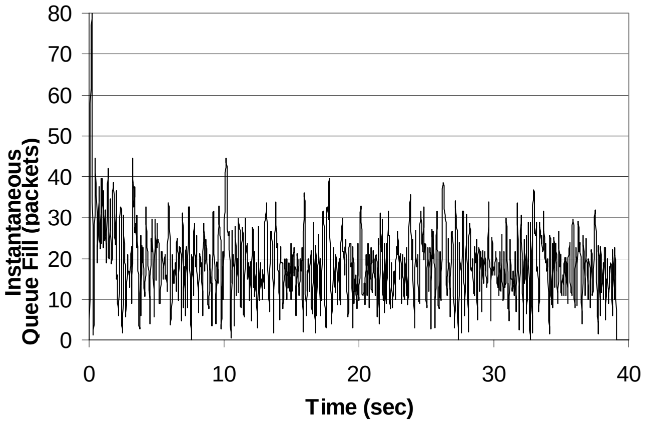

Figure 14 – Effect of max_p=0 on VBR queue fill

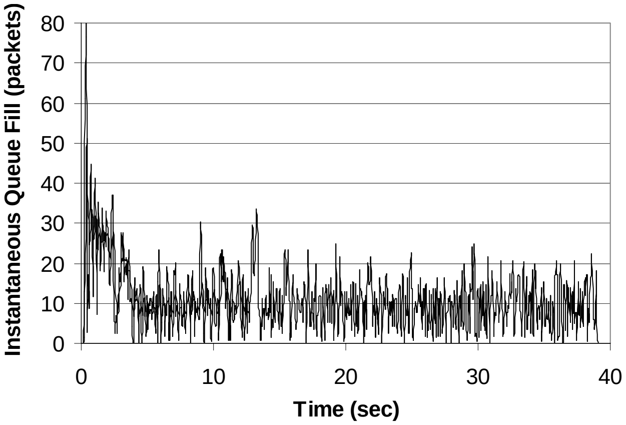

Figure 15 – Effect of max_p=1 on VBR queue fill

The same settings were repeated for variable rate traffic with max_p varying from 0 through 1. As seen in Figures 14 and 15, the queue fill now is affected by max_p. If max_p is very low, then packets are dropped less severely and it leads to line drops because the queue fills too quickly. On the other hand if max_p is very high then excessive packet drops are observed for the class that is affected. Hence for variable rate traffic, the max_p value should be set depending on the range of loads to be supported. However, regardless of how max_p is set, the effect is not as significant as is changing other parameters.

VI. SUMMARY OF DESIGN GUIDELINES

Given the results and observations from the previous sections, the following guidelines are recommended for deploying MAMT AQM to provide preferential treatment to emergency traffic.

1. Establish a minimum queue size that will provide zero dropping to emergency traffic, given the expected burstiness of the input traffic and the range of loads to be supported. This should be around 80 to 100 packets.
2. Select the averaging parameter w_q that will provide sufficiently low variance in queue fill but also not be so large as to make AQM lack robustness. Recommended values for w_q are in the range (0.002, 0.1).
3. Determine the spacing of the MAMT curves so that a desired balance is achieved between throughput for higher priority packets versus oscillations in the queue fill that will affect delay variation. A spacing of 15 packets should be sufficient to protect throughput for emergency traffic.
4. Finally, some adjustments in max_p values can be made to provide fine tuning to protect against line drops or excessive dropping of higher priority packets.

VII. CONCLUSION

The four contributions of this work are as follows.

(1) AQM is used to support emergency traffic requirements for availability and dependability. It is strategically placed where severe congestion is anticipated and is used to drop non-emergency packets enough to allow emergency traffic to proceed.

(2) MAMT AQM with coupled queues is proposed to support emergency traffic.

(3) Non-responsive, UDP-type traffic is studied in isolation to understand its particular impact on AQM performance. The characteristics of AQM itself are shown to have a significant impact on performance independent of whether the traffic is TCP or UDP.

(4) MAMT AQM is shown to perform effectively over a range of loads, as studied through fluid flow analysis and simulation. Oscillations in the queue fill can be affected by w_q, the spacing between AQM curves, the steepness of the curves, and the input traffic rate

and burstiness. Oscillations must be contained to avoid tail drops and excessive delay variation.

Although the nature of emergency response is inherently dynamic and unpredictable, strategic deployment of these proposed AQM functions could provide significant improvements. Internet-based services for emergency workers can be created to be effectively used and trusted to make resources available when needed.

References

[1] B. Brewin, Nation's Networks See Sharp Volume Spikes After Attacks, *Computerworld*, September 17, 2001.

[2] Computer Science and Telecommunications Board, *The Internet Under Crisis Conditions: Learning from September 11*, 2002.

[3] Internet Engineering Task Force, Internet Emergency Preparedness Working Group, http://www.ietf.org/html.charters/ieprep-charter.html.

[4] H. Folts, “Standards Initiatives for Emergency Telecommunications Service (ETS),” *IEEE Communications Magazine*, July 2002, pp. 102-107.

[5] Ken Carlberg, Ian Brown, Cory Beard, "Framework for Supporting ETS in IP Telephony", *Work-in-Progress, Internet Engineering Task Force, Internet Draft*, draft-ietf-ieprep-framework-09.txt, February 2004.

[6] R. Makkar, I. Lamadaris, J. H. Salim, N. Seddigh, B. Nandy, and J. Babiarz, “Empirical Study of Buffer Management Scheme for DiffServ Assured Forwarding PHB,” in *International Conference on Computer Communications and Networks (ICCCN 2000)*, 2000.

[7] D. D. Clark and W. Fang, “Explicit allocation of best-effort packet delivery service,” *IEEE/ACM Trans. Networking*, vol. 6, pp. 362–373, Aug. 1998.

[8] C. V. Hollot, Yong Liu, Vishal Misra, and Don Towsley, “Unresponsive Flows and AQM Performance,” *Proc. of IEEE INFOCOM ’03*.

[9] E. Bowen, C. Jeffries, L. Kencl, A. Kind, and R. Pletka. “Bandwidth allocation for non-responsive flows with active queue management,” *Int. Zurich Seminar on Broadband Communications, IZS 2002*.

[10] I. Brown, “A Security Framework for Emergency Communications,” Internet Engineering Task Force Internet Draft, Work in Progress, draft-ietf-ieprep-security-01.txt.

[11] Government Emergency Telecommunications Service (GETS), National Communications System, http://gets.ncs.gov/.

[12] The President’s National Security Telecommunications Advisory Committee, Network Group, *Internet Report: An Examination of the NS/EP Implications of Internet Technologies*, June 1999.

[13] 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.

[14] C. Beard, "Preemptive and Delay-Based Mechanisms to Provide Preference to Emergency Traffic," under review by *Computer Networks*. Available at http://www.sce.umkc.edu/~beardc.

[15] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, “Assured Forwarding PHB Group”, IETF RFC 2597, June 1999.

[16] S. Floyd and V. Jacobson, “Random early detection gateways for congestion avoidance,” *IEEE Trans. Networking*, vol. 1, no. 4, pp. 397-413, Aug. 1993.

[17] http://www.icir.org/floyd/red.html.

[18] M. May, J. Bolot, C. Diot, and B. Lyles, “Reasons Not to Deploy RED,” *Proc. of 7th. Int. Work-shop on Quality of Service (IWQoS’99)*, pp 260–262, June 1999.

[19] Ratul Mahajan, Sally Floyd, and David Wetherall, “Controlling High-Bandwidth Flows at the Congested Router,” *9th International Conference on Network Protocols (ICNP)*, November 2001

[20] S. Athuraliya, V. H. Li, S. H. Low and Q. Yin, “REM: Active Queue Management,” *IEEE Network*, May 2001.

[21] J. Aweya, M. Ouellette, and D. Y. Montuno, “A Control Theoretic Approach to Active Queue Management,” *Computer Networks*., vol. 36, issue 2–3, July 2001, pp. 203–35.

[22] S. Floyd, R. Gummadi, and S. Shenker, "Adaptive RED: an algorithm for increasing the robustness of RED," available at http://www.icir.org/floyd/papers/adaptiveRed.pdf, Aug. 2001.

[23] Yossi Chait, C.V. Hollot, V. Misra, D. Towsley, Honggang Zhang and John Lui, "Providing Throughput Differentiation for TCP Flows Using Adaptive TwoColor Marking and Multi-Level AQM", *Proceedings of IEEE INFOCOM 2002*.

[24] V. Misra, W. Gong, and D. Towsley, “Fluid based analysis of a network of AQM routers supporting TCP flows with an application to RED,” in *Proceedings of ACM/SIGCOMM, 2000*.

[25] R.W. Brockett, W.B. Gong, and Y. Guo, “Stochastic analysis for fluid queueing systems," in *Proceedings of IEEE CDC'99*, 1999, pp. 3077-3082.

[26] Ajay Mansata, Salil Talauliker, Manali Joshi, and Cory Beard, "Removal of the Noise Source Inherent to AQM," Technical Report, available at http://www.sce.umkc.edu/~beardc/AQM_Correction.pdf

[27] http://www.isi.edu/nsnam/ns.