Towards the Formation of Comprehensive SLAs between Heterogeneous Wireless DiffServ Domains

Mohsin Iftikhar and Bjorn Landfeldt
School of Information Technologies
University of Sydney
Sydney, NSW, Australia
mohsinif@it.usyd.edu.au, bjornl@it.usyd.edu.au

Mine Caglar
Department of Mathematics
Koc University
Istanbul, Turkey
mcaglar@ku.edu.tr

***Abstract*—Traffic patterns generated by multimedia services are different from traditional Poisson traffic. It has been shown in numerous studies that multimedia network traffic exhibits self-similarity and burstiness over a large range of time-scales. The area of wireless IP traffic modeling for the purpose of providing assured QoS to the end-user is still immature and the majority of existing work is based on characterization of wireless IP traffic without any coupling of the behaviour of queueing systems under such traffic conditions. Work in this area has either been limited to simplified models of FIFO queueing systems which do not accurately reflect likely queueing system implementations or the results have been limited to simplified numerical analysis studies. In this paper, we advance the knowledge of queueing systems by example of traffic engineering of different UMTS service classes. Specifically, we examine QoS mapping using three common queueing disciplines; Priority Queuing (PQ), Low Latency Queuing (LLQ) and Custom Queueing (CQ), which are likely to be used in future all-IP based packet transport networks. The present study is based on a long-range dependent traffic model, which is second order self-similar. We consider three different classes of self-similar traffic fed into a G/M/1 queueing system and construct analytical models on the basis of non-preemptive priority, low-latency queueing and custom queueing respectively. In each case, expressions are derived for the expected waiting times and packet loss rates of different traffic classes. We have developed a comprehensive discrete-event simulator for a G/M/1 queueing system in order to understand and evaluate the QoS behaviour**

of self-similar traffic and carried out performance evaluations of multiple classes of input traffic in terms of expected queue length, packet delay and packet loss rate. Furthermore, we have developed a traffic generator based on the self-similar traffic model and fed the generated traffic through a CISCO router-based test bed. The results obtained from the three different queueing schemes (PQ, CQ and LLQ) are then compared with the simulation results in order to validate our analytical models.

Keywords: *QoS, Self-Similar Queuing systems, 3G, 3GPP, UMTS, CDMA2000, GGSN, PDSN.*

I. INTRODUCTION

The IETF has standardized two main QoS frameworks IntServ [1] and DiffServ [2] to provide predictable and controllable behavior of IP networks. IntServ focuses on supporting individual applications by providing an architecture requiring per-flow traffic at every hop along an application's end-to-end path. The Resource Reservation Protocol (RSVP) is used to reserve resources in routers within an IS domain to provide particular QoS levels to different flows. As a counterpoint to the relative complexity and end-to-end nature of IntServ, the DiffServ domain does not reserve network resources on a per-flow basis, traffic is instead classified into a number of traffic groups. Each group is labeled appropriately by a particular value termed Differentiated Services Code Point (DSCP) and based on this value, each group is then treated independently by the DiffServ domain.

The dramatic increase in demand for wireless Internet access has lead to the introduction of new wireless architectures and systems including 3G, Wi-Fi and WiMAX. It is a likely development that mobile terminals will increasingly have capability to access many of these wireless networks types. Because of the scalable class-based traffic management mechanism, without using per-flow resource reservations, DiffServ is currently the most promising architecture to interwork the heterogeneous wireless access networks and the Internet to provide seamless global roaming and broadband access to the end-user [3-4]. The domain-based resource management feature of DiffServ makes it the most suitable platform for interconnecting heterogeneous wireless access networks because each domain can freely choose whatever policies are proper for internal resource management as long as its Service Level Agreements (SLAs) are met with neighboring domains.

II. ALL-IP DIFFSERV ARCHITECTURE FOR WIRELESS INTERNET

In this paper, we investigate the DiffServ architecture (shown in Fig. 1) for interconnecting heterogeneous wireless

access networks as given in [5]. In this all-IP DiffServ architecture, a number of nearby radio access networks (RANs) having the same interface are grouped into a wireless DiffServ domain and further, all the domains are connected through a DiffServ IP backbone to provide end-to-end Internet services to the mobile station (MS). It is assumed that in each DiffServ wireless domain, all network elements of RAN are enhanced to fulfill the functionality of a DiffServ IP router. The gateway and base stations operate as edge routers of the domain and are connected through core routers [5]. Further, the gateway is the interface to the Internet backbone. For example, GPRS support node (GGSN) is the gateway of the UMTS domain to the external DiffServ Internet; similarly, PDSN is the gateway of the domain to the external DiffServ Internet in CDMA2000. In the gateway, SLAs are implemented to specify the resources allocated by Internet Service Provider flowing from/into the domain. The gateway conditions the aggregate traffic for each service class according to SLA resource commitments. All DiffServ routers use different queuing and scheduling algorithms, to provide differentiated classes of services.

It has been shown that wireless data traffic exhibits self-similarity and long-range dependency [6-9]. While taking into account the self-similar nature of multiservice traffic, it is not mundane to build tight bound SLAs between heterogeneous QoS domains due to the high variability in the offered traffic.

To offer realistic SLAs based on tight bound QoS parameters, between a wireless DiffServ domain gateway (for example GGSN) and the DiffServ Internet backbone, the present study focuses on the performance evaluation of three different queueing schemes (PQ, CQ and LLQ) in a DiffServ domain with multiple classes of self-similar input traffic. The derived QoS parameters in terms of expected queue length, packet delay and packet loss rate forms the basis on which to build realistic SLAs and ultimately provide support to interwork heterogeneous wireless access networks. This in turn is necessary to provide seamless global roaming, fast handoff and end-to-end QoS to the end-user.

Cellular to IP QoS Mapping

3GPP has defined four QoS classes for UMTS; (1) Conversational, (2) Interactive, (3) Streaming and (4) Background. Traffic is classified and ordered on the basis of relative delay sensitivity [10]. On the other hand, DiffServ defines Expedited Forwarding (EF) per-hop behavior for premium service and the Assured Forwarding (AF) PHB for elastic but time constrained service in addition to the classical best effort class [11]. To extend IP services to the wireless domain, the UMTS QoS classes must be mapped to the DiffServ classes. According to 3GPP, UMTS-to-IP QoS mapping is performed by a translation function in the GGSN that classifies each UMTS packet flow and maps it to a

suitable IP QoS [12]. Normally, the conversational class can be mapped to EF PHB for very low-delay and low-loss service, streaming and interactive traffic to the AF PHB and background traffic to best-effort service [5]. Under such mapping, we present a model based on a G/M/1 queueing system by considering three different classes of self-similar traffic input and analyze it on the basis of PQ, LLQ and CQ. The work in this paper extends on an initial conference paper [13] and brings the following major contributions to the wireless traffic modeling;

- We present closed form expressions of packet delay and PLR for different classes under PQ, LLQ and CQ service disciplines.
- We build the Markov chain for all three systems.
- We develop a comprehensive discrete-event simulator for a G/M/1 queueing system in order to understand and evaluate the QoS behavior of self-similar traffic. The simulation study produces performance evaluation results of multiple classes of input traffic in terms of expected queue length, packet delay and packet loss rate.
- A traffic generator has been developed to realize our self-similar traffic model.
- We implement a Cisco-router based test bed, which serves to experimentally validate the simulation results. The results obtained from the three different queuing schemes (PQ, CQ and LLQ) provide a foundation for better resource allocation to different traffic classes based on different QoS parameters e.g. delay, queue length and packet loss rate.

The remainder of the paper is organized as follows. Section III and IV are devoted to explaining the self-similar traffic model with multiple classes and the formulation of Embedded Markov Chain along with the derivation of packet delays and PLR. We present simulation and test-bed results in Section V. Section VI gives an overview of related work. Section VII presents the applications of the work and finally, conclusion and future work is given in Section VIII.

III. SELF-SIMILAR TRAFFIC WITH SEVERAL CLASSES

In this section, we review the self-similar traffic model introduced in [14] and the associated interarrival time distributions with several classes.

A. Traffic Model

The traffic model captures the dynamics of packet generation while accounting for the scaling properties observed in telecommunication networks [14]. The traffic model is parsimonious (with few parameters to match measurements). The model is analytical (solvable when fed into queueing models), flexible (one model but many variants for different applications), implement-able (less time consuming for simulation) and exhibits absolute accuracy (critical for business case studies). The model is furthermore similar to on/off processes. It belongs to a particular class of self-similar traffic models called infinite source Poisson models. Our traffic model is long-range dependent and almost second-order self-similar as the auto-covariance function of its increments is equal to that of fractional Gaussian noise for sufficiently large time lags. The traffic can be approximated by an FBM or a Levy process when the rate of packet arrivals tends to infinity [14, 15]. Bordered by these self-similar and/or long-range dependent stochastic processes for data traffic, our packet generation model covers a wide range of statistical distributions through the choice of its parameters.

The traffic is found by aggregating the number of packets generated by several sources. Each source initiates a session with a Pareto distribution whose density is given by $g(r) = \delta b^{\delta} r^{-\delta-1}$, $r > b$, where δ is related to the Hurst parameter by $H = (3-\delta)/2$. The sessions arrive according to a Poisson process with rate λ and the packets generated by each source arrive according to a Poisson process with rate α locally throughout each session [14]. The traffic $Y(t)$ measured as the total number of packets injected to a router during [0, t] can be written as

$$Y(t) = \sum_{S_i \le t} U_i (R_i \wedge (t - S_i))$$

where U_i denotes the local Poisson process over session i, R_i and S_i denote the duration and the arrival time of session i, respectively, and the values of i denote an enumeration of the arriving sessions. Here, R_i is positive, S_i is real valued and U_i which counts the number of packets of session i is integer valued. As a result, $Y(t)$ corresponds to the sum of packets generated by all sessions initiated in [0,t] until a session expires when R_i is less than t-S_i, and until t if the session is active at that time. We consider the stationary version of this model based on an infinite past.

In the present study, we replicate the traffic model as many times as needed to represent different classes of traffic streams. Each stream has its own parameters and is assumed to be independent from the other(s). The packet sizes are taken to be fixed because each queue or traffic class corresponds to a certain type of application where the packets have fixed size or at least fixed service time distribution. Although the local packet generation is assumed to be Poisson over each session, the aggregated packet arrival process is clearly not Poisson. This aspect is consistent with the long-range dependence of the packet arrivals.

B. Interarrival Times for Several Classes

In contrast to other infinite source Poisson models or on/off processes, our model lends itself to the computation of the interarrival time distribution of consecutive packets under certain simplifications. The details of the derivation have been recently given in [16]. We have shown that the complementary cumulative distribution function of the interarrival time *T* for one class of packets is given by

$$\overline{F}_T(t) = P\{T > t\} = \frac{e^{-\lambda \mu_G}}{1-e^{-\lambda \mu_G}} \exp[-\lambda t(1-e^{-\alpha t})]\left[\exp[v(A_t)e^{-\alpha t}]\exp[v(B_t)(1-e^{-\alpha t})/(\alpha t)]-1\right].$$

where μ_G denotes the mean of the Pareto distribution and

$$v(A_t) = \lambda \int_{-\infty}^{0} \int_{t-s}^{\infty} g(r)\,dr\,ds = \lambda \int_{t}^{\infty} (r-t)g(r)\,dr = \lambda \int_{t}^{\infty} r\,g(r)\,dr - \lambda t \overline{G}(t) \qquad (1)$$

$$v(B_t) = \lambda \int_{-\infty}^{0} \int_{-s}^{t-s} g(r)\,dr\,ds = \lambda \int_{0}^{t} r\,g(r)\,dr + \lambda t\,\overline{G}(t) \qquad (2)$$

This can be differentiated and negated to find the probability density function $f_T(t)$ of *T*.

We consider three classes of traffic streams arriving at a router. Let T_i denote the interarrival time of class *i* packets, *i*=1, 2, 3. We have derived the distribution of the interarrival time between a type *i* and a type *j* packet when two types of packets arrive at the router in [16]. The generalization to three classes is trivial. Given that a type *i* arrival occurred and the next arrival is again type *i*, the density of the time until the next arrival is denoted by $f_{T_i}(t)$. We first overview the derivation of the cross interarrival time density for the arrival of a type 2 packet given that a type 1 arrival

occurred. If a type 1 packet arrived at the current time, this information has no implication on the number of active sessions of class 2 or 3. Then, we compute the complementary probability

$$\overline{F_2}^0(t) = P\{\text{no type 2 packets arrive in } t \text{ time units}\} \qquad (3)$$

given by

$$\overline{F_2}^0(t) = e^{-\nu_2(A_t)} e^{-\nu_2(B_t)} \exp[-\lambda_2 t(1-e^{-\alpha_2 t})] \exp[\nu_2(A_t)e^{-\alpha_2 t}] \exp[\nu_2(B_t)(1-e^{-\alpha_2 t})/(\alpha_2 t)]$$

where $\nu_2(A_t)$ and $\nu_2(B_t)$ are defined analogously as in (1) and (2). The detailed derivation has been given in [16]. The density function of the time until the arrival of a class 2 packet next is denoted by $f_2^0(t)$, which can be found through taking the derivative of the complementary distribution function $\overline{F}_2^0$. Here, $\overline{F_2}^0(t)$ differs from $\overline{F}_{T_2}(t)$ by the condition that a type 2 session is active is assumed in the latter probability whereas it may or may not be active in the first one. In other words, the first is an unconditional probability and the latter is conditional on the event that a type 2 arrival has occurred. The use of the density functions $f_{T_i}(t)$ and $f_i^0(t)$ in the Markov chain of the next section is as follows. For a transition to occur from a class 1 arrival to a class 1 arrival; the event "no type 2 or type 3 packets arrive in *t* time units" must occur, which has probability $\overline{F_2}^0(t)\overline{F_3}^0(t)$ where $\overline{F_3}^0(t)$ can be written analogously to (3). Then, the probability that a transition from a state involving an arrival of type 1 to another state also with an arrival of type 1 is found by using the fact that the next arrival will occur at time *t* with density $f_{T_1}(t)$ and with the condition that neither class 2 nor class 3 packets arrive in the mean time, which happens with probability $\overline{F_2}^0(t)\overline{F_3}^0(t)$. Hence, we can make use the product $f_{T_1}(t)\overline{F_2}^0(t)\overline{F_3}^0(t)$ to calculate the complete transition probability from a given state to another, when both states have an arrival of type 1. Along the same lines, the density $f_2^0(t)$ gets multiplied with $\overline{F}_{T_1}(t)\overline{F}_3^0(t)$ to make sure that a transition occurs from a class 1 arrival to a class 2 arrival and the time until the next arrival is *t*. In this case, the given condition is on type 1 packet. Therefore, we have the conditional probability $\overline{F}_{T_1}(t)$. Other combinations follow similarly. Although it does *not* denote a density

function, we use the notation $f_{T_{ij}}$ to denote a product of a density and two complementary probabilities when a class *i* packet is followed by a class *j* packet. That is, the notation used below is

$$f_{T_{ii}}(t) = f_{T_i}(t)\overline{F}_j^0(t)\overline{F}_k^0(t) \qquad f_{T_{ij}}(t) = f_{T_j}^0(t)\overline{F}_{T_i}(t)\overline{F}_{T_k}^0(t)$$

where we multiply the corresponding density with complementary probabilities to make sure that the desired transition occurs from type *i* arrival to type *i* or *j*, and $i, j, k \in \{1,2,3\}$.

IV. ANALYTICAL MODELS OF VARIOUS QUEUEING DISCIPLINES

We consider a model of three queues based on G/M/1 by taking into account three different classes of self-similar input traffic denoted by SS/M/1, and we analyze it on the basis of priority with no preemption. Let the service time distribution have rate μ_1, μ_2 and μ_3 for type 1, type 2 and type 3 packets, respectively, and let type 1 packets have priority over type 2 and type 3 packets, similarly type 2 packets have priority over type 3 packets.

A. SS/M/1 with Three Classes: Non Preemptive Priority Service

The usual embedded Markov chain [17] formulation of $G/M/1$ is based on the observation of the queueing system at the time of arrival instants, right before an arrival. At such instants, the number in the system is the number of packets that arriving packet sees in the queue plus packet in service, if any, excluding the arriving packet itself. We specify the states and the transition probability matrix *P* of the Markov chain with the self-similar model for three types of traffic.

Let $\{X_n : n \geq 0\}$ denote the embedded Markov chain at the time of arrival instants. As the service is based on priority, the type of packet in service is important at each arrival instant of a given type of packet to determine the queueing time. Therefore, we define the state space as:

$$S = \{(i_1, i_2, i_3, a, s) : a \in \{a_1, a_2, a_3\}, s \in \{s_1, s_2, s_3, I\}, i_1, i_2, i_3 \in Z_+\}$$

where a_1, a_2 and a_3 are labels to denote the type of the arrival, s_1, s_2 and s_3 are labels to denote the type of the packet in service, i_1, i_2 and i_3 are the number of packets in each queue including a possible packet in service, and *I* denotes the idle state in which no packet is either in service or being queued.

Some of the states in the state space *S* given above have zero probability. For example, $(i_1,0,i_3,a_1,s_2)$ is impossible. The particular notation is chosen for simplicity, although the impossible states could be excluded from *S*. Each possible state, the reachable states from each and the corresponding transition probabilities will be explicitly shown in the sequel.

B. States of the Embedded Markov Chain for Non-Preemptive Priority Service

The states of the Markov chain and the possible transitions with respective probabilities can be enumerated by considering each case. We will analyze the states with non-empty queues and those with at least one empty queue at the time of an arrival, separately.

States (i_1,i_2,i_3,a,s) *with* $i_1,i_2,i_3 \neq 0$ *and* $s \neq I$ *:*

We can divide the states and transitions into 81 groups. Because (*a, s*) can occur 3x3=9 different ways, and the next state (*p, q*) can be composed similarly in 9 different ways as $a, p \in \{a_1,a_2,a_3\}$ and $s,q \in \{s_1,s_2,s_3\}$. We will analyze the two states in detail; the others follow similarly.

Transition from $(i_1,i_2,i_3,a_1,s_1) \rightarrow (j_1,j_2,j_3,a_2,s_2)$

This is the case where a transition occurs from an arrival of type 1 to an arrival of type 2 such that the first arrival has seen a type 1 packet in service, i_1 packets of type 1 in the system (equivalently, total of queue 1 and the packet in service) and i_2 packets of type 2 and i_3 packets of type 3 in the system. The transition occurs to j_1 packets of type 1, j_2 packets of type 2 and j_3 packets of type 3 in the system with a type 2 packet in service. This transition has been shown in Fig. 2. Due to priority scheduling, an arrival of type 2 can see a type 2 packet in service in the next state only if all type 1 packets including the one that arrived in the previous state are exhausted during the interarrival time. That is why j_1 can take only the value 0 and exactly i_1+1 packets of type 1 are served. In contrast, the number of packets served from queue 2, say *k*, can be anywhere between 0 and i_2-1 as at least one type 2 packet is in the system, one being in service, when a new arrival occurs. The transition probability is

$$P\{X_{n+1} = (0, i_2 - k, i_3, a_2, s_2) \mid X_n = (i_1, i_2, i_3, a_1, s_1)\}$$

$$= P\{i_1 + 1 \text{ served from type 1, } k \text{ served from type 2 and a type 2 packet remains in service during } T_{12}\}$$

where we use the fact that the remaining service time of a type 1 packet in service has the same exponential distribution Exp (μ_1), due to the memory-less property of a Markovian service and we denote the interarrival time by T_{12}. Therefore, for $k = 0,\ldots,i_2 - 1$

$$P\{X_{n+1} = (0, i_2 - k, i_3, a_2, s_2) \mid X_n = (i_1, i_2, i_3, a_1, s_1)\}$$

$$= \int_0^\infty \int_0^t \int_{t-x}^\infty f_{S_2}(s)\, f_{S_1^{i_1+1}+S_2^k}(x)\, f_{T_{12}}(t)\, ds dx dt$$

where S_m^l : sum of *l* independent service times of type *m* packets, *m*=1,2; $l \in Z_+$. Note that S_m^l has an Erlang distribution with parameters (l, μ_m) as each service time has an exponential distribution, and the sum $S_1^{l_1} + S_2^{l_2}$ being the sum of several exponentially distributed random variables has a hypoexponential distribution. The density functions of all these distributions can easily be evaluated numerically.

Transition from $(i_1, i_2, i_3, a_1, s_1) \to (j_1, j_2, j_3, a_2, s_1)$

This is the case where a transition occurs from an arrival of type 1 to an arrival of type 2 such that the first arrival has seen a type 1 packet in service, i_1 packets of type 1 in the system (equivalently, total of queue 1 and the packet in service), i_2 packets of type 2 and i_3 packets of type 3 in the system. The transition occurs to j_1 packets of type 1 j_2 packets of type 2 and j_3 packets of type 3 in the system with a type 1 packet in service. This transition has been shown in Fig. 3. An arrival of type 2 sees a type 1 packet in service in the next state, which indicates that no type 2/type 3 packet has been served during this transition due to priority scheduling. In contrast, the number of packets served from queue 1, say *k*, can be anywhere between 0 and i_1 as at least one type 1 packet is in the system, the one being in service, when a new arrival occurs. The transition probability is

$$P\{X_{n+1} = (i_1 - k + 1, i_2, i_3, a_2, s_1) \mid X_n = (i_1, i_2, i_3, a_1, s_1)\}$$

$$= P\{k \text{ served from type 1, no packet served from type 2/type 3 and type 1 packet remains in service during } T_{12}\}$$

$$= \int_0^\infty \int_0^t \int_{t-x}^\infty f_{S_1}(s)\, f_{S_1^k}(x)\, f_{T_{12}}(t)\, ds dx dt$$

The above two transitions are summarized below.

Initial State	Reachable State	Transition Probability
$(i_1, i_2, i_3, a_1, s_1)$	$(0, i_2 - k, i_3, a_2, s_2), k = 0, \ldots, i_2 - 1$	$\int_0^\infty \int_0^t \int_{t-x}^\infty f_{S_2}(s) f_{S_1^{i_1+1}+S_2^k}(x) f_{T_{12}}(t)\, ds dx dt$
$(i_1, i_2, i_3, a_1, s_1)$	$(i_1 - k + 1, i_2, i_3, a_2, s_1), k = 0,1.....i_1$	$\int_0^\infty \int_0^t \int_{t-x}^\infty f_{S_1}(s) f_{S_1^k}(x) f_{T_{12}}(t)\, ds dx dt$

Similarly, we can enumerate all 81 cases.

States (i_1, i_2, i_3, a, s) *with* i_1 *or* i_2 *or* i_3 *equal to* 0 *or* $s = I$:

The states when one queue is empty i.e. ($i_1 = 0$ or $i_2 = 0$ or $i_3 = 0$) or when two queues are empty or when all queues are empty and the system is idle, i.e. $(i_1 = i_2 = i_3 = 0, s = I)$ can be considered similarly. The details can be found in [18].

C. Limiting Distribution and QoS Parameters for PQ Model

The steady state distribution π as seen by an arrival is obtained by solving $\pi P = \pi$, where P is the transition matrix of the Markov chain analyzed above. In practice, the queue capacity is limited in a router. So the Markov chain is finite and the steady state distribution exists.

To the best of our knowledge, no previous analytical expressions are available for the waiting time of a G/M/1 queue with priority. Our analysis relies on the limiting distribution of the state of the queue at the arrival instances, which can be computed using the analysis given above for our self-similar traffic model. In general, the following analysis is valid for any G/M/1 queueing system where the limiting distribution π at the arrival instances can be computed.

The expected waiting time for the high priority queue can be found as:

$$E[W_1] = \sum_{j_1=1}^{J_1-1} \sum_{j_2=0}^{J_2} \sum_{j_3=0}^{J_3} \frac{j_1}{\mu_1} \pi(j_1, j_2, j_3, a_1, s_1) + \sum_{j_1=0}^{J_1-1} \sum_{j_2=1}^{J_2} \sum_{j_3=0}^{J_3} \left(\frac{j_1}{\mu_1} + \frac{1}{\mu_2}\right) \pi(j_1, j_2, j_3, a_1, s_2) +$$

$$\sum_{j_1=0}^{J_1-1} \sum_{j_2=0}^{J_2} \sum_{j_3=1}^{J_3} \left(\frac{j_1}{\mu_1} + \frac{1}{\mu_3}\right) \pi(j_1, j_2, j_3, a_1, s_3)$$

where J_1, J_2 and J_3 are the respective capacities of each queue. This follows clearly from the fact that an arriving packet of higher priority will wait until all packets of the same priority as well as the packet in service are served. Depending on the type of the packet in service, we have the constituent expressions in the sum.

On the other hand, we obtain the expected waiting time of a packet for the low priority queues by analyzing the events that constitute this delay. The amount of work in the system at any time is defined as the (random) sum of all service times that will be required by the packets in the system at that instant. The waiting time of a type 2 packet can be written as:

$$W_2 = Z_1 + Z_2 + Z_3 + ... \qquad (4)$$

where Z_1 is the amount of work seen by the arriving packet in the system, Z_2 is the amount of work associated with high priority (i.e.type 1) packets arriving during Z_1, Z_3 is the amount of work associated with type 1 packets arriving during Z_2, and so on. As illustrated in Fig.4, the waiting time of an arriving packet of type 2 is indeed given by the total workload building in front of it. The arrows in the figure denote the arrival times of type 1 packets, and all the oblique lines have 45 degrees angle with the time axis. In this figure the waiting time is $W_2 = Z_1 + Z_2 + Z_3 + Z_4$ as an example.

Let M_j denote the number of type j arrivals over Z_i, j=1, 2,…. Then

$$W_2 = Z_1 + S_1^{M_1} + S_1^{M_2} + \cdots$$

where $S_1^{M_j}$ denotes the random sum of M_j independent service times of type 1 packets. Then,

$$E[W_2] = E[Z_1] + E[S_1]E[M_1] + E[S_1]E[M_2] + \cdots$$

since the service times and the arrival process are independent. For a stationary packet arrival process, we get

$$E[M_j] = E[E[M_j \mid Z_j]] = E[c_1 Z_j] = c_1 E[Z_j]$$

due to mentioned independence, where $c_1 > 0$ is a constant particular to the arrival process. That is, expectation of the number of arrivals in any period of time is proportional to the length of that period because of stationarity in time and

linearity of expectation. In our stationary self-similar traffic input process, c_1 is the expected number of arrivals per unit time which can be called the arrival rate, given by the product of the arrival rate of session arrivals, the arrival rate of packets over a session, and the expected session length [14].

Explicitly, $c_1 = \lambda\alpha\delta b/(\delta - 1)$. Hence, the expected waiting time reduces to

$$E[W_2] = E[Z_1] + E[S_1]c_1E[Z_1] + E[S_1]c_1E[Z_2] + \cdots$$

$$= E[Z_1] + \frac{c_1}{\mu_1}(E[Z_1] + E[Z_2] + \cdots)$$

$$= E[Z_1] + \frac{c_1}{\mu_1}E[W_2]$$

in view of (4). Therefore, we get

$$E[W_2] = \sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{J_3}(\frac{j_1}{\mu_1} + \frac{j_2}{\mu_2})\pi(j_1, j_2, j_3, a_2, s_1) + \sum_{j_1=0}^{J_1}\sum_{j_2=1}^{J_2-1}\sum_{j_3=0}^{J_3}(\frac{j_1}{\mu_1} + \frac{j_2}{\mu_2})\pi(j_1, j_2, j_3, a_2, s_2) +$$

$$\sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=1}^{J_3}(\frac{j_1}{\mu_1} + \frac{j_2}{\mu_2} + \frac{1}{\mu_3})\pi(j_1, j_2, j_3, a_2, s_3) + \frac{c_1E[W_2]}{\mu_1}$$

which implies that the traffic intensity $\frac{c_1}{\mu_1}$ must be less than 1. Similarly, the expected waiting time for a packet of type 3, which is the lowest priority queue can be found from:

$$E[W_3] = \sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3-1}(\frac{j_1}{\mu_1} + \frac{j_2}{\mu_2} + \frac{j_3}{\mu_3})\pi(j_1, j_2, j_3, a_3, s_1) + \sum_{j_1=0}^{J_1}\sum_{j_2=1}^{J_2}\sum_{j_3=0}^{J_3-1}(\frac{j_1}{\mu_1} + \frac{j_2}{\mu_2} + \frac{j_3}{\mu_3})\pi(j_1, j_2, j_3, a_3, s_2) +$$

$$\sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}\sum_{j_3=1}^{J_3-1}(\frac{j_1}{\mu_1} + \frac{j_2}{\mu_2} + \frac{j_3}{\mu_3})\pi(j_1, j_2, j_3, a_3, s_3) + (\frac{c_1}{\mu_1} + \frac{c_2}{\mu_2})E[W_3]$$

Another QoS parameter readily available from this description of the system is the packet loss rate (PLR) (due to a full queue) or equivalently the system availability. For each class of traffic, this is the sum of the steady-state probabilities of states where an arrival occurs for a full queue:

$$PLR_1 = \sum_{k=0}^{J_3} \sum_{j=0}^{J_2} \sum_{m=1}^{3} \pi(J_1, j, k, a_1, s_m)$$

D: SS/M/1 with Three Classes:Low Latency Queueing (LLQ) Service Discipline

We consider a model of three queues (one priority queue and two non-priority queues) based on G/M/1 by considering three different classes of self-similar traffic input. We analyze the system using LLQ as the scheduling discipline (the scheduler can serve non-priority queues only if there is no packet waiting in priority queue; further the scheduler serves non-priority queues in a round robin fashion according to specified reserved bandwidth by taking fixed number of bytes (packets) during each cycle; we specify the scheduler logic in such a way that the scheduler serves one packet from each non-priority queue during each cycle provided there is no packet waiting in priority queue) [19]. We develop the finite Markov chain for LLQ scheduling discipline; extending the previous work on infinite capacity system. The formulation is based on observation of the queueing system at packet arrival instants. At these instants, the number in the system is the number of packets that the arriving packet sees in the queue plus the packet in service, if any, excluding the arriving packet itself. Let $\{X_n : n \geq 0\}$ denote the embedded Markov chain at the time of arrival instants. We define the state space as:

$$S = \{(i_1, i_2, i_3, a, s) : a \in \{a_1, a_2, a_3\}, s \in \{s_1, s_2, s_3, I\}, i_1, i_2, i_3 \in Z_+\}$$

We generate the transition probability matrix P of the Markov chain by specifying the transition probabilities from all the states in the states space i.e. non-idle states, states with empty queues and arrival at full queue. We only write down one transition in detail:

Transition from $(i_1, i_2, i_3, a_1, s_1) \rightarrow (j_1, j_2, j_3, a_2, s_2)$

We consider the case in which a transition occurs from an arrival of type 1 to an arrival of type 2 such that the first arrival has seen a type 1 packet in service, i_1 packets of type 1 (equivalently, total of queue 1 and the packet in service), i_2 packets of type 2 and i_3 packets of type 3 in the system. The transition occurs to j_1 packets of type 1, j_2 packets of type 2 and j_3 packets of type 3 in the system with a type 2 packet in service. Due to LLQ scheduling, an arrival of type 2 can see a type 2 packet in service in the next state only if all type 1 packets including the one that

arrived in the previous state are exhausted during the interarrival time. That is why j_1 can take only the value 0 and exactly i_1+1 packets of type 1 are served. In contrast, the number of packets served from queue 2, say *k*, can be anywhere between 0 and i_2-1 as at least one type 2 packet is in the system, one being in service, when a new arrival occurs. Similarly, the number of packets served from queue 3 can be anywhere between 0 and i_3 due to RR scheduling between queue 2 and queue 3 and depending on the condition $(i_2 < i_3$ or $i_2 \geq i_3)$. This transition has been shown in Fig. 5. The transition probabilities are: if $i_2 < i_3$:

$$P\{X_{n+1} = (0, i_2 - k, i_3 - k, a_2, s_2) \mid X_n = (i_1, i_2, i_3, a_1, s_1)\}$$

$$= \int_0^\infty \int_0^t \int_{t-x}^\infty f_{S_2}(s)\, f_{S_1^{i_1+1}+S_2^k+S_3^k}(x)\, f_{T_{12}}(t)\, ds dx dt$$

where we use the fact that the remaining service time of a type 1 packet in service has the same exponential distribution Exp(μ_1), due to the memory-less property of a Markovian service, f_S is the density function for service time *S*, S_i^j is the sum of *j* i.i.d service times of type *i* packets, and we denote the density of the interarrival time from a type 1 to type 2 arrival multiplied with the probability that no other type of arrivals in between by $f_{T_{12}}$. Or if $i_2 \geq i_3$

$$P\{X_{n+1} = (0, i_2 - k, 0, a_2, s_2) \mid X_n = (i_1, i_2, i_3, a_1, s_1)\}$$

$$= \int_0^\infty \int_0^t \int_{t-x}^\infty f_{S_2}(s)\, f_{S_1^{i_1+1}+S_2^k+S_3^{i_3}}(x)\, f_{T_{12}}(t)\, ds dx dt$$

Similarly we can write down all possible states. The details are given in [18]

E. Limiting Distribution and QoS Parameters for LLQ Model

Steady state distribution π as seen by an arrival can be found by solving $\pi P = \pi$ using the transition matrix P of the Markov chain analyzed above. In practice, the queue capacity is limited in a router. So, the steady state distribution exists. To the best of our knowledge, no previous analytical expressions are available for the waiting time of a G/M/1 queue with LLQ. Our analysis relies on the limiting distribution of the state of the queue at the arrival instances, which can be computed using the analysis given above for our self-similar traffic model. In general, the following analysis is valid for any G/M/1 queueing system where the limiting distribution π at the arrival instances can be computed. The expected waiting time for the low latency queue (priority queue) can be found as

$$E[W_1] = \sum_{j_1=1}^{J_1-1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3}\frac{j_1}{\mu_1}\pi(j_1, j_2, j_3, a_1, s_1) + \sum_{j_1=0}^{J_1-1}\sum_{j_2=1}^{J_2}\sum_{j_3=0}^{J_3}(\frac{j_1}{\mu_1} + \frac{1}{\mu_2})\pi(j_1, j_2, j_3, a_1, s_2) +$$

$$\sum_{j_1=0}^{J_1-1}\sum_{j_2=0}^{J_2}\sum_{j_3=1}^{J_3}(\frac{j_1}{\mu_1} + \frac{1}{\mu_3})\pi(j_1, j_2, j_3, a_1, s_3)$$

The expression is exactly similar to the highest priority queue of PQ system. This follows clearly from the fact that an arriving packet of higher priority will wait until all packets of the same priority as well as the packet in service are served. Depending on the type of the packet in service, we have the constituent expressions in the sum. On the other hand, we obtain the expected waiting time for the non low latency queues (low priority queues) by analyzing the events that constitute this delay. The amount of work in the system at any time is defined as the (random) sum of all service times that will be required by the packets in the system at that instant. The expected waiting time for a packet arriving to queue 2 or queue 3 is same due to the symmetry of alternating service. We consider two factors (the impact of high priority queue and the effect of alternating service) to find out the expected waiting time of a packet arriving to non priority queues. By combining these two factors, we derive exact bounds on packet delay for non priority queues as $C_2 \le E[W_2] \le C_3$, where

$$C_2 = \sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=j_2+1}^{J_3}\pi(j_1, j_2, j_3, a_2, s_1)(\frac{j_1}{\mu_1} + \frac{j_2}{\mu_2} + \frac{j_2}{\mu_3}) + \sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{j_2}\pi(j_1, j_2, j_3, a_2, s_1)(\frac{j_1}{\mu_1} + \frac{j_2}{\mu_2} + \frac{j_3}{\mu_3}) +$$

$$\sum_{j_1=0}^{J_1}\sum_{j_2=1}^{J_2-1}\sum_{j_3=j_2+1}^{J_3}\pi(j_1, j_2, j_3, a_2, s_2)(\frac{j_1}{\mu_1} + \frac{1}{\mu_2} + \frac{j_2-1}{\mu_2} + \frac{j_2}{\mu_3}) + \sum_{j_1=0}^{J_1}\sum_{j_2=1}^{J_2-1}\sum_{j_3=0}^{j_2}\pi(j_1, j_2, j_3, a_2, s_2)(\frac{j_1}{\mu_1} + \frac{1}{\mu_2} + \frac{j_2-1}{\mu_2} + \frac{j_3}{\mu_3}) +$$

$$\sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=j_2+1}^{J_3}\pi(j_1, j_2, j_3, a_2, s_3)(\frac{j_1}{\mu_1} + \frac{1}{\mu_3} + \frac{j_2}{\mu_2} + \frac{j_2}{\mu_3}) + \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=1}^{j_2}\pi(j_1, j_2, j_3, a_2, s_3)(\frac{j_1}{\mu_1} + \frac{1}{\mu_3} + \frac{j_2}{\mu_2} + \frac{j_3-1}{\mu_3}) + \frac{c_1}{\mu_1}C_2$$

and

$$C_3 = \sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{J_3}\pi(j_1, j_2, j_3, a_2, s_1)(\frac{j_1-1}{\mu_1} + \frac{j_2}{\mu_2} + \frac{j_2}{\mu_3} + \frac{1}{\mu_1}) + \sum_{j_1=0}^{J_1}\sum_{j_2=1}^{J_2-1}\sum_{j_3=0}^{J_3}\pi(j_1, j_2, j_3, a_2, s_2)(\frac{j_1}{\mu_1} + \frac{j_2-1}{\mu_2} + \frac{j_2}{\mu_3} + \frac{1}{\mu_2})$$

$$+ \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=1}^{J_3}\pi(j_1, j_2, j_3, a_2, s_3)(\frac{j_1}{\mu_1} + \frac{1}{\mu_3} + \frac{j_2}{\mu_2} + \frac{j_2}{\mu_3}) + \frac{c_1}{\mu_1}C_3$$

F: SS/M/1 with Three Classes: Custom Queueing (CQ) Service Discipline

We consider a model of three queues based on G/M/1 by considering three different classes of self-similar traffic input. We analyze the system using Custom Queueing (CQ) as the scheduling discipline; in which the scheduler serves the specific number of queues in a round robin fashion by taking fixed number of bytes (packets) from each

queue [19]. We specify the scheduler logic in such a way that the scheduler serves two packets from queue no. 1, one packet from queue no. 2 and one packet from queue no. 3 during each cycle. We develop the finite Markov chain for CQ scheduling discipline; extending the previous work on infinite capacity system. The formulation is based on observation of the queueing system at packet arrival instants. At these instants, the number in the system is the number of packets that the arriving packet sees in the queue plus the packet in service, if any, excluding the arriving packet itself. Let the service time distribution have rates μ_1, μ_2 and μ_3 for class 1, class 2 and class 3 packets respectively. A cycle consists of $s_1^1 + s_1^2 + s_2 + s_3$ time units. Since the scheduler serves two packets from queue no. 1, one packet from queue no. 2 and one packet from queue no. 3 during each cycle; hence, we need to differentiate between the first and second packet of queue 1 of the same cycle and then we need to classify between class 1, class 2 and class 3 packets as well. Therefore s_1^1 is the first packet of queue 1, s_1^2 is the second packet of queue 1, s_2 is the packet of queue 2 and s_3 is the packet of queue 3. That's why the notation s_n^m can be used to differentiate between these four different kinds of packets of the same cycle, where $m = 1,2$ and $n = 1,2,3$. The subscript m will be used only when $n = 1$, where s_n^m is the service time required by class *n* packets.

Let $\{X_n : n \geq 0\}$ denote the embedded Markov chain at the time of arrival instants. We define the state space as:

$$S = \{(i_1, i_2, i_3, a, s) : a \in \{a_1, a_2, a_3\}, s \in \{s_1^1, s_1^2, s_2, s_3, I\}, i_1, i_2, i_3 \in Z_+\}$$

We generate the transition probability matrix P of the Markov chain by specifying the transition probabilities from all the states in the states space i.e. non-idle states, states with empty queues (i.e idle states) and arrival at full queue. We only write down one transition in detail:

Transition from $(i_1, i_2, i_3, a_1, s_1^1) \rightarrow (j_1, j_2, j_3, a_2, s_1^2)$

Here a transition occurs from an arrival of class 1 traffic to an arrival of class 2 traffic, such that the class 1 arrival has seen the first packet of class 1 traffic in service of some cycle, with i_1 packets of class 1 in the system (equivalently, total of queue 1 and the packet in service), i_2 packets of class 2 and i_3 packets of class 3 in the system. The transition occurs to a state, where the new arrival (class 2 packet) sees j_1 packets of class 1, j_2 packets of class 2 and

j_3 packets of class 3 in the system, with a second packet of class 1 in service of some cycle. Recall that there are two kinds of classification, one is between first and second packet of same cycle of class 1 (queue 1) and then between class 1, class 2 and class 3 packets (queue 1, queue 2 and queue 3). Since in the previous state an arrival of class 1 has seen the first packet of class 1 in service, and in the next state an arrival of class 2 sees the second packet of class 1 in service, it is implied definitely that the first packet (s_1^1) of class 1, which was in service in the previous state has completed its service. Now as the new class 2 arrival finds second packet of class 1 (s_1^2) in service, but because of memory-less property of exponential service time, we do not know that how many cycles have been completed. To make this idea more clearly, we assume that in the previous state the first packet of class 1 (s_1^1) which was in service belongs to some cycle called cycle *A*. Now as the new arrival finds the second packet of class 1(s_1^2) in service, there are many possibilities, the first possibility is that s_1^2 belongs to the same cycle *A*, in this case, only one packet of class 1 has been served and no packet has been served from queue 2 and queue 3. If s_1^2 belongs to the next cycle, for example cycle *B*, then definitely 3 packets have been served from queue 1, one packet has been served from queue 2 and one packet has been served from queue 3 as well. If s_1^2 belongs to the next cycle for example cycle *C*, then 5 packets have been served from queue 1, 2 packets have been served from queue 2 and 2 packets have been served from queue 3 during the interarrival time T_{12} in view of this round robin/polling service. Fig. 6 illustrates this transition. The maximum number of class 1 packets that can be served are i_1(if the total number of packets in queue 1 i.e. i_1 are odd), otherwise the maximum number of packets that can be served are $i_1 - 1$ (if the total number of packets in queue 1 i.e. i_1 are even), if the arriving packet is still to find a class 1 packet in service. However, j_1 includes the type 1 packet that arrived in the previous state. Hence, we have

$$j_1 = i_1 + 1 - k, \quad j_n = i_n - (\frac{k-1}{2}), n = 2,3$$

until either queue 2 and queue 3 are exhausted or only one packet (in case of odd number of packets) or two packets (in case of even number of packets) of class 1 remain in the system, the second packet (s_1^2)of class 1, being in service,

whichever occurs first. We consider queue 2 and queue 3 as a single queue and denote it as QUEUE I_2. So there are two possibilities:

(1) If $i_1 \geq I_2$ and $k = 1,3,....2i_n - 1,\ n = 2,3$ or when $i_1 < I_2$ and $k = 1,3,5,.....i_1(odd)$ or $i_1 - 1(even)$: The transition probability is:

$$P\{X_{n+1} = (i_1 - k + 1, i_2 - (\frac{k-1}{2}), i_3 - (\frac{k-1}{2}), a_2, s_1^2) \mid X_n = (i_1, i_2, i_3, a_1, s_1^1)\}$$

$= P\{k$ served from queue1, (k -1)/2 served from queue 2 and queue 3 each and 2nd packet of class1 remains in service during $T_{12}\}$

$$= \int_0^\infty \int_0^t \int_{t-x}^\infty f_{S_1^2}(s)\, f_{S_1^k + S_2^{\frac{k-1}{2}} + S_3^{\frac{k-1}{2}}}(x)\, f_{T_{12}}(t)\, ds\, dx\, dt$$

Recall that $f_{T_{12}}(t)$ denotes the probability density function for the interarrival of two packets where a type 1 packet is followed by a type 2 packet and we used the fact, that the remaining service time of a type 1 packet in service has the same exponential distribution Exp (μ_1), due to memory-less property of the Markovian service time.

(2) On the other hand, merely class 1 packets are served if queue 2 and queue 3 are exhausted. Therefore, If $i_1 \geq I_2$ and $k = 2i_n + 1.....i_1(odd)$ or $i_1 - 1(even)$, n = 2, 3 then we have:

$$P\{X_{n+1} = (i_1 + 1 - k, 0, 0, a_2, s_1^2) \mid X_n = (i_1, i_2, i_3, a_1, s_1^1)\}$$

$= P\{k$ served from type1, i_2 served from queue 2 and i_3 served from queue 3 and a type 2 packet remains in service during $T_{12}\}$

$$= \int_0^\infty \int_0^t \int_{t-x}^\infty f_{S_1^2}(s)\, f_{S_1^k + S_2^{i2} + S_3^{i3}}(x)\, f_{T_{12}}(t)\, ds\, dx\, dt$$

Similarly we can write down all possible states. The details are given in [18]

G. Limiting Distribution and QoS Parameters for CQ Model

Again, to the best of our knowledge, no previous analytical expressions are available for the waiting time of a G/M/1 queue with CQ. As in our model, during each cycle, the CQ scheduler serves 2 packets from queue 1, one packet from queue 2 and one packet from queue 3; so the scheduler serves 4 packets in total during each cycle. To make the analysis simple, we consider queue 2 and queue 3 as a single queue and call it as QUEUE 2 because the expected delay for class 2/class 3 packet will be same due to the symmetry of alternating service. It means that the scheduler will serve two packets from queue 1 and two packets from QUEUE 2 (one packet of class 2 and one packet of class 3) during

each cycle. We study queue 1 in detail. Consider the steady state distribution at the time of packet arrivals to queue 1. An arriving packet of class 1 will wait for the service completion of the one already in service plus the service times of packets in queue 1 and QUEUE 2 according to the round-robin fashion. There are two possibilities: $i_1 < I_2$ and $i_1 \geq I_2$.

a) The states (i_1, I_2, a_1, s_n^m), $m = 1,2$, n = 1,2,3 and $i_1 < I_2$

If n = 1, then the first packet of queue 1 (s_1^1) or the second packet of queue 1 (s_1^2) can be in service. Further we have to consider that the total number of packets already waiting in queue 1 including the one in service is odd or even. If the packet in service is (s_1^1) and the total number of packets waiting in queue 1 are even then the new arriving packet of class 1 waits $R_m + S_1^{i_1-1} + S_2^{\frac{i_1}{2}} + S_3^{\frac{i_1}{2}}$ time units in the queue and in case of odd number of packets in queue 1, the new arriving packet of class 1 waits $R_m + S_1^{i_1-1} + S_2^{\frac{i_1-1}{2}} + S_3^{\frac{i_1-1}{2}}$ time units in the queue. On the other side, if the packet which is already in service is (s_1^2) and the total number of packet waiting in queue 1 are even, then the new arriving packet of class 1 has to wait for $R_m + S_1^{i_1-1} + S_2^{\frac{i_1}{2}} + S_3^{\frac{i_1}{2}}$ time units and in case of odd number of packets waiting in the queue, the new arriving packet has to wait for $R_m + S_1^{i_1-1} + S_2^{\frac{i_1+1}{2}} + S_3^{\frac{i_1+1}{2}}$ Where R_m denotes the remaining service time of a packet in service which has the same exponential distribution as S_n^m. Similarly the arguments can be written corresponding to other possibilities as well.

b) The states (i_1, I_2, a_1, s_n^m), $m = 1,2$, n = 1,2,3 and $i_1 \geq I_2$

This is the case where many possibilities can occur depending on the arrival of QUEUE 2 (here class 2 and class 3 both) packets during the waiting time of the type 1 packet that has just arrived. If type 2 arrivals occur in the right periods, the arriving packet may wait maximum the amount of time as given in case a), depending on the value of m and n. For example, if i_1=5 and I_2=3 (total number of packets in queue 2 and queue 3); we know that our CQ scheduler will serve the two packets from queue 1 and 2 packets from QUEUE 2 (one class 2 and one class 3 packet) during the first cycle, however during the second cycle, the scheduler will serve packet no. 3 and 4 from queue 1 but when the third packet from QUEUE 2 goes into service, either the fifth packet in queue 1 or a packet arriving to QUEUE 2 (either class 2 or class 3) during all the service times up to that point will follow. There are other possible combinations as well and the argument goes on even longer for larger i_1. If there are no arrivals, the arriving packet has to wait

minimum $S_1^{i_1} + S_2^{I_2}$ time units. Using the minimum and maximum values, it is possible to form exact bounds on the waiting time in the queue and hence in the router. Putting cases a) and b) together, we can find exact bounds on the expected waiting time of class 1 packet as: $C_1 \le E[W_1] \le C_1^{/}$, where

$$
\begin{aligned}
C_1 = & \sum_{j_1=1}^{J_1-1} \sum_{j_2=0}^{\lfloor J_1/2 \rfloor} \sum_{j_3=0}^{\lfloor J_1/2 \rfloor} \pi(j_1, j_2, j_3, a_1, s_1^1)(1/\mu_1 + (j_1-1)/\mu_1 + j_2/\mu_2 + j_3/\mu_3) \\
& + \sum_{j_1=1}^{J_1-1} \sum_{j_2=\lceil J_1/2 \rceil}^{J_2} \sum_{j_3=\lceil J_1/2 \rceil}^{J_3} \pi(j_1, j_2, j_3, a_1, s_1^1)(1/\mu_1 + (j_1-1)/\mu_1 + \lfloor j_1/2 \rfloor/\mu_2 + \lfloor j_1/2 \rfloor/\mu_3) \\
& + \sum_{j_1=1}^{J_1-1} \sum_{j_2=0}^{\lfloor J_1/2 \rfloor} \sum_{j_3=0}^{\lfloor J_1/2 \rfloor} \pi(j_1, j_2, j_3, a_1, s_1^2)(1/\mu_1 + (j_1-1)/\mu_1 + j_2/\mu_2 + j_3/\mu_3) \\
& \sum_{j_1=1}^{J_1-1} \sum_{j_2=\lceil J_1/2 \rceil}^{J_2} \sum_{j_3=\lceil J_1/2 \rceil}^{J_3} \pi(j_1, j_2, j_3, a_1, s_1^2)(1/\mu_1 + (j_1-1)/\mu_1 + \lceil j_1/2 \rceil/\mu_2 + \lceil j_1/2 \rceil/\mu_3) \\
& \sum_{j_1=0}^{J_1-1} \sum_{j_2=1}^{\lfloor J_1/2 \rfloor} \sum_{j_3=0}^{\lfloor J_1/2 \rfloor} \pi(j_1, j_2, j_3, a_1, s_2)(1/\mu_2 + j_1/\mu_1 + (j_2-1)/\mu_2 + j_3/\mu_3) \\
& \sum_{j_1=0}^{J_1-1} \sum_{j_2=\lceil J_1/2 \rceil}^{J_2} \sum_{j_3=\lceil J_1/2 \rceil}^{J_3} \pi(j_1, j_2, j_3, a_1, s_2)(1/\mu_2 + j_1/\mu_1 + \lfloor j_1/2 \rfloor/\mu_2 + (\lfloor j_1/2 \rfloor+1)/\mu_3) \\
& \sum_{j_1=0}^{J_1-1} \sum_{j_2=0}^{\lfloor J_1/2 \rfloor} \sum_{j_3=1}^{\lfloor J_1/2 \rfloor} \pi(j_1, j_2, j_3, a_1, s_3)(1/\mu_3 + j_1/\mu_1 + j_2/\mu_2 + (j_3-1)/\mu_3) \\
& \sum_{j_1=0}^{J_1-1} \sum_{j_2=\lceil J_1/2 \rceil}^{J_2} \sum_{j_3=\lceil J_1/2 \rceil}^{J_3} \pi(j_1, j_2, j_3, a_1, s_3)(1/\mu_3 + j_1/\mu_1 + \lfloor j_1/2 \rfloor/\mu_2 + \lfloor j_1/2 \rfloor/\mu_3)
\end{aligned}
$$

and

$$
\begin{aligned}
C_1^{/} = & \sum_{j_1=1}^{J_1-1} \sum_{i_2=0}^{J_2} \sum_{j_3=0}^{J_3} \pi(j_1, j_2, j_3, a_1, s_1^1)(1/\mu_1 + (j_1-1)/\mu_1 + \lfloor j_1/2 \rfloor/\mu_2 + \lfloor j_1/2 \rfloor/\mu_3) \\
& + \sum_{j_1=1}^{J_1-1} \sum_{j_2=0}^{J_2} \sum_{j_3=0}^{J_3} \pi(j_1, j_2, j_3, a_1, s_1^2)(1/\mu_1 + (j_1-1)/\mu_1 + \lceil j_1/2 \rceil/\mu_2 + \lceil j_1/2 \rceil/\mu_3) \\
& + \sum_{j_1=0}^{J_1-1} \sum_{j_2=1}^{J_2} \sum_{j_3=0}^{J_3} \pi(j_1, j_2, j_3, a_1, s_2)(1/\mu_2 + j_1/\mu_1 + \lfloor j_1/2 \rfloor/\mu_2 + (\lfloor j_1/2 \rfloor+1)/\mu_3) \\
& \sum_{j_1=0}^{J_1-1} \sum_{j_2=0}^{J_2} \sum_{j_3=1}^{J_3} \pi(j_1, j_2, j_3, a_1, s_3)(1/\mu_3 + j_1/\mu_1 + \lfloor j_1/2 \rfloor/\mu_2 + \lfloor j_1/2 \rfloor/\mu_3)
\end{aligned}
$$

The expected delay for class 2 and class 3 packets is same because of the symmetry of alternating service. Similarly by following the above procedure we can write down the exact bounds on expected waiting time of class 2 (same for class 3) packet as: $C_2 < E[W_2] = C_2^{/}$, where

$$C_2 = \sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{J_3}\sum_{j_1=1}^{2j_2+2} \pi(j_1,j_2,j_3,a_2,s_1^1)(1/\mu_1 + (j_1-1)/\mu_1 + j_2/\mu_2 + j_2/\mu_3)$$
$$+\sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{J_3}\sum_{j_1=2j_2+3}^{J_1} \pi(j_1,j_2,j_3,a_2,s_1^1)(1/\mu_1 + (2j_2+1)/\mu_1 + j_2/\mu_2 + j_2/\mu_3)$$
$$+\sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{J_3}\sum_{j_1=1}^{2j_2+1} \pi(j_1,j_2,j_3,a_2,s_1^2)(1/\mu_1 + (j_1-1)/\mu_1 + j_2/\mu_2 + j_2/\mu_3)$$
$$\sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{J_3}\sum_{j_1=2j_2+2}^{J_1} \pi(j_1,j_2,j_3,a_2,s_1^2)(1/\mu_1 + 2j_2/\mu_1 + j_2/\mu_2 + j_2/\mu_3)$$
$$\sum_{j_2=1}^{J_2-1}\sum_{j_3=0}^{J_3}\sum_{j_1=0}^{2j_2} \pi(j_1,j_2,j_3,a_2,s_2)(1/\mu_2 + j_1/\mu_1 + (j_2-1)/\mu_2 + j_2/\mu_3)$$
$$\sum_{j_2=1}^{J_2-1}\sum_{j_3=0}^{J_3}\sum_{j_1=2j_2+1}^{J_1} \pi(j_1,j_2,j_3,a_2,s_2)(1/\mu_2 + (j_2-1)/\mu_2 + j_2/\mu_3 + 2j_2/\mu_1)$$
$$\sum_{j_2=0}^{J_2-1}\sum_{j_3=1}^{J_3}\sum_{j_1=0}^{2j_2} \pi(j_1,j_2,j_3,a_2,s_3)(1/\mu_3 + j_2/\mu_2 + j_2/\mu_3 + j_1/\mu_1)$$
$$\sum_{j_2=0}^{J_2-1}\sum_{j_3=1}^{J_3}\sum_{j_1=2j_2+1}^{J_1} \pi(j_1,j_2,j_3,a_2,s_3)(1/\mu_3 + j_2/\mu_2 + j_2/\mu_3 + (2j_2+2)/\mu_1)$$

and

$$C_2' = \sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{J_3} \pi(j_1,j_2,j_3,a_2,s_1^1)(1/\mu_1 + j_2/\mu_2 + (2j_2+1)/\mu_1 + j_2/\mu_3)$$
$$+\sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=0}^{J_3} \pi(j_1,j_2,j_3,a_2,s_1^2)(1/\mu_1 + j_2/\mu_2 + 2j_2/\mu_1 + j_2/\mu_3)$$
$$+\sum_{j_1=0}^{J_1}\sum_{j_2=1}^{J_2-1}\sum_{j_3=0}^{J_3} \pi(j_1,j_2,j_3,a_2,s_2)(1/\mu_2 + (j_2-1)/\mu_2 + j_2/\mu_3 + 2j_2/\mu_1)$$
$$\sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2-1}\sum_{j_3=1}^{J_3} \pi(j_1,j_2,j_3,a_2,s_3)(1/\mu_3 + j_2/\mu_2 + j_2/\mu_3 + (2j_2+2)/\mu_1)$$

V. SIMULATION AND TEST BED RESULTS

In this section, we present the simulation framework and test bed setup along with a comparison of simulation and test bed results.

A. Simulation Framework

A comprehensive discrete-event simulator for queueing systems was built to understand and evaluate the QoS behaviour of self-similar traffic. The simulation engine is highly modular by design allowing free customization of the traffic generator and the scheduling logic. This allows for the ready evaluation of any scheduling discipline under any

specific kind of input traffic. The key element for the scheduler logic is the `Scheduler` class. Here we used the template method design pattern [20]. This allows any scheduling algorithm to be loosely coupled but easily integrated, overriding the existing program skeleton. `PriorityScheduler, LLQScheduler and CQScheduler` were actually implemented to analyze the corresponding QoS behaviour.

A traffic generator was also written, which implements the traffic model described in Section III. This generator may also be readily over-ridden by another traffic model.

A number of other associated classes were written to facilitate program function and accuracy. These include:

- `Simulation`. This class served as the simulation engine – moving time forward and updating the event list etc.
- `RandomNumber`. A class for generating random number with specific distributions including: uniform, exponential, Poisson, Compound-Poisson and Pareto.
- `Packet`. A class used to store the system state as encountered by each packet.

The QoS results from the simulation studies with 95% confidence interval are presented. Gross et al. study a related issue in detail in [20] and conclude that care must be taken in simulations involving Pareto distributions as they can lead to large errors due to the heavy tail. It should also be noted though, that the bulk of empirical evidence [21-25] suggests that H ~ [0.7, 0.85] is the region of interest in network traffic. Fig. 7 and 8 show Queue Length vs Hurst Parameter and Packet Loss Rate (PLR) vs Hurst Parameter respectively for PQ, LLQ and CQ models. We can see the significant detrimental impact of increasing the Hurst Parameter (the degree of self-similarity) on the QoS offered. We can also note the characteristic of a PQ system, LLQ system and CQ system: as load increases, we see a significant increase in the Packet Loss Rate and Queue length of the lower priority queues.

B. Test Bed Description

In this subsection, we describe the interim results of the IP QoS tests running non-preemptive PQ, LLQ and CQ scheduling on a Cisco Modular Router 1841 and present a comparison with the simulation results.

A Cisco 1841 Modular Router with Cisco QoS features running Cisco IOS 12.4 was connected to two Linux workstations through dedicated 100 Mbps Ethernet links as shown in Fig. 9. We implemented a traffic generator on the Sender workstation, which simultaneously generated three different self-similar traffic streams over UDP. We

implemented three sinks SINK1, SINK2 and SINK3 on the Receiver workstation to receive the three different classes of traffic on different ports. We first implemented Priority Queueing in the Router to determine the queueing delays for corresponding traffic classes. Each output queue had a capacity of 10 packets and packet arrivals occurred according to the process described in Section III. For the higher priority class (class 1 packets), we set the session arrival rate to $\lambda_1 = 6s^{-1}$, the in-session packet arrival rate to $\alpha_1 = 50s^{-1}$ (the characteristic of VoIP traffic) and the service rate to $\mu_1 = 2500s^{-1}$. For the low priority queues (both queue 2 and queue 3), we set the session arrival rate $\lambda_2 = \lambda_3 = 50s^{-1}$, the in-session packet arrival rate to $\alpha_2 = \alpha_3 = 6s^{-1}$ and the service rate to $\mu_2 = \mu_3 = \mu_1$. We investigated the effects of varying the Hurst parameter (0.5<H<1) on various QoS parameters.

C. Cisco 1841 Router Configuration

We first implemented Priority Queueing in a Cisco Modular Router 1841 to provide differential treatment to the different classes of self-similar traffic. Priority Queueing's most distinctive feature is its scheduler. It supports a maximum of four queues: High, Medium, Normal and Low. If the High queue always has a packet waiting, the scheduler will always serve the packets from this queue. On the other hand, if the High queue does not have a packet waiting, but the Medium queue does, one packet is taken from the Medium queue – and then the process starts over at the High queue. The low queue only gets service if the High, Medium, and Normal queues do not have any packets waiting [19]. Any number of queues out of four can be configured on an interface; the scheduler simply serves these configured queues and skips others. As we have three kinds of traffic, we configured three queues; High, Medium and Normal at the output interface Fa0/1. As shown in Fig. 9, there are two interfaces Fa0/0 (input interface) and Fa0/1 (output interface). We need to classify different kinds of traffic at the input interface and assign them to the proper queue at the output interface on the basis of destination port number. We briefly cover the configuration steps here:

We defined the priority list, classified the traffic at input interface (Fa0/0) and assigned them to the proper queue at the output interface (Fa0/1). Next we specified the maximum size of each queue at the output interface before assigning the priority list 1 to the output interface (Fa0/1)**.** Further, for the verification of LLQ and CQ models, we implemented LLQ and then CQ on Cisco modular router. The reader is further referred to [19] for the details of PQ, LLQ and CQ configuration on Cisco routers.

D. Time Synchronization between Sending and Receiving Machine

In order to obtain an accurate measure of the one-way delay through the network, the clocks on the sending and receiving machines had to be synchronized. Network Time Protocol (NTP) [26] was used for this purpose, as it meets our accuracy requirements and there are numerous readily available implementations. To have an accurate time synchronization between the sending and receiving machine's clocks and not to interrupt with the self-similar traffic passing through the router, we used dedicated Ethernet ports over a cross-over cable for the NTP connection. We assigned an IP address 173.16.10.1 to the sending machine's ethernet card and an IP address 173.16.20.1 to the receiving machine's ethernet card as detailed in Figure 9. An NTP primary server, or stratum 1, was connected to a high precision reference clock and equipped with NTP software. Other computers (stratum 2s), equipped with similar software automatically queried the primary server to synchronize their system clocks. We made the sending machine as the NTP primary server in our network. The NTP primary server was connected to a high precision reference clock (au.pool.ntp.org) to synchronize its system's clock. Further, to achieve real time synchronization between the sender and receiver's clocks, a small program was written, to enable NTP to run as a background process.

E. Measurement of Queueing Delay for Multiple Classes of Self-Similar Traffic

All packets in a network experience delay from when the packet is first transmitted to when it arrives at its destination. Fig. 9 shows the different kinds of delay a packet experiences from source to destination. We explain them here, briefly:

(1) Serialization Delay: is the time it takes to encode the bits of a packet on to the physical interface and can be calculated by dividing the number of bits sent by link speed.

(2) Propagation Delay: is the time it takes a single bit to get from one end of the link to the other and can be calculated by using the formula: $\frac{linklength}{2.1\times10^{8} m/s}$

(3) Processing Delay: refers to the time taken by the router to examine the packet at the input interface and placing it in the output queue on the output interface

(4) Queueing Delay: consists of time spent in the queues inside the router—typically just in output queues in a router.

(5) Transmission Delay: is the delay that the scheduler takes to put the packet from output queue on to the link; it is same as serialization delay [19].

In our delay calculations, we can ignore the processing delay inside the input interface of the router and at the receiving machine as this is in order of few microseconds, several orders of magnitude smaller than the expected delay. The propagation delay through the network is also negligible and therefore ignored. Compensating for the serialization delay at the sending machine and transmission delay at the output interface of the router, we found the following queueing delay for the three different classes of self-similar traffic in our test bed experiments. Fig. 10, 11 and 12 show the mean delay for PQ, LLQ and CQ models respectively, in which the test bed results have been plotted with 95% confidence interval against simulation results (Refer to Table 1).

We see the significant detrimental impact of increasing the Hurst parameter (the degree of self-similarity) on the QoS offered. We also note the characteristics of PQ, LLQ and CQ systems: as the load increases, we see a significant increase in the delay for the lower priority queues, especially queue 3. The slight difference between test bed and simulation results is likely due to congestion at the NIC of the Receiver workstation, particularly when self-similarity increases.

VI. RELATED WORK

In this section, we give an overview of related work. During the last ten years, substantial work has been done to evaluate the performance of communication networks in the presence of self-similar traffic [27-36]. The offered queueing based results lack the capability of offering differential treatment to multiple classes of input traffic because majority of the analysis is based on FIFO scheduling and further the results are asymptotic. Further, we can notice that there has emerged a paradigm shift towards IP based solutions for wireless networking to support real time multimedia applications over mobile devices [37-41]. Because of this, researchers have recently focused on understanding the nature of wireless IP traffic and it has been shown that wireless data traffic also exhibits self-similarity and long-range dependency [42-50]. Here we just discuss the most relevant work. In [42], a FBM/D/1 queueing system has been used to analyze the performance of GGSN while taking into account self-similar input. The submitted approach enabled the determination of different probabilistic and time characteristics: upper and lower bounds of the GGSN service rate, the average queue length in the server buffer and average service time of information units. A QoS framework for heavy-

tailed traffic over the wireless Internet is proposed in [43]. A simulation study that has been conducted to analyze the performance of the Foreground-Background scheduler and Round-Robin (RR) scheduler and the resulting insight shows that a FB scheduler requires much less network resources to attain a given QoS. There are no analytical proofs of the simulation results. The aggregated connectionless traffic is modeled with Fractional Brownian Motion (FBM) in [44]. This study indicates three major contributions (1) characterization of connectionless traffic, (2) bandwidth allocation formula and (3) short-term traffic prediction. An aggregated traffic model for UMTS is presented in [45]. The key idea is based on customizing the batch Markovian Arrival Process (BMAP) such that different packet sizes of IP packets are represented by rewards. Modeling and simulation of the Cellular Digital Packet Data (CDPD) network of Telus Mobility (a commercial service provider) are performed by using the OPNET tool in [46]. The trace-driven simulations with genuine traffic trace exhibiting long-range dependent behaviour are used to evaluate the performance of the CDPD protocol. The results indicate that genuine traffic traces, compared to traditional traffic models such as Poisson models, produce longer queues. The area of wireless IP traffic modeling is still immature and the majority of the analysis [42-50] is merely based on the characterization of wireless traffic. Further, still much of the current understanding of wireless IP traffic modeling is based on Poisson models which can yield misleading results and hence poor network planning. To address the issue of providing differential and guaranteed treatment to multiple traffic classes with different QoS demands in a realistic way, there is a need to accurately determine end-to-end QoS parameters such as delay, jitter, throughput, packet loss, availability and per-flow sequence preservation.

VII. APPLICATIONS OF THE MODEL

Here we briefly present the prime applications of the models. With the tremendous growth in data traffic, the telecommunication industry is evolving its core networks towards IP technology. An all-IP DiffServ model is widely considered to be the most promising architecture for guaranteed QoS provisioning in NG wireless networks. This is largely due to its scalability, mobility support and the ability to inter-network heterogeneous radio access networks [3-4]. To transport UMTS services through IP networks without loosing end-to-end QoS provisioning, an accurate and consistent QoS mapping is required. According to 3GPP, UMTS-to-IP QoS mapping is performed by a translation function in the GGSN router that classifies each UMTS packet flow and maps it to a suitable IP QoS class [12]. Being able to accurately model the end-to-end behaviour of different classes of IP traffic (conversational, streaming,

interactive and background) passing through a DiffServ domain is essential to the guaranteed delivery of various QoS parameters. Several queueing tools have been developed that can be implemented in IP routers within different QoS domains including Priority Queueing (PQ), Custom Queueing (CQ), Weighted Fair Queueing (WFQ), Class Based Weighted Fair Queueing (CBWFQ) and Low-Latency Queueing (LLQ) [19]. In this paper, we have specifically considered the QoS behaviour of PQ, LLQ and CQ service disciplines. Work on the other tools is ongoing. Our models are directly applicable to the problem of determining the end-to-end queueing behavior of IP traffic through both Wired and wireless IP domains. Modeling accuracy is most crucial though, in resource-constrained environments such as wireless networks. For example, our model is directly able to analyze the behaviour of different QoS classes of UMTS traffic (which have been proven statistically self-similar and long-range dependent) passing through a DiffServ domain, in which the routers implement a particular queueing and scheduling combination. The models enable tighter bounds on actual behaviour so that over-provisioning can be minimized. It also enables translations of traffic behaviour between different kinds of QoS domains so that it is possible to map reservations made in different domains to provide session continuity. We have jointly considered traffic engineering and QoS issues. The fundamental themes of this study span traffic modeling, stochastic analysis and network design. It also provides significant insight and guidance for the design of NG-IP based networks.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we have contributed to the accurate modeling of wireless IP traffic behaviour, by presenting novel analytical models based on a G/M/1 queueing system under different classes of self-similar input traffic. We have analyzed it on the basis of non-preemptive PQ, LLQ and CQ scheduling schemes and derived explicit expressions for the expected waiting time and packet loss rate for multiple classes. We have also performed an extensive simulation study along with actual test-bed experiments to validate the accuracy of our models. The present study can be used as a guide for the efficient allocation of buffer space and bandwidth for individual traffic classes – with the aim of guaranteeing the QoS required by different applications while minimizing excessive allocation. Further, the model represents an important step towards the overall aim of understanding realistic (under self-similar traffic) end-to-end QoS behaviour (in terms of QoS parameters such as delay, jitter and throughput) of multiple traffic classes passing through heterogeneous wireless IP domains (IntServ, DiffServ and MPLS). Our future work will focus on determining

end-to-end QoS parameters by conducting simulation and experimental studies over different DiffServ domains implemented with different queueing and scheduling combinations.

ACKNOWLEDGEMENTS

The authors would also like to thank Khalid Hameed and Adeel Baig for their great help in the test bed implementation.

REFERENCES

[1] W. Stallings, "Integrated Services Architecture: The Next Generation Internet", International Journal of Network Management , 9 1999, pp. 38-43

[2] S. Blake et al. "An Architecture for Differentiated Services", IETF RFC 2475

[3] B. Moon and H. Aghvami, "DiffServ Extension for QoS Provisioning in IP Mobility Environments, " IEEE Wireless Comm. Vol. 10. no. 5, Oct 2003, pp. 38-44

[4] Y. Cheng and W. Zhuang, "DiffServ Resource Allocation for Fast Handoff in Wireless Mobile Internet, "IEEE Comm. Mag. Vol. 40, no. 5, May 2002, pp. 130-136

[5] Y. Cheng et al, "Efficient Resource Allocation for China's 3G/4G Wireless Networks", IEEE Communication Magazine, 2005. pp. 76-83

[6] R. Chakravorty, J. Cartwright and I. Pratt, "Practical Experience with TCP over GPRS", *in IEEE GlobeCom*, Nov. 2002

[7] D. Schwab and R. Bunt, "Characterizing the use of a Campus Wireless Network", *in IEEE INFOCOM*, March 2004

[8] X. Meng, S. Wong, Y. Yuan and S.Lu, "Characterizing Flows in Large Wireless Data Networks", *in ACM Mobicom*, September 2004

[9] A. Balachandran, G. M. Voelker, P. Bahl and P. Venkat Rangan, "Characterizing user behavior and network performance in a public Wireless LAN", *Sigmetrics Performance Evaluation. Review*, vol. 30. no. 1, 2002, pp. 195-205

[10] 3GPP, Universal Mobile Telecommunication System (UMTS); QoS Concepts and Architecture" TS23.107V6, March 2004

[11] G. Armitage, "Quality of Service in IP Networks", MTP, 2004, pp. 105-138

[12] R. Ben Ali and Y Lemieus, "UMTS-to-IP QoS Mapping for Voice and Video Telephony Service" IEEE Network, March/April 2005, pp. 26-32

[13] M. Iftikhar, B. Landfeldt and M. Caglar, "Traffic Engineering and QoS Control Between Wireless DiffServ Domains Using PQ and LLQ", in proc. of IEEE/ACM Mobiwac 07, 22nd Oct. 2007, Chania, Crete Island, Greece

[14] M. Caglar, "A Long-Range Dependant Workload Model for Packet Data Traffic", *Mathematics of Operations Research*, 29, 2004, pp. 92-105

[15] I. Kaj, "Limiting fractal random processes in heavy-tailed systems", In Fractals in Engineering, New Trends in Theory and Applications, Eds.J. Levy-Lehel, E. Lutton, *Springer-Verlag London,* 2005, pp. 199-218

[16] M. Iftikhar, T. Singh, B. Landfeldt and M. Caglar, "Multiclass G/M/1 Queueing System with Self-Similar Input and Non-Preemptive Priority, " to appear in Journal of Computer Communications 2008

[17] E.Cinlar, “Introduction to Stochastic Processes”, 1975, pp. 178

[18] M. Iftikhar and B. Landfeldt, “Markov Chain Formulation of G/M/1 Queueing System with Multiple Classes of Self-Similar Input on the basis of PQ, CQ and LLQ Service Disciplines, Technical Report Submitted, Nov. 2007, School of IT, University of Sydney

[19] W. Odom and M. J. Cavanaugh, “IP Telephony Self-Study Cisco DQoS Exam Certification Guide”, Cisco Press, 2004, pp. 3-314.

[20] D. Gross, J. Shortle, M. Fischer and D. Masi, “Difficulties in Simulating Queues with Pareto Service”, Proceedings of the 2002 Winter Simulation Conference, 2002

[21] W. Leland, M. Taqqu, W. Willinger and D. Wilson, “On the self-similar nature of Ethernet traffic (extended version)”, *IEEE/ACM Transactions on Networking*, vol. 2. no. 1, pp. 1-15, Feb. 1994.

[22] M. Crovella and A. Bestavros, “Explaining World Wide Web Traffic Self-Similarity”, *Tech. Rep. TR-95-015, Boston University, CS Dept, Boston*, MA 02215, Aug. 1995

[23] M. W. Garrett and W. Willinger, “Analysis, Modeling and Generation of Self-Similar VBR Video Traffic”, *ACM Computer Communication Review*, vol. 24, Oct. 1994, SIGCOMM 94 Symposium

[24] Kihong Park, Gi Tae Kim and Mark E. Crovella, “On the relationship between file sizes, transport protocols and self-similar network traffic”, in proc. of the International Conference on Network Protocols, pp. 171-180, Oct. 1996

[25] T. Tuan and K. Park, “Performance Evaluation of Multiple time scale TCP under self-similar traffic conditions”, Technical Report CSD-TR-99-040, Department of Computer Sciences, Purdue University, 1999. http://citeseer.ist.psu.edu/article/tuan99performance.html

[26] D. L. Mills, “Simple network time protocol (SNTP) version 4 for IPv4, IPv6 and OSI,” RFC 2030, IETF, Oct. 1996. http://www.ietf.org/rfc/rfc2030.txt

[27] Y. Zhou and H. Sethu, “Performance of shared output queueing in ATM switches under self-similar traffic,” *in Proc. of Applied Telecommunication Symposium*, Washington, D.C., USA, April 16-20, 2000

[28] A. Erramilli, O. Narayan and W. Willinger, “Experimental queueing analysis with long-range dependent packet traffic,” *IEEE/ACM Transactions on Networking*, vol. 4, no. 2, pp. 209-223, April 1996

[29] M. Zukerman et al, “Analytical Performance Evaluation of a Two Class DiffServ link”, *IEEE ICS*, 25-28 Nov. 2002, vol. 1, pp. 373-377

[30] S. Kasahara, “Internet traffic modeling: A Markovian approach to self-similar traffic and prediction of loss probability for finite queues,” *IEICE Transactions on Communications: Special Issue on Internet Technology*, vol. E84-B, no. 8, pp. 2134-2141, August 2001

[31] H. Yousefi’zadeh, “A neural-based technique for estimating self-similar traffic average queueing delay,” *IEEE Communications Letters*, 6 (10), pp. 429-421, 2002

[32] J. M. Chung, Z. Quan, “Impact of Self-Similarity on Performance Evaluation in DiffServ Networks”, *IEEE MWSCAS*, 4-7 Aug. 2002, vol. 2, pp. 326-329

[33] B. Tsybakov and N. D. Georganas, “Self-Similar traffic and upper bounds to buffer overflow in ATM queue”, *Performance Evaluation*, 36, 1998, pp. 57-80

[34] A. Adas and A. Mukherjee, "On Resource Management and QoS guarantees for long-range dependant traffic", *In Proc IEEE INFOCOM*, 1995, pp. 779-787

[35] M. Parulekar and A. Makowski, "Tail Probabilities for a Multiplexer with self-similar input", *In proc IEEE INFOCOM*, 1996, pp. 1452-1459

[36] I. Norros, "A Storage Model with self-similar input", *Queueing System*, 16, 1994, pp. 387-396

[37] J. Yang and I. Kriaras, "Migration to all-IP based UMTS networks", IEEE 1st International Conference on 3G Mobile Communications Technologies, 27-29 March, 2000, pp. 19-23

[38] P. Newman, Netillion Inc. "In Search of the All-IP Mobile Network", *IEEE Communication Magazine*, vol. 42, issue 12, Dec. 2004, pp. S3-S8

[39] G. Araniti, F. Calabro, A. Iera, A. Molinaro and S. Pulitano, "Differentiated Services QoS Issues in Next Generation Radio Access Network: a New Management Policy for Expedited Forwarding Per-Hop Behavior", *IEEE Vehicular Technology Conference*, VTC 2004-Fall, vol. 4, 26-29 Sept. 2004, pp. 2693-2697

[40] S. Uskela, "All IP Architectures for Cellular Networks", *2nd International Conference on 3G Mobile Communication Technologies*, 26-28 March 2001, pp. 180-185

[41] Jeong-Hyun Park, "Wireless Internet Access for Mobile Subscribers Based on GPRS/UMTS Network" *IEEE Communication Magazine*, vol. 40, issue 4, April 2002, pp. 38-39

[42] Y. Koucheryavy, A. Krednzel, S. Lopatin and J. Harju, "Performance estimation of UMTS release 5 IM-subsystem elements," *4th International Workshop on Mobile and Wireless Communication Networks, IEEE MWCN*, pp. 35-39, 9-11, September, 2002

[43] Z. Shao and U. Madhow, "A QoS framework for heavy-tailed traffic over the wireless Internet," *in proc. of MILCOM 2002*, vol. 2, pp. 1201-1205, 7-10 Oct. 2002

[44] I. Norros, "The management of large flows of connectionless traffic on the basis of self-similar modeling," IEEE International Conference on Communications, vol. 1, pp. 451-455, 18-22 June, 1995

[45] A. Klemn, C. Lindemann and M. Lohmann, "Traffic modeling and characterization for UMTS networks," IEEE Globecom, vol. 3, pp. 1741-1746. 25-29. Nov. 2001

[46] M. Jiang, M. Nikolic, S. Hardy and L. Trajkovic, "Impact of Self-Similarity on Wireless Data Network Performance", *IEEE ICC*, 2001, vol. 2, pp. 477-481

[47] J. Ridoux, A. Nucci and D. Veitch, "Characterization of Wireless Traffic based on Semi-Experiments", *Technical Report-LIP6*, December 2005

[48] Z. Sahinoglu and S. Tekinay, "On Multimedia Networks: Self-Similar Traffic and Network Performance", *IEEE Communication Magazine*, vol. 37, issue 1, Jan. 1999, pp. 48-52

[49] I. Norros, "On the use of Fractional Brownian Motion in theory of connectionless networks", *IEEE Journal on Selected Areas in Communications*, vol. 13. no. 6, August 1995, pp. 953-962

[50] P. Benko, G. Malicsko and A. Veres, "A Large-scale, passive analysis of end-to-end TCP Performances over GPRS", *in IEEE INFOCOM*, March 2004

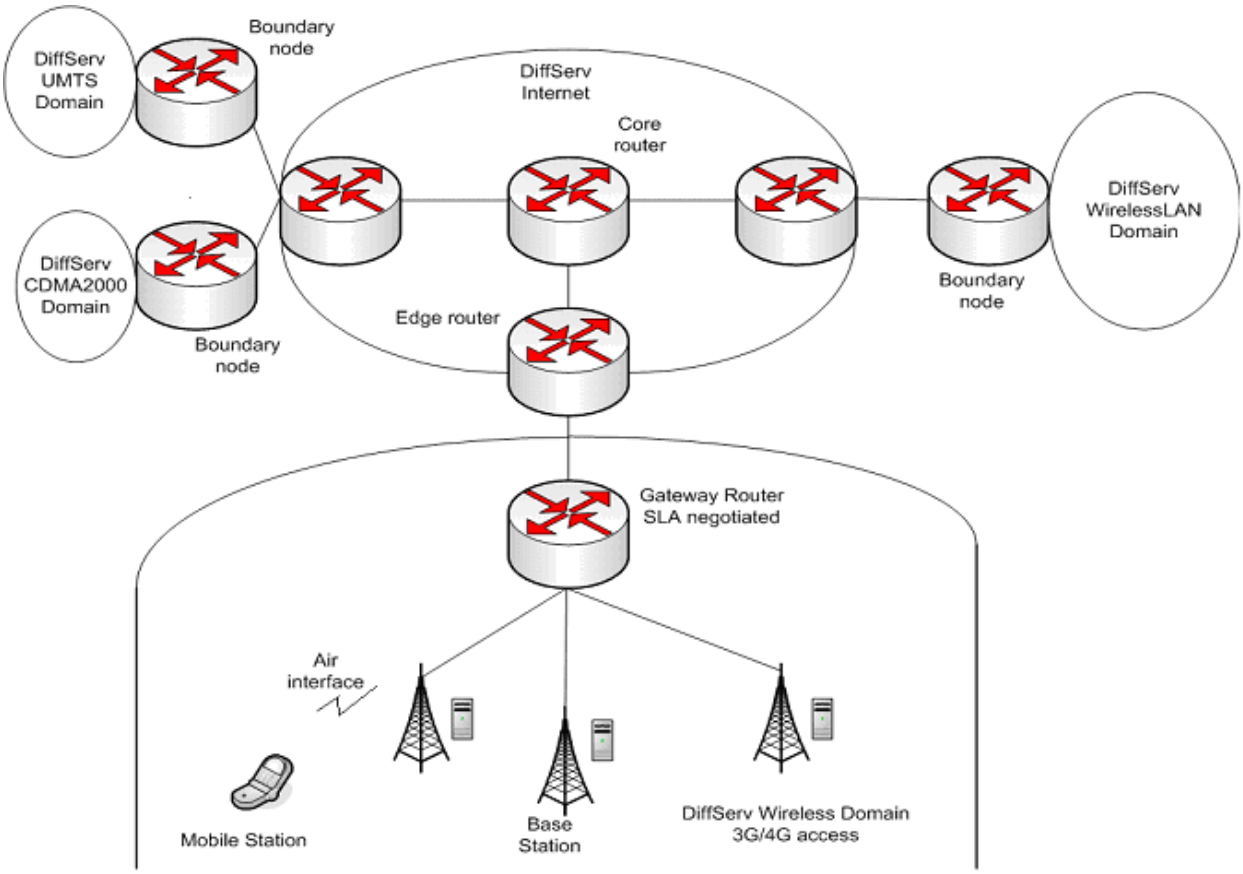


Fig. 1: An All IP DiffServ Architecture for 3G/4G Wireless Communications

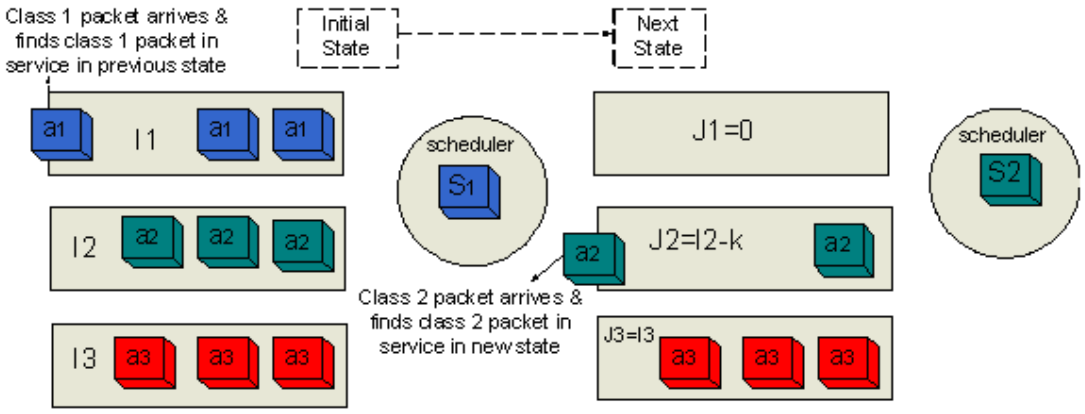


Fig. 2 An Example of Markov Chain Transition for PQ from $(i_1, i_2, i_3, a_1, s_1) \rightarrow (j_1, j_2, j_3, a_2, s_2)$

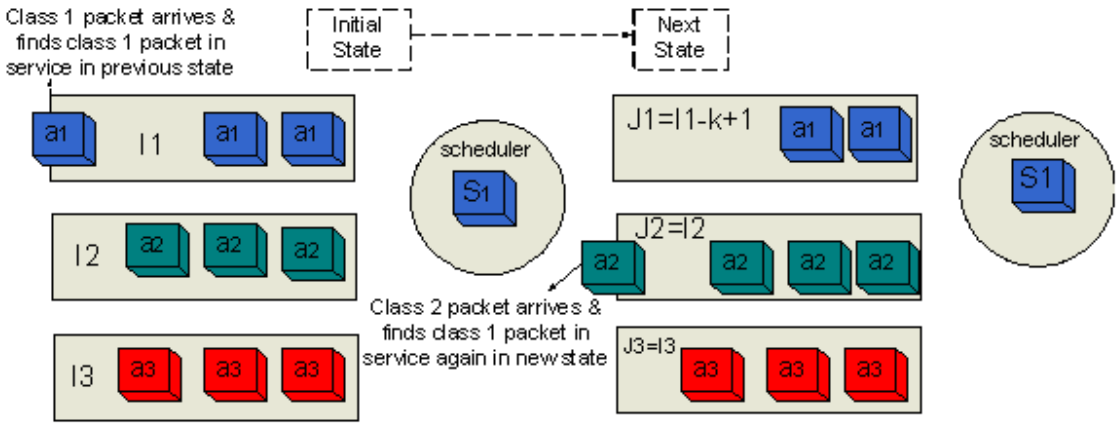


Fig. 3: An Example of Markov Chain Transition for PQ from $(i_1, i_2, i_3, a_1, s_1) \rightarrow (j_1, j_2, j_3, a_2, s_1)$

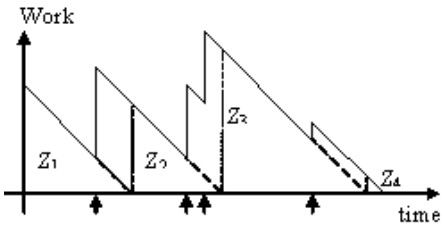


Fig. 4: Waiting time of a type 2 packet in terms of Zj's

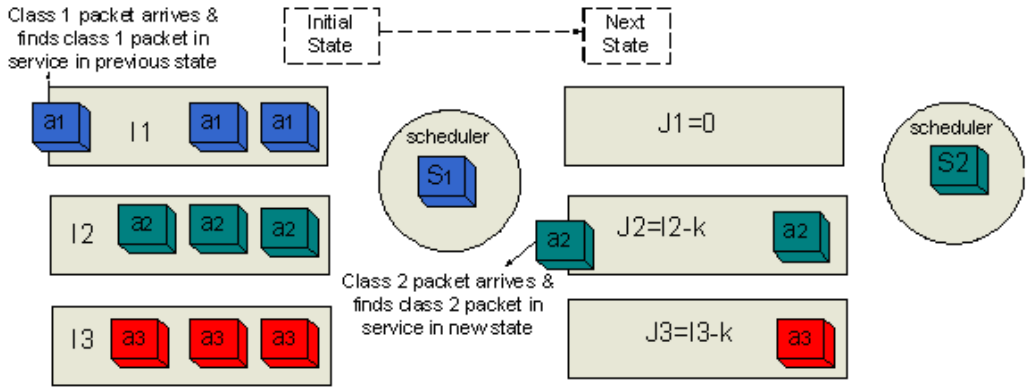
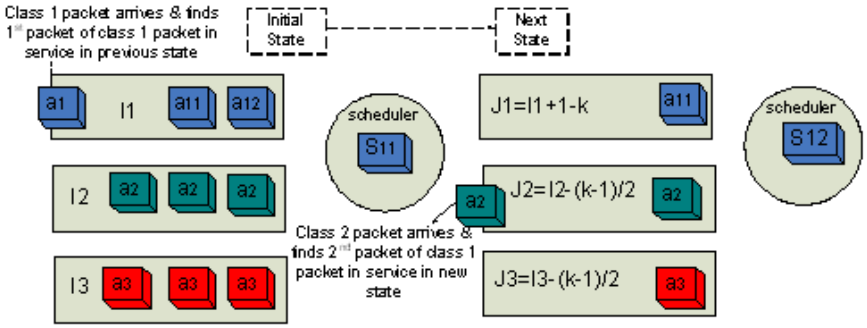
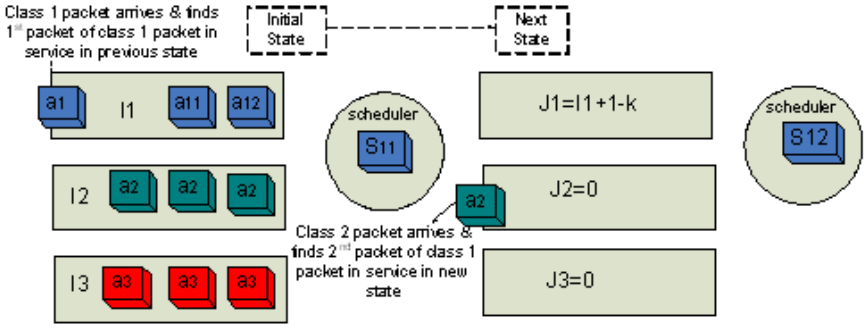


Fig. 5: An Example of Markov Chain Transition for LLQ from

$$(i_1, i_2, i_3, a_1, s_1) \rightarrow (j_1, j_2, j_3, a_2, s_2)$$



Possibility (a) when $i_1 < I_2, I_2 = (i_2 + i_3)$



Last Possibility when $i_1 \geq I_2, I_2 = (i_2 + i_3)$

Fig. 6: An Example of Markov Chain Transition for CQ from

$$(i_1, i_2, i_3, a_1, s_1^1) \rightarrow (j_1, j_2, j_3, a_2, s_1^2)$$

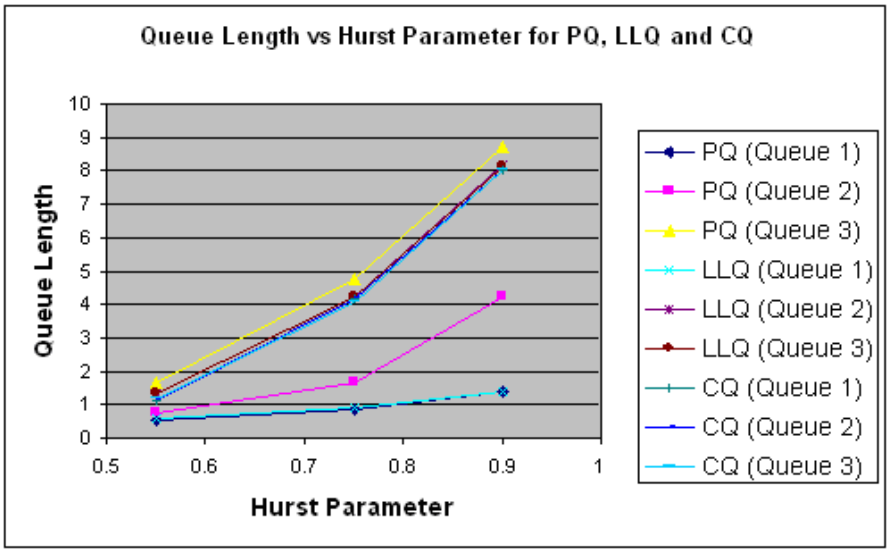


Fig. 7: Queue Length vs Hurst Parameter: Simulation Results for PQ, LLQ and CQ

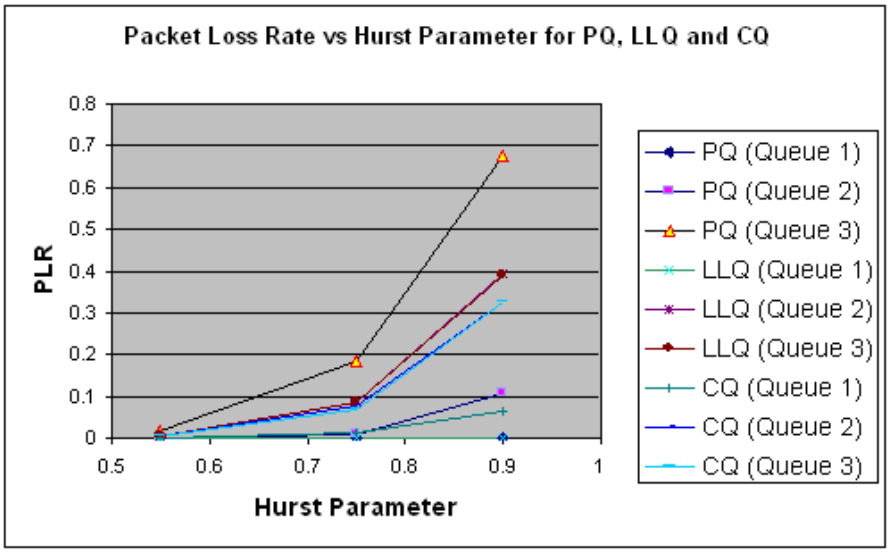


Fig. 8: Packet Loss Rate vs Hurst Parameter: Simulation Results for PQ, LLQ and CQ

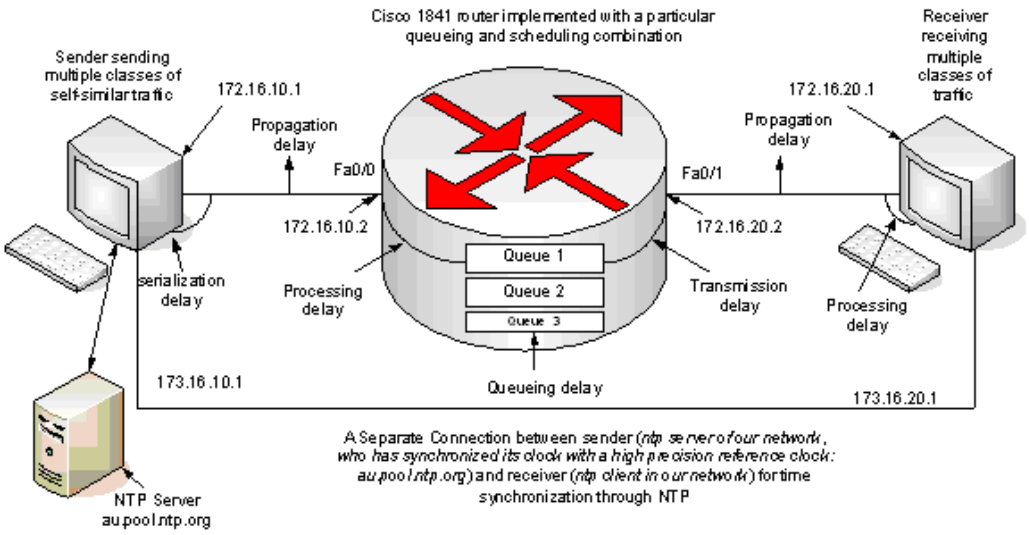


Fig. 9: Test Bed Setup

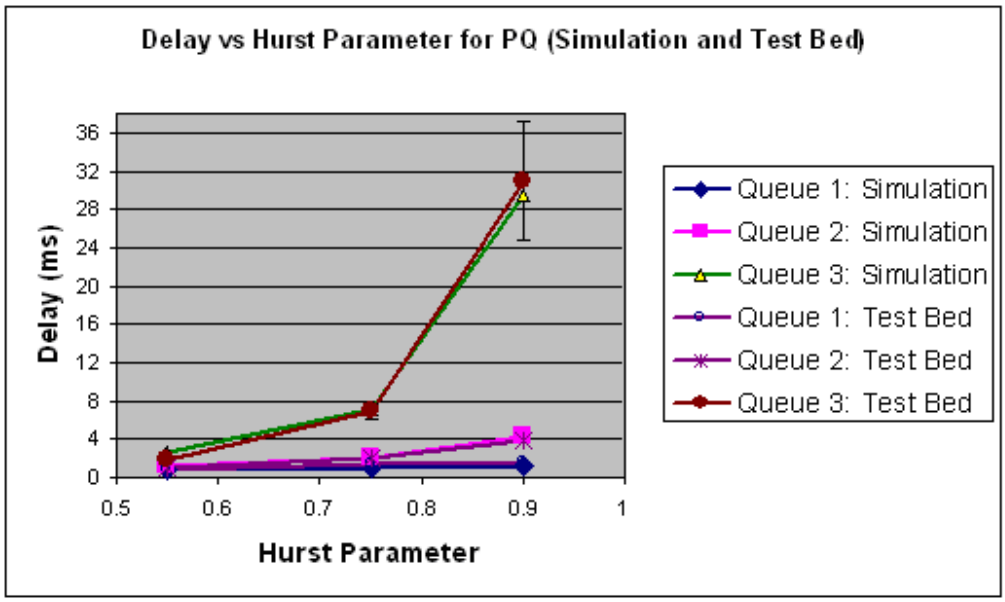


Fig. 10: Mean Delay vs Hurst Parameter for PQ, Simulation vs Test Bed Results

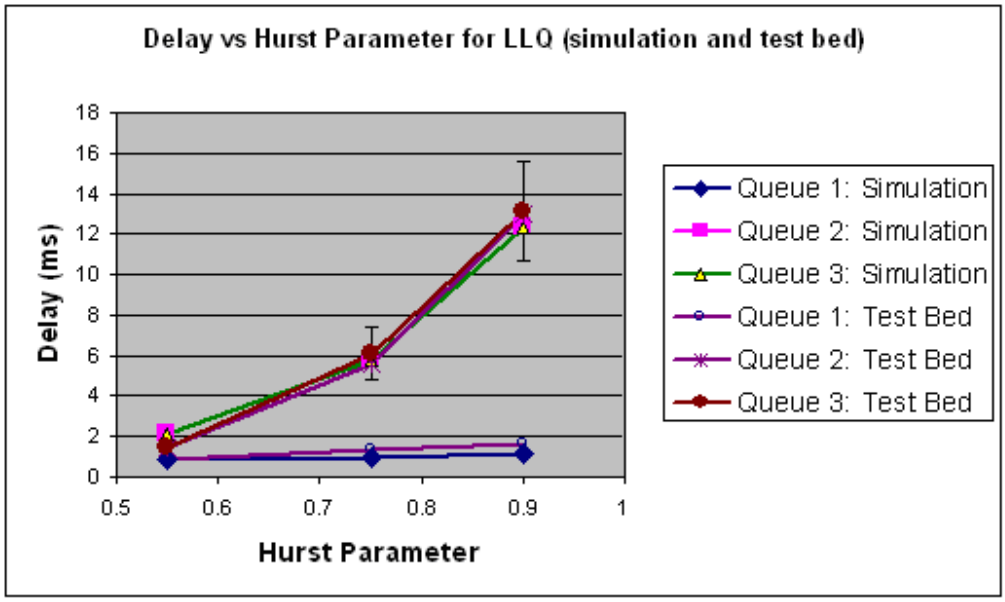


Fig. 11: Mean Delay vs Hurst Parameter for LLQ, Simulation vs Test Bed Results

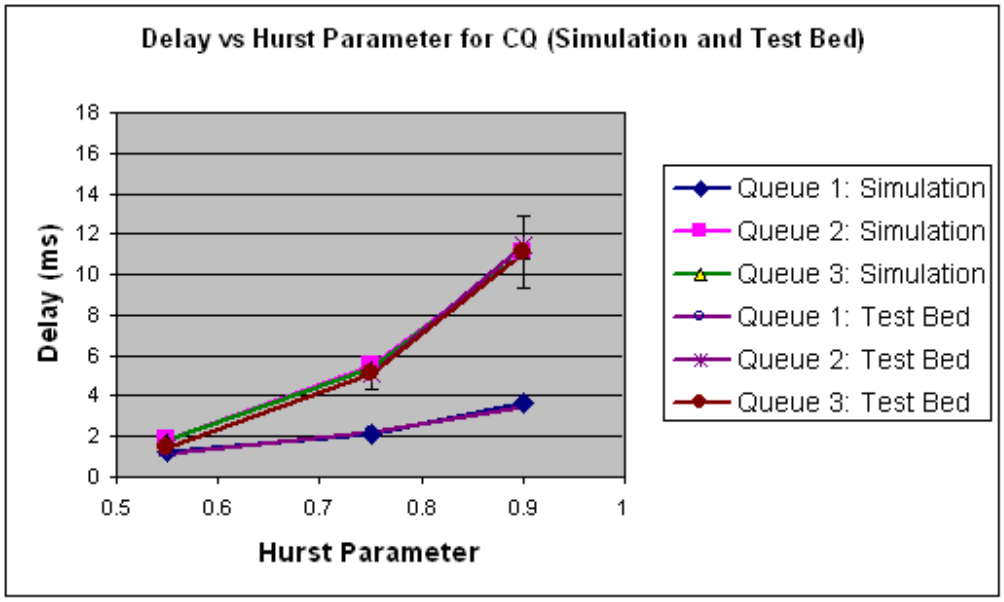


Fig. 12: Mean Delay vs Hurst Parameter for CQ, Simulation vs Test Bed Results