On the QoS Behavior of Self-Similar Traffic in a Converged ONU-BS Under Custom Queueing

Brownson Obaridoa Obele, Mohsin Iftikhar, and Minho Kang

Abstract: **A novel converged optical network unit (ONU)-base station (BS) architecture has been contemplated for next-generation optical-wireless networks. It has been demonstrated through high quality studies that data traffic carried by both wired and wireless networks exhibit self-similar and long range dependent characteristics; attributes that classical teletraffic theory based on simplistic Poisson models fail to capture. Therefore, in order to apprehend the proposed converged architecture and to reinforce the provisioning of tightly bound quality of service (QoS) parameters to end-users, we substantiate the analysis of the QoS behavior of the ONU-BS under self-similar and long range dependent traffic conditions using custom queuing which is a common queuing discipline. This paper extends our previous work on priority queuing and brings novelty in terms of presenting performance analysis of the converged ONU-BS under realistic traffic load conditions. Further, the presented analysis can be used as a network planning and optimization tool to select the most robust and appropriate queuing discipline for the ONU-BS relevant to the QoS requirements of different applications.**

Index Terms: **Gigabit ethernet passive optical network (GEPON), optical wireless converged networks, quality of service (QoS), queuing delay, worldwide interoperability for microwave access (WiMAX).**

I. INTRODUCTION

In the telecommunications industry, we can witness a growing trend in network providers indicating a paradigm shift from a network-centric-based approach to a customer-centric-based approach in terms of service provisioning. In customer-centric propositions, it is utmost important to provide services to customers relevant to the quality of service (QoS) requirements of their applications. Besides, the provisioning of guaranteed or expected QoS is closely tied to assorted QoS parameters such as delay, packet loss rate, throughput etc. Consequently, to provide profitable customer-centric services, service providers need to use accurate models that will enable them to properly optimize their network resources and to construct realistic and proactive service level agreements (SLAs). Ultimately, the meticulous and apprehensive traffic models will enable service providers to make higher return on investment (RoI) since it will enable them to conservatively over-provision their network resources more precisely. Lately, to achieve the goal of fixed-mobile convergence (FMC), researchers have focused on the convergence of fixed and mobile technologies. The convergence of optical and wireless technologies is desirous both from technological and marketing points of view. This is because of their ability to introduce fast, innovative and flexible services in prompt response to market demand, enabling operators to become or remain competitive. Network operators and service providers know that they can succeed only if they foster new markets, broaden their range of services and provide services at a quicker pace and at more competitive prices. Hence, they have to eliminate operational and service bottlenecks imposed by different technologies. An emerging solution is the removal of the barrier between wired and wireless networks through the convergence of optical and wireless networks and services. In this regard, the convergence of gigabit Ethernet passive optical network (GEPON) and worldwide interoperability for microwave access (WiMAX) has gained attention [1]–[3].

Manuscript received March 24, 2010; approved for publication by Suresh Subramaniam, Division III Editor, October 28 , 2010.

This work was supported by the IT R&D program of MKE / KEIT [2009-F-057-01, Large-scale wireless-PON convergence technology utilizing network coding].

B. O. Obele is with the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Building C, Room 415, 1 Oryong Dong, Buk-gu, Gwangju 500-712, South Korea, email: brownson @gist.ac.kr.

M. Iftikhar is with the Computer Science Department, King Saud University, Riyadh, Saudi Arabia, email: miftikhar@ksu.edu.sa.

M. Kang is with the Electrical Engineering Department, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, email: min hokang@kaist.ac.kr.

II. BACKGROUND FOR CONVERGING GEPON AND WIMAX

The convergence of GEPON and WiMAX is getting much attention because the respective technologies are quite similar and positively complement each other. Therefore, synergistically converging them makes very good business sense for a next-generation access network solution. For instance, although, fiber (GEPON) provides huge bi-directional bandwidth capacities, it is still relatively expensive to run fiber cables directly to subscriber homes and devices; which is where the cheap and quick deployment characteristics of WiMAX easily complements GEPON. Additionally, GEPON and WiMAX show a good match in bandwidth hierarchy because a WiMAX base station (BS) supports approximately 70 Mb/s bandwidth over a 20 MHz channel, while GEPON can typically provide 1 Gb/s bandwidth in both upstream and downstream, which is shared by a group of (typically 16) optical network units (ONUs) with each ONU getting an average of about 62.5 Mb/s [1]. Moreover, while wireless access technologies are known to generally offer limited bandwidth with high bit error and packet loss rates, optical access technologies are known to provide limitless bandwidth with extremely low bit error and packet loss rates. Further, while optical networks present difficulties in reconfiguration, maintenance and rewiring when topology changes, wireless networks are known to give little or no difficulties in such issues and so integration is sure to yield synergistic gains. Convergence enables integrated bandwidth allocation, packet schedul-

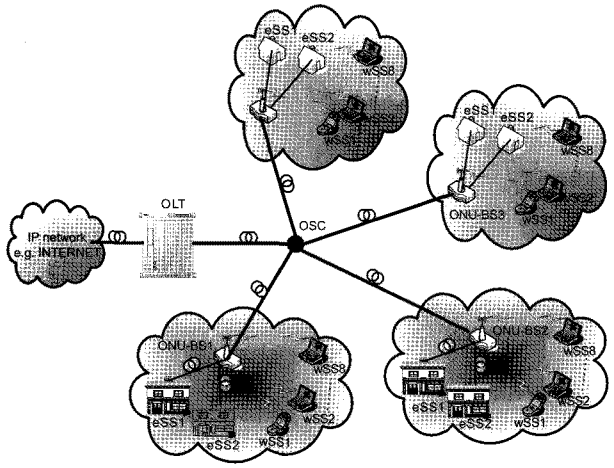

Fig. 1. WiMAX-GEPON convergence architecture.

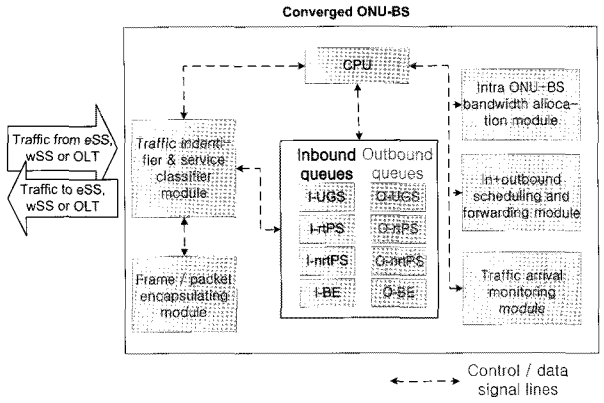

Fig. 2. Architecture of the converged ONU-BS.

ing, and network control and management which help to provide better QoS support and improve network throughput and performance as well as lower the service providers' operating expenses (opex). This convergence brings broadband access directly to subscriber devices and supports mobility as well.

Consequently, in [2], we proposed an architecture for converging GEPON and WiMAX and we addressed the technical challenges involved in such a convergence. In this paper, we extend our previous work [4] on priority queuing (PQ) [5] to provide an analytical modeling of the converged network under custom queuing (CQ) [6]. Meanwhile, some recent high quality studies and Internet measurements [7]–[11] have proven that in modern networks (both wired and wireless), actual data traffic exhibit self-similar and long range dependent (LRD) characteristics; attributes that simple Poisson models, which have been relied on for several years, fail to capture; thus leading to wrong estimates of delay, packet loss rate and other QoS parameters. Accordingly, this results in poor network planning and over-provisioning when service providers rely on such Poisson-based models. Thus, to overcome the limitations of current state-of-the-art work in traffic modeling, we present an analytical model based on self-similar and LRD traffic conditions for our proposed converged architecture. Also, unlike existing work (which only considers outbound data traffic) we consider both inbound and outbound data traffic. It is extremely important to consider inbound traffic (which is peer-to-peer (p2p)); because most of the inbound traffic observed in real networks is p2p. Further, most existing work is based on the simple first-come-first-serve (FCFS) scheduling discipline which cannot provide differential treatment to different kinds of traffic.

To the best of our knowledge, this work presents for the first time the analysis of the queuing behavior of a converged network under self-similar and LRD data traffic conditions using a common and widely-used queuing discipline that is likely to be used in future and real networks—CQ. Our analytical framework is based on G/M/1 queuing system which takes into account multiple classes of traffic input exhibiting LRD and self-similarity. We derive exact QoS parameters including the expected queue length, expected waiting time in queue (queuing delay), end-to-end delay and the packet loss rate; all per QoS traffic class for the CQ scheduling logic. We also develop the finite queue Markov chain for the corresponding CQ discipline. The derived analytical expressions are validated through extensive numerical analysis and simulation study (the numerical results closely match the simulation results). The numerical analysis and simulation experiments have been conducted to investigate and understand the behavior of self-similar traffic in the converged network and particularly to see how QoS parameters are affected. We clarify that this work does not delve into the specific characteristics of WiMAX or GEPON traffic but on their aggregate behavior in the system. This work extends our previous work on priority queuing [4], helps to predict and analyze the behavior of traffic in the converged network and leads to derive accurate and exact parameters such as packet delay and packet loss rate, thus facilitating the provisioning of guaranteed QoS to the end-users.

III. THE CONVERGED ONU-BS ARCHITECTURE

The converged architecture is shown in Fig. 1 and its detail has been presented in [2]. The figure shows that the optical line terminal (OLT) is located at the central office and connects the converged network to an IP core network. The original GEPON ONU, however, is replaced with an ONU-BS, which results from the convergence of the functional entities of a WiMAX BS and a GEPON ONU. The ONU-BS has Ethernet ports for wired connections and WiMAX radios for wireless connections to WiMAX fixed or mobile subscriber stations (SSs). In the architecture, two types of subscribers are defined—eSS and wSS as shown in Fig. 1. The eSSs are the original GEPON subscribers while the wSS are the original WiMAX fixed or mobile subscribers. Communication between two users (eSS or wSS) of different ONU-BSs is always through the OLT. Between the OLT and the ONU-BS, data frames are encapsulated in Ethernet packet data units (PDUs); while, within the ONU-BS, data frames are encapsulated either as Ethernet or WiMAX PDUs. The packets are in Ethernet format over the wired interface between the ONU-BS and the eSS, while over the air interface, which is between the ONU-BS and the wSS, they are encapsulated as WiMAX PDUs. Frames are formatted to suite their destination or target output port; for instance, a frame originating from an eSS and destined for a wSS will be encapsulated as a WiMAX PDU; whereas a frame originating from a wSS and destined for either the OLT or an eSS will be encapsulated in

Ethernet format. Frames originating from an eSS and destined for the OLT are left unchanged; so also are frames originating from a wSS and destined for another wSS connected to the same ONU-BS. Frames from the OLT to the ONU-BS are in Ethernet format.

The main functional entities of the converged ONU-BS are shown in Fig. 2. The multiple classes of incoming traffic into the ONU-BS are received first by the traffic identifier and service classifier module (TISCM). The TISCM checks the incoming frames and based on their destination, and determines if they require encapsulation or not. It also uses the destination address to determine the set of queues they belong to: Inbound or outbound queues. The inbound queues are for frames destined for wSSs and eSSs associated with that ONU-BS while the outbound queues are for packets destined for destinations reachable through the OLT, that is, eSSs and wSSs connected to other ONU-BSs or nodes connected to the IP network. Therefore, depending on their destination and priority class, packets are queued in one of 8 queues—4 inbound queues: Inbound unsolicited grant service (I-UGS), inbound real time polling service (I-rtPS), inbound non-real time polling service (I-nrtPS), and inbound best effort (I-BE), and 4 outbound queues: Outbound UGS (O-UGS), outbound rtPS (O-rtPS), outbound nrtPS (O-nrtPS), and outbound BE (O-BE). The intra ONU-BS bandwidth allocation module determines bandwidth allocation to the SSs and to the ONU-BS queues. The queues can be served according to any queuing scheme of choice. The processor serves packets from one set of queues only, at any given time. For instance, while the ONU-BS is waiting for a grant from the OLT, only the inbound queues are served, and when a grant arrives, their service is suspended for outbound queues to be served. The outbound queues are then served for the duration of the grant. We clarify that the arrival of a grant cannot preempt an inbound packet that is already in service because service is done on a packet-by-packet basis. On the expiration of an allocated grant, the processor resumes servicing of the inbound queues. The scheduling and forwarding module handles this functionality. For more details, the reader is referred to [2].

IV. ANALYTICAL MODELING OF THE ONU-BS

The traffic model considered in this paper has been used and studied in [12]. It is similar to an on/off process and captures the dynamics of packet generation while accounting for scaling properties observed in telecommunication networks. It belongs to a particular class of self-similar traffic models called infinite source Poisson models. A common feature of such models is their heavy-tailed distribution of sessions that occur at the flow level and arrive according to a Poisson process. Furthermore, the local traffic injection process over each session is a distinguishing feature. In addition, the Hurst parameter is implicit in the distribution of the sessions. The traffic model is LRD and almost second-order self-similar as the auto-covariance function of its increment is equal to that of fractional Gaussian noise for sufficiently large time lags [13]. Injected traffic is obtained by aggregating packets generated by several sources. In the framework of a Poisson point process, the model represents an infinite number of potential sources. Similar to [13], each source initiates a session with a heavy-tailed distribution, in particular 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 are taken to arrive according to a Poisson process with rate λ. Locally, at the ONU-BS, packets generated by each source arrive according to a Poisson process with rate α throughout each session. The local packet generation process could be taken as a compound Poisson process which would then represent packet sizes as well [13]. We consider the stationary version of this model based on an infinite past. Further, the packet sizes are assumed 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 local packet generation is assumed to be Poisson over each session, consistent with the long-range dependence of the packet arrivals, the aggregated packet arrival process is clearly not Poisson.

Specifically for our converged GEPON-WiMAX network, a model of eight queues based on G/M/1 queuing system is considered, which takes into account multiple classes of self-similar traffic input. The packets arriving into the converged ONU-BS are queued on a first-come-first-queued basis into one of eight finite-sized queues depending on their traffic class and destination (inbound or outbound) and the scheduling inside each queue is first come first serve (FCFS). For simplicity, we denote the queues as: type-1, representing UGS queues; type-2, for rtPS; type-3, for nrtPS; and type-4, for BE. A type-p ($p = 1, 2, 3, 4$) packet can either be inbound or inbound. We let the service time distribution have rates μ_1, μ_2, μ_3, and μ_4, respectively for the four traffic classes. We assume that the arrival rate of traffic from the eSS and wSS is the same and experience an equal amount of propagation delay to get to the ONU-BS but is different for each traffic class. For simplicity, we do not consider the specific characteristics of WiMAX or GEPON traffic. Moreover, due to the nature of self-similar traffic, whether we consider the traffic arrival rate at eSS or at wSS, the aggregate behavior of the traffic arriving into the ONU-BS will always be self-similar and LRD. Therefore, in this paper, we consider that the arrival rate of packets into the ONU-BS depends only on the traffic class of each arriving packet and not on their individual sources (wired Ethernet for eSS, or wireless WiMAX for wSS). Consequently, the arrival rate of packets into the ONU-BS is simply represented by λ_p packets per second, for type-p packets. We assume that all the packets are unicast (or anycast, where all members of the anycast group are in the same local area network (LAN)), which means that each packet is usually destined for a specific host only. Next, we specify the transition probabilities from all the states in the state space, which forms the transition probability matrix P of the Markov chain. Further, we derive closed form expressions of the expected queuing delay in the ONU-BS's queues and the packet loss rate; all per QoS class using CQ discipline.

A. Self-Similar Traffic Interarrival Distribution

In [13]–[15], the interarrival time distributions between packets of same class have been derived. The distributions of cross interarrival time between different types of packets were derived on the basis of results for a system where there is only one

type/class of packet. By following the same methodology, we find the interarrival distribution and cross interarrival distributions between four different types of packets for our self-similar traffic model as follows.

$$f_{T_{ij}}(t)=f_j^0(t)\overline{F_i^0}(t)\overline{F_k^0}(t)\overline{F_l^0}(t) \quad (1)$$

$$f_{T_{ii}}(t)=f_{T_i}(t)\overline{F_j^0}(t)\overline{F_k^0}(t)\overline{F_l^0}(t) \quad (2)$$

where $i,j,k,l = 1,2,3,4$, $f_{T_{ij}}(t)$ is the density of time until a j arrival given an i arrival, $f_{T_i}(t)$ is the density of time until the next type i arrival, and $\overline{F_i^0}(t) = \Pr\{\text{no type } i \text{ packets arrive in } t \text{ time units}\}$. For more details of the interarrival time distribution calculations, the reader is referred to see [13]–[15].

B. QoS Behavior Under CQ

For our proposed converged architecture, we have considered a common queuing discipline, CQ, to analyze the QoS behavior of traffic. In PQ [4], high priority queues get absolute service priority over low priority queues, which is a major drawback. CQ addresses the biggest drawback of PQ by providing a queuing tool that does service all queues, even during times of congestion. The CQ scheduler reserves an approximate percentage of overall link bandwidth to each queue. CQ approximates the bandwidth percentages, as opposed to meeting an exact percentage, due to the simple operation of the CQ scheduler. The CQ scheduler performs round robin service on each queue, beginning with queue 1. It takes packets from a queue, until the byte/packet count specified for that queue has been met or exceeded. After the queue has been serviced for that many bytes/packets, or the queue does not have any more packets, the scheduler moves on to the next queue and repeats the process. Hence, CQ scheduler essentially guarantees a minimum bandwidth for each queue, while allowing queues to have more bandwidth as well under the right conditions. The readers are referred to [5] and [6] for detailed discussions on CQ discipline. For simplicity, we specify the scheduler logic in such a way that the scheduler serves two packets from the UGS queue, two packets from the rtPS queue, one packet from the nrtPS queue and one packet from the BE queue during each service cycle. Accordingly, when there is a grant, two packets are served from each of the O-UGS and O-rtPS queues; and one packet is served from each of the O-nrtPS and O-BE queues in each outbound service cycle. Whereas, when there is no grant, two packets are served from each of the I-UGS and I-rtPS queues; and one packet is served from each of the I-nrtPS and I-BE queues in each inbound service cycle. We develop the finite Markov chain for the CQ scheduling discipline. Our analysis for CQ relies on the limiting distribution of the state of the queue at the arrival instants, which can be computed using the analysis given above for our self-similar traffic model. We observe the CQ system at the instants of packet arrivals. At such instants, the number in the system is the number of packets that the arriving packet sees in the queues plus the packet in service, if any, excluding the arriving packet itself. An inbound service cycle will consist of $s_{1in}^1 + s_{1in}^2 + s_{2in}^1 + s_{2in}^2 + s_{3in}^1 + s_{4in}^1$ time units, while an outbound service cycle will consist of $s_{1out}^1 + s_{1out}^2 + s_{2out}^1 + s_{2out}^2 + s_{3out}^1 + s_{4out}^1$ time units. Where for instance, s_{1out}^1 denotes the servicing of the first packet of the O-UGS queue, s_{1out}^2 denotes the servicing of the second packet of the O-UGS queue, etc. And so, without loss of generality, the notation s_p^m can be used to differentiate between the six different kinds of packets that can be in service during each inbound or outbound service cycle, where $m = 1,2$ and $p = 1in, 2in, 3in, 4in$ for an inbound service cycle; and $m = 1,2$ and $p = 1out, 2out, 3out, 4out$ for an outbound service cycle. We note that for this CQ model, the size of allocated grants (and grant requests) will be multiples of the outbound service cycle so that a service cycle maybe completed before a grant is terminated. Similarly, for an inbound service cycle, the arrival of a grant cannot preempt the completion of an inbound service cycle that has already begun. That is, the arrival of a grant can only preempt the next inbound service cycle and not an inbound service cycle that has already begun. We let $\{X_n : n \geq 0\}$ denote the imbedded Markov chain at the time of these arrival instants and we define the state space under CQ as

$$\begin{aligned} X = \Big\{ & (in_1, in_2, in_3, in_4, out_1, out_2, out_3, out_4, A_P, S_P^m, \\ & d, G) : A_P \in \{A_1, A_2, A_3, A_4\}, d \in \{d_{in}, d_{out}\}, \\ & S_P^m \in \{s_{1in}^1, s_{1in}^2, s_{2in}^1, s_{2in}^2, s_{3in}^1, s_{4in}^1, s_{1out}^1, s_{1out}^2, s_{2out}^1, \\ & s_{2out}^2, s_{3out}^1, s_{4out}^1, I\}, G \in \{0,1\}, in_1, in_2, in_3, in_4, out_1, \\ & out_2, out_3, out_4 \in Z_+ \Big\}. \end{aligned} \quad (3)$$

C. The Embedded Markov Chain States for CQ

The transition probability matrix P of the Markov chain is generated by specifying the transition probabilities from all the states in the state space, which includes idle, non-idle, and full queue states. We enumerate the states of the Markov chain, their possible transitions, and the probabilities of such transitions. We elaborate on the analysis of one possible transition; the analysis for all other possible transitions follows similarly. We analyze for the transition from $(in_1, in_2, in_3, in_4, out_1, out_2, out_3, out_4, A_1, s_{1in}^1, d_{in}, 0)$ to $(in_1^*, in_2^*, in_3^*, in_4^*, out_1^*, out_2^*, out_3^*, out_4^*, A_3, s_{2in}^2, d_{out}, 0)$.

This is a case where a transition occurs from a type-1 arrival to a type-3 arrival, such that the type-1 destined to join the inbound queue, in_1, sees in_1, in_2, in_3, and in_4 packets in the inbound queues; out_1, out_2, out_3, and out_4 packets in the outbound queues; the first packet of I-UGS in service for some inbound service cycle; and no grant. The state transitioned to a state where a grant is similarly not active; the arrival, a type-3 destined for the outbound queue, out_3^*, sees the second type-2 inbound packet in service; and finds $in_1^*, in_2^*, in_3^*, in_4^*, out_1^*, out_2^*, out_3^*, out_4^*$ packets in the queues. Due to our CQ scheduling logic and the ability of outbound traffic to preempt inbound traffic when a grant arrives, the next-state type-3 arrival can only see an inbound type-2 packet in service if and only if, there is still no grant and the servicing of two inbound type-1 packets and the first inbound type-2 packet, i.e, s_{1in}^1, s_{1in}^2, and s_{2in}^1, have definitely been completed. Regarding the second inbound type-2 packet that is found in service, i.e., s_{2in}^2, it could either belong to the same inbound service cycle as the first inbound type-1 packet that was met in service in the previous state or it could belong to another inbound service cycle in the case where several (the exact number we do not

know due to the memory-less property of exponential service times) inbound service cycles were exhausted during the inter-arrival time T_{13}. In the case where the packet served in the previous state and the packet found in service in the next state belong to the same inbound service cycle, say cycle-A, two inbound type-1 packets and one inbound type-2 packet were served during T_{13} with no packets served from the inbound type-3 and type-4 queues. More so, because no grant arrived during T_{13}, no outbound packets were served. If on the other hand, s^2_{2in} belongs to the next inbound service cycle, say cycle-B, then surely four inbound type-1 packets, three inbound type-2 packets, one inbound type-3 packet and one inbound type-4 packet were served during T_{13}. And if s^2_{2in} belongs to cycle-C, then six inbound type-1 packets, five inbound type-2 packets, two inbound type-3 packets and two inbound type-4 packets were served during T_{13}, etc. Hence, during T_{13}, the maximum number of inbound type-1 packets that could have been served is in_1 (if the number of packets in in_1 is even), else it is $in_1 - 1$ (if the number of packets in in_1 is odd). While the maximum number of inbound type-2 packets that could have been served will be either in_2 (if the number of packets in in_2 is odd) or $in_2 - 1$ (if the number of packets in in_2 is even). Because in_* includes the inbound type-1 packet that arrived in the previous state, we have

$$in_1^* = in_1 + 1 - k, \qquad in_2^* = in_2 - (k-1),$$
$$in_3^* = in_3 - \frac{(k-2)}{2}, \qquad in_4^* = in_4 - \frac{(k-2)}{2}$$

until the inbound type-1, type-3, and type-4 queues are exhausted, or only one packet (in case of even number of packets) or two packets (in case of odd number of packets) of type-2 remain in the system, the second type-2 packet (s^2_{2in}) being in service, whichever occurs first. First, we consider queue 3 and queue 4 as a single queue and denote it as queue in_{34}. So, there are two possibilities.

1) If $in_{34} \leq in_2 \geq in_1$ and $k = 2, 4, \cdots, in_1$ (in case of odd no. of packets in queue 1) or $in_1 - 1$ (if even no. of pkts) and for queue 3 and 4, $k = 2, 4, \cdots, 2in_m, m = 3, 4$ or when $in_{34} > in_2 < in_1, k = 2, 4, \cdots, in_2$ (even) or $in_2 - 1$ (odd). Therefore, the transition probability is

$$\Pr\Big[X_{n+1} = (in_1 - k + 1, in_2 - (k-1), in_3 - (\frac{k-2}{2}), in_4 - (\frac{k-2}{2}), out_1, out_2, out_3, out_4, A_3, s^2_{2in}, d_{out}, 0) \Big| X_n = (in_1, in_2, in_3, in_4, out_1, out_2, out_3, out_4, A_1, s^1_{1in}, d_{in}, 0)\Big]$$
$$= \Pr\{k \text{ served from } in_1, k-1 \text{ served from } in_2, (k-2)/2 \text{ served from } in_3 \text{ and } in_4 \text{ each, no grant during } T_{13} \text{ and the 2nd type-2 pkt remains in service during } T_{13}\}$$
$$= \overline{F^0_G}(t) \int_0^\infty \int_0^t \int_{t-x}^\infty f_{s^2_{2in}}(s) f_{s^k_{1in} + s^{k-1}_{2in} + s^{\frac{k-2}{2}}_{3in} + s^{\frac{k-2}{2}}_{4in}}(x) f_{T_{13}}(t)\, ds dx dt. \quad (4)$$

2) On the other hand, only class 2 packets are served if queue 1, queue 3, and queue 4 are exhausted. Therefore, If $in_1 \leq in_2 \geq in_{34}$ and $k = in_1 + 1, \cdots, in_2$ (even) or $in_2 - 1$ (odd) for queue 1 and for queue 3 and 4, $k = 2in_m + 2, \cdots, in_2$ (even), or $in_2 - 1$ (odd), $m = 3, 4$; in which case we have the transition probability as follows.

$$\overline{F^0_G}(t) \int_0^\infty \int_0^t \int_{t-x}^\infty f_{s^2_{2in}}(s) f_{s^{in_1}_{1in} + s^{k-1}_{2in} + s^{in_3}_{3in} + s^{in_4}_{4in}}(x) f_{T_{13}}(t)\, ds dx dt. \quad (5)$$

In the same way, the transition probabilities for all other possible transitions can also be easily obtained.

D. Derivation of the QoS Parameters for CQ

To simplify our analysis, we consider type-3 and type-4 queues as a single queue and call it type-34 (read as type-three-four) queue. We can do this because due to the symmetry of alternating service, the expected delay for type-3 and type-4 packets will be the same. Considering type-3 and type-4 queues as a single queue, the CQ scheduler will serve two packets from the type-1 queue, two packets from the type-2 queue, and 2 packets from the type-34 queue (in reality, one packet each from type-3 and type-4 queues) during each inbound or outbound service cycle. For a type-1 arrival, it will wait for the packet it meets in service (if any) plus the service time of type-1, type-2, and type-34 packets due to the round robin service fashion already described. In addition, if the type-1 is inbound, then its waiting time will also be affected by grants that arrive before it is served; while for outbound packets, their delay will be affected by the time spent waiting for grants to arrive. For convenience, we use a reduced/simplified form of our general state space, which considers that type-3 and type-4 queues have been combined into a single type-34 queue. The reduced state space is

$$X^* = (J_1, J_2, J_{34}, J_G, A_P, S^m_P, G), \quad P = 1, 2, 3, 4,\ m = 1, 2,\ G = 0, 1 \quad (6)$$

where, when G = 0 (no grant), J_1, J_2, and J_{34} are the respective queue occupancy of the inbound queues,J_G is the aggregate occupancy of the outbound queues, and S^m_P denotes the mth type-p inbound packet in service. Whereas, when G = 1 (there is a grant), J_1, J_2, and J_{34} are the respective queue occupancy of the outbound queues, J_G is the aggregate occupancy of the inbound queues / symbolic representation of the time until the next grant and S^m_P denotes the mth type-p outbound packet in service. A_P is used to denote a type-p arrival.

E. Expected Waiting Time (Queuing Delay) for CQ

The waiting time expressions for type-1 and type-2 packets under our CQ scheduling discipline are the same (i.e., $E[W_1^{in/out}] = E[W_2^{in/out}]$) due to the symmetry of alternating service, where $E[\cdot]$ is the expectation of a random variable. Similarly, the waiting time expressions for type-3 and type-4 packets are also the same (i.e., $E[W_3^{in/out}] = E[W_4^{in/out}]$) due to the symmetry of alternating service of those queues. We consider the waiting time for a type-1 (same for a type-2) inbound packet in detail; the analysis for an outbound type-1 (same for an outbound type-2) follows quite similarly. For a type-1 arrival, its waiting time will depend on the packet it meets in service (if any) plus the service time of type-1, type-2 and type-34 packets that will be served before it due to the round robin

service fashion already described. If the arrival is inbound and there is a grant (or grants arrive before it is served), the waiting time will depend on the total (possibly random) size of the grants or aggregate outbound queues, whichever is smaller, because inbound queues can only be served when there is no grant; and our model allows for an ONU-BS to terminate its allocated grant when it has no outbound packets. Due to the round robin and symmetrical service fashion of the queues, the size of each queue relative to the others is an important consideration; equally important is the evenness or oddness of the number of packets in the queue. We consider when $J_2 \leq J_1 \geq J_{34}$ and when $J_2 > J_1 < J_{34}$.

a) The states $(J_1, J_2, J_{34}, J_G, A_1, s_P^m, G)$, $P = 1, 2, 3, 4$, $m = 1, 2$, $G = 0, 1$, and $J_2 > J_1 < J_{34}$: If P=1, then either the first inbound (G=0) / outbound (G=1) type-1 packet, s_1^1, or the second inbound/outbound type-1 packet, s_1^2, is in service for that inbound/outbound service cycle. We consider an inbound type-1 arrival. If s_1^1 is in service and the number of inbound type-1 packets in queue is even, then a newly arriving inbound type-1 packet will wait for $R_m + s_{out}^{J_{out}} + s_1^{J_1-1} + s_2^{J_1} + s_3^{\frac{J_1}{2}} + s_4^{\frac{J_1}{2}}$ time units while if the number of packets is odd, it will wait for $R_m + s_{out}^{J_{out}} + s_1^{J_1-1} + s_2^{J_1-1} + s_3^{\frac{J_1-1}{2}} + s_4^{\frac{J_1-1}{2}}$. Whereas, if it is s_1^2 in service with even number of type-1 packets in queue, then the arriving packet will wait $R_m + s_{out}^{J_{out}} + s_1^{J_1-1} + s_2^{J_1} + s_3^{\frac{J_1}{2}} + s_4^{\frac{J_1}{2}}$ time units, while if the number of packets in the queue is odd, it will wait $R_m + s_{out}^{J_{out}} + s_1^{J_1-1} + s_2^{J_1+1} + s_3^{\frac{J_1+1}{2}} + s_4^{\frac{J_1+1}{2}}$, where R_m denotes the remaining service time of a packet in service which has the same exponential distribution as s_p^m; and $s_{out}^{J_{out}}$ denotes the service time associated with the aggregate number of outbound packets that will be served if grants arrive before the newly arrived packet is served. Similarly, arguments can easily be written for other possibilities as well.

b) The states $(J_1, J_2, J_{34}, J_G, A_1, s_P^m, G)$, $P = 1, 2, 3, 4$, $m = 1, 2$, $G = 0, 1$, and $J_2 \leq J_1 \geq J_{34}$: There are many possibilities that can occur depending on the size of the various queues relative to each other, such as, the arrival of more packets into the type-2 and type-34 queues while the type-1 arrival is still waiting in queue, the arrival of grants which causes the queuing delay of inbound packets to increase, and the non-arrival of grants which causes the queuing delay of outbound packets to increase. While the new arrival is still waiting, if new packets arriving into the type-2 and type-34 queues occur at the right periods, the newly arriving inbound type-1 packet may wait a maximum of the amount of time given in case a), depending on the values of m and p. For example, if $J_1 = 5$, $J_2 = 3$, and $J_{34} = 3$ (total number of packets in the type-3 and type-4 inbound queues); our CQ scheduler will serve two packets from the inbound type-1 queue, two packets from the type-2 queue, and two packets from the type-34 queue (one each from the type-3 and type-4 queues) in the first inbound service cycle. However, during the second inbound service cycle, the CQ scheduler will serve the 3rd and 4th packets from the type-1 queue, but when the 3rd packet from the type-2 queue goes into service, either the 4th packet from the type-2 queue (if new packets arrived into the type-2 queue) or the 3rd packet from the type-34 queue (if new packets did not arrive into the type-2 queue) will enter into service next. Similarly, when the 4th packet from the type-2 queue or the 3rd packet from the type-34 queue enters into service, either the 4th packet from the type-34 queue (if new packets arrived into the type-34 queue) or the 5th packet from the type-1 queue (if new packets did not arrive into the type-34 queue) will enter into service next. There are many more possibilities as well, including when grants arrive or don't arrive and the argument goes on even longer for larger J_1 . If no arrivals occur into the inbound type-2 and type-34 queues and no grants arrive, the newly arriving type-1 inbound packet will wait a minimum of $s_1^{J_1} + s_1^{J_2} + s_1^{J_3} + s_1^{J_4}$ time units.

Using the minimum and maximum values, it is possible to form exact bounds on the queuing delay in the ONU-BS. Putting cases a) and b) together, we can obtain the exact bounds on the expected waiting time for an inbound type-1 (same for an inbound type-2) packet as $C_1^{in} \leq E[W_1^{in}] \leq C_2^{in}$ given in (10) and (11). Following the same procedure, we can easily write down the exact bounds of the expected waiting time in the ONU-BS for an inbound type-3 (same for an inbound type-4) packet as $C_3^{in} \leq E[W_3^{in}] \leq C_4^{in}$, where C_3^{in} and C_4^{in} are as given in (12) and (13). Same goes for the outbound packets. Due to the symmetry of alternating service, the expected delay for type-1 and type-2 outbound packets are the same. For the same reason, the expected delays for type-3 and type-4 outbound packets are also the same.

F. Packet Loss Rate under CQ

We only consider packet losses that occur when arrivals of specific traffic class (inbound or outbound) meet their respective queues full. This kind of packet loss rate (PLR) can be readily extracted from the description of the CQ system given previously. We note that in real networks (especially wireless networks); packet loss also occurs for many other reasons. So, assuming a system where packet loss is due to a full queue only, the PLR for each traffic class and bound can be obtained as the sum of the steady-state probabilities of states where an arrival occurs for a full queue. We only show the expression for type-1 arrivals. All others follow similarly.

$$\mathrm{PLR}_1^{in} = \sum_{i=0}^{J_4} \sum_{j=0}^{J_3} \sum_{k=0}^{J_2} \sum_{l=0}^{J_{out}} \sum_{m=1}^{4} \sum_{n=0,1} \pi(J_1, k, j, i, l, A_1, s_m, n) \tag{7}$$

$$\mathrm{PLR}_1^{out} = \sum_{i=0}^{J_4} \sum_{j=0}^{J_3} \sum_{k=0}^{J_2} \sum_{l=0}^{J_{in}} \sum_{m=1}^{4} \sum_{n=0,1} \pi(J_1, k, j, i, l, A_1, s_m, n) \tag{8}$$

G. End-to-End Delay Under CQ

In actual networks, most of the packets will be outbound and so in this section we only consider the end-to-end delay of outbound packets. The end-to-end delay of an outbound type-p packet will depend on the time spent by a type-p packet in the ONU-BS queues until it is served and the delay experienced by the packet from the time it is served in the ONU-BS to the

$$
\begin{aligned}
C_1^{in} = & \sum_{j_1=0}^{J_1-1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3}\sum_{j_4=0}^{J_4}\sum_{j_{out}=1}^{J_{out}} \left(\frac{j_{out}}{\mu_{out}}+\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,j_{out},A_1,s_{out}^{j_{out}},1) \\
& + \sum_{j_1=1}^{J_1-1}\sum_{j_2=0}^{\lfloor j_1\rfloor}\sum_{j_3=0}^{\lfloor j_1/2\rfloor}\sum_{j_4=0}^{\lfloor j_1/2\rfloor} \left(\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_1^1,0) \\
& + \sum_{j_1=1}^{J_1-1}\sum_{j_2=\lceil j_1\rceil}^{J_2}\sum_{j_3=\lceil j_1/2\rceil}^{J_3}\sum_{j_4=\lceil j_1/2\rceil}^{J_4} \left(\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lfloor j_1/2\rfloor}{\mu_3}+\frac{\lfloor j_1/2\rfloor}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_1^1,0) \\
& + \sum_{j_1=1}^{J_1-1}\sum_{j_2=0}^{\lfloor j_1\rfloor}\sum_{j_3=0}^{\lfloor j_1/2\rfloor}\sum_{j_4=0}^{\lfloor j_1/2\rfloor} \left(\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_1^2,0) \\
& + \sum_{j_1=1}^{J_1-1}\sum_{j_2=\lceil j_1\rceil}^{J_2}\sum_{j_3=\lceil j_1/2\rceil}^{J_3}\sum_{j_4=\lceil j_1/2\rceil}^{J_4} \left(\frac{j_1}{\mu_1}+\frac{\lceil j_1\rceil}{\mu_2}+\frac{\lceil j_1/2\rceil}{\mu_3}+\frac{\lceil j_1/2\rceil}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_1^2,0) \\
& + \sum_{j_1=0}^{J_1-1}\sum_{j_2=1}^{\lfloor j_1\rfloor}\sum_{j_3=0}^{\lfloor j_1/2\rfloor}\sum_{j_4=0}^{\lfloor j_1/2\rfloor} \left(\frac{j_2}{\mu_2}+\frac{j_1}{\mu_1}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_2^1,0) \\
& + \sum_{j_1=0}^{J_1-1}\sum_{j_2=\lceil j_1\rceil}^{J_2}\sum_{j_3=\lceil j_1/2\rceil}^{J_3}\sum_{j_4=\lceil j_1/2\rceil}^{J_4} \left(\frac{1}{\mu_2}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor+1}{\mu_2}+\frac{\lfloor j_1/2\rfloor+1}{\mu_3}+\frac{\lfloor j_1/2\rfloor+1}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_2^1,0) \\
& + \sum_{j_1=0}^{J_1-1}\sum_{j_2=1}^{\lfloor j_1\rfloor}\sum_{j_3=0}^{\lfloor j_1/2\rfloor}\sum_{j_4=0}^{\lfloor j_1/2\rfloor} \left(\frac{j_2}{\mu_2}+\frac{j_1}{\mu_1}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_2^2,0) \\
& + \sum_{j_1=0}^{J_1-1}\sum_{j_2=\lceil j_1\rceil}^{J_2}\sum_{\lceil j_3=j_1/2\rceil}^{J_3}\sum_{\lceil j_4=j_1/2\rceil}^{J_4} \left(\frac{1}{\mu_2}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lfloor j_1/2\rfloor+1}{\mu_3}+\frac{\lfloor j_1/2\rfloor+1}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_2^2,0) \\
& + \sum_{j_1=0}^{J_1-1}\sum_{j_2=0}^{\lfloor j_1\rfloor}\sum_{j_3=1}^{\lfloor j_1/2\rfloor}\sum_{j_4=0}^{\lfloor j_1/2\rfloor} \left(\frac{j_3}{\mu_3}+\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_3,0) \\
& + \sum_{j_1=0}^{J_1-1}\sum_{j_2=\lceil j_1\rceil}^{J_2}\sum_{j_3=\lceil j_1/2\rceil}^{J_3}\sum_{j_4=\lceil j_1/2\rceil}^{J_4} \left(\frac{1}{\mu_3}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lceil j_1/2\rceil}{\mu_3}+\frac{\lceil j_1/2\rceil+1}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_3,0) \\
& + \sum_{j_1=0}^{J_1-1}\sum_{j_2=0}^{\lfloor j_1\rfloor}\sum_{j_3=0}^{\lceil j_1/2\rceil}\sum_{j_4=1}^{\lceil j_1/2\rceil} \left(\frac{j_4}{\mu_4}+\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_3}{\mu_3}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_4,0) \\
& + \sum_{j_1=0}^{J_1-1}\sum_{j_2=\lceil j_1\rceil}^{J_2}\sum_{j_3=\lceil j_1/2\rceil}^{J_3}\sum_{j_4=\lceil j_1/2\rceil}^{J_4} \left(\frac{1}{\mu_4}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lfloor j_1/2\rfloor}{\mu_3}+\frac{\lfloor j_1/2\rfloor}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_4,0) \qquad (10)
\end{aligned}
$$

time it arrives at the OLT, so-called propagation delay of optical signal over the passive optical network (PON) segment. We assume a tree topology as shown in Fig. 1. We also assume that single fiber spans between the OLT and the optical splitter-combiner (OSC), so the average propagation delay, $W_{\text{OSC-OLT}}$, is the same both in the upstream and downstream. It is also the same for all ONU-BSs associated with that OLT. Further, we assume that single fibers span between each ONU-BS and the OSC. Although, this implies same propagation delay in both directions per ONU-BS, the delay is different for each of the N ONU-BSs, assuming a 1:N network. This is because it is well known that the propagation delay in fiber is deterministic and depends upon the length of the fiber. Further, in actual networks, the ONU-BSs are not necessarily placed at equal distances away from the OSC. Hence, for an outbound type-p packet of ONU-BS i, we obtain the end-to-end delay for CQ as

$$E[W_P^{e2e}] = W_{\text{ONUBS}_i-\text{OSC}} + W_{\text{OSC - OLT}} + E[W_P^{\text{out}}]. \qquad (9)$$

The lower bound for an arriving type-1 packet (similar for a type-2), is obtained as (10). And for the upper bound of an arriving type-1 packet (similar for a type-2), we have (11). Whereas the lower bound for an arriving type-3 packet (similar for a type-4), is obtained as (12). And for the upper bound of an arriving type-3 packet (similar for a type-4), we have (13).

V. NUMERICAL AND SIMULATION RESULTS

In this section, the results of simulation experiments conducted to understand and evaluate the QoS behavior of self-similar traffic in the converged ONU-BS are presented. Also, we present numerical results that closely match the simulation results, verifying the validity of the closed form expressions presented in this paper. The simulation engine was built in a well structured manner to allow free and easy customization to suit any desired scheduling logic. The key element for the scheduler logic in the simulator is the Scheduler class. Here, we used the template method design pattern given in [16]. This allows any scheduling algorithm to be loosely coupled but easily integrated, overriding the existing program skeleton. Custom queuing scheduler was actually implemented to analyze the corre-

$$
\begin{aligned}
C_2^{in} = & \sum_{j_1=0}^{J_1-1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3}\sum_{j_4=0}^{J_4}\sum_{j_{out}=1}^{J_{out}}\left(\frac{j_{out}}{\mu_{out}}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lfloor j_1/2\rfloor}{\mu_3}+\frac{\lfloor j_1/2\rfloor}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,j_{out},A_1,s_{out}^{J_{out}},1) \\
& +\sum_{j_1=1}^{J_1-1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3}\sum_{j_4=0}^{J_4}\left(\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lfloor j_1/2\rfloor}{\mu_3}+\frac{\lfloor j_1/2\rfloor}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_1^1,0) \\
& +\sum_{j_1=1}^{J_1-1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3}\sum_{j_4=0}^{J_4}\left(\frac{j_1}{\mu_1}+\frac{\lceil j_1\rceil}{\mu_2}+\frac{\lceil j_1/2\rceil}{\mu_3}+\frac{\lceil j_1/2\rceil}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_1^2,0) \\
& +\sum_{j_1=0}^{J_1-1}\sum_{j_2=1}^{J_2}\sum_{j_3=0}^{J_3}\sum_{j_4=0}^{J_4}\left(\frac{1}{\mu_2}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor+1}{\mu_2}+\frac{1+\lfloor j_1/2\rfloor}{\mu_3}+\frac{1+\lfloor j_1/2\rfloor}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_2^1,0) \\
& +\sum_{j_1=0}^{J_1-1}\sum_{j_2=1}^{J_2}\sum_{j_3=0}^{J_3}\sum_{j_4=0}^{J_4}\left(\frac{1}{\mu_2}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lfloor j_1/2\rfloor+1}{\mu_3}+\frac{\lfloor j_1/2\rfloor+1}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_2^2,0) \\
& +\sum_{j_1=0}^{J_1-1}\sum_{j_2=0}^{J_2}\sum_{j_3=1}^{J_3}\sum_{j_4=0}^{J_4}\left(\frac{1}{\mu_3}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lfloor j_1/2\rfloor}{\mu_3}+\frac{\lfloor j_1/2\rfloor+1}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_3,0) \\
& +\sum_{j_1=0}^{J_1-1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3}\sum_{j_4=1}^{J_4}\left(\frac{1}{\mu_4}+\frac{j_1}{\mu_1}+\frac{\lfloor j_1\rfloor}{\mu_2}+\frac{\lfloor j_1/2\rfloor}{\mu_3}+\frac{\lfloor j_1/2\rfloor}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_1,s_4,0) \qquad (11)
\end{aligned}
$$

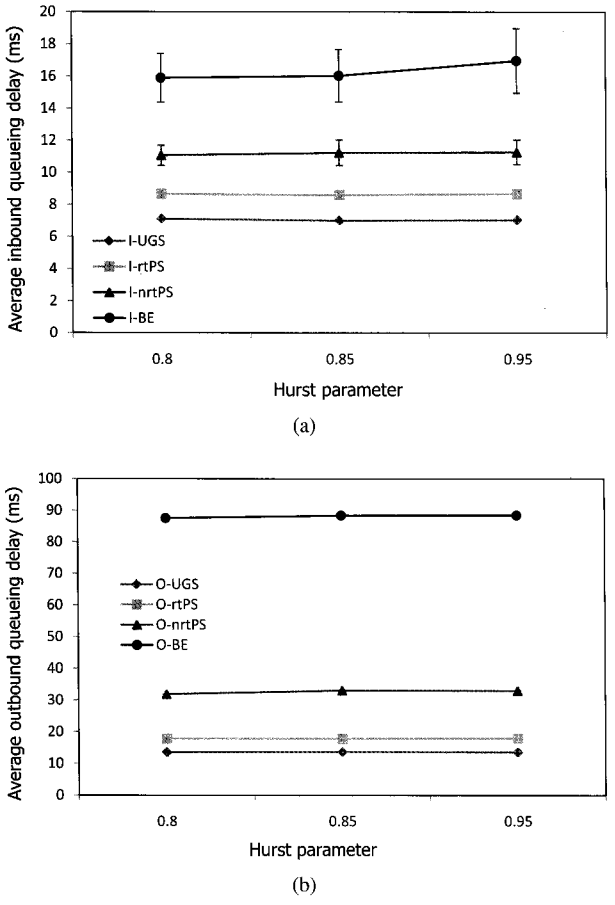

Fig. 3. Queuing delay for PQ discipline: (a) Inbound and (b) outbound.

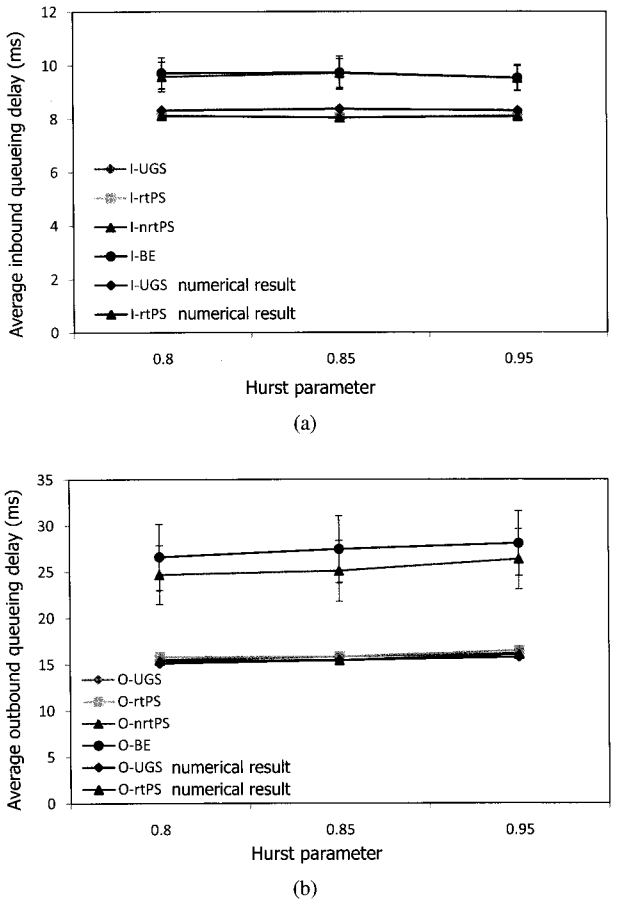

Fig. 4. Queuing delay for CQ discipline: (a) Inbound and (b) outbound.

sponding QoS behavior. In [16], it is advised that extreme care must be taking in simulations involving Pareto distributions because they can lead to large errors due to their heavy tail characteristics. A traffic generator was written, which implements the traffic model described in Section IV. This generator may also be readily over-ridden by any other traffic model of choice. 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.

$$
\begin{aligned}
C_3^{in} = & \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{J_4}\sum_{j_{out}=1}^{J_{out}} \left(\frac{j_{out}}{\mu_{out}}+\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,j_{out},A_3,s_{out}^{j_{out}},1) \\
& + \sum_{j_1=1}^{2j_3+2}\sum_{j_2=0}^{(2j_3+2)}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{j_3} \left(\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_1^1,0) \\
& + \sum_{j_1=2j_3+3}^{J_1}\sum_{j_2=2j_3+3}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=j_3+1}^{J_4} \left(\frac{1}{\mu_1}+\frac{2j_3+1}{\mu_1}+\frac{2j_3+2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_3}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_1^1,0) \\
& + \sum_{j_1=1}^{2j_3+1}\sum_{j_2=0}^{(2j_3+2)}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{j_3} \left(\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_1^2,0) \\
& + \sum_{j_1=2j_3+2}^{J_1}\sum_{j_2=2j_3+3}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=j_3+1}^{J_4} \left(\frac{1}{\mu_1}+\frac{2j_3}{\mu_1}+\frac{2j_3+2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_3}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_1^2,0) \\
& + \sum_{j_1=0}^{2j_3}\sum_{j_2=1}^{(2j_3+2)}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{j_3} \left(\frac{j_2}{\mu_2}+\frac{j_1}{\mu_1}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_2^1,0) \\
& + \sum_{j_1=2J_3+1}^{J_1}\sum_{j_2=2j_3+3}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=j_3+1}^{J_4} \left(\frac{1}{\mu_2}+\frac{2j_3}{\mu_1}+\frac{2j_3+1}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_3}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_2^1,0) \\
& + \sum_{j_1=0}^{2j_3}\sum_{j_2=1}^{2j_3+1}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{j_3} \left(\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_2^2,0) \\
& + \sum_{j_1=2j_3+1}^{J_1}\sum_{j_2=2j_3+2}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=j_3+1}^{J_4} \left(\frac{1}{\mu_2}+\frac{2j_3}{\mu_1}+\frac{2j_3}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_3}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_2^2,0) \\
& + \sum_{j_1=0}^{2j_3}\sum_{j_2=0}^{2j_3}\sum_{j_3=1}^{J_3-1}\sum_{j_4=0}^{j_3} \left(\frac{j_3}{\mu_3}+\frac{j_4}{\mu_4}+\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_3,0) \\
& + \sum_{j_1=2J_3+1}^{J_1}\sum_{j_2=2j_3+1}^{J_2}\sum_{j_3=1}^{J_3-1}\sum_{j_4=j_3+1}^{J_4} \left(\frac{j_3}{\mu_3}+\frac{j_3}{\mu_4}+\frac{2j_3}{\mu_1}+\frac{2j_3}{\mu_2}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_3,0) \\
& + \sum_{j_1=0}^{2j_3+2}\sum_{j_2=0}^{2j_3+2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=1}^{j_3+1} \left(\frac{j_4}{\mu_4}+\frac{j_3}{\mu_3}+\frac{j_1}{\mu_1}+\frac{j_2}{\mu_2}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_4,0) \\
& + \sum_{j_1=2j_3+3}^{J_1}\sum_{j_2=2j_3+3}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=j_3+2}^{J_4} \left(\frac{1}{\mu_4}+\frac{2j_3+2}{\mu_1}+\frac{2j_3+2}{\mu_2}+\frac{j_3}{\mu_3}+\frac{j_3}{\mu_4}\right)\pi(j_1,j_2,j_3,j_4,A_3,s_4,0) \qquad (12)
\end{aligned}
$$

- 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.

Additionally, a specialist numerical algorithm [17] was implemented for computing the variance to combat the numerical instability in the aggregation of the QoS statistics. The numerical solution of the queuing systems modeled in this paper amounts to calculating the transition probabilities of the corresponding Markov chain to generate the transition probability matrix $\mathbf{P}$. The steady-state distribution π can then be obtained by solving the left-eigenvalue problem: $\pi\mathbf{P} = \pi$. The effects of various grant policies, varying the traffic arrival patterns (packet and session arrival rates) as well as the effects of varying the degree of self-similarity (Hurst parameter: $0.5 < H < 1$) on the QoS parameters were extensively investigated. Considering that most of the available empirical evidence [18] suggest that $H \sim [0.7, 0.85]$ is the region of interest in network traffic, we show the effect of three different Hurst parameters (i.e., 0.80, 0.85, and 0.95) on the queuing delay (Figs. 3 and 4) and then on the PLR (Figs. 5 and 6). Due to the nature of conversational traffic (UGS traffic in our paper), its packet size is very small, compared to the sizes of other traffic types. On the other hand, the average arrival rate of conversational traffic is much higher. Other traffic loads have been given as a performance reference and they can be finely tuned so as to reflect real traffic. However, in this paper, we do not focus on the behavior of a particular traffic type and so we simply use a reasonable (not too large and not too small) packet arrival rate to show how much each queue is built up. For the results shown, the highest priority queues (I-UGS and O-UGS), session arrival rate was set to 3 per second and the in-session packet arrival rate to 25 per second. For all the other queue types, we set the session arrival rate to 25 per second and the in-session packet arrival rate to 3 per second. To be able to extract the numerical solution, the queue size was chosen such that each queue has a maximum capacity to hold 10 packets because having bigger queues that hold more number of packets makes it a bit harder to extract the numerical solution of the large Markov chain accurately. The grant policy is round-robin and a grant is active at each ONU-BS for 40% of the ONU-BS's CPU time. In other words, 40% of the time, out-

$$C_4^{in} = \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{J_4}\sum_{j_{out}=1}^{J_{out}} \left(\frac{j_{out}}{\mu_{out}} + \frac{2j_3+2}{\mu_1} + \frac{2j_3+2}{\mu_2} + \frac{j_3}{\mu_3} + \frac{j_3}{\mu_4}\right)\pi(j_1, j_2, j_3, j_4, j_{out}, A_3, s_{out}^{j_{out}}, 1)$$
$$+ \sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{J_4} \left(\frac{1}{\mu_1} + \frac{2j_3+1}{\mu_1} + \frac{2j_3+2}{\mu_2} + \frac{j_3}{\mu_3} + \frac{j_3}{\mu_4}\right)\pi(j_1, j_2, j_3, j_4, A_3, s_1^1, 0)$$
$$+ \sum_{j_1=1}^{J_1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{J_4} \left(\frac{1}{\mu_1} + \frac{2j_3}{\mu_1} + \frac{2j_3+2}{\mu_2} + \frac{j_3}{\mu_3} + \frac{j_3}{\mu_4}\right)\pi(j_1, j_2, j_3, j_4, A_3, s_1^2, 0)$$
$$+ \sum_{j_1=0}^{J_1}\sum_{j_2=1}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{J_4} \left(\frac{1}{\mu_2} + \frac{2j_3}{\mu_1} + \frac{2j_3+1}{\mu_2} + \frac{j_3}{\mu_3} + \frac{j_3}{\mu_4}\right)\pi(j_1, j_2, j_3, j_4, A_3, s_2^1, 0)$$
$$+ \sum_{j_1=0}^{J_1}\sum_{j_2=1}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=0}^{J_4} \left(\frac{1}{\mu_2} + \frac{2j_3}{\mu_1} + \frac{2j_3}{\mu_2} + \frac{j_3}{\mu_3} + \frac{j_3}{\mu_4}\right)\pi(j_1, j_2, j_3, j_4, A_3, s_2^2, 0)$$
$$+ \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}\sum_{j_3=1}^{J_3-1}\sum_{j_4=0}^{J_4} \left(\frac{1}{\mu_3} + \frac{2j_3}{\mu_1} + \frac{2j_3}{\mu_2} + \frac{j_3-1}{\mu_3} + \frac{j_3}{\mu_4}\right)\pi(j_1, j_2, j_3, j_4, A_3, s_3, 0)$$
$$+ \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}\sum_{j_3=0}^{J_3-1}\sum_{j_4=1}^{J_4} \left(\frac{1}{\mu_4} + \frac{2j_3+2}{\mu_1} + \frac{2j_3+2}{\mu_2} + \frac{j_3}{\mu_3} + \frac{j_3}{\mu_4}\right)\pi(j_1, j_2, j_3, j_4, A_4, s_4, 0) \quad (13)$$

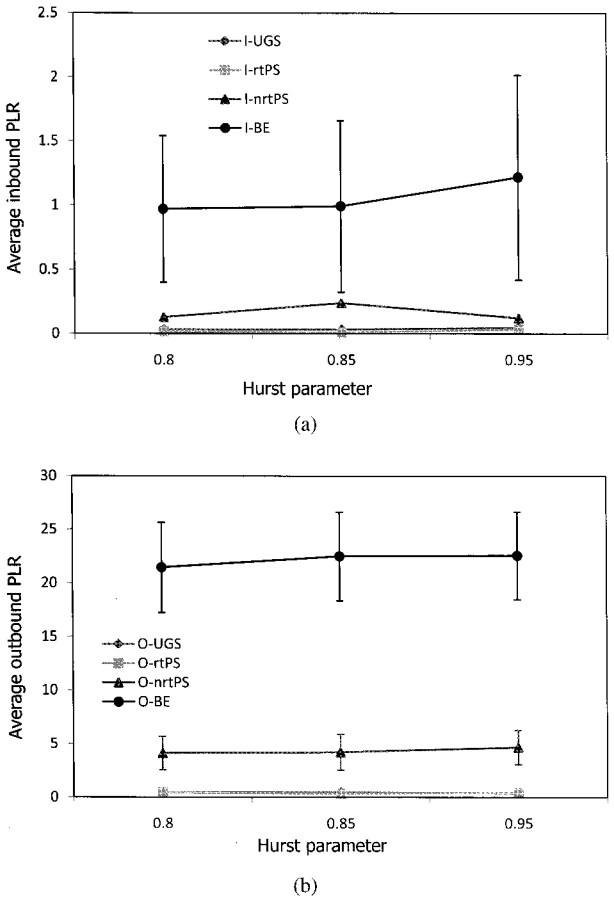

Fig. 5. PLR for PQ discipline: (a) Inbound and (b) outbound.

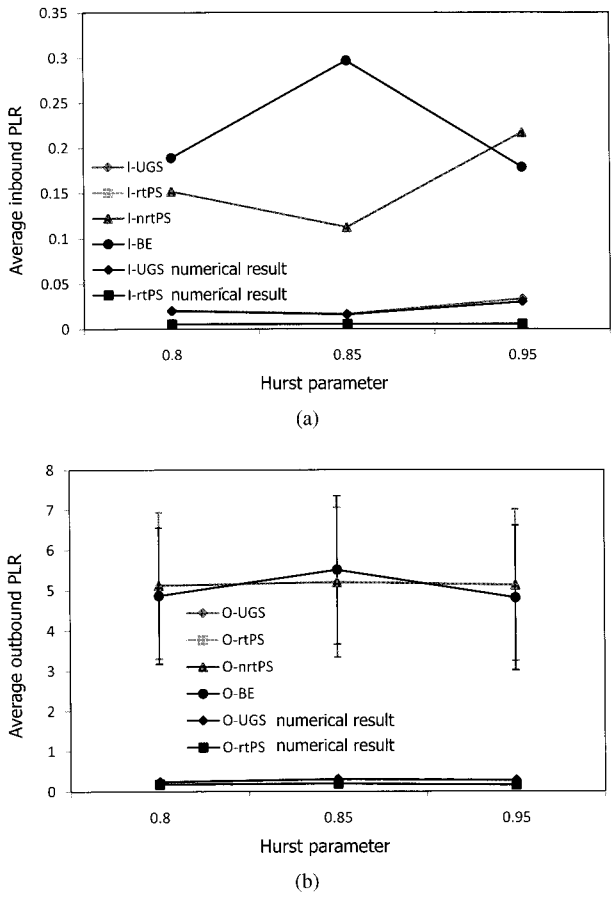

Fig. 6. PLR for CQ discipline: (a) Inbound and (b) outbound.

bound queues are being served while 60% of the time inbound queues are being served.

The results, Figs. 3–6, show the effect of increasing the degree of self similarity (Hurst parameter) on various QoS parameters. As the degree of self-similarity increases, we can notice that the lower priority queues are seen to do significantly better under CQ than under PQ discipline. The outbound queues are observed to perform worse because the CPU spends more time serving the inbound queues. The numerical results closely match with simulation results thus validating the closed form ex-

pressions. The results therefore show the characteristics of the PQ and CQ scheduling logics under each degree of self similarity. Further, we can also observe that as the degree of self-similarity (based on Hurst parameter) increases, there is slight difference in the queueing delay particularly for high priority queues (queue 1 and queue 2). Further, it can also be noted that in some cases, as the value of Hurst parameter increases, we can see a decrement in the QoS parameter value such as PLR. This is not surprising because of the fact noted by authors in [19]. The authors in [19] have shown through an extensive study conducted using the real traffic traces, that Hurst parameter alone is not sufficient to predict the queuing performance and sometimes can shown an inversion fact. Although each data point in the figures was obtained from a hundred simulation runs and a 0.95 confidence interval, we do not show all the error bars because in some cases, it distorts the clarity of the figures. The error bars varying the results is not surprising because in [16], after a detailed study, the authors concluded that care must be taken in simulations involving Pareto distributions as they can lead to large errors due to their heavy tail trait.

VI. RELATED WORK

In this section, we give an overview of the related work. Over the past few years, a lot of interest has been generated both by the academia and service providers in the convergence of wireless and wired technologies. Some of these interests which are well summarized in [1] have been specifically related to converging GEPON/EPON and WiMAX access networks. In [1] and [3] the authors essentially proposed architectures for the convergence of optical-wireless access network and the scope of work covered in them is substantially different from the scope of work that we present in this paper, especially in terms of the depth of queuing analysis. Our prior work [2] and [4], to the best of our knowledge, is the first effort to explicitly consider wired subscribers and consequently, the scheduling of inbound and outbound packets. Further, this paper is an extension of our prior work [4]. In [4], we conducted an analysis of the QoS behavior of the converged architecture under PQ; whereas, in this paper, we have concentrated on CQ. Moreover, while other studies only provide simulation based performance evaluation of the converged networks, we have provided a detailed analytical framework (fully supported and verified by numerical analysis and simulation results) under realistic traffic load and queuing conditions combined with a more robust queuing discipline.

The revelation that Ethernet and wireless data traffic exhibit self-similarity and long range dependency [7]–[11], has triggered a lot of work on the evaluation of the performance of communication networks under self-similar and long range dependent data traffic [2], [13]–[16], [18]–[23]. We have provided an extensive and detailed analytical framework for the converged network, which is substantially different from the prior work covered in [13]–[16], [18]–[23], especially in terms of the use of grant messages and the dynamics of inbound and outbound packets. Unlike most of the prior work, our queuing based results are capable of providing differential treatment to multiple classes of traffic which is a fundamental requirement in real network deployments for the provisioning of guaranteed QoS to the end-user.

VII. CONCLUSION

In this paper, we have contributed to the accurate modeling of converged optical-wireless networks, specifically a converged ONU-BS by providing a novel analytical models based on a G/M/1 queuing system under different classes of self-similar input traffic and CQ discipline, which is likely to be used in next-generation networks. We have developed an analytical framework, derived and presented closed-form mathematical expressions for the expected waiting time in the ONU-BS queues, end-to-end delay and the packet loss rate for the multiple classes of traffic under the CQ scheduling logic. We have also performed extensive simulation experiments to investigate and understand the behavior of self similar traffic in the converged access network and to observe how QoS parameters are affected. In addition, we have provided numerical analysis which closely match the simulation results; hence validating the closed-form mathematical expressions presented. This work can be used as a guide for the efficient allocation of network resources such as appropriate bandwidth allocation to individual traffic classes for the purpose of guaranteeing the QoS required by different applications while minimizing any excess provisioning of the network resources.

REFERENCES

[1] G. Shen, R. S. Tucker, and C. J. Chae, "Fixed mobile convergence architectures for broadband access: Integration of EPON and WiMAX," *IEEE Commun. Mag.*, vol. 45, no. 8, pp. 44–50, Aug. 2007.

[2] B. O. Obele and M. H. Kang, "Fixed mobile convergence: A self-aware QoS architecture for converging WiMAX and GEPON access networks," in *Proc. IEEE NGMAST*, Wales, UK, Sept. 2008, pp. 411–418.

[3] Y. Luo, S. Yin, T. Wang, Y. Suemura, S. Nakamura, N. Ansari, and M. Cvijetic, "Qos-aware scheduling over hybrid optical wireless networks," in *Proc. OFC/NFOEC*, Mar. 2007, pp. 1–7.

[4] B. O. Obele, M. Iftikhar, S. Manipornsut, and M. H. Kang, "On the analysis of the behavior of self-similar traffic in a QoS-aware architecture for integrating WiMAX & GEPON," *IEEE/OSA J. Optical Commun. Netw.*, vol. 1, no. 4, pp. 259–273, Sept. 2009.

[5] W. Odom and M. J. Cavanaugh, *Cisco DQOS Exam Certification Guide (IP Telephony Self-Study)*. Cisco Press, 2003.

[6] Configuring custom queuing. [Online]. Available: http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122cgcr/fqos_c/qcfcq.htm

[7] W. E. Leland, M. S. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[8] W. Willinger, M. S. Taqqu, and A. Erramilli, "A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks," *Stochastic Netw.: Theory and Appli.*, pp. 339–366, 1996.

[9] W. Carey, "Internet traffic measurement," *IEEE Internet Comput. Mag.*, vol. 5, no. 6, pp. 70–74, Nov. /Dec. 2001.

[10] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[11] M. Crovella and A. Bestavros, "Explaining world wide web traffic self-similarity," Boston University, CS Dept, Boston, MA 02215, Tech. Rep. TR-95-015, Aug. 1995.

[12] M. Caglar, "A long-range dependent workload model for packet data traffic," in *Proc. Mathematics of Operations Research*, vol. 29, 2004, pp. 92–105.

[13] M. Iftikhar, "Quality of service management in IP networks with self-similar traffic input," *Ph.D. dissertation submitted to the School of Information Technologies*, University of Sidney, 2008.

[14] M. Iftikhar, T. Singh, B. Landfeldt, and M. Caglar, "Multiclass G/M/1 queuing system with self-similar input and non-preemptive priority," *Elsevier J. Comput. Commun.*, vol. 31, no. 5, pp. 1012–1027, Mar. 2008.

[15] M. Iftikhar, B. Landfeldt, and M. Caglar, "Traffic engineering and QoS control between wireless diffServ domains using PQ and LLQ," in *Proc. ACM MobiWac*, Greece, Oct. 2007, pp. 120-129.
[16] D. Gross, J. Shortle, M. Fischer, and D. Masi, "Difficulties in simulating queues with Pareto service," in *Proc. WSC*, 2002.
[17] D. Kunth, *The Art of Computer Programming, Semi Numerical Algorithms. vol. 2.* 3rd ed., Addison-Wesley, Boston, pp. 232.
[18] K. Park and M. Crovella, "On the relationship between file sizes, transport protocols and self-similar network traffic," in *Proc. Int. Conf. Network Protocols*, Oct. 1996, pp. 171–180.
[19] R. Ritke, X. Hong, and M. Gerla, "Contradictory relationship between Hurst parameter and queueing performance (extended version)," *Springer J. telecommun.*, vol. 16, no. 1-2, pp. 159–175, Jan. 2001.
[20] L. Kleinrock, *Queuing Systems, Volume II: Computer Applications*. John Wiley & Sons, Inc.
[21] P. Ulanovs and E. Petersons, "Modeling methods of self-similar traffic for network performance evaluation," *Scientific Proc. RTU*, series 7, vol. 2, Telecom. and Electronics, pp. 40–49, 2002.
[22] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in *Proc. ACM Comput. Commun. Rev.*, vol. 24, Oct. 1994, pp. 269–280.
[23] S. Kasahara, "Internet traffic modeling: A Markovian approach to self-similar traffic and prediction of loss probability for finite queues," *IEICE Trans. Commun.: Special Issue on Internet Technol.*, vol. E84-B, no. 8, pp. 2134–2141, Aug. 2001.

Brownson Obaridoa Obele received his B.T. degree in Computer Engineering (first class honors) from the Rivers State University of Science and Technology (RSUST), Nigeria in 2001 and his Ph.D. in Information and Communication Engineering from the Korea Advanced Institute of Science and Technology (KAIST), South Korea in 2010. He is currently with the wireless communications and networking laboratory, School of Information and Communications of Gwangju Institute of Science and Technology (GIST), also in South Korea, as a Post Doctoral Researcher. After receiving his bachelor's degree, he worked with Allstates Trust Bank Plc (now Ecobank Plc), Nigeria as a Regional Communications Network Engineer between July 2001 and December 2003. In January 2004, he joined Global Communications Networks (GCN) Limited, Nigeria as a VSAT/Communications Engineer and rose to the position of Head of Engineering—South. He was with GCN until August 2006, when he left for South Korea to further his studies. He is Cisco certified CCNA and CCNP. Further, his biography has been selected for publication in "Who is Who in the World 2011 edition." His research interests include QoS, IPTV, IP traffic modeling and engineering, mobility management protocols, wireless and wired convergence networks, wireless communication, and networking technologies.

Mohsin Iftikhar received his B.S. in Electrical Engineering from the University of Engineering and Technology Lahore, Pakistan, M.S. in Telecommunications from the University of New South Wales, Australia and Ph.D. from University of Sydney, Australia in 1999, 2001, and 2008, respectively. Currently, he is working as an Assistant Professor at Computer Science Department, King Saud University, Riyadh, Saudi Arabia. He has published several papers in international conferences and journals and has been awarded several industrial and academic prizes/awards including (Siemens Prize for solving an industry problem 2006, Networks and Systems Prize in research project work, school of IT, 2007). He has been awarded an Endeavour Postgraduate Fellowship by Australian Government in 2008. Further, he has been nominated for inclusion in "Who is Who in the World 2010 edition." He is a Member of IEEE, ACM, and IET. His research interests includes QoS, IP traffic modeling, Markov chains, self-similar traffic modeling, network calculus, queuing theory, and polling models.

Minho Kang received his B.S., M.S., and Ph.D. degrees in Electrical Engineering from Seoul National University, University of Missouri-Rolla, and the University of Texas at Austin in 1969, 1973, and 1977, respectively. From 1977 to 1978, he was with AT&T Bell Laboratories, Holmdel, NJ. From 1978 to 1989, he was a Department Head and a Vice-President at the Electronics and Telecommunications Research Institute (ETRI), Korea. In addition, from 1985 to 1988, he served as the Electrical and Electronics Research Coordinator at the Korean Ministry of Science and Technology. During 1990–1998, he was an Executive Vice-President at Korea Telecom (KT) in charge of R&D, quality assurance, and overseas business development groups. In 1999, he joined the Information and Communications University (ICU) now KAIST-ICC, Korea as a Professor, and is presently the Director of the Optical Internet Research Center, a position he has held since 2000. He is also presently the Vice President of KAIST IT Convergence Campus. He served as the Study Group Chairman at the Asia Pacific Tele-Community of Bangkok during 1996-1999, is a Member of the National Academy of Engineering of Korea and a Senior Member of IEEE. He is an author of Broadband Telecommunications Technology, published in 1993 and revised in 1996 by Artech House. He is also an Associate Editor of IEEE Optical Communications and Networks Magazine.