An Analytical Approach for Dynamic Wavelength Allocation in WDM-TDMA PONs Servicing ON-OFF Traffic

John S. Vardakas, Ioannis D. Moscholios, Michael D. Logothetis and Vassilios G. Stylianakis

Abstract—Optical access systems are now considered a feasible alternative to the predominant broadband access technologies, while, at the same time, Passive Optical Networks (PONs) are viewed as an attractive and promising type of fiber access systems. In this paper we present and analyze three basic dynamic wavelength allocation scenarios, for a hybrid Wavelength Division Multiplexing (WDM) - Time Division Multiple Access (TDMA) PON. We propose new teletraffic loss models for calculating call-level performance measures, like connection failure probabilities (due to unavailability of a wavelength) and call blocking probabilities (due to the restricted bandwidth capacity of a wavelength). The PON accommodates bursty service-classes of ON-OFF traffic. The proposed models are extracted from one-dimensional Markov chains which describe the wavelengths' occupancy in the PON, and two-dimensional Markov chains which describe the bandwidth occupancy inside a wavelength. The accuracy of the proposed models is validated through simulation and is found to be quite satisfactory. Moreover, these models are computationally efficient, because they are based on recursive formulas.

Index Terms—wavelength division multiplexing; passive optical networks; dynamic wavelength allocation; blocking probability; connection failure; delay; ON-OFF Traffic.

I. INTRODUCTION

THE wide spread of bandwidth-intensive applications and the expansion of the multimedia services that are offered to end users require the provision of an access network capable of providing broadband communication with Quality of Service (QoS) guarantees [1]. A solution is offered by new generation fiber-based access technologies which are under swift development. However, the penetration of the optical fiber down to the premises of the end users brings up economical considerations, since the deployment of a new infrastructure requires capital expenditures in optical fiber installation and excessive-cost equipment [2]. The Passive Optical Network (PON) is a cost-effective fiber-based access solution, given that the exploitation of a PON minimizes the total length of the installed optical fiber and reduces the network cost by using inexpensive passive elements [3].

The basic building blocks of a PON are a centralized Optical Line Terminal (OLT), located in the central office, and a number of Optical Network Units (ONUs) located at the users' premises up to 20 km far from the OLT [2] (Fig. 1). A Passive Optical Splitter/Combiner (PO-SC) broadcasts traffic from the OLT to the ONUs (downstream direction) and transmits traffic from the ONUs to the OLT through one fiber (upstream direction). In common PON configurations where a Time Division Multiplexing (TDM) scheme is employed, downstream traffic destined towards a specific ONU, although it is received by all ONUs, it is accepted by the appropriate ONU only, thanks to the packet header. In the upstream direction, a Time Division Multiple Access (TDMA) scheme prevents data collision among different ONUs. The specifications of these TDM-PONs have been discussed by both ITU-T and IEEE. ITU-T recommends the Asynchronous Transfer Mode (ATM)-PON and its extensions, the Broadband PON (BPON), and the Gigabit PON (GPON) [4]. In addition, the Ethernet PON (EPON) has been discussed by the IEEE 802.3ah study group [5].

The main disadvantage of the aforementioned TDM-PONs is the inability to increase the transmission speed. The Wavelength Division Multiplexing (WDM) technique is a graceful PON evolution that exploits the high bandwidth capacity of the optical fiber by allocating a dedicated wavelength to an ONU and therefore increases the number of end users [6]-[8]. In WDM-PONs, the PO-SC is replaced by an Array Waveguide Grating (AWG); this is a passive optical device which (de)multiplexes a number of wavelengths in order for multiple wavelengths to be employed either in upstream or downstream directions [7].

Wavelength allocation schemes can be divided into two schemes, static and dynamic, and result in inter-wavelength statistical multiplexing [9]. In the Static Wavelength Allocation (SWA) scheme, a specific wavelength is dedicated to each ONU; therefore the number of ONUs is limited by the number of the supported wavelengths. The main limitation of this scheme refers to the inability of the network to support additional ONUs. The Dynamic Wavelength Allocation (DWA) scheme assigns a wavelength to an ONU based...
on the requested bandwidth, thus allowing each ONU to share the network resources in a more efficient way. Its main drawback is that ONUs should transmit at any of the supported wavelengths. This is a cost-prohibitive solution, since the transceivers installed at each ONU must operate at different frequencies. To provide inexpensive solutions to this disadvantage, several techniques have been proposed [10]–[12].

In this paper we study a hybrid WDM-TDMA PON, which is a promising PON configuration for a low-cost access network and full-service offerings [13]. Each ONU supports user-connections (calls) of different service-classes, by allocating a wavelength dynamically and multiplexing the service-classes through a TDMA scheme on this wavelength. More precisely, calls arrive to an ONU according to a Poisson process. At each ONU a number of users is accommodated, and gain access to the PON’s resources, through the establishment of a new ONU-OLT connection, or by using an existing connection. An ONU-OLT connection is realized through the allocation of an available wavelength in the PON. We present basically three scenarios for DWA in the upstream direction. Each scenario introduces a different wavelength release procedure according to the bandwidth occupancy of the wavelength, and achieves different call-level performance. The first scenario is the “Primary Wavelength Release” (P-WR) scenario. The second scenario, named “Delay Wavelength Release” (D-WR) scenario, performs better than the first scenario in terms of connection failure probability (CFP, due to unavailability of a wavelength), but introduces delay in servicing of some calls. In between is the CFP performance of the third scenario, which combines the P-WR and D-WR in order to reduce the mean number of calls that experience delay. The third scenario is named “Primary-Delay Wavelength Release” (PD-WR) scenario. For each scenario, we examine the call level performance of the hybrid WDM-TDMA PON with DWA and propose a teletraffic loss model, in order to calculate the CFP and the other performance measures related to the specific scenario.

Calls under service are of ON-OFF type; i.e. they alternate between periods of transmission (ON) and silent periods (OFF). More precisely, accepted calls enter the system via state ON and may alternate between states ON and OFF, or remain always in state ON. After the transfer of a call to state OFF, the call releases the bandwidth held in state ON, so that this bandwidth becomes available to new arriving calls. When a call attempts to return to state ON, it requests its bandwidth. If it is available, a new ON-period (burst) begins; otherwise burst blocking occurs and the call remains in state OFF. Call blocking occurs when a call cannot enter the system, due to lack of bandwidth. We use the formulas presented in [14] to determine the Burst Blocking Probability (BBP) and the Call Blocking Probability (CBP) that occurs after the establishment of the OLT-ONU connection, due to the restricted bandwidth capacity of a wavelength. Note that the BBP is the probability that an accepted call remains in state OFF but no call loss happens, while the CBP is the probability that a new call is blocked and lost.

The proposed loss models are extracted either from one-dimensional Markov chains which describe the wavelengths occupancy in the PON, or two-dimensional Markov chains (due to the ON-OFF traffic model) which describe the bandwidth occupancy inside a wavelength. For instance, the CFP determination is obtained by solving one-dimensional Markov chain, while it involves parameters which are obtained from the two-dimensional Markov chain. The proposed models are computationally efficient, because they are based on recursive formulas. The accuracy of all models is validated through simulation and is found to be quite satisfactory.

To the best of our knowledge, call-level performance analysis in WDM PONs has not been published by other authors. Before our present study, in which we emphasize not only on DWA schemes but also on the bursty nature of traffic, as it is denoted by the ON-OFF model, in [15] and [16] we have provided analytical models for the calculation of blocking probabilities in WDM-TDM PONs under stream and elastic traffic. Several other studies have been appeared in the literature providing an analytical solution to the DWA problem in WDM PONs; most of them consider WDM-EPONs [9], [17], [18]. Inter-ONU scheduling algorithms are proposed which are based on the 802.3ah protocol. The basic philosophy behind is that in each ONU both a transmission window and a wavelength are allocated.

The remainder of this paper is organized as follows. In Section II we describe the basic network architecture and its operation. In Section III we provide an overview of the multi-rate ON-OFF traffic model. In the following sections IV, V and VI, we analyze the three DWA scenarios: each section corresponds to each scenario. Section V is splatted into two subsections, A and B, which correspond to two functions involved in the wavelength release procedure for the D-WR scenario. Section VII is the evaluation section, where both analytical and simulation results are presented and discussed. We conclude in Section VIII. In Appendix we provide a list of the notation used in the paper.

II. The Network Model

We consider a WDM-TDM PON, as it is shown in Fig. 2. The PON is connected to a wide area or a metropolitan network through the OLT. In the upstream direction (under consideration), the PON supports W wavelengths (typically in the 1310 nm switching window). N ONUs are connected to the OLT through the AWG. In the case where W < N, DWA is needed; otherwise an ONU could allocate a wavelength statically. Each ONU accommodates K service-classes; the calls of each service-class are of ON-OFF type, when they are in service.

Service-class k (k = 1, ..., K) calls arrive to an ONU according to a Poisson process with mean arrival rate λk. They are conveyed through a connection established between ONU and OLT that is realized through one (only) wavelength in the PON. The service of a service-class k call is realized by allocating a specific number of time-slots, bk, in a frame for the entire duration of the call (TDMA scheme). The total number of time-slots in a frame defines the fixed bandwidth capacity, C, of the wavelength (static bandwidth allocation inside the wavelength).

When no other calls from this ONU are in service, the connection establishment of the ONU to the OLT is realized by the first arriving call from any service-class. If the OLT finds a free wavelength, it is assigned to the ONU; the ONUs are assumed to be color-free (i.e. any PON wavelength can be assigned to an ONU). Thus, a connection is established and the call is conveyed; otherwise the call is blocked and lost.
We consider the connection between an ONU and the OLT through one wavelength as a single link of \( C \) bandwidth units (b.u.), which accommodates the \( K \) service-classes. A b.u. corresponds to a time-slot. Service-class \( k \) calls are accepted for service while being in state ON and may constantly remain in state ON for their entire duration, or alternate between states ON and OFF. Throughout state ON they constantly remain in state ON for their entire duration, or may be blocked and lost. If a wavelength is not assigned to an ONU upon a service-class \( k \) call arrival, the ONU firstly sends to the OLT a control packet to request a free wavelength. If a free wavelength is found, the OLT responds with a control packet which contains the frequency of the assigned wavelength, and adjusts a receiver to this frequency. At the re-connection of this control packet, the ONU adjusts a transmitter to this frequency.

In all cases we calculate the Connection Failure Probability (CFP), which occurs due to the unavailability of a free wavelength in the link between the AWG and the OLT. We also calculate the delay that OFF-calls suffer from the time instant the wavelength is released until the re-connection of the ONU to the OL, for the P-WR and PD-WR scenarios. Each DWA scenario introduces an alternative wavelength release procedure, based on the bandwidth occupancy of the wavelength that is assigned to an ONU. In the P-WR scenario, a wavelength is released when all calls (either in state ON or OFF) have terminated. The CFP performance of this scenario could be improved by accelerating the wavelength release procedure/rate. This is achieved by the D-WR scenario: A wavelength is released when no calls in state ON exist in the system, but do exist in state OFF, when the occupied fictitious bandwidth is up to a percentage of the fictitious capacity; this percentage is defined as an increase or a constant function of the number of occupied wavelengths in the PON e.g. if there are 1 out of 4 wavelengths seized in the PON at the time-point of possible wavelength release, this percentage could be 3%, arbitrarily chosen, if there are 2 seized wavelengths, out of 4, this percentage could be higher, say 5%; if there are 3 seized wavelengths, out of 4, this percentage could be 8%, while if there are 4 seized wavelengths out of 4, this percentage could be 12%; alternatively, it could be 4%, arbitrarily chosen, irrespectively from the number of seized wavelengths). Such a relation allows greater ONU connectivity by delaying the servicing of some OFF-calls; they are obliged to postpone their service until the next ONU connection (upon a new call arrival).

Under a constant function, the wavelength release procedure is independent from the number of occupied wavelengths; the desirable wavelength release procedure would have a low rate when the wavelength occupancy is low and a high rate when the wavelength occupancy is high. Under an increase function, the lower the number of occupied wavelengths, the lower mean service rate of the wavelength; they are both result in smaller number of calls that suffer delay. This number is further reduced by the PD-WR scenario (at the expense of CFP).

Fig. 2. A WDM-TDMA PON servicing ON-OFF traffic.
III. Overview of the Multirate ON-OFF Model

The following analysis refers to a single link of capacity \( C \) b.u. that accommodates \( K \) service-classes of Poisson arriving calls, which, under service, behave according to the ON-OFF model of Fig. 2. This analysis is applicable to the WDM-TDMA PON, where each wavelength is equivalent to the single link.

According to [18], [19], a new service-class \( k \) call will be accepted in the system, if it satisfies the following both constraints:

\[
j_1 + b_k \leq C \quad \text{and} \quad j_1 + j_2 + b_k \leq C^* \tag{1}
\]

where \( j_1 \) and \( j_2 \) represent the occupied bandwidth of the real and the fictitious link, respectively. The first constraint guarantees that the bandwidth of the new call together with the occupied bandwidth does not exceed the capacity of the real link. The second constraint restrains the system to accept new calls, when most of system calls are in state OFF. If we denote by \( \Omega \) the set of the permissible states, then the distribution \( \vec{j} = (j_1, j_2) \), denoted as \( q(\vec{j}) \) can be calculated by the following two-dimensional recursive formula:

\[
\sum_{i=1}^{2} \sum_{k=1}^{2} b_{i,k,s} p_{i,k,s} q(\vec{j} - B_{i,k}) = j_i q(\vec{j})
\]

where

\[
\vec{j} \in \Omega \Rightarrow \left\{ j_1 \leq C \cap \left( \sum_{s=1}^{2} j_s \leq C^* \right) \right\}
\]

The parameter \( s \) refers to the links \( s=1 \) indicates the real link, \( s=2 \) the fictitious link, while \( i \) refers to the ON-OFF states \( i=1 \) specifies the state ON, \( i=2 \) specifies the state OFF. Also,

\[
b_{i,k,s} = \begin{cases} b_k & \text{if } s = i \\ 0 & \text{if } s \neq i \end{cases} \quad \tag{4}
\]

and \( B_{i,k} = (b_{i,k,1}, b_{i,k,2}) \) is the i,k row of the \( (2K \times 2) \) matrix \( B \), with elements \( b_{i,k,s} \). Finally, \( p_{i,k} \) is the utilization of the \( i^{th} \) link by service-class \( k \):

\[
p_{i,k} = \frac{e_{i,k}}{\mu_{i,k}} = \begin{cases} \frac{\lambda_k}{1-\sigma_k} & \text{for } i = 1 \\ \frac{\lambda_k}{1-\sigma_k} & \text{for } i = 2 \end{cases}
\]

where \( e_{i,k} \) is the total arrival rate of service-class \( k \) calls to the \( i^{th} \) state and \( \mu_{i,k} \) is the mean service time of service-class \( k \) calls in the \( j^{th} \) state, exponentially distributed. The CBP of service-class \( k \) is given by the following formula:

\[
P_{b_k} = \sum_{\vec{j}|\{j_1+b_{k,1}j_2+b_{k,2}\geq C^*}\} G(\vec{j})
\]

where \( G = \sum_{\vec{j}\in\Omega} q(\vec{j}) \) is the normalization constant. The calculation of the BBP is based on the summation of the departure rates over the burst blocking state-space \( \Omega^* \):

\[
\vec{j} \in \Omega^* \Rightarrow \left\{ C-b_k+1 \leq j_1 \leq C \cap \left( \sum_{s=1}^{2} j_s \leq C^* \right) \right\}
\]

Therefore, the BBP is given by the normalized summation of the rates from above state-space, over the rates from all states of the entire state-space \( \Omega \), as it is defined in Eq. (3)

\[
(\vec{1}): \quad P_{b_k} = \sum_{\vec{j}|\{j_1+b_{k,1}j_2+b_{k,2}\leq C^*}\} G(\vec{j})
\]

IV. The Primary Wavelength Release (P-WR) Scenario

In the P-WR scenario, an ONU-OLT connection is terminated when all calls both in the ON and OFF states are terminated. The CFP determination for the P-WR scenario is based on the knowledge of the occupied wavelength distribution. To this end, we construct the one-dimensional Markov chain of the system, and formulate the state transition diagram of Fig. 3, where each state \( w \) \( (w=0,1,\ldots, W) \) represents the number of the occupied wavelengths in the PON. The transition rate from one state to the next upward state represents the rate of the connection establishment; it is a function of the number of ONUs that have not been connected to the OLT yet. Therefore, when the system is in state \( [w] \), an arriving call will transfer it to the state \( [w+1] \); this transition will occur \( (N-(w-1)) \) times per unit time (Fig. 3), where \( N-(w-1) \) is the number of ONUs which have not been connected yet and \( \lambda = \sum_{k=1}^{K} \lambda_k \) is the total arrival rate of all service-class calls from one ONU.

The downward transition rate defines the mean rate \( R_{P} \) by which a wavelength is released; its determination depends on the applied wavelength-release scenario. The transition from state \( [w] \) to state \( [w-1] \) is realised \( wR_{P} \) times per unit time. \( (R_{P} \text{ can be seen as the mean service rate of a wavelength}) \). In the present P-WR scenario, the wavelength is released after the departure of the system of the last ON-call of any service-class, while, simultaneously, no calls are present in the system in state OFF. The rate \( R_{P} \) is equal to the sum of the rates from state \( (b_k,0) \) for all service-classes that lead to state \( (0,0) \), given that the system is in state \( (b_k,0) \); therefore:

\[
R_{P} = \sum_{k=1}^{K} \mu_{1k} y_{1k}(b_k,0)(1-\sigma_k)q(b_k,0)
\]

\[
= \sum_{k=1}^{K} \mu_{1k} y_{1k}(b_k,0)(1-\sigma_k)q(0,0)\frac{q(0,0)}{q(0,0)}
\]

where \( y_{1k}(\vec{j}) \) is the mean number of service-class \( k \) calls in the wavelength when the system is at state \( \vec{j} = (j_1, j_2) \); \( q(\vec{j}) = q(j_1, j_2) \) is the occupancy distribution inside a wavelength and is given by Eq. (2); \( i(0) \) is the conditional probability that i b.u. are occupied only in the real link, given that the wavelength is occupied; and \( G = \sum_{\vec{j}\in\Omega} q(\vec{j}) \) is the normalization constant.
The probability \( P(w) \) that \( w \) wavelengths are occupied in the PON can be derived from the global-balance rate equations of the state transition diagram of Fig. 3:

\[
(N-(w-1))\lambda P(w-1) + ((w+1)R_p)P(w+1) = (N-w)\lambda + wR_p)P(w)
\]

where \( P(w) = 0 \) for \( w < 0 \) and \( w > N \).

By summing up Eq. (11) side by side, from \( w = 0 \) to \( w - 1 \) (this procedure is known as the ladder method [20]), we get the following recurrence formula:

\[
P(w) = \frac{\lambda}{R_p} \frac{N-(w-1)}{w} P(w-1), \quad w = 1, \ldots, W \tag{12}
\]

Equation (12) stands for the local-balance rate equation of the state transition diagram (Fig. 3), and denotes that the system has a product form solution [20]. Through consequent applications of Eq. (12) and the normalization condition \( \sum_{w=0}^{W} P(w) = 1 \), we derive the occupied wavelength distribution in the PON:

\[
P(w) = \left( \frac{\lambda}{R_p} \right)^w \prod_{z=1}^{w} \frac{[N-(z-1)]}{w!} P(0) \tag{13}
\]

where \( P(0) = \left[ \sum_{w=0}^{W} \left( \frac{\lambda}{R_p} \right)^w \prod_{m=1}^{w} \frac{[N-(m-1)]}{m!} \right]^{-1} \)

The CFP is determined by Eq. (13) when \( w = W \), i.e. \( P(W) \), since a connection between an ONU and the OLT cannot be established (connection failure occurs) only when all \( W \) wavelengths are busy.

The proposed formula, Eq. (13), is computationally efficient in realistic applications, since the typical values of parameters \( W \) and \( N \) are small numbers (e.g., \( W = 16 \), \( N = 24 \)). For large values of these parameters one can solve Eq. (12) recursively by setting \( P(0)=1 \) and normalizing all values over the \( \sum_{w=0}^{W} P(w) \).

As far as the CBP and BBP inside a wavelength are concerned, they are determined by Eq. (6) and Eq. (8), respectively.

V. THE DELAY WAVELENGTH RELEASE (D-WR) SCENARIO

Under this scenario, we recall that the wavelength is released after the departure from the system of the last ONU-call (the occupied b.u. in the real link become zero), while at the same time the occupied fictitious b.u. are up to a percentage of the fictitious capacity. We shall firstly assume that this percentage is a constant function of the number of occupied wavelengths in the PON, and afterwards, we shall assume that it is an increase function. This consideration facilitates the inductive derivation of formulas for the CFP determination and the other performance measures.

A. Constant function

Let us assume that the constant function is \( mC^\ast \), where \( m \) is an arbitrarily chosen parameter \( \in (0,1] \) (for easy computation, we choose the product \( mC^\ast \) to be integer). All other assumptions of the D-WR scenario are the same with those of the P-WR scenario. We construct the system Markov chain in the same way as in the P-WR scenario (Fig. 3), but the mean service rate of the wavelength becomes \( R_D \) and is determined by summing up all rates from state \( (b_k,j_2) \) to the wavelength-release state \( (0,j_2) \), for each service-class \( k \) \( (k = 1, \ldots, K) \) and for all \( j_2 \in [0, (mC^\ast)] \); therefore:

\[
R_D = \sum_{k=1}^{K} \sum_{j_2=0}^{mC^\ast} \mu_{k,y_{k,1}}(b_k,j_2)(1-\sigma_k)q(b_k,j_2) \tag{14}
\]

Equation (13) provides the occupied wavelength distribution \( P(w) \) for the D-WR scenario, by substituting \( R_P \) with \( R_D \). Then, the CFP is determined as \( P(W) \).

We now proceed to define the delay that occurs when bursts remain in state OFF, while waiting for the next ONU-connection upon a new call arrival. In order to calculate this delay, we first have to determine the interarrival time between two consecutive connection requests. As it is shown in Fig. 4, the interarrival time of ONU-OLT connection-establishment requests equals to the sum of the wavelength service time plus an idle period, where no calls are in service. This idle time, from the termination of the connection until the next ONU-OLT connection-establishment request is the delay that we have to calculate. Having determined the mean wavelength service rate \( R_D \), the wavelength (mean) service time in Fig. 4 is \( 1/R_D \). Since the call's interarrival time is by average equal to \( 1/\lambda \), we can find the mean number of calls which arrived within the wavelength service time, by dividing this time by \( 1/\lambda \). Then, by increasing this number by one, we count the next call arrival, which coincides with the new ONU-OLT connection-establishment request. Thus, the ONU-OLT connection-establishment request interarrival time is expressed as a multiple of the mean call's interarrival time:

\[
\frac{1}{\lambda} = \left( \left\lfloor \frac{1}{R_D} \right\rfloor + 1 \right) \frac{1}{\lambda} \tag{15}
\]

where \( \Lambda \) is the connection establishment request arrival rate and \( \lfloor x \rfloor \) denotes the smallest integer not exceeding \( x \).

Since connection failure may occur, the mean arrival rate of the successful connection requests \( \Lambda_s \) is reduced, as follows:

\[
\Lambda_s = \Lambda(1-P(W)) \tag{16}
\]

According to Fig. 4, the (average) delay that a burst suffers until the successful re-connection of the same ONU to the OLT is given by:

\[
T_D = \frac{1}{\Lambda_s} - \frac{1}{R_D} \tag{17}
\]

The number \( N_{D,k} \) of service-class \( k \) calls that suffer this delay in state OFF of state \( j = (b_k,j_2) \), where \( j_2 \in [1, (mC^\ast)] \) can be calculated by:

\[
N_{D,k} = \sum_{j_2=1}^{mC^\ast} y_{k,j_2} \tag{18}
\]

Under the D-WR scenario, the CFP is anticipated to be lower than that of the P-WR scenario, since the service rate \( R_D \) is higher than the service rate \( R_P \) (the established ONU-OLT connections are terminated more rapidly than in the D-WR scenario). However, this benefit is at the expense of the delay suffered by some calls.
B. Increase function

The increase function is proposed, because the rapid termination of connections (caused by the constant function) is not vital when the number of occupied wavelengths in the PON is low, given that in that case free wavelengths do exist for additional ONU-OLT connections. When \( w \) wavelengths are seized, the occupied b.u. in the fictitious link is less than or equal to \( \lceil (mC^*/W)w \rceil \), where \( \lceil x \rceil \) denotes the smallest integer not exceeding \( x \), and \( w = 1, \ldots, W \). Due to this relation, the mean service rate of a wavelength, \( R_D \), is a function of the number \( w \) of the occupied wavelengths in the PON. Therefore, it can be derived from Eq. (14) by changing the upper bound of the second summation:

\[
R_D(w) = \sum_{k=1}^{W} \sum_{j=0}^{k} \mu_{1k}y_{1k}(b_k,j_2)(1-\sigma_k)q(b_k,j_2) = \sum_{k=1}^{W} \sum_{j=0}^{k} \mu_{1k}y_{1k}(b_k,j_2)(1-\sigma_k)q(b_k,j_2) \frac{G_q(0,0)}{\bar{\sigma}}
\]

For small values of \( w \), it is possible that \( \lceil (mC^*/W)w \rceil = 0 \); then the internal summation of Eq. (19) equals \( \mu_{1k}y_{1k}(b_k,0)(1-\sigma_k)q(b_k,0) \), which means that Eq. (19) becomes identical to Eq. (10); i.e. in this case of the D-WR scenario calls do not experience delay (like the P-WR scenario).

Needless to say, a different increase function could be chosen for the D-WR scenario. Note that the proposed increase function was chosen so that it results in the constant function of \( mC^* \), when \( w = W \).

The CFP calculation, under the D-WR scenario with the increase function, is based on the construction of the Markov chain of Fig. 5, which is similar to that of Fig. 3; they differ in the downward transition rates only. To solve this chain and obtain the occupied wavelength distribution, \( P(w) \), we follow the same standard procedure, as in section IV.

\[
P(w) = \frac{(\lambda)^w}{w!} \prod_{z=1}^{W} \frac{[N-(z-1)]}{R_D(z)} P(0) \tag{20}
\]

where \( P(0) = \sum_{k=0}^{W} \frac{\lambda_k^w}{w!} \prod_{m=1}^{k} \frac{[N-(m-1)]}{R_D(m)} \).

Then, the CFP is determined as \( P(W) \) from Eq. (20), where \( R_D(w) \) is obtained from Eq. (19).

As far as the delay that some calls suffer is concerned, it can be calculated by using the method which is presented in the previous subsection. The only difference is that this delay is now a function of the number \( w \) of the occupied wavelengths in the PON. Equation (21) gives the distribution of this delay for \( w = 1, \ldots, W \):

\[
T_D(w) = \frac{1}{\Lambda_s} - \frac{1}{R_D(w)} \tag{21}
\]

where \( \Lambda_s \) is given by Eq. (16).

Having determined the distribution of the delay by Eq. (21), the average of the delay is given by:

\[
T_D = \frac{\sum_{w=1}^{W} T_D(w)}{W} = \frac{1}{\Lambda_s} - \frac{1}{\sum_{w=1}^{W} R_D(w)} \tag{22}
\]

where \( T_D(1) \) and \( T_D(W) \) are the minimum and the maximum values of the delay distribution, respectively.

To calculate the mean number \( N_{D,k} \) of service-class \( k \) calls in state OFF that experience delay per established connection, we sum up over the number of wavelengths, the average number of service-class \( k \) calls in state OFF, of state \( j = (b_k,j_2) \) for all \( j_2 \in [1, (mC^*/W)w] \):

\[
N_{D,k} = \frac{\sum_{w=1}^{W} \sum_{j_2=1}^{\lceil (mC^*/W)w \rceil} y_{2k}(b_k,j_2)}{W} \tag{23}
\]

The division by \( W \) is due to the fact that \( N_{D,k} \) is determined per established ONU-OLT connection.

VI. THE PRIMARY-DELAY WAVELENGTH RELEASE (PD-WR) SCENARIO

Under the PD-WR scenario, a connection is terminated after the departure from the system of the last ON-call, given that the fictitious link is empty and the number of occupied wavelengths in the PON is less than or equal to \( W - W_T \), where \( W_T \) is a pre-defined threshold. If the number of occupied wavelengths is greater than \( W - W_T \), the wavelength is released after the departure from the system of the last ON-call, given that the occupied fictitious bandwidth is up to a percentage of the fictitious capacity. In other words, if \( w \leq W - W_T \), the P-WR scenario is applied, whereas if \( w > W - W_T \), then the D-WR scenario is applied. To simplify the analysis, we consider the D-WR scenario with the so-called constant function; therefore, the mean service rate of a wavelength is given by:

\[
R_{PD}(w) = \begin{cases} 
\sum_{k=1}^{W} \mu_{1k}y_{1k}(b_k,0)(1-\sigma_k)q(b_k,0), & \text{if } w \leq W - W_T \\
\sum_{k=1}^{mC^*} \sum_{j_2=0}^{\lceil (mC^*/W)w \rceil} \mu_{1k}y_{1k}(b_k,j_2)(1-\sigma_k)q(b_k,j_2), & \text{if } w > W - W_T 
\end{cases} \tag{24}
\]

The CFP can be determined through Eq. (20) as \( P(W) \), while the \( R_D \) is substituted by \( R_{PD}(w) \) from Eq. (24). When \( w > W - W_T \), bursts, found in state OFF at the time instant that the ONU-OLT connection terminates, experience delay, which can be calculated through Eq. (17), when
substituting $R_D$ by $R_{PD}$, because, in that case, the D-WR scenario is applied.

Finally, the mean number $N_{PD,k}$ of service-class $k$ calls per established connection that experience delay, can be calculated by summing the average number of service-class $k$ calls in state OFF of state $j = (b_1, j_2)$ for all $j_2 \in [1, (mC^*)]$, over all possible numbers of occupied wavelengths, which exceeds $W - W_T$:

$$N_{PD,k} = \sum_{W=W_T}^{W} \sum_{j_2=1}^{(mC^*/W)} y_{2k}[b]$$

(25)

**VII. Evaluation And Discussion**

We study the performance of the proposed scenarios and evaluate the accuracy of the corresponding analytical models through application examples, where the analytical results are compared with simulation results. To this end, we consider the WDM-TDM PON of Fig. 2, with $N = 24$ ONUs. The network supports $W = 16$ wavelengths with (real) bandwidth capacity $C = 90$ b.u. The fictitious bandwidth capacity of each wavelength is $C^* = 100$ b.u. For presentation purposes, each ONU accommodates only $K = 2$ service-classes, with the following traffic characteristics:

$(b_1, b_2) = (20, 16)$ b.u.,

$(\mu_{11}, \mu_{21}) = (0.009, 0.008)$ sec.,

$(\mu_{21}, \mu_{22}) = (0.01, 0.01)$ sec.,

$(\sigma_1, \sigma_2) = (0.95, 0.9)$.

For the applicability of the proposed WR scenarios, we use the parameter $m = 0.2$ and $W_T = 4$ wavelengths. We simulate this network for the three WR scenarios while considering varying offered-traffic load, by using the Simscript II.5 simulation tool [21]. The simulation measures of CFP, CBP, BBP, delay and mean number of calls per service-class experiencing delay are obtained as mean values from 8 runs with confidence interval of 95%. For each simulation run we assume 3,000,000 originating calls, while we have assumed a stabilization time that corresponds to the first 300,000 calls, in order for the system to reach the steady state.

Starting with the P-WR scenario, we present analytical and simulation results for the CFP, CBP and BBP versus the call arrival rate in Tables I, II and III, respectively. For simplicity, we assume that the arrival rate is the same for the two service-classes. The analytical CFP results are obtained by Eq. (13), which requires the use of Eq. (10), Eq. (9) and Eq. (2). The analytical CBP results are obtained from Eq. (6) through the use of Eqs (2)-(5), while the analytical BBP results are obtained from Eq. (8), which involves the use of Eqs (2)-(7) and Eq. (9). The comparison between the analytical and the corresponding simulation results of Tables I, II and III shows that the accuracy of the proposed models is completely satisfactory.

In Tables IV and V we present the CFP results under the D-WR with the constant function and the increase function, respectively. Table VI contains the results of the PD-WR scenario. As the comparison between the analytical and simulation results shows, the models accuracy is quite satisfactory for this call-level performance index. The small deviation of the CFP results is due to the fact that we have to neglect the transitions between specific states of the multirate ON-OFF model, because we terminate the ONU-OLT connection sharply. More specifically, under the D-WR scenario with the constant function, OFF to ON transitions,
determined over those calls that may experience delay, as Eq. (22) reveals. According to Figs. 7, 8 and 9 the accuracy of the proposed analytical expressions for the delay is completely satisfactory. In order to facilitate the comparison of the three scenarios, we comparatively present the analytical results of Figs. 7, 8, and 9 in Fig. 10. The D-WR scenario with the increase function achieves the lowest delay for small arrival rates (<2 calls/sec), while the same scenario with the constant function achieves the lowest delay for relatively high arrival rates (>2 calls/sec). The delay performance of the PD-WR scenario is similar to that of the D-WR scenario with the constant function, when the arrival rate is low (<2 calls/sec); however, further increase of the arrival rate has a negative effect on the delay. In general, as these figures reveal, the delay decreases with the increase of the arrival rate, down to a minimum value, while further increase of the arrival rate results in the increase of the delay; this behavior, as well as the oscillation under this increase depicted in the figures, is explained by studying the monotony of the delay as a function of the call’s arrival rate. For example, with the aid of Mathematica [22] one can find that the entire set of the solution of the equation \( \frac{\partial T}{\partial \lambda} = 0 \) constitutes a genuinely increasing discrete function. One can also find the ranges where \( \frac{\partial T}{\partial \lambda} > 0 \) or \( \frac{\partial T}{\partial \lambda} < 0 \) in order for the oscillations to be explained. Note that the delay generally increases with the total arrival rate: higher values of the arrival rate result in higher CFP; therefore an ONU has to wait longer in order to successfully establish a new connection. Also, from Eq. 17 one can observe that the delay is a function of the mean arrival rate of the successful connection requests \( \Lambda_s \) and the mean wavelength service rate \( R_D \). The \( R_D \) is a decreasing function of \( \lambda \), thus the service time \( 1/R_D \) increases with the increase of \( \lambda \). This is due to the fact that when the arrival rate is high, the ON-OFF system tends to high occupancy states; in this case the probability that the system is at state (0,0), which is the wavelength release state, is low and therefore the wavelength is occupied for a longer period (longer service time). As far as \( \Lambda_s \) is concerned, it is a function of \( \lambda \) and the CFP \( P(W) \) (see Eq. 16): \( \Lambda \) is an increasing function of \( \lambda \) (see Eq. 15), while the factor \( (1 - P(W)) \) decreases with the increase of \( \lambda \): a result that is expected (see also Table I). The combination of these reasons is responsible for the oscillations depicted in Figs. 7, 8, and 9.

Moreover, the delay generally increases with the total arrival rate: higher values of the arrival rate result in higher CFP; therefore an ONU has to wait longer in order to successfully establish a new connection. Note that from Eq. 17 the delay is a function of the mean arrival rate of the successful connection requests \( \Lambda_s \) and the mean wavelength service rate \( R_D \). The parameter \( R_D \) is a decreasing function of \( \lambda \), thus the service time \( 1/R_D \) increases with the increase of \( \lambda \). This is due to the fact that when the arrival rate is high, the ON-OFF system tends to high occupancy states; in this case the probability that the system is at state (0,0),
which is the wavelength release state, is low and therefore the wavelength is occupied for a longer period (longer service time). As far as \( \lambda_s \) is concerned, it is a function of \( \Lambda \) and the CFP \( P(W) \): is an increasing function of \( \lambda \) (see Eq. 15), while the factor \( (1 - P(W)) \) decreases with the increase of \( \lambda \), as it comes out the result of Table I; the combination of these variations is responsible for the oscillation that is depicted in Figs. 7, 8, and 9.

The last parameter that we examine, while comparing the efficiency of the proposed scenarios, is the mean number of service-class \( k \) calls per established connection that experience delay (in state OFF). In our application example, this number is 1 call of each service-class per established connection for the D-WR scenario with the constant function, while for both the D-WR scenario with the increase function and the PD-WR scenario is 0.0625 and 0.25 calls per established connection, for the 1st and the 2nd service-class, respectively. The corresponding simulation results are 0.995 and 0.993 for the two service classes, respectively, for the D-WR scenario with the constant function, having considered service-classes with \( (b_1, b_2) = (20, 16) \) b.u., while \( mC^* = 20 \) b.u., there is no room for more than 1 calls from each service-class to delay in state OFF.

For further investigation of the mean number of calls that experience delay, among the proposed scenarios, according to the call's arrival rate, we modify our application example, so that \( (b_1, b_2) = (2, 1) \) b.u. and \( (C, C^*) = (9, 10) \) b.u. The results (analytical only) are presented in Fig. 11, 12 and 13, for the D-WR scenario with the constant, the same scenario with the increase function and the PD-WR, respectively. In the same figures we also present analytical results of the delay, by using the right vertical axis. As these figures show, the PD-WR scenario performs best, while the D-WR scenario with the increase function performs better than the same scenario with the constant function, for both service-classes.

In the following, considering our initial application example, we investigate the effect of the parameter \( m \) on the CFP and the delay for the D-WR scenario with the constant function. In Fig. 14 we present the analytical CFP results (left vertical axis) as well as the delay results (right vertical axis) versus the arrival rate, for several values of \( m \). The increment of \( m \) results in a higher wavelength service-rate \( R_D \) (Eq. 14); therefore the wavelengths are released...
VIII. CONCLUSION

We present a call-level performance analysis in a hybrid WDM-TDMA PON which supports service-classes of ON-OFF traffic. The PON allocates dynamically a wavelength to each ONU-OLT connection according to different scenarios for the wavelength release procedure. The analysis is based on Markov chains, either one-dimensional or two-dimensional, and leads to recursive CFP, CBP, BBP determination, as well as to the estimation of the average delay that some OFF-calls suffer until the ONU is re-connected to the OLT. We also determine the mean number of OFF-calls that experience this delay. The analytical models are evaluated through extensive simulation and their accuracy was found to be quite satisfactory. Based on the results of the models, we comprehensively studied the performance of each scenario. Our study provides essential guidelines for the PON configuration in respect of which scenario is suitable, according to the users' requirements: the P-WR scenario is preferable for high delay-sensitive applications (which can tolerate connection failures). If the CFP is an important issue while the delay is not so vital, then the D-WR scenario with the constant function could be applied. Alternatively,
the network designer could apply the D-WR scenario with the increase function, or the PD-WR scenario, for a moderate network performance. A decisive factor for the selection of one of these two cases could be the number of OFF-type calls that experience delay. In conclusion, the proposed scenarios and the corresponding models are especially useful to network designers, when installing a new PON, for optimal network performance.

**APPENDIX**

Notation and explanation of terms used in the analysis of the WDM-TDMA PON.

\[ N \]: Number of ONUs

\[ W \]: Number of wavelengths

\[ K \]: Number of service classes

\[ \lambda_{i,k} \]: Arrival rate of service-class \( k \) calls

\[ \mu_{i,k} \]: Service rate of service-class \( k \) calls in state \( i \)

\[ \sigma_i \]: Probability to transit from state OFF to ON

\[ b_i \]: Bandwidth requirement of service-class \( k \)

\[ C \]: Real capacity of a wavelength

\[ C^\ast \]: Fictitious capacity of a wavelength

\[ p_{i,k} \]: Utilization of the \( i-th \) link by service-class \( k \)

\[ y_{i,k} \]: Average number of service-class \( k \) calls in state \( i \)

\[ R_x \]: Service rate of a wavelength in \( x \)-WR scenario (\( x = P \), PD)

\[ T_x \]: Delay in the \( x \)-WR scenario (\( x = D \), PD)

\[ N_{x,k} \]: Mean number of service-class \( k \) calls in the \( x \)-WR scenario (\( x = D \), PD)

\[ \lambda \]: Connection establishment request arrival rate

**ACKNOWLEDGMENT**

This work was supported by the Research Program Carathéodory of the Research Committee of the University of Patras, Greece.

**REFERENCES**


