Adding Mobile Base Station Support to PONs

T. Orphanoudakis[1], E. Kosmatos[1], J. Angelopoulos[2], A. Stavdas[1]
[1]*Department of Telecommunications Science and Technology, University of Peloponnese, Tripolis, 22100, Greece*
Tel: +30 2710372241, Fax: + 30 2710372160, Email: {fanis, vkosmatos, astavdas}@uop.gr
[2] *Technological Education Institute of Piraeus, P. Rali & Thivon 250, Aigaleo, Greece*
Tel: + Tel: +30 210 5381338, , Email: jaggel@teipir.gr

Abstract: The continuing deployment of fiber in the local loop, presents an obvious opportunity to also provide backhauling to mobile base stations in the vicinity of Passive Optical Networks (PONs). To exploit this synergy, one must fully exploit the existing PON MAC tools and also ascertain that certain extreme performance requirements arising from the mobile traffic support can be met by the shared PON medium. This dual objective is addressed in this paper with performance simulations evaluating the most demanding scenario in terms of latency: i.e. that of the hard-handover exchange of messages via one PON over the core network. Also MAC fine-tuning techniques that can improve latency as well as system utilization in the presence of mobile traffic are presented and assessed.

Keywords: PON, MAC, DBA, handover, TDMA, EPON, GPON, mobile backhaul

1. Introduction

The introduction of 3GPP Long Term Evolution (LTE) technology to the mobile network thereby creating a denser station network, in conjunction with burgeoning mobile traffic driven by 3/4G systems is exhausting current microwave-based backhaul networks. In addition, the fixed pipe principle lacks the flexibility required by the new situation and is not as cost-effective as Passive Optical Network (PON) systems deployed for the support of residential or commercial customers. This combination, if properly exploited, can help both the mobile and the fixed operator increase revenue bringing forward the economic break-even point by allowing them to reach faster the critical mass of traffic that balances the initial cash outlays.

The purpose of this paper is to investigate the critical issues arising in this environment, assess the traffic-handling capabilities of TDMA PONs under such a mixed initial traffic scenario and provide design and deployment guidelines to both manufacturers and operators as to the fine tuning of the PON MAC parameters.

2. TDMA PON Architecture and traffic allocation

The typical case of the proposed mixed mobile and fixed user backhaul is depicted in Figure 1 where some Optical Network Units (ONUs) support residential or professional users and small businesses, while one or two serve mobile Base Stations (BSs). This architecture offers distinct advantages for operators pressed with the need to raise the capacity of the backhaul by taking advantage of an already deployed TDMA PON serving a neighborhood in the vicinity of a base-station at a fraction of the cost of a dedicated fiber.

The BS can not be treated as a normal business user connected to the PON with a static Service-Level Agreement (SLA), neither can the mobile operator see this link as a fixed pipe as was the case up to now. Such an approach would not do justice to the bandwidth management properties of the PON and would not fully exploit existing multiplexing gain opportunities. Instead, special tools need to be provided to take advantage of the capabilities of the MAC to the benefit of both operators. This task is assigned to a novel functional unit charged with carrying out the extended bandwidth management with interfaces towards both the BS and the PON as shown in Figure 1. This unit coordinates the SLAs between the two operators and handles the relevant alarms and other events. It is the objective of the related project to design the whole PON wireless backhaul architecture and the full functionality.

TDMA PONs behave quite different to the fixed pipe backhaul links and their main feature is the Dynamic Bandwidth Allocation (DBA) of their MAC protocol. Under its guidance, packets are marshalled one behind the other in perfect and gapless succession multiplexed towards a single input port of the Optical Line Termination (OLT).

When referring to the TDMA PON, there are two dominant standards that can be used for the mixed backhaul network: Gigabit PON (GPON) [1] and Ethernet PON (EPON) [2] (and recently the evolution of both to support 10 Gb/s rates). Both foresee the support of different QoS levels embedded in TDMA PONs for a successful performance, but operators must be well aware of the idiosyncrasies of priorities and MAC functions particularly since the differences between EPONs and GPONs in this respect are not trivial, though based on the same general principles. There is no space to dwell on the way the PON MAC operates and the reader can find relevant information in [1] for the GPON and [2], [4], [5] for EPON.

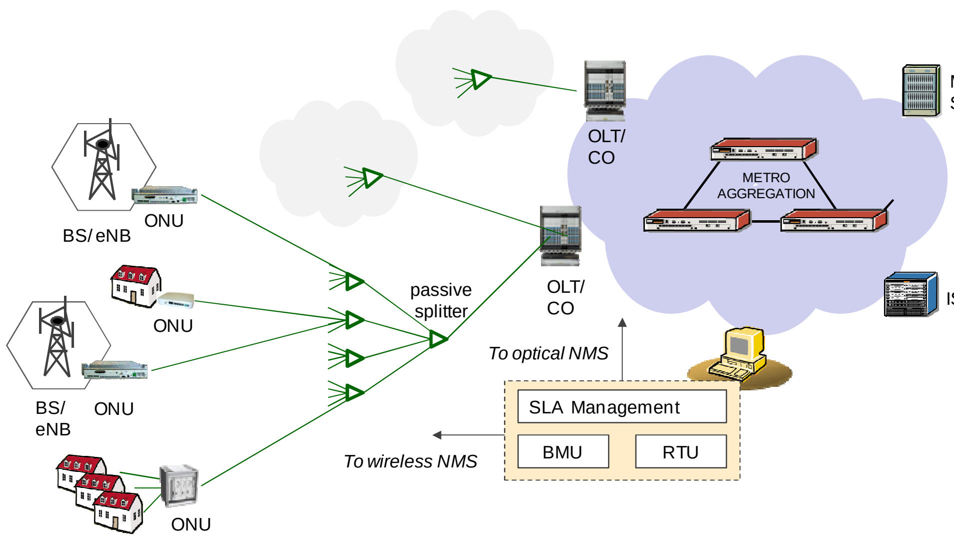

Figure 1: Typical mixed PON architecture

With a PON serving time-sensitive traffic, a trade-off emerges between excessive over-provisioning leading to poor resource utilization and higher utilisation risking loss of QoS ([1],[3],[4]). Service prioritization can however protect sensitive traffic without significant over-provisioning in conjunction with the Dynamic Bandwidth Allocation (DBA). DBA works by first having the ONUs request service indicating their queue length in a report field, following this the OLT allocates enough upstream transmission grants to allow them to relieve the full content of their queues.

Periodic polling is the solution to serving ONU queues that were empty on the previous grant service round, limiting the worst case latency to the polling period as described in [3]. Thus, polling involves granting a transmission interval to an ONU on the basis of time passed and not on known queued traffic. Frequent polling results in wasted bandwidth, a large polling interval, on the other hand, increases latency (i.e. the time waiting for the first grant when arriving at an empty queue, since non-empty queues can always transmit requests). The only way to guarantee a low level of latency is to avoid the delay of request/grant by resorting to pre-arranged periodic Unsolicited Grants (UGs).

3. The traffic management functionality

A central issue in this environment is how to respond to the different timescales of the traffic change, and for this two hierarchical levels are envisaged. One handles the pre-arranged SLAs and the slow long-term changes (BW Management Unit-BMU) and the other handles the fast changes in an autonomic manner (Real-Time Measurement Unit-RTU), as shown in Figure 1. The first objective of the new functionality is to let each operator have control of his side of the negotiation. Thus, the mobile operator will issue bandwidth requests and the PON operator will respond with what is in a position to satisfy. In this way a service bandwidth framework will be established defining the upper and lower limits of the traffic agreement. However, the autonomic part will rapidly set short term parameters within this framework that enable better exploitation of the joint systems.

The autonomic subsystem RTU is based on the concept of the Two-Rate, Three-Color Marker [13]. It consists of a traffic measurement and characterisation using 2 Leaky Buckets (LBs) one for the Peak Information Rate (PIR) with a small bucket size P (the size of about 5 average packets, i.e. about 5kbytes) and one for the Reserved Information Rate (RIR) with a bucket size R (about 5% of the total ONU buffer size). The peak and reserved rates are set by the BMU. They may differ depending on the hour of the day, or the day of the week, based on historical data collected by the BMU. Flows constrained by the (RIR,R) LB set must always be served by the PON and they are counted as green. While not using their full rate, spare capacity can be used for other PON users. Packets exceeding the (RIR,R) LB but not the (PIR,P) are counted as yellow and will be served only if spare capacity is available. Finally, packets outside the (PIR,P) constraint are counted as red and will be blocked.

The RTU obtains parameters from BMU and reports back to it periodically and after certain events. When the number of yellow packets exceeds a threshold in any given hourly period, a renegotiation is initiated by the BMU between the two operator units and a new higher RIR may be agreed. Similarly, prolonged lower usage may lead to lowering the reserved BW. Foreseeable events that will require increased BW (e.g. games, festivals, concerts, rallies etc) in the vicinity of the station, can be manually programmed into the BMU and create new reservations of limited duration that are passed into the RTU for monitoring. When the BS linking to the PON is first commissioned, the initial SLA parameters can be tentative values inserted by the mobile operator based on experience, but are gradually updated by the measurements taken via the RTU, and in this fashion, the RTU creates the working profile of each BS according to the hour of the day and the day of the week and even the month of the year (useful for BS residing is summer holiday resorts or office areas etc). The relevant data are kept in a database of historical traffic values and can be used in deciding and even predicting demand.

4. Handling low-latency mobile traffic in the TDMA PON

As explained in section 2, TDMA PONs rely on priorities to offer the required QoS to delay-sensitive traffic. EPONs support 8 priority levels following the 802.1P approach and a somewhat restrictive native Ethernet support (i.e. Ethernet frames must be supported as a whole) [2], while GPONs allow breaking up frames into smaller parts encapsulated in special frames [1], [3]. This allows much lower levels of latency and delay than is possible in EPONs for the same level of efficiency, as will become clear in the simulation results of the next section. In GPON terms, the traffic classes are 5 and are called TCONTs (Traffic Containers) [12]. TCONT traffic is intended for the emulation of leased-line services (i.e. matching the need of real-time applications for guaranteed bandwidth and minimum latency) and is serviced only by unsolicited periodic grants (UGs). The other TCONTs support Variable Bit Rate (VBR) traffic and, since they are serviced with lower priority, as the TCONT number increases, so does the experienced delay and throughput of the traffic.

It is worth dwelling on the last TCONT5, which is a combined class of two or more of the other four TCONTs. The characteristic is that no target TCONT queue is specified (only ONU) and it is now left to the ONU to choose which queue to service (also called intra-ONU scheduling in [4], [5]). The use of this approach (sometimes referred to as using *colorless grants* or *intra-ONU scheduling*) is left to the system designer and its activation (when implemented) is left to the operator.

When using a TDMA PON as a mobile traffic backhaul system, the latency introduced by the MAC polling approach is obviously much higher than in dedicated links. Though in almost any other case this delay is of no concern, a situation arises in the mobile traffic support that warrants investigation. This is the latency imposed by the MAC polling delay upon the string of hard handover messages ([9], [10]) from a base station situated in one ONU, to another station supported by a different ONU or a different PON. The performance of such a scenario will be investigated in the next section. It is also of interest to compare EPONs and GPONs in the handling of this.

The scheduling interval D_m dominates delay performance and is implied in both systems [3]. Obviously the PON MAC has to periodically visit all active ONU queues (i.e. those known from previous requests to hold unserviced packets), but also to poll all the other ONUs in case packets have subsequently arrived in their queues. This necessity leads to the concept of the mean scheduling/polling period D_m, which is the time interval until the grant schedule goes back to the same ONU. The scheduler decides the allocations after collecting the newer information about packet arrivals, which is also completed in one such period.

The DBA scheme has been well studied in (Broadband PON) BPONs [6], GPONs [1], and EPONs [4], [5] and it has been shown that the mandate of the latter for the support of integral Ethernet frames, gives a somewhat worse delay performance, unless bandwidth efficiency is sacrificed to some extent. The D_m parameter must be kept low enough in order to keep latency and delay variations low as well. In GPON the scheduling period can be quite small (integral multiples of 125μs enabling even the support of TDM services [12]) and together with the low protocol overhead and the fragmentation of frames easily achieves low latency and delay variation, while in the EPON this can only be achieved by selecting a low D_m at the expense of efficiency. This has been shown in several studies [3], [4]. The rate of UGs is critical since they also dominate the latency when an urgent packet arrives in an empty queue and has no other way of requesting service than a polling grant.

An adequate rate of UGs is the only way the operator can actually guarantee service to a contracted peak rate R_{p1} and a strict delay bound, which can be derived as a function of D_m. The scheduling period D_m is used to calculate the bytes to be allocated to each queue to achieve the desired service rate. Hence, considering the case where unsolicited grants cover the sustainable rate R_{s2} of the second Class of Service CoS2 (TCONT2), the total number of unsolicited grants for the i th ONU (UGi) in bytes is expressed as follows: UGi $=(R_{p1i} + R_{s2i})* D_m$ (rates expressed in bytes/s). The remaining unallocated part of each scheduling period D_m is distributed dynamically in a weighted manner and a service weight w_i can be used [5] to enforce proportional sharing of the upstream transmission window among ONUs to guarantee the portion reserved for CoS3 queues (TCONT3). Finally, CoS4 is served as best effort, i.e. whenever unallocated slots exist.

A typical hard handover scenario in Global System for Mobile Communications (GSM) [9] and Universal Mobile Telecommunications System UMTS [10] consists of a sequence of 4 or 5 upstream single packet messages and the whole exchange must be completed well within the service interruption time allowed by the mobile standards. The specification for International Mobile Telecommunications (IMT) [7] gives a maximum service interruption in a handover of 40ms while [10] specifies 50ms. This includes all causes of delay (protocol processing, air interface and propagation) but the new aspect of PON MAC protocol delay, due to the queuing involved until a grant becomes available, is, of course, completely unaccounted for in the standard. It is reasonable to assume that only a small portion of this delay budget can be consumed by the TDMA PON. While one author suggested a value of as low as 2ms, [11], it does not seem expedient to put such a tight restriction, since some margin has undoubtedly been built into the service interruption value given by the standards. Although it is not our intention to give a hard limit, it is reasonable to note that exceeding 10ms seems inadvisable and, obviously, the lower the value, the safer. In any case our objective is to quantify the distribution of this delay in assisting a relevant standardisation effort.

5. Performance Assessment by Simulation

In this section computer simulation results are presented to investigate the traffic behaviour of the proposed backhaul solution and, on this basis, suggest best practices for operators in allocating bandwidth and fine-tuning the unsolicited grants (UG). The available margins and trade-offs between latency and utilization will also be presented. The handover scenario is selected because it is the most demanding in terms of time-critical service (together with voice communication, representing relatively low bandwidth flows with hard real-time requirements) and it must be supported through the highest priority MAC service and with carefully selected unsolicited grant provision.

The simulation set-up employs 16 ONUs one of which exclusively serves a mobile BS while the others carry residential traffic. Three classes of service (CoS) are simulated of which the highest priority, COS1, is served in GPON by TCONT1 and by the top priority in EPON, while the other two by TCONT2 and 3 respectively. The 4th class (best effort) is not represented here, as its study offers no useful conclusions. The traffic mix characteristics per ONU type are shown in Table 1. For all ONUs CoS1 traffic is considered to account for 20% of its total offered load and is modelled either as Constant Bit Rate (CBR) voice traffic or control message traffic in the case of the wireless BS, or data traffic modelled following an on-off model with a low burst factor $BF=(T_{on}+T_{off})/T_{on}$ in the case of residential users.

Table 1: Simulated traffic load profiles per ONU

	CoS1 (TCONT1)		CoS2 (TCONT2)		CoS3 (TCONT3)	
	ONU load (%)	Profile	ONU load (%)	Profile	ONU load (%)	Profile
Residential ONUs	20	on-off, BF=3	25	on-off, BF=5	55	on-off, BF=5
BS ONU	10 10	CBR signalling	25	on-off, BF=5	55	on-off, BF=5

All other traffic sources (CoS2 and CoS3 traffic) are considered as highly bursty data sources following an on-off model. The total simulated time was 15sec (resulting in a total of around $4*10^6$ simulated packets) for each simulation scenario, while each scenario was repeated 20 times (noting insignificant deviations in the results in different repetitions).

The total delay for the typical signalling exchange of a handover scenario was measured and the pdf of this delay is depicted in Figure 2 for a total load of 40% (no significant difference is observed at higher loads because this takes the highest priority and it is not affected by the total load). The signalling exchange consisted of 5 upstream single packets modelling the two way handover protocol message exchange, which had to endure the access delay of the PON backhaul and an additional processing delay before a response message is generated (random processing delay following a Poisson distribution was assumed with a mean of 1ms - any further propagation delay was considered insignificant). Values near or above 10ms would risk unacceptable service interruption. As can be seen in Figure 2, the impact of the D_m parameter is dominant as expected from previous studies of TDMA PON delay. In reality only the value of D_m=0.75ms provides a small enough tail to give confidence in the mixed architecture studied in this paper. This D_m value is 6 times the frame size of GPON and can also be easily programmed into the EPON but at the penalty of some inefficiency. This is due to the way EPON is designed to carry whole Ethernet frames leaving an unused space remainder (USR) at the end of each upstream allocation. Lowering D_m decreases the mean upstream transmission length, thus increasing this waste. There is no need to repeat the interesting investigation of this EPON idiosyncrasy, which has been extensively studied (e.g. [3], [8]), however, to give a quantitative indication of this effect here, we provide for comparison in the inset of Figure 2 the values of U_{loss}, (i.e. of the throughput lost in EPON because of USR as a percentage of that of a GPON) for each D_m value and the same loading. The parameter U_{loss} is defined as $U_{loss}=(U_{GPON}-U_{EPON})/U_{GPON}$ and quantifies the percentage of bandwidth lost in the EPON versus the GPON under same operating conditions due to higher overhead and lack of fragmentation to fully fill frames. It is worth noting that the GPON can still improve on latency by using an even lower D_m=0.5 without noticeable inefficiency.

Next, attention in this simulation study was directed to the impact of grant allocation, investigating four alternative policies in order to make clear to the operators the available options and the consequent performance trade-offs. First, in each scenario two modes regarding the target of the allocations were differentiated: coloured (targeting individual ONU queues since the grant identifiers specify individual queues), but also colourless (i.e. ignoring the actual queue identifier and letting the ONU allocate according to local priority criteria).

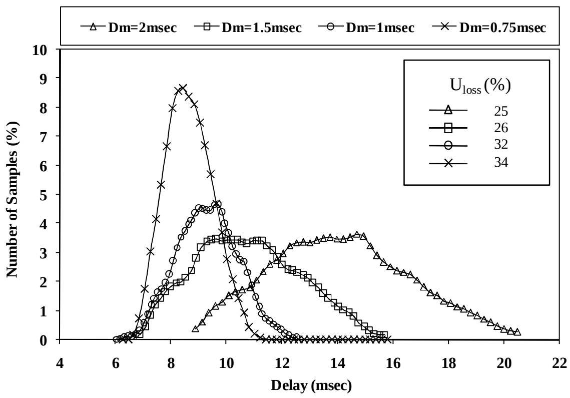

Figure 2: PDF of signalling delay

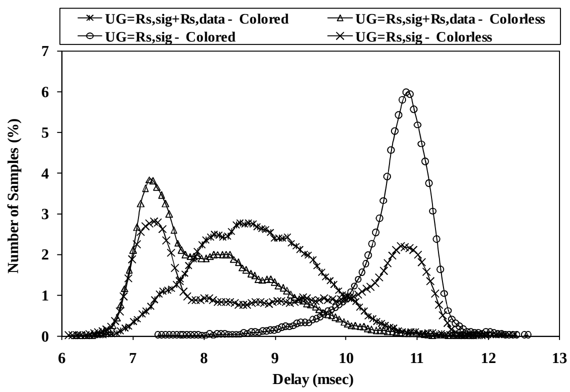

Figure 3: Comparison of coloured v. colourless mode

At this point, two polling policies were also distinguished: in the first one, the polling (by means of UG) is set to just that required for the expected signalling rate of the first priority (while the rest used requests made possible by these UGs). In the other, UGs are issued at a rate equal to the sum of signalling plus the sustainable rate contracted by the SLA. This, of course, refers to the top priority, since the other classes grab the opportunity of first class grants to send requests for their own queues provoking corresponding grants at a second round, as is the normal PON MAC approach.

To understand the incentive for an elaborate UG policy, one must approach the issue from the operator's perspective. What is really needed is a way for the operator to predict the volume of signalling traffic in order to pre-allocate enough bandwidth via polling to guarantee a lower than the maximum tolerable latency. Since in real life, the generation of signalling traffic is unpredictable, it is natural to consider over-provisioning, thus allocating the UG rate equal to the expected worst-case peak rate of the signalling traffic. However, in that case, the available bandwidth is under-utilized during periods of low signalling load.

It then follows naturally that an improvement can be reached by multiplexing signalling and non-signalling high priority traffic into the same queue. In that case, the unused bandwidth of signalling traffic is allocated to data traffic, resulting in better efficiency, without at all compromising the critical latency and delay of the signalling traffic, nor that of first priority traffic, since the weak point of the PON lies in initiating transmissions from an initially empty queue (realized only by means of a blind polling) and not for continuing service on a queue (which takes place by the chain of requests). It is like a chain smoker who needs no fire to light one cigarette after the other, but will need new light once he breaks the chain and extinguishes the last one. In the PON the new light comes from UGs. Exploiting this idiosyncrasy of the TDMA PON, we propose this strategy (i.e. $R_{s,sign}$+$R_{s,data}$) by pre-allocating bandwidth equal to the sustainable rate of signalling traffic plus the sustainable rate of high priority traffic.

Figure 3 depicts the pdf of signalling delay at a high total offered load of 90% and a D_m of 0.75ms and allows the comparison of the colourless mode and the UG allocation policies. As it can be seen, in the case of colourless mode, the vast majority of signalling delay values are positioned around the value of 7.5 msec. In contrast, in case of coloured mode, the distribution of delay is highly affected by the strategies adopted by the network operator, reaching prohibitive values of around 11 msec. A smaller D_m is warranted for safer handover signalling, but in the EPON, the maximum loading would have to be reduced if smaller D_m values are used. Another observation, is that the policy of ($R_{s,sign}$+$R_{s,data}$) outperforms the mode ($R_{s,sign}$) in terms of signalling delay both under low and high loads.

As seen from the same Figure, the colourless policy gives consistently better results in all cases and its adoption is recommended. This is to be expected since the OLT has limited knowledge of the local situation in comparison with a centralised multiplexer which instantly knows all queue lengths. As one cannot have this knowledge, one should at least delegate the remote multiplexing enacted by the PON MAC protocol controller to the local ONU, (unfortunately with the limited scope of the local queues) thus improving performance. This is particularly useful among the different priority queues of the ONU resulting in the obviously useful effect of high priority queues "stealing" grants directed to lower priority which forces the latter to report the same packet again in their request suffering no real harm since they are delay-tolerant. In addition, due to this fact, in colourless mode the performance of signalling traffic is isolated from other lower priority queues of the same ONU and the traffic generated from other ONUs as well, and therefore the performance of signalling traffic is only dependent on the selected UG policy.

6. Conclusions

The widespread deployment of PON systems at a time when wireless backhaul is been upgraded, provides a serendipity that cannot be missed as it offers a smooth migration path, both in technical as well as financial terms. However, the TDMA aspect presents certain peculiarities and careful traffic management by the operator is needed. As demonstrated in this paper, the colourless policy together with the adoption of a polling allocation covering the expected peak signalling traffic plus the sustainable part of the demanding high-priority traffic, carries distinct advantages in terms of latency and delay bounding for sensitive traffic without sacrificing efficiency under high load. A simulation study has quantified the distribution of the delay added by the PON in the demanding case of a hard handover and showed the importance of the granting frequency and policy in keeping it under control.

Abbreviations

BMU	Bandwidth Management Unit	OLT	Optical Line Termination
BPON	Broadband Passive Optical Network	ONU	Optical Network Unit
BS	Base Station	PIR	Peak Information Rate
BW	Bandwidth	PON	Passive Optical Network
CBR	Constant Bit Rate	RIR	Reserved Information Rate
CoS	Class of Service	RTU	Real-Time Measurement Unit
DBA	Dynamic Bandwidth Allocation	SLA	Service-Level Agreement
EPON	Ethernet Passive Optical Network	TCONT	Traffic Containers
GPON	Gigabit Passive Optical Network	TDMA	Time Division Multiple Access
GSM	Globale Système Mobile	UG	Unsolicited Grant
IMT	International Mobile Telecom/tions	UMTS	Universal Mobile Telecom/tions System
LB	Leaky Bucket	USR	Unused Space Remainder
LTE	3GPP Long Term Evolution	VBR	Variable Bit Rate
MAC	Media Access Control	WDM	Wavelength-Division Multiplexing

References

[1] J. Angelopoulos, H.-C. Leligou , T. Argyriou, S. Zontos , E. Ringoot, T. Van Caenegem, "Efficient transport of packets with QoS in an FSAN-aligned GPON", IEEE Communications Magazine, vol. 42, pp. 92–98, 2004.

[2] IEEE Std 802.3ah-2004, 2004.

[3] T. Orphanoudakis, H.-C. Leligou, E. Kosmatos and J. D. Angelopoulos, "Performance evaluation of GPON vs EPON for multi-service access" Int. J. Commun. Syst. 2009; 22:187–202.

[4] A. Shami et al., "Jitter Performance in Ethernet Passive Optical Networks," IEEE J. of Lightwave Technology, vol. 23, No. 4, pp. 1745-1753, April 2005.

[5] C. M. Assi, Y. Ye, S. Dixit, M. A. Ali "Dynamic Bandwidth Allocation for Quality-of-Service Over Ethernet PONs", IEEE J. on Sel. Areas In Commun., Vol. 21, No. 9, Nov. 2003

[6] J.D.Angelopoulos, G.C. Boukis, I.S.Venieris, "Delay priorities enhance utilization of ATM PON Access Systems," Computer Communications Journal, Vol. 20,pp. 937-949, (1997).

[7] ITU-R M.2134, "Requirements related to technical performance for IMT-Advanced radio interface(s)", 2008.

[8] S. Lallukka , P. Raatikainen , "Link utilization and comparison of EPON and GPON access network cost", Proceedings of Globecom 2005, St. Louis, MO, U.S.A., 28 Nov.–2 Dec. 2005.

[9] GSM 04.18 Digital Cell Telecommunications System (Phase 2+); Mobile Radio Interface Layer 3 Specification, Radio Resource Control Protocol.

[10] 3GPP TS 23.009 V5.6.0, Handover procedures, Technical Specification,3rd Generation Partnership Project, Technical Specification Group Core Network.

[11] J. Chung, J. Park, B. Choi, Y. Park ,J. Yoo, B.-W. Kim, "Application of 2W-PON to the backhaul of a wireless network", Journal of Optical Metworking, Vol. 7, No. 11, Nov. 2008, pp, 905-913.

[12] ITU-T G.984.3, "Gigabit-capable Passive Optical Networks (G-PON): Transmission convergence layer specification", March 2008.

[13] RFC 4115, "A Differentiated Service Two-Rate, Three-Color Marker with Efficient Handling of in-Profile Traffic", July 2005.