New Unified PON-RAN Access Architecture for 4G LTE Networks

Kostas Ramantas, Kyriakos Vlachos, Anastasios N. Bikos, Georgios Ellinas, and Antonis Hadjiantonis

***Abstract*—This work presents a new converged access architecture for LTE mobile backhaul networks. In the proposed architecture, evolved NodeBs (eNBs) are interconnected with local ring-based wavelength-division-multiplexed (WDM) passive optical networks (PONs), which aggregate and efficiently transport traffic to the evolved packet core (EPC). The proposed WDM-PON ring design supports a dynamic setup of virtual circuits for inter-base-station communication, over a dedicated λ_{LAN} channel. It also supports load balancing, by dynamically reallocating and sharing the capacity of the downstream wavelengths. The reservation mechanism is arbitrated by the optical line terminal, which monitors the traffic imbalances of downstream channels and orchestrates the setup of subwavelength "transient flows."**

***Index Terms*—FiWi; Optical burst switching; Optical-wireless convergence; Passive optical networks (PONs).**

I. Introduction

The newest standardized mobile telecommunication system, universally recognized as 4G, is being prototyped for increased capacity and reliable wireless communications. Next-generation wireless access architectures, namely Mobile WiMAX and LTE, are two competing technologies that are expected to achieve data rates beyond 100 Mbits/s per end user. On the other hand, the demand for high-access bandwidth is also expected to grow continuously, due to the increased expansion of innovative and high-bandwidth applications such as Web 2.0, mobile TV, and streaming content. Thus, current backbone standards are expected to become less effective for building mobile access networks. Specifically, legacy technologies such as circuit-switched T1/E1 wireline or microwave used for existing 3G network infrastructures cannot scale to the capacity requirements of new 4G access architectures [1]. Thus, mobile operators are investing heavily in upgrading their backhaul infrastructure, with fiber-optic deployments to the LTE base stations ("fiber to the cell").

Passive optical network (PON) technology, whose access capacity has been steadily increasing, is a viable alternative for next-generation fiber access [2] and can also be considered for building mobile access networks. The mobile backhaul portion of 4G telecommunication networks, or the radio access network (RAN), interconnects the evolved packet core (EPC) with the edge section of the wireless domain and transports traffic from individual base stations (BSs) to the access gateway (AGW). However, it must be noted that with peak per-edge cell downlink throughput of 1 Gbit/s and uplink of 500 Mbits/s in the case of LTE Advanced, 4G base stations are expected to be densely populated to achieve high spectral efficiency and require high bandwidth and cost-effective backhauling [3]. This makes next-generation PONs (NG-PONs) strong candidates for implementing mobile backhaul due to their high capacity, longer reach, and economical deployment.

The new wireless communications standards, LTE Advanced and Mobile WiMAX, are the two routes of the evolution toward 4G and beyond. LTE Advanced can be seen as an enhancement to LTE, offering a clear upgrade path to mobile carriers. This makes it more cost effective for vendors to offer LTE and then upgrade to LTE Advanced [4]. Furthermore, LTE and LTE Advanced will also make use of additional spectrum and multiplexing to achieve higher data speeds. Coordinated multi-point (CoMP) transmission will also allow more system capacity to help handle the enhanced data speeds, which is a necessity for the optical–wireless architecture convergence. Additionally, unlike WiMAX, LTE uses an evolution of the existing Universal Mobile Telecommunication System (UMTS) infrastructure, used by more than 80% of mobile operators [1]. Thus, it is not necessary to build a new network infrastructure, making LTE more popular with operators worldwide.

Due to their compelling advantages, many works have addressed the need for building access architectures for LTE networks, such as [5,6]. However, very few address the issue of efficient inter-communication of LTE base

Manuscript received February 24, 2014; revised August 25, 2014; accepted August 25, 2014; published September 23, 2014 (Doc. ID 206809).

K. Ramantas (e-mail: ramantas@ceid.upatras.gr) is with the Computer Engineering and Informatics Department, University of Patras, Greece, and Iquadrat Informatica S.L., Barcelona, Spain.

K. Vlachos and A. N. Bikos are with the Computer Engineering and Informatics Department and the Research Academic Computer Technology Institute, University of Patras, Greece.

G. Ellinas is with the Department of Electrical and Computer Engineering, University of Cyprus, Cyprus.

A. Hadjiantonis is with the Department of Engineering, University of Nicosia, Cyprus.

http://dx.doi.org/10.1364/JOCN.6.000890

stations. In this work we propose a new unified PON-RAN architecture for LTE mobile backhaul networks, employing ring-based wavelength-division-multiplexed (WDM) PONs. Mobile backhaul networks are perfect candidates for exploiting the high capacity, inherent resilience, and ubiquity offered by WDM rings, as the—comparatively higher—infrastructure cost due to the use of WDM components is amortized to a much larger number of mobile clients. However, employing current generation WDM-PON networks results in the inefficient use of resources, as wavelength capacity cannot be reallocated or shared between optical network units (ONUs). Furthermore, communication among ONUs is performed via the optical line terminal (OLT), unnecessarily increasing delay and wasting capacity of both the upstream and the downstream channels. In this work, we propose a new converged access architecture for LTE mobile backhaul networks, where evolved NodeBs (eNBs) are interconnected with local ring-based WDM PONs as shown in [7,8], which aggregate and efficiently transport traffic to the EPC, supporting all-optical inter-communication and full meshing of LTE base stations.

The rest of the paper is organized as follows. In Section II background information and related work are discussed. The new converged architecture is described in Section III, while in Section IV we present an end-to-end quality-of-service (QoS) framework. In Section V, a resource reservation protocol is detailed for setting up all-optical virtual circuits at the λ_{LAN}. In addition, a downstream wavelength sharing scheme is described and used to support load balancing between ONUs/eNBs. Finally, in Section VI, the proposed architecture is evaluated with extensive simulation experiments, followed by concluding remarks in Section VII.

II. Related Work

PON technology is a viable solution for next-generation fiber access networks. The most popular variant considered is usually time division multiplexed-passive optical networks (TDM-PONs). In TDM-PONs a single wavelength is shared by all ONUs in the upstream direction based on a time division multiple access (TDMA) algorithm that is arbitrated by the OLT. In order to increase access capacities, 10GE-PONs have been proposed and recently standardized [2]. They offer compatibility with pre-existing PON deployments and share existing passive components, resulting in a smooth network evolution. However, for future high-bandwidth applications and higher split ratios, more efficient NG-PON architectures must be considered that employ WDM technology to offer capacities beyond 10 Gbps and longer reach albeit at a higher cost [9]. WDM-PONs can be considered as an evolutionary scenario of existing TDM-PONs employing a dedicated wavelength per ONU for OLT/ONU communication, to offer increased capacities. One of the most important challenges to be met by next-generation optical access networks is energy efficiency [10], while QoS support and efficient resource utilization are also important. Finally, there is a large body of research on survivable access architectures. Ring-based architectures typically employ double rings to offer protection [11], while tree-based PONs reroute wavelengths via backup fibers using AGWs [12].

Recently, research work has focused on the integration of NG-PONs and NG-WBANs, to build converged architectures that combine the merits of both wired and wireline access technologies. Such converged infrastructures would enable the deployment of new, innovative, high-bandwidth services, which support mobility and end-to-end QoS guarantees. One of the challenges that must be met by future converged architectures is the efficient inter-communication of LTE base stations. In [5], the authors detail the implementation methodology on how to efficiently integrate optical and wireless access technologies. They also propose a WDM-PON ring-based converged architecture, which supports ONU inter-communication with a dedicated λ_{LAN} channel. The proposed access architecture is further enhanced in [13], where a PON ring, which supports all-optical inter-BS communication, is presented, along with a new wavelength sharing scheme. In [6] a tree-based converged architecture is proposed, with the ability of all-optical communication of all BSs that belong to the same PON. Finally, hybrid optical–wireless architectures that employ reconfigurable WDM technologies, such as GROWnet [14], have been considered in the literature, to offer flexibility for varying traffic demands.

III. Converged LTE/PON Architecture

The LTE network architecture consists of an all-IP core network, called the EPC, and new, enhanced base stations called eNBs [1]. The eNBs are connected by means of the S1 interface to the EPC, whose logical components are the mobility management entity (MME), the serving gateway (S-GW), and the packet data network gateway (P-GW), together also known as the AGW. LTE also introduced support for inter-BS connectivity via the X2 interface, to support handover operations. Recent studies have estimated traffic traversing the X2 interface to reach 4%–10% of traffic traversing the S1 interface [6]. Thus, it is important for efficient converged architectures to support at least partial meshing of eNBs, so that X2 traffic does not flow through the AGW, which would waste resources and significantly increase packet delay.

It is generally accepted that fiber deployment to cell towers (“fiber to the cell”) is the only future-proof solution to build mobile backhauls, which will scale to the increased capacity requirements of future NG-WBAN technologies [1]. Converged architectures based on PONs have the added benefit of reusing passive components after upgrading the active infrastructure, providing a cost-effective network upgrade path.

A. Access Architecture and Ring Design

The proposed ring-based WDM-PON access architecture, initially proposed in [15,7] (see Figs. 1 and 2), employs

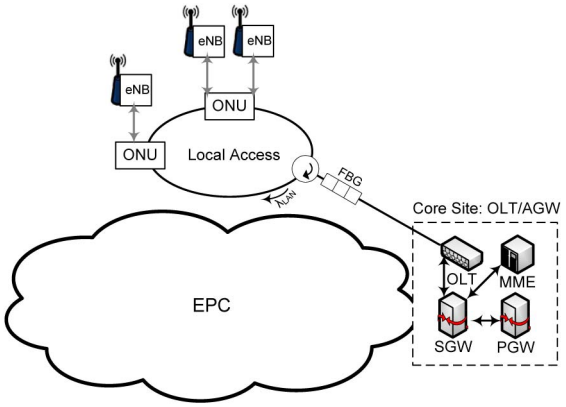

Fig. 1. Proposed converged architecture.

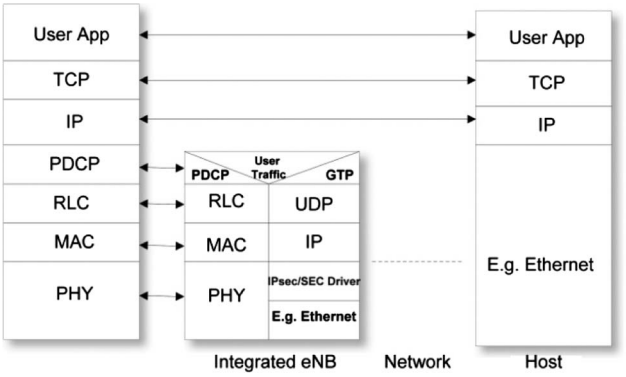

Fig. 3. Block design of the eNB.

WDM rings of ONUs to interconnect LTE BSs in the same local access area. Each WDM ring is connected to the corresponding OLT via a bidirectional 10–20 km feeder fiber and a passive three-port circulator. OLTs are collocated with the AGWs at the core sites of the EPC.

Each ONU terminates traffic from/to one or more LTE eNBs, which are assumed to be directly interconnected via Ethernet interfaces. Each ONU is assigned two wavelengths, namely a dedicated wavelength for downstream/upstream traffic from/to the OLT and another one, denoted as λ_{LAN}, which is shared by all ONUs across the ring, for inter-ONU communication (see Fig. 2). The former carries traffic that belongs to the S1 interface (i.e., interconnects eNBs to the EPC), and the latter carries traffic from the X2 interface (i.e., interconnects base stations).

Unlike the converged architecture proposed in [5], a new ONU design allows bypassing intermediate ONUs, thus avoiding unnecessary termination of λ_{LAN} traffic in intermediate ONUs. The new ONU design supports all-optical meshing of eNBs that belong to the same ring, offering all-optical subwavelength connectivity [6]. Transmission is unidirectional at the ring: both upstream and downstream signals are transmitted in the same direction.

A fiber Bragg grating (FBG) reflects back the λ_{LAN} wavelength from the upstream signal heading to the OLT and allows it to recirculate around the ring.

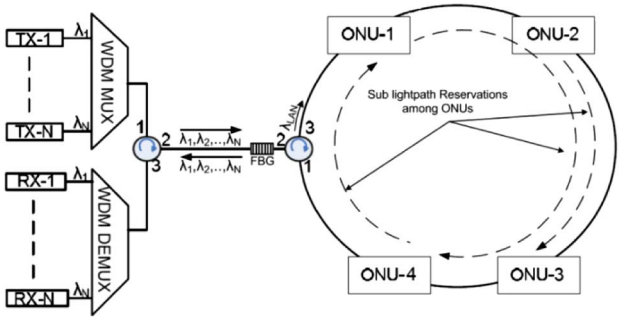

Fig. 2. WDM-PON ring-based architecture.

B. LTE eNB Design

In Fig. 3, we present a standard eNB architecture, found in modern LTE network deployments. Hardware vendors typically employ FPGAs or ASICs to implement the PHY and baseband processing, DSPs for lower layer protocols (i.e., MAC and RLC), and CPUs or network processors for the upper layers of the protocol stack. The MAC sublayer is responsible for QoS-aware downstream/upstream packet scheduling. For downlink traffic, the scheduler decides which packets to be sent to the intended user equipment (UE). Uplink scheduling results in resource grants being sent to UEs. Each UE is responsible for determining which data to transmit within the granted resources. A QoS-aware MAC scheduler at the eNB aims to distribute the available air interface resources to the UEs within the cell, supporting QoS guarantees.

The bearer information must be carried on all system interfaces and mapped to preconfigured QoS parameters regarding priority, packet delay, and packet loss (see Section IV). This includes RAN elements that might be prone to congestion-related losses or excess packet forwarding delay. The eNBs are assumed to support a common standard interface to interconnect with the ONU. For example, the eNB implementation may use two gigabit Ethernet (GigE) interfaces to transport X2 and S1 interface traffic, or a single 10 GigE interface in which X2/S1 traffic is multiplexed. These are interconnected with the corresponding Ethernet interfaces of the ONUs.

C. ONU Design

Figure 4 displays the proposed ONU block design. Each ONU is equipped with a pair of lasers and receivers; λ_i is used for OLT-to-ONU$_i$ downstream/upstream communication and λ_{LAN} for inter-ONU communication (termed as LAN traffic). Subwavelength sharing of λ_{LAN} involves the aggregation of packets per destination ONU in a separate virtual output queue (VOQ). Aggregated packets are then transmitted in burst mode over the λ_{LAN} wavelength. A 1×2 optical switch is used to either extract bursts at the destination ONU or transparently forward them

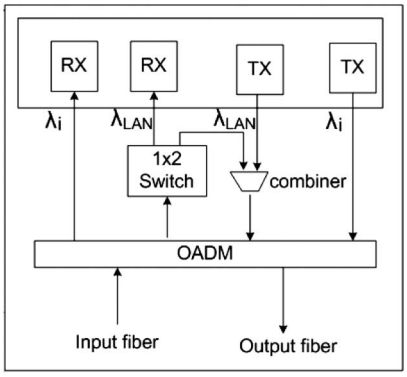

Fig. 4. Block design of the optical network unit.

to the ONUs that follow next on the ring. Burst-mode transmission technology is very mature, as it has already been used in commercial TDM PONs, where ONUs transmit bursts of packets during timeslots preallocated by the OLT. By extracting bursts that reach their destination via the optical switch (destination stripping), the common wavelength can be spatially reused by downstream nodes, leading to an increased capacity. It must be noted that fast switches and burst-mode transceivers are also the building blocks of optical burst switching (OBS), which has been extensively researched for sub-wavelength capacity provisioning in future backbone networks and metropolitan rings [16]. However, it is worth noting that commercially available fast optical switches for OBS networks (with nanosecond switching times) exhibit a limited extinction ratio that may impact the scalability of the proposed scheme, especially the maximum number of ONUs, due to the concatenation of many optical switches in a relatively short distance. Nevertheless, there are switching devices (based on SOAs) that have been shown to possess the potential of high extinction ratio, i.e., >70 dB [17]. Evidently, when moving to a high number of ONUs and long-reach networks, crosstalk (as well as other physical layer impairments) may impact the "goodput" performance of the proposed scheme. In order to quantify such performance issues, especially for a high number of ONUs and long-reach networks, a combined network-physical layer simulation should be carried out.

Clearly, the proposed architecture, if combined with a modular backplane-based ONU design, would offer the mobile carriers a clear upgrade path, since supporting more eNBs would only require adding more line cards to the ONUs. In case an upgrade of the backhaul link capacity is needed (e.g., moving from 1 Gbps to 2.5 Gbps to 10 Gbps wavelengths), only the corresponding active components (i.e., lasers and transceivers) of the ONUs would have to be upgraded, reusing already deployed passive components and eNBs.

IV. QoS Framework for the Converged Architecture

End-to-end QoS is one of the significant benefits offered by LTE networks, unlike 3G and HSPA, where even voice traffic is subject to network related uncertainties. LTE supports end-to-end QoS by setting up logical links between the UE and the gateway, termed "EPS bearers." Each bearer corresponds to a unique QoS identifier (QCI) and specific QoS parameters (i.e., delay, loss, and bandwidth). A bearer has two or four QoS parameters, depending on whether it concerns a real-time or best-effort service:

- QoS class indicator (QCI).
- Allocation and retention priority (ARP).
- Guaranteed bit rate (GBR) (for real-time services).
- Maximum bit rate (MBR) (for real-time services).

The QCI parameter specifies the treatment of IP packets received on a specific bearer by each network node. QCI values impact several node-specific parameters, such as link layer configuration, scheduling, and queue management. Bearers are divided into two broad categories: GBR, where blocking is preferred over packet dropping, and non-GBR, where packet dropping may be experienced. Typically, GBR traffic as well as non-GBR traffic that corresponds to carrier-provided services is distinguished and prioritized, so that it has the required packet forwarding treatment end to end. On the other hand, lower-priority non-GBR traffic (such as Web browsing, P2P traffic, and online video) may suffer congestion-related losses at the backhaul network. GBR bearers reserve a minimum amount of bandwidth end to end, and always consume resources regardless of whether it is used or not. GBR bearers should not experience congestion-related packet losses on the radio link or the RAN. User service flows (e.g., Web browsing sessions and file transfers) in LTE networks are bound to specific bearers based on network operator policies, typically defined with an IP five-tuple. The IP tuple contains at least the source and destination IP, source and destination port, and protocol identification.

A. QoS Mapping Scheme

Each IP packet entering the LTE network is provided with a tunnel header on the different system interfaces [18]. This tunnel header contains the bearer identifier (QCI) so that the network nodes can associate the packet with the required QoS parameters. The EPS bearer is not visible at the backhaul transport elements (i.e., ONUs and OLTs) that only have access to the IP header and/or the Ethernet frame header. Thus, a mapping process must be performed to translate bearer-level QoS to transport-level QoS, as proposed in [19]. Using this function, packets on a bearer associated with a specific QCI are marked with a specific IP *differentiated services code point* (DSCP) value for forwarding in the transport network. The DSCP field in the IP header is used for packet classification in the Diffserv model, and it is supported by all enterprise network elements. The traffic forwarding treatment (i.e., queue management and packet scheduling) is performed on individual packets, based on the DSCP value.

In the proposed architecture, for each downlink packet entering the AGW that does not have the DSCP

header set, the AGW calculates a DSCP value based on a predefined QCI-to-DSCP map table. The UEs are responsible for tagging the uplink packets. However, the DSCP IP header is not available to layer 2 devices. Thus, packet priorities are also encoded in the three 802.1p priority bits of the Ethernet frame, also called classes of service (CoS) bits. The prioritization specification 802.1p works at the media access control (MAC) framing layer and establishes eight priority levels, whose actual mappings to QCI parameters are again determined by the AGW. It must be noted that as long as the LTE network uses up to eight different QoS classes, 1:1 mapping is possible. In deployments with more than eight LTE classes, more than one QOS class has to be collapsed to a single layer 2 priority level.

QoS mapping allows uniform congestion management in all system interfaces. We can identify two potential bottleneck points in LTE networks:

- The air interface between the eNB and the UE. With limited bandwidth available and interference from adjacent cells, the air interface is considered a constrained resource.
- The backhaul and aggregation networks, which support the S1 interface between the evolved packet core (EPC) and the eNB.

In each system interface, appropriate scheduling algorithms and traffic queues are employed to enforce the QoS constraints. The eNB is responsible for the congestion management of the air interface, for both downlink (i.e., eNB to UE) and uplink (i.e., UE to eNB) packets. ONUs and OLTs are responsible for congestion management of the backhaul connections. In this work, we will focus on congestion management at the network backhaul, assuming a 1:1 mapping between traffic queues at the eNBs and the ONUs, based on the QoS mapping framework presented. This means that a packet on a specific bearer can expect to be stored at the same QoS traffic queue end to end.

B. Traffic Shaping

QoS-aware scheduling is integral for fulfilling the QoS characteristics associated with the different bearers. Traffic shaping at the EPC is not sufficient to provide end-to-end QoS, as it lacks vital information about the network edge and the backhaul network, which are often the main bottlenecks. Thus, traffic shaping functions (i.e., packet scheduling, queue management) need to be implemented by the backhaul transport elements, i.e., the OLTs and ONUs. To distribute the transport network resources between the established bearers, the LTE RAN implements uplink and downlink scheduling functions.

In the proposed architecture, we employ two GBR traffic classes (denoted as Q0 and Q1) and two non-GBR traffic classes (namely Q2 and Q3), which are common for both air and wireline interfaces. Specifically, Q0 is reserved for real-time voice traffic with very low delay requirements (i.e., <20 ms) and Q1 for real-time video calls with low delay requirements (i.e., <50 ms). Q2 corresponds to carrier-provided assured forwarding traffic (such as IPTV) and is given priority over Q3, which corresponds to best-effort traffic (such as Web browsing, P2P traffic, and online video). Assuming R the access rate of the backhaul connections, and R_i the data rates of the traffic classes, we assume that W_i are the (normalized) scheduling weights so that

$$\sum (W_i * R_i) < R. \tag{1}$$

To increase statistical multiplexing gains, the QoS scheduler groups the high-priority bearers in a single high-priority class, with a weight of $W_h = W_0 + W_1$. Traffic policing will guarantee that no GBR flows will be accepted that will overflow the GBR class, i.e., $R_1 + R_2 < W_h * R$. An appropriate QoS-aware scheduling discipline, such as weighted fair queuing or weighted round robin, is then employed to schedule GBR and non-GBR traffic, given the corresponding weights. Packets that belong to GBR flows are then selected for transmission with the earliest deadline first (EDF) policy, according to their specific QCI. Specifically, voice packets have a 20 ms deadline, while video packets have a 50 ms deadline. It has been shown that EDF policy guarantees that all deadlines will be met as long as the system capacity is not exceeded [20]. On the other hand, for non-GBR traffic, packets that belong to the Q2 class (i.e., assured forwarding) have absolute priority over packets belonging to the Q3 class.

V. Dynamic Bandwidth Allocation and Load Balancing in PON Rings

In the proposed converged architecture, inter-ONU communication is performed by sharing all-optically the λ_{LAN} wavelength. The proposed design, presented in Section III, avoids terminating λ_{LAN} traffic at the intermediate ONUs, which would waste resources and unnecessarily increase packet delay, and supports all-optical meshing of ONUs/eNBs that belong to the same ring. In this section, the focus is on resource reservation and scheduling at the λ_{LAN} channel. Additionally, a new load-balancing scheme for the proposed WDM-PON ring architecture is presented, which exploits the common λ_{LAN} channel. Specifically, it will be shown that efficient ONU inter-communication allows offloading excess traffic from congested ONUs, and redirecting it through one (or more) uncongested one(s). Thus, the network can react to short-term traffic changes by dynamically reallocating and sharing the capacity of the downstream wavelengths. This mitigates one of the main drawbacks of the current generation of WDM-PONs, whose wavelength capacity cannot be reallocated or shared between ONUs, resulting in the inefficient use of resources. In the proposed architecture, ONUs are interconnected in local rings that function as distributed load-balanced optical switches, aggregating traffic from the eNBs and efficiently transporting it to the EPC.

A. LAN Traffic Resource Reservation

The proposed PON-RAN access architecture employs centralized TDM arbitration for subwavelength sharing of the λ_{LAN}, as in standard TDM-PONs (see Algorithm 1). The OLT gathers bandwidth requests from ONUs and schedules data transmission in a round-robin fashion with MPCP signaling, which is defined in IEEE 802.3ah standard [21], to avoid collisions. MPCP signaling includes the transmission of REPORT messages from ONUs, which include the length of up to eight QoS queues, and GATE messages from the OLT, which contain up to four granted transmission periods. It must be noted that unlike standard TDM-PONs, the λ_{LAN} is employed for ONU/eNB meshing instead of (point-to-point) ONU/OLT communication. However, it is still possible to employ MPCP for inter-ONU point-to-multipoint communication using the logical links concept, which is inherently supported in the 802.3ah standard. In the proposed LAN traffic reservation protocol, each ONU is assigned one logical link per end destination. Since the logical link is the responsible entity for receiving GATE messages and replying with REPORTs, each ONU will receive and send multiple GATE/REPORT pairs, one per destination ONU.

Algorithm 1 LAN Resource Reservation Algorithm

1. The OLT compiles GATE messages from the grant allocation table.
2. GATE messages (one per ONU logical link) are forwarded in interleaved fashion to both sender and receiver ONUs.
3. GATE messages are sent "just in time" for ONUs to set up their 1×2 switches and transmit or receive data in the granted period.
4. After data transmission, each transmitter ONU sends a REPORT message to the OLT.
5. After receiving REPORTs from all ONUs, a scheduling algorithm for the following polling cycle is executed.
6. The grant allocation tables are updated, and the process is repeated.

As mentioned in the previous section, data packets destined for inter-ONU/eNB communication are aggregated in VOQs. GRANT and REPORT messages are transmitted over each ONU's dedicated channel in an interleaved fashion, as in IPACT [22], with a polling cycle $t_{\text{cycle}} = 2$ ms. A centralized traffic scheduler at the OLT coordinates transmissions, by notifying via GRANT messages each ONU when to start data transmission. Receiver ONUs are also notified to accordingly set up their 1×2 optical switch, in order to extract bursts of packets that reach their destination. 1×2 switches are set up to the "forward" state by default, so it is not necessary to notify intermediate hops of data transmissions. This concept is similar to the (centralized) just in time (JIT) scheduling in OBS networks [16], as the 1×2 switches have to be set up "just in time" for the burst.

To simplify the design of the scheduler and avoid floating point operations, time is discretized in fixed timeslots

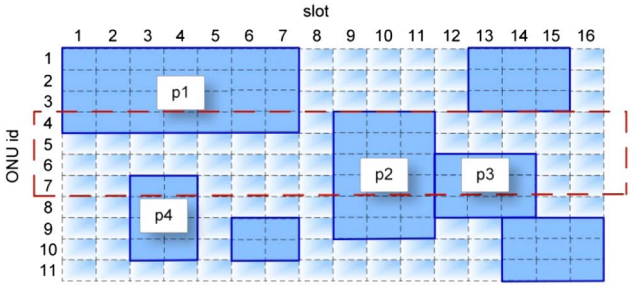

Fig. 5. TDM grid structure for traffic scheduling.

of duration τ; thus the utilization profile of each ONU's outgoing link is divided in a reservation window of 64 timeslots. This leads to a TDM grid structure, with TDM slot IDs at the X axis and ONU IDs at the Y axis (see Fig. 5). Current reservations are shown with p1, p2, etc., while the dashed red rectangle shows a potential new connection request from ONU 4 to ONU 7. The traffic scheduler must allocate a set of free continuous slots on every link across the path from source (s) to destination (d) to satisfy this request. Such a series of slots exists only during the eighth timeslot. The best known algorithm in the literature to efficiently solve the aforementioned scheduling problem is first fit (FF) scheduling [23]. FF processes the bandwidth requests in arbitrary order, attempting to "pack" the requests from left to right in the grid structure. However, FF is known to cause a high degree of fragmentation, due to the naive approach in selecting requests to be scheduled. Thus, the well-known first fit descending (FFD) policy is also evaluated, in order to enhance the FF algorithm performance. In the FFD implementation, the bandwidth requests are first sorted [an O(N*Log N) operation] and then scheduled with the FF algorithm starting with the lowest ONU id. Finally, a more "relaxed" version of the FFD was also developed, sorting the bandwidth requests solely by ONU ids (FFO). The latter can be implemented with O(N) complexity, by keeping per-ONU queues to store incoming bandwidth requests.

B. Downstream Capacity Sharing

In the proposed access architecture, subwavelength λ_{LAN} connections can be set up from congested ONUs to one or more uncongested ONUs, to facilitate downstream capacity sharing. For example, let us assume that λ_i is a heavily loaded wavelength channel dedicated to ONU_i and λ_k is a lightly loaded channel dedicated to ONU_k. A fraction of the downstream traffic destined to $\text{ONU}_i/\text{eNB}_i$ can be transported via λ_k and terminated at ONU_k. This traffic subflow, termed as *transient traffic flow*, is stored in the corresponding VOQ buffer of ONU_k and is retransmitted over the λ_{LAN} to reach its destination ONU, i.e., ONU_i. Finally, the destination ONU_i forward the excess traffic to the destination eNB_i via the X2 interface. It must be noted that traffic redirection is typically performed for low-priority "best-effort" traffic (e.g., P2P file transfers, Web traffic)

appropriately marked by the AGW with the corresponding CoS value. Here, a downstream wavelength sharing scheme is proposed, to support load balancing in the WDM-PON ring architecture. The proposed scheme is implemented at the OLT, which has access to the traffic profile of all downstream wavelengths. It includes a periodic polling-cycle operation that is divided in two phases. In the first phase, overloaded and lightly loaded wavelengths (or equivalently congested/uncongested ONUs) are identified. Further, the excess traffic of congested ONUs and unused capacity of uncongested (or donor) ONUs are quantified. In the second phase, the λ_{LAN} channel is exploited to absorb temporal traffic overloads by redirecting traffic from congested ONUs through uncongested ONUs. An algorithm is devised to dynamically match congested to uncongested ONUs. In what follows, the two phases of the proposed scheme are detailed:

1) First Phase: Identification of Donor and Acceptor ONUs: In the first phase, the OLT periodically performs estimates of underlying traffic parameters and calculates the effective bandwidth C_{EB} of all downstream channels using the Norros formula [24]:

$$C_{\text{EB}} = \mu + \left[B^{H-1}k(H)\sqrt{-2\alpha\mu \ln e}\right]^{\frac{1}{H}}, \tag{2}$$

where $K(H) = H^H(1-H)^{1-H}$, B is the buffer size (in bits), μ is the traffic mean rate (in bps), H is the Hurst parameter of the traffic, and a is the coefficient of variance. Effective bandwidth is a statistical estimate of the capacity required in order to satisfy a given QoS constraint (typically a buffer overflow probability). Lightly loaded channels (which correspond to donor ONUs) are the ones with $C_{\text{EB}} < C$, and heavily loaded ones (which correspond to congested ONUs) are those with $C_{\text{EB}} \geq C$, where C is the nominal wavelength capacity. Then, the OLT calculates a traffic counter (TC, in bps) for each ONU. The function of the TC is to quantify the excess traffic or the unused capacity of congested or uncongested ONUs respectively.

For uncongested (donor) ONUs, the TC corresponds to the unused downstream capacity. Assuming that N subflows are redirected through ONU_k and C_j is the effective bandwidth of each flow, the following equation must hold so that the donor ONU_k is not overloaded:

$$\sum_{N} C_j + C_{\text{EB}}^k \leq C. \tag{3}$$

This is a conservative estimate, as it does not account for statistical multiplexing gains from flow aggregation, providing a safety margin in case the underlying traffic profile changes. Thus, the donor capacity can be derived as the difference between the effective bandwidth of the downstream channel and the nominal wavelength capacity, i.e., $TC_k = C - C_{\text{EB}}^k$. Each time a new *transient flow* is redirected by a donor ONU, its TC is decremented accordingly.

For congested ONUs (where $C_{\text{EB}} \geq C$), TC corresponds to the effective bandwidth of the excess traffic component that must be redirected. For each congested ONU, downstream traffic is progressively split in a number of transient flows, until all excess traffic has been redirected. Each transient flow corresponds to a fraction of the overall mean rate, so that it can be matched and redirected over one donor ONU. Traffic splitting and redirection depends on the unused capacity of the associated donor ONUs (quantified with the TC) as well as the availability of the λ_{LAN}. The λ_{LAN} utilization profile is known to the OLT, since it is the OLT that arbitrates bandwidth allocations for the ONUs. Assuming ONU_i is a congested ONU, ONU_k is a donor ONU, and $A_{k,i}$ is the available bandwidth in the λ_{LAN} channel for the links in the path between ONU_k and ONU_i, the following equation must hold:

$$C_{i,k} = \min(TC_k, TC_i, A_{k,i}), \tag{4}$$

where $C_{i,k}$ is the effective bandwidth of the transient flow from (congested) ONU_i to (donor) ONU_k.

2) Second Phase: Dynamic Traffic Redirection: In the second phase, traffic redirection is implemented by setting up *transient flows* to be redirected from congested ONUs through uncongested ones, starting from the lowest priority classes. This requires a mechanism to efficiently split downstream traffic into multiple components (or subflows). In this work, a well-known stochastic hash-based technique is used for traffic splitting [25]. This technique involves calculating hash-based fingerprints on a packet-by-packet basis and classifying each packet to a subflow based on its fingerprint value. The fingerprint is calculated by applying a hash algorithm to the packet header five-tuple, i.e., source and destination address, source and destination port, and protocol number. Fingerprint values are mapped to traffic subflows that correspond to a fraction of the overall downstream traffic. For example, downstream traffic with mean rate m can be split into an excess traffic component, which corresponds to a fraction p of the overall traffic, with mean rate $m_{ex} = p \bullet m$. If no single donor ONU is able to accommodate the excess traffic component, the latter can be split into multiple subflows.

At the beginning of the second phase, after acceptor/donor ONUs have been identified and excess/unused capacities have been quantified, the proposed scheme attempts to match donor and acceptor ONUs. It is evident that matching a congested ONU with the closest uncongested ONU(s) in the upstream direction minimizes the number of hops spanned by the transient traffic flows, which results in maximizing the spatial reusability gains of the λ_{LAN}. Thus, for each congested ONU, the OLT considers all upstream donor ONUs sequentially. Transient flows are set up from donor ONUs to acceptor ONUs until all excess bandwidth is redirected or the λ_{LAN} capacity is exhausted. The basic algorithmic steps, which are repeated periodically at the OLT, are illustrated in Algorithm 2. It must be noted that changes to the underlying traffic profile result in changes to the *transient flow* setup.

Algorithm 2 Dynamic Traffic Redirection Algorithm

ForEach congested ONU_i
  ForEach upstream donor ONU_k
    $C_{i,k} = \min(TC_k, TC_i, A_{k,i})$
    If $(C_{i,k} > 0)$ **Then**
      Set up transient flow from ONU_k to ONU_i
      Update λ_{LAN} utilization profile
      Update TC_i, TC_k
      If $(TC_i==0)$ **Then**
        Break {no remaining excess traffic}
      EndIf
    Else{λ_{LAN} capacity exhausted}
      Report failure
      Break
      EndIf
EndFor
EndFor

VI. Performance Evaluation

In this section the unified LTE/PON architecture is evaluated, along with the proposed QoS framework and the dynamic bandwidth allocation/reallocation scheme. The proposed architecture and schemes were implemented in ns-2 simulator framework, and evaluated through extensive simulation experiments. Specifically, an LTE backhaul network with eight ONU/eNBs was simulated, interconnected in a fiber ring with a 2 km diameter. An OLT was connected to the ring with a 20 km bidirectional feeder fiber. The LTE air interface implementation was derived from an ns-2-based LTE simulator, presented in [26]. The LTE simulator also supports QoS-aware scheduling for both the air interface and the S1 interface. The scheduling algorithm and QoS mapping scheme presented in Section IV were implemented and evaluated via simulation experiments. In what follows three sets of simulation experiments are presented, to evaluate different aspects of the converged architecture and proposed schemes. In all simulation experiments, communication between each source–destination pair of ONUs (i.e., LAN traffic flowing through X2 interface) and between each OLT–ONU pair is modeled as separate traffic sources that generate packets according to a self-similar process with packet size 1 KB and $H = 0.7$, unless stated otherwise. ONU traffic from/to the OLT is denoted as downstream or upstream traffic, respectively, and ONU traffic load as downstream or upstream load. Traffic load is expressed as a percentage of wavelength capacity, which is the maximum amount of traffic that can flow through the links of the fiber ring. The wavelength capacity was set to 1 Gbps, and the buffer size of ONUs to 3 MB. Finally, the polling-cycle time for LAN traffic, t_{cycle}, was set to 2 ms.

A. Inter-ONU (LAN) Traffic

In the first set of experiments the effect of the proposed unified ONU architecture on inter-ONU (or LAN) traffic is evaluated. First, the proposed architecture is compared with a simple reference design that does not incorporate the λ_{LAN} wavelength. In the reference design all inter-ONU traffic flows through the OLT, via the dedicated upstream/downstream wavelengths. Figure 6 displays the delay of the inter-ONU traffic (or LAN traffic) for two architectures: the legacy one termed as *Reference Arch.*, in Fig. 6, where all LAN traffic is transmitted via the OLT, and the proposed one, termed as *New Arch.*, where all traffic uses the λ_{LAN} wavelength and the ONU design shown in Fig. 4. Results shown are obtained varying the LAN traffic load and the upstream/downstream load (denoted with ρ). Without loss of generality, the results concern a single source–destination pair. Packet delay at the proposed ONU design is dominated by the queuing delay at the VOQs. At light loads the average queuing delay equals $t_{cycle}/2$, as packets wait for their designated timeslot, which is essentially the overhead of the MPCP signaling. For a typical value of $t_{cycle} = 2$ ms this is translated to a 1 ms overhead, which is well below the tolerances of inter-ONU traffic. It can be seen that the packet delay in the new architecture is stable and remains significantly lower than the packet delay of the reference architecture, especially at high loads. When the upstream/downstream channels are congested, packet delay at the reference architecture increases significantly.

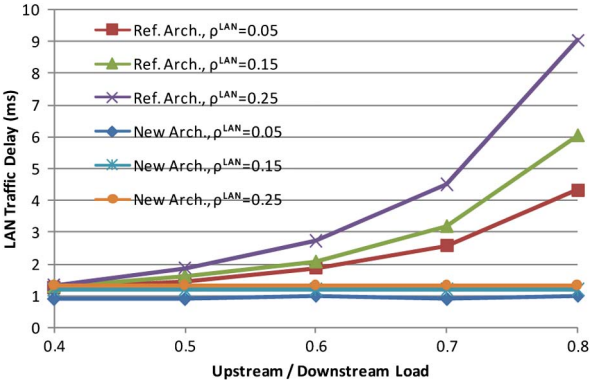

Fig. 6. Inter-ONU (LAN) traffic delay versus load.

In the following experiment, the proposed dynamic resource reservation scheme is simulated, for subwavelength sharing of the λ_{LAN}. As mentioned in the previous section, resource reservation for LAN traffic is arbitrated by the OLT via MPCP signaling, with a polling-cycle operation and $t_{cycle} = 2$ ms. In this experiment, the goal is to evaluate the performance of the scheduling policies. Figure 7 compares the packet loss ratios of FF, FFD, and FFO scheduling policies presented in Section V for varying LAN traffic loads, while packets that failed to be scheduled were dropped. For reference, the performance of an "ideal" scheduler is included, by removing the slot continuity constraint. It can be seen that FFD is the best performing heuristic and therefore is employed in the rest of this work.

B. QoS Framework for Upstream/Downstream Traffic

In the second set of measurements the QoS framework for upstream/downstream traffic, presented in Section IV,

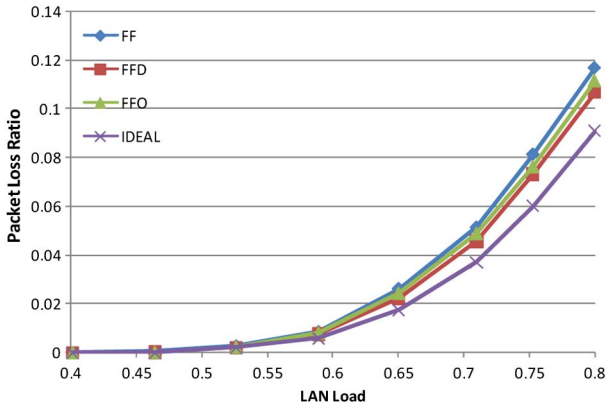

Fig. 7. Packet loss ratio versus load, for different scheduling policies.

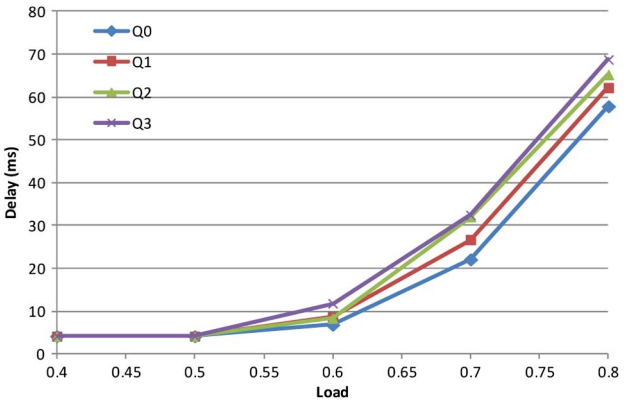

Fig. 9. Delay versus load for different priority classes without QoS mapping.

is evaluated. As mentioned previously, four traffic classes are assumed, Q0–Q3, which correspond to voice traffic, video traffic, interactive traffic, and best-effort traffic. The traffic load is split in all four classes, in proportions of 10%, 10%, 40%, and 40% respectively. Voice traffic is generated with a ns-2 RealAudio traffic generator, with burst time (or ON period) 200 ms, IDLE time (or silence period) 500 ms, and packet size 200 bytes. Constant bit rate (CBR) encoding was assumed for video traffic, and a packet size of 1 KB. Finally, traffic that belongs to classes Q2 and Q3 is modeled with a self-similar process, with packet size 1 KB and $H = 0.7$.

It must be noted also that the sending rate and/or packet inter-arrival times are derived from the traffic load and the packet size.

An important characteristic of the proposed QoS framework is mapping of traffic classes in wired and wireless networks. Figure 8 shows the delay of traffic classes for varying traffic loads, when QoS mapping is supported. Figure 9 repeats the measurements without support for QoS mapping, where all priority classes collapse to a single traffic class at the eNB. It can be seen that for fulfilling the QoS characteristics associated with the different bearers, QoS mapping is vital. In Fig. 9, differences in packet delay among traffic classes are not significant, and priority inversions are possible.

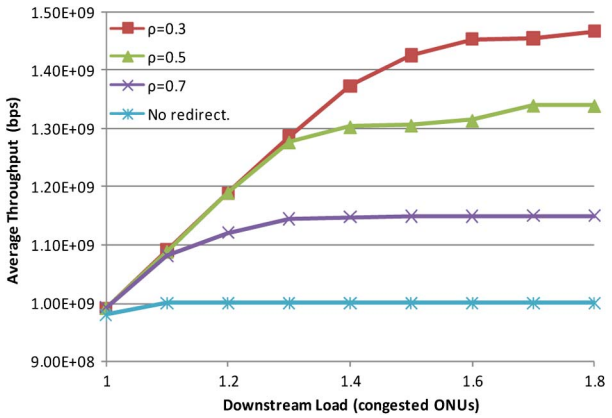

Fig. 10. Average throughput of congested ONUs versus downstream load. Uncongested ONU downstream loads are denoted with ρ, and the LAN load is set to 0.

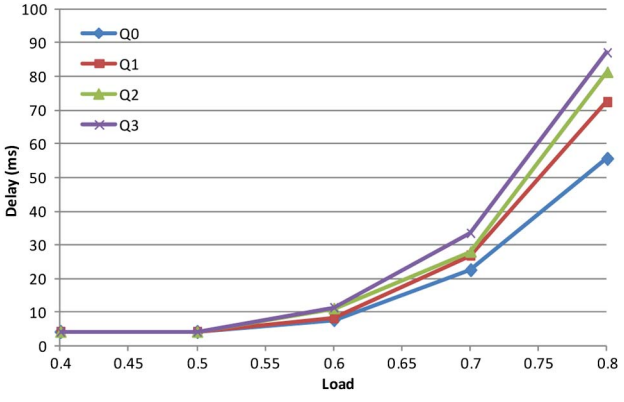

Fig. 8. Delay versus load for different priority classes and QoS mapping.

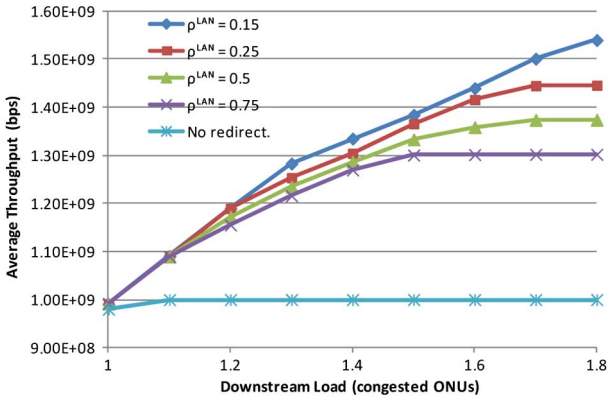

Fig. 11. Average throughput of congested ONUs versus downstream load for different LAN loads (marked ρ^{LAN}), and uncongested ONU downstream load set to 0.1.

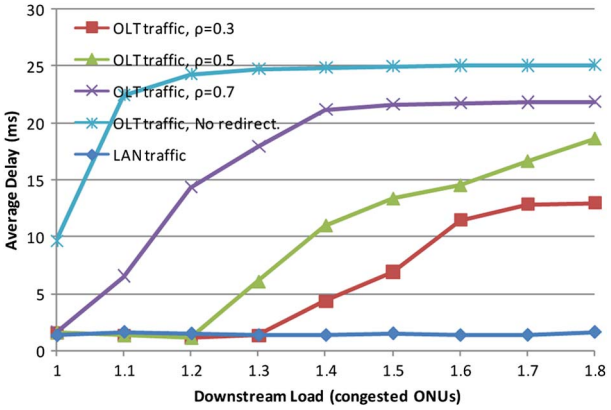

Fig. 12. Downstream traffic and LAN traffic delay versus downstream traffic load, for different uncongested ONU loads (marked ρ).

C. Load-Balancing Scheme

In the third set of measurements, the proposed scheme for dynamic bandwidth allocation and sharing of downstream wavelengths is evaluated. This set involves communication among all ONUs that belong to the LTE backhaul. Among the eight ONUs, four were randomly selected with input load ≥1.0 to simulate traffic overloads (congested ONUs) and four with input load ≤0.7 (uncongested ONUs). Figures 10 and 11 display the aggregated throughput of congested ONUs versus the downstream load. In Fig. 10, the downstream load of uncongested ONUs (denoted with ρ) is varied, while the LAN load is set to 0. In Fig. 11, the LAN load is varied, while the downstream load of uncongested ONUs is kept constant ($\rho = 0.1$). For reference, the case of no redirection is shown in both figures.

It can be seen in both figures that the proposed architecture can effectively take advantage of the unused capacity of uncongested ONUs, as long as there is sufficient bandwidth availability in the λ_{LAN} channel. In particular, for the case of 1.8 loads, the throughput of congested ONUs reaches 1.57 Gbps for an uncongested ONU downstream load of $\rho = 0.1$ as can be seen in Fig. 10. This is a significant improvement, keeping in mind that the transient traffic would have been lost if traffic redirection was not implemented. As the uncongested ONU load increases to $\rho = 0.7$ (and the unused capacity decreases accordingly) the congested ONU throughput gradually decreases to 1 Gbps. The same behavior is also recorded with the gradual increase of LAN load, as shown in Fig. 11.

Finally, Fig. 12 illustrates the end-to-end delay measured as the time that a packet leaves the OLT (or the originating ONU) until it reaches the destined congested ONU. In particular, Fig. 12 displays the delay of the LAN traffic as well as the delay of the downstream traffic (denoted as OLT traffic) for different loads of uncongested ONUs ($\rho = 0.3$, 0.5, and 0.7), which are compared with the case of no redirection. From Fig. 12, it can be seen that the delay of the downstream traffic increases but never exceeds the case of no redirection. Thus, the QoS for delay-sensitive services does not deteriorate when traffic redirection of excess traffic is employed. Furthermore, it can be seen that although the unused capacity of the λ_{LAN} channel is exploited for downstream traffic redirection, the delay of inter-ONU traffic remains unaffected and just marginally increases with the increase of the downstream load.

VII. Conclusion

In this work a new unified PON-RAN architecture for LTE mobile backhaul networks was proposed, employing ring-based WDM PONs. The proposed architecture supports all-optical inter-BS communication as well as load balancing, via the dynamic reallocation and sharing of downstream wavelengths. Lightpaths for inter-ONU communication are dynamically set up, bypassing intermediate ONUs, thus avoiding unnecessary electronic processing. A dynamic bandwidth allocation scheme was also proposed for resource reservation and sharing of LAN traffic. It was shown through simulation that the proposed architecture significantly improves the delay characteristics of inter-eNB traffic, offloading upstream/downstream wavelengths. Finally, the proposed load-balancing scheme effectively takes advantage of unused capacity, by redirecting traffic overloads to uncongested wavelengths.

Acknowledgments

This work is supported by the Cyprus Research Promotion Foundation's Framework Programme for Research, Technological Development and Innovation 2009-2010 (DESMI 2009-2010), co-funded by the Republic of Cyprus and the European Regional Development Fund, and specifically under Grant TII/EIIIKOI/0609(BIE)/07.

References

[1] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution: HSPA and LTE for Mobile Broadband*, 2nd ed. Academic, 2008.

[2] M. D. Andrade, G. Kramer, L. Wosinska, J. Chen, S. Sallent, and B. Mukherjee, "Evaluating strategies for evolution of passive optical networks," *IEEE Commun. Mag.*, vol. 49, no. 7, pp. 176–184, 2011.

[3] I. F. Akyildiz, D. M. Gutierrez-Estevez, and E. C. Reyes, "The evolution to 4G cellular systems: LTE-advanced," *Phys. Commun.*, vol. 3, no. 4, pp. 217–244, Dec. 2010.

[4] S. Parkvall, A. Furuskar, and E. Dahlman, "Evolution of LTE toward IMT-Advanced," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 84–91, 2011.

[5] M. Ali, G. Ellinas, H. Erkan, A. Hadjiantonis, and R. Dorsinville, "On the vision of complete fixed-mobile convergence," *J. Lightwave Technol.*, vol. 28, no. 16, pp. 2343–2357, 2010.

[6] C. Ranaweera, E. Wong, C. Lim, and A. Nirmalathas, "Next generation optical-wireless converged network architectures," *IEEE Netw.*, vol. 26, no. 2, pp. 22–27, 2012.

[7] H. Erkan, A. D. Hossain, R. Dorsinville, M. A. Ali, A. Hadjiantonis, G. Ellinas, and A. Khalil, "A novel ring-based WDM-PON access architecture for the efficient utilization

of network resources," in *Proc. IEEE ICC*, Beijing, China, 2008, pp. 5175–5181.

[8] K. Ramantas, K. Vlachos, G. Ellinas, and A. Hadjiantonis, "Efficient resource management via dynamic bandwidth sharing in a WDM-PON ring-based architecture," in *Proc. IEEE ICTON*, Coventry, UK, 2012.

[9] D. Breuer, F. Geilhardt, R. Hülsermann, M. Kind, C. Lange, T. Monath, and E. Weis, "Opportunities for next-generation optical access," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. s16–s24, 2011.

[10] S. Lambert, J. Montalvo, J. A. Torrijos, B. Lannoo, D. Colle, and M. Pickavet, "Energy efficiency analysis of next-generation passive optical network (NG-PON) technologies in a major city network," in *Proc. Int. Conf. on Transparent Optical Networks (ICTON)*, Cartagena, Spain, 2013.

[11] H. Erkan, G. Ellinas, A. Hadjiantonis, R. Dorsinville, and M. A. Ali, "Native Ethernet-based self-healing WDM-PON local access ring architecture: A new direction for supporting simple and efficient resilience capabilities," in *Proc. IEEE ICC*, Cape Town, South Africa, May 2010, pp. 1–6.

[12] E. S. Son, K. H. Han, J. H. Lee, and Y. C. Chung, "Survivable network architectures for WDM PON," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf. (OFC/NFOEC)*, Anaheim, CA, 2005, paper OFI4.

[13] K. Ramantas, K. Vlachos, G. Ellinas, and A. Hadjiantonis, "A converged optical wireless architecture for mobile backhaul networks," in *Proc. Optical Network Design and Modeling (ONDM)*, Brest, France, 2013, pp. 155–160.

[14] L. Kazovsky, S.-W. Wong, T. Ayhan, K. M. Albeyoglu, M. R. N. Ribeiro, and A. Shastri, "Hybrid optical–wireless access networks," *Proc. IEEE*, vol. 100, no. 5, pp. 1197–1225, May 2012.

[15] A. D. Hossain, R. Dorsinville, A. Hadjiantonis, G. Ellinas, and M. Ali, "A simple self-healing ring-based local access PON architecture for supporting private networking capability," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, Nov. 2007, pp. 2193–2198.

[16] Y. Chen, C. Qiao, and X. Yu, "Optical burst switching: A new area in optical networking research," *IEEE Netw.*, vol. 18, no. 3, pp. 16–23, 2004.

[17] S. Tanaka, S. H. Jeong, S. Yamazaki, A. Uetake, S. Tomabechi, M. Ekawa, and K. Morito, "Monolithically integrated 8:1 SOA gate switch with large extinction ratio and wide input power dynamic range," *IEEE J. Quantum Electron.*, vol. 45, no. 9, pp. 1155–1162, Sept. 2009.

[18] H. Ekstrom, "QoS control in the 3GPP evolved packet system," *IEEE Commun. Mag.*, vol. 47, no. 2, pp. 76–83, 2009.

[19] W. Lim, P. Kourtessis, K. Kanonakis, M. Milosavljevic, I. Tomkos, and J. Senior, "Modeling of LTE back-hauling through OFDMA-PONs," in *Proc. Optical Network Design and Modeling (ONDM)*, Brest, France, 2013, pp. 240–245.

[20] R. I. Davis and A. Burns, "A survey of hard real-time scheduling for multiprocessor systems," *ACM Comput. Surv.*, vol. 43, no. 4, pp. 1–44, Oct. 2011.

[21] "Ethernet in the First Mile," IEEE 802.3ah, June 2004.

[22] M. Assi, S. Dixit, and M. Ali, "Dynamic bandwidth allocation for quality-of-service over Ethernet PONs," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 9, pp. 1467–1477, 2003.

[23] C. Chekuri and S. Khanna, "On multidimensional packing problems," *SIAM J. Comput.*, vol. 33, no. 4, pp. 837–851, 2004.

[24] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 953–962, 1995.

[25] Z. Cao, Z. Wang, and E. Zegura, "Performance of hashing-based schemes for Internet load balancing," in *Proc. IEEE INFOCOM*, Tel Aviv, Israel, 2000, pp. 332–341.

[26] Q. L. Qiu, J. Chen, L. D. Ping, Q. F. Zhang, and X. Z. Pan, "LTE/SAE model and its implementation in NS 2," in *Proc. 5th Int. Conf. on Mobile Ad-hoc and Sensor Networks*, Fujian, China, 2009, pp. 299–303.