Abstract—There is an emerging wide interest to transition from legacy WAN transport technologies to Ethernet-based technology. The current round of carrier Ethernet standards will successfully equip Service Providers (SPs) with the required tools to provide carrier-grade scalability and to provision and engineer connection-oriented point-to-point (P2P) packet trunks across a native Ethernet infrastructure. Building on these standards, this paper demonstrates how to support and implement full traffic engineering in a global-scale, two-tiered native Ethernet-over-WDM optical networking architecture. To achieve these objectives, several networking innovations are presented and developed including: 1) a GMPLS-based unified control plane that offers a tighter integration between layer-1 (optical transport layer) and layer-2 (Ethernet layer), 2) a fully distributed integrated routing and signaling framework for dynamically provisioning Ethernet Switched Paths (ESPs) at any bandwidth granularity including both full wavelength and finer granularity (sub-lambda) ESPs in an integrated Ethernet-Optical networking environment, and 3) a novel notion of an integrated Link-State Advertisements strategy that is consistent with a fully integrated routing and signaling protocol.

Index Terms—Optical Unified Control Plane, Optical Signaling and Routing Algorithms, GigE-over-WDM.

I. INTRODUCTION

ETHERNET has been the dominant LAN technology with the majority of all data traffic terminating on an Ethernet port. Recent dramatic advances in Ethernet and optical networking technologies fueled by Gigabit Ethernet and 10 Gigabit Ethernet are pushing Ethernet footprint into the Metro and Wide Area Networks (MANs/WANs). Ethernet Services are evolving beyond the LAN and are becoming a dominant technology in the MANs/WANs. However, for Ethernet to evolve as a next generation networking technology and rival Frame Relay, ATM and Private Lines, it has to evolve to support voice, video, and data at wire speed with a predictable quality of service (QoS). The Metro Ethernet Forum (MEF) has defined this evolution as “Carrier Ethernet”, which has several attributes including scalability, stringent QoS, fault management, and resiliency. The simplicity, low cost, and ubiquitous of Ethernet has resulted in increased interest among industry and standards in carrier-grade Ethernet as a lower-cost alternative to traditional WAN transport infrastructure [1-9].

The fundamental problem is that the majority of today’s SP infrastructures are built on legacy circuit-based infrastructures (SONET/SDH), and private line services are the basis for the majority of frame relay, ATM, and IP services. In addition to the higher cost and potential interworking problems associated with supporting multiple technologies, the speeds of legacy layer-1 access solutions are rather limited, and the cost of dedicated high-speed WAN access solutions is prohibitively high.

Enterprise data traffic, meanwhile, is nearly all Ethernet. But once it departs the corporate LAN and heads onto the MAN/WAN, it is converted and encapsulated into one or more different protocol and framing formats, only to be translated back into Ethernet once it reaches its destination. These conversions are unnecessary—if native Ethernet can be transported end to end. The fact that well over 95% of all data traffic either originates or terminates as Ethernet, coupled with its simplicity, ubiquity, and natural support for IP services, provides a compelling case for Ethernet to evolve as a convergence transport solution for next generation networks.

But before Ethernet can be adopted as a global carrier-grade transport infrastructure, it must overcome several technical hurdles and, at least, achieve carrier grade quality. To understand the challenges facing Ethernet as a global multiservice transport infrastructure, one has to breakdown the generic term “Ethernet technology” into its key three elements, namely, Ethernet interfaces, Ethernet services, and Ethernet transport [10]. Ethernet interfaces and services are widely deployed. However, Ethernet as a transport technology still faces several key challenges including, scalability, QoS guarantees, and resilience. Ethernet as a transport technology is not a requirement for delivery of Ethernet services. In fact, Ethernet WAN services are currently transported across SONET/SDH-based transport networks (Ethernet-over-SONET) or MPLS-Based networks (Ethernet-over-MPLS).

Specifically, Ethernet has faced a number of scaling challenges due to its nature. These include the flat addressing structure and lack of routing hierarchy, the
problems associated with Ethernet’s connectionless behavior such as MAC learning, flooding/broadcasting processes, and the use of spanning Tree Protocol (STP) to route traffic. These issues combined have placed limits on the upper bounds of sizing a native Ethernet network.

To address the list of outstanding issues described above, a flurry of carrier Ethernet standards and initiatives are underway including the IEEE 802.1ad (Provider Bridge) [7], the IEEE 802.1ah (Provider Backbone Bridges) [8], the IEEE 802.1aq (Connectivity Fault Management) [9] and, more recently, the “Provider Backbone Transport” (PBT) initiative, which is an innovative Ethernet technology currently being introduced into the relevant industry standards bodies (IEEE, ITU, IETF, etc) [5]. PBT is a technique for providing traffic and network engineering of P2P backbone paths over 802.1ah networks. This is achieved by replacing the MSTP or 802.1aq control plane with a PBT management plane or external control plane to populate the switch bridging tables. Another alternative being introduced into the relevant industry standards bodies is the use of GMPLS a control plane to control PBT’s bridge-forwarding tables. GELS (GMPLS Ethernet Label Switching) BOF is part of current work within the IETF, trying to extend the GMPLS control plane to configure VLAN-aware Ethernet switches to establish P2P and point-to-multipoint (P2MP) ESPs [11-12].

All of the above initiatives and standards, however, have mainly focused on addressing and engineering just the Ethernet logical connectivity (layer-2 logical topology) of the network. The fact remains that this engineered logical connectivity must be built on top of either static single wavelength-based or a fully dynamic wavelength division multiplexed (WDM)-based physical topology. In a Metro network, which is mostly the near-term target of these standards, the core nodes are connected via fixed single/multiple wavelength(s) pipes in which the provisioned ESPs are groomed and overlaid on top of a static physical topology. In this case, the target is to merely optimize layer-2 (L-2) logical connectivity, as the underlying physical transport layer is static, which is an adequate objective.

However, in the case of an all Ethernet-based, global-scale network, which is the target of this work, the underlying physical topology must be dynamically reconfigurable WDM-based optical transport layer (layer 1). In this case, the objective is to optimize the combined topology and resource usage information at both the physical (L-1) and logical (L-2) layers. This would require a unified control plane that manages both GigE and optical switches (contrary to the GELS BOF initiative, which proposes an external control plane to manage only GigE switches at the logical topology).

We have recently proposed a novel native Ethernet-based networking architecture and switching paradigm to scale metro Ethernet networks into a global multi-services infrastructure [13-16]. Specifically, we have proposed a truly native end-to-end layer-2 MAC frame-based Optical Ethernet infrastructure seamlessly stretching from enterprise LAN to Metro to Global. By combining the simplicity and cost effectiveness of Ethernet technology with the ultimate intelligence of WDM-based optical transport layer, Optical Ethernet (Ethernet-over-WDM) could evolve as a next generation global multi-service infrastructure. The proposed “Ethernet-over-WDM” model is truly a two-layer model where native Ethernet frames are mapped directly over WDM. It offers advantages over existing L-2 and MPLS solutions in that it divorces the Ethernet from legacy transport mechanisms like SONET/SDH and other L-2 protocols.

Building on the current round of carrier Ethernet standards, this paper demonstrates how to support and implement full traffic engineering in a global-scale two-tiered native Ethernet-over-WDM optical networking model. To achieve this ambitious objective, four salient features must be implemented:

1) Unlike PBT, where STP, MAC learning, and broadcasting/flooding are partly disabled (VLAN IDs spaces are allocated to PBT, MSTP, or 802.1aq), these functionalities are totally disabled in the model proposed here. Note that Ethernet standards currently allow turning off these Ethernet functionalities. Furthermore, the proposed model is generic and is not limited to supporting just 802.1ad and 802.1ah networks. A novel converged Layer 2/1 networking model in terms of control plane functionality must be developed.

2) A fully distributed integrated routing and signaling framework for dynamically provisioning ESPs at any bandwidth granularity including both full wavelength and finer granularity (sub-lambda) ESPs in an integrated Ethernet-Optical networking environment.

3) An innovative integrated Link-State Advertisement (LSA) strategy (a single-level updating scheme) that is consistent with a fully integrated routing and signaling protocol. As will be shown, this is a challenging task since two-level topological updates information (physical and logical) is needed. However, if two independent updates are used, the updating scheme is not integrated and extra redundancy is added.

Specifically, the proposed model supports an GMPLS-based unified control plane that manages both GigE and optical switches and offers a tighter integration between layer-1 (optical transport layer) and layer-2 (Ethernet layer), leading to the collapse of the two layers into a single integrated layer managed and traffic engineered in a unified manner. The unified control plane is optical layer-based, where most of the networking functionality and intelligence is delegated to L-1 (the optical layer). In this case, optical switches, rather than GigE switches, manage and control network resources (both physical and logical). Consequently, the
optical layer of the proposed model can concurrently provision/restore both full lambdas and finer granularity ESPs. Note that the proposed optical layer-based unified control plane is generic and transparent to the higher layer protocols. The implementation procedures are identical regardless of the kind of Provider edge (PE) devices that are attached to the optical cloud including IP/MPLS routers, ATM switches, etc.

Traditionally, GMPLS-based signaling standards proposed by IETF, such as Resource ReSerVation Protocol with Traffic Engineering extensions (RSVP-TE), were developed to provision full wavelength channels at the optical layer only. Likewise, MPLS-based signaling standards developed by IETF were also developed to provision label-switched paths (LSPs) at the logical layer only. However, there are no integrated signaling standards that can concurrently provision both full lambdas and finer granularity ESPs/LSPs at a single layer.

To the best of our knowledge, this is the first integrated signaling scheme that can concurrently setup and teardown both full wavelength and finer granularity ESP requests solely at the optical layer. The proposed integrated signaling protocols go beyond those being developed within GMPLS by the IETF and OIF. Provisioning of diverse traffic granularities entirely on the optical layer’s terms, as this work will show, introduces numerous new challenges including additional control plane complexities.

It is important to emphasize that, although the development of a GMPLS-based unified control plane has been a central issue in IETF during the past 10 years, the main objective was to develop an integrated L-1/L-3 control plane (peer model control plane) where managing and controlling network resources were delegated to L-3 (the edge IP/LSR router). This model never materialized due to the router’s major scalability problem, which was caused by the significant amount of state and control information that must flow between the IP and optical layer. To avoid falling in the same trap, in the model proposed here, the function of managing and controlling network resources (both physical and logical) is rather delegated to L-1 optical switches.

The remainder of this paper is organized as follows: Section II presents the network model, and Section III presents the hybrid optical node architecture. Section IV introduces the notion of a fully intelligent optical networking layer, and Section V presents the unified layer 1-2 control plane that manages both GigE and optical switches. The integrated routing and signaling framework is presented in Section VI. Simulation results are discussed in Section VII and Section VIII offers summary and conclusions.

II. THE NETWORK MODEL

The proposed “Ethernet-over-WDM” is a truly two-layer model that has no intermediate layers (e.g., IP/MPLS/ATM/SONET). However, the important functionalities provided by these layers (traffic engineering in ATM, routing/restoration in IP/MPLS, and multiplexing and fast restoration in SONET) must be retained by the proposed model. Thus, these functionalities must be distributed between the model’s two layers (layer-1 and layer-2). However, typical L-2 functionalities have been disabled (STP, MAC learning, and broadcasting/flooding). In this case, these functionalities, along with most of all other networking functionalities and intelligence must now be passed on down to the optical layer (the only available option). This would require three optical networking innovations: 1) a fully intelligent and agile optical transport layer; 2) a novel hybrid optical node architecture; and 3) an integrated control plane that manages both layers (later-1 and layer-2), which must be owned by the optical layer rather than by the GigE switches.

In the network model considered here, client GigE switches (might also be owned by the SP who owns the Optical Transport Network (OTN), e.g., PE switches) are attached to a fully intelligent OTN. The OTN consists of multiple hybrid optical nodes interconnected via WDM links in a general mesh topology. The edge GigE switches are clients of the OTN and are connected to their peers over dynamically switched lightpaths spanning potentially multiple optical nodes.

III. A HYBRID OPTICAL NODE ARCHITECTURE

In order to efficiently utilize the capacity of each wavelength channel (lightpath), the traffic of several independent lower-speed ESPs must be multiplexed onto a single lightpath. The process of combining low-rate traffic streams onto high-capacity optical channels (lightpaths) is known in the literature as “traffic grooming” [17-18]. To support traffic grooming, the cross-connect fabric of each optical node should have the capability of switching traffic at the wavelength granularity as well as at finer granularities. Therefore, a hybrid switching solution that capitalizes on existing electronic switch fabrics (GigE switches) as well as all-optical switch fabrics, by making use of each switch’s functionality and capability, appears to be the most appropriate for building the optical nodes of the proposed Optical Ethernet.

The backbone node architecture of the proposed model is composed of 3 key modules: (1) All-optical switch fabric (Optical Cross-Connect (OXC)): Performs pure optical switching without wavelength conversion capabilities, where the granularity of switching is the entire wavelength. (2) Backbone electronic switch fabric (GigE switch): Capable of multiplexing, de-multiplexing, and switching low-speed traffic streams (ESPs) onto the wavelength capacity. The backbone GigE switch is attached to the optical switch fabric through an array of transceivers and can generate and terminate the traffic to/from a lightpath. The number of
wavelength channels that can be terminated/generated into/from the electronic switch is a function of the transceiver array size. (3) OXC-Controller: A non traffic-bearing IP/MPLS-based intelligent control plane module managing both optical and logical domains (layers-1 and -2).

IV. A FULLY INTELLIGENT AGILE OPTICAL NETWORKING LAYER

To realize the “ultimate vision” of an agile, fully intelligent optical networking layer capable of supporting integrated routing and signaling algorithms for real-time provisioning of ESPs at any bandwidth granularity (on a per-call basis), the following two salient features must be implemented: 1) Most of the networking functionalities and intelligence must be migrated down to the optical layer including switching, protection, traffic engineering, OAM&P, provisioning of both full lambda and sub-lambda ESP requests, and selective restoration, all supported entirely on the optical layer’s terms; 2) The optical layer must own and manage both the physical connectivity and resources (layer-1 optical resources) and logical connectivity and resources (layer-2 Ethernet resources). Thus, both the logical and physical topologies now belong to a single administrative domain managed and controlled by the optical layer, leading to the creation of a unified control plane with the optical layer running a single integrated routing/signaling protocol instance.

V. A UNIFIED CONTROL PLANE FOR MANAGING BOTH GIGE AND OPTICAL SWITCHES

To implement the proposed optical layer-based unified control plane, the OXC controller, in addition to its conventional functionality of managing the optical layer resources including full wavelengths and fibers on physical links, must now support databasing and provisioning of finer granularity sub-lambda connection requests (ESPs) as well. This is achieved by simply augmenting the OXC controller with an additional resource usage database for sub-lambda connection requests. Thus, each OXC controller is assumed to maintain a single topology and resource usage database for both layers [19-20].

The OXC controller is now responsible for creating, maintaining and updating both the physical and logical connectivity tables. The responsibility of the edge GigE switch (which had most of its typical functionalities disabled) is, then, simply to request a service from the OTN (both full-lambdas and sub-lambdas) and the latter is responsible for providing this service. Each client/provider edge GigE switch registers its unique MAC/VLAN ID or a combination of both with the corresponding attached OXC-controller. The OXC controllers then run an IP routing protocol amongst themselves to determine all destinations (of all registered edge switches attached to the core network) reachable over the optical network.

The OXC controllers are assumed to communicate with each other over a data communications network (DCN). Using either out-of-band or in band signaling, the DCN will serve as the packet transport network for all the signaling messages required for connection set-up and tear-down in the optical transport network. Thus, the data plane is composed of native Ethernet frames (without any encapsulation) which flow between backbone GigE switches, while the control plane consists of (G)MPLS-based messages, which are being exchanged (out-of-band) between controllers. Note that the two planes are completely segregated.

While GigE switches currently do not support control plane routing and/or signaling protocols, such protocols can run at the external OXC controllers, which are connected to the switches’ control ports.

External control plane routing and signaling could process RSVP-TE messages on behalf of these switches, perform bandwidth management through connection admission control protocols, and then issue SNMP commands to configure the GigE switch fabrics.

Note that layer-2 is neither aware of, nor does it maintain any topology and/or resource usage information about the optical layer or its own layer. Furthermore, there is no exchange of information between the boundaries of the two layers except for that of the simple UNI. This means that the two layers are topologically and functionally isolated. The service provider infrastructure is now fully isolated from the customer broadcast domain and VLANs partitioning.

In addition, most of the burden has been off loaded from the PE devices to the OXC controllers residing within the optical layer.

VI. AN INTEGRATED ROUTING AND SIGNALING FRAMEWORK

Provisioning of connection requests requires routing algorithms for path selection, and signaling mechanisms to request and establish connectivity within the network along a chosen path. The source-based path selection approach used here is based on a fully adaptive integrated routing scheme, introduced recently by the authors [21]. The model proposed in [21] uses the combined topology and resource usage information at both the IP (layer-3) and optical layers, but has been modified in this work to utilize the combined topology and resource usage information at both the Ethernet (layer-2) and optical layers.

Two types of distributed signaling protocols have been investigated in the literature, namely forward and backward reservation signaling protocols [22-23]. In the forward scheme, reservations are made in the forward direction where the source node sends a reservation message towards the destination along the selected path.
In the backward scheme, the source node sends a probe message to the destination without reserving any resources. This message will collect information about resource usage along the path. Upon receiving the probe message, the destination node then sends a reservation messages to the source node along the same path, where each intermediate node starts reserving the appropriate network resources. To conform to GMPLS-based signaling standards being proposed by IETF, where RSVP-TE has been selected as the signaling protocol for optical networks, this work adopts the backward reservation scheme.

A. Generic Protocol Description

In the proposed link-state approach, each optical node must maintain the complete network state information, including the combined topology and resource usage information at the Ethernet and optical layers. Nodes exchange such information with each other through Link-State Advertisements (LSAs). All the nodes must be informed when the state of the network changes.

As mentioned above, a forward-probing and backward-reserving scheme is used to setup a connection. Upon receiving a connection request, the source node performs explicit routing, i.e. the source node selects the full route and assigns a wavelength (in the case of setting up new lightpath). In the case of multi-fiber environment, as it is assumed here, if the assigned wavelength is available on more than one outgoing fiber owned by an intermediate node along the path, the decision on which fiber to select is decided locally by that node.

The source node then sends a PROBE message, which contains the explicit path, towards the destination along the selected path. As the PROBE message travels downstream through the intermediate nodes along the path, it simply probes the involved nodes on the availability of their resources that the path requires. No reservation is taking place during the probing. If the PROBE message is processed at all intermediate nodes successfully and reaches the destination, the latter reserves (allocates) its resources and sends an ACK message upstream towards the source node. An ACK message is actually a reservation message, which means that if a node generating it finds its resources in question still available, it will allocate them to this request (rendering them unavailable for further requests). When an ACK reaches the source of the call, it too will check if its resources are still available and, if so, will admit the call in the network.

If a node processing a PROBE finds the resources in question unavailable, it will stop the signaling by sending a NACK in the upstream direction, to the source. If a node processing the ACK message finds the resources in question unavailable, it will send two messages: a NACK upstream to the source informing it of the failed signaling, and a downstream RELEASE message to the destination informing the downstream nodes, that have already reserved resources for this request, to release them.

While the implementation of the generic signaling scheme described above is rather straightforward when independently provisioning full wavelength ESPs at the optical layer or finer granularity ESPs/LSPs at the logical layer (i.e. when a sequential provisioning approach is followed), it becomes more challenging and complicated when an integrated approach (that can concurrently setup both full wavelengths and finer granularities ESPs at the optical layer alone) is assumed. To illustrate this argument through an example, consider that a sub-lambda ESP request is to be provisioned. Assume that the integrated routing algorithm invoked by the source node returns a hybrid path, where the call is to be routed over a combination of existing and newly created lightpaths. In this case, the source node might need to (possibly remotely) setup new lightpaths in addition to reserving bandwidth on existing lightpaths, in which case the probed/reserved resources at the involved nodes are not always of the same type (i.e. logical or physical), but rather a function of the explicit path.

One simple approach to resolve this problem is to use two independent generic PROBE and RESERVATION signaling messages, one for setting up new lightpath(s) at the optical layer and the other one for reserving the requested ESP's bandwidth on existing lightpath(s) at the logical layer (layer-2). However, the choice to follow a truly integrated signaling (i.e. avoid the introduction of new signaling messages, while retaining a single forward-probing backward reserving scheme) invokes the need to introduce further intelligence in the way signaling messages are formatted and processed.

B. Signaling Under An Integrated Approach: Innovations and Challenges

The main characteristics of the proposed signaling scheme are: 1) it supports a single instance of integrated routing and signaling protocols running solely across the optical layer, 2) the integrated routing approach is a fully adaptive scheme that uses the combined topology and resource usage information at both layers (1-2) to select an explicit path. A Data Communications Network (DCN) is assumed to serve as the packet transport network for all the signaling and network state updates messages. The DCN is assumed to be an exact replicate of the physical topology and a control channel operating at a lower speed than data channels is reserved for signaling.

To devise an integrated signaling protocol that utilizes a single unified probe and reservation message as well as to concurrently provision both full- and sub-lambda ESP requests at the optical layer, the following assumptions and innovations are introduced:

1. In contrast to conventional provisioning of full wavelength channels at the optical layer where
signaling messages are processed at each node along the entire path (on a hop-by-hop basis), signaling messages for provisioning sub-lambda ESP requests are only processed at the source of the existing lightpath (LP).

2. In contrast to conventional provisioning of full wavelength channels at the optical layer where each node along the path owns its local resources including its outgoing physical links all the resources along the path are assumed to be owned by the source that sets up the LP. Thus, the source node of an existing LP owns all the logical link resources, e.g. all outgoing physical links, wavelength channel residual capacity, and ports comprising the entire lightpath. By owning, it is meant that the “owner node” which allocates/de-allocates the call resources (a full lambda and/or sub-lambda), has the most updated status information about these links (physical link (PL) and/or logical link (LL)), and that any changes along the path can be advertised and updated only by the owner node. This implies that when a new LP is setup, the ownership of the entire LP resources, (i.e. local physical resources of each intermediate node along the path), is transferred to the source of the new LP, and when a LP is torn down, the ownership of these physical resources is transferred from the source of the torn-down LP to the intermediate nodes that originally owned them only when these intermediate nodes have received: a) the tear-down update from the torn-down LP source, or b) a TEAR_DOWN message that explicitly allows the transfer (this addresses the two-ownership challenge, which is explained in Section VI.D.2).

3. If a signaling message is to traverse a logical link (single LP), the message is assumed here to be forwarded along the static shortest path (SP) between the endpoints of the lightpath and need not follow the exact data path (out-of-band signaling). This means that intermediate nodes along the SP receiving a signaling message only forward it further, without processing it since they are not involved in the signaling process. In addition, since NACK messages impose no intelligence in processing at intermediate nodes, they, too, are assumed to be forwarded instead of being processed until they reach the source of the call.

4. If an ESP is to be provisioned over a path whose last logical link is an existing LP, the PROBE message need not traverse the entire path; it is forwarded only up to the source of the last existing LP. This implies that a connection request provisioned over a single existing lightpath requires no signaling, as the source of the call owns the resources the call setup requires.

5. When a node receives a PROBE message, it needs to identify the probed resources. This is achieved by examining its previous (upstream) and next (downstream, if applicable) links in the path contained in the PROBE message.

6. Due to the added complexity and cost it would introduce, global synchronization among network nodes when exchanging LSAs is not assumed. In other words, time-stamping the control messages is not necessary for deadlock-free protocol operation.

It is important to emphasize that all the assumptions and ideas presented in this work are the result of running extensive simulation experiments that initially revealed the challenges, deadlocks, and complexity of the protocol.

As an illustrative example, Figure 1 shows how messages are processed for four different cases. Assume that SP(a, b) denotes the static shortest path (short, in terms of distance) between nodes a and b. Figure 1a shows that signaling messages are processed at each node (on a hop-by-hop basis) when a connection is being provisioned over a newly created single LP (purely RWA) along path S-1-2-D. Figure 1b shows that no signaling is required when a sub-lambda connection is being provisioned over a single existing LP (purely logical). Figure 1c shows that signaling messages are processed only at the sources of LPs when a sub-lambda ESP request is purely logically provisioned over two existing LPs (logical links LL1 and LL2). Note that the PROBE message reaches only up to the source of LL2, where an ACK is generated, and that signaling over LL1 is performed over SP(S,3) (shown with one intermediate unmarked node), rather than the lightpath route.

Figure 1d considers the case of hybrid provisioning where a call is being serviced over three segments from S to D; the first segment is an LP to be setup with source S and destination 1 over physical link PL1; the second segment 1-2-3-4-5 (LL1) is an existing LP with source 1 and destination 5; the third segment is another LP to be setup with source 5 and destination D, over PL2, PL3 and PL4. Note that the PROBE message is processed only at nodes S, 1, 5, 6, 7 and D, and that the LL1 segment is signaled over SP(1,5) (shown with one
intermediate unmarked node, which simply forwards, rather than processes, the messages).

As mentioned above, when a node receives a PROBE message, it identifies the probed resources by examining the previous (upstream) and next (downstream) links. For instance, as shown in Figure 1d, when node 5 receives the PROBE message, since the path downstream link is a PL (PL2) and the upstream link is an LL (LL1), it checks on the availability of both PL2 and a transmitting port, as this node will be the source of a newly created lightpath to serve the call. On the other hand, when node 1 receives the PROBE message, since the path downstream link is an LL (LL1) and the upstream link is a PL (PL1), it will check on both the availability of LL1 and a receiving port, as it will be the destination of a lightpath to be created to serve the call.

C. Connection Release

When a call is terminated, the network frees up the resources allocated to the call. Since sub-lambda ESP requests are assumed, the LPs are torn down when the last call using them departs the network. The signaling for a call release may consists of a single message, RELEASE BW, or possibly two messages, RELEASE BW and TEAR DOWN. The RELEASE BW message contains the call ID. A node processing (or generating) such a message, de-allocates the resources the call was using and then propagates the message downstream to the next LP’s source until it reaches the source of the last LP along the path. The source nodes then update the status of their lightpaths by increasing their residual bandwidth by an amount that corresponds to the call’s released bandwidth. RELEASE BW message is transmitted (over the shortest path) to, and processed at all the sources of the involved LPs.

In addition, at each of these LP sources, the need for a TEAR_DOWN message transmission is also examined. Thus, the node always checks the LP’s residual capacity (after releasing the call’s bandwidth). If the residual capacity is equal to the full wavelength channel capacity (last call to use the LP), the node releases the hardware (e.g. optical switch ports) associated with the LP and then generates a TEAR_DOWN message that propagates downstream to all nodes comprising the LP to be torn down over the actual data path.

D. Updating

1) Link-State Advertisements (LSAs): In the proposed link-state approach, each optical node must maintain complete network state information, including combined topology and resource usage information at both the Ethernet and optical layers. Nodes exchange such information with each other in a distributed manner via LSA messages. Whenever the state of the network changes, all the nodes must be informed. In other words, all nodes must maintain a synchronized and identical topology and link state information. Therefore, a change in the network status such as a call arrival/departure (setting-up/tearing-down one or more lightpaths or LSPs) may result in broadcasting of update messages to all nodes in the network. In this work an update can be triggered by a timer (timer-triggered) or by a change (change-triggered). Under the first scheme, a node updates the network periodically whenever a local timer expires (unsynchronized with remote node timers). Under the second scheme, the updates (changes created at the source node due to a change in the network status) are sent out to all other nodes whenever the number of changes reaches a predefined threshold.

To realize the updating part of the protocol, every node maintains a list of changes that occurred and involve its local resources (i.e. its outgoing physical and logical links). In this list, changes are added whenever they occur, provided that they do not cancel out or alter an old change already in the list. In other words, if a new change is to be added, it should not be relevant to an old change that exists in the list. If it is, as shown in Figure 2, one of two actions is taken: if the new change cancels out the old change, the new change is not added and the old change is removed (Figure 2, top); if the new change alters an existing old change, the new change is not added and the old change is replaced by one that reflects the final state after alteration (Figure 2, bottom).

The LSA messages have a unique ID that is tied to the originator node (of local significance). When a node receives an LSA it first has to check if it already processed this LSA. This is done by comparing the received LSA’s ID with a collection of received <LSA_ID, originator> combinations it keeps. If it is an already processed LSA, it is discarded. If, however, it is a fresh LSA, the node adds the <LSA_ID, originator> combination to the list of received LSAs and processes it.

2) Updating Challenges: The absence of global synchronization, the fully distributed scheme, and the
two-level topological environment, introduce several challenges on how to properly update the network status information between the nodes. In fact, two major challenges, namely, the two-level (physical and logical) topological information updates challenge and the two-ownership challenge, face the implementation of an integrated updating strategy that is consistent with a fully integrated signaling protocol. The first challenge arises due to provisioning of both full lambda and sub-lambda ESP requests at the optical layer. In order to exchange such updating information among different network nodes, two-level (physical and logical) topological information updates are needed. However, if two independent updates are used, the updating scheme is not integrated, and extra redundancy is added.

To address this challenge, we propose a single-level updating scheme where the sole updating entity is the LL availability (i.e. lightpath status). Information about the physical links can then be extracted from the lightpath state. The assumption is that physical resources in the network are not added or deleted as frequently as logical resources and can be thus considered constant. For instance, it is reasonably assumed that any increase in the number of supported wavelengths on WDM links, or additions of extra fibers will occur very infrequently, allowing the network physical resources set to be considered fixed. What frequently changes in the status of an operational network is its logical resource availability, which is essentially a translation of resources from PLs to LLs and vice versa. Thus, if PRS(t) and LRS(t) are the sets of Physical and Logical Resources as a function of time, the full view of the network status can be obtained by extracting the PLs used by the LLs from the initial physical resource status (network status before any traffic, considered as fixed). Then:

$$PRS(t) = PRS(initial) - LRS(t)$$

The second challenge arises due to the absence of global synchronization and message time-stamps. As mentioned above, when an existing lightpath is torn-down the ownership of its physical resources is transferred from its source to the intermediate nodes that originally owned them (assumption 2 in section VI.B). Initial runs of the protocol, however, revealed an inherent deadlock under which different nodes updated the same resources differently to the rest of the network, without any guarantee that the correct update would be sent last. Although infrequent, this takes place when an LP is torn-down and, before its deletion is advertised by its source node, a new LP is setup with the following characteristics: a) it is originating from one of the first LP’s intermediate nodes, and b) its first PL is one that the torn down LP just released. Then, the two sources (of the torn down LP and the newly setup LP) update the intersection of their physical resources differently.

To further illustrate this problem, consider the network shown in Figure 3 and assume that $t_0 < t_1 < t_2 < t_3 < t_4$. Suppose there exists an LP1 from 1 to 3 on $\lambda_1$ and that at $t_0$ it starts to be torn down. At $t_1$, node 2 receives the TEAR_DOWN message and releases its OXC switch, which makes its outgoing physical link (from 2 to 3, on $\lambda_1$) available. Then, assume that node 2 receives a request to setup a second LP2 to node 4 and at $t_2$ it completes the setup of the new LP2 (2→3→4) on $\lambda_2$ successfully. At $t_3$, node 2 advertises the existence of LP2 and later, at $t_4$, node 1 advertises the tear down of LP1. It is clear now that the rest of the nodes in the network will be misinformed about the status of link 2→3 on $\lambda_1$ as they will consider it available in their routing procedures.

To address the two-ownership challenge, we impose the following constraint: Unless a node is the original source of a torn-down lightpath (the node that initiates the tear-down action), it may not immediately use its released physical resources that become now available via a TEAR_DOWN message. Instead, it has to wait until the deletion update is received from the source of the torn-down lightpath (just like the rest of the network nodes). In the case that the torn lightpath was never advertised by its source to begin with, the TEAR_DOWN messages raise a special flag bit which entitles the intermediate nodes to consider their physical resources available immediately (see assumption 2 in section VI.B).

**VII. PERFORMANCE EVALUATION**

A custom-built C++ discrete-event simulator is constructed to conduct simulation experiments. The performances of the proposed integrated signaling scheme are evaluated through the simulation of several network topologies that demonstrated similar conclusions. We present results for the 16-node, and 25 bidirectional link NSF16 network representing a backbone US network. The core GigE switches are assumed to have enough interfaces and process all the traffic that can potentially pass through them. This assumption can be relaxed to account for the cases
where GigE switches have limited processing capabilities. All links are assumed to have 2 fibers in each direction, and each fiber is assumed to have 4 wavelengths. The capacity of each wavelength channel is fixed at 2.5 Gbps.

Control messages have a fixed length of 100 bytes and are exchanged over a dedicated control channel of 100Mbps. Call requests arrive according to a Poisson process (1 every 5 minutes) and their durations follow an exponential distribution, which varies to reflect different loads. ESP requests are normally distributed around 400Mbps with a standard deviation of 200Mbps. Message processing and forwarding times are fixed at 1ms and 100µs, respectively. The request processing time at a source is assumed to be 5ms and the optical switch cross-connection execution time is assumed fixed at 15 ms. Results were obtained with the signaling component enabled as well as disabled. The latter implies that no signaling effects were taken into account as message-related time delays were all set to zero.

Figure 4 presents the timer-triggered scheme when nodes update periodically every 1-minute, 5-minutes (equal to the arrival rate) and 1-hour, and compares them with the no signaling case. The results present the effect of contention, which is most severe at low loads and when updates are less frequent. Figure 5 shows the same comparison, but when the network updates are change-triggered (using thresholds at 1, 4 and 10 changes). Again, contention is severely degrading network performance at low loads, whenever updating is not done frequently enough for nodes to obtain an accurate network status.

The worst performance of Figs. 4 and 5 is exhibited by the case when updates are change-triggered with 10 changes gathered in the list. This is because when setting the updating threshold at 10 changes, there is hardly any updates at all (especially at lower loads), as new changes are most likely to cancel previous changes and thus reduce (instead of increase) the changes list size. This is in contrast to the best performance achieved when the nodes update for every single change. In fact, updating every single change performs as good as when the signaling component is disabled.
because nodes are always informed about the most accurate network status.

To verify the contention observation, Figure 6 presents the contention probability of a blocked call (if no blocking exists, the value is zero). It is clearly seen that blocking up to 300 erlangs is solely due to contention under any scheme. The penalty of improved blocking performance in the previous figures is depicted in Figure 7 where the number of exchanged messages to implement the various updating schemes is presented. It is clearly seen that whenever blocking probability is improved, it is because more messages were exchanged and the most recent network status has been conveyed to the network nodes via LSAs.

VIII. CONCLUSIONS

Building on the current round of carrier Ethernet standards, this paper has demonstrated how to support and implement full traffic engineering in a global-scale two-tiered native Ethernet-over-WDM optical networking architecture. To implement this vision, three key innovative components have been presented and devised: 1) a novel converged Layer 2/1 networking model in terms of control plane functionality, 2) a fully distributed integrated routing and signaling framework for dynamically provisioning ESPs at any bandwidth granularity including both full wavelength and finer granularity (sub-lambda) ESPs in an integrated Ethernet-Optical networking environment, and 3) a novel notion of an integrated Link-State Advertisements strategy that is consistent with a fully integrated routing and signaling protocol.

The proposed optical layer-based unified control plane is generic and transparent to the higher layer protocols. The implementation procedures are identical regardless of the kind of PE devices that are attached to the optical cloud including IP/MPLS routers, ATM switches, etc. The proposed model is shown to off load most of the PEs burden to the OXC controllers residing at the optical layer. The ramifications of this are far-reaching and certain to impact the service providers who offer layer-2 and layer-3 MPLS-based VPNs. This is because off loading the burden from the PE devices to the OXC controllers virtually eliminates the typical PE’s major scalability bottleneck, that has placed limits on the upper bounds of sizing L-2 and L-3 VPNs.

REFERENCES

Antonis Hadjiantonis received the BSEE and MSEE from the City College of the City University of New York in 1998 and 2000, respectively. In 2005 and 2006 he earned the M.Phil. and the Ph.D., both in Electrical Engineering from the Graduate School and University Center of the City University of New York.

He is currently a Senior Researcher with SignalGenerix, Inc., and an Assistant Professor with the Department of Sciences and Engineering at Intercollege, both in Cyprus. His research interests include the vertical integration in multi-layer networking environments, routing and signaling algorithms, and first mile access network architectures.

While at CUNY, Dr. Hadjiantonis received the prestigious Caurell Dissertation Fellowship Award for his outstanding research in optical networking.

Mohamed A. Ali received his Ph.D. degree in electrical engineering from the City College of the City University of New York in 1989. He joined the faculty of electrical engineering at the City College of New York in 1989, where he is currently a Professor. He has published over 100 refereed journal papers, invited talks, conference proceedings, and several book chapters. His current research interests include PON-based broadband access networks, WDM-based Optical Networking and architecture, Ethernet Networking Technologies and switching, IP/MPLS routing and signaling, MPLS-based VPNs, Traffic Engineering, Next-Generation data-centric MAN, and WAN Networking Paradigm, and modeling and simulation of high speed networks and systems.

Dr. Ali is a recipient of the NSF CAREER Award.

Haidar Chamas is a distinguished member of technical staff at Verizon Communications working on the security testing and evaluation of Verizon’s broadband and Ethernet services. He has over 13 years of experience in Verizon where he had directed and led the development and roll-out of remote access services (dial-up, DSL, and IP-VPN), IP multimedia services and 3rd generation fast-packet services (FR and ATM).

Previously, he was a director in the Enterprise Solutions Group responsible for the end-to-end implementation of Verizon’s NGEN switched Ethernet services across optical and MPLS networks nationwide. In addition, Haidar is a co-inventor of several patents and had received numerous awards and grants.

William Bjorkman is a Senior Technical Consultant at Verizon Communications working on the deployment of next generation Ethernet services and carrier class Ethernet networks. He has over 20 years of experience in high speed data and digital transmission networks. Previously, Bill was Director of System Technology in Nokia Telecommunications and Director of Broadband Data Services in NYNEX Science & Technology. He is a frequent contributor to Metro Ethernet Forum (MEF) technical specifications, editor of MEF 6 (Ethernet Services Definitions Technical Specification) and co-chair of the MEF Technical Committee. In addition, Bill is a co-inventor of several patents and had received numerous awards.

Stuart Elby is the Vice President of Network Architecture responsible for setting Verizon Telecomm and Verizon Business’ network architecture vision encompassing broadband access, optical transport, fast packet and Ethernet switching, IP/MPLS routing, and emerging voice over IP and video technologies. Stuart is also responsible for defining service specific architectures to support product line marketing, specifying network element requirements, coordinating Verizon’s Standards activities, and leading collaborative R&D activities with universities and industry partners.

Prior to joining Verizon in 1993, Dr. Elby was a Research Associate at NSF’s Center for Telecommunications Research at Columbia University where he was responsible for leading research in optoelectronic devices, all-optical network architectures and developing early WDM / ATM platforms. He was co-director of a multi-university research program on all-optical packet switching networking, and collaborated with Teachers’ College in the development and deployment of a multi-media educational network for primary and secondary schools.

Dr. Elby received a BS degree from the University of Rochester, NY, in 1982 and a MSEE, M.Phil, and Ph.D. from Columbia University in 1989, 1992, and 1994, respectively.

Ahmad Khalil received the BE and MS degrees in Mechanical Engineering from the Lebanese University (1997) and Ohio University (2000) respectively. He also received in May 2005 a Ph.D. degree in Electrical Engineering from the Graduate School and University Center of The City University of New York (CUNY). Currently, Ahmad is an Assistant Professor in the engineering science program at LaGuardia College (CUNY). His current research interests are in the area of data/optical networking, traffic grooming and multicast in optical networks, EPON as well as ad-hoc networks. While at CUNY, Ahmad was awarded the Robert E. Gilleece Fellowship.

Georgios Ellinas holds a B.S., M.S., M.Phil, and a Ph.D. in electrical engineering from Columbia University. George is currently an Assistant Professor of Electrical and Computer Engineering at the University of Cyprus. Prior to joining the University of Cyprus George was an Associate Professor of Electrical Engineering at City College of the City University of New York. Before joining the academia, George was a senior network architect at Tellium Inc (2000-2002) and a senior research scientist in Telcordia Technologies’ (formerly Bellcore) Optical Networking Research Group. He has authored and co-authored more than 80 journal and conference papers, and is also the holder of 29 U.S. and international patents on optical networking.

Nasir Ghan is an Associate Professor in Computer Engineering at Tennessee Tech University, where he is actively involved in a wide range of externally-funded research activities. He has published over 60 refereed journal and conference papers, several book chapters, various standardization contributions, and has two patents granted. Dr. Ghan is an associate editor of the IEEE Communications Letters journal and has also guest-edited special issues of IEEE Network, IEEE Communications Magazine, and Cluster Computing. Dr. Ghan is a recipient of the NSF CAREER Award and is a Senior Member of the IEEE. His current research interests include network virtualization and services, hybrid inter-networking, grid-computing, metro/access networks, and survivability. Dr. Ghan received the BSc from the University of Waterloo, Canada, the MSEE from McMaster University, Canada, and the Ph.D. degree in computer engineering from the University of Waterloo, Canada.