DynaSPOT: Dynamic Services Provisioned Optical Transport Test-bed – Achieving Multi-Rate Multi-Service Dynamic Provisioning using Strongly connected Light-trail (SLiT) Technology

Ashwin Gumaste, Nasir Ghani, Paresh Bafna, Akhil Lodha, Anuj Agrawal, Tamal Das and Si Qing Zheng

Abstract: We report on the DynaSPOT (Dynamic Services Provisioned Optical Transport) test-bed – a next-generation metro ring architecture that facilitates provisioning of emerging services such as Triple Play, Video-on-Demand (VoD), Pseudo Wire Edge to Edge Emulation (PWE3), IPTV and Data Center Storage traffic. The test-bed is based on the recently proposed Strongly connected Light-trail (SLiT) technology that enables the triple features of dynamic provisioning, spatial sub-wavelength grooming and optical multicasting – that are quintessential for provisioning of the aforementioned emerging services. SLiT technology entails the use of a bidirectional optical wavelength bus that is time-shared by nodes through an out-of-band control channel. To do so, the nodes in a SLiT exhibit architectural properties that facilitate bus function. These properties at the network side include ability to support the dual signal flow of drop and continue as well as passive add, while at the client side include the ability to store data in order to support time-shared access. The latter (client side) improvisation is done through a new type of transponder card – called the *trailponder* that provides for storage (electronic) of data and fast transmission (burst-mode) onto the SLiT. Further in order to efficiently provision services over the SLiT, there is a need for an efficient algorithm that facilitates meeting of service requirements. To meet service requirements we propose a dynamic bandwidth allocation algorithm that allocates data time-slots to nodes based on a *valuation* method. The valuation method is principally based on an auctioning scheme whereby nodes send their valuations (bids) and a controller node responds to bids by sending a grant message. The auctioning occurs in the control layer, out-of-band and ahead in time. The novelty of the algorithm is the ability to take into consideration the dual service requirements of bandwidth request as well as delay sensitivity. At the hardware level, implementation is complex – as our trailponders are layer-2 devices that have limited service differentiation capability. Here, we propose a dual VLAN tag and GFP based unique approach that is used for providing service differentiation at layer-2. Another innovation in our test-bed is the ability to support multi-speed traffic. While some nodes function at 1 Gbps, and others function at 2.5 Gbps (using corresponding receivers), a select few nodes can support both 1 Gbps and 2.5 Gbps operation. This novel multi-speed support coalesced with the formerly mentioned multi-service support is a much needed boost for services in the metro networks. We showcase the test-bed and associated results as well as descriptions of hardware subsystems.

I. INTRODUCTION

The shift of revenues from traditional voice services to a multitude of VoIP, *Video-on-Demand* (VoD), *Pseudo Wire Edge-to-Edge Emulation* (PWE3), Triple play and Data-Center storage traffics is a strong motivation for new, low-cost and dynamic optical layer solutions in metro environments. Conventional SONET/SDH hierarchy is now being replaced by more data-centric packet aware technologies like GigE lightpaths and *Resilient Packet Rings* (RPR) with IP/MPLS overlay. GigE is not efficient (on account of its requirement of end-to-end wavelength granularity) nor is it dynamic, while RPR is expensive (due to OE and EO conversions at every node). A new approach is required that provides efficient grooming of sub-wavelength traffic preferably at the optical-layer while enabling necessary dynamic bandwidth provisioning thus facilitating emerging services.

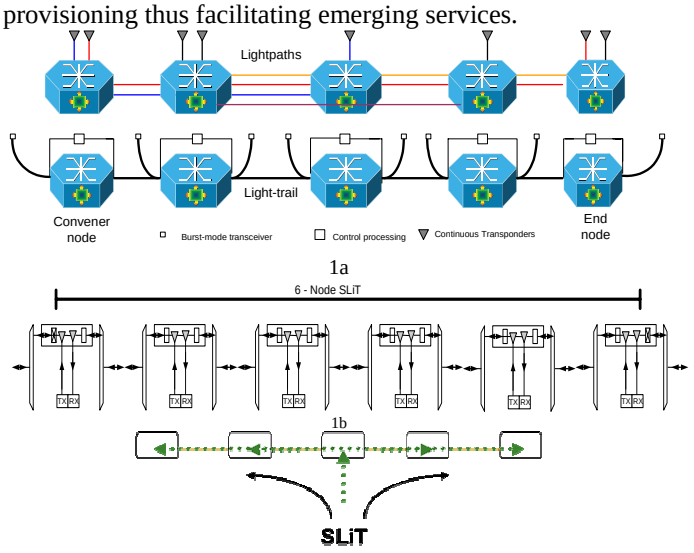

Fig. 1.
(1a) Comparison of lightpaths and light-trails [10].
(1b) A 6 node SLiT.
(1c) A node transmitting to both East and West directions

We report a solution that enables sub-wavelength all-optical (spatial) grooming, multi-service support at multiple line-rates using mature and available technology. The proposed solution is deployed through a metro ring WDM test-bed called DynaSPOT – Dynamic Services Provisioned Optical Transport, and is built on the concept of Strongly connected Light-trail (SLiT) technology [1-3].

A light-trail [4,5] is a unidirectional optical bus that is provisioned through an Out-Of-Band (OOB) control channel facilitating spatial sub-wavelength [5] grooming of traffic along multiple nodes.

A SLiT [1] is a bidirectional implementation of a light-trail with the node architecture first proposed in [1] for implementation in metro rings. SLiT bandwidth is arbitrated by a *controller* and nodes time-share the bandwidth to provision *connections*. Connections are provisioned using burst-mode optics (like in light-trail [5]). In the DynaSPOT test-bed we show how to dynamically provision these connections with a particular emphasis on services. The time-penalty incurred due to queueing of data while facilitating time-sharing is made proportional to the particular service latency type by a new transponder subsystem – the *trailponder*.

Another salient feature of the DynaSPOT test-bed is the ability to use the same SLiT for connections at multiple line-rates, doing so in a dynamic fashion. This feature enables nodes with diverse transponder inventory to use the same SLiT, thus reducing capital expenditure and maximizing efficiency. A final focus of the DynaSPOT test-bed is the ability to provision a multitude of services, both delay sensitive and bandwidth intensive.

The rest of the paper is organized as follows: Section II gives a primer to SLiT technology and builds a platform that entails advantages of this technology in metro networks. Section III describes the DynaSPOT test-bed, while Section IV shows the results obtained from the experiments performed over the DynaSPOT test-bed. Section V summarizes the paper.

II. SLiT TECHNOLOGY PRIMER – ADVANTAGES AND OPEN ISSUES

In this section we present a primer on SLiT technology that serves as a preamble to the rest of the paper in particular to the DynaSPOT test-bed. SLiT technology has emerged from our recently proposed concept of light-trails [4-5]. Light-trails are a generalization of a lightpath (or optical wavelength circuit), such that multiple nodes can take part in communication along the path. A light-trail is analogous to a unidirectional shared wavelength bus. Bandwidth within the bus is shared by arbitration amongst the constituent nodes through an out-of-band control channel. The control channel is optically dropped, electronically processed and then reinserted back into the network at every node-site (see Fig. 1a). Light-trails have been shown to lead to efficient sub-wavelength optical grooming (which we will define later as spatial sub-wavelength grooming), dynamic provisioning (of bandwidth to nodes), facilitate optical layer multicasting and be built using low-cost and mature technology. To support the above features, light-trail nodes have node architectures that have characteristics to support bus functionality on a per-wavelength basis. Shown in Fig. 1a is a light-trail and its comparison to the well established lightpaths (point-to-point to optical circuits). As is seen, a light-trail is a wavelength bus that is regulated between two extreme nodes, called the convener node and the end node, with the direction of the flow from the convener node to the end node. To set up the light-trail the wavelength is blocked between the two extreme nodes, while the intermediate nodes allow for all-optical pass-through as well as support of the bus functionality. Setting up, tearing down and dimensioning (growing the light-trail) is carried about through optical switch (re)configuration. Since, conventional (typically mechanical) and contemporary optical switches are slow (typically requiring several milliseconds) to change state, we assume that formation (set up) tear down and dimensioning of the light-trail is infrequent and is semi-permanent.

Once, a light-trail is set up nodes can communicate to one-another by establishing time-differentiated *connections*. These connections are short duration transmissions of data over the light-trail. Connections are set up and torn down over the light-trail without any optical switching. The only constraint on the connections is that no two connections can coexist over the light-trail at the same time – which if it would happen, would lead to collisions. To avoid collision and guarantee fairness, nodes communicate to each other and synchronize their transmissions (with respect to each other) through the out-of-band control channel.

Over an n-node light-trail, a maximum of ${}^{n}C_{2}$ connections are possible. Provisioning of connections, since it does not require any optical switching is called soft-provisioning, while provisioning of the light-trails due to requirement of optical switching is called hard provisioning.

The feature of unidirectional bus functionality in light-trails leads to optical multicasting, while the characteristic of being able to provision connections over a light-trail without optical switching leads to dynamic provisioning. The dynamic provisioning feature results in a property that we term as *spatial sub-wavelength* support and is now defined: since, multiple nodes time-share the light-trail bandwidth, which in effect means that nodes time-share a wavelength resulting in each node achieving an effective bandwidth that is sub-wavelength; further since these nodes are spatially separated along the light-trail, this leads to our notion of spatial sub-wavelength grooming.

We have in [4, 6] shown the advantages of light-trails over Gigabit Ethernet and Resilient Packet Rings, applied light-trails as an effective candidate technology to Storage Area Networking (SAN) [5] and compared light-trails to optical burst switching [4].

Despite the above mentioned advantages showcased by light-trails, there are a few drawbacks that affect performance as well as increase cost. A primary drawback of light-trail technology is that there is uneven per-span utilization in the unidirectional bus. As can be seen from Fig. 1, the convener node in the light-trail can set up *n-1* connections (to *n-1* prospective downstream destination nodes), while the second node can set up *n-2* connections, and so on. This implies that the span between nodes N_2 and N_3 would result in higher utilization than the span between nodes N_1 and N_2 and so on. This unbalanced utilization implies wastage of bandwidth.

A second disadvantage of light-trails is that they are unidirectional and hence they do not naturally support duplex communication. The unbalanced utilization of spans coupled

with the asscoiated delay jitter also has led to unfair allocation of bandwidth to nodes and several schemes [7-9] have been proposed for bandwidth allocation and fairness within light-trails.

To alleviate the above mentioned problems while maintaining the advantages of light-trails we in [1] extended the basic unidirectional light-trail to the bidirectional *Strongly connected Light-Trail* or SLiT. A SLiT is a bidirectional version of a light-trail that allows communication between nodes over a single wavelength in duplex fashion. Hence an *N*-node SLiT is able to support a maximum of N^2 connections (i.e. N^2 source-destination pairs), with the constraint that no two connections can co-exist at the same time.

To support duplex communication over a single wavelength, we have proposed in [2] a unique node architecture that facilitates bidirectional support. In addition in [2] we have proposed a new protocol that guarantees fairness and leads to efficient bandwidth utilization within the SLiT. This protocol, which we have implemented in our DynaSPOT test-bed below and which we will describe in detail, takes into consideration service requirements of delay and bandwidth hence facilitating delay sensitive and bandwidth intensive services to be provisioned over the shared wavelength SLiT.

Conceptually an East-West SLiT is shown in Fig. 1b, where nodes communicate to other nodes in both directions. Also shown in Fig. 1c is an example of a node transmitting in both Eastward as well as Westward direction. Using the passive optical bus, a node can also receive data from other nodes that are either Eastwards or Westwards of itself.

To support the SLiT, a node has an architecture that is shown in Fig. 2. Composite WDM signal from the network enters the node at either of the two Arrayed Waveguides (AWGs) that can act as both multiplexer and de-multiplexer depending on whether the signal is entering the node or leaving it. The composite WDM signal is de-multiplexed into its constituent wavelengths. Each wavelength is fed to a SLiT Optical Retrieval Section (SORS) that allows the node access to the SLiT signal. The SORS consists of two ON/OFF optical shutters (slow moving optical switches) which are separated by two passive couplers (both in 2x1configuration). The two couplers are 3dB (50/50) type, i.e. power at any input port is split in half to the other two output ports.

One of the two couplers is called Drop Coupler (DC) and the other coupler is called Add Coupler (AC). Signal that enters the node from either direction (West/East) is dropped for local processing at the drop coupler. The DC also forwards a copy of the signal (using optical splitting property) to nodes further downstream in the SLiT, thus resulting in *drop and continue* operation. The second coupler (AC) allows the node to send in signal into the SLiT. The AC does so using passive properties, i.e. allows addition of signal without any switching operation and this is called as *passive add* operation. Signal sent into the AC is split into two copies, one sent into the Eastward direction and the other sent into the Westward direction (as shown in Fig. 1c).

Since nodes time-share the SLiT bandwidth, it implies that when a node establishes a connection (over the SLiT), other nodes have to queue their data and wait for transmission rights over the SLiT. A new sub-system is required that enables nodes to queue data and then efficiently transmit (and likewise receive) whenever the protocol allows the node rights to form a connection. We propose a subsystem called a *trailponder* that stores data in a format that plausibly supports layer-2 data without affecting layer-2 functionality. The trailponder is similar to a transponder in the function that it transmits and receives signals between the network side (SLiT) and the client-side. However, the trailponder has the additional onus of storing data in a format that allows effective communication (explained in the next section).

A second function of the trailponder is to efficiently utilize the SLiT bandwidth. To do so, the trailponder must be able to send data (with minimal delay), once the node is given rights to form a connection. This means that the transmitter (laser) must be able to switch "ON" in a very short time interval. Conversely, the receiver in the trailponder must also be able to "latch-on" to an incoming signal with minimal preprocessing or training bits. To provide this kind of fast ON/OFF and rapid reception capability, the trailponder is designed using burst-mode optics [11] as shown in the architecture of Fig 2.

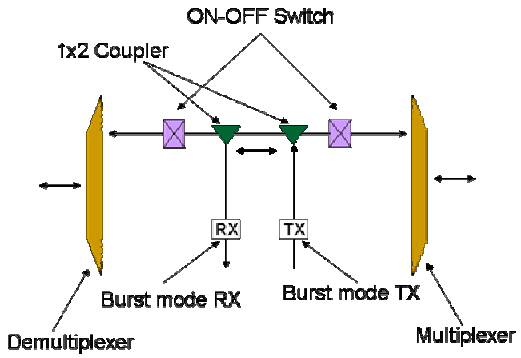

Fig. 2. SLiT node Architecture

Control Protocol and timing issues:

The control protocol for bandwidth arbitration in the SLiT is designed based on an auctioning algorithm. The SLiT bandwidth is assumed to be the object that is being bid for by multiple nodes in the SLiT (that act as bidders). The data channel (SLiT) is assumed to be time-slotted with slots of large duration (typically 0.3-5ms). It is not necessary to synchronize all the nodes with respect to each other. Slot boundaries are assumed to be only loosely synchronized. Loose synchronization is achieved by the out-of-band control channel that is optically dropped and electronically processed at every node before being inserted back into the network. Due to this OEO function on the control channel, it is fair to assume (and hence implement) the control channel to be tightly synchronized with respect to all the nodes (sharing common clock). Since the control card at every node is responsible for transmission of data (and hence formation of a

connection) over the SLiT, it is assumed that the trailponder, which actually sends and receives data is also pseudo-synchronized. The data time-slots are selected large enough so that the auction algorithm can result in a convergent bidding process. In each SLiT, a particular node is selected as the *controller* node. This node acts as an arbiter for bids in the SLiT. The bid is sent as a *valuation* that is a single numerical quantity representing both the bandwidth intensity and delay sensitivity of the data that is stored at the node (in its trailponder). In every data time-slot, each node sends a valuation to the controller node through the control channel. The controller receives valuations from every node in the SLiT and then selects the node that sent the highest valuation. It then sends a *grant* message to this node (that sent the highest valuation). This node then sets up a connection (over the SLiT) in the *next* data time-slot. In this way bandwidth is dynamically allocated to nodes in a SLiT. The process of computing valuations is explained in the next section.

The entire procedure of sending valuations by the nodes to the controller, computing the node with the highest valuation and signaling back to the successful bidding node, is done *ahead-in-time* and through the out-of-band control channel. We have in another article [16] shown that the procedure of computing valuations leads to proportional fairness [12].

III. DESCRIPTION OF THE DYNASPOT (DYNAMIC SERVICES PROVISIONED OPTICAL TRANSPORT) TEST-BED

The DynaSPOT test-bed is a 4-node open-optical metro-ring that can support single wavelength communication at both 1 Gbps and 2.5 Gbps. It is in the future expandable to support WDM communication at 100 GHz channel spacing and bit-rates up to 10 Gbps. The test-bed is built to support metro applications and makes use of WDM optics at 100 GHz spacing and a channel count of 40. SLiT technology forms the corner-stone of the test-bed.

In the present version of the test-bed the SLiT is statically set up while connections over the SLiT are dynamically set up based on the valuation protocol that we described previously.

Our objective of the test-bed is (1) to demonstrate dynamic bandwidth allocation within a SLiT, (2) allocation of sub-wavelength flows to each node in the SLiT, (3) achieving multi-rate communication within the same SLiT (using different *type* of trailponder cards, i.e. at different bit-rates and (4) provisioning metro services such as VoIP, Triple Play, VoD, PWE3 and data centers.

The following design choices are involved in the DynaSPOT test-bed:

SLiT: A four-node SLiT is created to facilitate sub-wavelength dynamic service provisioning. A node can communicate to any of the other 3 nodes using SLiT principles of communication (all-optical, time-sharing of bandwidth) mentioned in Section II. An out-of-band control channel is used for arbitration. At each node we have designed and implemented a trailponder with support one or more plausible bit-rates. Each trailponder is provided with a PowerPC embedded in an FPGA (*Xilinx Virtex 2Pro*) for arbitration and control purposes. Nodes can establish connections with other nodes that have compatible receivers – i.e. a source node can communicate with destination node(s) under the condition that the destination node has a receiver at the same line-rate as the source-node. In this way multiple nodes can time-share the SLiT at different line-rates.

For example, in the 4-node SLiT, nodes N_1 and N_3 communicate at 2.5 Gbps while nodes N_2 and N_4 communicate at 1 Gbps. Further, node N_1 can receive information at both 2.5 Gbps as well as 1 Gbps.

The transmitter, receiver and traffic profile that is provisioned at the nodes is shown in Table 1. The SLiT is assumed to be time-slotted with data time-slots of 400 *us* duration separated by guard-bands of 10 *us*. It is possible to change the duration of the time-slots in the range of 300 *us* to 5 ms depending on specific user requirement.

A general guideline for slot duration selection is that larger the time-slots greater the average delay, and an overall betterment of efficiency (at higher loads). Our choice of smaller data time-slots (of 400 *us*) is based on the logic that we desire to support delay sensitive services and also the fact that at low to medium loads the efficiency of the system (construed as average utilization) is comparable to the efficiency obtained when the time-slots are of larger duration. This is because, with large time-slots, at low to medium loads the nodes do not have enough data in the buffer (within the trailponders) to occupy entire slot width.

Controller

	N_1	N_2	N_3	N_4
Voice	x	x	x	
Video	x		x	x
Data	x	x	x	x
Storage/DCN	x		x	

1GbTX		x		x
1GbRX	x	x		x
2.5GbTX	x		x	
2.5GbRX	x		x	

TABLE 1. Node configuration and Service provisioning in the DynaSPOT test-bed

Node architecture: The node architecture used in DynaSPOT is shown in Fig. 3 and the critical subsystem used for provisioning traffic – the *trailponder* is shown in Fig. 4. The conceptual layout of the test-bed is shown in Fig. 5 while the photographs of the test-bed are shown in Fig. 6 and Fig. 7. Specifically shown in Fig. 6 is a single node trial version of the DynaSPOT test-bed.

The node architecture has evolved from a Reconfigurable Optical Add-Drop Multiplexer (ROADM), with incoming WDM signal de-multiplexed by an AWG (Arrayed Wave Guide). Our AWGs have a special slot to de-multiplex and multiplex signal at 1550 nm (non-ITU wavelength) to support off-the-shelf PON optics which are

presently used in the test-bed. Each constituent wavelength (in the C-band) is fed to a SLiT Optical Retrieval Section that consists of a series of two ON/OFF (optical) switches (with 1 dB insertion loss and 5 ms switching time) and two passive optical couplers (3dB) in 2x1 configuration.

Incoming signal (in either direction) is dropped and continued at the drop-coupler while local signal can be added into the SLiT passively through the add-coupler. The two couplers have 50/50 splitting/combining ratio. The two optical switches are in the ON state at all intermediate nodes in the SLiT. At the extreme nodes the switches on the outer end are in the OFF state, i.e. for the East-most node the East-most switch is in the OFF state, while for the West-most node the West-most switch is in the OFF state. This implies that a node cannot be part of two SLiTs on the same wavelength even if the two SLiTs are graphically non-overlapping. This feature is another major differentiator from light-trails [13]. Two SLiTs on the same wavelength can coexist if and only if they do not have any common nodes between them, while in light-trails; two light-trails can coexist if they have a single common node between them.

In addition, a single Variable Optical Attenuator (VOA) is used to stabilize optical power-level in the SORS. The VOA enables features of flattening gain tilt as a result of skewed bidirectional amplification. The VOA is connected to the arbiter. The arbiter runs a simple gain-control algorithm to stabilize signal power.

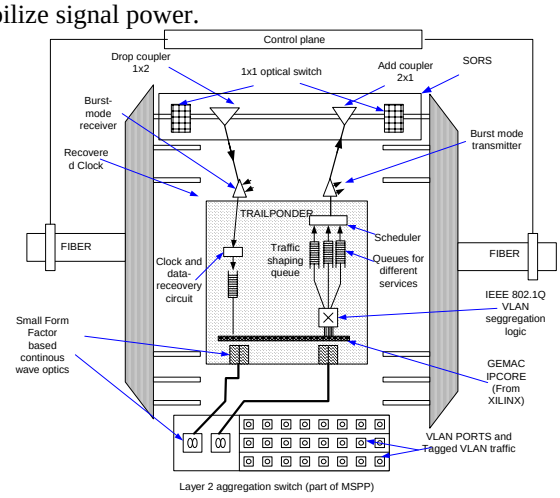

Fig. 3. Node architecture in the DynaSPOT test-bed.

To add and drop signal, to and from the time-shared SLiT as well as to facilitate service provisioning, we use the trailponder. The trailponder (as shown in Fig. 4) consists of burst-mode optics (laser and receiver) [1, 2] and associated electronics (memory, processor) and is triggered through a control card. The trailponder is analogous to a transponder – it facilitates client signals to be transmitted (and received) over the SLiT. The OE (and EO) trailponder card has the added function of storing data as well as scheduling stored data in an optimal manner to facilitate service provisioning. The trailponder stores data in a way that the client side layer-2 equipment (typically an aggregation switch) is oblivious to the storage and scheduling of data over the shared medium SLiT. To do so, it has to manipulate the layer-2 protocol which will be described subsequently.

Apart from providing access to the SLiT, the trailponder is a key device that enables services to be inculcated into our valuation based bandwidth provisioning algorithm. Flows corresponding to different services are fed into the trailponder as shown in Fig. 4. The trailponder then differentiates services based on the following technique. It uses either a VLAN based differentiator that segregates incoming packets (from a client) based on VLAN tags (explained later), or uses the *type* field for differentiation if the transmission format is based on Generic Framing Procedure (GFP), ITU G.7041. Packets based on service types are stored in corresponding *service buffers*. Whenever a node is granted a data time-slot for transmission, the data is then sent into the SLiT. To do so, the trailponder uses a *TX_EN* signal that enables the burst-mode laser. Once the laser is enabled, the trailponder maps the stored data into Ethernet frames or GFP payload. To do so, it selects a corresponding MAC address using the GEMAC IP CORE from Xilinx. The data from the buffers is now perfectly aligned into an Ethernet frame. The multiple service buffers are emptied in the following priority order: storage→voice→video→data. The total memory allocated for buffering in a trailponder is 2Mb (1Mb for TX, 1Mb for RX). Further, there are 4 service buffers of size 200, 200, 200 and 400 kb. The 400 kb buffer is used for data (due to it being bandwidth intensive) while the remaining three buffers are used for voice, video and storage/data center/pseudo-wire traffic. The 1Mb buffer at receiver side is used for traffic engineering (a future function that is not yet implemented in the DynaSPOT test-bed). s

Control card:

This is built using a Virtex2 FPGA board (XUP V2P) and is connected to all the nodes through a 10/100 Ethernet switch. For sake of simplicity it is assumed that the control channel is non-blocking, implying that when two nodes send requesting signals on the control channel the two signals (in form of Ethernet packets) do not collide. In practice, the control channel would be a dedicated wavelength (at 1510 nm) which would be time-slotted with miniature control slots (compared to the data time-slots). However, to maintain simplicity and due to collocation of the nodes (in the test-bed), we simply use Fast Ethernet based control channel that is connected to a single control card. This single control card acts as an arbiter and is provisioned at node N_3.

Connection provisioning:

This sub-section discusses how connections are provisioned within the SLiT. At the beginning of each data time-slot, every node sends a request to the arbiter node (node N_3 in our case) to form a connection. The request is sent as a utility *valuation*. The valuation is computed by the trailponder based on a method that is described later. The trailponder computes valuation (in every data time-slot) and sends this to the arbiter

through a trailponder Fast Ethernet interface. To do so, the computed valuation (a numerical quantity in [0,1]) is mapped to an Ethernet frame. This frame is sent to the arbiter using a Windowed Automatic Repeat Request (ARQ) protocol: as part of this protocol, the trailponder sends its valuation to the arbiter. If the arbiter receives the valuation correctly, (without any transmission-line errors that are detected based on single bit parity); then the arbiter sends the same valuation back to the trailponder in another Ethernet frame. If the returned valuation is within a certain time-window since the trailponder first sent data, then the trailponder knows that the arbiter has correctly received its request for bandwidth in the next time-slot (connection formation). If however, the initiating trailponder does not receive the acknowledgement frame within the specified window, it then resends the valuation. Typically the window period is set to half data time-slot length. Upon receiving the valuations from all the trailponders the arbiter decides which node would transmit in the next slot based on the highest utility valuation. This node is *granted* permission to set up a connection (of max duration 400 *us*). Between two data time-slots there is a guard-band of 10 *us*. This is necessary to *reset* the burst-mode receiver logic bias.

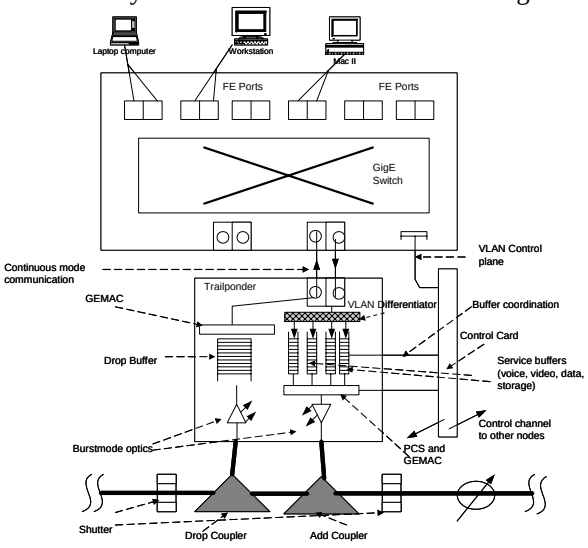

Fig. 4. Trailponder (for VLAN support) and control card

VLAN/GFP type based service differentiation: As shown in Fig. 4 traffic from the client-side arrives into the trailponder in a continuous-mode fashion (typically through 10/100/1000 Mbps interfaces). This traffic is aggregated by a layer-2 switch and then sent to the trailponder. The trailponder receives layer-2 frames (Ethernet) with VLAN (IEEE 802.1Q) tags and either sends them into the network as Ethernet frames (at 1 Gbps), or maps these into GFP frames (at 2.5 Gbps) depending on the line-rate of the burst-mode transmitter. While transmitting data, the trailponder preserves the mapping between ingress VLAN tag and egress GFP-type field. These frames (Ethernet/GFP) are queued in service buffers. For the former case of end-to-end Ethernet traffic, the frames enter the trailponder FPGA through the Xilinx GEMAC that is engineered to capture VLAN tags as well as to perform PCS (Physical Coding Sub-Layer) functions. As part of PCS function, the header of the incoming Ethernet frame is stripped off. The header is separately stored on an on-board FPGA cache memory, while the data part is stored into a FIFO that is created in an SDRAM that is available on the Xilinx Virtex 2pro 1152 board. Interconnection between the SDRAM and the FPGA is managed by one of the two on-board (FPGA) PowerPCs, and the actual interface is operated by a memory access module embedded in the FPGA.

Since VLAN tags have 3-bits allocated for service type, the FPGA is able to store the incoming data (within the Ethernet frame) into one of the 4 appropriate buffers (as shown in Fig. 3 and Fig. 4). The information is stored into a buffer depending on the match of the buffer id (000=voice, 001=video, 011=storage/datacenter/pseudo-wire and all other=data) to the corresponding VLAN tag or GFP-type.

Utility Valuation and Provisioning:

In this sub-section we discuss how valuations are computed. Valuation is a quantitative measure reflecting how much a node requires the next data time-slot in order to meet both its bandwidth and delay needs. The challenge in computing valuation is that bandwidth intensity and delay sensitivity are difficult to normalize with respect to each other. What we are interested in is the net utility that the network would get if a node is allowed to provision bandwidth while meeting the node's requirement. From the work of Shenker [14] we know that from a utility standpoint, a sigmoidal like utility function best describes the time-variant need of real-time services. This implies that a function that has a sigmoidal-like curvature as a function of time is ideal for real-time services like voice, video etc. Likewise Lee, Shroff and Mazumdar [15] have shown that concave utility functions serve as optimal allocation strategy for bursty data traffic.

Hence the valuation that we compute is based on both utilities – bandwidth as well as delay and is now shown: For every buffer, the trailponder computes a value called *time-to-service*, defined as the time remaining before which a buffer must be scheduled (into the SLiT) or else the longest waiting packet (of that service) would be timed-out (service latency not met). If $a_i^j(t)$ is the time-to-service for the j^{th} buffer at node N_i at time *t*, then $b_i(t) = \min_j \left(a_i^j(t) \right)$ is called the delay criticality [10] and represents the minimum time before which the node must be serviced, or else packets from at least one buffer would be timed out.

The trailponder also computes the buffer activity period as follows: Let $c_i^j(t)$ be the time elapsed since the first packet entered buffer *j* at node N_i at time *t*; then, $d_i(t) = \max_j \left(c_i^j(t) \right)$ represents the activity period of the trailponder at node N_i. The trailponder can now compute service valuation (resulting from the delay sensitive services) as:

$$s_i(t) = \frac{b_t(t)}{1 + d_i(t)} \qquad (1)$$

In [16] we have shown that probability distribution function of the above can be reduced to a sigmoidal like function.

Similarly the trailponder also computes a value of *buffer utilization*, defined as the ratio of the number of bits in the buffer to the total buffer capacity. Hence, if $y_i(t)$ is the total number of bits in all the buffers at node N_i at time *t*, and *Y* is the total size (max capacity) of the buffers combined, then buffer utilization (valuation) is computed as:

$$z_i(t) = \frac{y_i(t)}{Y} \qquad (2)$$

Again, as shown in [16] the above function ($z_i(t)$) for Poisson and Pareto arrivals has a concave distribution.

Since both service valuation and buffer utilizations are entities yielding ratios in the range [0,1] the trailponder passes on the maximum of the two ratios to the control card. Hence the valuation that a node sends is given by:

$$val_i(t) = \max[z_i(t), s_i(t)] \qquad (3)$$

Each control card then sends this valuation to the arbiter. The valuation is analogous to the utility that the node has for the bandwidth (in the next data time-slot) and hence we also call the valuation as a utility valuation. The control card also sends a *destination* list to the arbiter – that consists of all possible destination MAC addresses of packets stored in the buffer. The multiplicity in destinations facilitates optical multicasting feature in the SLiT bus and also supports multi-rate/speed communication.

Multi-Rate/Speed support:

When a connection is provisioned at a certain line-rate the source and destination node are assumed to have the requisite laser/receiver at that line rate. It may however happen that a node has 2 receivers at different line rates (1 Gbps and 2.5 Gbps) connected passively (in drop and continue fashion). In such a case, the node must switch OFF the receiver that is not in sync with the line-rate of the connection. This is done ahead in time through the OOB control packet (*grant*) that is sent by the arbiter node and triggers the corresponding receiver bias OFF. This works as follows: if N_1 is the source node and N_2 is the destination node, then and if N_2 has receivers for both 1 Gbps and 2.5 Gbps while N_1 has a transmitter at 1Gbps, then N_2 has to switch OFF its receiver at 2.5 Gbps. When N_1 gets a grant message from the arbiter to form a connection, the arbiter also tells N_2 (through an Ethernet frame in the control channel) that N_1 would be transmitting at 1Gbps line-rate. N_2 then switches OFF its receiver at 2.5 Gbps. The assumption here is that the arbiter has global knowledge about which node has what line-rate capabilities. The assumption is valid since the arbiter acts as a central point of intelligence connected to all the nodes through the control channel.

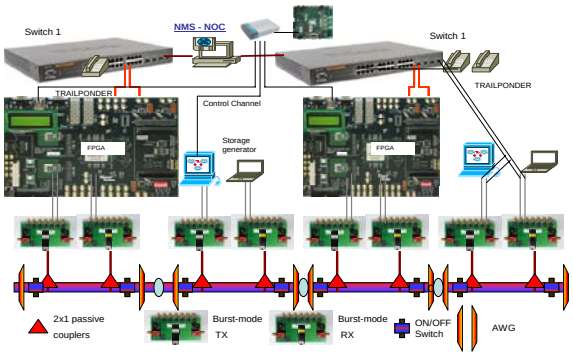

Fig. 5. Conceptual layout of the test-bed.

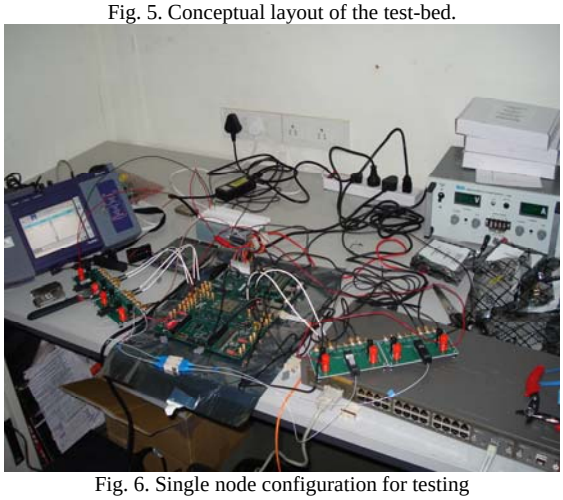
Fig. 6. Single node configuration for testing

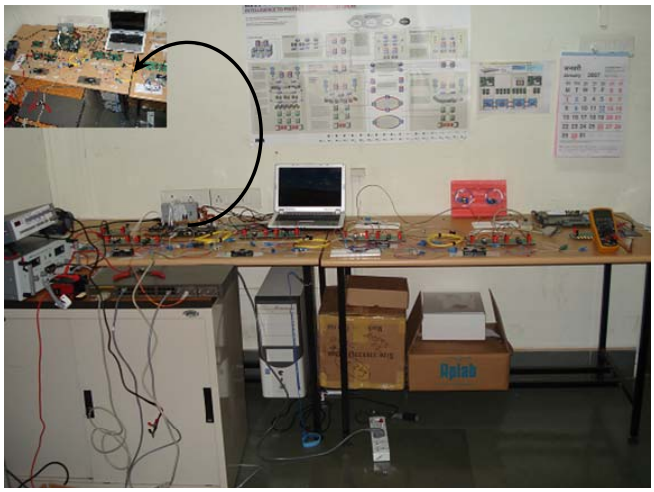
Fig. 7. 4-node DynaSPOT test-bed using SLiT technology.

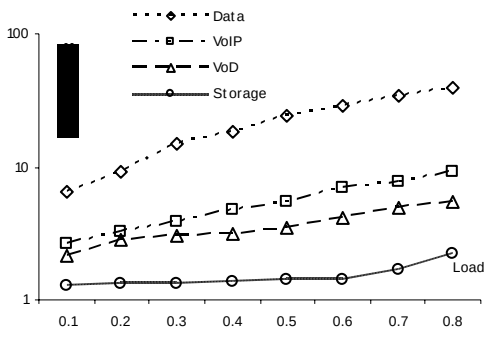

Fig. 8. Delay as a function of load

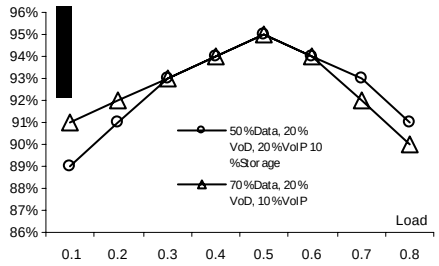

Fig. 9. Efficiency of the system.

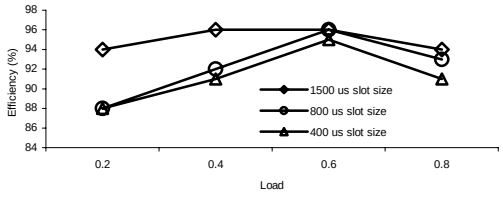

Fig. 10. Efficiency within the SLiT for different data slot-sizes.

IV. EXPERIMENT AND RESULTS

We performed an experiment with a 4-node SLiT at 1550 nm in the DynaSPOT test-bed. The SLiT supported line-rates of 1 Gbps (GigE) as well as 2.5 Gbps (GFP). The setup assumed four traffic types –VoIP, VoD, data and storage/data-center/pseudo wire; feeding into an aggregation switch at all the nodes. All incoming (client-side) Ethernet traffic was attached with VLAN tags while GFP traffic was differentiated based on the type field. Traffic was varied by a generator through a layer-2 aggregation switch as well as a Gigabit Ethernet VLAN tester that generated tagged VLAN based packets. 2 nodes were connected to the aggregation switch while 2 other nodes were connected to the Gigabit Ethernet VLAN tester. Traffic intensity (load) could be varied at both generators. Measurements were performed by increasing traffic from 100 Mbps to 800 Mbps for GigE and 200 Mbps to 2 Gbps for GFP.

Shown in Fig. 8 is the average end-to-end delay profile for data, VoIP, VoD and storage traffic as a function of load. The profile of traffic is shown in Table 1. VoIP traffic was generated as part of the VLAN tester as well as by emulating a point-to-point skype connection. The VoD model is representative of a Video Hub Office (VHO) and several Video Serving Offices (VSOs). Video on demand traffic was emulated by connecting one node to a video server (through the Ethernet switch) while the other nodes acted as recipients of video traffic. In Fig. 8 we observe that the average end-to-end delay is well within the acceptable service latency requirements even when the SLiT is heavily loaded with duplex VoIP traffic and sensitive storage traffic.

For storage/data-center traffic we assume a largely dynamic and extremely delay sensitive traffic characteristic (delay tolerance <10 ms). To emulate storage traffic we create end-to-end pseudo wires that connect two hard drives. The pseudo wire traffic is then mapped into Ethernet frames based on RFC 3985 PWE3 (Pseudo Wire Edge to Edge Emulation). Dynamism is brought about in the network by a C# applet that controls each hard disk and that requests for data transfer from one hard disk to another (over the SLiT). The rate of requests and amount of data transfer can all be varied to observe performance.

Shown in Fig. 9 is the efficiency of the system with two different traffic mixes, one with 10 % storage traffic and the other with no storage traffic. In Fig. 9 we observe that with more data traffic (and less dynamic – storage traffic), efficiency is better at lower and medium loads but degrades at heavy loads due to data-burstiness. Not shown in the figure but also observed is that the fall in efficiency is primarily because data time-slots are scantily utilized for highly delay sensitive storage traffic.

Shown in Fig. 10 is the effect of slot size on efficiency. It is seen that larger the slot-size better the efficiency. The complete picture, i.e. the effect on delay is shown in Fig 11. Shown in Fig. 11 is the effect of slot-size increase on efficiency and the corresponding increase in delay. We here define a parameter *p* that is the ratio of increase in efficiency to the corresponding increase in delay, normalized over the corresponding increase in slot-size. Hence, *p* would be unity if a 10 % increase in slot-size results in a 10 % increase in efficiency and a 10 % increase in delay. As can be seen, *p* is not linear for increase in slot-size (starting from 400 *us*). This shows that by using larger slot-sizes we can achieve better efficiency but we have severe penalty in terms of delay.

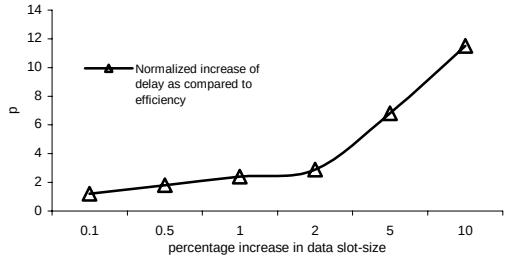

Fig. 11. Normalized increase of delay as a function of efficiency for increase in data time-slot size.

V. CONCLUSION

We demonstrate the DynaSPOT test-bed to support sub-wavelength grooming, dynamic service provisioning (VoIP, data, VoD and storage) and multi-rate communication (1 Gbps and 2.5 Gbps) over a single wavelength using SLiT technology. Services such as VoIP, data, video on demand, storage (data-center) are provisioned over our test-bed using a dynamic bandwidth allocation protocol. An architectural overview and experimental results are presented. Dynamic bandwidth allocation is based on a novel auctioning algorithm that computes bids as valuations reflecting bandwidth and delay needs of the services is shown and implemented in the test-bed. Compliance to high efficiency and delay requirements of the provisioned services are shown. In summary, we demonstrate different features of SLiT technology showcasing it as an enabler for next generation metro applications, by providing the required features for service provisioning and using a low-cost and evolutionary (from ROADM) set up.

Acknowledgement: Authors thank TCS-TRDDC for financial support, R. Nachane (JDS Uniphase), and R. Subramaniam, N. Varma (Xilinx/Avnet) for their equipment support , and Prof. Krithi Ramamritham (IIT Bombay) for his encouragement and Biswanath Mukherjee (UC-Davis) for some of his inputs on our test-bed

References

[1] A. Gumaste, S. Jain and S. Q. Zheng, "*SLiT: Strongly connected Light-trail solution for Cost Efficient and Dynamic Optical Networking*," 22nd IEEE/OSA Optical Fiber Communications Conference OFC 2006 Anahiem CA

[2] A. Gumaste, N. Ghani, P. Bafna, A. Lodha, S. Srivastava, T. Das and S. Zheng, "Achieving Multi-Rate Dynamic Sub-Wavelength Service Provisioning in Strongly connected Light-trails (SLiTs)," POST DEADLINE PAPER *IEEE/OSA OFC Optic Fiber Conference*, Anaheim, CA March 2007

[3] P. Bafna, A. Gumaste and N. Ghani, "Delay Sensitive Smoothed Round Robin Scheduler (DS2R2) for Light-trail and SLiT Networks," *IEEE/OSA OFC 2007, Anaheim CA TuG.*

[4] A. Gumaste and I. Chlamtac, "Light-trails: An Optical Solution for IP Transport, "*OSA Journal on Optical Networking* May 2004, 864-891.

[5] A. Gumaste and S. Zheng, "Next Generation Optical Storage Area Networks: The Light-trails Approach," *IEEE Communications Magazine* Mar. 2005 Vol. 21. No. 3. pp. 72-79 and references therein

[6] A. Gumaste and S. Q. Zheng, "Optical Implementation of Resilient Packet Rings using Light-trails," *21st Prof. of Optical Fiber Conference/National Fiber Optic Engineers Conf. NFOEC/OFC* 2005, CA

[7] A. Gumaste and S. Zheng, "Dual Auction (and Recourse) Opportunistic Protocol for Light-trail Network Design," *IEEE Wireless and Opt. Commun. Conf.(WOCN)* Bangalore, India 2006.

[8] A. Gumaste and P. Palacharla, "Heuristic and Optimal Assignment Techniques for Light-trail Ring WDM Networks," *Elsevier's Computer Communications Journal (CCJ)*, March 2007, pp 21-32

[9] W. Zhang, G. Xue and K. Thulasiraman**, "**Dynamic light trail routing and protection issues in WDM optical networks," IEEE Globecom 2005 Dec 2005

[10] A. Gumaste et al, "On Control Channel for Service Provisioning in Light-trail WDM Networks," 43rd *IEEE Intl Conf. on Commun. ICC 2007* June 2007, *Glasgow UK*.

[11] Y. Ota and R. Swartz, *IEEE Journ. of Lightwave Tech. 1990 Vol. 8 No 12.*

[12] F. Kelly, "Models for a Self Managed Internet," *University of Cambridge UK*, online version.

[13] A. Gumaste and I. Chlamtac, "Mesh Implementation of Light-trails: A Solution to IP Centric Communication in the Optical Domain," 13th IEEE Intl Conf on Computer Communications and Networks ICCCN, Dallas TX Oct 2003.

[14] S. Shenker, "Fundamental Design Issues for Future Internet," *IEEE Journ. Of Select Areas in Commun. JSAC, Vol. 13, No 7*, Dec 1995. pp 1176-1185

[15] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, ``Non-convex Optimization and Rate Control for Multi-class Services in the Internet," *IEEE/ACM Trans. on Networking*, vol. 13, no. 4, Aug. 2005 pp. 827 – 840

[16] A. Gumaste et al, "Two Stage Auction Algorithm for Fair Bandwidth Allocation and Topology Growth in Light-trail based WDM Networks," submission to *IEEE JSAC 2007*