

Distributed Grooming in Multi-Domain IP/MPLS-DWDM Networks

Q. Liu¹, T. Frangieh², F. Xu¹, C. Xie¹, N. Ghani¹, T. Lehman³, Chin Guok⁴

¹University of New Mexico, ²Virginia Tech, ³University of Southern California, ⁴Lawrence Berkeley National Lab

Abstract: This paper studies distributed multi-domain, multi-layer provisioning (grooming) in IP/MPLS-DWDM networks. Although many multi-domain studies have emerged over the years, these have primarily considered “homogeneous” network layers. Meanwhile, most grooming studies have assumed idealized settings with “global” link state across all layers. Hence there is a critical need to develop practical distributed grooming schemes for real-world networks consisting of multiple domains and technology layers. Along these lines, a detailed hierarchical framework is proposed to implement inter-layer routing, distributed grooming, and setup signaling. The performance of this solution is analyzed in detail using simulation studies and future work directions are also high-lighted.

Keywords: Multi-domain grooming, multi-domain IP-DWDM, inter-domain routing

I. INTRODUCTION

Modern backbone infrastructures have evolved to support a broad range of networking technologies [1]. From a functional view, these networks are commonly segmented into multiple domains running at different technology layers/link granularities. Namely, domains are commonly delineated using geographic, administrative, or cost boundaries. For example, backbone cores have mostly migrated to scalable *dense wavelength division multiplexing* (DWDM) transport using *optical cross-connect* (OXC) and *reconfigurable add-drop multiplexer* (ROADM) devices [1]. These networks can provision “lightpath” circuits, yielding multi-gigabit capacities between metro/regional domains. In turn, metro/regional networks generally operate at finer traffic granularities and use electronic switching such as *IP multi-protocol label switching* (IP/MPLS), *time division multiplexing* (TDM) SONET/SDH, or new “carrier-grade” Ethernet.

In concert with the above, growth in client-layer application continues to drive up bandwidth demand. As a result, there is a pressing need for provisioning connections across multiple *heterogeneous* domains/layers. For example, many commercial *Ethernet private line* (EPL) and *virtual LAN* (VLAN) services require transport across underlying SONET and/or DWDM infrastructures [2]. Similarly, “e-science” applications are demanding largescale connectivity across multi-domain *research and education* (RE) infrastructures for peta-exabyte file transfers [3]. As such, these trends are pushing an evolution towards automated *multi-domain, multi-layer* provisioning, also termed as *vertical-horizontal* integration [1]. However, even though some generic standards are taking shape here [1], related algorithm design/performance evaluation studies are largely lacking.

To date, most work on multi-domain networks has focused on single network layer, e.g., IP/MPLS, DWDM, etc [4]-[17]. The key concepts explored here include topology abstraction,

distributed path computation, and crankback signaling. However the extension of these schemes to multi-layer settings, e.g., in which domains operate at different link granularities, is not entirely straightforward owing to the inherent grooming dimension. Now even though a wide range of grooming studies have been conducted for IP-DWDM, SONET-DWDM, and IP-SONET networks, realistic multi-domain settings have not been addressed. Instead, most grooming studies have assumed a single underlying DWDM topology (domain) along with complete state knowledge (topology, resources) across all higher layers, see [18]. Indeed, it is not realistic for a single entity to maintain global state in distributed multi-carrier settings, owing to obvious scalability and confidentiality concerns.

In light of the above, novel multi-domain/multi-layer grooming solutions must be developed to route “sub-rate” IP/MPLS flows over coarse DWDM network domains. These schemes must operate in distributed settings with limited global state and scale to achieve a level of “optimality”, i.e., where optimality can be inferred as the path chosen in the idealized case of a “flat” network (with no domain-level partitioning and global state [1]). Typically, these objectives can be conflicting. To address this challenge, the work herein extends upon the multi-domain DWDM solutions developed by the authors in [12]. Specifically, novel algorithms are introduced for topology abstraction, resource dissemination, *traffic engineering* (TE) path computation/grooming, and setup signaling. The paper is organized as follows. Section II reviews existing work on multi-domain network provisioning and subsequently Section III details the proposed distributed grooming schemes. Performance evaluation results are then presented in Section IV along with conclusions in Section V.

II. BACKGROUND

As mentioned above, current multi-domain studies have only looked at integration between homogeneous network layers, e.g., IP-IP or DWDM-DWDM. As such these solutions lack extensibility to multi-domain/multi-layer settings and do not address the inherent grooming dimension [1]. In order to get a better sense of the related work, a brief survey is presented.

Early studies on multi-domain packet- and cell-switched networks have focused on *topology abstraction* schemes [4]. Here, a designated routing controller node in the domain condenses domain resource and topology state into an “abstracted” graph with reduced vertices and links. This “abstracted” link state is then flooded to routing controllers in other domains to build a “global” *aggregated graph*, i.e., hierarchical inter-domain routing. Now earlier studies in *asynchronous transfer mode* (ATM) networks have used peer group summarization to achieve very high state reduction, orders magnitude in nature [5]. Subsequent work for IP

quality of service (QoS) networks has also tabled various star, mesh, tree, and spanner graph abstractions. When coupled with path computation strategies (such as widest-shortest, shortest-distance, etc) the results here show very good routing scalability and reduced route fluctuations [4]. More recent efforts have also applied topology abstractions for condensing both delay and bandwidth metrics, e.g., using information-theoretic and line segmentation techniques to bound distortion [6],[7]. Overall results here show very good gains in terms of higher success rates and lower crankback, etc.

Additional work in the area of multi-domain IP/MPLS networking has also applied *signaling* crankback techniques for connection setup. For example, [8] details a *compute while switching* (CWS) scheme in which per-domain computation is used to setup an initial feasible route. Data transmission is then started on this route and control plane crankback initiated in parallel to search for a more optimal route (requiring new RSVP-TE attributes). Results here show very high setup success, on par with global state, as the scheme essentially mimics exhaustive search methods.

Meanwhile, *lightpath routing and wavelength assignment* (RWA) in multi-domain DWDM optical networks has also been studied. For example, [9] details a “domain-by-domain” scheme where border gateway nodes maintain complete alternate routes across all-optical and opto-electronic domains. Simulations results here show good blocking performance, although path dissemination is not studied. Meanwhile [10] studies multi-segment DWDM networks and tables three inter-domain RWA schemes, i.e., *end-to-end* (E2E), *concatenated shortest path* (CSP), and *hierarchical routing* (HIR). Here the E2E scheme assumes a “flat” globalized graph, the HIR scheme assumes a hierarchical graph with segments summarized as nodes, and the CSR scheme only uses local information for segment-by-segment routing. Results with specialized mesh-torus topologies show significant blocking reduction with the E2E and CSR schemes. However no intra or inter-domain routing or path state dissemination is done.

More recently researchers have also adapted topology abstraction for multi-domain DWDM networks. Here, DWDM links pose very different constraints (versus the delay and bandwidth metrics of IP/MPLS links), including wavelengths, timeslots, converters, risk groups, etc. For example, [11] tables a simple-node abstraction scheme but does not address broader signaling and wavelength conversion needs—the latter being a key necessity at domain boundaries performing regeneration and bit-level *service level agreement* (SLA) monitoring. Meanwhile [12] develops full-mesh and star abstractions for all-optical and opto-electronic domains using modified RWA schemes and couples them with inter-domain routing update strategies and distributed RWA schemes. Findings show much-improved blocking performance (and lower signaling loads) with full-mesh abstraction, albeit routing overheads are higher. Finally, [13] presents a theoretical study of partial information models for domains with border node conversion. Lightpath selection is treated as a Bayesian decision and findings show that scalable information models (i.e., logarithmic growth per wavelengths)

achieve a good tradeoff with loss (Bayes error rate). However this treatment is mostly theoretical and only handles bus topologies. Additionally, inter-domain routing and RWA schemes are not studied here.

Finally, some studies have also looked at multi-domain *and* multi-layer networking. For example, [14] tables a *multi-segment* graph model for SONET-DWDM domains and outlines three path selection schemes—centralized (full-knowledge), domain-by-domain (local knowledge), and hierarchical source routing (partial inter-domain knowledge). The latter only propagates domain state for specified granularities, but distributed state aggregation (i.e., topology abstraction, inter-domain routing/dissemination) is not detailed. The results show much lower blocking with increased levels of inter-domain state propagation. Finally, [15] studies Ethernet-DWDM integration using multi-domain signaling to simultaneously provision lightpaths and groom “sub-rate” Ethernet circuits. However underlying DWDM networks are assumed to be static and hence routing is only done at the Ethernet layer. Overall, performance results with varying routing update strategies and crankback show minimal increases in setup latencies and signaling loads.

By and large, the design of multi-domain/multi-layer grooming schemes for real-world distributed settings with partial/aggregated global state is a very complex, and largely unaddressed, problem. Along these lines, a comprehensive solutions framework is now presented.

III. DISTRIBUTED GROOMING SOLUTION

A comprehensive framework is now presented for achieving *distributed grooming* of IP connection/flows in multi-domain IP-DWDM networks. This work builds upon the “single-layer” multi-domain DWDM framework developed by the authors in [12]. Note that full wavelength conversion is assumed at all optical DWDM domains. This precludes the need for wavelength-selective RWA and allows all links to be treated as bandwidth entities (simplified). Although further studies can look to incorporate more restrictive all-optical link types (as per [12]), this is a fairly germane assumption for several reasons. Foremost, most DWDM networks already field a large amount of wavelength conversion. Additionally, numerous studies have shown very close blocking performances between partial and full conversion in both single [16] and multiple [17] domains.

The proposed framework incorporates all key components of multi-domain and multi-layer operation [1]. First, hierarchical inter-domain link-state routing is used to disseminate domain-level state across domains and maintain “global” network visibility. Specifically, multi-granularity topology abstraction is used to condense domain state (Section III.A). Next, distributed multi-layer path computation (grooming) and signaling schemes are defined to leverage this compressed state and achieve distributed grooming (Section III.B). Finally, signaling schemes are defined to reserve route resources and also setup special “tunneled” links (between IP/MPLS border nodes) by extracting any underlying DWDM

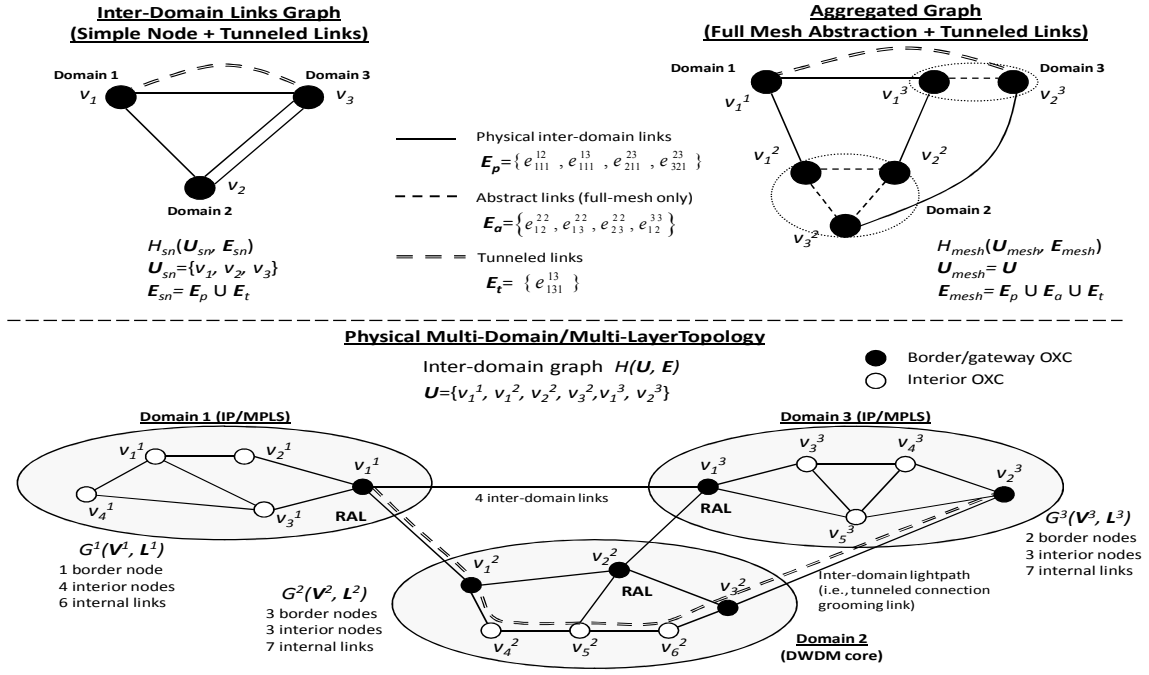


Figure 1: Hierarchical multi-layer/multi-domain graph model: physical, abstract, and tunneled links

lightpath sequences, Section III.C. These links are then advertised across domains to allow multiple higher-layer IP flows to share underlying lightpath capacities. As such, these tunneled links are a vital provision for achieving “distributed” grooming efficiencies. Note that the proposed framework is fully compatible with existing standards and can readily be instantiated using the *generalized multi-protocol label switching* (GMPLS) framework, e.g., two-layer *open-shortest path first-traffic engineering* (OSPF-TE) link-state routing, *reservation protocol* (RSVP) signaling, and *path computation element* (PCE) [1],[8],[19]. Before presenting the proposed framework, however, the required notation is first introduced (all set and vector entities in bold).

Consider a multi-domain/multi-layer network comprising of D domains, with the i -th domain having n^i nodes and b^i border/gateway nodes, $1 \leq i \leq D$, Figure 1. This network is modeled as a collection of domain-level sub-graphs, denoted as $G^i(V^i, L^i)$, $1 \leq i \leq D$, where $V^i = \{v_1^i, \dots, v_{n^i}^i\}$ is the set of physical domain nodes and $L^i = \{l_{jk}^i\}$ is the set of physical *intra-domain* links in domain i ($1 \leq i \leq D$, $1 \leq j, k \leq n^i$), i.e., l_{jk}^{ii} is the link from v_j^i to v_k^i with capacity c_{jk}^{ii} and granularity Δ_{jk}^{ii} . In particular, the latter quantify denotes the minimum unit of allocable bandwidth on a link, e.g., megabits for IP/MPLS links, wavelengths channel bit-rate for DWDM links. Furthermore, without loss of generality, it is assumed that all link connectivity is bi-directional, i.e., there are two opposite direction links between connected nodes. In addition, the set B^i represents the set of border nodes of domain i . And without loss of generality, it is assumed that these nodes are numbered as the *first* group of nodes in their respective domain, i.e., $B^i = \{v_1^i, \dots, v_{b^i}^i\}$.

In order to build a “global” view, a condensed higher-level topology is built and propagated via a *hierarchical* inter-domain routing protocol running between all domain border nodes. Namely, this topology is given by the graph $H(U, E)$, where $U = \sum_i \{B^i\}$ is the set of border nodes across all domains ($1 \leq i \leq D$) and E is the set of links, Figure 1. This graph contains all physical border nodes and inter-domain links but does not necessarily have full connectivity—which is achieved via appropriate topology abstractions (detailed next). Furthermore three link types are identified here (Figure 1), *physical inter-domain links* E_p , *abstract domain links* E_a , and *tunneled inter-domain links* E_t , i.e., $E = E_p \cup E_a \cup E_t$. Physical inter-domain links interconnect border nodes across domains and are denoted as $E_p = \{e_{kmn}^{ij}\}$, where e_{kmn}^{ij} is the n -th link interconnecting node v_k^i in domain i with v_m^j in domain j with spare available capacity c_{kmn}^{ij} and granularity Δ_{kmn}^{ij} . Meanwhile abstract links are only applicable to full-mesh topology abstraction and summarize the traversal cost of a domain (detailed in Section III.A). These link entities and are denoted as $E_a = \{e_{jk}^{ii}\}$, where e_{jk}^{ii} is the abstract link between border nodes v_j^i and v_k^i in domain i and c_{jk}^{ii} is its *computed* available capacity. Finally, “tunneled” links interconnect IP/MPLS border nodes and represent underlying physical DWDM lightpath connections, i.e., $E_t = \{e_{kmn}^{ij}\}$, where e_{kmn}^{ij} is the n -th tunneled connection link between border node v_k^i in domain i and v_m^j in domain j (and c_{kmn}^{ij} denotes the spare capacity on this link). These link types play a vital role in inter-domain grooming and one such link is illustrated in Figure 1 as $v_1^1 - v_2^3$. Further details on the overall hierarchical routing and grooming algorithms are now presented.

A. Hierarchical Routing and Topology Abstraction

Topology abstraction summarizes domain level state and is performed by a designated *border* node in each domain, termed as the *routing area leader* (RAL) [12]. This entity transforms the physical topology into a reduced abstract graph with fewer links. This abstract link state is then flooded to border nodes in other domains/layers (via hierarchical inter-domain routing) in order to maintain a synchronized global network view. For this study, a *full-mesh* abstraction approach is used, although other schemes can also be considered in future work. Namely, this scheme summarizes the domain traversal costs between all border node pairs in a domain, i.e., $O(|\mathbf{B}^i|(|\mathbf{B}^i|-1))$ abstract links. Specifically, the available capacity on each abstract link e_{jk}^{ii} is computed as the average of the bottleneck capacities of the K -shortest paths between the respective border node pairs, denoted as $\{\mathbf{p}_{jkm}^i\}$ where \mathbf{p}_{jkm}^i is the m -th path vector (node-sequence) between border nodes v_j^i and v_k^i in domain i , $1 \leq m \leq K$. Overall, this full-mesh abstraction can be represented as $\mathbf{H}(\mathbf{U}, \mathbf{E}) \rightarrow \mathbf{H}_{\text{mesh}}(\mathbf{U}_{\text{mesh}}, \mathbf{E}_{\text{mesh}})$ where \mathbf{U}_{mesh} is the set of border nodes and $\mathbf{E}_{\text{mesh}} = \mathbf{E}_p \cup \mathbf{E}_t \cup \mathbf{E}_a$, $\mathbf{E}_a = \sum_i \{\mathbf{E}_a^i\}$ and \mathbf{E}_a^i is the above-computed set of abstract links for domain i ($1 \leq i \leq D$), Figure 1. This approach provides good domain visibility, albeit at the cost of significant RAL compute complexity and inter-domain routing loads. Namely border nodes (running the hierarchical inter-domain routing protocol) must maintain state for $O(|\mathbf{B}^i|^2)$ abstract links for each domain i , in addition to state for actual physical and tunneled inter-domain links, i.e., $O(|\mathbf{E}_p| + |\mathbf{E}_t| + \sum_i |\mathbf{B}^i|^2)$.

Note that many studies have used more basic *simple node* abstraction [4],[11],[12] to compare the relative performance of full-mesh abstraction. This abstraction simply reduces a domain to one single “virtual” node emanating all physical inter-domain links and provides no visibility of domain-internal resource levels, see Figure 1. Hence the above transformation is represented as $\mathbf{H}(\mathbf{U}, \mathbf{E}) \rightarrow \mathbf{H}_{\text{sn}}(\mathbf{U}_{\text{sn}}, \mathbf{E}_{\text{sn}})$ where \mathbf{U}_{sn} is the set of virtual nodes representing each domain and $\mathbf{E}_{\text{sn}} = \mathbf{E}_p \cup \mathbf{E}_t$ is the set of physical and tunneled inter-domain links, i.e., $\mathbf{E}_a = \Phi$. Hence the inter-domain routing state overheads here are bounded by $O(|\mathbf{E}_p| + |\mathbf{E}_t|)$.

Now consider the dissemination of “global” link state across multiple layers, i.e., physical, abstract, and tunneled ($\mathbf{E}_p, \mathbf{E}_a, \mathbf{E}_t$). As mentioned earlier, this is done by running a hierarchical link state routing protocol between domain border nodes, e.g., second level of OSPF-TE [19]. Specifically, *link-state advertisement* (LSA) updates are generated using *significance change factor* (SCF) triggering policies [12]. Namely, LSA updates for physical or tunneled (lightpath connection) links are flooded to all neighboring nodes only if the *relative* change in available bandwidth on the node’s link exceeds the SCF value and the duration since the last update exceeds a *hold-down timer* (HT). Meanwhile, routing updates for abstract domain-level links (for full-mesh abstraction only) are handled slightly differently as these pertain to *computed* entities and not actual underlying physical links or lightpath sequences. Specifically, topology abstractions are regularly computed at the expiry of the HT at the RAL. After these computations, an SCF threshold approach is again used to

extract a subset of abstracted links for update propagation, i.e., “two-step” method, akin to that used in [12] for “flat” single-layer DWDM domains.

B. Multi-Domain Grooming/Path Computation

Distributed grooming implements path computation for “higher-layer” IP connection/flows across multiple domains. This is a very challenging problem as the end-to-end path can traverse multiple layers (e.g., IP/MPLS-to-DWDM-to-DWDM-to-IP/MPLS) and internal domain state is very limited as per the abstraction type (Section III.A). Along these lines, a *hierarchical* distributed computation approach is proposed, in line with the RSVP signaling standards [19]. Namely, the source IP/MPLS node first queries its closest border node (or RAL) to compute a *loose route* (LR) domain sequence to the destination, i.e., “skeleton path”. This queried border node essentially acts like a PCE [8]. If LR computation is successful here, the source node initiates RSVP-TE *explicit route* (ER) expansion signaling along the “skeleton” route to resolve all links and reserve resources (i.e., IP/MPLS link bandwidth, DWDM link wavelengths), Section III.C.

Now consider the actual LR path computation which is done using the source domain RAL node’s copy of the inter-domain graph, i.e., $\mathbf{H}_{\text{sn}}(\mathbf{U}_{\text{sn}}, \mathbf{E}_{\text{sn}})$ or $\mathbf{H}_{\text{mesh}}(\mathbf{U}_{\text{mesh}}, \mathbf{E}_{\text{mesh}})$. Here, two grooming strategies are proposed, namely *unified* and *layer-by-layer* (LBL) path computation. The former approach treats all links in the same manner regardless of their granularity types (IP or DWDM). As a result, this scheme tends to favor high bandwidth (i.e., low cost) DWDM links over lower bandwidth IP/MPLS links. Conversely, the latter technique attempts to route (i.e., groom) as many connections as possible on to existing (physical, tunneled) IP/MPLS layer links in order to save DWDM layer resources. Namely it first searches for a feasible path at the “IP/MPLS sub-graph” level. This graph is generated by copying all IP/MPLS links (i.e., equivalent granularity levels Δ_{kmn}^i) to a temporary graph and then running Dijkstra’s shortest path algorithm. If an “all-IP/MPLS” path can be found here, the computed LR is selected and RSVP-TE signaling initiated. Conversely, if such a path cannot be found, the LBL scheme copies the DWDM layer links to the temporary sub-graph and then re-runs the above unified path computation algorithm. In general, this strategy will work well with more tunneled links.

Carefully note that both of the above grooming schemes (unified, LBL) can further implement *minimum hop count* or *minimum distance* routing. Specifically, in the former case, the link costs in Dijkstra’s shortest-path algorithm are set to unity to achieve resource minimization. Conversely in the latter case, link costs are set inversely-proportional to the free/available link capacity in order to achieve load balancing [4],[12]. Also note that wavelength selection issues do not arise at the DWDM layer since all links are assumed to have full conversion capabilities. Nevertheless, wavelength selection in all-optical and/or partial conversion DWDM domains has been studied in [17] and this can be incorporated in future studies.

C. Setup Signaling and Tunneled Link Extraction

Setup signaling represents the final step in the grooming process. Namely, the sourcing node uses the above-computed LR sequences to initiate RSVP-TE *PATH* signaling message to resolve explicit end-to-end paths. Here, downstream border nodes receiving these *PATH* messages perform ER expansion on the incoming LR sequence to resolve explicit *intra-domain* node sequence across the domain, both IP/MPLS and DWDM. Specifically, this expansion is done by searching the K -shortest paths to the domain egress node, $\{p_{jkm}^i\}$ Section III.A, and selecting the sequence with highest amount of available capacity. This essentially implements a *widest-shortest* approach [4] on the intra-domain topology and is designed to achieve resource minimization.

As mentioned in the start of Section III, tunneled links are also extracted during the signaling phase. These links pertain to underlying DWDM lightpath segments and play a vital role in improving distributed grooming efficiencies. Specifically, these links are generated by having IP/MPLS border nodes check all RSVP-TE *RESV* messages propagating back towards the source. If the downstream node for the IP/MPLS node is a DWDM node, the lightpath sequence to the egress IP/MPLS border node is extracted and advertised as a “direct” tunneled link between the two respective domains. For example in Figure 1, the IP/MPLS border node v_1^1 will extract and generate a tunneled link to IP/MPLS border node v_2^3 by checking the lightpath sequence $v_1^2 - v_4^2 - v_5^2 - v_6^2 - v_3^2$ across the DWDM domain 2. This link is then advertised at the inter-domain level as a direct IP/MPLS-to-IP/MPLS link. Finally, a tunneled link is only taken down when its usage drops to zero, i.e., by issuing RSVP-TE takedown sequence.

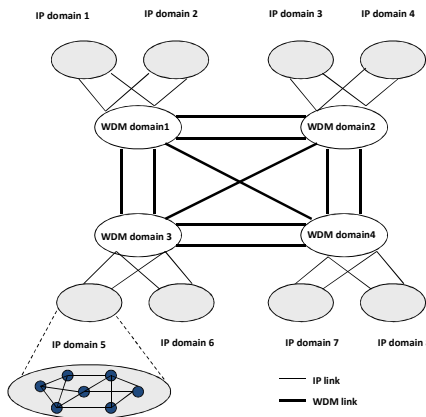


Figure 2: Sample multi-domain/multi-layer network topology

IV. PERFORMANCE EVALUATION

The performance of the proposed distributed multi-domain IP/MPLS-DWDM grooming framework is tested using the *OPNET Modeler*TM simulation tool. Namely, a sample multi-domain/multi-layer topology is designed here, consisting of 8 stub (client) IP/MPLS domains hosted by 4 core DWDM domains, Figure 2. Average domain sizes are set to 6-10 nodes in each case. This topology is chosen to reflect real-world

operational settings in which “sub-rate” IP/MPLS nodes are usually at network stub and high-bandwidth nodes are at network core. At the inter-domain level, this network topology has 16 physical inter-domain IP/MPLS links and 10 DWDM links. Furthermore, each IP/MLS and DWDM link is assigned 10 Gb/s and 100 Gb/s capacity, respectively (i.e., 10 wavelengths per link). Meanwhile the traffic granularities of these links are set to 50 Mb/s for IP/MPLS and 10 Gb/s (wavelength capacity) for DWDM, respectively. Note that border DWDM OXC nodes connecting to higher-layer IP/MPLS domains are assumed to be grooming capable and can hence aggregate traffic from stub domains. It is also assumed that all DWDM nodes are fully conversion-capable. Moreover, each IP/MPLS domain has two inter-domains links connecting it to its core DWDM domain, i.e., dual-homed to achieve scalability and reliability.

For the purposes of evaluation, all connections are generated between randomly-selected domains using a 70-30% intra/inter-domain ratio. A node can only generate traffic to a node of the same type, i.e., IP/MPLS-to-IP/MPLS or DWDM-to-DWDM request, and for the case of DWDM-to-DWDM, no grooming is performed. Furthermore, within a given domain, the respective source and destination nodes are also chosen randomly via a uniform distribution. For routing updates, all intra/inter-domain routing hold-down timers (*HT*, *IHT*) are set to 5 seconds and the SCF routing update threshold is set to 10%. In addition, the maximum number of virtual connection links are varied from 0 (i.e., no grooming) to 32. Finally, all runs are averaged over 100,000 connections and mean holding times are set to 600 sec (exponential). Request inter-arrival times are also exponential and vary with load.

The inter-domain request *bandwidth blocking rate* (BBR) is first measured for the unified and LBL schemes with varying numbers of tunneled links, Figure 3 (minimum distance load balancing, simple node abstraction, and tunneled link counts $|E_t|$ denoted as “VC” for “virtual connection” links). Foremost, these results show a clear graded relationship between the number of tunneled links and inter-domain BBR for the LBL scheme. Namely, performance increases significantly for larger tunneled link counts, e.g., at 126 Erlang, 4 and 8 tunneled links reduce the inter-domain BBR by 60% and 90%, respectively, versus no grooming. Conversely, the unified grooming scheme is less sensitive to the number of tunneled links. This is due to the fact that this approach favors higher-capacity DWDM links when running load-balancing Dijkstra’s path computations, i.e., generally larger available capacities yield lower link weights (inverse of available capacity). Hence increasing the number of tunneled links in this unified approach does not necessarily lower blocking. Note that the results for the LBL scheme in Figure 3 also indicate a “leveling off” of in BBR reduction after 16 tunneled links, e.g., at 144 Erlang, 32 tunneled links only reduce inter-domain BBR by 5% versus 16 tunneled links.

As a further comparison, the inter-domain BBR performance of the unified grooming approach (using hop-count routing) is also presented in Figure 4. Namely, both the DWDM and IP/MPLS link weights are set to unity here,

implying that both link types are equally likely to be chosen. Akin to the LBL results in Figure 3, the results here also show notable improvements with increased tunneled link counts.

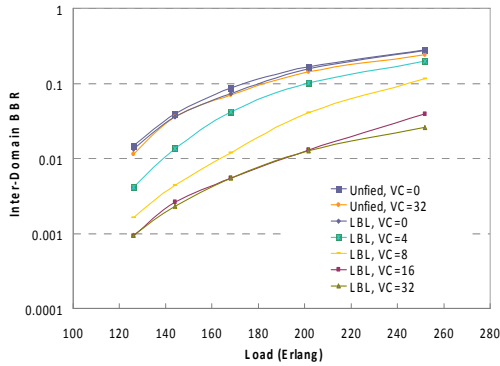


Figure 3: Inter-domain BBR (min. distance load-balancing)

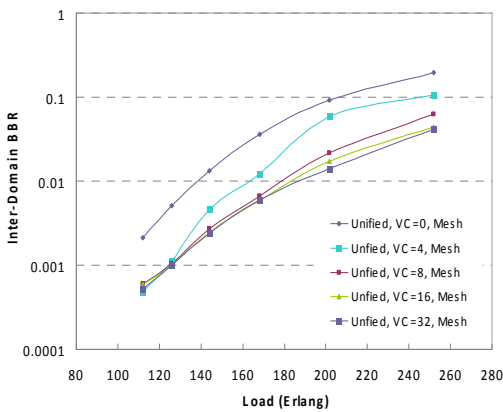


Figure 4: Inter-domain BBR (min. hop resource minimization)

Next, the impact of advanced topology abstraction algorithms is studied. As mentioned earlier in Section III.A, full-mesh abstraction provides more accurate domain state (versus simple-node abstraction), and hence may assist in bypassing of congested domains. Along these lines, full-mesh abstraction is evaluated for the unified grooming scheme, and the results are shown in Figure 5. Here it is seen that the addition of abstract link state yields good BBR reduction in blocking versus simple-node abstraction for both 0 (i.e., no grooming) and 32 tunneled links. For example, at 126 Erlang loads, full-mesh abstraction reduces the inter-domain BBR by about 50% versus simple node for 32 tunneled links. However, the relative separation between these two abstraction schemes is lower for the LBL scheme, Figure 6. This observation can be explained by the fact that the LBL grooming scheme first searches for a path at the IP/MPLS layer, essentially disregarding DWDM-layer abstract link state to an extent. Moreover, with increased tunneled link counts, the difference between full-mesh and simple-node abstraction decreases further (with the LBL scheme). Overall, this behavior is expected as the probability of finding a path in IP/MPLS layer becomes higher and therefore the next step of LBL path computation, i.e., incorporating DWDM layer to run Dijkstra’s algorithm, is less likely to be invoked.

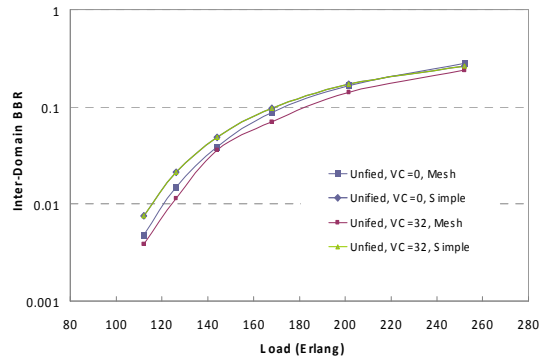


Figure 5: Inter-domain BBR (min. distance load-balancing)

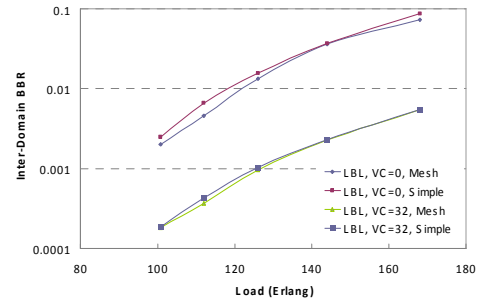


Figure 6: Inter-domain BBR (min. distance load-balancing)

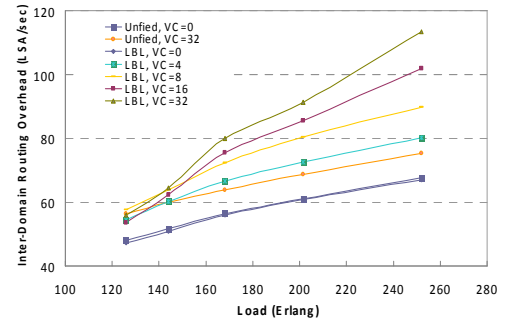


Figure 9: Inter-domain routing overheads (LSA/sec)

Finally, the scalability of the hierarchical link-state routing approach is also gauged by measuring *inter-domain* routing message loads in terms of link-state updates/sec (LSA/sec), Figure 9. Overall, these results show a very direct dependence upon network load, close to a linear relationship. More importantly, the LBL scheme yields notably higher routing loads across all load ranges versus the unified approach. This increase is due to the fact that this strategy “perturbs” capacity levels across many more (IP/MPLS physical and tunneled) links, thereby yielding more threshold crossovers. Nevertheless, LBL routing loads still do not grow linearly with tunneled link counts, e.g., at 200 Erlang load the LBL routing load with 4 tunneled links is 72 LSA/sec, whereas it is 81 LSA/sec for 8 tunneled links and 85 LSA/sec for 16 tunneled links. These results indicate acceptable routing scalabilities and this can be further improved by incorporating

advanced “gradient-based” update triggering strategies [20]. Overall, the maximum number of tunneled links will reflect a design choice by network operators, allowing them to achieve a desired tradeoff between routing scalability and inter-domain BBR performance.

V. CONCLUSIONS

This paper presents a comprehensive framework for connection grooming in distributed multi-domain/multi-layer networks. Namely, a hierarchical routing setup is first developed to help build and maintain condensed partial “global” views, complete with physical and tunneled link state information. Commensurate, inter-domain routing/triggering policies are also defined. Subsequently, two distributed grooming schemes are proposed based upon unified and layer-by-layer path computation. Detailed performance evaluation results show much-improved bandwidth-blocking performance with layer-by-layer grooming, albeit at the cost of increased routing overheads. These results also show that the use of advanced full-mesh topology abstraction techniques can yield further improvements in blocking performance. Future studies will look at larger multi-domain topologies and also study the impact of partial wavelength conversion in DWDM domains. Distributed survivability methodologies will also be studied for these grooming scenarios.

REFERENCES

- [1] N. Ghani, *et al*, “Control Plane Design in Multidomain /Multilayer Optical Networks,” *IEEE Communications Mag.* Vol. 46, No. 6, June 2008, pp. 78-87.
- [2] A. Kasim (Editor), *Delivering Carrier Ethernet: Extending Ethernet Beyond the LAN*, McGraw Hill, November 2007.
- [3] N. Rao, *et al*, “Ultra Science Net: Network Testbed for Large-Scale Science Applications,” *IEEE Communications Magazine*, November 2005, Vol. 3, No. 4, pp. S12-S17.
- [4] F. Hao, E. Zegura, “On Scalable QoS Routing: Performance Evaluation of Topology Aggregation,” *IEEE INFOCOM 2000*, pp. 147-156.
- [5] I. Iliadis, “Optimal PNNI Complex Node Representations for Restrictive Costs and Minimal Path Computation Time”, *IEEE/ACM Trans. on Networking*, Vol. 8, No. 4, August 2000.
- [6] B. Awerbuch, Y. Shavitt, “Topology Aggregation for Directed Graphs,” *IEEE/ACM Transactions on Networking*, Vol. 9, No. 2, February 2001, pp. 82-90.
- [7] K. Liu, K. Nahrstedt, S. Chen, “Routing with Topology Abstraction in Delay-Bandwidth Sensitive Networks,” *IEEE/ACM Trans. on Networking*, Vol. 12, No. 1, February 2004, pp. 17-29.
- [8] F. Aslam, *et al*, “Interdomain Path Computation: Challenges and Solutions for Label Switched Networks,” *IEEE Communications Mag.*, Vol. 45, No. 10, Oct. 2007, pp. 94-101.
- [9] X. Yang, B. Ramamurthy, “Inter-Domain Dynamic Routing in Multi-Layer Optical Transport Networks,” *IEEE GLOBECOM 2003*, San Francisco, CA, December 2003.
- [10] Y. Zhu, A. Jukan, M. Ammar, “Multi-Segment Wavelength Routing in Large-Scale Optical Networks,” *IEEE ICC 2003*, Anchorage, AL, May 2003.
- [11] S. Sanchez-Lopez, *et al*, “A Hierarchical Routing Approach for GMPLS-Based Control Plane for ASON,” *IEEE ICC 2005*, Seoul, Korea, June 2005.
- [12] Q. Liu, *et al*, “Hierarchical Routing in Multi-Domain Optical Networks”, *Computer Communications*, Vol. 30, Issue. 1, December 2006
- [13] G. Liu, *et al*, “On the Scalability of Network Management Information for Inter-Domain Light-Path Assessment,” *IEEE Trans. on Networking*, Vol. 13, No. 1, Jan. 2005, pp. 160-172.
- [14] Y. Zhu, A. Jukan, M. Ammar, W. Alanqar, “End-to-End Service Provisioning in Multi-Granularity Multi-Domain Optical Networks,” *IEEE ICC 2004*, Paris, France, June 2004.
- [15] A. Hadjiantonis, *et al*, "Evolution to a Converged Layer 1, 2 in a Global-Scale, Native Ethernet Over WDM-Based Optical Networking Architecture," *IEEE JSAC*, Vol. 25, No. 5, June 2007, pp. 1048-105.
- [16] H. Zang, J. Jue, B. Mukherjee, “A Review of Routing and Wavelength Assignment Approaches for Wavelength- Routed Optical WDM Networks”, *Optical Networks Magazine*, Vol. 1, No. 1, Jan. 2000.
- [17] Q. Liu, N. Ghani, “Topology Abstraction Schemes in Multi-Domain Full Wavelength Conversion DWDM Networks,” *IEEE Symposium on Advanced Networks & Telecommunications Systems (ANTS) 2007*, Bombay, India, December 2007.
- [18] R. Dutta, A. Kamal, G. Rouskas (Editors), *Traffic Grooming for Optical Networks: Foundations, Techniques and Frontiers*, Springer, Boston 2008.
- [19] G. Bernstein, B. Rajagopalan, D. Saha, *Optical Network Control-Architecture, Protocols and Standards*, Addison Wesley, Boston 2003.
- [20] Q. Liu, *et al*, “Inter-Domain Routing Scalability in Optical DWDM Networks,” *IEEE ICCCN 2008*, US Virgin Islands, August 2008.