This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2010 proceedings

Enhanced Crankback Signaling for Multi-Domain Traffic Engineering

F. Xu[1], M. Esmaeili[1], C. Xie[1], N. Ghani[1], M. Peng[2], Q. Liu[3]
[1]*University of New Mexico,* [2]*Wuhan University,* [3]*Oak Ridge National Laboratory*

Abstract: *Multi-domain traffic engineering is a major focus area for carriers today and crankback signaling offers a very promising and viable alternative here. Although some initial crankback studies have been done, there is still significant latitude for improving multi-domain crankback performance. To address these concerns, this paper studies realistic IP/MPLS multi-domain networks and proposes a novel solution for joint intra/inter-domain signaling crankback. Namely, dynamic intra-domain link-state routing information is coupled with inter-domain path/distance-vector routing state to improve the overall search process. Mechanisms are also introduced to limit setup signaling overheads/delays. The performance of the proposed solution is then analyzed using simulation and compared against other crankback techniques as well as hierarchical inter-domain routing strategies.*

Keywords: *Crankback signaling, multi-domain networks*

I. INTRODUCTION

Traffic engineering (TE) in IP-based *multi-protocol label switching* (MPLS) and optical *generalized MPLS* (GMPLS) networks is a very well-studied problem area. Here a wide range of constraint-based routing solutions have been proposed, but most have focused on single "domain" settings in which a provisioning entity has complete "network-wide" topology/resource views, e.g., single *autonomous system* (AS) running link-state routing [1],[2]. However, as user application demands grow, there is a strong desire to achieve TE provisioning across multiple domains, i.e., inter-AS TE, particularly for higher-end applications such as *voice over IP* (VoIP), packet video transport, *virtual private network* (VPN) extension, etc. Owing to obvious scalability and confidentiality concerns here, it is clear that this must be achieved in a distributed, decentralized manner.

To address these challenges, a diverse set of provisions have emerged to help improve multi-domain TE support, both at the IP/MPLS and underlying optical GMPLS layers [1]-[5]. On the standards side, many ubiquitous routing protocols already provide varying levels of inter-domain visibility, e.g., next-hop/path-vector dissemination in *exterior gateway protocol* (EGP) and hierarchical link-state dissemination in two-level *open-shortest-path-first* (OSPF-TE). Furthermore, the new IETF *path computation element* (PCE) [3] framework also defines multi-domain path computation strategies.

Meanwhile on the research side, a host of multi-domain TE schemes have been studied. A key focus here is to address the tradeoff between inter-domain visibility and control plane complexity (i.e., dissemination, computation) [1]. For example, some have developed hierarchical link-state routing solutions to increase inter-domain visibility [6]-[13]. The major contributions here are graph-theoretic topology abstractions for compressing domain-level state in IP/MPLS [6]-[9] and optical *dense wavelength division multiplexing* (DWDM) networks [10]-[13]. However, even though hierarchical routing delivers good performance, associated routing overheads are very high. Hence these schemes will likely be problematic in real-world settings where carriers tend to prefer distance/path-vector protocols, e.g., *border gateway protocol* (BGP) variants. These latter types only provide hop-count and reachability state and most operational versions do not support any *quality-of-service* (QoS) parameters, e.g., delay, bandwidth, etc. Hence to address these concerns, alternate "per-domain" computation schemes have been proposed for multi-domain TE [14]-[16], leveraging crankback signaling to overcome lower inter-domain visibility. Although some good blocking gains are also seen here, related signaling overheads/setup delays can be significant [15].

In light of the above, there is a clear need (and significant scope) to develop more advanced multi-domain crankback solutions and gauge their performance against counterpart hierarchical routing schemes. Along these lines, this paper proposes an enhanced crankback solution for multi-domain networks using the standard *resource reservation* (RSVP-TE) protocol. A key aim of this study is to address realistic scenarios where individual domains have full internal visibility via link-state routing, e.g., OSPF-TE, but otherwise generally limited "next-hop" inter-domain visibility, e.g., BGP routes. Two levels of crankback are defined—at the intra and inter-domain levels—and active crankback history (failure state) is also leveraged. Note that even though the work herein focuses on bandwidth provisioning IP/MPLS networks, it can readily be extended to optical wavelength-routing networks.

The paper is organized as follows. Section II first presents a survey of the latest work on multi-domain TE provisioning, including standards and research-based activities. Subsequently, Section III presents the proposed enhanced intra/inter-domain crankback signaling solution. Detailed performance analysis is then conducted in Section IV and the results compared against those from counterpart hierarchical inter-domain routing schemes. Finally, conclusions and future research directions are highlighted in Section V.

II. BACKGROUND

The IETF has defined a range of TE capabilities for multi-domain provisioning. Foremost, the PCE framework has been introduced to decouple path computation from signaling by defining domain-level computational entities. These entities can either reside in a standalone manner or be co-located with nodes and have access to the internal domain resource/policy databases. At the inter-domain level, these PCE entities can interact in a distributed manner to resolve end-to-end routes and two approaches have been defined for varying levels of "global" state, i.e*., per-domain* and *PCE-based* [3],[15]. The former schemes compute TE paths in a "domain-to-domain" manner and are most germane for limited inter-domain

978-1-4244-6404-3/10/$26.00 ©2010 IEEE

visibility. Meanwhile the latter schemes rely upon the head-end PCE to computes a *partial* or *loose* route to the destination (domain sequence) and are more suited for increased inter-domain visibility. However, even after path computation, blocking can occur during signaling along a chosen route. Hence new RSVP-TE crankback extensions have also been defined to re-try alternate routes [4]. Specifically, varying levels of multi-domain crankback have been outlined, i.e., *local domain*, *intermediate domain*, *source-based*.

Researchers have also studied various multi-domain TE schemes, broadly grouped as *hierarchical routing* or *per-domain* strategies, i.e., PCE-based [3]. In the latter, local domain topology/ resource state is condensed to generate an "abstracted" graph with fewer vertices and links, e.g., at a designated controller in each domain. This state is then flooded to other domains using hierarchical *link-state* routing between border gateways to build a "global" aggregated graph. For example, earlier work in peer group summarization for *asynchronous transfer mode* (ATM) networks has shown very high levels of state reduction [1]. Subsequent studies on multi-domain IP/MPLS networks have also proposed a variety of graph abstractions (e.g., star, mesh, tree, spanner graphs, etc) to compress bandwidth [6], bandwidth-delay [7],[8], and even diversity/survivability [9] information. When coupled with various computation heuristics (widest-shortest, shortest-widest, generalized cost, etc) these schemes yield very good blocking reduction and lower setup signaling overheads.

Now topology abstraction/hierarchical routing has also been applied in multi-domain DWDM networks, i.e., to summarize wavelength/converter/risk-group information. For example [10] outlines simple-node abstraction for all-optical domains. Meanwhile [11] and [12] develop full-mesh and star schemes for more realistic multi-domain settings with partial (boundary) conversion. Distributed *routing and wavelength assignment* (RWA) algorithms are also defined to leverage this "global" state. Findings show good inter-domain blocking reduction with full-mesh abstraction (about 20-40% lower than single node abstraction), albeit routing overheads are much higher, almost 200-300% higher. Further abstractions for multi-domain optical survivability are also presented in [13].

However, topology abstraction entails significant link-state routing overheads at the inter-domain level, e.g., second level of *open-shortest-path-first* (OSPF-TE) [1],[12]. Hence the adoption of this approach may be limited in real-world settings where more scalable distance/path-vector protocols are already well-entrenched. Along these lines, a handful of studies have proposed *signaling-based* crankback strategies for "per-domain" path computation (akin to PCE classification [5]). The goal here is to have individual domains compute their own traversing segments to build a concatenated end-to-end path. For example, [14] defines a basic *"per-domain"* (PD) crankback scheme which probes egress domain nodes for traversal routes and upon failure, notifies upstream border nodes. Overall results show higher request blocking rates and crankback delays, particularly when compared to PCE-based strategies utilizing pre-determined inter-domain routes. Meanwhile, [15] and [16] detail a modified *compute while switching* (CWS) scheme for MPLS networks. First, a similar crankback procedure to [14] is used to compute an initial inter-domain route, i.e., by probing egress nodes specified by interior and/or exterior gateway protocols. If this search is successful, transmission is started and *simultaneous* crankback is initiated to search for a shorter route, e.g., since per-domain computation generally does not yield the shortest route. If a shorter route is found, data switchover is performed. Results here show good setup success rates as the scheme essentially mimics an exhaustive search. However, CWS entails very high signaling overheads/delays (not analyzed) and requires non-standard extensions to RSVP-TE attributes. Moreover, hitless post-setup flow switchovers may be difficult, especially in GMPLS settings. Finally, [17] addresses end-to-end path delays in multi-domain settings and presents two next-hop domain selection strategies. The first selects the next-hop as the "nearest" egress border node in the domain whereas the other uses tailored inter-domain *round-trip time* (RTT) measurements, i.e., pre-computed global state. Overall the latter heuristic is shown to yield slightly higher carried load and less crankbacks. However, it requires adoption of a very specialized virtual coordinates system [17].

Overall, the above crankback solutions embody some good initial contributions. However, added innovations are possible for multi-domain settings, e.g., such as novel schemes to limit crankback overheads/delays, improved use of intra/inter-domain crankback (failure) history, and judicious application of available distance/path-vector routing state.

III. ENHANCED CRANKBACK SOLUTION

An enhanced multi-domain crankback solution is now presented based upon standard IETF protocols. The solution assumes realistic settings with full link-state routing at the intra-domain level (e.g., OSPF-TE) and more scalable path/distance-vector routing at the inter-domain level (e.g., BGP). Furthermore, each domain is assumed to have at least one PCE entity with full access to interior and exterior routing databases. This entity plays a key role in the crankback process as it helps resolve next-hop domains (egress border gateways). Meanwhile, all setup signaling is done using the recent crankback framework defined for RSVP-TE [4].

Overall three key innovations are introduced to enhance multi-domain crankback, i.e., 1) dual intra/inter-domain crankback counters to limit signaling complexity/delay, 2) full crankback history usage to improve re-try domain traversal, and 3) intelligent per-domain search strategies leveraging existing "next-hop" EGP state. The details are now presented.

A. Multi-Domain Crankback Operation

Before detailing the scheme, the requisite notation is introduced. Consider a multi-domain network comprising of D domains, with the i-th domain having n^i nodes and b^i border/gateway nodes, $1 \le i \le D$. This network is modeled as a set of domain sub-graphs, $\boldsymbol{G}^i(\boldsymbol{V}^i,\boldsymbol{L}^i)$, $1 \le i \le D$, where $\boldsymbol{V}^i = \{ v^i_1, v^i_2, \dots \}$ is the set of domain nodes and $\boldsymbol{L}^i = \{ l^{ii}_{jk} \}$ is the set of *intra-domain* links in domain i ($1 \le i \le D$, $1 \le j,k \le n^i$), i.e., l^{ii}_{jk} is the link from v^i_j to v^i_k with available capacity c^{ii}_{jk}. A physical inter-domain link connecting border node v^i_k in domain i with

border node v^j_m in domain j is further denoted as l^{ij}_{km} and has available capacity c^{ij}_{km}, $1 \le i,j \le D$, $1 \le k \le b^i$, $1 \le m \le b^j$. Also, $\boldsymbol{B}^i$ denotes the set of border nodes in domain i. Now consider the relevant RSVP-TE message fields. The path route is given by a node vector, $\underline{\boldsymbol{R}}$. Meanwhile, other fields are also defined for crankback [4], and include an exclude link vector, $\underline{\boldsymbol{X}}$, to track crankback failure history as well as dual intra/inter-domain crankback counters, h_1 and h_2 (whose usage will be detailed shortly). Note that [4] only defines a single counter field but bit masking can be used to generate two "sub-counters".

An overview of per-domain computation is first given for the case of *non-crankback* operation, i.e., no resource failures. Consider a source node receiving a request for x units of bandwidth to a destination node in another domain. This source queries its PCE to determine an egress link to the next-hop domain, e.g., using the *PCE-to-PCE protocol* [3],[5]. The PCE then determines the next-hop domain to the destination domain (detailed in Section III.B) and returns a domain egress border node/link to this domain. Note that this information also contains the *ingress* border node in the downstream domain. Upon receiving the PCE response, the source uses its local OSPF-TE database to compute an *explicit route* (ER) to the specified egress border node. This step searches the k-shortest path sequences over the *intra-domain* feasible links (i.e., $c^{ii}_{jk} \ge x$) and chooses the one with the lowest "load-balancing" *cost*, i.e., individual link costs inversely-proportional to free link capacity, i.e., $1/c^{ij}_{km}$. This method is used as it outperforms basic hop count routing, see [6],[11]. Granted that an ER path is found, it is inserted in the path route vector, $\underline{\boldsymbol{R}}$, and RSVP-TE *PATH* messaging is then initiated (along the expanded route) to the ingress border node in the next-hop domain. Here, each intermediate node checks for available bandwidth resources on its outbound link and pending availability, propagates the message downstream. The above procedure is repeated at all next-hop domain border nodes until the destination domain. When the *PATH* message finally arrives at the destination domain, the border node (or PCE) expands the ER to the destination. Upon receiving a fully-expanded *PATH* message, the destination then initiates upstream reservation, i.e., *RESV* message.

Now clearly insufficient bandwidth resources can cause downstream *PATH* setup failure at any route link. Hence various levels of crankback have been outlined in [4], and two types are adapted here, *intra-domain* (local) and *inter-domain* (intermediate). Now although other studies have also considered such strategies, e.g., [14]-[16], a key differentiating factor here is the use of the dual counters (h_1, h_2) to limit the number of re-try attempts and incorporation of crankback (failure) history to prune both intra and inter-domain links. Specifically, the above counters are initialized to pre-specified limit values (H_1 and H_2, respectively) in the initial *PATH* message and then decremented during crankback to limit setup delays/overheads for longer and less efficient paths.

Now in terms of crankback operation, two key steps are defined here, i.e., *notification* and *re-computation*. The former refers to the (upstream) signaling procedures executed upon link resource failure at an intermediate node, whereas the latter refers to the actual re-routing procedures to select a new route. Now in general, resource failures can occur at *three* different node types, i.e., domain ingress border nodes, domain egress border nodes, and interior nodes. However, in the proposed scheme, only the former performs *re-computation* whereas the latter two simply perform crankback *notification*. These steps are now detailed further.

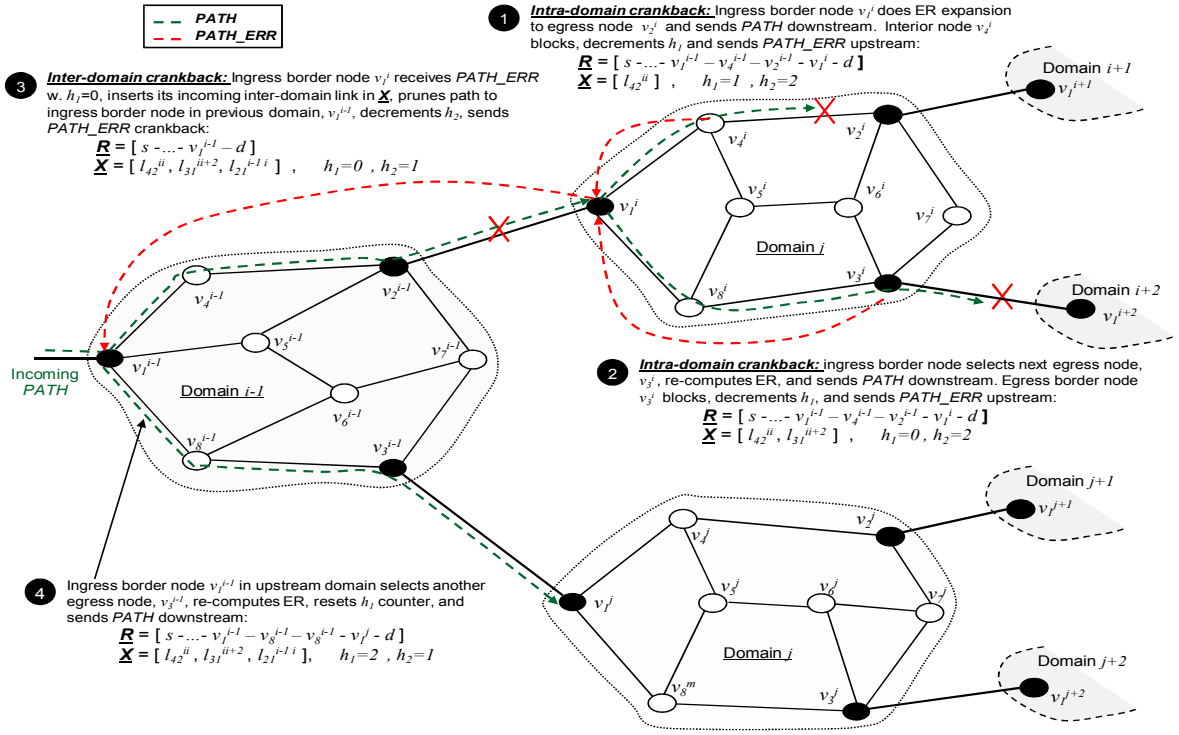

Figure 1: Enhanced intra/inter-domain crankback scheme (H_1=2, H_2=2)

Crankback Notification: Upstream notification is done when there is insufficient bandwidth at an intra-domain link (at an intra-domain node) or an inter-domain link (at an egress border node) on an already-expanded ER. Here the *PATH* message is terminated and its appropriate fields updated and copied to an upstream *PATH_ERR* message to the domain's ingress border node. Specifically, the intra-domain counter h_1 is decremented and the congested link is noted in $\underline{\boldsymbol{X}}$. Note that if blocking occurs in the source domain, the *PATH_ERR* is sent to source. Hence the notifying node runs the following:

```
if (insufficient resources on outbound link)
    Decrement intra-domain counter h1, extract route
    vector R and exclude link vector X from PATH

    Add congested outbound link to exclude route X

    Remove all nodes in route vector R up to ingress
    border node (prune congested intra-domain segment)

    Generate PATH_ERR, copy h1, R, X fields and send to
    upstream ingress border node
```

An example of crankback notification is shown in Figure 1 for interior and egress border nodes (H_1, H_2=2). For example, consider bandwidth blocking on the link $l^{ii}{}_{42}$, i.e., step 1. Here, the interior node v_4^i prunes the route vector $\underline{\boldsymbol{R}}$ to the domain ingress node, adds the blocked link to the exclude route vector $\underline{\boldsymbol{X}}$, decrements the intra-domain counter h_1, and sends all this information back to the ingress node v_1^i via a *PATH_ERR* message. A similar procedure is also shown for blocking at the egress border node v_3^i (i.e., step 2, Figure 1).

Crankback Re-Computation: Meanwhile, path re-routing is done by ingress border nodes receiving a *PATH_ERR*. Note that for special case of a source domain (i.e., non-ingress border node), the receiving source node relays the *PATH_ERR* to its PCE for processing. Now two types of crankback re-computations can be done. First consider "intra-domain" crankback. If the arriving h_1 counter has not expired, a next-hop domain/egress border node is selected by the ingress border node (or PCE) for ER expansion. Here the exact sequence of next-hop domains tried is pre-computed to try *successively longer* inter-domain routes (i.e., via multi-entry distance vector table, detailed in Section III.B). Now the enhanced scheme makes full use of crankback history to avoid congested links. Foremost, all congested inter-domain links in $\underline{\boldsymbol{X}}$ that egress from the domain are removed from consideration, i.e., only consider "non-congested" next-hop domain egress links. Additionally, all intra-domain links listed in the exclude link vector $\underline{\boldsymbol{X}}$ are also precluded from *local* ER computation. Finally, the route vector $\underline{\boldsymbol{R}}$ is also searched to make sure that an upstream domain is not traversed twice, i.e., no "domain-level" loops. Regardless, it still may not be possible to initiate/establish a domain-traversing route for various reasons, i.e., h_1 counter expired, LR expansion failure to selected egress node, or all egress border links in exclude link vector $\underline{\boldsymbol{X}}$, etc. In these cases, the ingress border node initiates a more globalized "inter-domain crankback" response via a *PATH_ERR* message to the ingress node in the *upstream* domain in the *PATH* route vector $\underline{\boldsymbol{R}}$ (or source node if upstream domain is source domain) with the h_1 counter reset to H_1. To improve crankback history tracking, the ingress border node also inserts its own *ingress link* in the exclude route vector of the *PATH_ERR* message, i.e., in order to avoid future re-tries on this link. Note that "inter-domain crankback" is only initiated if the h_2 value is non-zero, otherwise the request is dropped (i.e., *PATH_ERR* to source). In summary, an ingress border node runs the following:

```
/* Attempt intra-domain re-routing */
if (h1 not expired)
    Select next-hop domain/egress link using multi-entry
        distance vector table s.t. next-hop domain is not in R
        and egress link is not in X

    if (next hop egress node found)
        Make copy of local network graph (via IGP database),
        prune all local failed links listed in X, compute new ER
        to egress border node

        if (LR expansion successful)
            Initiate PATH signaling to new egress node
            intra_domain_crankback_done=1;

/* Attempt inter-domain re-routing */
if (!intra_domain_crankback_done & h2 not expired)
    Decrement inter-domain counter h2, extract route vector R
    and exclude route vector X from PATH

    Add ingress inter-domain link to exclude link vector X

    Remove all nodes in route vector R up to previous
    domain's ingress border node

    Copy h2, R, X fields, reset h1=H1, generate PATH_ERR and
    send to previous domain's ingress border node
else
    Copy h1, h2, R, X fields,generate PATH_ERR,send to source
```

Sample crankback re-computation operation is also shown in Figure 1. For example when blocking occurs on link $l^{ii}{}_{42}$, the ingress border node v_1^i re-tries egress node v_3^i. When this second intra-domain attempt fails at the egress link $l^{ii+2}{}_{31}$, ingress node v_1^i receives a *PATH_ERR* with a zero h_1 counter. In response, it marks its ingress link $l^{i-1\,i}{}_{21}$ as congested, prunes the route to the ingress border node in prior domain *i-1*, v_1^{i-1}, and sends a *PATH_ERR* (step 3, Figure 1). This upstream node then initiates a re-try to a new egress border node v_3^{i-1} (step 4, Figure 1). Note that if the prior domain is the source domain, the *PATH_ERR* is simply sent to the source.

Overall, the above scheme limits the number of intra-domain attempts to H_1 per domain and the number of inter-domain attempts to H_2, i.e., maximum of H_1H_2. Also, note that future considerations can also incorporate delay metrics.

B. Next-Hop Domain Computation

As mentioned earlier, a key provision in the enhanced crankback scheme is the use of existing inter-AS state (i.e., BGP databases) to improve the search process. This is achieved by pre-computing a *multi-entry* distance vector table at all domain border nodes (or PCE) to list up to *K* next-hop domains/egress links to each destination domain. Namely, at domain *i*, the *k*-th table entry to a destination domain *j*, $T^i(j,k)$, is computed as the egress inter-domain link (to the next-hop domain) on the *k*-th shortest "domain-level" hop-count path to

domain j ($1 \le i, j \le D$, $i \ne j$, $1 \le k \le K$). Clearly the number of entries to a destination will be upper-bounded by the minimum of K and the maximum number of inter-domain links egressing from the domain. Now consider the actual computation of this table at a border node (or PCE) in domain i. The algorithm first uses the inter-AS (EGP) path database to build a "simple node" graph model [6] of the global network, $\boldsymbol{H}(\boldsymbol{U},\boldsymbol{E})$, where $\boldsymbol{U}$ is the set of domains $\{\boldsymbol{G}^i\}$ reduced to vertices and $\boldsymbol{E}$ is the set of physical inter-domain links $\{l^{ij}_{km}\}$, $i \ne j$. An iterative shortest-path scheme is then used to compute multiple routes to all other destination domains over $\boldsymbol{H}(\boldsymbol{U},\boldsymbol{E})$ as follows:

```
Generate simple-node abstraction of global topology via
EGP database information, i.e., H(U, E)
/* At domain i, loop across all possible destination domains */
for j = 1 to D
   if (j ≠ i)
      Make temporary copy of graph H(U,E), i.e., H'(U,E)
      /* Compute up to K table entries */
      for k=1 to K
         Compute shortest-path from domain i to j in H'(U,E)
         if (shortest path route found)
            Save route line from domain i in k-th table entry
            T^i(j,k), i.e., link from domain i vertice in H'(U,E)
            Prune above-selected link from H'(U,E)
         if (domain i becomes disconnected)
            break k-loop
```

The above scheme basically loops over all destination domains $j \ne i$ (index j) and computes up to K next-hop egress links (index k) over a temporary copy of $\boldsymbol{H}(\boldsymbol{U},\boldsymbol{E})$, i.e., $\boldsymbol{H'}(\boldsymbol{U},\boldsymbol{E})$. At the k-th iteration, the scheme computes the shortest "domain-level" hop-count path to the destination domain using $\boldsymbol{H'}(\boldsymbol{U},\boldsymbol{E})$, and if found, stores the egress link from the source domain in $T^i(j,k)$. This link is then pruned from $\boldsymbol{H'}(\boldsymbol{U},\boldsymbol{E})$ and the procedure repeated to compute the next shortest "domain-level" hop-count path. The procedure is terminated if all K entries are filled or the vertice for domain i in $\boldsymbol{H'}(\boldsymbol{U},\boldsymbol{E})$ becomes disconnected. Hence the next-hop domain selection procedure during crankback re-computation (Section III.A) simply searches these k table entries, $T^i(j,k)$, to a destination domain j in increasing order. This sequentially drives the crankback along fixed domain sequences of increasing length, but with provisions to avoid congestion (via $\underline{\boldsymbol{X}}$). Overall, these tables will be quite static if EGP state changes are infrequent.

IV. PERFORMANCE EVALUATION

The performance of the enhanced multi-domain crankback solution is tested by developing specialized models in *OPNET Modeler*™. A sample 10-domain topology is used, Figure 4, with 25 inter-domain links where each domain has about 10 interior nodes and about 4-5 border nodes. Moreover, multi-homed interconnection is done for some domains to reflect realistic settings. All link capacities are set to 10 Gbps and connection requests sizes are varied from 200 Mbps–1 Gbps in increments of 200 Mbps, i.e., fractional Ethernet. All connections are generated between random nodes in randomly domains. Each run is averaged over 250,000 connections and mean holding times are set to 600 sec (exponential). Meanwhile, request inter-arrival times are also exponential and varied with load. Finally, K=5 next-hop domain entries are computed in the distance vector table, although the number searched is limited by H_2.

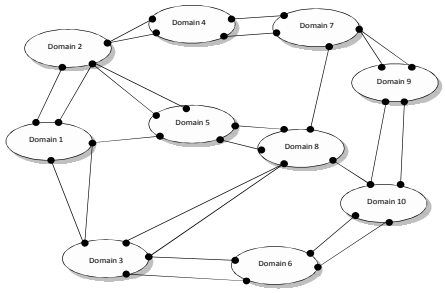

Figure 2: 10-domain topology (only border nodes shown)

A key objective here is to compare against hierarchical inter-domain routing with topology abstraction, i.e., simple node, full-mesh [6],[7],[11]. Consider the details. In full-mesh abstraction, the PCE computes "abstract links" to condense trans-domain routes, $O(|\boldsymbol{B}^i|(|\boldsymbol{B}^i|-1))$ state. Here, the capacity of an abstract link is derived as the mean bottleneck capacity of the k-shortest paths between the respective border nodes [11]. These links (along with physical inter-domain links) are then advertised using a second level of OSPF-TE between border nodes [1]. Namely, link updates are generated using *significance change factors* (SCF) and hold-off timers [1], and the respective values are set to 10% (SCF) and 200 sec (hold-off timer). This inter-domain link state is then used to build a "global" topology for computing/ expanding end-to-end *loose-routes* (LR). Meanwhile in simple node abstraction, all domains are condensed to virtual nodes and only *physical* inter-domain links are advertised. The exhaustive-search *per-domain* (PD) crankback scheme of [14] is also tested here.

First, the inter-domain *bandwidth blocking rates* (BBR) are plotted for the various schemes in Figure 3 ("HR" for hierarchical routing, "CB" for crankback, "PD" for scheme in [14]). Moreover, several configurations are tested for the enhanced crankback scheme, including intra-domain only (H_1=3/H_2=0), inter-domain only (H_1=0/H_2=2), and joint (H_1=2/H_2=2, H_1=3/H_2=3). Results show best crankback performance when running joint intra/inter-domain crankback, whereas inter-domain-only crankback gives the lowest results, e.g., H_1=0. More importantly, moderate intra/inter-domain crankback counts yield very competitive results. For example, BBR reduction tends to level off after H_1=2/H_2=2, and is notably better than that with the more exhaustive PD scheme [14] (no congested intra-domain link pruning or intelligent next-hop domain selection). Added runs (not shown) with the enhanced scheme for even higher crankback counts, e.g., H_1=5/H_2=5, indicate worsening BBR due to increased route lengths and bandwidth fragmentation. Meanwhile, when compared with hierarchical routing, the enhanced crankback scheme notably outperforms simple-node abstraction but not more advanced full-mesh abstraction. Nevertheless, associated crankback messaging overheads are over an order magnitude lower than hierarchical routing message loads.

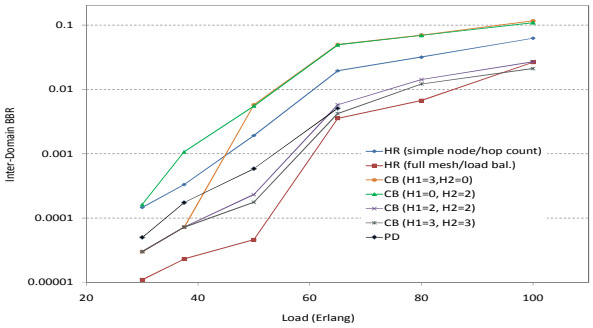

Figure 3: BBR (crankback, hierarchical routing)

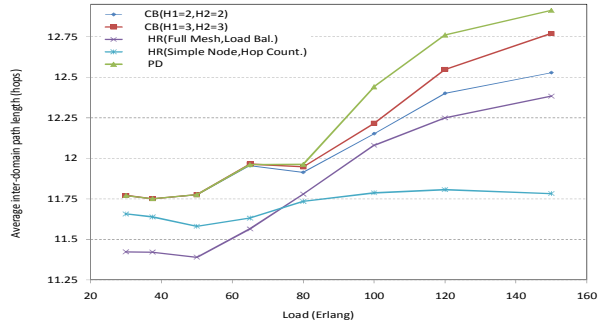

Figure 4: Average inter-domain path length

Next, the resource usage/efficiencies of the respective schemes are gauged by plotting the average inter-domain path lengths, Figure 4. Here, it is seen that increased *inter-domain* crankback (i.e., H_2=3, exhaustive PD search [14]) result in the highest utilizations, particularly at higher loads. However, lower levels of crankback (i.e., H_2=2) can almost match the usage levels of hierarchical routing with full-mesh abstraction. As expected, simple node abstraction has the lowest overheads of all (although its BBR performance is not optimal).

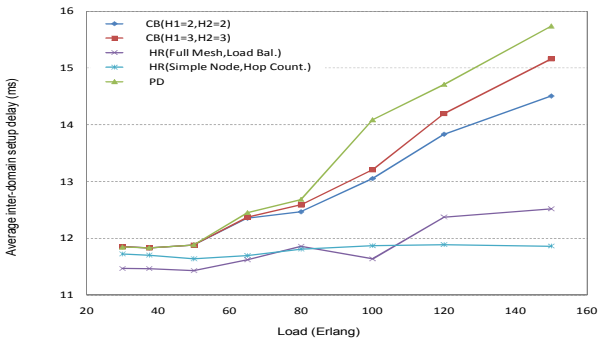

Figure 5: Average connection setup delay

Finally, connection setup delays are also plotted in Figure 4 (for 0.5 ms link delays and 0.05 ms node processing delays). Here, per-domain crankback yields15-20% higher levels versus hierarchical routing, which is generally acceptable. Overall, the above findings show that moderate levels of joint intra/inter-domain crankback driven by parsed distance vector state can yield very good overall performance.

V. CONCLUSIONS

This paper studies crankback signaling in multi-domain settings and introduces several key innovations. Foremost, improved next-hop domain selection strategies are introduced to drive the search process with minimal inter-domain routing overheads. In addition, a dual counter scheme is used to limit the number of intra/inter-domain crankback attempts. Finally, crankback history in the form of link congestion state is leveraged at both the intra and inter-domain levels. Results show much-improved blocking performance with the proposed scheme, i.e., as compared with more exhaustive "end-to-end" crankback schemes. In many cases, the performance even approaches that of more complex hierarchical inter-domain routing strategies (with topology abstraction). Future studies will look at extending this work for survivability support.

VI. ACKNOWLEDGEMENTS

This research has been supported by the US *Department of Energy* (DOE) Office of Science under Award ER25828 and the US *National Science Foundation* (NSF) under Award 0806637. The authors are very grateful for this support.

REFERENCES

[1] N. Ghani, *et al*, "Control Plane Design in Multidomain /Multilayer Optical Networks," *IEEE Communications Magazine,* Vol. 46, No. 6, June 2008, pp. 78-87.

[2] R. Zhang, J. Vasseur, "MPLS Inter-Autonomous Systems Traffic Engineering (TE) Requirements," *IEFT RFC 4226*, November 2005.

[3] J. Ash, J. Le Roux, "A Path Computation Element (PCE) Communication Protocol Generic Requirements," *IETF RFC 4657*, September 2006.

[4] A. Farrel, *et al*, "Crankback Signaling Extensions for MPLS and GMPLS RSVP-TE," *IETF Request RFC 4920*, July 2007.

[5] P. Torab, *et al*, "On Cooperative Inter-Domain Path Computation," *IEEE ISCC 2006*, Sardinia, Italy, June 2006.

[6] F. Hao, E. Zegura, "On Scalable QoS Routing: Performance Evaluation of Topology Aggregation," *IEEE INFOCOM 2000*.

[7] T. Kormaz, M. Krunz, "Source-Oriented Topology Aggregation with Multiple QoS Parameters in Hierarchical Networks," *ACM TOMACS,* Vol. 10, No. 4, October 2000, pp. 295-325.

[8] K. Liu, *et al*, "Routing with Topology Abstraction in Delay-Bandwidth Sensitive Networks," *IEEE/ACM Transactions on Networking*, Vol. 12, No. 1, February 2004, pp. 17-29.

[9] A. Sprintson, *et al*, "Reliable Routing with QoS Guarantees for Multi-Domain IP/MPLS Networks," *IEEE INFOCOM 2007*, Alaska, May 2007.

[10] S. Sanchez-Lopez, *et al*, "A Hierarchical Routing Approach for GMPLS-Based Control Plane for ASON," *IEEE ICC 2005*, Korea, June 2005.

[11] Q. Liu, *et al*, "Hierarchical Routing in Multi-Domain Optical Networks", *Computer Comm.* Vol. 30, No. 1, December 2006.

[12] Q. Liu, C. Xie, T. Frangieh, N. Ghani, A. Gumaste, N. Rao, T. Lehman, "Inter-Domain Routing Scalability in Optical DWDM Networks," *IEEE ICCCN 2008*, US Virgin Islands, August 2008.

[13] D. Truong, B. Thiongane, "Dynamic Routing for Shared Path Protection in Multi-Domain Optical Mesh Networks," *OSA Journal of Optical Net.*, Vol. 5, No. 1, January 2006, pp. 58-74.

[14] S. Dasgupta, J. C. de Oliveira, J. P. Vasseur, "Path-Computation-Element-Based Architecture for Interdomain MPLS/GMPLS Traffic Engineering: Overview and Performance," *IEEE Network*, Vol. 21, No. 4, July/August 2007, pp. 38-45.

[15] F. Aslam, *et al*, "Interdomain Path Computation: Challenges and Solutions for Label Switched Networks," *IEEE Communications Magazine,* Vol. 45, No. 10, Oct. 2007, pp. 94-101.

[16] F. Aslam, *et al*, "Inter-Domain Path Computation Using Improved Crankback Signaling in Label Switched Networks," *IEEE ICC 2007*, Glasgow, Scotland, June 2007.

[17] C. Pelssner, O. Bonaventure, "Path Selection Techniques to Establish Constrained Interdomain MPLS LSPs", *Proc. of IFIP Intn^l Networking Conference*, Coimbra, Portugal, May 2006.