An Adaptive Inter-domain PCE Framework to Improve Resource Utilization and Reduce Inter-domain Signaling

Mohit Chamania\textsuperscript{a}, Xiaomin Chen\textsuperscript{a}, Admela Jukan\textsuperscript{a}, Franz Rambach\textsuperscript{b}, Marco Hoffmann\textsuperscript{b}

\textsuperscript{a}Technische Universität Carolo-Wilhelmina zu Braunschweig, Germany, \{chamania, chen, jukan\}@ida.ing.tu-bs.de
\textsuperscript{b}Nokia Siemens Networks GmbH & Co. KG, Munich, Germany, \{franz.rambach,marco.hoffmann\}@nsn.com

Abstract

Upcoming broadband commercial and scientific applications are now demanding high bandwidth pipes across multiple domains with guaranteed Quality of Service (QoS). Recent research initiatives such as the Path Computation Element (PCE) framework are focusing on the development of scalable multi-domain QoS provisioning frameworks, especially within the emerging carrier grade transport technologies based on layer-2 tunnels. QoS provisioning across multiple domains requires that QoS parameters for available transit paths inside a domain be advertised in the inter-domain routing algorithms, while the dynamic inter- and intra-domain connections vary the available resource and hence require frequent inter-domain updates. The signaling load on the other hand hampers the scalability of the inter-domain routing mechanisms. We propose the use of an adaptive partitioning framework, which can effectively use network resources and at the same time stabilize the advertised domain topologies and thus path advertisements. Our method partitions network resources by pre-reserving resources for inter-domain transit traffic, and uses policies to modify the resource partitioning in order to maintain the available transit capacity between specified thresholds. We show by simulations that the proposed mechanism can reduce inter-domain signaling load by 10-20% and reduce overall blocking inside a domain by creating a trade-off between available resources for intra-domain connections and inter-domain transit connections. The reduction in inter-domain signaling and blocking can be used as a building block to design scalable QoS routing systems for carrier-grade transport networks.

Key words: multi-domain, QoS, PCE, adaptive partitioning
1. Introduction

Recently, the ever-growing demands of broadband commercial and scientific applications with QoS guarantees have fueled significant research and standardization efforts for the development of dynamic carrier grade transport technologies. Current transport technologies such as Multi Protocol Label Switching with Transport Profile (MPLS-TP) and Provider Backbone Bridging with Traffic Engineering (PBB-TE) can provision layer-2 tunnels with guaranteed QoS on the fly with inherent support for additional services such as OAM and path protection [1, 2]. Research initiatives are now focusing on the design of scalable path provisioning architectures for multi-domain carrier-grade transport networks. Of note is the Path Computation Element (PCE) [3] architecture, which can be used to compute optimal constrained paths across multiple domains. The PCE uses its Traffic Engineering Database (TED) to compute optimal paths inside a domain, and PCE’s of different domains can interact with each other using the PCEP [4] protocol to compute optimal inter-domain paths along a specified domain chain.

The domain chain required by the PCE can be pre-configured or obtained using inter-domain routing mechanisms and affects the quality of the path computed by the PCE. Pre-configuration of domain chains can only be done in small networks and inter-domain routing mechanisms are required for large multi-domain networks. Existing routing mechanisms in IP networks focus on scalability of the multi-domain routing architecture and therefore exchange only reachability information with other domains while revealing minimal information about transit resources inside the domain. The lack of transit path information makes these algorithms unsuitable for QoS routing, and extensions to popular routing schemes such as BGP [5] and topology aggregation schemes [6] have been proposed which advertise transit parameters to support QoS routing in multi-domain networks. However, in a dynamic system, frequent changes in the available transit resources due to set up/teardown of inter- as well as intra-domain connections can lead to frequent inter-domain updates which hamper the stability and consequently the scalability of the inter-domain routing system. Traditional techniques used to counter these drawbacks such as threshold or time driven update policies are also not sufficient as they can lead to outdated inter-domain information leading to connection blocking.

In this work, we exploit a trade-off between data plane resources and control plane signaling load to create stable inter-domain topologies. We partition the resources in the data plane used for inter-domain transit and intra-domain traffic so as to ensure that fluctuations in intra-domain traffic does not affect the available resources for inter-domain transit. The use of static partitioning of network resources can lead to over- or under-utilization of the reserved inter-domain resources and the available resources for transit inside a domain would

---

A short summarized version of this paper was presented at the Second Symposium on Advanced Networks and Telecom Systems (ANTS) 2008 in Mumbai, India, in December 2008.
still change with the set up/teardown of every inter-domain connection. We propose a new adaptive partitioning scheme which can be used to create stable inter-domain topologies. The adaptive partitioning scheme modifies the capacity available for transit so as to ensure that the available capacity for transit across a domain remains between pre-defined bounds and these bounds are then advertised via the inter-domain routing protocols. The use of the adaptive partitioning ensures that the pre-defined bounds are not violated frequently, and as the inter-domain routing protocol advertises these bounds, new inter-domain advertisements are only required when the bounds are violated. The advertised bound also give a deterministic range for the advertised capacity as compared to existing threshold or time triggered inter-domain advertisements, where the update policies of the remote domains are not known, and thus the error in the advertised capacity cannot be determined.

This paper is organized as follows: In section 2 we briefly discuss the related work. Section 3 describes the adaptive partitioning concept and the framework for incorporating adaptive partitioning in the PCE framework. The simulation study of the proposed framework is presented in section 4 and section 5 concludes the paper.

2. Related Work

The PCE architecture has been standardized for the GMPLS architecture with the capability to compute multi-domain as well as multi-layer paths. The PCEP protocol [4] has been proposed to allow clients to request paths from the PCE, and has also been extended to allow for multi-domain computation along a domain chain computed by the BRPC [7] algorithm. A reference architecture for multi-layer path computation using the PCE framework is presented in [8], where a Virtual Network Topology manager (VNTM) is introduced in the lower layer with a PCE to set up paths in lower layers when sufficient capacity is not available in the higher layers. The PCE architecture can also be deployed in a hierarchical architecture similar to the ASON[9] architecture. The hierarchical architecture is useful for small networks, where optimal inter-domain paths can be computed in a hierarchical fashion. However, given that each domain can have customized routing policies and have different confidential agreements with different domains, the hierarchical architecture is best suited for single provider multi-domain networks.

Pre-reservation of resources was first introduced in multi-service systems to ensure resources for higher priority services in the network. This concept was demonstrated in [10] where pre-reservation schemes were coupled with IP traffic prediction models. Pre-reservation of resources on inter-domain peering links between domains was also studied in [11, 12] where extra capacity was reserved between border nodes of neighboring domains to reduce inter-domain signaling. In this work, we focus on the advance reservation of layer-2 resources inside the domain. Resources are reserved in the form of Layer-2 tunnels, and all inter-domain connections are then routed through these tunnels. The use of Layer-2
tunnels allows for flexible control of the route to be used between border nodes inside a domain as well as the reserved capacity inside the domain.

3. Adaptive Advance Reservation and Inter-Domain Routing

Existing networks use one of three mechanisms, namely topology aggregation, path vector protocols or partial topology information dissemination for inter-domain routing [13] and each of these schemes can advertise transit resource information to neighboring domains in theory. In dynamic networks, the available transit capacity between a pair of border nodes can change due to dynamic inter/intra domain connections causing frequent inter-domain updates. Updates in inter-domain transit parameters trigger re-computation of optimal domain chains in remote domains, and a high frequency of inter-domain updates can hamper the stability of the inter-domain routing system. In our proposal, adaptive partitioning of data-plane resources in the form of Layer-2 transit tunnels is used to negate the effect of intra-domain traffic fluctuations on available transit resources. The adaptive partitioning schemes also attempt to maintain the available capacity in these transit tunnels between pre-configured bounds. These bounds are then included in inter-domain advertisements for transit capacity, and the adaptive partitioning schemes ensure that these bounds are not violated frequently, thus negating the need for frequent inter-domain advertisements in a dynamic network scenario. We now describe the different aspects of the proposed architecture in detail.

3.1. Adaptive Partitioning for Transit

In our proposal, we attempt to partition network resources along defined paths between border nodes for use for intra- and inter-domain connections respectively. Typical networks use either static partitioning of resources or do not partition resources for inter-domain transit connections. Partitioning has the advantage of reducing the effect of dynamic intra-domain connections on the resources available for inter-domain transit. However, static partitioning can lead to inefficient data plane utilization, and available inter-domain resources are still affected by dynamic inter-domain connections. We employ an adaptive partitioning scheme which attempts to maintain a stable inter-domain topology by maintaining the available transit capacity between specified bounds. Adaptive partitioning is implemented by pre-reserving capacity between border nodes in the form of layer-2 tunnels. The network of layer-2 tunnels between the border nodes in a domain forms a virtual mesh topology which is the abstract topology used for inter-domain transit. The reservation of capacity in advance ensures that available transit capacity is not affected instantaneously by intra-domain traffic and all inter-domain connections are reserved inside these tunnels. To counter the effects of dynamic inter-domain connections, we modify the capacity of the transit tunnels after inter-domain connection set up/teardown to maintain the available capacity in the tunnel between pre-defined thresholds ($min, max$). The $min$ threshold indicates the minimum amount of capacity always available
for inter-domain transit, and a tunnel capacity increase is triggered whenever the available capacity falls below the min threshold. The max threshold is introduced to release excess available capacity in the transit tunnel to be used by intra-domain resources, and a tunnel capacity decrease is triggered whenever the available capacity in the transit tunnel increases beyond the max threshold.

3.2. Inter-domain Advertisements

In our proposal, we advertise the thresholds for available capacity with the total available capacity along with other possible QoS parameters such as delay, jitter limits, number of hops etc. The bandwidth bounds correspond to the min and max thresholds set in the adaptive partitioning policy and the use of pre-reserved tunnels implies that QoS bounds on parameters like delay and number of hops for transit connections are known. The adaptive partitioning mechanism attempts to maintain the available capacity between the specified thresholds and therefore fewer inter-domain capacity updates are required as compared to traditional advertisements which advertise actual available capacity instead of capacity bounds. Traditional systems typically use time triggered or threshold triggered update policies in order to control the triggering rate. However, as this policy is not known to other domains, the accuracy of these measurements cannot be determined, whereas the advertised capacity bounds in our method can be used by remote domains as a probability measure for the success of a connection set up.

3.3. Policies for Advance Reservation

Policies are used to trigger capacity updates for the transit tunnels. These policies have the objective of maintaining the available capacity of the tunnel between the specified thresholds. In this work we use three different types of policies for the control of transit tunnel capacity triggering, which are now described in detail.

Basic Triggering Policy

The basic triggering policy triggers a capacity update whenever the available capacity violates the min or max thresholds and attempt to modify the capacity of the transit tunnel to a fixed reset value. The reset value is configured by setting \( \alpha = 0.5 \) in Algorithm 1. The capacity available inside the tunnel is defined by the variable \( C_T \), and in order to increase this capacity, we reserve additional capacity for the tunnel. The maximum additional capacity is given by the available capacity on the links along the tunnel path and is defined as \( C_{\text{path}} \). If the available capacity along the path is not enough, the policy greedily reserves capacity for the transit tunnel in order to bring the available capacity as close to the reset capacity as possible. The performance of the policy is determined by its ability to keep the available capacity between the thresholds, the effect in terms of blocking on intra- and inter-domain traffic and the frequency of capacity update triggers. The parameters form a non-trivial trade-off, as the range of acceptable available capacity can affect the frequency of capacity update triggers as well as the intra-domain blocking. A very high
frequency of triggers is not desirable as it disrupts other connections in the switch, and can technically be reduced by increasing the range of acceptable available capacity in a tunnel. However, a higher range can lead to more intra-domain blocking as more resources may be committed to the transit tunnels. Similarly the choice of the min threshold can influence the stability of the inter-domain system, with the probability of a policy threshold violation being less likely with a lower min threshold. We study the effects of the various parameters further in the results section.

Algorithm 1: Threshold based policy to trigger capacity updates

Parameters: The current available capacity on tunnel $C_T$, current available capacity along tunnel path $C_{path}$, $\alpha$, ($0 \leq \alpha \leq 1$)

$$reset = min + \alpha \ast (max - min)$$

if ($C_T < min$) then
  if ($C_{path} > (reset - C_T)$) then
    Increase tunnel capacity by ($reset - C_T$)
  else
    Increase tunnel capacity by $C_{path}$
  end
else if ($C_T > max$) then
  Decrease tunnel capacity by ($C_T - reset$)
end

Aggressive Reservation/Release Policy

In the basic triggering policy, as the reset capacity is set to the mean of the min and max thresholds and thus the capacity increase in case of the min threshold violation is equal to the capacity decrease in the case of a max threshold violation. If however, the reset capacity was set closer to the max threshold, the policy would reserve more capacity when the min capacity is violated as compared to the capacity released when the max threshold is violated. Such a policy would exhibit aggressive reservation as compared to the basic policy, while a policy with reset capacity set closer to the min threshold would reduce transit tunnel capacity more aggressively. The reset point in this policy is therefore defined as shown in Algorithm 1, where $\alpha$ is a constant introduced to control the behavior of the policy. For a value of $\alpha > 0.5$, the policy exhibits aggressive reservation while $\alpha < 0.5$ indicates aggressive de-reservation.

3.4. PCE Based Inter-domain Architecture

Fig. 1 depicts the extended PCE based architecture to support advance reservation. The traffic engineering database (TED) used by the PCE to compute paths is divided into the intra-domain TED and the inter-domain TED. The PCE uses the intra-domain TED to compute paths inside a domain, while PCEs of different domains talk to each other and use the inter-domain TED to compute transit paths between border nodes for inter-domain requests. Inter-domain routing advertisements can be initialized at the border nodes as seen typically in the case of BGP or can also be initialized by the PCE which has a complete view of the transit topology of the domain.
The Transit Tunnel Manager (TTM) is introduced in the architecture to govern the initialization and operation of the transit tunnels inside a domain. As seen in Fig. 1, the TTM collects information about the status of the transit tunnels from the corresponding border nodes, and is responsible for triggering capacity updates determined by the corresponding policies. The transit tunnels are established in the form of Layer-2 tunnels between border nodes with reserved capacity, and the border node at the ingress is responsible for admitting incoming inter-domain connections in these tunnels. It is assumed that the control plane in the domain runs a layer-2 routing protocol such as OSPF-TE or IS-IS to gather link state information, and this information is used to update the intra-domain TED. However, in order to avoid any intra-domain traffic on the transit tunnels, the transit tunnels are not advertised as forwarding adjacencies in the intra-domain routing protocol. The border nodes send transit tunnel updates to the TTM after every inter-domain path set up/tear-down in the form of SNMP messages to the TTM which is responsible for updating the inter-domain TED in the PCE. The PCE then uses this information to choose candidate transit tunnels during inter-domain path computation requests. On the event of a policy violation, the TTM can send capacity reserve/release commands to the corresponding ingress border nodes, which can then initiate signaling in the control plane to update the transit tunnel capacity. While capacity release can be triggered directly, the TTM consults the PCE to verify capacity availability before triggering capacity increase.
4. Results

In this section, we study the performance of the proposed PCE based inter-domain provisioning architecture using a Java based event-driven simulator. We compare the performance of our method with the no-partitioning schemes in terms of inter-domain update frequency and blocking of intra-domain connections. While current carrier networks typically over-provision their network to ensure blocking-free operation, we present blocking results as they indicate the relative performance of the adaptive partitioning scheme.

In order to understand the performance of the various factors affecting the performance of the proposed policies, we introduce inter-domain transit load as well as intra-domain load in a single domain modeled on the NSFNet [14] topology. Each link in the network is assumed to be 100 Gb/s Ethernet. Connections for both intra- and inter-domain requests arrive according to a Poisson process with exponentially distributed holding times and an average holding time of 0.5 days for all connections. Intra-domain connection requests are assumed to be uniformly distributed over all possible (source, destination) pairs while transit connections requests are randomly distributed over (source, destination) pairs from the set of four border nodes (Seattle, Houston, Ann Arbor, Princeton) which are connected in a full mesh by transit tunnels. Each intra-domain connection requests a 1Gb/s connection while bandwidth demand for a transit connection is randomly distributed between 1 - 5 Gb/s with 1 Gb/s granularity.

In the first study, we observe the performance of the proposed system implementing the basic triggering policy with the traditional no-partitioning systems. In order to support QoS routing, only paths advertised for inter-domain transit can be used for incoming transit connections, and therefore the paths used for transit are fixed beforehand, with both the no-partitioning system and the adaptive partitioning system using the same paths in the domain. Fig. 2 presents the performance of different adaptive partitioning policies with increasing min threshold with the range of the policy (max − min) kept constant. As can be seen in Fig. 2(a), it is clear that the adaptive partitioning schemes create a trade-off between available data plane resources for intra-domain and transit connections. Policy based schemes show significant performance increase in transit blocking, which increases with the min threshold. Reduction in transit blocking is also accompanied with increased intra-domain blocking as seen in Fig. 2(a). We compare the ratio of inter-domain capacity updates generated by the adaptive partitioning scheme as compared to the traditional threshold based scheme with no partitioning in Fig. 2(b). The threshold variation for the traditional scheme is set to \((\pm (\frac{\text{max} - \text{min}}{2}))\) to match with the thresholds of the adaptive partitioning schemes. It can be seen that the adaptive partitioning schemes significantly reduce inter-domain updates at high loads, and the percent reduction in the number of updates increases with the increase in the min threshold of the basic partitioning policy. At low loads, all adaptive partitioning schemes reduce the inter-domain signaling load by approximately 10%. The total connection blocking observed in the network (intra-domain + transit) is
shown in Fig. 2(c), where it is observed that the adaptive partitioning schemes reduce the total connection blocking by reserving capacity in advance for the transit connections. However, this parameter is dependent on the relative ratio of the transit and the intra-domain loads and in cases with very low transit loads, it is seen that the adaptive partitioning schemes over-provision capacity for transit and therefore exhibit higher overall blocking.

In the next study, we compare the performance of basic triggering policies with the same min threshold and different ranges \((\text{max} - \text{min})\). As can be seen from Fig. 3(a), the intra-domain blocking increases and transit blocking decreases with the increase in range. A bigger range also implies that a policy is less likely to be triggered and therefore at high loads, a policy with lower range can get triggered more frequently. A higher frequency of triggering may be advantageous in racing conditions between intra-domain and transit capacity at high loads and some cases with lower range therefore exhibit better transit blocking performance than policies with higher range as seen in Fig. 3(a), where the policy with \((\text{min}, \text{max})\) thresholds set to \((2\text{Gbps}, 8\text{Gbps})\) exhibits lower transit blocking at some higher load scenarios. The blocking performance in these cases is influenced by both range as well as the frequency of triggering of the adaptive partitioning scheme. The total blocking (Fig. 3(c)) observed also shows a non-uniform trend at higher loads, while at lower loads, the total blocking generally reduces with the increase in range. The non-uniform effect at high loads is seen due to the existence of a racing condition between intra-domain connections and the capacity increase for the transit tunnels. We also compare the number of advertisements with no partitioning policies in Fig. 3(b). As expected, the signaling loads decreases with the increase in the threshold size of a domain. As the range of the adaptive partitioning policies vary in this study, we compute the average inter-domain signaling for no-partitioning schemes with different thresholds corresponding to the ranges of different policies \((\pm (\text{max} - \text{min})/2)\), and use these values to compute the signaling load ratio for the adaptive partitioning schemes with the no-partitioning schemes. It is observed that the signaling performance first increases with the increase in range in the case of policy \((2, 10)\), and then degrades with the increasing threshold size. The decrease in performance of the adaptive partitioning scheme with a high range is observed due to existence of racing conditions, which degrade the performance of adaptive policies with high ranges.

We also study the performance of the aggressive reservation/release policy with different values of \(\alpha\) for a constant \((\text{min}, \text{max})\) threshold pair. The constant \(\alpha\) determines if the sensitivity of the proposed partitioning policy towards reservation/release of capacity from the tunnel, and therefore influences the rate at which capacity is released/reserved. As can be seen from Fig. 4(a), the transit blocking reduces with increase in \(\alpha\) which makes the policy more aggressive towards reservation, while intra-domain blocking is seen to reduce with decrease in \(\alpha\). The inter-domain signaling rate is approximately equal at low loads, but is seen to decrease with increasing \(\alpha\) at high loads, as a higher value of \(\alpha\) can reserve capacity more aggressively, and consequently leading to fewer cases where available capacity in the tunnel is below the \(\text{min}\) threshold. It is also observed
(a) Intra-domain and transit connection blocking measured against increasing load expressed as (intra-domain, transit) load in Erlangs.

(b) Normalized inter-domain routing advertisement rate measured against increasing load expressed as (intra-domain, transit) load in Erlangs. Inter-domain routing advertisement rate is normalized against the measured rate for the No partitioning policy.

(c) Total connection blocking measured against increasing load expressed as (intra-domain, transit) load in Erlangs.

Figure 2: Performance of basic triggering policy with increasing min threshold. Value pairs on the X axis in figures indicate (intra-domain, transit) load in erlangs, while the basic triggering policies are indicated by the (min, max) thresholds in Gbps.
(a) Intra-domain and transit connection blocking measured against increasing load expressed as (intra-domain, transit) load in Erlangs.

(b) Normalized inter-domain routing advertisement rate measured against increasing load expressed as (intra-domain, transit) load in Erlangs. Inter-domain routing advertisement rate is normalized against the measured rate for the No partitioning policy.

(c) Total connection blocking measured against increasing load expressed as (intra-domain, transit) load in Erlangs.

Figure 3: Performance of basic triggering policy with increasing max threshold and fixed min threshold. Value pairs on the X axis in figures indicate (intra-domain, transit) load in erlangs, while the basic triggering policies are indicated by the (min, max) thresholds in Gbps.
(a) Intra-domain and transit connection blocking measured against increasing load expressed as (intra-domain, transit) load in Erlangs

(b) Normalized inter-domain routing advertisement rate measured against increasing load expressed as (intra-domain, transit) load in Erlangs. Inter-domain routing advertisement rate is normalized against the measured rate for the No partitioning policy

(c) Total connection blocking measured against increasing load expressed as (intra-domain, transit) load in Erlangs

Figure 4: Performance of Aggressive reservation/release policy for different values of $\alpha$. Value pairs on the X axis in figures indicate (intra-domain, transit) network load in erlangs, while the Aggressive reservation/release policies are indicated by the $(\text{min}, \text{max}, \alpha)$ tuples with thresholds in Gbps.
that while a very low $\alpha$ (0.2) can lead to marginally higher overall blocking, the overall blocking does not vary significantly with the variation in $\alpha$.

From this study, we see that the proposed adaptive partitioning schemes can significantly reduce inter-domain signaling and reduce overall blocking, and reduce inter-domain blocking for a minor increase in intra-domain blocking. Parameters such as range and $min$ threshold can be used by operators to identify the target operational range, and $\alpha$ can be used as a parameter for fine adjustments in the transit/intra-domain blocking values without significantly disturbing the overall blocking of the system. It was observed that adaptive partitioning policies with high ranges could suffer from decreased performance at high loads, and aggressive triggering mechanisms using a high $\alpha$ could be used to boost performance of these schemes.

5. Conclusion

We proposed a new inter-domain PCE framework with adaptive partitioning of domain resources using layer-2 tunnels. We demonstrated via simulations that the proposed adaptive partitioning schemes create a trade-off between intra-domain and transit connection blocking, and can be used to decrease overall blocking in the network while reducing inter-domain path advertisements by 10 - 20 percent. We proposed two policies to control the adaptive partitioning scheme and studied the performance of the partitioning schemes for different parameters which are essential in the design of the proposed framework. The results show that the proposed system can be used as a building block for inter-domain provisioning frameworks with QoS support.

References


