Distributed Explicit Partial Rerouting (DEPR) Scheme for Load Balancing in MPLS Networks

Sherif Ibrahim Mohamed
shf_ibrahim@yahoo.com

Khaled M. F. Elsayed, *senior member IEEE*
khaled@ieee.org

Department of Electronics and Communications Engineering
Faculty of Engineering, Cairo University, Giza, Egypt 12613

Abstract

Traffic engineering is one of the important enhancements introduced by the deployment of Multiprotocol Label Switching (MPLS) into IP-networks. Minimizing congestion is a central goal of traffic engineering as it degrades the overall network performance. We propose the Distributed Explicit Partial Rerouting (DEPR) scheme as a reactive congestion removal for MPLS networks. DEPR removes congestion by rerouting the LSPs crossing the congested link to a partial route set up around the congested link (unlike other methods that re-establish the full route from the source to the destination). The scheme is shown to adapt quickly to network dynamics and to scale well with large networks. It also enhances the overall network throughput and delay performance.

1. Introduction

The IP protocol was designed from the beginning with interconnection in mind, thus offering flexibility and scalability [1]. IP networks provide a very simple best effort service. Therefore, IP networks need to be enhanced in the areas of availability, dependability and quality of service (QoS) [2] in order to provide mission critical network environments.

IP networks are enhanced with the deployment of Multiprotocol Label Switching (MPLS) [3, 4]. MPLS introduces a new connection-oriented forwarding paradigm in IP networks based on fixed length labels. It allows sophisticated routing control capabilities, fast forwarding, QoS support, reliability, and traffic management to be introduced into IP networks. MPLS provides IP networks with the ability to forward packets over arbitrary non-shortest paths (explicit routes) and to provide high speed "tunnels" [5]. MPLS introduces traffic engineering [6, 7] capabilities to IP-based networks. The main objective of traffic engineering is to optimize the network performance through an efficient utilization of network resources.

In this paper we assume the context of an MPLS-based IP network and develop the DEPR scheme for MPLS networks. The main concept in DEPR is to remove congestion by rerouting the LSPs passing through the congested link to a partial route set up around the congested link(s). This is in contrast to other methods that re-establish a full route from the source up to the destination. The algorithm is shown to adapt quickly to the network dynamics and to scale well with large networks.

2. Related Work

Gallager [8] developed the theory for perfect load balancin g through the formulation of the minimum delay routing problem (MDRP). Some heuristic techniques were developed to approximate the Gallager perfect load balancing conditions in an attempt to increase the possibility of practical adoption into production networks.

Within the context of MPLS networks, several methods were developed to remove congestion and load balance the network, similar to minimum interference routing algorithm (MIRA) [9], dynamic load balancing algorithm (DYLBA) [10], MPLS adaptive traffic engineering (MATE) [11] and fast acting traffic engineering algorithm (FATE) [12]. However, these techniques focused on load balancing by shifting traffic from congested links to less utilized links using centralized rerouting. These techniques share the property that rerouting is done based on end-to-end basis which means rerouting is done completely for all of the path starting from the source up to the destination.

For large networks this would be inefficient due to the following reasons:

- Establishing a new alternate route from the source node up to the destination requires the whole topology state information to be available at the source nodes making the solution non scalable.
- False rerouting at source nodes would occur due to incurred delays in transferring network state information to source nodes
- Slow adaptation to congestion especially with fast network dynamics.

3. The DEPR Congestion Removal Scheme

The main idea of load balancing is to shift traffic from the high congested links to the low congested or underutilized links. Widest path algorithms tend to prefer shifting traffic towards the path with the maximum free bandwidth without taking link capacities or utilizations into consideration. This would congest the path to which the traffic is shifted.

The connection-oriented nature of MPLS networks makes route establishment take some time (set up time) which is proportional to the path length. This neither scales well with large networks nor fits well with network dynamics rapid changes.

For speeding reaction to congestion, all the network nodes (including the core nodes) participate in the DEPR scheme. DEPR runs in a distributed manner where each node periodically monitors the congestion of its outgoing links. We consider links to be monitored in a unidirectional manner in the sense that link (i,j) is different from link (j,i). In case multiple direct links are found congested, the DEPR scheme starts with the most congested link, then goes sequentially to the next successor congested link.

Definitions:

- *The exploration region*: is the network part where a given node explores all the possible alternate paths for the LSPs passing through the congested link. The maximum path length around the congested link is defined as the exploration region diameter.
- *Reacting node*: is the upstream node of the congested link that detects the congestion and reacts to it.
- *Partial rerouting path (PRP)*: is the alternate path established by the reacting node for traffic rerouting.
- *Label information base (LIB)*: is a table used at each MPLS node for labelling packets before being forwarded.
- *Congestion threshold*: the value of link utilization beyond which a link is considered congested.
- *Upstream and downstream nodes*: are respectively the start and the end nodes of the congested link.

DEPR could be applied either on the flow level or the LSP level, LSP level will be used as it would be more practical.

We explain the DEPR main idea by the comparative example in Figure 1. Assume link 5-6 has become congested. Instead of rerouting a passing through LSP traffic from its source node (1) to its destination node (10) using the full alternate route (1-11….18-10), the DEPR scheme at the reacting node (5) reroutes part of that LSP (5-6-7) around the congested link using a partial alternate route (5-19-20-7). This is why it is called a partial rerouting scheme.

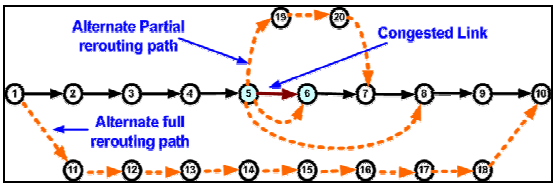


Figure 1: Full and partial rerouting

Another aspect that makes partial rerouting more attractive is that shifting traffic away from the congested link may have bad impact on the links of the new alternate path, even if it appears better. It is clear that, the less the length of the alternate path, the smaller the potential bad impact on the network links. In addition, it is not necessary to have the full network state information at each node but only within the node's exploration region. The diameter of the exploration region depends on the required length of the discovered alternate routes (1, 2, or 3 hops). This leads to faster reaction than other reactive schemes. Also, more accurate rerouting is possible due to less incurred delays in transferring the network state information. Additionally, it scales well with large networks.

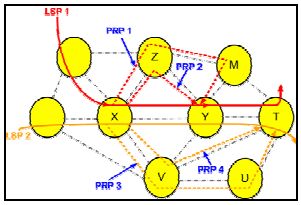


Figure 2: DEPR rerouting using methods 1 and 2

3.1. Methods for Selection of the Partial Re-routing Paths

When a reacting node decides to perform partial re-routing around a congested link, it is possible to have various ways of selecting the end node of the alternate path. It is necessary for the reacting node to know the original successor nodes over which the rerouted LSP passes on its way to the final destination, so that it could continue its original path at that node. Consider the example in Figure 1. Assume (5-6) is a congested link through which LSP passes and that the destination of the LSP is node 10. One of the nodes (6, 7, 8, 9, 10) could be selected to be the end node where the partial rerouting path (PRP) ends.

For simplicity and fast reaction to congestion the PRP is taken to end at nodes at a maximum length of 2 hops downstream the original path from the reacting node, this is to limit the signaling information to both the end nodes of the congested link. In addition, the maximum length of the PRP is limited to 4 hops to avoid wasting resources. With these constraints, the valid PRP's are [5-6] and [5-

19-20-7]. Accordingly, we have two methods for partial rerouting.

3.1.1. Method 1

In this method both the origin and destination nodes of the partial rerouting path should be the two end nodes of the congested link as shown in Figure 2 for PRP 1 and PRP 2 where link (X-Y) is the congested link.

The PRP maximum length is set to 4 hops to increase the probability of finding non congested alternate path. The longer the PRP the more network resources are consumed and the more delays that would be introduced. This is why we limit the maximum length of the PRP to 4 hops. If multiple alternate paths exist then the shortest is the best.

3.1.2. Method 2

In this method one of the edge nodes of the PRP is the congested link upstream node, and the other node is a neighbor node of the congested link downstream node with the constraint that it lies on the original path. As shown in Figure 2, link X-Y is congested in direction X-Y, we say X is the upstream node and Y is the downstream node, T is a neighbor node of the downstream node Y and lies on LSP 2 original path, so both of PRP 3 and PRP 4 are valid

We prefer to use method 2 rather than method 1 because of the lower consumed resources and the less incurred delays.

3.2. Load Balancing

The main idea of load balancing is to shift traffic from the high congested links to the low congested or underutilized links. By dividing the value of the shifted flows traffic rate by the PRP links capacities provides the potential increase in their utilizations after the traffic is shifted to them. Link utilization in addition to the link free bandwidth gives a good measure for deciding traffic shifting.

In DEPR each node monitors the utilization of its direct links. If a link is found congested, the node establishes a partial path for rerouting some LSPs traffic. If multiple partial paths exist, then the preference is given to the one with the lowest utilization.

Before shifting traffic from the congested link to the PRP, it is necessary to compare the utilization of the congested link to a value expressing the utilization of the whole partial rerouting path which may be composed of multiple links each of a different utilization value. Accordingly we identify the following:

- A congested link is a path of 1 hop
- Path utilization is taken to be the maximum link utilization among its constituent links.

3.2.1. Partial Rerouting Path Eligibility Rule

A PRP is considered eligible if PRP Utilization < Congested link utilization. However, the above condition does not necessarily mean that the PRP (to which traffic is shifted) is not congested (i.e. we could have PRP Utilization > Threshold). The above condition increases the possibility to find a valid PRP even if congested as explained in the next section.

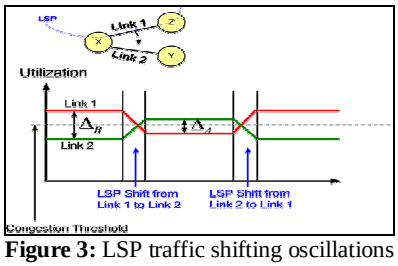


Figure 3: LSP traffic shifting oscillations

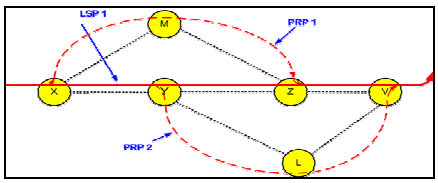


Figure 4: Example showing autonomous synchronization of nodes

3.2.2. Load Balancing and Oscillation Avoidance

As show in Figure 3, link 1 is congested while link 2 is underutilized. Node X decides to shift the LSP traffic from link 1 to link 2. Due to LSP traffic shifting, link 2 becomes congested while link 1 becomes non-congested. After sometime node X will decide to take remedial action and shift the traffic (assuming same LSP) back to link 1. These oscillations degrade the network performance. Figure 3 illustrates the oscillations of the LSP traffic shifting between the links.

To avoid oscillations, let us define the following symbols:

$\Delta U_{before}^{link1-link2}$ The absolute utilization difference between link 1 and link 2 before traffic shifting action.

$\Delta U_{after}^{link1-link2}$ The absolute utilization difference between link 1 and link 2 after traffic shifting action.

To prevent oscillations, rerouting should be performed under the following condition:

$$\Delta U_{after}^{link1-link2} < \Delta U_{before}^{link1-link2} \quad (1)$$

This condition not only prevents the occurrence of oscillations but also ensures good load balancing. It means traffic shifting occurs only if the difference between the links' utilization becomes smaller (indicating loads are balanced).

In the general case, link 2 in the above example is replaced with partial rerouting path (PRP) so equation (1) takes the general form:

$$\Delta U_{after}^{link1-PRP} < \Delta U_{before}^{link1-PRP} \quad (2)$$

3.3. Handling of LSP Labels

Through this section, we will refer to the term LSP as the original path used by the packets before being rerouted using a PRP. Switching packets from the LSP to the PRP necessitates some changes in the nodes LIB table. The general idea of the DEPR scheme is to be distributed, scalable and fast reactive to congestion. That is why our objective is to restrict the DEPR changes or reactions to only the end nodes of the congested link keeping the PRP end node away from these changes.

To do that, the packets at the reacting node should carry a stack of two labels, the outer label forwards them along the PRP, while the inner label is used only at the PRP end node to forward these packets along their normal way of the unchanged LSP part.

3.4. Node Synchronization

Although the DEPR scheme works in a distributed manner, it does not require synchronization between nodes when establishing PRPs. We will show through the following example how the synchronization problems would occur and solved autonomously. The problem arises when two successive nodes decide to reroute the same LSP.

Assume in Figure 4, the links X-Y and Y-Z are congested. According to the DEPR scheme, nodes X and Y detect the congestion and establish PRP 1 and PRP 2 respectively to reroute LSP1. However, this will not cause a problem as discussed below.

i) Case node Y switched traffic before X

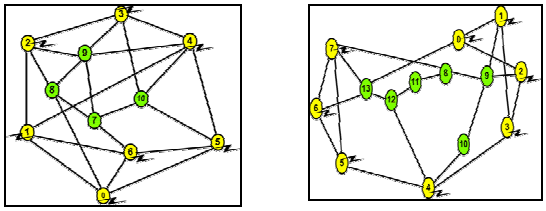
In this case, only the LIB tables of node Y are changed to implement the forwarding. When node X starts to switch the LSP traffic to PRP 1, the LIB table of nodes Z and V are still valid, therefore PRP 1 packets could be forwarded at Z. Node Y will have useless entries for the LSP and PRP 2, which will not affect the overall operations.

ii) Case node X switched traffic before Y

In this case, node Y will not affect node X and will have useless LIB entries as before. We notice that it is always the case that a node dominates its downstream nodes. This behavior is advantageous because as in the above example node X reaction would relief the congestion of the two links.

To remove the useless entries resulting from such behavior, the reacting node should send a CLEAR message to its downstream node. The downstream node should remove the LSP related entries if and only if it has already established a PRP for that LSP. This ensures leaving all necessary related LSP entries at node Z; otherwise PRP1 traffic will have a problem upon arriving at node Z.

4. Performance Evaluation



(a) COST239 Network Topology **(b) NSFNET Network Topology**

Figure 5: The simulated network topologies

We evaluate the performance of the DEPR scheme using discrete-event simulation. The simulations are carried out using the COST239 and NSFNET network topologies shown in Figure 5. We use the NS2 [13] network simulator tool to study and obtain the results of applying DEPR scheme and compare them with the performance of shortest path algorithms. The links are all bidirectional with equal capacity of 1Mbps, equal cost metric of a value 1, equal propagation delay of a value 10 μsec, equal buffer size of 1000 bytes, and about 50% of topology nodes are responsible for generating and receiving traffic as specially marked with a in Figure 5. The propagation delays are chosen to be small so as to study the effect of DEPR scheme on the queuing delays.

Multiple connections are selected to arrive at the network in accordance with a Poisson distribution with an average inter-arrival time of 2 seconds. For each arrival event, multiple connections are generated and started at the same event time. The number of connections arriving at the network per each event is randomly selected between 5 and 9 connections with uniform distribution. The source-destination pair per each connection is randomly generated from the allowed nodes following a uniform distribution.

Each connection lasts for a certain time uniformly distributed between (15, 20) sec. The traffic type of the connections is UDP traffic with an average rate that is uniformly distributed in between (90, 180) kbps. Each connection traffic pattern follows a Pareto distribution with a mean burst time of 0.5 sec, a mean idle time of 0.05 sec, and a packet size of 200 bytes. During the simulation the congestion threshold is set to be 70%.

To evaluate the efficiency of the delay enhancement due to using DEPR compared to the shortest path algorithms, equation (3) is used.

$$l_x^D = \frac{D_x^{sh} - D_x^{DEPR}}{D_x^{sh}} \qquad (3)$$

where, x is the network input load in Mbps, l_x^D is DEPR system delay efficiency at network input load x, D_x^{sh} is system delay using shortest path algorithms at input load x, and D_x^{DEPR} is system delay using DEPR scheme at input load x. To evaluate the DEPR gain with respect to system throughput compared to the shortest path algorithms, equation (4) is used.

$$l_x^w = \frac{w_x^{DEPR} - w_x^{sh}}{x} \qquad (4)$$

where, l_x^w is DEPR system throughput efficiency at network input load x, w_x^{sh} is system throughput using shortest path algorithms at input load x, and w_x^{DEPR} is system throughput using DEPR scheme at input load x.

Figures [6-(a), 7-(a)] illustrate the DEPR system delay reduction percentage with respect to COST239 and NSFNET topologies at different input loads. For the COST239 topology, it ranges between 30% and 70% and for the NSFNET topology it ranges between 15% and 45%. It is evident that there is a good enhancement in the overall system delay when using DEPR scheme. It is cleat that the COST239 results are better than in the NSFNET since the former topology is more connected.

Figures [6-(a), 7-(b)] plot the DEPR throughput gain percentage with respect to both COST239 and NSFNET topologies at different input loads, where it reaches 24% at higher loads for COST239 topology and 15% for the NSFNET topology. The gain increased due to the fact that load balancing has reduced the dropped packets.

4.1. Effect of Changing DEPR Congestion Threshold

Figures [8-(a), 8-(b)] plot the DEPR system delay reduction and the system throughput gain with different congestion threshold values (30%, 50%, 70%) for the COST239 topology. The benefit of making DEPR react at lower values of link congestion is to study DEPR load balancing in the normal state of the network without links being congested. It is evident that there is a slight difference in both the system delay reduction and the system throughput gain indicating no need for reaction as long as the link is not heavily utilized. At the same time this reduces the processing load of the scheme. When congestion threshold is 30% or 50%, DEPR performance is superior to 70% threshold. The performance at 30% and 50% thresholds are very close to each other.

4.2. Effect of Changing Monitoring Interval Length

Figures [9-(a), 9-(b)] plot the DEPR system delay reduction and the system throughput gain for the COST239 topology using different monitoring intervals. The monitoring interval is the periodic time at which the DEPR scheme at the node scans the surrounding links to react to congestion. Different monitoring intervals are selected (2 seconds, 10 seconds, and 20 seconds). It is evident from both figures that there is a big difference in both the delay reduction and the throughput gain when the monitoring interval is varied. The fast dynamic nature of the network traffic requires the increase of the periodic monitoring rate so as to react to congestion at the proper time, increasing both the delay reduction and the throughput gain.

5. Conclusions

We presented the DEPR scheme for load balancing in MPLS networks. The results showed the great improvements to both the overall system throughput and the overall system delay in comparison with the shortest path algorithms.

We showed that DEPR has many advantages over other techniques such as its ability of fast reaction to congestion due to the new introduced partial rerouting technique. DEPR is also scalable and can be easily deployed in large networks. DEPR performance depends on the network connectivity degree; the more the network is connected the higher the performance. The diameter of the node's exploration region controls the DEPR load balancing speed of conversion. The longer the diameter the lower the reaction speed but the higher the probability to find alternate routes.

It is recommended not to lower DEPR congestion threshold below 50% to reduce the DEPR processing overheads. Reducing the monitoring period makes the DEPR scheme adapts fast to rapidly changing network dynamics.

References

[1] A. Tanenbaum, "Computer Networks", Third Edition, New Jersey: Prentice Hall, pp. 412, 1996.

[2] Rosenberg J, Schulzrinne H, "Internet Telephony Gateway Location", Proceedings of IEEE Infocom'98, vol. 2, pp 488-496, 1998.

[3] E. Rosen, et al., "MPLS architecture", work in progress, draft-ietf-mpls-arch-07, July 2000.

[4] E. Rosen, A. Viswanathan, and R. Callon, "RFC 3031: Multiprotocol label switching architecture", Jan. 2001.

[5] G. Armitage, "MPLS: The Magic Behind the Myths", IEEE Communications Magazine, vol. 38, no. 1, pp. 124-131, January 2000.

[6] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell and J. McManus, "RFC 2702: Requirements for Traffic Engineering over MPLS", Sept. 1999.

[7] D. Awduche, "MPLS and Traffic Engineering in IP Networks", IEEE Communications Magazine, vol. 37, no. 12, pp. 42-47, Dec. 1999.

[8] R. Gallager, "A minimum Delay Routing Algorithm Using Distributed Computations", IEEE Transactions on Communications, vol. 25, no. 1, pp. 73-85, January 1977.

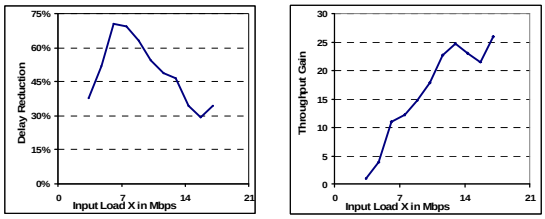
[9] K. Kar, M.Kodialam, and T.V. Lakshman. "Minimum Interference Routing of Bandwidth Guarnateed Tunnels with MPLS Traffic Engineering Applications." IEEE Journal on Selected Areas in Communications, vol. 18, no. 12, pp. 2566-2579, December 2000.

[10] R. Battiti and E. Salvadori., "A Load Balancing Scheme for Congestion Control in MPLS Networks", In IEEE Symposium on Computers and Communications (ISCC 2003), pp. 951, June 2003.

[11] I. Widjaja, A. Elwalid, "MATE: MPLS Adaptive Traffic Engineering", IEEE INFOCOM 2001- The Conference on Computer Communications, no. 1, pp. 1300-1309, April 2001.

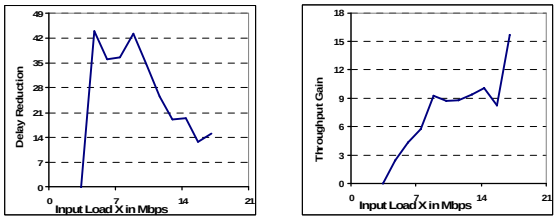
[12] F. Holness, C. Phillips, "Fast Acting Traffic Engineering (FATE) within MPLS Networks", MPLS World Congress 2001, Building a New IP Architecture, February 6-9th, 2001.

[13] NS2 web site and documentation, http://www.isi.edu/nsnam/ns, http://www.isi.edu/nsnam/ns/ns-documentation.



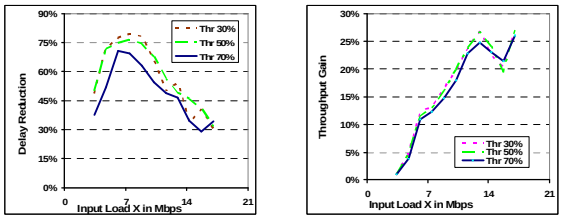
(a) (b)

Figure 6: (a) delay reduction, and (b) throughput gain using DEPR versus shortest path algorithm for the COST239 topology



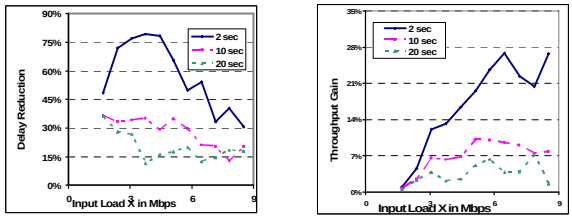
(a) (b)

Figure 7: (a) delay reduction, and (b) throughput gain using DEPR versus shortest path algorithm for the NSFNET topology



(a) (b)

Figure 8: DEPR system (a) delay reduction and (b) throughput gain for the COST239 topology with different congestion thresholds



(a) (b)

Figure 9: DEPR system (a) delay reduction and (b) throughput gain for COST239 topology with different monitoring intervals