Survivability-enhancing routing scheme for multi-domain networks

Li, Xiaohua; Ruepp, Sarah Renée; Fagertun, Anna Manolova; Dittmann, Lars

Published in:
Globecom 2008

Link to article, DOI:
10.1109/GLOCOM.2008.ECP.430

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Li, X., Ruepp, S. R., Manolova, A. V., & Dittmann, L. (2008). Survivability-enhancing routing scheme for multi-domain networks. In Globecom 2008 (pp. 1-5). IEEE. DOI: 10.1109/GLOCOM.2008.ECP.430

Survivability-Enhancing Routing Scheme for Multi-Domain Networks

X.Li, S. Ruepp, L.Dittmann, and A. V. Manolova
Network Competence Area, DTU Fotonik, Technical University of Denmark
Oersteds plads, building 343, 2800 Kgs. Lyngby, Denmark

Abstract — **We present a routing solution which eliminates the inherent path exploration problem of BGP and thereby enhances survivability in multi-domain networks. The path exploration problem is caused by the dependency among paths learned from neighboring domains. We propose to solve this issue by using two domain level disjoint paths. Our work is based on the aggregated representation of transit domains. The aggregated scheme we use is obtaining the shortest path between each pair of border routers for the associated domain. We also propose to use a single node to represent the destination domain, thereby the size of routing table is 2*(**n-1**). To implement our solution, domain level source routing is used and a SDRP header is added to the delivered packet. This avoids re-calculation and path exploration when repairing inter-domain link failures. Inter-domain link failures must be repaired at domain level, while intra-domain link failures can be repaired by neighboring nodes, border routers or at domain level. The intra-domain repair methods are compared by simulation and based on the results border router repairing is recommended.**

Keywords: aggregated representation, disjoint paths, inter-domain routing, multi-domain survivability

I. Introduction

Today's Internet is constructed of a vast amount of inter-connected autonomous systems (*ASes* , also called domains in this paper) administrated by different Internet Service Provider (ISP). Border Gateway Protocol (BGP), a path vector protocol, is the de-facto routing protocol used to exchange reachability information among domains. The path exploration phenomenon is the main reason of slow convergence of BGP deployed in multi-domain networks. The root cause of path exploration is the path dependency among neighboring *ASes* [1]. In other words, the selected paths in one domain are dependent on the path selection made in the neighboring *ASes* . In the worst case, the theoretical upper bound of alternate paths which may have to be explored after a failure is O(n!) [2].

Several solutions are proposed to reduce or limit the influence caused by the path exploration problem. Forward edge sequence number (fesn) introduced in [1] attaches a fesn list according to the AS path in the BGP announcements. This fesn list is used to distinguish failure events and invalidate multiple routes which match this list to limit the path exploration. The ghost-flushing proposal described in [3] improves the convergence time by propagating bad news quicker than good news. These solutions are palliative and cannot solve the problem thoroughly. In [4], a new routing protocol based on disjoint path calculation is proposed where a complex aggregated scheme obtaining minimum total cost of two link disjoint paths is adopted. The aggregated scheme contains two components, one is represented by M_i , an array of domain D_i storing the minimum cost between each pair of border routers and the other component includes a set of arrays represented by M_i^{jl} . M_i^{jl} stores the minimum cost for each two border routers after reversing links in the shortest path between border routers B_i^j and B_i^l and negating their cost. Through the representation of M_i and M_i^{jl} , these two paths are completely link disjoint even when they traverse the same domain. However, depending on the network's topological characteristics, it may be possible to find a shorter link disjoint path at domain level rather than finding a link disjoint path pair within the same sequence of domains. Here domain level disjoint means the inter-domain links in the two paths are disjoint, whereas inside a domain they may share some intra-domain links (caused by the topological characteristics, e.g., they traverse the same domain). Hence we use a simple aggregated scheme which only contains the set of array M_i .

Intra-domain link failures can be repaired by neighbor nodes, border routers of the associated domain or at domain level. We develop a new survivability-enhancing routing solution based on the simple aggregated scheme for minimum cost between each pair of border routers. We also propose to use a single node to represent the destination domain while applying the aggregated scheme proposed in [7] to the source and transit domains to decrease the size of the routing table. Our solution uses domain level source routing and a Source Demand Routing Packet (SDRP) header is added to the packets [5]. Inside domains standard Internal Gateway Protocol (IGP) is deployed.

The remainder of this paper is organized as follows: Sec. 2 describes the aggregated scheme used in this paper. Sec. 3 illustrates the formation of the routing table. We describe the source routing details in Sec. 4. Sec. 5 presents the implementation details and Sec.6 discusses the survivability issues in a multi-domain network scenario based on the simulation results. Finally, conclusions and future work are outlined in Sec. 7.

II. Aggregated Scheme

For scalability and commercial reasons, each domain cannot know the internal topology of other domains in multi-domain networks. However, calculating disjoint paths requires information about the traversal properties of each domain. We use the Aggregated Representation (AR) described in [7] to

978-1-4244-2324-8/08/$25.00 © 2008 IEEE.

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2008 proceedings.

calculate the minimum cost between each pair of border routers. We use B_i^j to denote border router of domain i (D_i). Thus domain D_i can be represented by a two-dimensional symmetric matrix M_i , where the element $m_i^{j,l}$ denotes the cost of the shortest path between border router B_i^j and B_i^l . Fig. 1 depicts an illustrative example of AR. For the topology shown in fig. 1, the elements of matrixes M_1, M_2 , M_3 , and M_4 are presented in fig. 2. The elements of M_i are calculated by the Dijkstra algorithm. Suppose domain D_i has n border routers, matrix M_i will contain n^2 elements.

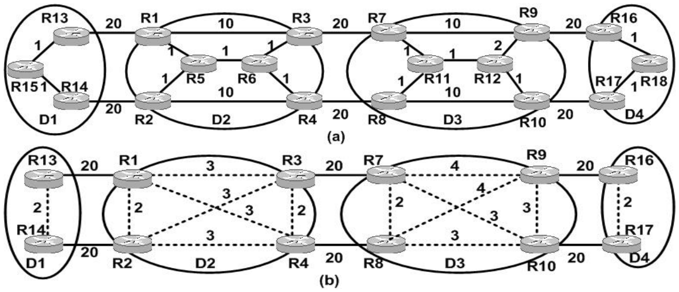

Fig. 1: An illustrative example of AR. (a) original topology (b) virtual topology after using AR to represent domains. Solid lines represent inter-domain links whereas dashed line represent virtual links obtained by AR.

$$\begin{array}{c|cc} M_1 & R_{13} & R_{14} \\ \hline R_{13} & / & 2 \\ R_{14} & 2 & / \end{array} \quad \begin{array}{c|cccc} M_2 & R_1 & R_2 & R_3 & R_4 \\ \hline R_1 & / & 2 & 3 & 3 \\ R_2 & 2 & / & 3 & 3 \\ R_3 & 3 & 3 & / & 2 \\ R_4 & 3 & 3 & 2 & / \end{array} \quad \begin{array}{c|cccc} M_3 & R_7 & R_8 & R_9 & R_{10} \\ \hline R_7 & / & 2 & 4 & 3 \\ R_8 & 2 & / & 4 & 3 \\ R_9 & 4 & 4 & / & 3 \\ R_{10} & 3 & 3 & 3 & / \end{array} \quad \begin{array}{c|cc} M_4 & R_{16} & R_{17} \\ \hline R_{16} & / & 2 \\ R_{17} & 2 & / \end{array}$$

Fig. 2: Elements of Matrixes *M1, M2, M3, and M4.*

III. Formation of the Routing Table

With BGP the reachability information is carried in route updates containing announcements or withdrawals of reachable destinations. A border router stores the set of all feasible paths, then it runs a route selection process and only the selected best route is advertised to its neighbor *ASes* [8]. We assume that each domain knows the reachability information of destination networks directly attached to other domains. Link state information of both inter-domain links and virtual links obtained by the AR is propagated by border routers throughout the multi-domain network. We use $l_{inter-domain}$ to denote the number of inter-domain links and $|B_i|$ to denote the number of border routers for domain D_i . The total number of links, both real and virtual, is calculated by equation (1).

$$\text{Total\_links} = l_{inter-domain} + \sum_i \frac{|B_i| * (|B_i| - 1)}{2} \qquad (1)$$

To propagate link state information, a flooding mechanism and fully-meshed internal Border Gateway Protocol (iBGP) connections [8] are adopted. Flooding is used to propagate information among neighboring domains while iBGP connections are used to propagate information among border routers of the same domain. This requires a modification in the BGP protocol to include link-state information dissemination. The exact modification of the BGP packet format is out of the scope of this paper. Proposals for BGP extensions to carry path-related information (a path can be seen as a link between two nodes) exist in the literature [13]. After inter-domain link state databases are synchronized, two domain level disjoint paths are calculated using the Suurballe algorithm [6] and the routing table is formed. Each border router maintains two link state databases, one for inter-domain routing and the other for intra-domain routing. We propose using a single node to represent the destination domain, thus the size of the routing table is $2*(n-1)$. Each entry in the routing table contains the following fields: destination domain, a binary field to distinguish the primary and backup paths, the domain level path and the total cost.

IV. Domain Level Source Routing

The OSPF protocol [10] uses the Dijkstra algorithm for path computation and only stores the next hop information instead of the whole path. However, when using the Suurballe algorithm this information is not sufficient. Fig. 3 presents an illustrative example. The shortest path from R5 to R10 is R5->R6->R4->R8->R11->R12->R10 marked by solid arrowed lines in fig. 3(a). Intermediate routers R6, R4, R8, R11 and R12 also choose the same route to the destination. Fig. 3(b) depicts the case of using the Suurballe algorithm. Two disjoint paths from R5 to R10 are marked by solid arrowed lines in black and gray color, respectively. However, for node R6 the two disjoint paths (marked by black and gray dashed lines) do not follow the same links. Thus, it is insufficient to store the next hop information and perform hop-by-hop routing when disjoint paths have been calculated with the Suurballe algorithm. Hence, we use domain level source routing to specify the packet's route.

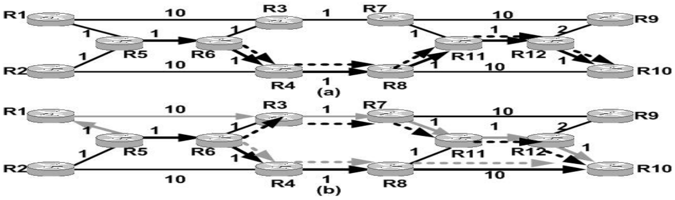

Fig. 3: An illustrative example. (a) Example of OSPF protocol (b) Example when using the Suurballe algorithm

Source routing is implemented by encapsulating the entire packet in a SDRP, and a SDRP header is added [5]. With our proposal the SDRP header stores domain level hops, i.e. the border routers on the way from source to destination. Hops inside a domain are not specified in the SDRP.

According to our proposal, the working path in a multi-domain network is specified in three steps:

1) Use IGP routing inside the source domain to reach the closest border router with the original packet header.

2) Use domain level source routing with an SDRP header.

3) Use IGP routing inside the destination domain with the original packet header.

V. Detailed Description of the Implementation

In this section, we describe the implementation in detail. For simplification, two assumptions are made:

(1) There are no parallel links between two different routers or domains.

(2) When a border router announces routes, no export policy caused by the commercial relationships, is applied between *ASes* .

Fig. 4 presents an illustrative example for domain level disjoint path calculation. For the topology depicted in fig. 1 (a), fig. 4(a) shows two domain level disjoint paths for data traffic from R15 to R18. The primary and backup paths are marked by arrowed black and gray lines, respectively. Fig. 4(b) illustrates two physical paths and these two paths share two intra-domain links: R5<->R6 and R11<->R12.

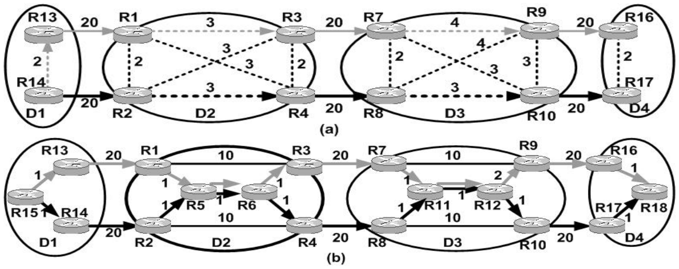

Fig. 4: An example of domain level disjoint paths (a) two domain level disjoint paths for traffic from R15 to R18, they are marked by arrowed black and gray lines, respectively. (b) Two physical paths from R15 to R18

A. Routing Table

For the multi-domain network shown in fig. 1(a), the routing table of border router R1 is the following:

Destination domain	Primary[1]	Next hops along the path	Cost
D_1	1	D_1	20
D_1	0	R_2 D_1	22
D_3	1	R_3 D_3	23
D_3	0	R_4 D_3	23
D_4	1	R_4 R_8 R_{10} D_4	46
D_4	0	R_3 R_7 R_9 D_4	47

The topology contains 4 domains, thus the size of the routing table is 2*(4-1) =6.

B. Procedure after an inter-domain link failure

Fig. 5 presents the procedure after an inter-domain link failure and this procedure is executed only by border routers.

VI. Survivability in a Multi-Domain Network

In this section we discuss different survivability schemes in a multi-domain network and present our simulation results. In [12], failures are classified into 18 cases according to the failure position, type of the failed element and the presence of parallel links. We focus on single link failures. Inter-domain link failures can only be repaired at domain level, whereas intra-domain link failures can be repaired by the two neighboring nodes of the failed link, the border router of the domain including the failed link, or the border router of the source domain possibly traversing other domains. We use OPNET [14] as the simulation tool and the following statistics are used to compare the different repair mechanisms for intra-domain link failures:

```
Received LSA packet indicating inter-domain link failure
  If this information is new then
    If it is received by flooding then
       Use iBGP connections to propagate LSA
       Flooding the LSA except the received interface
    Else
       Use flooding to propagate LSA
    End if
    For all $i \in \{Route\ entries\}$ do
      If failed link is contained in the working path then
         Switch to the other path
      End if
    End for
  Else
     Discard the LSA packet
  End if
```

Fig. 5: Process procedure after receiving LSA packet indicating inter-domain link failure

1: Packet loss:

$$Packet\ Loss = Packets_{generated} - Packet_{received} \qquad (2)$$

2: ETE delay:

$$ETE\ delay\ = t_{packet\_received} - t_{packet\_sent} \qquad (3)$$

3: Hops from source to destination

4: Total cost from source to destination

5: Re-convergence time of domains: if traffic originated from a domain can reach its destination again after failure occurs, we assume that the domain has converged.

A. Repairing intra-domain link failure in a multi-domain network

1) Neighbor nodes repairing

When a neighbor node detects the failure, it will calculate a new shortest path between itself and the downstream node in order to circumvent the failed link. To increase recovery speed, we propose that only the neighbor nodes update their topology. Packets which do not pass through the failed link still follow their previous route whereas those passing through the failed link use the new shortest path. Neighbor nodes repairing may introduce one or two routing loops between itself, the upstream node or the downstream node. Fig. 6 depicts an illustrative example. For traffic from R1 to R4 the shortest path without

[1] 1: Field “Primary” is used to distinguish the primary path and backup path, “1” means the primary path whereas “0” means the backup path. [1]

failure is R1->R5->R6->R4 marked by arrowed black lines in fig. 6. If link R5<->R6 fails, a new path R5->R1->R3->R6 is used to replace the failed link. The traffic now should follow the expected path R1->R5->R1->R3->R6->R4, marked by gray lines. The loop R1->R5->R1between R5 and the upstream node appears in the path, because neighbor nodes use the new shortest path to forward packets while other nodes still use their old routing table. In practice with normal OSPF infinite loops between R1 and R5 which are caused by the asynchronous routing tables in the nodes in the network are formed. To break infinite loops source routing between the two neighbor nodes is implemented and hops in this route are "strict" [5]. For the neighbor node repair scheme the re-convergence time is: $T_{\text{Re-convergence time}} \approx T_{\text{Detect time}}$.

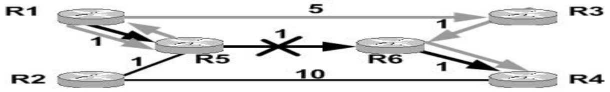

Fig. 6: The shortest path from R1 to R4 without failure is marked by arrowed black lines. Path obtained by neighbor node repairing after link R5<->R6 fails is marked by gray lines

2) Border routers repairing

When border routers of the domain where the link fails repair the failure, there is no topology change at the viewpoint of other domains and data traffic still follows the previous route. However, the real path is modified by IGP. In this case, the re-convergence time is: $T_{\text{Re-convergence time}} \approx T_{\text{Convergence time of } D_i}$

3) Domain level repairing

With domain level repairing, the border router of the source domain is the repairing node. Failure information, reflected by AR links which are re-calculated after the associated domain converges, is propagated throughout the network, and the border router of the source domain switches to the backup path. In this case, the re-convergence time is:

$$T_{\text{Re-convergence time}} \approx T_{\text{Convergence time of } D_i} + T_{\text{Propagation time from } D_i \text{ to } D_s}.$$

B. Comparison of the repairing methods

In our model 10 hops in the SDRP header are defined, which means a data packet can traverse 10 domains. We create a simple multi-domain network containing 5 domains resulting in the topology depicted in fig. 7 (a). Data traffic is created from R6 to R18 and we fail the intra-domain link R11<->R12. Statistics obtained by the three different repairing methods mentioned above are compared below.

a) Packet loss

The "Packet loss" is defined by equation (2) and fig. 8 depicts the packet loss experienced by R18 (i.e., the destination node). We assume that the links are error-free and the "Packet loss" is only caused during network convergence and re-convergence. At the beginning of the simulation the multi-domain network needs time to converge, hence all packets sent to the destination are lost. In fig. 8, ten packets are lost before network convergence, then this value is kept until link R11-R12 fails. Before the network re-converges the "Packet Loss" is increased. We mark the moments "Network converges", "Link fails" and "Network re-converges" of domain level repairing. For neighbor node repairing, the detection time is so short that the curve looks almost smooth, whereas domain level repairing always has the largest packet loss which means the source domain re-converges slowest.

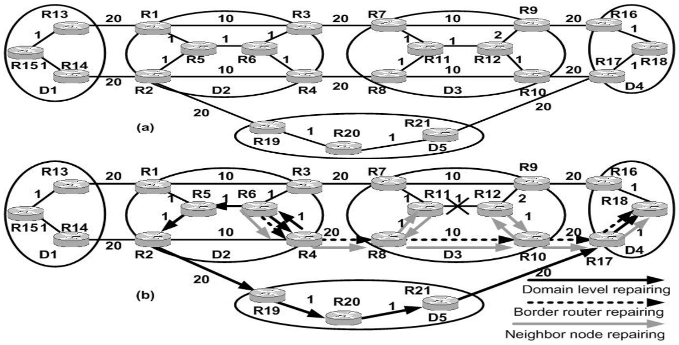

Fig. 7: An illustrative example of different methods to repair intra-domain link failure (a) The original topology (b) New paths after link R11<->R12 fails.

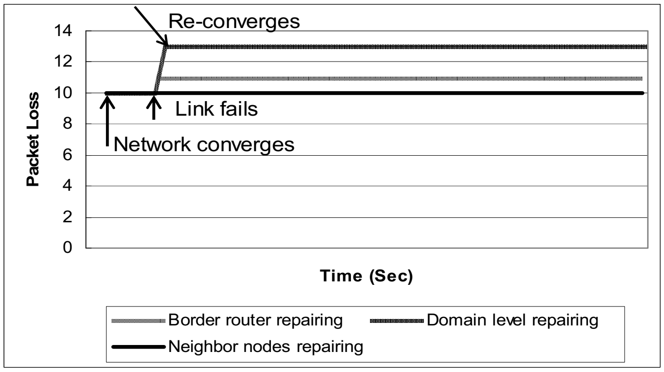

Fig. 8: Comparison of "Packet Loss" experienced by R18.

b) ETE delay

The "ETE delay" is compared in fig. 9. Neighbor node repairing has the largest ETE delay whereas domain level repairing has the smallest. Neighbor nodes repairing always has larger ETE delay than border router repairing because it traverses a longer path, however, there is no general result between border router repairing and domain level repairing, because the paths they traverse are topology-dependent.

Table 1 summarizes the comparison results of the three repair methods. Considering the trade-off among the factors compared in table 1, we recommend border router repairing.

C. Repairing inter-domain link failures

The procedure for repairing inter-domain link failures was depicted in fig. 5, and is similar to repairing intra-domain link failure at domain level. The re-convergence time of each domain is topology dependent and domains closer to the failed link re-converge quicker because the failure propagation time is smaller. Fig. 11 presents an illustrative example. For the topology shown in fig. 7 (a), we create three traffic streams from R10 to R17, R6 to R18, R15 to R16. Fig.10 depicts the primary paths of these streams which are marked by dashed, gray and black lines, respectively. Inter-domain link R10<->R17 now fails, and the failed link is included in the primary paths of traffics from R10 to R17 and R6 to R18. Fig. 11

TABLE I
COMPARISON OF THREE METHODS TO REPAIR INTRA-DOMAIN LINK FAILURE

Items	Neighbor nodes repairing	Border router repairing	Domain level repairing
Propagation scope of the failure	Smallest, limited to two neighboring nodes	Medium, limited to the associated domain	Largest, throughout the network
Re-convergence time	smallest	medium	largest
Packet loss	smallest	medium	largest
ETE delay	Largest[1]	Medium[1]	Smallest[1]
Hops (routers)	9[1]	5[1]	9[1]
Total cost	56[1]	52[1]	47[1]
Route loop	May introduce	No	No
Concern other domains	No	No	Yes
Recover node failure	No	Yes	Yes

Superscript [1] means the value is topology-dependent

depicts the "Packet Loss" calculated at the connection end nodes. There is no change in the curve of R16 because the primary path of the traffic does not include the failed link. The source domain of the traffic from R10 to R17 is closer to the failed link than traffic from R6 to R18, resulting in a smaller packet loss in R17 than R18, which is illustrated in fig. 11.

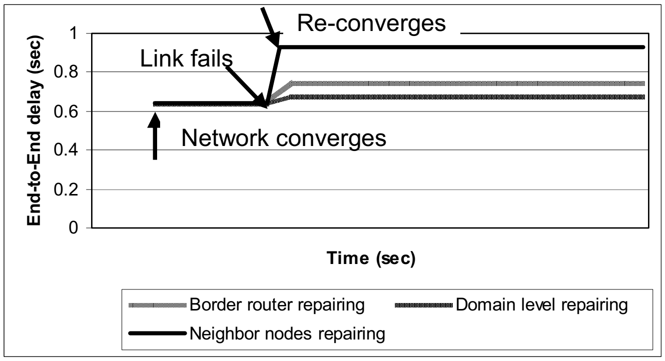

Fig. 9: Comparison of statistic "ETE delay" for three repairing methods.

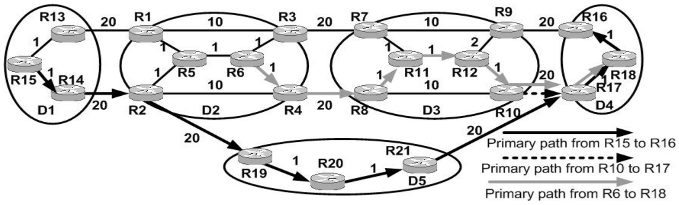

Fig. 10: The primary paths of traffics from R10 to R17, R6 to R18 and R15 to R16 are marked by dashed gray and black lines, respectively.

VII. CONCLUSIONS AND FURTURE WORK

This paper presents a routing solution which can eliminate the path exploration problem in multi-domain networks. We use the Suurballe algorithm to compute two domain level disjoint paths, and thus, solve the path dependency which is the root cause of path exploration for BGP. The aggregated scheme we use is obtaining the shortest path between each pair of border routers for the transit domains. To implement our solution, domain level source routing is used and a SDRP header is added to the packets. Considering the trade-off amongst re-convergence time, packet loss and ETE delay we recommend border router repairing for intra-domain link failures. With our proposal neither re-calculation nor path exploration are performed in case of an inter-domain link failure. For future work, we plan to include export policies and parallel links in the multi-domain network scenario.

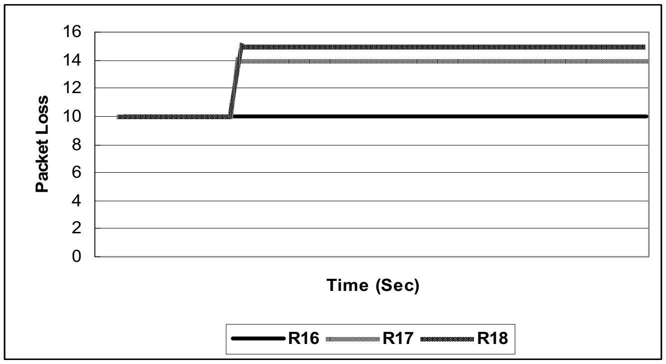

Fig. 11: Packet losses calculated at the nodes in the destination domain.

REFERENCES

[1] J. Chandrashekar, Z.Duan, , Z.-L Zhang and J.Krasky, "Limiting path exploration in BGP" in INFOCOM 2005, 13-17 March 2005

[2] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet Routing Convergence", IEEE/ACM Trans. Netw., vol. 9, no. 3, p.293-306, 2001.

[3] A. Bremler-barr, Y. Afek and S. Schwarz, "Improved BGP Convergence via Ghost Flushing", in INFOCOM 2003. Vol. 2, Issue 30, p.927-937, March-3 April 2003

[4] A. Sprintson, M. Yannuzzi, A. Orda, X. Masip-Bruin, "Reliable Routing with QoS Guarantees for Multi-Domain IP/MPLS Networks", INFOCOM 2007: p.1820-1828. March 2007

[5] D. Estrin, T. Li, Y. Rekhter, K. Varadhan and D. Zappala, "Source Demand Routing: Packet Format and Forwarding Specification (version 1), IETF RFC 1940, May 1996

[6] J. Suurballe and R. Tarjan, "A Quick Method for Finding Shortest Pairs of Disjoint Paths", Networks, Vol.14, No.2, p.325-336, 1984

[7] S. Uludag, K. S. Lui, K. Nahrstedt, and G. Brewster, "Comparative Analysis of Topology Aggregation Techniques and Approaches for the Scalability of QoS Routing.", Technical report TR05-010, DePaul University, Chicago, USA, May 2005

[8] Y. Rekhter, "A border Gateway Protocol 4 (BGP-4)", IETF RFC1771, March 1995

[9] T. Bates, R. Chandra and E. Chen, "BGP Route Reflection – An Alternative to Full Mesh IBGP", IETF RFC 2796, April 2000

[10] J. Moy, "OSPF Version 2", IETF RFC 2328, April 1998

[11] L. Subramaniam, S. Agarwal, J. Rexford and R. H. Kaze, "Characterizing the Internet Hierarchy from Multiple Vantage points", IEEE INFOCOM 2002, New York, June 2002

[12] D. Larrabeiti, R. Romeral, I. Soto, M. Uruena, T. Cinkler, J. Szigeti, J. Tapolcai, "Multi-domain issues of resilience", Transparent Optical Networks, 2005, Vol. 1, Issue , p. 375 - 380, 3-7 July 2005

[13] A. Muchanga, L. Wosinska, "Requirements for Interdomain Routing in Optical Networks," In Proc. of NFOEC, March 2005.

[14] OPNET Modeler. www.opnet.com

978-1-4244-2324-8/08/$25.00 © 2008 IEEE.