An MPLS Extension for Fast Recovery from Gateway Failures

Dimitri Staessens, Didier Colle, Mario Pickavet, Piet Demeester

Dept. of Information Technology (INTEC), Ghent University - IBBT - IMEC
G. Crommenlaan 8, bus 201, 9050 Gent, Belgium
{dimitri.staessens, didier.colle, mario.pickavet, piet.demeester}@intec.ugent.be

Abstract. When we examine today's internet architecture, we notice that the IP layer network is composed of different (sub)networks, managed by different service providers, operating different architectures, providing different services, pursuing different business paradigms. In order to provide survivable inter-domain connections to ensure connectivity in case of the most prevalent failures, different strategies can be followed [1]. All these schemes need multiple gateways to cover for gateway failures. In this paper, we recapitulate on a static and a dynamic multi-layer multi-domain recovery mechanism with reasonable capacity requirements in intermediate domains. Then, we comment on a current option to protect against IP-gateway outages and finally propose a mechanism based on Multi-Protocol Label Switching (MPLS), called Fast Gateway Switchover (FGS) in order to speedup gateway failure recovery.

1. Introduction

Today, optical backbone networks provide the physical transmission medium for multimedia, voice and internet traffic. With the introduction of WDM technology, the amount of data traversing links of these networks is tremendous [2]. Fast and scalable network recovery techniques are therefore of utmost importance in order to compensate for traffic loss that single link or node failures can induce [3]. Up to now, most resilience mechanisms are developed for single-domain environments, requiring total topological information of the network. When end-to-end traffic traverses different networks, end-to-end resilience can only be provided by a strategy that takes this situation into account.

Fig. 1 gives a simple schematic of a multi-domain network. Clearly, if we want to connect the client in network A to the client in network F, we need to traverse other networks. One option may be to connect through domains B and D; another option may be to set up a connection through networks C and D. When we want to set up survivable connections in such an environment, there may be extra requirements. For instance, if we want to provide end-to-end protection against all single node failures, we will need two distinct gateways for each domain. If there would be only one gateway in a domain, its failure would lead to the separation of its domain from the rest of the network.

1 Domains are sometimes referred to as Autonomous Systems (AS) or sub-networks. In this paper, we will use the term "domain".
When we have two gateways per domain, there are still some options for setting up working and backup paths or connections. For instance, we could set up disjoint paths through the same domains (e.g., both the working as the backup connection use the domains A-B-D-F), or we could set up the connections through different domains (e.g., A-B-D-F for the working path and A-C-D-F for the backup connection). One may intuitively feel that the second option is reliable; however, this is not the case. In the first case, where both connections run through the same domains, disjointness can be guaranteed (the operators of each domain know the topology), whereas in the second case, one needs to establish certainty that both domains are physically disjoint, i.e., be sure that there are no cross-domain Shared Risk Link Groups (SRLGs). A simple example of a cross-domain SRLG can be given by several links managed by different operators that use the same bridge over a river.

In the following sections we will first concentrate on the interconnection domain and then take a closer look on the source and destination IP domains.

2. Multi-domain IP-over-WDM resilience schemes

In what follows, we assume that the intermediary domains (like B and D in our introductory example) are based on WDM technology. We will discuss two options to provide a survivable leased-wavelength service between two IP-domains interconnected by this transport network, based on static and dynamic recovery [1]. The options will be presented for a single interconnection domain, however, the extension towards multiple interconnection domains can be made by applying one of these options in each interconnection domain.

Fig. 2 shows the hierarchy of recovery options within multi-domain (IP/MPLS) networks, interconnected by an Optical Transport Network (OTN). To provide end-to-end protection in such an environment, we can try to use existing recovery techniques within the different sections achieve end-to-end protection network node or link failure.

The different recovery sections and target domain, gateway-recovery (at both ends) as sections, we can apply a specific Fast Reroute (FRR) [7] in intermediate domain. An or multilayer networks is given.

In the following two sub schemes for the mentioned IP

2.1. Static recovery

First of all, we make a for this case, there is no need to its path will not disturb any concrete decisions on which for every inter-domain conn every inter-domain connect gateway for some connection; balance the load on the gateway.

It is possible to reduce lightpath to preempt the bac
within the different sections and provide proper coordination between them in order to achieve end-to-end protection and ensure inter-domain connectivity in case of single network node or link failures.

![Diagram of Multi-domain recovery](image)

**Fig. 2.** Multi-domain recovery.

The different recovery sections in Fig. 2 are: IP/MPLS recovery in both the source and target domain, gateway-to-gateway recovery, which is again subdivided in gateway recovery (at both ends) and intermediate network recovery. Within each of these sections, we can apply a specific scheme for protection against failures, like RSVP TE Fast Reroute (FRR) [7] in the IP domains and some path or link protection in the intermediate domain. An overview of single-domain recovery options in single and multilayer networks is given in [5].

In the following two subsections we will give a static and a dynamic cooperation scheme for the mentioned IP-over-WDM scenario.

### 2.1. Static recovery

First of all, we make a formal choice between working and backup IP connections. In this case, there is no need to protect the backup lightpath optically, since a failure along its path will not disturb any working connections [1]. The operators should make concrete decisions on which gateways are primary and which ones are used as backup for every inter-domain connection. The reason why gateways should be declared for every inter-domain connection is that a gateway must be able to serve as primary gateway for some connection while serving as backup for another. In this way we can balance the load on the gateways.

It is possible to reduce capacity requirements by allowing the primary backup lightpath to preempt the backup lightpath (Fig. 3). This is referred to as *common pool*
capacity sharing [5]. This is because the backup lightpath will only be used when an unrecoverable (from the interconnection domain point of view) failure in the working lightpath occurs, i.e. a gateway failure. In case of a failure in the working lightpath, the backup IP link is rendered invalid, but the working IP connection remains intact.

![Diagram of static inter-domain protection mechanism with backup preemption]

Fig. 3. A static inter-domain protection mechanism with backup preemption.

2.2. Dynamic recovery

In this scenario, we assume that the underlying optical network has Automatically Switched Optical Network (ASON) [6] capabilities. In this case, the optical layer tries to reroute all traffic among the available links and routers after a failure. These lightpaths aren’t protected again optically. We choose to set up only one (working) path, and, in case of a gateway failure, route the traffic over another gateway, requesting another path. This scenario is highly specific for multi-domain survivability, since the traffic originating (from the interconnection provider point of view) in the failing gateway can still be recovered, which is not possible in single-domain networks (Fig. 4). It should be noted that in order to provide this type of recovery in real-world situations, a reliable protocol needs to be developed with extensive signaling between IP domain and backbone network. In the source domain, an efficient way of rerouting traffic to the other gateway should be provided.
Capacity requirements for different methods for providing multi-domain recovery over an optical interconnection domain have been studied in detail in [5]. Following the conclusion that the static scheme with preemption required least capacity and this scheme has reasonable optimization potential, we have developed a mathematical programming model [13] in order to get an idea of a lower bound for the studied capacity requirement.

As the conducted study of inter-domain resilience schemes shows, the most difficult failures to be resolved are gateway failures, because they need to be resolved in both adjacent domains. In the following sections we will give an overview of current methods for recovery of a gateway failure in an IP network, and then propose an MPLS extension that could speed up this recovery. We will provide the study for a single-layer IP/MPLS domain. Further study will be conducted towards multi-layer IP domains, especially G/MPLS domains and towards the interactions required for providing full end-to-end survivability.

3. **Standard IP recovery with 2 gateways**

First we take a look at the standard recovery strategy of a statically configured IP network. In the network depicted in Fig. 5, we have two clients (10.0.0.101–102) connected to a network with 8 IP-routers (10.0.0.11–18) and 2 gateways (10.0.0.1–2). The routing information contained in an IP packet is its source (10.0.0.101) and destination (e.g. 157.193.40.33) address. When the client 10.0.0.101 wants to send packets to a remote destination, e.g. 157.193.40.33 it will be forwarded in the network according to a default value in each router’s routing table. This will be the default gateway, 10.0.0.1 in our example. The routing tables
are constructed and updated using an Interior Gateway Protocol (IGP), like Open Shortest Path First (OSPF, [10]). In our OSPF example, packets will be forwarded along the following (shortest) path:

```
10.0.0.101 - 10.0.0.15 - 10.0.0.12 - 10.0.0.1 - ...
```

Similarly, packets originating from the other client will follow the path:

```
10.0.0.102 - 10.0.0.17 - 10.0.0.13 - 10.0.0.11 - 10.0.0.1 ...
```

![Fig. 5. IP network recovery](image)

When router 10.0.0.1 fails, we will (manually) need to update the default value in the routing tables for the other gateway (being 10.0.0.2), or wait for reconfiguration of every router using e.g. the Dynamic Host Configuration Protocol (DHCP, [11]) after leases have expired. For explicit details we refer the reader to [3]. Dynamic discovery protocols, like IRDP [12] (ICMP Router Discovery Protocol) provide some alleviation of the problem, but they incur configuration and processing overhead in the clients and routers. In any case, very fast recovery times (within 50ms) will not be possible. We can therefore consider the gateway a single point of failure.

4. Virtual Router Redundancy Protocol (VRRP)

In order to address the stated problem of the gateway being a single point of failure inside a static IP-based network, VRRP has been developed. This protocol puts a group of gateways (called a VRRP group) on the same virtual IP address and designates one as master, the other as backups. All members of a VRRP group are also members of an IP multicast group. The priority and state of the gateways within its VRRP group can be configured using the VRRP states: master, backup, or null. If a failure occurs in the master, it will be detected and recovery from a gateway advertisement interval, without failover, is still not fast enough for high-reliability applications.

5. Using MPLS tunnels

In order to provide fast reachability features in multiprotocol tunnels, tunneled connections must be set up, to each gateway, and the multiprotocol tunneled approach typically means the use of an MPLS LSP. Especially the computation of the optimal path requires major effort. In the case of the IETF network, it may take some time for the best path to be discovered.
multicast group. The master sends VRRP advertisements over IPv4 to all other gateways within its VRRP group (using the reserved multicast address), communicating the priority and state of the master virtual router. All clients and routers in the network can be configured using the virtual address (e.g. 10.0.0.1), and in case of a gateway failure, they do not need to update their routing tables. This speeds up (and simplifies) recovery from a gateway failure, but recovery time will be dependent on the VRRP advertisement interval, which will be in the order of seconds. This recovery mechanism is still not fast enough for real-time content delivery.

5. Using MPLS tunnels

In order to provide fast recovery in case of gateway failures, MPLS tunnels can be used to provide routes to the backup gateway (Fig. 6). This means that for every connection that is set up, two Label Switched Paths (LSPs) must be calculated (one to each gateway), and the client must know when to switch over to the backup LSP. This approach typically means the calculation of end-to-end LSPs from source to destination.

Most of the work towards multi-domain survivability is done in this direction [8]. Especially the computation of end-to-end LSPs across multiple domains has received major attention by the IETF, [9]. However, if an end-to-end LSP fails in a remote network, it may take some time until the head-end detects this failure and switches over to the backup LSP.
6. A proposed Extension of MPLS: Fast Gateway Switchover (FGS)

To alleviate these problems, we propose a distributed scheme based upon MPLS label swapping, with a dedicated label for each gateway. We will focus in our examples on networks with two gateways, but extensions towards multiple gateways are of course possible. The only things that should be configured are the labels and the priority order in which the gateways will be used.

The scheme works as follows: in normal operation of the IP/MPLS network, all IP packets heading for a remote destination get labeled with the label of the primary gateway. The Label Switched Routers (LSRs) can set up their forwarding tables based on their IP routing table (constructed using OSPF), or the operator can configure those tables manually, should he prefer other than shortest paths towards the gateways. If the forwarding table is based on the IP routing table, packets with the label according to the primary gateway are forwarded on the same interface as packets destined for the primary gateway (its IP address), packets with the secondary gateway label will be forwarded towards the secondary gateway, and so on.

The recovery operation is as follows: whenever an LSR receives a packet labeled with the label of a backup gateway, it sets a flag denoting the highest priority gateway that is available and swaps all higher priority gateway labels for this backup gateway's label. The router that detects the gateway failure starts swapping packet labels and sets the flag upon detection, and thus triggers the recovery. In this way, packets are forwarded towards the secondary gateway without interruption. The flag is required to stabilize the recovery process and to re-establish the network to normal operational status after the failure is recovered. After it receives a message from a higher priority gateway that it is up again, the flag is updated, and the LSR swaps labels according to this flag. (So, the lower priority labels in the network will still be forwarded towards the lower priority gateway).

The recovery time is dependent only on how long it takes for the LSR directly upstream the failing gateway to detect the failure. Note that only gateway failures are handled this way, all link failures (even those between an LSR and a gateway) should be handled by FRR, using NHOP backup tunnels and internal router failures should be handled by means of NNHOP backup tunnels [7].
We now illustrate the working of the recovery based on our example topology. As default gateway we select 10.0.0.1 and assign it a white label. The backup gateway is thus 10.0.0.2. We assign the backup gateway a black label. We also assume that the routers set up the LSPs (more accurately, their MPLS forwarding tables) automatically using their IGP (OSPF) routing table. In Fig.7, we show the basic workings of the protocol in normal, failure-free operation. When an outbound IP packet originates from client 10.0.0.101, it forwards it according to its IGP routing table towards LSR 10.0.0.15. This label switched router detects an outbound IP packet, and labels it white (towards the primary gateway) and forwards it to LSR 10.0.0.12. This router forwards the packet (based on the MPLS label) towards the gateway 10.0.0.1. Since 10.0.0.1 is the destination, it strips the label. Similarly, for outbound packets originating in 10.0.0.102, the packets will be forwarded towards 10.0.0.17, 10.0.0.13, 10.0.0.11, and finally outbound through gateway 10.0.0.1.
We now consider the failure of the primary gateway 10.0.0.1 (Fig. 8).

In this case, router 10.0.0.12 will be the first to detect the failure. Thus, it just sets its flag and starts swapping white labels for black labels (until the gateway 10.0.0.1 sends a message it’s up again). Its MPLS forwarding table says black labels should be forwarded towards 10.0.0.11, and 10.0.0.11 receives a black label. Let’s assume it hasn’t detected the failure yet. 10.0.0.11 sets its flag and starts swapping white labels for black labels (upon reception of the black label), and forwards the packet coming from 10.0.0.101 towards the gateway 10.0.0.2, after stripping the label. (If it already detected the failure, swapping would have started at the moment of detection).

The packet stream from same manner (Fig. 9); e.g., swaps it for a black one, at label.

When router 10.0.0.12 detect this and stop swap 10.0.0.11 will still be safe.

Note that when the operational. We need to key

7. Conclusions

In this paper, we have Gateway Switchover (FGS) speed of the recovery deploy detect that the primary IP- scenario, the scheme can be

Future work includes st the protocol. It will also t
The packet stream from the other client (10.0.0.102) will be forwarded in the same manner (Fig. 9); eventually 10.0.0.11 will receive a white labeled packet, swaps it for a black one, and forwards it towards 10.0.0.2 after stripping the (black) label.

When router 10.0.0.1 comes back online, 10.0.0.12 and 10.0.0.11 will detect this and stop swapping. All black labels forwarded from 10.0.0.12 to 10.0.0.11 will still be forwarded towards 10.0.0.2.

Note that when the original gateway comes back online, both gateways are operational. We need to keep this in mind in choosing a multi-domain recovery scheme.

7. Conclusions

In this paper, we have proposed a distributed MPLS-based scheme, called Fast Gateway Switchover (FGS), which provides fast recovery from IP gateway failures. The speed of the recovery depends only the time the first LSR with FGS capability needs to detect that the primary IP-gateway is down. While we concentrated on an IP/MPLS scenario, the scheme can be used with any packet-switched technology.

Future work includes standardization of the message format and implementation of the protocol. It will also be interesting to study the time it takes to flush all lower
priority labeled packets out of the network (through the backup gateway) when a higher priority gateway recovers.

Acknowledgments

This work has been funded by the EC through the IST integrated project NOBEL phase II.

References

[13] D. Staessens et al., “Providing Survivable Interdomain Connections over an Optical Backbone Network”, to be presented at the 8th INFORMS Telecommunications Conference, Dallas, TX, March 30 – April 2, 2006

Abstract:

Future transparent WDM physical nature has a sig requires fulfillment of cor of network dimensioning a applications. In this contr [1] and pursue the approach that is constrained to a incorporated with differ dependence, linearity of c dependence.

Thus, constrained RWA computational complexity Computational complexity affects the network perf allowable setup time and performance. As tradition demand and emerging app become more adaptive.

We investigate an iterati constraints by a simple complexity, the model ai Therefore, it is also suit short computation time performance comparing w

This talk includes results EU 6th Framework project
V Workshop in G/MPLS Networks

Edited by José L. Marzo

Girona, 30-31 March 2006