Avoiding disruptions during maintenance operations on BGP sessions

Pierre Francois
Dept CSE
Université catholique de Louvain
Belgium
pierre.francois@uclouvain.be

Pierre-Alain Coste
France Telecom R&D
pierrealain.coste@orange-ftgroup.com

Bruno Decraene
France Telecom R&D
bruno.decraene@orange-ftgroup.com

Olivier Bonaventure
Dept CSE
Université catholique de Louvain
Belgium
Olivier.Bonaventure@uclouvain.be

***Abstract*— This paper presents a solution aimed at avoiding Losses of Connectivity when an eBGP peering link is shut down by an operator for a maintenance. Currently, shutting down an eBGP session can lead to transient Losses of Connectivity even though alternate path are available at the borders of the network. This is very unfortunate as ISPs face more and more stringent Service Level Agreements, and maintenance operations are predictable operations, so that there is time to adapt to the change and preserve the respect of the Service Level Agreement.**

I. Introduction

Service Level Agreements (SLA) established between ISPs and their customers are more and more stringent, which can be explained by the success of End-to-End performance demanding markets like VoIP and on-line gaming. The VPN services market follows the same trend, and a bad performance VPN service is critical for a provider as it is of its sole responsability.

In such a context, End-to-End performances in Today's Internet are unstable, and Losses of Connectivity (LoC) lasting tens of seconds can occur. It has been confirmed recently through measurements that such LoC were often correlated with routing changes and that the topology of an ISP, its routing policies, and its iBGP configurations have an impact on the extent of those LoC [1].

Also, many of those routing changes are caused by manual, maintenance operations. For example, shutting down a BGP peering link in an ISP network is a common maintenance operation that has to be performed on a daily basis by its operators.

So, due to the way things work with BGP, Service Providers can reach the limit of their SLA by simply performing a maintenance operation. Scheduling maintenance operations when the network is less utilized, or when customers are less expected to notice the failure has sometimes been decided as a better-than-nothing management solution to reduce complaints from the customers. However, such periods are usually nights in Regional ISPs, which tends to increase maintenance costs. Worse, those periods are not easy to identify in Tier-1 ISPs, and can vary according to the customer. When the maintenance affects multiple customers at a time, finding the best maintenance window is an organizational nightmare.

The iBGP is not fitted for current performance needs, and the introduction of Route Reflectors to solve scalability issues bound to iBGP full-meshes has worsened the extent of routing failures in the case of manual shutdown. Though, customers are right to complain about the poor performance they perceive under maintenance operations, especially when they pay for redundant access to the ISP. In addition, large recovery times are unfortunate when caused by predictable events.

To solve that problem, a "make before break" solution should be available when a peering link is shut down. The goal of this paper is to propose and evaluate an operational behaviour that should be followed in order to perform a zero packet loss convergence in the case of a peering link maintenance. Firstly, we present some data on the frequency of maintenance operations performed in a Tier-1 ISP, demonstrating the need for such a mechanism. We will then present an analysis of the causes of transient Loss of Connectivity (LoC), and we will illustrate them with a lab experiment. After that, we will present a technique that provides loss free convergence when a link is manually shut down. Next, we will present some slight modifications to the BGP protocol stack making this operation automatic. Finally, some measurements results showing the gain of the solution will be exposed. Those measurements where performed by manually introducing commands to routers to simulate the automatic process that we propose.

II. Motivation

In this section, we present some data on the frequency of maintenance operations performed in some ISPs. These data show that maintenance operations can have a significant impact on the service provided to the customers of an ISP. This motivates the introduction of techniques that can help to avoid any disruption of service in the case of planned modification of a network topology or of one link with a customer, peer, or provider.

The number of these planned maintenances greatly depends on the network considered and especially on its growth in term of bandwidth and new services provided to the customers.

For example in a European Tier-1 Internet ISP each router is reloaded every 6 months. In addition each eBGP session is on average shut down twice a year (peer unconfiguration,

link upgrade, max-prefix limits, malformed updates...). So an average customer is impacted by 8 maintenance operations per year on any given path, as each path flows accross two eBGP peering links. Note that this does not account operations performed on the customers' routers.

In a major VPN Service Provider 80% of PE "failures" are due to planned maintenance. For every minute of PE unavailability, 46% (28 seconds) is due to planned maintenance operations performed on the PE. In Figure 1, we can see a plot of the percentage of PE unavailability caused by maintenance operations for 12 months in this VPN Service Provider. One ratio shows, for each month, the number of seconds of PE unavailability implied by maintenance operations divided by the total number of seconds of PE unavailability. The second ratio shows, for each month, the number of maintenance events provoking a PE unavailability divided by the total number of events that led to a PE unavailability.

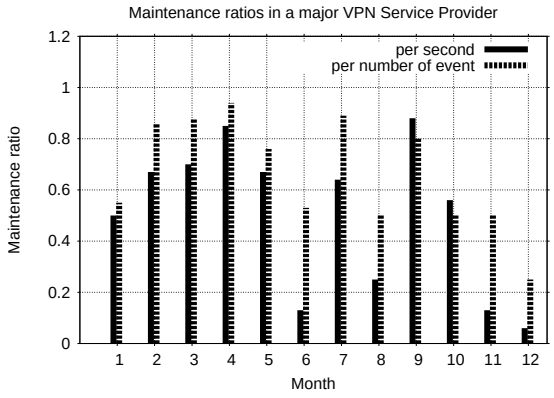

Fig. 1. Maintenance Ratios

Finally, as a third example, in a commercial Internet network, 58% of router failures happen in the 22H-06H maintenance window and most of these failures are due to maintenance operations.

During these maintenance operations, routing protocols behave as if it was a sudden link or router failure. As a consequence, the service is disrupted as in the case of a sudden failure. The duration of the disruption vary according to the number of affected prefixes and the design of the networks undergoing the topological change.

To understand why this problem is inconvenient for a service provider, let us consider that a customer pays for a redundant connectivity with a provider, via two or more interdomain links. Such a customer can typically request for a VPN service with a strong Service Level Agreement. In this context, when the operators of the provider have to upgrade a router or shut down a peering link for any reason, the customer will notice a loss of connectivity that is only due to the way BGP handles manual operations.

Though, scheduling maintenance operations to fixed time windows, when the traffic is supposed to be low, is not practical in an international network spanning through one or more continents [2]. We thus need a mean to do a peering link shutdown that does not jeopardize the reachability across the network, even transiently, so that maintenance operations can be performed at any time.

We can split maintenance operations affecting peering links in two types. The first type of maintenance covers the cases where the forwarding plane is not actually affected by the event. In some router designs, it is possible to reboot the control plane of a router or restart a BGP process without affecting the fowarding planes that depends on them. For these cases, graceful restart extensions as standardized in [3] are much more appropriate then what is proposed in this paper, as they permit to avoid a BGP convergence during the router control plane reboot or a BGP process restart. The second type of maintenance covers the cases where the forwarding plane is affected by the operation. This is the case for example for router line card upgrade, BGP shutdown due to unconfiguration, and link bandwidth increase. On a platform that does not support graceful restart extensions, it is also required to trigger a graceful convergence before rebooting a router or restarting a BGP process.

III. Losses of Connectivity Caused by Maintenance Events

Today, when a BGP session is manually shutdown, transient losses of connectivity (LoC) can still occur. In this section, we first review the cause of such LoC, and we illustrate them by the means of an example and a lab experiment.

A. The causes of transient LoC

The first reason is when a session shutdown is **performed abruptly** by a router. Packets arriving to the router via this peering link might be dropped, if the router has been configured to do so, by using a Reverse Path Forwarding Check [4].

Secondly, some routers can transiently **lack of alternate paths** to some prefixes at the moment of the shutdown. This happens in the typical case where potential alternate paths are not selected as best routes by the egress routers and route reflectors, so that they are not propagated through the network.

Finally, some networks still use Pervasive BGP. Thus, a BGP lookup is performed by each router on the path of a packet, from its Ingress point towards its Egess point. In such a context, **transient inconsistency** among the BGP tables of the routers can lead to packet loops. Note that a packet caught in a loop will be dropped as soon as its TTL reaches 0.

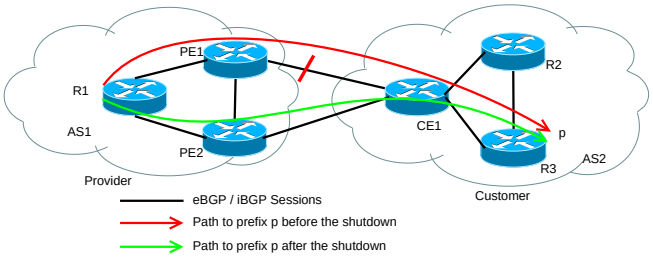

Fig. 2. A dually connected client

To illustrate these problems, let us look at the simple topology depicted in Figure 2. As a first approach, we will consider that a **full-mesh** of iBGP sessions is used in both networks.

A shutdown of the link between $PE1 \leftrightarrow CE1$ is performed on $PE1$. A BGP Cease Notification message will be sent by $PE1$ to $CE1$. Also, $PE1$ will re-execute its BGP Decision Process by removing all the routes that were received from $CE1$, and will select alternate path via $PE2$ when available. $CE1$ will also update its routing tables to forward all its packets along $CE1 \rightarrow PE2$.

Meanwhile, packets to be forwarded along $PE1 \leftrightarrow CE1$ are dropped, until alternate paths are selected and the FIB of all the routers are updated for all the prefixes that they reached via $PE1 \leftrightarrow CE1$.

For traffic engineering purposes, MED or agreement on communities with associated Local-Pref can be used by AS2. For example, the routers of AS1 can be forced to reach a given prefix p of $AS2$ via one specific link, let us say $PE1 \leftrightarrow CE1$. Thus, $PE2$ will not select as a best path the path via $PE2 \rightarrow CE1$. As it is not a best path, it will not be propagated inside the iBGP topology of $AS1$. As a consequence, when the shutdown is performed on $PE1$, $PE1$ does not have an alternate path towards p. $PE1$ will send a withdraw towards its iBGP peers for its previously advertised route towards p. When $PE2$ receives the path withdraw for p, it will run its BGP Decision Process and select its own path via $CE2$. Then, $PE2$ will advertise this path towards $PE1$. The LoC between $PE1$ and p will only be recovered once $PE1$ performs a FIB update taking this path into account.

If Pervasive BGP is used in the network, the fact that $PE1$ now has an alternate path towards p does not mean that $PE1$ recovered the reachability of p. Let us assume that the shortest intra domain path from $PE1$ towards $PE2$ is via $R1$. If $R1$ has not updated its BGP lookup table yet for p, it still forwards packets to p towards $PE1$, and a forwarding loop is taking place. If MPLS was used in the network, then $PE1$ encapsulates packets towards $PE2$, so that the path followed by these are loop free even if the intermediate routers do not have consistent BGP lookup tables.

The introduction of **route reflectors** also has a negative impact on the recovery after planned maintenance. Let us change the topology of Figure 2, and consider that $PE1$ and $PE2$ are clients of $R1$. In that case, when the shutdown is performed on $PE1$, the following steps must be performed. $PE1$ sends a withdraw to its route reflector $R1$ for a prefix p. $R1$ runs its decision process and finds no alternate route for p. $R1$ sends a withdraw towards $PE2$. $PE2$ runs its decision process and select the alternate path via $PE2 \rightarrow CE1$. $PE2$ advertises this new path towards $R1$. $R1$ runs its decision process, and propagates the path towards $PE1$. $PE1$ runs its Decision Process and finally recovers the reachability of p. Note that upon a BGP session shutdown, there can be plenty of prefixes impacted, so that the recovery process explained above can take seconds.

B. Lab experiment

To evaluate the LoC in a realistic setting, we reproduced in a lab a network topology composed of two interconnected ASes. This topology has similar characteristics of network topologies found in large ASes and small customer networks. The topology is shown in figure 3. In the bottom of the figure, a small customer AS is modelled as being composed of five routers $(lr10 - lr14)$. All the links shown in the customer AS have their IGP weight set to 10. The customer AS has two peering links ($lr10 - lr1$ and $lr11 - lr2$) with its provider in the bottom of the figure. The provider AS is composed of nine routers ($lr1 - lr9$). In this AS, all links have their IGP weight set to 10, except the link between $lr4$ and $lr8$ whose IGP weight is set to 30.

Reproducing this topology with real routers would have either forced us to use low-end routers or routers running Linux or FreeBSD and open-source routing daemons on PCs. Unfortunately, the problem with these two approaches is that the main CPU of the router or the PC serves both for the routing protocol and the packet forwarding. When the CPU is running BGP, it does not forward packets and vice-versa. Instead, we opted for a platform where both IP and MPLS forwarding are performed by a dedicated ASIC.

We choose a Juniper M7i running JUNOS 7.x. This router is a commercial router running the same BGP code as in large ISP networks. Furthermore, the M7i can be configured to act as a set of virtual routers [5]. With a single M7i, we were able to reproduce the topology shown in figure 3 with 14 logical routers. The CPU of the M7i is shared between all the logical routers (e.g. to run the BGP processes), but the M7i has a dedicated ASIC to switch packets between logical routers so all packets forwarding is performed in hardware.

Using these logical routers has some side effects compared to other approaches such as using real routers, using a PC-based testbed or using simulation. Compared to real routers, the main advantage of logical routers is their cost. With real routers, it would have been impossible for us to use a similar topology. Compared to a PC-based testbed, the main advantages of the M7i are that it uses production-quality BGP code and contains an ASIC for packet forwarding. Thanks to this ASIC, the M7i completely supports MPLS and IP forwarding without using the central CPU. The main drawback is that the CPU is shared among all logical routers. This implies that CPU bounded processes will be slower on the logical routers than on a testbed with 14 real routers. Note that care has been taken not to overload the physical router memory during the experiment. Finally, compared to simulations, the main advantage of the utilisation of logical routers is that the testbed uses the same BGP code as in a real network. Simulations would have allowed us to evaluate the behaviour in larger networks.

We only present results for the case of a failure of a peering link between a provider and a customer. Note that the results are equivalent for the case of a shared cost relationship between the neighboring ASes. Indeed, the transient LoC are

due to a lack of alternate paths throughout iBGP topologies, which happen as well for paths crossing such a type of peering links. Due to space limitations, we do not provide detailed results for this case.

The measurements were conducted by attaching an Agilent router tester to $lr9$ and to $lr12$. The router tester is able to generate both BGP routes and real packets. To avoid overloading the memory of the M7i, we configured the router tester to advertise 6000 prefixes to $lr9$. Those 6000 prefixes are distributed through the iBGP and eBGP sessions to all logical routers. Those prefixes were extracted with their path attributes from a real BGP routing table. Besides during the tests the CPU load was monitored to check that the CPU was not the bottleneck. This loads the router in order to force the BGP processes to deal with a large number of routes like in the operational networks.

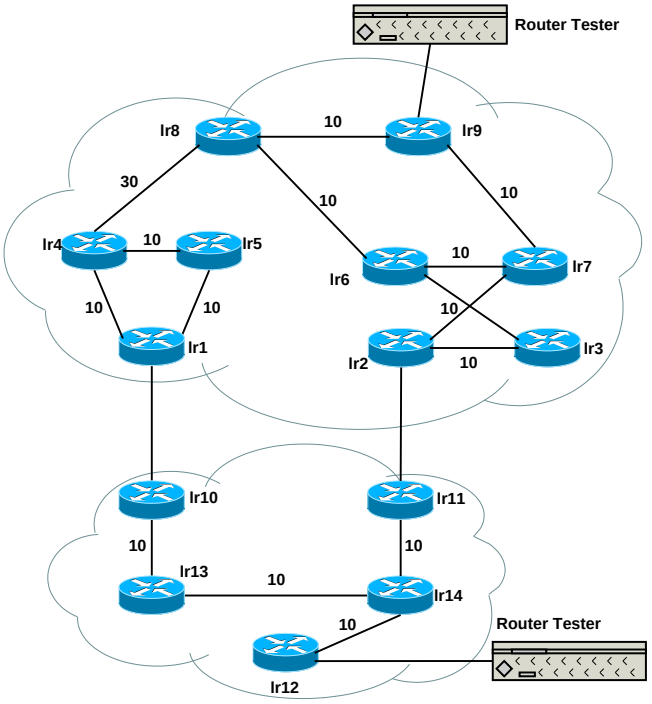

Fig. 3. Topology of the experiment

1) Context of the experiment: This physical topology is not sufficent to determine how a network will react to the shutdown of BGP peering links. The organisation of iBGP inside each AS must also be specified. In a small AS, iBGP is usually organised as a full-mesh of iBGP sessions. In a large AS, one or several route reflectors are used to distribute the BGP routes inside the AS. For these measurements, we considered the following iBGP organisations :

- A full mesh of iBGP sessions inside the top and the bottom AS.
- A redundant route reflector (RR) topology in each AS. On the provider side, [lr4, lr5] form a first cluster and have both the same cluster-ID. [lr6, lr7] form a second cluster and have both a same cluster-ID, yet different from the first cluster. The two top logical routers $[lr8, lr9]$ and the two bottom right $[lr2, lr3]$ each have an iBGP "client" session with $[lr6, lr7]$. $lr1$ has an iBGP "client" session with $[lr4, lr5]$. On the customer side, the two bottom logical routers $[lr13, lr14]$ are the RR for the other logical routers $[lr10, lr11, lr12]$.
- A redundant Hierarchical RR topology in the ISP AS with a different cluster ID for each RR in both AS. On the provider side, $lr8$ and $lr9$ are the hierarchical RR for both pairs of RR $[lr4, lr5]$ and $[lr6, lr7]$. $[lr2, lr3]$ each have iBGP sessions with $[lr6, lr7]$ and $lr1$ has an iBGP "client" session with $[lr4, lr5]$. On the customer side, the two bottom logical routers $[lr13, lr14]$ are the RR for the other logical routers $[lr10, lr11, lr12]$.

Besides the iBGP organisation, a second factor to be considered is the criteria used by BGP to select the best path to reach each destination. The BGP decision process is composed of seven steps [6], [7]. For the lab experiment, we focus on the two main criterias that are used to select the best exit peering link among the ones connecting two neighboring ASes. First, we consider that the import filters of each AS were configured to set the same local-pref values on all routes received from the neighbour AS, and let routers use the IGP tie break rule to select their best paths. In this case, $lr9$ will prefer to send its packets towards the customer AS via $lr2$ as its internal path towards $lr2$ is shorter than its internal path towards $lr1$. Second, we consider that the import filters were tuned to set a higher local-pref value to prefer the peering link $lr2 - lr11$. In both cases, all packets received by $lr9$ or $lr12$ will use the $lr11 - lr2$ peering link. Note that the results obtained with this case would have been equivalent if the tie-break was performed by letting the customer AS set a lower MED value to the routes advertised on the $lr11 - lr2$ peering link, as routers of the provider prefer customer routes with a lower MED among the ones with the same local-pref.

C. Measurements

Once BGP was stable, we configured the router testers to let $lr9$ and $lr12$ send packets to each other. One IP packet of 64 bytes was sent per msec from one router tester to the other, which makes a transmitted rate of around $0,50Mbit/s$.

The router tester is able to count each packet loss thus providing an exact measurement of the perceived loss of connectivity (LoC). The formula used to calculate the loss of connectivity is the following one:

$$LoC = \frac{Nb\_of\_packets\_lost \times Size\_of\_one\_packet}{Transmitted\_rate}$$

This simple formula has been used because we noticed that the lost packets were contiguous. In addition the TTL of the packets has been limited to 25 in order to quickly drop the packets that are caught in a transient loop. Each test is run five times and the tables show the mean values.

For each test the eBGP session is shut down on $lr11$. Note that the link $lr2 - lr11$ remains physically up so that the LoC

iBGP organization	LoC with Pervasive BGP	LoC with MPLS	Stream direction
Full-mesh	1.879	0.000	$lr12 \rightarrow lr9$
	0.248	0.000	$lr9 \rightarrow lr12$
RR	2.058	0.000	$lr12 \rightarrow lr9$
	1.220	0.983	$lr9 \rightarrow lr12$
HRR	1.707	0.000	$lr12 \rightarrow lr9$
	1.146	3.367	$lr9 \rightarrow lr12$

TABLE I
LOSS OF CONNECTIVITY (SECONDS) - IP AND MPLS FORWARDING PLANE - IGP USED AS TIE-BREAK

iBGP organization	LoC with Pervasive BGP	LoC with MPLS	Stream direction
Full-mesh	1.982	0.000	$lr12 \rightarrow lr9$
	2.298	1.948	$lr9 \rightarrow lr12$
RR	2.155	0.000	$lr12 \rightarrow lr9$
	2.407	1.709	$lr9 \rightarrow lr12$
HRR	2.183	0.000	$lr12 \rightarrow lr9$
	2.180	3.266	$lr9 \rightarrow lr12$

TABLE II
LOSS OF CONNECTIVITY (SECONDS) - IP AND MPLS FORWARDING PLANE - LOCAL-PREF USED AS TIE-BREAK

is only due to the BGP behaviour as the physical path via the link remain valid.

Let us analyse the results for the case where the IGP tie-break is the rule that permits to select the best path (Table I).

When looking at the results in Table I, we can see that not a single packet is lost for the packet stream **from lr12 to lr9**, i.e. the stream from the customer to the provider, when MPLS is used in the customer network. In fact, with all the iBGP topologies tested in the customer network, $lr12$ never lacked of a route towards $lr9$:

- When the iBGP topology is a full-mesh, $lr11$ has the path to $lr9$ via $lr10 \rightarrow lr1$ in its Adj-Rib-In. When the shutdown is performed on $lr11$, $lr11$ updates its FIB and forwards packets destined to $lr9$ in a MPLS tunnel towards $lr10$. Transiently, the path from $lr12$ to $lr9$ is $lr12 \rightarrow lr11 \rightarrow lr10 \rightarrow lr1 \ldots lr9$. When $lr12$ receives the withdraw from $lr11$, $lr12$ directly pushes the packets in a tunnel towards $lr10$.
- When the topologies with route reflectors are used, $lr12$ and $lr11$ also have the path via $lr10 \rightarrow lr1$ in their Adj-Rib-In, because they always have an iBGP session with route reflector $lr13$, whose best path is via $lr10 \rightarrow lr1$.

The fact that the routers on the customer do not lack of routes **does not depend on wether MPLS or Pervasive BGP is used** in the network. This means that the same analysis can be performed for the Pervasive BGP scenarii. However, we see in the results that some LoC has been perceived for the traffic from $lr12$ to $lr9$, which was not perceived when using MPLS. This can be explained by **forwarding loops** occuring on the link $lr11 \leftrightarrow lr14$ when the shutdown is performed. In all the scenarii presented in this paper, forwarding loops can happen in the customer network until $lr14$ has updated its FIB to forward packets destined to $lr9$ towards $lr13$. Before that, $lr11$ deviates packets to $lr14$, which forwards them back to $lr11$. When the TTL of those packets reaches 0, the packets are dropped, which explains the LoC.

Now, if we look at the results for the traffic from **lr9 to lr12**, we can see that some LoC has occured when using Pervasive BGP and MPLS :

In the case of an iBGP full-mesh all the routers have the alternate path via $lr1 \rightarrow lr10$ in their Adj-Rib-In before the shutdown. Indeed, $lr1$ has selected this path as its best one and propagated it in the Provider AS. When the shutdown is performed in $lr11$, and a BGP Cease message is sent to $lr2$, $lr2$ selects $lr1$ as its new nexthop for $lr12$.

When MPLS is used, those packets are tunneled to $lr1$ so that no forwarding loops occur. By chance, with the platform used to perform the test, $lr11$ accepted packets from $lr2$ for a while even if a BGP session shutdown had been issued in the router. This gives $lr2$ the time to update its FIB and avoid the link $lr2 \rightarrow lr11$ before $lr11$ drops packets received on that link. However, this behaviour is not mandatory and one could see packets being dropped on different platforms or when an unicast RPF check is configured on the peering link.

When Pervasive BGP is used, the same scenario happens, but $lr2$ does not use a tunnel to forward packets to $lr1$. Transiently, routers $lr8, lr9, lr6, lr7$ and $lr3$ still forward packets destined to $lr12$ on their Shortest Path towards $lr2$, while $lr2$ forwards them back to those routers. Transient fowarding loops thus occur, which resulted in an average LoC of 248 msec. As $lr2$ is the first router to be aware of the shutdown, it its the first router to adapt to the change, so that the ordering of the FIB updates that would imply a loop free convergence will never be applied by the routers.

In the case of a topology with route reflectors (without hierarchy) some transient LoC can occur. Indeed, $lr2$ only peers with route reflectors $lr6$ and $lr7$. Their best paths is via $lr2$ so that $lr2$ does not know about the alternate path via $lr1$. When the shutdown is performed, $lr2$ will send a Path Withdraw to $lr6$ and $lr7$, and start dropping packets as it has no alternate path towards $lr12$. Upon the reception of the path withdraw, $lr6$ and $lr7$ will send the alternate path via $lr1$ to $lr2$. Indeed, those route reflectors are peers of route reflectors $lr4$ and $lr5$, which already selected this path as their best path, so that this path was already present in the Adj-Rib-In of $lr6$ and $lr7$.

When MPLS is used, the connectivity is recovered once $lr2$ or $lr9$ receive the alternate paths from their route reflectors. The packets are then tunneled towards $lr1$, which forwards them to the customer. Note that, in most of the cases, $lr2$ is the first to actually reroute packets towards $lr1$. Indeed, $lr2$ reroutes them once it receives the alternate path from one of its route reflectors, as it has no other path available. The situation is slightly different in $lr9$, as this router has the old path in its Adj-Rib-In until all of its route reflectors send a Path Update message to $lr9$. Until then, $lr9$ still has one copy of the old

route, and keeps using it as this old route is better than the alternate path that it received.

When Pervasive BGP is used, the connectivity is recovered once both $lr9$ and $lr8$ have switched to the new path. This requires that both route reflectors $lr6$ and $lr7$ have sent path updates to $lr9$ and $lr8$, so that no copy of the old route remains in the Adj-Rib-In of $lr9$ and $lr8$. Until $lr9$ and $lr8$ have switched to the new path, packets towards $lr12$ loop between routers that still use the old path and routers that have already switched to the new one, which results in a LoC.

Let us now analyse the results in for the cases where the local pref value of the routes is the decisive rule to select the best path (Table II).

In the case of an iBGP full-mesh some LoC happens due to a transient lack of path towards $lr12$. Indeed, $lr1$ cannot select and advertise its path via $lr1 \rightarrow lr10$ as the local pref rule is considered before the IGP tie-break. When $lr2$ receives the BGP Cease message from $lr11$, it will not have an alternate path towards $lr12$ until $lr1$ receives the withdraw from $lr2$, that will let $lr1$ select its own path towards $lr12$ and propagate it in the Provider network.

In the case of a topology with route reflectors (without hierarchy) some transient LoC can occur. Compared to the case where the IGP metric is used as a tie-break between the two possible BGP nexthops for $lr12$, we can see that none of the four route reflectors know about the alternate path via $lr1 \rightarrow lr10$. Indeed, due to the usage of local-pref, $lr1$ itself does not select its external path so that it is not propagated. $lr1$ will propagate this path once it has received a path withdraw message from its route reflectors, $lr4$ and $lr5$. These route reflectors will send this path withdraw once both other route reflectors, $lr6$ and $lr7$, have sent a path withdraw. Until all those obsolete advertisements have been withdrawn, $lr1$ will not know about the unavailability of this path. This explains the longer convergence time compared to the other scenarii.

Similar results were obtained when using the topology with route reflectors deployed by forming a hierarchy. This time, the observed convergence time was longer with MPLS then with Pervasive BGP. The gain in convergence time obtained thanks to the avoidance of forwarding loops with MPLS was counterbalanced by the longer time required to install a new BGP route with a MPLS label in the FIB.

Note that our goal is not to compare the convergence times obtained with MPLS against the ones obtained with Pervasive BGP. However, we can make the general observation that, when an iBGP convergence happens without transient lack of alternate paths among the routers, using MPLS helps in avoiding the forwarding loops implied by the transient inconsistency of the BGP tables of the routers. On the other hand, when transient lack of alternate paths can occur among some routers, a rapid installation of received paths is required for a fast convergence to take place. In some platforms, installing BGP routes with MPLS labels can take a longer time than installing pure IP routes, so that the convergence time can be longer.

IV. The solution

In this section, we present a mechanism to perform a convergence that anticipates the maintenance of a peering link without loosing packets. The solution depends on the policies that are applied by the ISP. In this paper we only consider usual peering relationships and policies, i.e. Customer-Provider peerings and Shared Cost peerings.

We consider that the local pref assigned by the AS Border Routers to incoming BGP routes fall in three distinct ranges $[provider_{min}, provider_{max}]$, $[peer_{min}, peer_{max}]$ and $[client_{min}, client_{max}]$ such that client routes are preferred to routes received from Shared Cost peers which are themselves preferred to routes received from providers. Using ranges allows the ISP to define preferences within each class of routes. Also, only routes received over client peering links are propagated to Shared Cost peers and providers.

Note that the solution can be easily extended to other kinds of policies like for example backup peerings for customer-to-provider peerings and shared-cost peerings.

We firstly present the solution when a shutdown of a Customer-Provider link is performed at the provider side. After that, we examine the case of a shutdown performed at the customer side. Next, we look at the problem of shutting down a Shared-Cost peering link. It is to be noticed that in all those cases, there is always at least one peering link to backup the peering link being shut down.

Due to space limitations, we could not present the solution for the cases when this property is not verified in this paper.

The idea underlying the scheme is the following. Firstly, a link remains up while routers adapt to its scheduled removal. As we have seen, this is not sufficient to avoid all packet loss, because some BGP routers can lack of alternate path. Thus routers are also allowed to keep using the paths via the link until they find alternate ones. These compromised paths will be made less preferable to other paths, so that the convergence process will take place and alternate paths will be spread across the network.

A. Shutting down a $Provider \rightarrow Customer$ link

Let us assume that a peering link is shut down in an ISP AS. This link goes between an internal router PE and a client router CE of a neighbouring client AS. In order to avoid packet loss, two problems have to be solved. The first one is to ensure that the routers within the local AS stop using the routes towards the destinations that were reached via $PE \rightarrow CE$ without loosing packets. The second one is to ensure that the routers behind CE stop using the routes for the destinations that they used to reach via $CE \rightarrow PE$ without loosing packets. As a consequence the traffic must be uninterrupted in both directions thus ensuring a convergence without packet loss.

1) Outgoing problem: The simple topology shown in Figure 4 will be used to understand the problem. In this topology all the routers in the AS are fully meshed at the iBGP level. We assume that the link between $PE1$ and $CE1$ will be shut down by an operator of AS1, a provider of AS2. Let us

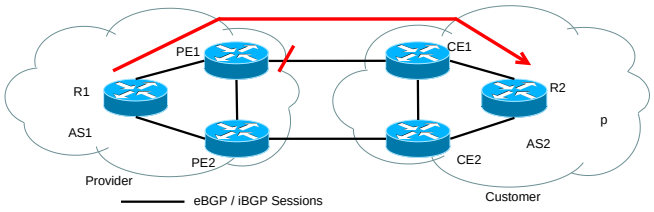

Fig. 4. A dually connected client

assume that $PE1$ keeps the link up for a while, and behaves as if the eBGP session was down. Consequently, $PE1$ sends withdraw messages to its iBGP peers for the paths received from $CE1$ that it selected as its best paths. The recovery will be performed by using the paths along $PE2 \rightarrow CE2$.

A transient unreachability for a prefix p may still occur because routers may have only known about the route for p via $PE1 \rightarrow CE1$ in their Adj-Rib-In, and thus will drop packets destined to p until they receive the alternate path. For example, if there is an agreement between AS1 and AS2 to let AS2 perform incoming traffic engineering by using communities or MED, $PE2$ may prefer the path via $PE1 \rightarrow CE1$ over the path received on its own eBGP peering link with $CE2$, so that secondary path will not be propagated towards $R1$ and $PE1$. When $R1$ receives the withdraw from $PE1$ for the path towards p, it will start dropping packets for this prefix. When $PE2$ receives a withdraw for a prefix p from $PE1$, it will select its external route for p and propagate it on its iBGP sessions. The recovery will only be accomplished when $R1$ receives the update from $PE2$.

This example illustrates that ensuring the forwarding along the link being shut down is not sufficient to provide a zero packet loss convergence. In addition, routers should avoid the paths that will become invalid, as soon as possible, while allowing those routers to still use these paths if they do not have alternate paths. To do that without modifying BGP, we must use a two step approach. First, we must render the affected paths less preferable than any other available path. Thus, the attributes of those paths must be modified to impact their quality at the very first step of the BGP decision process.

To do that, we can set the local-pref attribute of those paths to 0, and let the router performing the shutdown propagate updates for those paths. In the example above, $PE1$ should do this. When the other routers have switched to the alternate paths, $PE1$ will withdraw the old paths. This operation will have no impact on the forwarding since those paths are no longer used for forwarding.

BGP does not currently allow one ASBR $PE1$ to explicitely detect wether a distant node $PE2$ has actually stopped using an obsolete path. An operator may thus have to rely on information such as the number of prefixes affected by the maintenance operation, or on the traffic that is being forwarded on the link to be shut down to find the time required by each step of the procedure.

This solution is adequate in the context of VPN, where

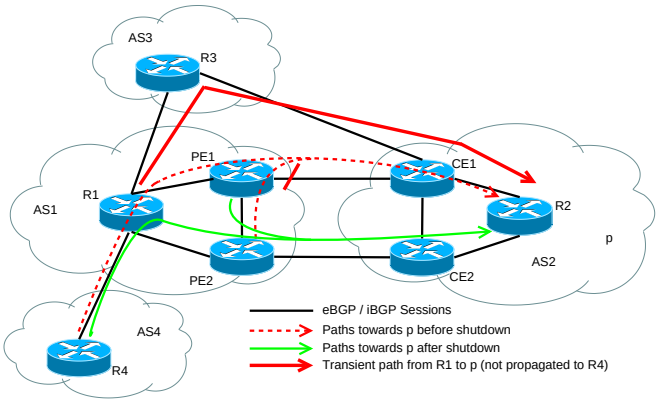

Fig. 5. A dually connected client

the possible alternate paths always come from the same neighbouring ASs. It also works in the case where the client is a single-homed AS that is dually connected to its provider.

However, the solution is not safe enough in the case of regular internet traffic, where some alternate paths for client routes could be advertised by providers. The topology depicted in Figure 5 will be used to illustrate this issue when using the solution above. In this case, AS3 is a provider of AS1, and AS4 is a shared-cost peer of AS1.

During the first step, $PE1$ will send an update with a local-pref attribute of 0 toward $R1$ for the path towards prefix p. For the same reason as described in the previous example, $R1$ may not know about the path via $PE2 \rightarrow CE2$. In this case, it becomes thus very possible that $R1$ selects an alternate path via $R1 \rightarrow R3$, which would be the only one available. But according to the usual peering relationships among neighbouring ASs, this provider route cannot be propagated to the providers or shared cost peers of AS1. Consequently, $R3$ will send a withdraw to $R4$ for the prefix p and the recovery will only be performed once R1 receives the alternate client route, selects it as its best route, and sends an update to $R4$. Meanwhile $R4$ will transiently drop the packets destined to p.

Hence a local-pref of 0 cannot be used in the first step anymore. However, a local pref that is lower than any local pref assigned to client routes within the ISP and higher than any local-pref assigned to the routes receives from providers and shared cost peers can be used. As a result routers will keep using the routes until new clients routes are propagated. Thus only path updates messages will be propagated to providers and shared cost peers, instead of abrupt withdraws.

2) Incoming problem: The routers on the neighbouring AS have to be forced to stop using the link being shutdown. In the example topology of Figure 4, the routers in AS2 should stop using routes passing through $CE1 \rightarrow PE1$.

The first solution is basically to contact the operators of $AS2$ and let them use the same technique as described in the preceding section. Although it works, it is not very convenient because it requires synchronizing operating teams. Furhtermore, maintenance is generally performed during the least disturbing time periods for the client (during the night for

example). The maintainance operation may require the client to assign dedicated human resources for this task, which is unfortunate. It may also induce additional financial cost for the provider and the client. Moreover this task becomes a real scheduling nightmare when the maintenance affects multiple clients at a time, e.g. in the case of a linecard removal or a whole shutdown for a typical provider edge router.

A simpler solution is to have the provider agree with each client on a community that would be dedicated to routes that have to be avoided by the client. When the provider performs the shutdown, it will re-advertise its routes by tagging them with this community. On the client side, the routers will have been pre-configured to set a local-pref value of 0 to all the routes tagged with this community. After a while, the router on the provider side will send a Cease Notification to actually withdraw the routes and shut down the session.

This solution is applicable without modifying BGP. However, is not very fast as it requires the PE to re-advertise all its routes towards the CE. In addition, it may create a large amount of update messages.

Another solution is to implement a new BGP message which would simply mean that the session will be shut down within a given amount of time, and that the CE should adapt to it by using the techniques described above. This message can be for instance an eBGP Cease Notification message as defined in [8], with a new sub code or a dynamic capability.

B. Shutting down a $Customer \rightarrow Provider$ link

When a $Customer \rightarrow Provider$ peering link is shut down at the customer side, the proposed behaviour of the routers is similar to the one proposed when the provider performs the shutdown. We will thus briefly resume the behaviour that should be applied.

1) Outgoing problem: When a peering link between a customer and its provider is shut down at the customer side, the router where the shutdown command is issued must set a local pref of 0 for the routes that it received over the impacted link, in order to reroute the traffic that was going from the Customer towards the Provider. This will force routers on the customer side to select paths received over other peering links with providers.

2) Incoming problem: When the graceful shutdown is performed by using an agreement between the customer and the provider, the local-pref value that has to be set by the provider must be lower than any local pref assigned to client routes within the ISP, and higher than any local-pref assigned to the routes received from providers and shared cost peers. This will force the routers on the provider side to select paths via other peering links with clients. After a while, the local-pref value will be set to 0 in order to let routers of the provider select alternate paths via other peering links (shared-cost or provider peering links), for the prefixes for which no alternate paths via customers could be found. Finally, the eBGP session can be shut down.

C. Shutting down a Shared-Cost peering Link

The simpler solution in the case of a Shared Cost peering link shutdown is to also set a local-pref value of 0 on the routes received over this link to solve the **outgoing problem**, and to also use an agreement on a dedicated community to solve the **incoming problem**.

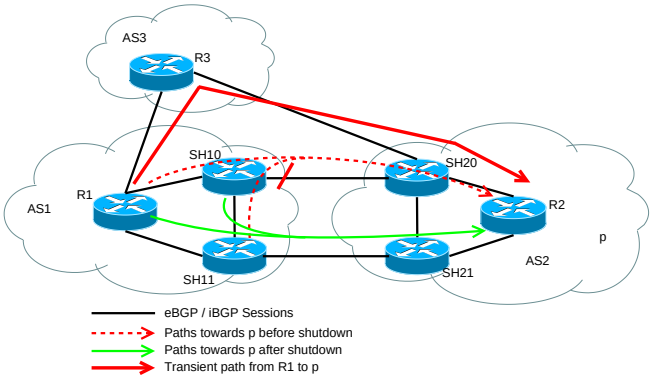

Fig. 6. Dually connected Shared-Cost peers

However, this could cause a transient utilization of provider links to reroute the traffic even if, after the convergence of the network, alternate paths through other Shared-Cost links will be used. For example, in Figure 6, let us assume that AS1 and AS2 have a Shared Cost peering relationship. Once again, it is possible that $SH11$ prefers a path via $SH10 \rightarrow SH20$ to reach p, so that it does not propagated its own alternate path into AS1. When the link between $SH10$ and $SH20$ is to be shut down on $SH10$, $SH10$ would send local pref updates to 0 for its path towards p. $R1$ would then select the only alternate path that it knows at that time, which is a provider path via AS3. Finally, when $SH11$ selects and propagate its own path via $SH11 \rightarrow SH21$, $R1$ will prefer this path via this Shared Cost peering link and reroute again.

This scenario does not provoke packet loss as for the Provider-Customer link shutdown case. Indeed, according to the usual peering relationship model, routes received over shared cost links or links with providers are only propagated to clients. Thus, a router switching from one kind of route to the other will not send abrupt withdraw to its peers. However, the operators of AS1 might not want to transiently use alternate paths via providers if some paths via shared cost peerings are available. Besides, it would limit the unuseful exploration of some paths and if Pervasive BGP is used it would therefore reduce the LoC caused by forwarding loops during this exploration.

The reassignment of the local-pref to the routes becoming invalid should then be done with a value that is lower than all the ones assigned to Shared-Cost routes, but higher than the local-pref values assigned to provider routes. As a second step, a local-pref value of 0 should be re-assigned to those routes to face the case where no alternate path can be found for some prefixes over other Shared-Cost peering links.

D. Using other attributes to perform a loss free convergence

There are cases where the local pref attribute is not sufficient to select a best path, and a tie-break is performed between paths having the same local pref value on the basis of the IGP distance of their respective nexthops. In such cases, the IGP distance of the nexthop could be increased, or the MED attributes of the affected paths could be lowered to let routers switch to the alternate paths.

However, as the local-pref based solution covers the broadest scenario space, due to the position of the local-pref rule in the BGP decision process, we recommend to always tune the local-pref attribute, for the sake of simplicity.

E. Bringing a BGP peering link up

Unfortunately, when an operator brings an eBGP peering link up, transient LoC can also occur. Indeed, when the ASBR on which the peering link will be brought up will start propagating paths through the iBGP topology, some other ASBRs may start considering such new paths as better than their own external paths, because these win a tie-breaking rule that precedes the IGP distance tie-break. As a consequence, such ASBRs will withdraw the paths that they had initially advertised. If a router receives such withdraw before being aware of the new paths, it may lack of paths towards the concerned prefixes and start dropping packets.

As an exemple, let us consider the scenario illustrated in Figure 4. Let us assume that the peering link between $PE1$ and $CE1$ was down, and is brought back up. Let us assume that $PE1$ learns a path towards p that has a MED value lower then the one of the path towards p via the peering link between $PE2$ and $CE2$. As a consequence, when $PE1$ advertises this path in iBGP, $PE2$ will switch to that path and withdraw its own path. If $R1$ processes the withdraw message from $PE2$ before receiving the update message from $PE1$, it transiently has no path towards p and starts dropping packets.

Handling such situations with operational procedures is difficult because the emitters of the withdraw messages are distant from the location of the link up operation.

When the iBGP topology is a full-mesh, enabling "advertise best external" features [9] on the ASBRs avoids the problem, as ASBRs will advertise on their iBGP sessions their best external paths even if their actual best path is a path learned over iBGP. In our example, this means that when $PE2$ learns the new path via $PE1$, it switches to this path but will not withdraw, over its iBGP sessions, the path via its own eBGP peering link.

In the general case where Route Reflectors are used, using such features helps in reducing the LoC, but is not sufficient in theory to ensure a packet lossfree convergence.

Let us illustrate this problem in figure 7. Let us assume that the peering link $PE_2 \leftrightarrow CE_1$ is brought up, and that a path towards prefix p is advertised by CE_1 to PE_2. As illustrated in the figure, let us consider that AS Path prepending is done on the path for p advertised by CE_1 over $PE_1 \leftrightarrow CE_1$, while it is not on the path for p advertised by CE_1 over $PE_2 \leftrightarrow CE_1$. When RR_3 receives the path to p via $PE_2 \rightarrow CE_1$,

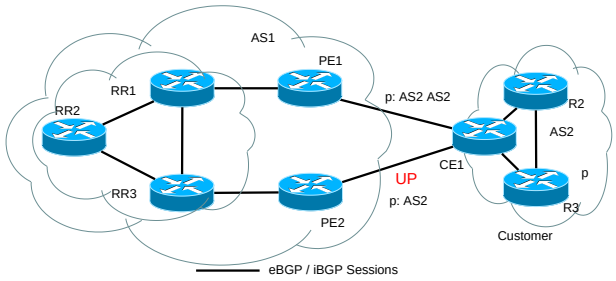

Fig. 7. A dual-homed Stub

it selects it as its best path. As this path was learned from a client peer in the Route Reflector hierarchy, RR_3 will advertise this path to its non-client iBGP peers. When RR_1 processes this advertisement, it runs its BGP decision process, selects the path via $PE_2 \rightarrow CE_1$, and, by respecting [10], sends a withdraw message for the path p via $PE_1 \rightarrow CE_1$ to RR_1 and RR_3. Indeed, a route reflector RR_1 as defined in [10] is supposed not to advertise its best path for a prefix p to a non-client peer RR_2 if this best path was learned from another non-client peer RR_3. In theory, RR_2 could receive such a withdraw message from RR_1 before receiving the update message from RR_3, so that RR_2 could transiently lack of a path towards p.

These LoC situations are difficult to reproduce in practice, but could happen in theory. However, as this paper focuses on procedures that operators can apply to reduce the LoC, we cannot recommend some that would work with the current BGP routing suite without being way too impractical.

V. Measurements

We simulated the solution exposed for the above by using the test topologies that were designed to quantify the loss of packets happening during planned maintenance operations. We performed a manual shutdown of a $Customer \rightarrow Provider$ link, by issueing BGP commands by hand in the routers, to reproduce the automatic behaviour that we propose. The results show the improvement brought by the solution. It shows that LoC are totally avoided in MPLS networks, and much reduced in networks using Pervasive BGP.

The solution has been implemented using BGP policies and communities exchanged between $lr2$ and $lr11$. $lr11$ performs the maintenance operation thus: $lr11$ sends a community attribute to $lr2$ while advertising the routes coming from $lr2$ with a local-pref attribute of 0. When $lr2$ receives the community, it re-advertises the routes with a local-pref attribute of 0. The link is then shut down after enough time is given to both AS to converge.

The results are shown in the tables below. In each table, the LoC time is measured for both directions. The second column results are for the actual BGP behaviour and the third column results are the results achieved with the policies for planned maintenance operations.

All the results are in seconds with an accuracy of 2ms in the worst case.

iBGP organization	Current BGP Behaviour	Planned Behaviour	Stream direction
Full-mesh	1.879	0.000	lr12 → lr9
	0.248	0.064	lr9 → lr12
RR	2.058	0.000	lr12 → lr9
	1.220	0.256	lr9 → lr12
HRR	1.707	0.000	lr12 → lr9
	1.146	0.198	lr9 → lr12

TABLE III
LOSS OF CONNECTIVITY (SECONDS) - PERVASIVE BGP - IGP USED AS TIE-BREAK

iBGP organization	Current BGP Behaviour	Planned Behaviour	Stream direction
Full-mesh	1.982	0.000	lr12 → lr9
	2.298	0.390	lr9 → lr12
RR	2.155	0.000	lr12 → lr9
	2.407	0.333	lr9 → lr12
HRR	2.183	0.000	lr12 → lr9
	2.180	0.212	lr9 → lr12

TABLE IV
LOSS OF CONNECTIVITY (SECONDS) - PERVASIVE BGP - LOCAL PREF USED AS TIE-BREAK

iBGP organization	Current BGP Behaviour	Planned Behaviour	Stream direction
Full-mesh	0.000	0.000	lr12 → lr9
	0.000	0.000	lr9 → lr12
RR	0.000	0.000	lr12 → lr9
	0.983	0.000	lr9 → lr12
HRR	0.000	0.000	lr12 → lr9
	3.367	0.000	lr9 → lr12

TABLE V
LOSS OF CONNECTIVITY (SECONDS) - MPLS FORWARDING PLANE - IGP USED AS TIE-BREAK

iBGP organization	Current BGP Behaviour	Planned Behaviour	Stream direction
Full-mesh	0.000	0.000	lr12 → lr9
	1.948	0.000	lr9 → lr12
RR	0.000	0.000	lr12 → lr9
	1.709	0.000	lr9 → lr12
LP\HRR	0.000	0.000	lr12 → lr9
	3.266	0.000	lr9 → lr12

TABLE VI
LOSS OF CONNECTIVITY (SECONDS) - MPLS FORWARDING PLANE - LOCAL PREF USED AS TIE-BREAK

The absolute figures are highly hardware and software dependant: a smaller/older router would have experience longer LoC. The purpose of these tests is not to measure an exact LoC due to the BGP convergence, but to evaluate the relative gain of using a BGP planned maintenance procedure.

The results show that the use of the solution for planned maintenance operations reduces a lot the LoC time.

For the Pervasive BGP case (Tables III and IV), there is still some LoC for the packet stream from $lr9$ to $lr12$, that are due to transient loops during the iBGP convergence. This means that, during the convergence phase, even if all the routers always have a valid path towards all the prefixes that are impacted by the shutdown, some transient inconsistencies among the FIB of the routers led to forwarding loops, which led to packets being dropped. However, the LoC is very much reduced. Indeed, with the solution, forwarding loops are the only component of the LoC, as lacks of alternate paths do not occur anymore.

No LoC occured for the packet stream from $lr12$ to $lr9$. In fact, only $lr11$ and $lr14$ updated their FIB during the convergence phase inside the customer network, to use the peering link $lr10 \rightarrow lr1$ instead of $lr11 \rightarrow lr2$. As we modified the **export** policy of $lr11$ to let it propagate local pref updates in the network, $lr14$ always updated its FIB before $lr11$. $lr11$ updated its FIB when the actual shutdown command had been issued. This explains why transient loops did not occur on the customer side.

For the MPLS forwarding case (Figures V and VI), not a single packet is lost when using the solution. With MPLS, when an Ingress Router receives a packet, it will perform a lookup in its BGP table to find the Egress point for this packet. The packet is then tunnelled towards the Egress point, so that the intermediate routers will not perform a BGP lookup to forward this packet. Thus, even if some BGP routing tables could have been inconsistent during the convergence, as one single router is doing a lookup in a BGP table for each packet, the packets will reach one of both egress routers, which will forward the packet towards the other network. When the solution is not used, packets of the stream from $lr9$ to $lr12$ can be lost, because some routers temporarily lack of a path towards the destination of the stream. Packets of the stream from $lr12$ to $lr9$ were not lost, because the routers of the customer always have an alternate path towards the destination during the convergence phase. These last measurements confirmed that the packet lost for this stream in the IP forwarding mode where only caused by forwarding loops.

VI. RELATED WORK

To the best of our knowledge, this is the first proposed solution to the problem of packet loss during maintenance operations on BGP peering links. Several solutions have been proposed to support maintenance operations for other components of ISP Networks. In [11], [12], a mechanism is proposed to shutdown or install intra-domain links and routers without loosing packets. Other solutions for the intra-domain problem have been proposed in [13], [14]. MPLS make-before-break solutions have been proposed in [15] to provide the feature in MPLS Networks using RSVP.

Graceful Restart extensions are current topics of various working group of the IETF. These solutions apply for maintenance operations that do not jeopardize the forwarding of packets, typically when rebooting the control plane of a router [16], [3]. These solutions solve differents problems, as their

goal is to let packets be forwarded along the same paths during the maintenance. Here, we solve the problem of converging to alternate paths without loosing packets when the initial paths are made invalid by the maintenance operation.

In [17], a solution is proposed to quickly deviate packets on an alternate path once a sudden failure of a peering link occured. In the case of a long lasting failure, the network must adapt to the topological change. If the peering link is protected with such a solution, the convergence is not urgent. So, the solution presented here can also be used as a complement to this technique to avoid packet loss in the case of an urgent failure of a protected link.

Other proposals to improve the convergence of BGP have been proposed to fasten the convergence, notably in [18], [19]. While these techniques improve the recovery time of BGP in the case of sudden failures, what is proposed in this paper is to take advantage of the non urgent nature of the convergence process following maintenance operations to perform a packet loss free convergence.

VII. Conclusion

Respect of Service Level Agreements with tight performance constraints are key quality evaluation aspects in the Internet connectivity and VPN markets. In such a context, network topology changes due to hardware and software upgrades and maintenance are common operations that lead to transient Losses of Connectivity. This is considered as very unfortunate by ISPs as these events are predictable, so that it can be considered as of common sense that they should not be harmful to the reachability throughout the network.

In this paper, we illustrated with a review of ISP operational data that maintenance operations are common events in an ISP network. We also showed with a lab experiment that typical iBGP topology designs lead to Losses of Connectivity due to transient lacks of alternate paths within the routers. Then, we presented a solution to perform a make-before-break iBGP convergence in the case of a manual shutdown of an eBGP peering link. The solution prevents packets from being dropped in networks using an encapsulation scheme, and reduces a lot the Losses of Connectivity when Pervasive BGP is used. The solution handles typical policies applied within commercial Autonomous Systems, and can be easily extended to other policies. We manually applied the solution by introducing BGP commands in a lab network, and measured the gain in terms of packet loss. The results show that the solution succeeds in avoiding packet loss, despite its simplicity, and we argue that the shutdown command of an eBGP session should provide an optional parameter such as "[graceful]" so that the solution could be applied automatically by routers when the operator performs a maintenance.

Acknowledgments

This paper has been supported by Cisco Systems within the ICI project, and by France Telecom within project ER46-126-699. Bruno Decraene is partially supported by the IST AGAVE Project. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of Cisco Systems or France Telecom.

References

[1] F. Wang, Z. M. Mao, J. Wang, L. Gao, and R. Bush, "A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance," in *Proc. of ACM SIGCOMM*, September 2006.

[2] V. Gill, "Panel on BGP," April 2006, presented at Infocom 2006, http://www.ieee-infocom.org/2006/panelist/infocom-panel2-vijay.pdf.

[3] S. R. Sangli, E. Chen, R. Fernando, J. Scudder, and Y. Rekhter, "Graceful Restart Mechanism for BGP," Internet Engineering Task Force, Request for Comments 4724, Januay 2007.

[4] F. Baker and P. Savola, "Ingress Filtering for Multihomed Networks," Internet Engineering Task Force, Request for Comments 3704, March 2004.

[5] M. Kolon, "Intelligent Logical Router Service," 2004, http://www.juniper.net/solutions/literature/white_papers/200097.pdf.

[6] J. Stewart, *BGP4 : Interdomain Routing in the Internet*. Addison Wesley, 1999.

[7] Y. Rekhter, T. Li, and S. Hares, "A Border Gateway Protocol 4 (BGP-4)," Internet Engineering Task Force, Request for Comments 4271, January 2006.

[8] E. Chen and V. Gillet, "Subcodes for BGP Cease Notification Message," Internet Engineering Task Force, Request for Comments 4486, April 2006.

[9] P. R. Marques, "BGP Network Design," September 2004, RIPE 49.

[10] T. Bates, R. Chandra, and E. Chen, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)," Internet Engineering Task Force, Request for Comments 4456, April 2006.

[11] P. Francois and O. Bonaventure, "Avoiding transient loops during IGP Convergence in IP Networks," in *Proc. IEEE INFOCOM*, March 2005.

[12] P. Francois, O. Bonaventure, M. Shand, S. Previdi, and S. Bryant, "Loop-free convergence using ordered FIB updates," March 2006, internet draft, draft-francois-ordered-fib-01.txt, work in progress.

[13] A. Atlas and A. Zinin, "Analysis and Minimization of Microloops in Link-state Routing Protocols," October 2005, Internet draft, draft-ietf-rtgwg-microloop-analysis-01.txt, work in progress.

[14] ——, "Basic Specification for IP Fast-Reroute: Loop-free Alternates," March 2007, internet draft, draft-ietf-rtgwg-ipfrr-spec-base-06, work in progress.

[15] Z. Ali, J. Vasseur, and A. Zamfir, "Graceful Shutdown in GMPLS Traffic Engineering Networks ," September 2006, internet draft, draft-ietf-ccamp-mpls-graceful-shutdown-00.txt, work in progress.

[16] J. Moy, P. Pillay-Esnault, and A. Lindem, "Graceful OSPF Restart," Request for Comments 3623, November 2003.

[17] O. Bonaventure, C. Filsfils, and P. Francois, "Achieving sub-50 milliseconds recovery upon bgp peering link failures," in *CoNEXT'05: Proceedings of the 2005 ACM conference on Emerging network experiment and technology*. New York, NY, USA: ACM Press, 2005, pp. 31–42.

[18] D. Pei, M. Azuma, N. Nguyen, J. Chen, D. Massey, and L. Zhang, "BGP-RCN: Improving BGP convergence through Root Cause Notification," *Computer Networks*, vol. 48, no. 2, pp. 175–194, June 2005.

[19] J. Chandrashekar, Z. Duan, Z. Zhang, and J. Krasky., "Limiting path exploration in BGP," in *IEEE INFOCOM*, Miami, Florida, March 2005.