Core Backbone Convergence Mechanisms and Microloops Analysis

Abdelali Ala
Abdelmalik Essâadi University
Faculty of Sciences
Information and Telecom Systems Lab
Tetuan, Morocco
+212 6 65 24 08 28

Driss El Ouadghiri
Moulay Ismail University
Faculty of Sciences
Mathematics and Science Computer
Meknes, Morocco
+212 6 02 23 61 55

Mohamed Essaaidi
Abdelmalik Essâadi University
Faculty of Sciences
Information and Telecom Systems Lab
Tetuan, Morocco
+212 6 61 72 59 92

***Abstract*— In this article we study approaches that can be used to minimise the convergence time, we also make a focus on microloops phenomenon, analysis and means to mitigate them. The convergence time reflects the time required by a network to react to a failure of a link or a router failure itself. When all nodes (routers) have updated their respective routing and forwarding databases, we can say the network has converged. This study will help in building real-time and resilient network infrastructure, the goal is to make any evenement in the core network, as transparent as possible to any sensitive and real-time flows. This study is also, a deepening of earlier works presented in [10] and [11].**

Keywords-component: FC(Fast-convergence); RSVP(ressource reservation protocol); LDP (Label Distribution Protocol); VPN(Virtual Private Network); LFA (loop free alternate); MPLS (Multiprotocol Label Switching); PIC(Protocol independent convergence); PE(Provider edge router); P(Provider core router).

I. INTRODUCTION

Mpls/vpn backbones are widely used today by various operators and private companies in the world, high to medium-sized companies build their own Mpls/vpn backbone or use services of an operator . Real time applications like voice and video are more and more integrated to end user applications, making them ever more time sensitive.

Operators are offering services like hosting companies' voice platforms, VoIP call centers, iptv...Etc. All these aspects make the convergence time inside the backbone a challenge for service providers.

However, the global convergence time is an assembly of several factors including: link or node failure detection, IGP failure detection, LSP Generation, SPT Computation, RIB update, local FIB creation and distribution ...updates signaling...etc.

Based on analysis and statistics of large backbone possibilities we have delimited our convergence target as follows:

[PE to P] convergence, in other terms [PE to core] must be under sub-second, hopefully under 50 msec, even on highly loaded PE, the convergence time should be almost independent of vpnv4, 6PE, 6VPE or igp prefixes number…[P to PE] and [P to P] convergence must stay under sub-second and consistent in both directions: [core to PE], [PE to core].

From the customer point of view: the overall [end-to-end] convergence should stay under 1 sec (no impact on most time sensitive applications). A lot of approaches can be used to minimise the convergence time, our approach consists on enhancements and optimizations in control and forwarding plane. While a lot of things can also be made at the access, the scope of our work is the core backbone.

Not only a backbone design must take into account criterion like redundant paths at each stage, but redundancy at the control plane only, does not make a lot of sense if, in the forwarding plane, backup paths are not pre-computed. We can say that a backbone meets a good convergence design if at each segment of the tree structure; we are able to calculate the time it takes for flows to change from the nominal path to the backup one.

On the other hand, temporary microloops may occur during the convergence interval, indeed, after a link or node failure in a routed network and until the network re-converges on the new topology, routers several hops away from the failure, may form temporary microloops. This is due to the fact that a router's new best path may be through a neighbor that used the first router as the best path before failure, and haven't had yet a chance to recalculate "and/or" install new routes through its new downstream. We can understand microloops are transient and self-corrected, however depending on their duration, the CPU load on the control plan may increase to 100%, so in addition to mitigation methods presented in this article, some cpu protection mechanisms are also discussed. The approach used in this article is theory against lab stress and result analysis. The aim of the study is to give an accurate idea of gains and drawbacks of each method, and show when one or the other method more fits the network topology.

II. FAST CONVERGENCE MODELS

In an attempt to construct a model for IGP and BGP protocols, we must take into account the following components:

- Time to detect the network failure, e.g. interface down condition.
- Time to propagate the event, i.e. flood the LSA across the topology.
- Time to perform SPF calculations on all routers upon reception of the new information.
- Time to update the forwarding tables for all routers in the area.

And then modelise the IGP Fast Convergence by a formula which is the sum of all the above components:

IFCT = (LFD + LSP-GIF + SPTC + RU + DD)

And BGP Fast Convergence model as:

BFCT= IFCT + CRR

Where:

IFCT = IGP Fast Convergence Time

LFD = Link Failure Detection (Layer 1 detection mechanisms)

LSP-GIF = LSP Generation, Interval and Lifetime

SPTC = SPT Computation

RU = RIB Update

DD = *Distribution* Delay

BFCT = BGP Fast Convergence Time

CRR = CEF Recursive Resolution for BGP Prefixes

III. Link Failure Detection Mechanism

The ability to detect that a failure has happened is the first step to towards providing recovery, and therefore, is an essential building block for providing traffic protection. Some transmission media provide hard-ware indications of connectivity loss. One example is packet-over-SONET/SDH where a break in the link is detected within milliseconds at the physical layer. Other transmission media do not have this ability, e.g. Ethernet (note that the fast detection capability has been added to optical Ethernet).

When failure detection is not provided in the hardware, this task can be accomplished by an entity at a higher layer in the network. But there is disadvantage to that, using IGP hello as example: We know that IGPs send periodic hello packets to ensure connectivity to their neighbors. When the hello packets stop arriving, a failure is assumed. There is two reasons why hello-based failure detection using IGP hellos cannot provide fast detection times:

- The architectural limit of IGP hello-based failure detection is 3 seconds for OSPF and 1 second for ISIS. In common configurations, the detection time ranges from 5 to 40 seconds.
- Since handling IGP hellos is relatively complex, raising the frequency of the hellos places a considerable burden on the CPU.

IV. Bidirectional Forwarding Detection (BFD)

The heart of the matter lies in the lack of a hello protocol to detect the failure at a lower layer. To resolve this problem, Cisco and Juniper jointly developed the BFD protocol. Today BFD has its own working group (with the same name IETF [BFD]). So what exactly is BFD ?

BFD is a simple hello protocol designed to provide rapid failure detection for all media types, encapsulations, topologies, and routing protocols. It started out as a simple mechanism intended to be used on Ethernet links, but has since found numerous applications. Its goal is to provide a low-overhead mechanism that can quickly detect faults in the bidirectional path between two forwarding engines, wether they are due to problems with the physical interfaces, with the forwarding engines themselves or with any other component. But how can BFD quickly detect such a fault ?

In a nutshell, BFD is exchanging control packet between two forwarding engines. If a BFD device fails to receive a BFD control packet within the detect-timer:

*(Required Minimum RX Interval) * (Detect multiplier)*

Then it informs its client that a failure has occurred. Each time a BFD successfully receives a BFD control packet on a BFD session, the detect-timer for that session is reset to zero. Thus, the failure detection is dependent upon received packets, and is independent of the receiver last transmitted packet. So we can say that expected results depend on the platform and how the protocol is implemented, but available early implementations can provide detections in the range of tens of milliseconds.

V. MPLS LDP-IGP Synchronization

A. *Feature Description*

Packet loss can occur when the actions of the IGP (e.g. ISIS) and LDP are not synchronized. It can occur in the following situations:

- When an IGP adjacency is established, the router begins forwarding packets using the new adjacency before the LDP label exchange ends between the peers on that link.

If an LDP session closes, the router continues to forward traffic using the link associated with the LDP peer rather than an alternate pathway with a fully synchronized LDP session.

To solve the first point, the following algorithm is being used: If there is a route to the LDP peer, IGP adjacency is held down, waiting for LDP synchronization to be completed; in other words, waiting for labels exchange to be completed. By default, adjacency will stay down for ever if LDP does not synchronize. This default behavior is tunable via configuration command “mpls ldp igp sync hold-down <duration in ms>” to specify the maximum amount of time the adjacency will stay down. At expiration of this timer, the link will be advertised, but with metric set to maximum in order to avoid using this link. If there is no route to the LDP peer, IGP adjacency is brought up, but with a metric set to the maximum value in order to give a chance for the LDP session to go up. In this

case, once the LDP session goes up and finishes labels exchange, the IGP metric reverts back to its configured value.

To solve the second point, the feature will interact with IGP to modify link metric according to LDP session state. As soon as LDP session is going down, the IGP metric of the related link is set to its maximum. Then, others nodes on the network can compute a new path avoiding to use this link.

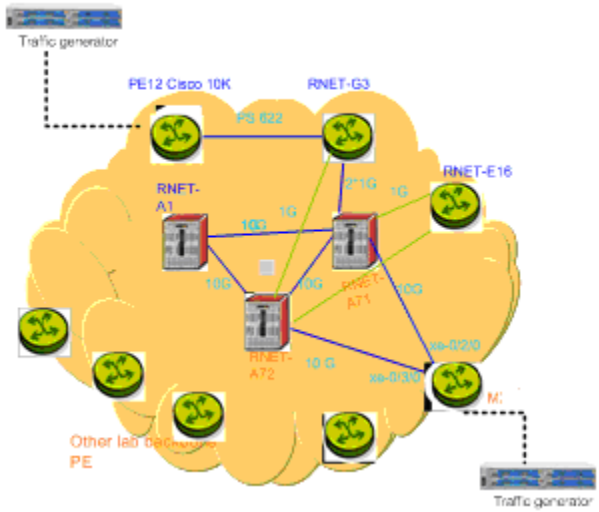

Figure 1. Lab setup diagram

B. *Test Description*

On the M1 router, we configure the ldp-synchronization under isis protocol, interface xe-0/2/0.0 (timer set to T1 sec) and under ldp protocol: (timer set to T2 sec). The timer under the ISIS section will set how much time ISIS will stay sending the infinite metric once it has been warned by LDP that its sessions are up. The timer under LDP section will set how much time LDP wait to warn the IGP once its sessions are up; by default this timer is equal to 10 sec.

```
M1-RE0# run show configuration protocols isis
traceoptions {
    file isis size 5m world-readable;
    flag ldp-synchronization send receive detail;
    flag lsp-generation detail;
----------truncated----------
interface xe-0/2/0.0 {
    ldp-synchronization {
        hold-time  "T1";}
    point-to-point;
    level 2 metric 10;
}
interface xe-0/3/0.0 {
    ldp-synchronization {
        hold-time  "T1"; }
    point-to-point;
    level 2 metric 100;

M1-RE0>show configuration protocol ldp
track-igp-metric;
----------truncated----------
igp-synchronization holddown-interval "T2";

M1-RE0>show configuration interfaces xe-0/2/0
description "10 GIGA_LINK_TO_PPASS_P71 through Catalyst TenGigabitEthernet2/5";
vlan-tagging;
mtu 4488;
hold-time up 5000 down 0;  / time here is in milliseconds /
```

While isis adjacency is operational, the ldp session is turned down (deactivation of xe-0/2/0.0 under ldp protocol on the MX side).

We look at the debug file on the MX and the isis lsp received on PE12 rising to infinite the isis metric toward RNET-A71.

```
PE-10K#show isis database M1-RE0.00-00 detail
S-IS Level-2 LSP M1-RE0.00-00
LSPID                 LSP Seq Num  LSP Checksum  LSP Holdtime  ATT/P/OL
M1-RE0.00-00   0x00000B71  0x7FFE       65520         0/0/0
  Area Address: 49.0001
  NLPID:        0xCC 0x8E
  Router ID:    10.100.2.73
  IP Address:   10.100.2.73
  Hostname: M1-RE0
  Metric: 16777214   IS-Extended RNET-A71.00
  Metric: 100        IS-Extended RNET-A72.00
  Metric: 100        IP 10.0.79.56/30
  Metric: 10         IP 10.0.79.52/30
```

After the expiration of (the configured hold-down timer) we can see that the metric is updated and set to the initial value.

```
PE-10K#show isis database M1-RE0.00-00 detail
IS-IS Level-2 LSP M1-RE0.00-00
LSPID                 LSP Seq Num  LSP Checksum  LSP Holdtime  ATT/P/OL
M1-RE0.00-00   0x00000B72  0x8FE2       65491         0/0/0
  Area Address: 49.0001
  NLPID:        0xCC 0x8E
  Router ID:    10.100.2.73
  IP Address:   10.100.2.73
  Hostname: M1-RE0
  Metric: 10         IS-Extended RNET-A71.00
  Metric: 100        IS-Extended RNET-A72.00
```

The duration of the infinite metric must cover the necessary time for a full labels exchange after the rising of the ldp session.

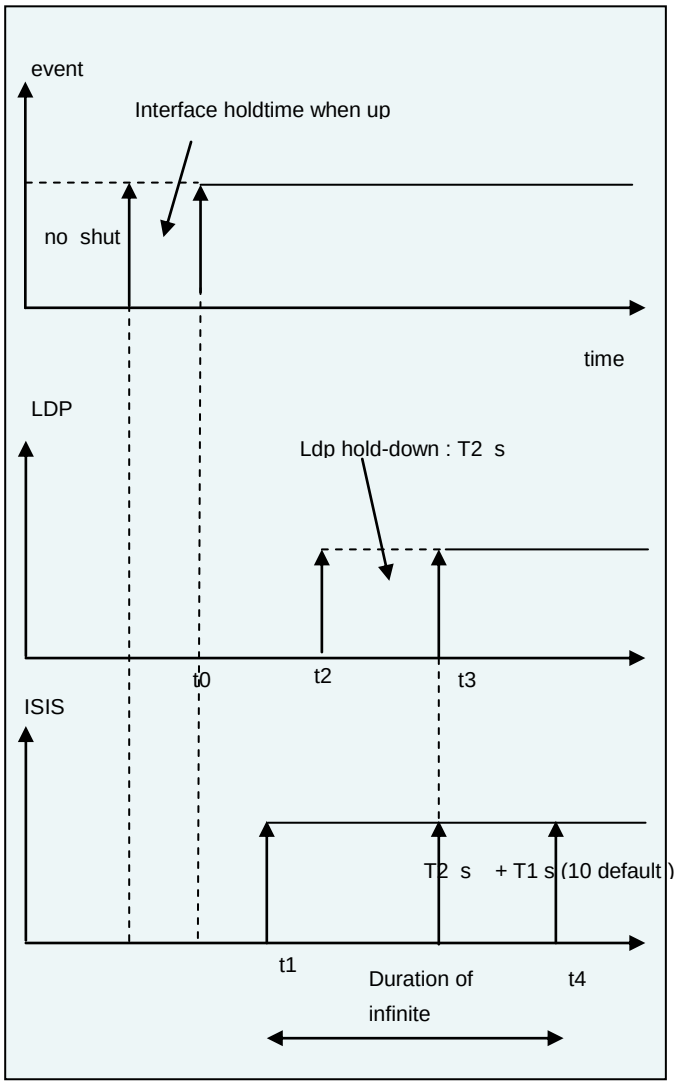

Figure 2. ldp-igp synchronization chronogram

VI. ISIS BACKOFF ALGORITHM

A. TUNING EXPLAINED

ISIS runs a Dijkstra-algorithm to compute the tree followed by a computation of the routing table. If the receipt of a modified LSP does affect the tree, an SPF (shortest path first calculation) is run; otherwise a simple PRC (partial route calculation) is run. An example of evenement that will trigger only a PRC is the addition of a loopback on a distant node (this does not change the tree, just one more IP prefix leaf is on the tree)

The PRC process runs much faster than an SPF because the whole tree does not need to be computed and most of the leaves are not affected.

However, by default, when a router receives an LSP which is triggering an SPF or a PRC, it does not start it immediately, it is waiting for a certain amount of time (5.5 seconds for SPF & 2 seconds for PRC).Lowering this initial "wait time" would significantly decrease the needed convergence time.

On the other hand, it is necessary to leave enough time to the router to receive all LSPs needed for computing the right SPF, so there is a lower limit not to be exceeded. Otherwise, If SPF computation starts before having received all important LSP, you may need to run another SPF computation a bit later. Then, overall convergence would not be optimal.

Between the first SPF (or PRC) and followings ones, the router will also wait for some times, default values are (5.5 seconds for SPF and 5 seconds for PRC). However the maximum amount of time a router can wait is also limited

(10 seconds for SPF *and* 5 seconds for PRC).

B. FEATURE USAGE IN OUR STUDY

The worst case, to take into consideration while choosing the initial wait time, is a node failure. In this situation, all neighbors of the failing node will send LSP reporting the problem. These LSP will be flooded through the whole network. Some studies indicate that 100 ms is enough for very large and wide networks.

So here our chosen values:

```
spf-interval 1 150 150
prc-interval 1 150 150
spf-interval <M> <I> <E>
prc-interval <M> <I> <E>
M = (maximum) [s]
I = (initial wait) [ms]
E = (Exponential Increment) [ms]
```

The same parameters have been applied on all routers to keep a consistency and same behavior on all nodes.

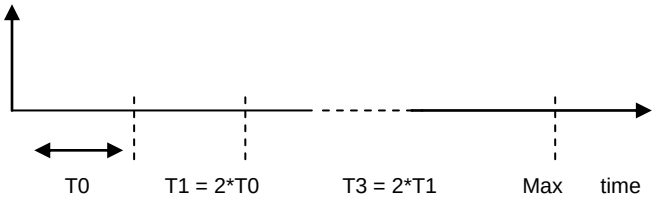

Figure 3. isis backoff algorithm timing

150 ms as initial waiting time for the first SPF calculation, then if there is a trigger for another SPF, the router will wait 300 ms, then wait 600 ms if there is a following one, until the max-value of 1000 ms. the waiting timer will stay equal to 1 second for as much as there is no trigger of a new calculation. In case there is no trigger during 1 second, the wait time is reset to the initial value and start as described in the *"Fig. 3"*.

C. MAIN GAIN FROM THIS TUNING

Simulations indicate that the most important gain is due to the first waiting timer decreased from default value to 150ms.

VII. BGP-4 SCALABILITY ISSUES (PROBLEM STATEMENT)

The BGP-4 routing protocol has some scalability issues related to the design of Internal BGP (IBGP) and External BGP (EBGP) peering arrangements.

IBGP and EBGP are the basically the same routing protocol just with different rules and applications.

- EBGP advertises everything to everyone by default.
- IBGP does not advertise "3rd-party routes" to other IBGP peers, this is because there is no way to do loop detection with IBGP

The RFC 4456 states that any BGP-4 router with EBGP peers must be fully meshed with all the other BGP-4 routers with EBGP peers in the same AS. This rule effectively means that every IBGP peers must be logically fully meshed. So you must have all BGP-speaking routers in your AS peer with each other. Below is a graphical example of a full-meshed 16-router . For more details see [15].

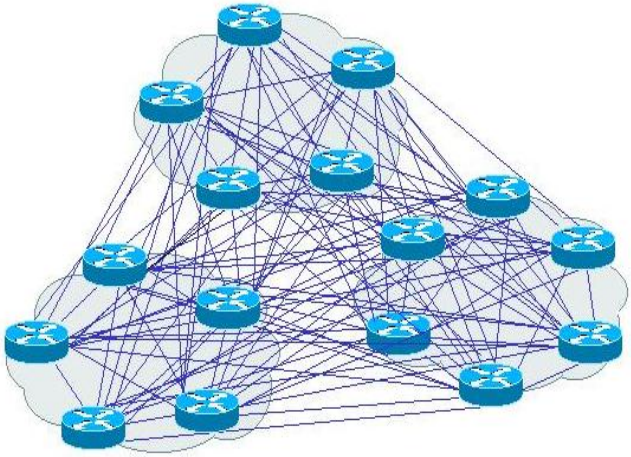

Figure 4. Example of full-meshed 16-IBGP routers

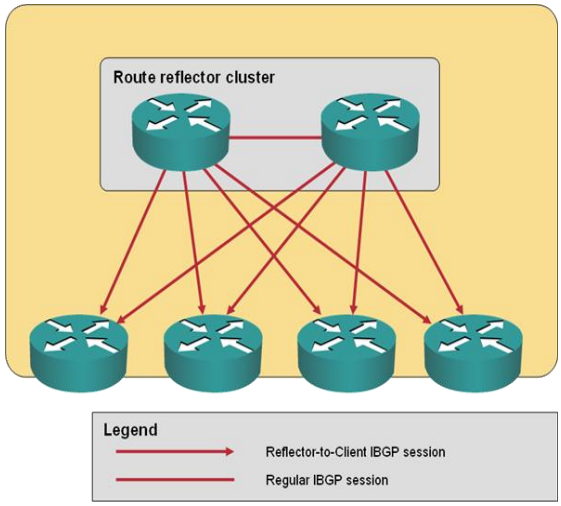

Figure 5. Example of Route reflectors cluster

There are resource constraints when you scale a network to many routers, globally, if we have: n BGP speakers within an AS, that requires to maintain: *[n*(n-1)/2]* BGP session per router. Another alternative in alleviating the need for a "full-mesh" is to use of "Route Reflectors" the *"Fig. 5"* above . They provide a method to reduce IBGP mesh by creating a concentration router to act as a focal point for IBGP sessions. The concentration router is called a Route Reflector Server. Routers called Route Reflector Clients have to peer with the RR Server to exchange routing information between themselves. The Route Reflector Server "reflects" the routes to its clients.

It is possible to arrange a hierarchical structure of these Servers and Clients and group them into what is known as clusters. Below is a diagram that illustrates this concept.

VIII. Route-Reflectors Impact On The Convergence

If we estimate the typical total number of customer's vpn routes transported inside an operator backbone to be something like 800 000 routes, each Route reflector have to learn, process the BGP decision algorithm to choose best routes, readvertise best ones, while maintaining peering relationships with all its client routers, the route-reflector CPU and memory get certainly consumed, and as a consequence, slows down route propagation and global convergence time.

A. *Test methodology*

The methodology we use to track this issue is to preload the route reflector by using a simulator acting as client routers (or PE routers), and then, nearly simultaneously, we clear all sessions on the route-reflector, then start the simulated sessions. Then we monitor convergence by issuing 'sh ip bgp vpnv4 all sum' commands while recording every 5 seconds all watched parameters (memory and CPU utilization for various processes).

When all queues are empty and table versions are synchronized, we consider the router has converged, (finished updating all its clients by all routes it knows). All these tests are performed several times to ensure they are reproducible. Results could slightly differ but accuracy is kept within ± 5%.

The goal is to find a tolerated convergence time for route reflectors, then we must limit the number of peering and number of routes per peering to respect the fixed threshold.

IX. BGP Constrained Route Distribution

A. *Feature description*

By default within a given iBGP mesh, route-reflectors will advertise all vpn routes they have to their clients (PE routers), then PE routers use Route Target (RT) extended communities to control the distribution of routes into their own VRFs (vpn routing and forwarding instances).

However PE routers need only hold routes marked with Route Targets pertaining to VRFs that have local CE attachments.

To achieve this, there must be an ability to propagate route target membership information between iBGP meshes and the most simple way is to use bgp update messages, so that Route Target membership NLRI is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes. The [AFI, SAFI] value pair used to identify this NLRI is (AFI=1, SAFI=132).

As soon as route-reflectors Receive Route Target membership information they can use it to restrict advertisement of VPN NLRI to peers that have advertised their respective Route Targets.

B. *Main findings of our study*

When we use Route-Target-constraints, The PEs receive considerably less routes. But, because in an operator backbone VRFs are spread everywhere geographically, they touch almost all route-reflectors, therefore:

- Route-Target-constraints does not help reducing the number of routes handled by route reflectors.

The only gain is that, instead of each RR sending its entire table, it's going to prefilter it before it send it to each of its PEs, which means less data to send, and less data to send, means being able to send faster, provided that there is no cpu cost due to pre-filtering on the route-reflectors side.

X. BGP Fast Convergence Mechanisms

A. *BGP Next Hop tracking*

By default within a given iBGP mesh, route-reflectors will advertise all vpn routes they have to their clients (PE routers), then PE routers use Route Target (RT) extended communities to control the distribution of routes into their own VRFs (vpn routing and forwarding instances).

XI. BGP Prefix Independent Convergence (PIC)

It provides the ability to converge BGP routes within sub-seconds instead of multiple seconds. The Forwarding Information Base (FIB) is updated independently of a prefix to converge multiple numbers of BGP routes with the occurrence of a single failure. This convergence is applicable to both core and edge failures and with or without MPLS.

A. *Setup description*

Let us consider the test setup in "Fig. 6". The simulator is injecting M and N vpn routes respectively from PE2 and PE3, PE2 end PE3 advertise injected routes respectively to route-reflector RR1 and RR2, PE1 imports the M and N VPN routes, each vpn prefixes uses as bgp next-hop either the IGP loopback of PE2 or PE3. The simulator attached to PE1 generates traffic toward those learned routes, we locate the best path chosen by PE1 in the it's forwarding table, then we cut the corresponding interface. Numbers M and N are increased progressively (by hundreds of thousands prefixes to make the impact more visible).

First phase: interface 0 fails down. It is detected and all FIB entries with this interface are deleted.

Second phase: IGP convergence occurs and new output interface is set to interface 1 for all VPN prefixes, hence a traffic disruption.

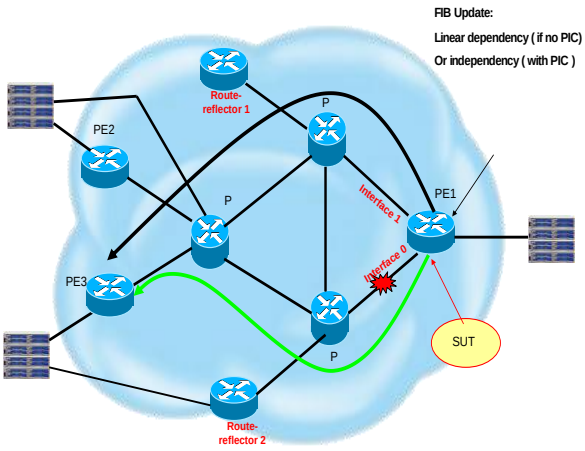

Figure 6. Lab setup diagram

B. *Feature description*

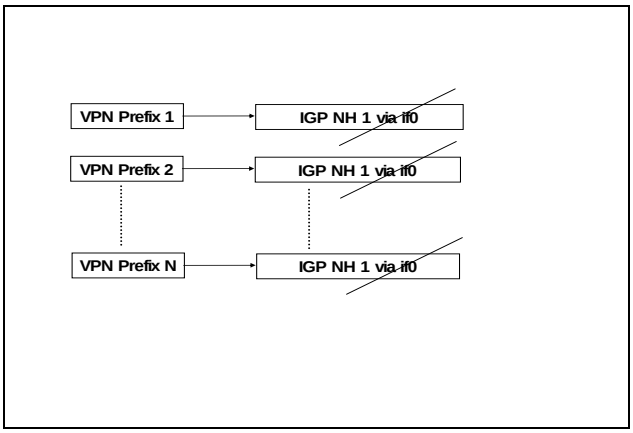

Figure 7. Forwarding table, rewriting of indexation toward interface 0

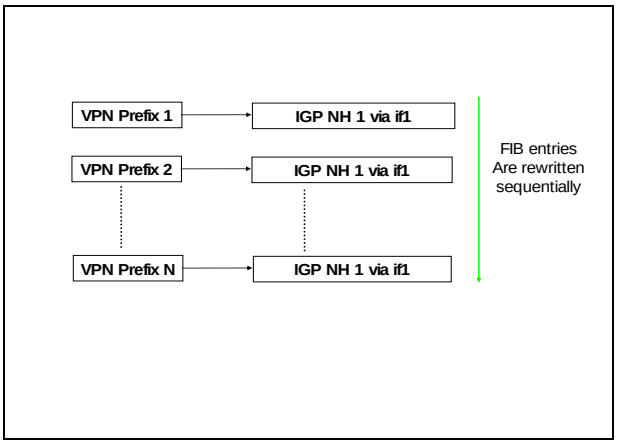

Figure 8. Forwarding table, rewriting of indexation toward interface 1

Third phase: all VPN prefixes attached to the NH1 are rewritten in the FIB with the new interface if1.

LoC = (IGP convergence) + (N * FIB Rewriting time)

Let us now analyze the behavior (with PIC feature): An intermediate Next-hop (called loadinfo) is created, and the content of the forwarding table modified as described below:

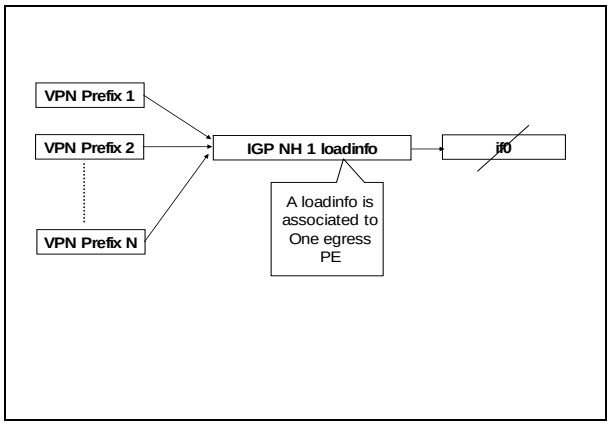

Figure 9. Forwarding table, structure modified when using the feature

First phase: if0 fails down. It is immediately erased but the loadinfo structure is not:

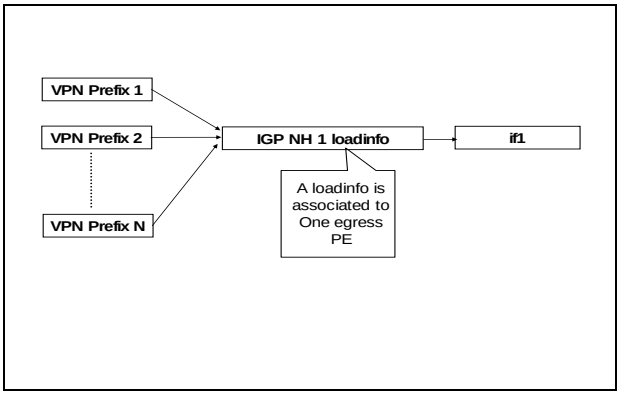

Figure 10. Forwarding table, deletion and rewriting concerns only one Next-hop

Second Phase: IGP convergence occurs and as soon as the new path via if1 is deduced, loadinfo is updated.

LoC = IGP convergence "only"

XII. Loop Free Alternate (LFA)/IPFRR

A. *Feature description*

This feature describes such a mechanism that allows a router whose local link has failed to forward traffic to a pre-computed alternate path. The alternate path stays used until the router installs the new primary next-hops based upon the changed network topology.

When a local link fails, a router currently must signal the event to its neighbors via the IGP, recompute a new primary next-hop for all affected prefixes, and only then install those new primary next-hops into the forwarding plane. Until the new primary next-hops are installed, traffic directed towards the affected prefixes is discarded. This process can take hundreds of milliseconds. The goal of IP Fast Reroute (IPFRR) is to reduce failure reaction time to 10s of milliseconds by using a pre-computed alternate next-hop in the event that the currently selected primary next-hop fails, so that, the alternate can be rapidly used when the failure is detected. A network with this feature experiences less traffic loss and less micro-looping of packets than a network without IPFRR. There are cases where traffic loss is still a possibility since IPFRR coverage varies, but in the worst possible situation a network with IPFRR is equivalent with respect to traffic convergence to a network without IPFRR. [2].

B. *Configuring the feature*

A loop-free path is one that does not forward traffic back through the router to reach a given destination. That is, a neighbor whose shortest path to the destination traverses the router is not used as a backup route to that destination. To determine loop-free alternate paths for IS-IS routes, a shortest-path-first (SPF) calculation is run on each one-hop neighbor.

```
M1-RE1> show configuration protocols isis
traceoptions {
    file ISIS_DEB1;
    flag lsp;
}
lsp-lifetime 65535;
overload;
level 2 {
    authentication-key "$9$2P4JDjHm5z3UD69CA0O"; ## SECRET-DATA
    authentication-type simple;
    no-hello-authentication;
    no-psnp-authentication;
    wide-metrics-only;
}
interface xe-0/2/0.0 {
    point-to-point;
    link-protection;
    level 2 metric 100;
}
interface xe-0/3/0.0 {
    point-to-point;
    link-protection;
    level 2 metric 10;
```

As a consequence the backup path through Rnet-A71 is precomputed and installed on the the forwarding table

```
M1-RE1>show route forwarding-table table CUST-VRF-AGILENT_PE_10 destination 1.0.0.1/32 extensive
Routing table: CUST-VRF-AGILENT_PE_10.inet [Index 5]
Internet:

Destination:  1.0.0.1/32
  Route type: user
```

```
Route reference: 0                    Route interface-index: 0
Flags: sent to PFE
Nexthop:
Next-hop type: composite              Index: 7094     Reference: 2
Next-hop type: indirect               Index: 1048581  Reference: 50001
Next-hop type: unilist                Index: 1050156  Reference: 2
Nexthop: 10.0.79.57
Next-hop type: Push 129419            Index: 502443   Reference: 1
Next-hop interface: xe-0/3/0.0        Weight: 0x1
Nexthop: 10.0.79.53
Next-hop type: Push 127258            Index: 7093     Reference: 1
Next-hop interface: xe-0/2/0.0        Weight: 0x4000 <- alternate path
```

See "Fig. 1" for lab setup

C. TEST CONDITIONS

From the lab setup described above, we announce 500000 routes, by 50k routes per vrf (vpn routing instances) from 10 different PE. The M1 receives the 50k routes in 10 different routing-instances, by 50K for each.

From the Simulator (an Agilent chassis) connected to the M1 we generate traffic consisting of 500K packets sized to 64 bytes:

- This flow use as a source an ip address varying randomly within the interval [10.0.9x.1/32 to 10.0.9x.254/32] while x=1 for vrf 1, 2 for vrf 2 etc until N for vrf N.
- This flow use as a destination an address varying sequentially within the interval [x.0.0.1/32 to x.0.195.80/32] while x=1 for vrf 1, 2 for vrf 2 etc until N for vrf N.
- We Chose isis metrics on the setup to make the Rnet-A72 the best IGP link, we shut this best link and observe the behavior of traffic curve as received on the Simulator connected to PE12

Shutdown of best link (bleu curve) , we see little negligible Impact on outgoing traffic from the MX

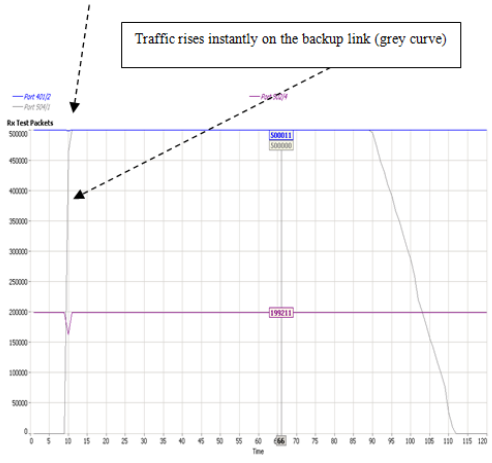

Figure 11. curve of vpn traffic with LFA

The Bleu curve is the forwarded traffic on the nominal link, the grey curve is the forwarded traffic on the backup link. The backup link have been mirrored to a free port and connected to the simulator to see the apparition and the disappearing of traffic on it.

As a comparison you can look at the traffic curve without the feature, it resembles to the diagram on the "Fig. 12". You can notice the duration of "Next-hops" rewriting of vpn prefixes toward the backup link in the forwarding table.

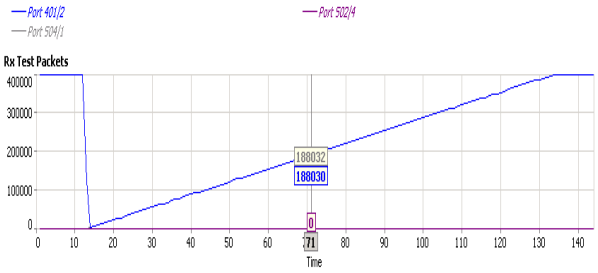

Figure 12. curve of vpn traffic without LFA

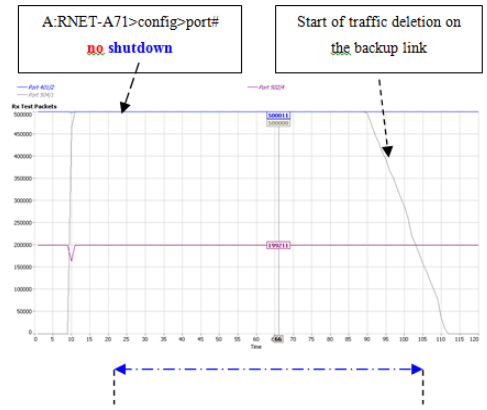

Figure 13. curve of vpn traffic with LFA, traffic retrieving on the nominal link

On the other hand, when we "de-shut" the best link, as in "Fig.13" we see that the traffic stays on the non-best link for more than 80 seconds, before going back to the best.

XIII. LDPOVERRSVP

The ldp over rsvp principle can be illustrated like in the "Fig. 14". Only core routers P1,P2 and P3 are enabling RSVP TE, ldp however they are configured to prefer rsvp tunnels to ldp one's.

The edge routers PE1 end PE2 are enabling only LDP with P1 and P3.

PE1 end PE2 are VPN and use MP-iBGP to signal vpn labels.

A. CONTROL PLAN ESTABLISHMENT

Let us consider PE2_FEC representing prefixes coming from CE2.

1. Establish RSVP tunnel-1-3 from P1 to P3, the label distributed to P2 from P3 is LR2, and the label distributed from P2 to P1 is LR1

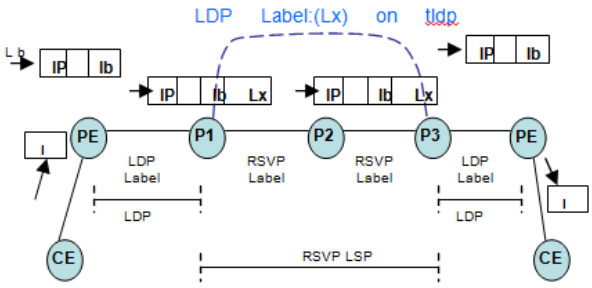

Figure 14. LDP over RSVP principle

2. Establish a targeted ldp session between P1 and P3
3. Enable IGP shortcut on P1, the egress path for PE2_FEC will be the tunnel-1-3.
4. PE2_FEC triggers the establishment of LSP on PE2, and the label mapping message will be sent to P3, let us consider this label is L2.
5. After P3 receives the label mapping message, it forwards that message to P1 through the targeted LDP session, let us consider this label is Lx
6. P1 receives the label mapping message, and finds out that the egress fo the route is tunnel-1-3.Then the LSP from PE1 to PE2 is transmitted I RSVP TE. The external label is LR1.
7. P1 continues to send Label mapping message to PE1, the label is L1.
8. PE1 generates Ingress
9. MP-BGP sends private network route of CE2 from PE2 to PE1, the label of private network is Lb.

At this stage the establishment of LSP between PE1 and PE2 is complete. This LSP traverses the RSVP TE area

(P1 ~~ P3).

B. *Forwarding plane process*

The forwarding process of packets is as follows:

We describe here the forwarding process of data from CE1 to CE2, if needed do the symmetrical reasoning regarding flows from CE2 to CE1:

1. After PE1 receives packets from CE1, it tags the BGP label Lb of private network and then it tags LDP label L1 of the provider network
2. (Lb,L1) label of PE1 is received on P1, replace L1 with Lx (the label sent to P1 through the targeted ldp session, and then tag tunnel label LR1 of RSVP TE, the label of packet becomes (Lb,Lx,Lr1).
3. From P2 to P3, with the RSVP TE transparently transmitting packets, the LR1 is replaced by LR2, that is, the packets received by P3 are tagged with the following labels (Lb,Lx,LR2)
4. Upon arriving P3, the LR2 is first stripped and then comes out Lx, and the label of LDP which is replaced by L2.The packet is then sent to PE2 and the label becomes (Lb,L2)
5. After the packet reaches PE2, L2 is first stripped and then the Lb. After that, the packet is sent to CE2

C. *LSP protection , one to one backup method*

Each P creates a detour (tunnel) for each LSP, the detour will play the role of a protecting LSP :

If the router P2 fails, P1 switches received traffic from PE1, along the detour tunnel [P1,P5] using the label received when P1 created the detour .

The detour is calculated based on the shortest IGP path from P1 to the router terminating the protected LSP, let us say: PE2. In this case the protecting LSP will avoid the failed router P2 (node protection).

At no point does the depth of the label stack increases as a consequence of taking the detour.

While P1 is using the detour, traffic will take the path [PE1-P1-P5-P6-P7-PE2]

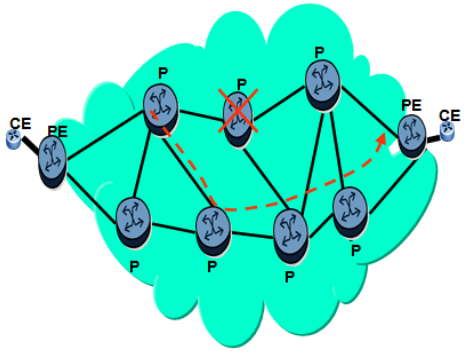

Figure 15. LDP over RSVP backup method

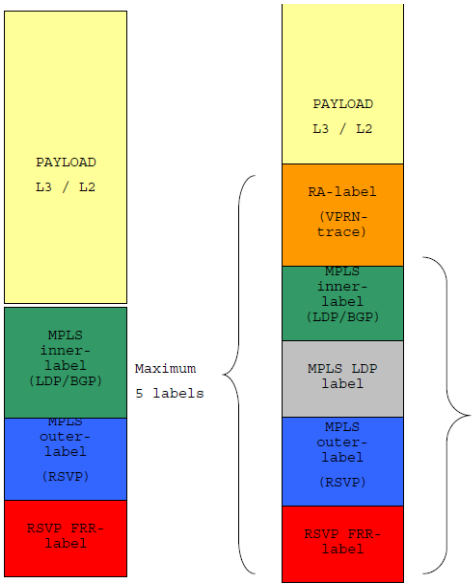

Figure 16. LDPoRSVP labels stack during FRR

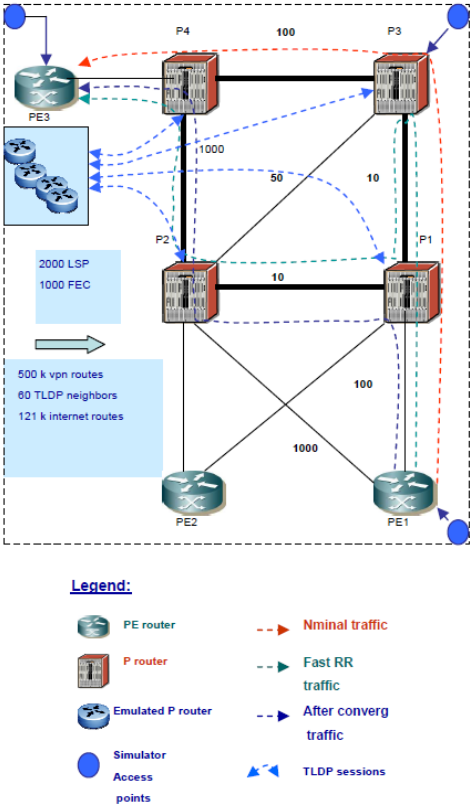

Figure 17. LDPoRSVP Lab setup

D. LDPoRSVP LABEL STACK DURING FRR

Nota: when deploying LDPoRSVP and enabling FRR (facility) as protection mechanism keep the 4 potential MPLS labels into account for MTU definition

E. LAB SETUP AND TESTS SCOPE

Here are described the implementations made in our lab, the CSPF (constrained shortest path first) was simplified to only shortest igp:

- Inter-P trafic will be encapsulated in a tunnel.
- No impact on all PE configuration, Only P routers are concerned by (LDPoRSVP).
- The tunnel is a TLDP session, between each P, so full mesh of: [n x P] routers.
- Each TLDP session is using an LSP which is dynamic.
- Signalling protocol for LSP is RSVP-TE , using cspf.
- CSPF is a modified version of SPF algo(Dijkstra) , used in ISIS.
- CSPF algorithme finds a path which satisfy constraints for the LSP (we simplify to only one constraint: the igp shortest path).
- Once a path is found by CSPF, RSVP uses the path to request the LSP establishment.

F. LAB TEST METHOD :

On each P router, we check that a (detour LSP is precalculated, presignaled for each LSP). We load heavily the P routers with:

- BGP vpn routes , internet routes
- IGP (ISIS) routes
- LDP labels
- TLDP sessions
- RSVP sessions

We generate traffic consisting of hundred thousands of packets in both directions, PE1 to PE3 see (Fig.2), note that

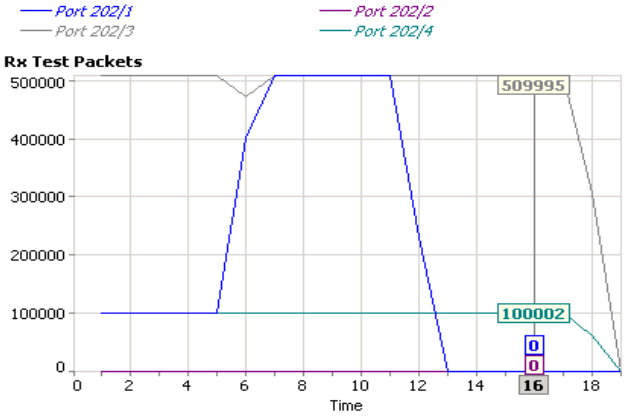

Figure 18. Received packets curve

The grey curve represents recived packets, we notice a small traffic fall.

TABLE 1 .LDPoRSVP Traffic measurement

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)
All Ports	10735999	14376986	702239958	942649764	295.680	396.905	n/a	129.61
202/4->202/3, StreamGroup 10	1760000	1752236	123200000	122656520	51.874	51.645	7764	151.55
202/4->202/3, StreamGroup 9	1760000	1752192	123200000	122653440	51.874	51.644	7808	151.57
202/4->202/3, StreamGroup 7'	1760000	1752192	123200000	122653440	51.874	51.644	7808	151.67
202/4->202/3, StreamGroup 7	1760000	1752182	123200000	122652740	51.874	51.643	7818	151.61
202/4->202/3, StreamGroup 7"	1760000	1752110	123200000	122647700	51.874	51.641	7890	151.64
202/1	0	3681475	0	243226522	0.000	102.411	n/a	71.15
202/2	0	0	0	0	0.000	0.000	n/a	
202/3	1759999	8936135	73919958	625529450	31.124	263.381	n/a	151.60
202/4	8976000	1759376	628320000	73893792	264.556	31.113	n/a	140.29

In "Fig. 18" the chosen igp metrics will force then nominal path to be :[PE1-P1-P3-P4-PE3] (the red path). We cut the link [P4–P3] : either by shuting the physical port or by removing the fiber from the port, we measure the convergence time through the number of lost packets related to the ratio: *(sent /received) packets per second.*

We check that, when the link [P4–P3] goes down, the P3 router, instead of waiting the igp convergence, instantly uses the precomputed backup link [P3-P1-P2-P4] (the green or detour path), then after the igp converges, the traffic goe, without impact, through the link [PE1-P1-P2-P4-PE3] (the blue path).

We check fast reroute performance at different load conditions: firstly we start with few LSPs then we increase the number progressively: (500, 1000, 2000 …)

G. TEST RESULTS:

We see that mainly: convergence time stays between 20 msec < t < 100 msec independently of number of LSPs. We notice some issues regarding scalability of LDP FECs. The "on purpose" studied case in the Fig.4 shows that during the fast-reroute phase, traffic goes back to the sender before taking the good (remaining) path. This topology case would exist in a backbone design, so the sizing of the link must take into account the potential and transcient traffic load.

XIV. LDP FASTREROUTE

It's a mechanism that provides a local protection for an LDP FEC by pre-computing and downloading to the "forwarding plane hardware": both a primary and a backup NHLFE (Next Hop Label Forwarding Entry) for this FEC.

The primary NHLFE corresponds to the label of the FEC received from the primary next-hop as per standard LDP resolution of the FEC prefix in RTM (routing table manager). The backup NHLFE corresponds to the label received for the same FEC from a Loop-Free Alternate (LFA) next-hop.

- LFA next-hop pre-computation by IGP is described in [2].
- LDP FRR relies on using the label-FEC binding received from the LFA next-hop to forward traffic for a given prefix as soon as the primary next-hop is not available.

In case of failure, forwarding of LDP packets to a destination prefix/FEC is resumed without waiting for the routing convergence.

The RTM module (routing table manager) populates both primary and backup route and the "forwarding hardware" should populate both primary and backup NHLFE for the FEC.

A. ROUTES AND LFA COMPUTATION REMINDER

Assuming : a,b,c,d,e,f,g represent the igp metrics on each node link:

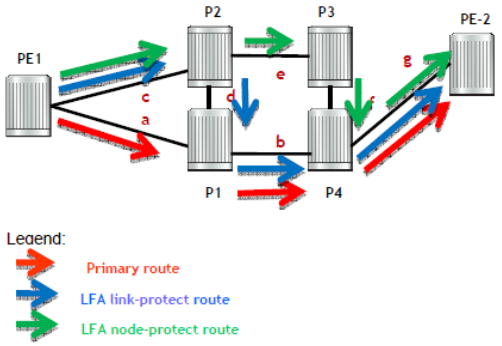

Figure 19. LFA concept reminder

The primary route will be via P1, assumed that:

$$a < (c + d) \text{ and } (a + b) < (c + e + f)$$

The LFA route via P2 and P1 protects against failure of link PE1-P1:

- Loop Free Criterion (computed by PE1): The cost for P2 to reach P4 via P1 must be lower than the cost via routes PE1 then P1, assumed that: $d < (a + c)$
- Downstream Path Criterion (to avoid micro-loops): The cost of reaching P4 from P2 must be lower than the cost for reaching P4 from PE1, assumed that: $d < a$

The LFA route via P2 and P3 protects against the failure of P1, node-protect condition for P2, assumed that:

$$(e + f) < (d + b)$$

B. THE SPF ALGORITHME BEHAVIOR

1. Attempt the computation of a node-protect LFA next-hop for a given prefix
2. If not possible, attempt the computation of a link-protect LFA next-hop.
3. If multiple LFA next-hops for a given primary next-hop are found, pick the node-protect in favor of the link-protect.
4. If there is more than one LFA next-hop within the selected type, pick one based on the least cost.
5. If more than one have the same cost, the one with the least (outgoing interface: OIF) index is selected.

Both the computed primary next-hop and LFA next-hop for a given prefix are programmed into the routing table management.

C. LDP FASTREOUTE: LAB SETUP AND TEST METHOD:

The work have being done on the setup of "Fig. 22" and results are reported on tables: 2, 3.

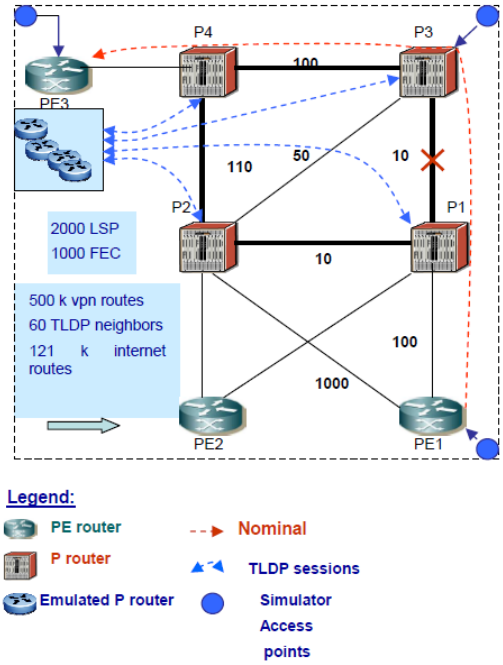

Figure 22. LDP Fastreroute Lab setup

```
P1# show router isis routes alternative 10.0.222.5/32  Route Table
Prefix[Flags]                    Metric          Lvl/Typ    Ver.
  NextHop                        MT              AdminTag
Alt-Nexthop                      Alt-Metric      Alt-Type
 ------------------------------------------------------------------------------
10.0.222.5/32                    11130           2/Int.     4950  P3
  10.0.79.21                     0               0
  10.0.70.49 (LFA)               11140           nodeProtection
 ------------------------------------------------------------------------------
No. of Routes: 1
Flags: LFA = Loop-Free Alternate nexthop
```

Table 2. Example of LFA precomputation

```
P1# show router isis lfa-coverage
===============================================
LFA Coverage
===============================================
Topology       Level  Node          IPv4              IPv6
-------------------------------------------------
IPV4 Unicast   L1     0/0(0%)       3257/3260(99%)    0/0(0%)
IPV4 Unicast   L2     27/28(96%)    3257/3260(99%)    0/0(0%)
-----------------------------------------------
```

Table 3. LFA Lab coverage pourcentage

D. *LAB TEST METHOD:*

Same as described before (2.6.1) except that, here we cut the inter P link [P1-P3], the backup path is [P1-P2-P4]. we measure the convergence time through the number of lost packets related to the ratio: *(sent /received) packets per second*.

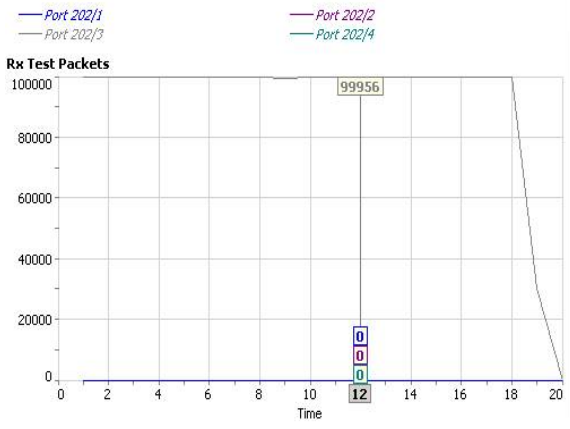

Figure 23. received traffic curve

A. *LDP FAST-REROUTE TEST RESULTS:*

We see that mainly: the convergence time stays around 5 ms. This makes the LDP fast-reroute more attractif, however it doesn't offer a 100% topology coverage.

Table 4 .LDP-FRR Traffic measurement

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)
All Ports	3660000	1829434	204960000	128060380	81.984	51.224	n/a	144
202/1	0	0	0	0	0.000	0.000	n/a	
202/2	0	0	0	0	0.000	0.000	n/a	
202/3	1830000	1829434	76860000	128060380	30.744	51.224	n/a	144
202/4	1830000	0	128100000	0	51.240	0.000	n/a	
202/4->202/3, StreamGroup 9	1830000	1829434	128100000	128060380	51.24	51.224	566	144

XV. RSVP-TE AND LDP-FRR COMPARAISON OUTCOMES

RSVP-TE gains:

- Fast convergence « P » (detour LSP is precalculated, presignaled for each LSP)
- A convergence time around: 20 msec < t < 100 msec
- RSVP-TE drawbacks:
- additional level of routing complexity; requires P-P trunk support rsvp, TLDP sessions, additional cpu load (rsvp msg)

LDP(/IP) FRR gains:

- local decision, no interop issues with other vendors
- very simple configuration (just turn it on)
- better scaling compared to full-mesh RSVP model
- less overhead compared to RSVP soft-refresh states

LDP(/IP) FRR drawbacks:

lower backup coverage: depending on topologies may vary between: 65 to 85%, indeed, the source routing paradigm: LDP will always follows IP route, so if a candidate backup router has its best route through originating node, this candidate node cannot be chosen as backup.

While the conceptual restriction of LDP(/IP) FRR is efficient against loops, it doesn't allow a 100% coverage of all topologies, however we can reach a good compromise by a mixture of both, RSVP shortcuts will be deployed if and where LDP(/IP) FRR cannot offer coverage.

XVI. IGP MICRO-LOOPS

In standard IP networks, except when using source routing, each router takes its own routing decision (hop by hop routing). When the topology changes, during the convergence time, each router independently computes best route to each destination.

Because of this independence, some routers may converge quickly than others, the difference in convergence time may create temporary traffic loops, that's what we call "microloops".

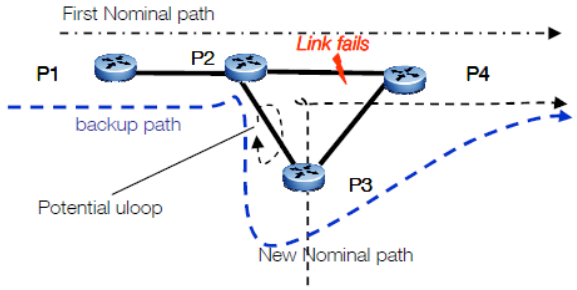

Figure 24. Microloop birth

Micro-loops can be triggered by any topology change that causes the network to converge like: link down, link up,metric change ..etc.

Given the "Fig. 24" above, when the link P2-P4 fails :

P2 detects failure and converges path to P3, as P3 is using P2 as its nominal path, if P2 has converged but P3 didn't yet, there is a creation of a micro-loop between both nodes, until P3 convergence is achieved.

A. *MICRO-LOOPS LOCALIZATION*

When a topology change occurs between 2 nodes A & B, and given the IGP metric as in the the figure 25, a microloop can occur :

- Between A and his neighbors (local loop)
- Between B and his neighbors (local loop)
- A router upstream of A and one of his neighbors (remote loop)
- A router upstream of B and one of his neighbors (remote loop)

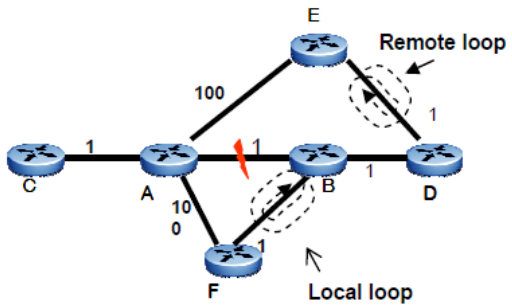

Figure 25. Microloops dispersion

B. *CONSEQUENCES OF MICRO-LOOPS*

1) BANDWIDTH CONSUMPTION ESTIMATION:

Given the illustration below:

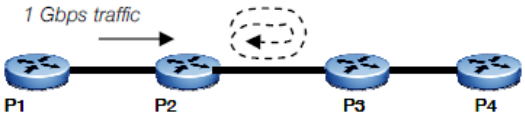

Figure 26. Microloop and bandwidth

Figure 26. Microloop and bandwidth

Given 1 gigabit traffic coming from P1, as soon as this traffic enters in the loop, each second , 1Gb additional data is introduced in the loop.

Time	P1-P2 link	P2-P3 link
0sec	1Gb	1Gb
1sec	1Gb	2Gb
2sec	1Gb	3Gb
3sec	1Gb	4Gb

Looping traffic will consume bandwidth on the affected link(s) until:

- The link comes congestion
- TTL of looping packet starts to expire
- The network has converged

The bandwidth consumption will depend on a lot of parameters:

- Amount of traffic injected per second in the loop
- Packet size
- TTL of packets
- RTD (round-trip delay time) of links
- Packet switching time

To illustrate this, have a link with an RTD of 20 ms, a monohop loop occurring on this link and a packet with "an initial TTL of 255" entering in the loop.

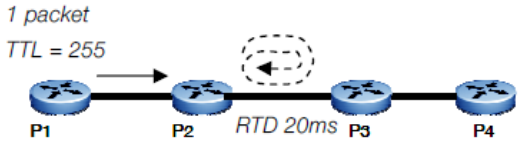

Figure 27. bandwidth consumption

Each time the packet crosses P2 and P3, the TTL is decreased by one, we consider that this packet will do 127 round trip over the loop, so it will take 2540ms for the packet to expire.

The bandwidth consumption depends also on the packet size, consider 1 Gbps of traffic injected in the loop with a packet size of 500 bytes , it means that each second, 250k packets are injected in the loop.

Time	P2-P3 link
0sec	250k packet
1 sec	500k packet
2 sec	750k packet
2,5 sec	750k packet **+ 125k packet (injected) – 250k packet (expired) = 625k packet**
2,7 sec	625k + 50k (new injected) – 50k (expiring)

In general we can say that:

*(BW consumed by loop)=(BW injected) *(TTL/2)*(RTD)*

Than for :

- 1Gbps, injected in a loop with 20ms RTD in loop, TTL=255, loop maximum rate is 2,5 Gbps, than max link BW usage of : 3,5 Gbps.
- 4Gbps, injected in a loop with 3ms RTD in loop, TTL = 250, loop maximum rate is 1,5 Gbps, than max link BW usage of : 5,5 Gbps.

2) OVERLOAD IN ROUTERS CPU

If large amount of mpls traffic loops between two nodes A and B; at each hop, the ttl of mpls pkt decreases by 1. When the ttl of the mpls pkt expires, this pkt is dropped by the control plane hardware (the routing engine) and not by the forwarding plane hardware.

Depending the duration of the loop, the amount of mpls ttl-expiring packets arriving to the control plane, the CPU load may increase to 100%.

Mpls ttl expired packets come to the routing-engine mixed with other important packets: igp (ISIS or OSPF) , bfd, bgp ..etc and all routing control packets, (mpls and non mpls). As a consequence: bfd, the most sensitive one, may go down firstly, and carry along all level3 protocol depending on it.

3) CAUTION ON QOS MODELS

If some quality of service models are used, and some types of packets are prioritized, have this type of packets entering in a loop, and depending on the loop duration, the amount of prioritized traffic, they may consume all the bandwidth and force control (routing) packets to be dropped. That is why, it is a wise design to put the control packets on the top priority, even above voive or other sensitive applications.

4) MICROLOOPS PROPAGATION

A level 3 loop occurring between two points A and B, and as explained in paragraph 4.2.2, may trigger a convergence again, potentially other microloops can appear far on other routers, generating cpu load. The overall network will undergo a phenomena we can define as a "loop propagation". Obviously, the cpu load will stay 100% until micro loops disappear and convergence stabilize.

XVII. MICROLOOPS LAB SETUP AND TEST METHOD

Given the Figure 28, firstly we confirmed we can produce loops by configuring different isis convergence timers to facilitate loops appearance, then is a second stage, in order to have more control, we created manual loops between P1-P3 and P1-P2.

We used a simple way to create loops: given a vpnA on a PE1 connected to P1 and a vpnB on a PE2 connected to P2:

- On PE1 vpnA have a static route to a destination [a.b.c.d/mask] with PE2 loopback as the next-hop.
- On PE2 vpnB have a static route to the same destination with PE1 loopback as the next-hop.
- PE1 and PE2 know loopback of each other through isis.

Using a traffic simulator we inject 10Millions packets having the destination [a.b.c.d/mask], and to accelerate the effect on CPU we put the TTL of all packet to values randomly equal to 2 or 3.

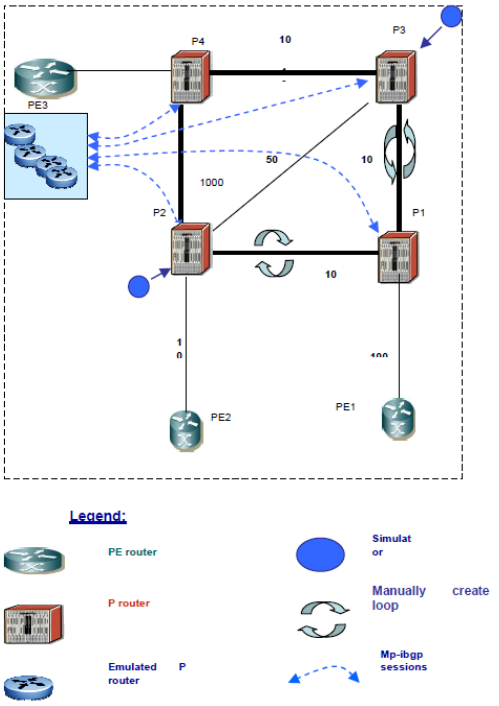

Figure 28. Microloops lab setup

A. MICROLOOPS AND TRAFFIC PROTECTION

5) MICROLOOPS AND LFA

As explained, LFA computes an alternate nexthop that is used when a local failure appears, however the alternate nexthop may not be the converged backup nexthop.

Given the case of "Fig. 29":

- H is the LFA node
- E is the converged nexthop, the backup calculated node after the link [A-B] broke down

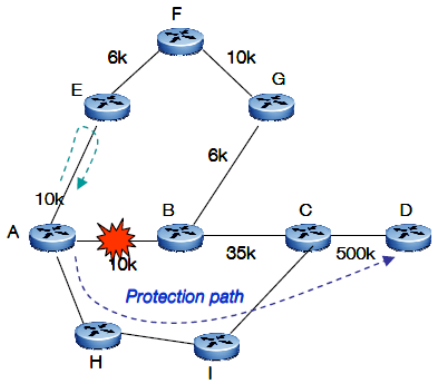

Figure 29 – Microloops and LFA

When failure occurs, local router switch traffic to LFA node, traffic is safe. When convergence is achieved on local node, traffic is switched from LFA node to backup nexthop:

- Traffic will be safe if backup node (and subsequent nodes) have converged
- Otherwise , traffic may go in microloop

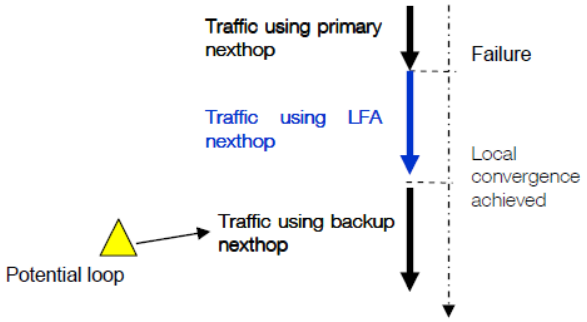

Figure 30 – potetial loop with LFA

6) MICROLOOPS AND IGP/LDP SYNCHRO

Setting high metric when IGP and LDP gets out of synchronization and getting back to nominal metric (LDP/IGP coming back in synchronization) can cause microloops (remote or local). Same effect expected as the failed link comes up, when the feature IGP/LDP synchronization in not implemented at all.

7) CPU-PROTECTION MECHANISMS

Depending on the router manufacturer, several CPU protection mechanisms may be implemented:

Ability to put a port overall rate that measures the arrival of all control packets sent to the CPU for processing, giving the possibility to selectively discard out-of-profile-rates. Ability to create per protocol queues and guarantee selective high priority for important packets. A dedicated study would assess the efficiency of one or the other protection mechanism and proof their robustness by testing under worst conditions.

XVIII. CONCLUSION

In this paper we presented the most important features wich can contribute in convergence enhancement; it is not aimed at detailing all existing features

We focused on methods that can be used to precompute backup paths on the forwarding plane, we presented features like: Prefix independent convergence and loop free alternate, test results and gains obtained in comparison to the situation with and without these features .

We presented a comparative study of RSVP-TE versus LDP (/IP) Fast reroute, it appears that: with RSVP-TE, the detour LSP is precalculated, presignaled for each LSP, the convergence time is around: 20 msec < t < 100 msec. However it has drawbacks like additional level of routing complexity, requiring That P-to-P trunks support rsvp and full mesh TLDP sessions, additional cpu load, due to rsvp messages. With LDP(/IP) FRR we have local decisions, hence no interop issues with other vendors, a simple configuration (just turn it on),a better scaling compared to full-mesh RSVP model and less overhead compared to RSVP soft-refresh states. However LDP (/IP) FRR has an important drawback: A lower backup coverage because of the source routing paradigm

Finally, we analyzed micro-loops phenomenon, bandwidth and CPU consumption; we studied their birth mechanisms and propagation, and initiated a reflexion on means to mitigate them.

Overall, it is clear that the control of the convergence in its globality is not an easy task, but our measurements and simulations indicate that with good design and choice of tuning features, we are confident a sub-second to tens of milliseconds convergence time can be met.

REFERENCES

[1] Nuova Systems, K. Kompella Juniper Networks, JP. Vasseur Cisco Systems, Inc., A. Farre Old Dog Consulting. Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE).

[2] A. Atlas, Ed BT, A. Zinin, Ed. Alcatel-Lucent. Basic Specification for IP Fast Reroute: Loop-Free Alternates (RFC 5286) September 2008

[3] E. Oki,T. Takeda NTT, A. Farrel Old Dog Consulting. Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions.April 2009.

[4] L. Andersson Nortel Networks Inc., P. Doolan Ennovate Networks, N. Feldman IBM Corp, A. Fredette PhotonEx Corp, B. Thomas Cisco Systems Inc. LDP Specification (RFC 3036). January 2001

[5] D. Awduche Movaz Networks, Inc., L. Berger D. Gan Juniper Networks, Inc. T. Li Procket Networks, Inc. V. Srinivasan Cosine Communications, Inc. G. Swallow Cisco Systems, Inc. RSVP-TE: Extensions to RSVP for LSP Tunnels (RFC 3209). December 2001.

[6] D. Awduche, J. Malcolm, J. Agogbua,M. O'Dell, J. McManus UUNET MCI Worldcom (RFC-2702) September 1999.

[7] E. Rosen, Y. Rekhter. BGP/MPLS IP Virtual Private Network (VPNs) (RFC-4364)

[8] Ina Minei, julian Lucek Juniper Networks, MPLS-Enabled Applications, Emerging Developments and New Technologies .September 2008.

[9] L. AnderssonNortel Networks Inc, P. Doolan Ennovate Networks N. Feldman IBM Corp, A. Fredette PhotonEx Corp, B. Thomas Cisco Systems, Inc. (RFC-3036). January 2001

[10] Abdelali Ala, Driss El Ouadghiri, Mohamed Essaaidi: Convergence enhancement within operator backbones for real-time applications. iiWAS 2010: 575-583.

[11] Ala, A. Inf. & Telecom Syst. Lab., Abdelmalik Essaadi Univ., Tetuan, Driss El Ouadghiri, Mohamed Essaaidi: Fast convergence mechanisms and features deployment within operator backbone infrastructures.

[12] P. Pan, Ed. Hammerhead Systems, G. Swallow, Ed. Cisco Systems, A. Atlas, Ed. Avici Systems (RFC-4090).May 2005.

[13] T. Bates, R. Chandra, D. Katz, Y. Rekhter. Multiprotocol Extensions for BGP-4 (RFC-2858).June 2000.

[14] Y. Rekhter, E. Rosen. BGP MPLS Carrying Label Information in BGP-4 (RFC 3107).May 2001.

[15] Y. Rekhter, T. Li, S. Hares, A Border Gateway Protocol 4 (BGP-4) (RFC-4271). January 2006.

[16] Alia K. Atlas (edit BT), A. Zinin, Ed. Alcatel-Lucent. IP Fast Reroute: Loop-Free Alternates (RFC 5286).September 2008

[17] P.Marques, R.Bonica from Juniper Networks, L.Fang, L.Martini, R. Raszuk, K.Patel, J.Guichard From Cisco Systems, Inc.

Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs). (RFC 4684).November 2006.

[18] Susan Hares, NextHop Technologies Scaling MPLS Software to Meet Emerging VPN Demands. January 2004.

[19] Zhuo (Frank) Xu Alcatel-Lucent SRA N0.1. Designing and Implementing IP/MPLS-Based Ethernet Layer 2 VPN Services.2010.

AUTHORS PROFILE

Ala Abdelali is a phd student at Information and Telecom Systems Lab, Faculty of Sciences Abdelmalek Essâadi university, Tetuan Morocco.He obtained his first engineer degree since September 1989 in Belgium, then his "D.E.A" from the university of Paris XI since September 1992. Then he worked ten years as support and telecom network designer in several IT companies and telecom operators. His research area is: architecture, core IP/MPLS/VPN design and network engineering.

Driss El Ouadghiri is a research and an associate professor at Science Faculty, Moulay Ismail University, Meknes, Morocco, since September 1994. He was born in Ouarzazate, Morocco. He got his "License" in applied mathematics and his "Doctorat de Spécialité de Troisième Cycle" in computer networks, respectively, in 1992 and 1997 from Mohamed V University, Rabat, Morocco. In 2000 he got his PhD in performance evaluation in wide area networks from Moulay Ismail University, Meknes, Morocco. He is a founding member, in 2007, of a research group e-NGN (e-Next Generation Networks) for Africa and Middle East. His research interests focus on performance evaluation in networks(modelling and simulation), DiffServ architecture (mechanisms based active queue management) and IPv6 networks. He spent at INRIA Sophia-Antipolis, in the MISTRAL team, two long trips to scientific research in 1995 and 1996. Also, he had a post-Doctoral research at INRIA-IRISA of Rennes, in the ARMOR team, for a year from October 2000 to October 2001.

Prof Mohamed Essaaidi is Currently director of ENSIAS. He is IEEE Senior Member, he received the "Licence de Physique" degree, the "Doctorat de Troisième Cycle" degree and the "Doctorat d'Etat" degree in Electrical Engineering and with honors, respectively, in 1988, 1992 and 1997 from Abdelmalek Essaadi University in Tetuan, Morocco. He is a professor of Electrical Engineering in Abdelmalek Essaadi University since 1993. He is the founder and the current Chair of the IEEE Morocco Section since November 2004. Prof. Essaaidi holds four patents on antennas for very high data rate UWB and multiband wireless communication networks (OMPIC 2006, 2007, 2008). He has also co-organized several competitions aiming at fostering research, development innovation in Morocco and in the Arab World (Moroccan Engineers Week 2006, 2007 and "Made in Morocco" and ASTF "Made in Arabia" Competitions in 2007 and 2009).