Technical Paper

Demystifying and Evaluating Router Resiliency

Do you need a redundant router at every POP/CO?

Applying Agilent's *Journal of Internet Test Methodologies* to the verification and measurement of router resiliency

Introduction

Responding to carrier demands to improve IP network reliability, several router vendors recently announced new technologies - such as MPLS Fast Reroute, Graceful Restart, Hitless Upgrade and Non Stop Routing. In this paper, we demystify resiliency technologies and show how to evaluate performance claims with some practical test scenarios.

Agilent Technologies

Router Resiliency is Number One

Recent surveys by Infonetics Research[1] and BTexact Technologies[2] of major global and North American carriers revealed that router reliability and stability constitute the number one selection criteria of service providers today. Similarly, Network Magazine's technology scorecard[3] highlighted "fault tolerant routing technology" as being equal-first in importance to public network architects. These surveys confirmed the findings of a Yankee Group report[4] that scored Software Reliability of existing router solutions with a lowly "D+". Service providers are more worried than ever about the resiliency of their IP networks!

BTexact's report suggests that carriers need the same level of reliability in their IP networks that they enjoy in their telephony networks. Operators are asking their router vendors for three-minute reboot times, hitless upgrades (in-service router software updates without packet loss), redundant hardware with hitless failover (automatic switchover upon failure), and no more than two hours of downtime in 40 years.

Router resiliency is becoming more important to providers for three reasons: the potential for cost savings, the ability to offer new revenue-generating services, and the increasing costs of network outages.

Cost Savings and Differentiated Services

Telephony switches with 99.999% availability have been the norm for many years. However, to offer IP services with a sufficient level of reliability, service providers have had to increase network resiliency by deploying redundant routers at every Point of Presence (PoP) or Central Office (CO).

With the convergence of voice and data over an IP/MPLS core, high levels of availability are now a necessity in carrier core IP networks. High availability is also needed to enable service providers to offer enterprises VPN, VLAN and VoIP services and to support mission-critical intranet and Internet eBusiness applications.

Router manufacturers and protocol software vendors recently announced new technologies to improve the availability of IP networks. Many of these technologies claim to significantly reduce packet loss or route reconvergence time during failover, while others offer "hitless" in-service software updates. Moreover, a few vendors contend that their routers can provide 99.999% carrier-class reliability. If this is true, then an estimated 30-40% cost savings could be achieved by eliminating redundant routing systems. This would free valuable space in POPs and COs and could allow providers to offer enhanced and differentiated services at lower cost.

The True Cost of IP Network Outages

Many of the costs to service providers of outages and downtime are the tangible, easily measurable costs:

- Direct costs of service restoration and equipment upgrades
- Loss of revenue during outages
- Penalties associated with customer SLAs

However, the "real" costs include some losses that are harder to quantify but may be far greater:

- Lost revenue from dissatisfied customers moving to competitors or taking new business to competitors
- The cost of a tarnished image: the lessened ability to credibly market future "premium" differentiated services and position them against competitors

It is estimated[5] that U.S. companies lost $100 billion due to network outages in 1999.

The Cause of Network Failures

Outages are caused by a variety of events — see the table in a Merit Network's study[6] for a revealing breakdown of network failure causes. In 1998, following a fiber cut and subsequent repair, a route update storm caused a second outage to a large U.S. carrier network[7]. During the flood of network update messages, the network's routers were too busy relearning routes to forward traffic.

Demystifying Router Resiliency

A number of different technologies have been developed to enable protection of links, interfaces and nodes within an IP network. These technologies provide mechanisms at different protocol layers to improve network availability. In this section, we compare and contrast these technologies to each other and to existing technologies, and we explain how each can contribute to providing a highly resilient, carrier-class IP network.

Layer 1 and 2 Link Protection

SONET/SDH Automatic Protection Switching (APS) is a layer-1 mechanism for protection of links against fiber cuts and excessive bit error rates. Originally designed for voice services, APS offers 50-millisecond restoration using diverse network paths. For 100% restoration, "1+1" protection is employed; traffic is simultaneously sent on working and standby paths, and the receiver selects the path to use. An alternative scheme that offers partial restoration but uses fewer resources (and is therefore cheaper) is called "1:N" protection. A single standby path, carrying only an idle signal, protects multiple working paths. Failover and restoration can be signaled using a protocol carried on the SONET/SDH K1 and K2 overhead bytes. Some carrier-class routers now integrate APS into optical line interface cards to eliminate the need for additional SONET network elements and to add protection against line card failures. There are also transport-layer mesh restoration schemes based on APS that have already been deployed.

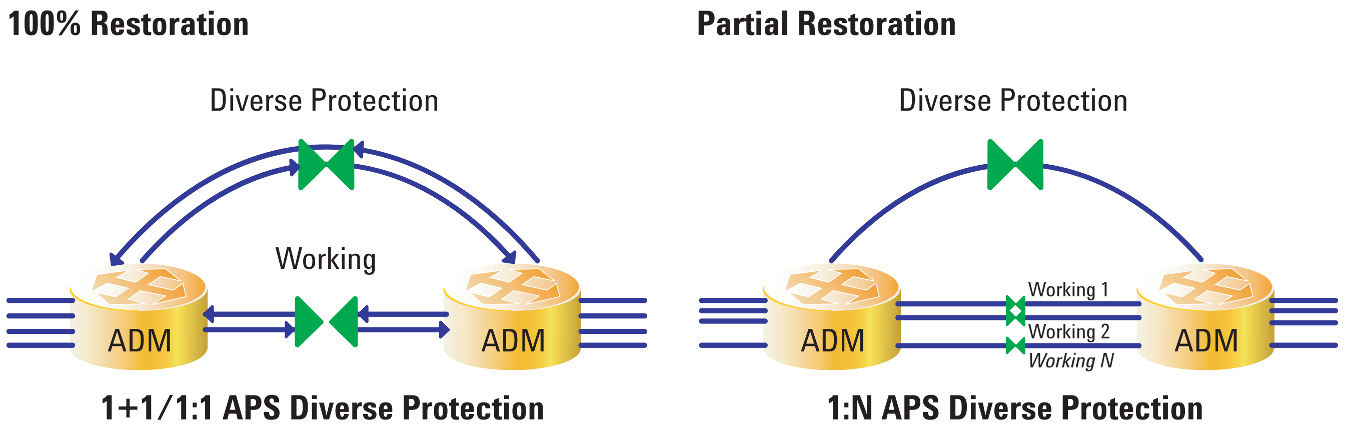
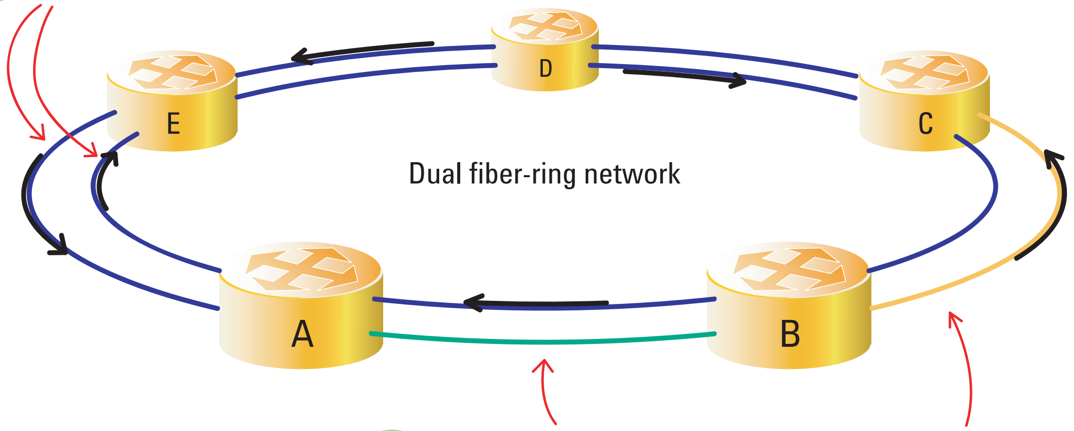


Figure 1: SONET/SDH Automatic Protection Switching

Resilient Packet Ring wastes less bandwidth because unlike SONET, RPR...

1 Carries traffic on both fiber rings at once in different directions

2 Uses the shortest path between points



3 Reuses time slots so data can go from A to B (green) and different data can go from B to C (orange) using the same RPR time slot

Figure 2: RPR Runs Rings Around SONET

Resilient Packet Ring[8] (RPR) is a relatively new technology that combines the benefits of SONET's reliability and 50ms restoration with Ethernet's efficiency and low cost. Currently being standardized in IEEE 802.17, RPR is gaining support and momentum because it is based on technology familiar to carriers and because it increases bandwidth utilization of regional and metro fiber rings. At least seven vendors already have RPR or RPR-like products, including Packet Ring interfaces for metro aggregation routers that enable carriers to optimize their existing SONET/SDH backbones for new data services.

In contrast to SONET Protection Switching and RPR, multiplexed link technologies, known by names such as "Composite Links", are layer-2 technologies that provide redundancy for point-to-point links. They can also offer recovery from router interface failures to effectively protect the "whole" link. Used in carrier-class routers, multiple POS or Ethernet circuits can be bonded together (like Multilink PPP[9] or Multilink Frame Relay[10]) into a higher-speed virtual trunk for load balancing and redundancy. Upper layer protocols view the composite link as a single logical interface, thereby providing the bandwidth utilization benefits of statistical multiplexing. Hardware modules can be added or removed while in service — similar to ATM's Inverse Multiplexing (IMA) mechanism.

APS, RPR and multiplexed link technologies protect links but not nodes (routers). In contrast, MPLS Fast Reroute (FRR) is a technology that can protect against both link and router failures by providing a label-switched path (LSP) tunnel with a pre-established backup LSP. When a failure is detected, the router immediately upstream from the failure quickly switches all traffic to the backup tunnel. At the same time, it notifies the head-end router at the start of the LSP so that a new, optimal route can be established. Thus Fast Reroute does not avert IP route reconvergence, but it does provide a temporary detour for traffic to circumvent lengthy outages and potential forwarding loops during reconvergence. Several vendors already support FRR[11]. The InteropNet Labs MPLS Initiative[12] announced the world's first multi-vendor Fast Reroute interoperability test at the NetWorld+Interop show during September 2002. Manufacturers are currently claiming LSP failover times ranging from 200 milliseconds down to as little as 5 milliseconds. A common "standard" is 50 or 60 milliseconds, in line with SONET/SDH restoration times. In contrast, IP shortest path first (SPF) rerouting can take in the order of several seconds[13] or much longer in large networks.

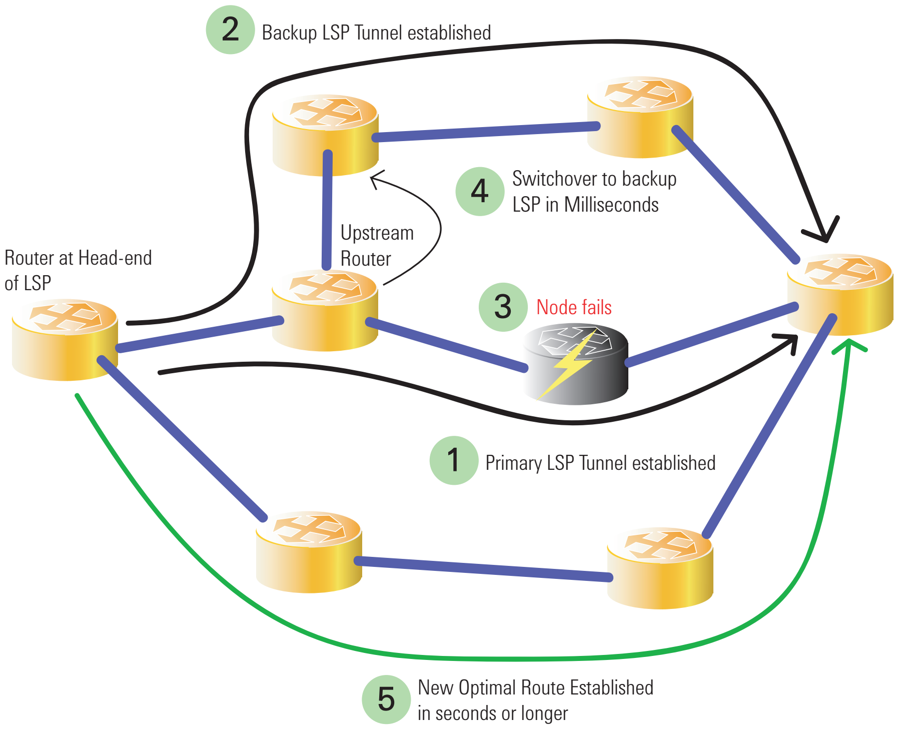

2 Backup LSP Tunnel established
4 Switchover to backup LSP in Milliseconds
Upstream Router
Router at Head-end of LSP
3 Node fails
1 Primary LSP Tunnel established
5 New Optimal Route Established in seconds or longer

Figure 3: MPLS Fast Reroute

MPLS Fast Reroute is not a complete solution for node protection. It cannot help if a network's ingress or egress router fails, and it is doubtful whether service providers would cooperate to provide backup LSP tunnels across their network borders. Nor does Fast Reroute prevent the need for route reconvergence following a failure (plus a second reconvergence after node recovery), or the associated storm of route update messages and potential route flapping. Furthermore, some implementations may still be proprietary and therefore may not interoperate in a multi-vendor network or between peer networks for carrier interconnect.

Hot Standby Control Cards

SONET APS, RPR and even the Fast Reroute mechanism of MPLS take a finite amount of time to failover so some packet loss is almost unavoidable. Moreover, backup links or paths may already be saturated with high-priority traffic or may not be able to offer the same Quality of Service as the primary LSP tunnel. In addition, if the fault is caused by a router failure, these mechanisms may not prevent the lengthy process of route reconvergence, which can disrupt the entire network with a storm of routing protocol update messages.

An alternative or complementary solution is to make a high-availability router that is resilient to faults, just like a traditional telephony switch.

Today's carrier-class routers typically offer redundancy for all or most hardware components. Power supplies, fans, switch fabrics and line cards can all be backed up by “hot standby” technology. Control cards (also known as route processors or routing engines), which process routing protocol information and maintain routing tables, can also be backed up — however, most of the “hot standby” control cards that are deployed today are less than ideal.

Typically, the backup control card is idle during normal operation of the primary control card. When the primary card fails, all of the session and state information is lost. The backup card goes through a lengthy procedure involving reboot, layer-2 configuration and link reestablishment, TCP and routing session reestablishment with adjacent routers, and reconvergence.

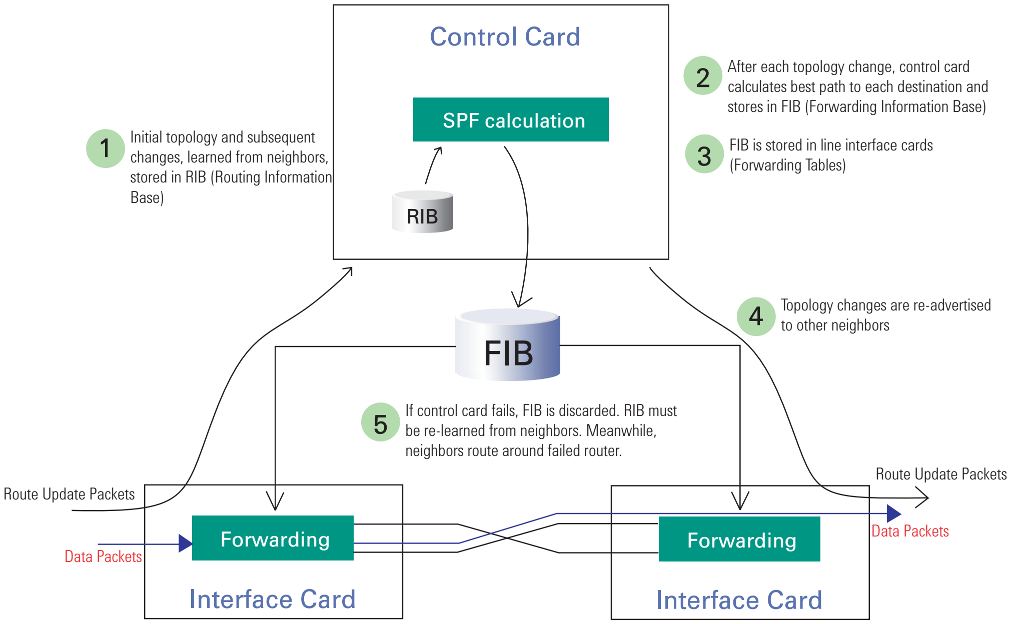

Control Card
SPF calculation
RIB
FIB
1
Initial topology and subsequent changes, learned from neighbors, stored in RIB (Routing Information Base)
2
After each topology change, control card calculates best path to each destination and stores in FIB (Forwarding Information Base)
3
FIB is stored in line interface cards (Forwarding Tables)
4
Topology changes are re-advertised to other neighbors
5
If control card fails, FIB is discarded. RIB must be re-learned from neighbors. Meanwhile, neighbors route around failed router.
Route Update Packets
Data Packets
Forwarding
Interface Card
Forwarding
Interface Card
Route Update Packets
Data Packets

Figure 4: Normal Router Operation

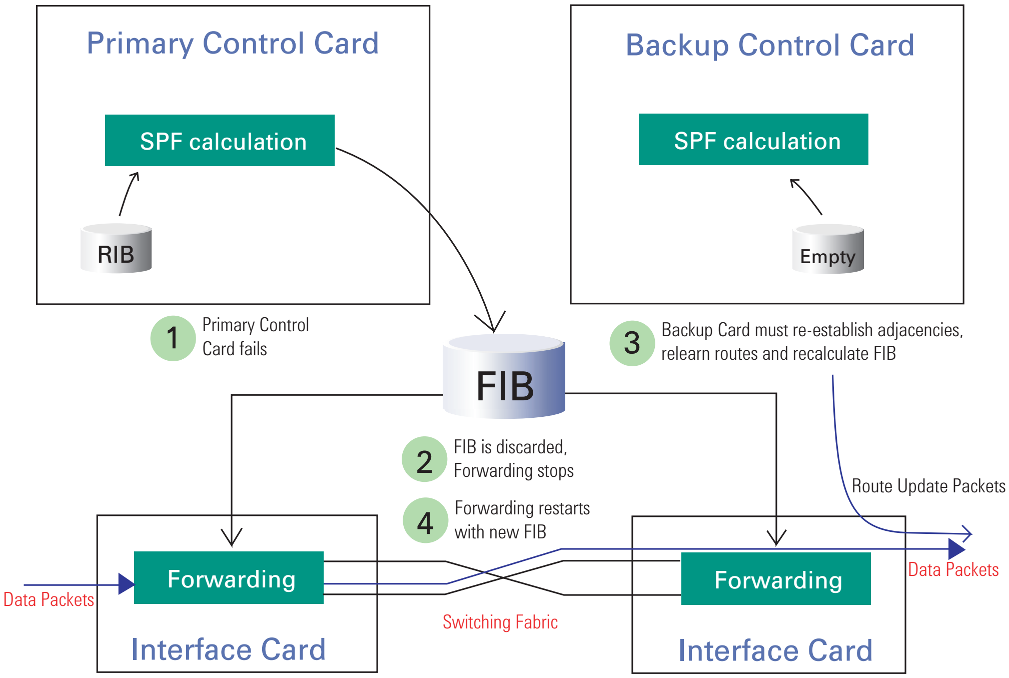


Figure 5: Hot Standby Control Cards

During reconvergence, the backup card relearns the routing topology from neighbor routers and rebuilds its routing table (also known as the Routing Information Base or RIB). This procedure involves the exchange of potentially thousands of route update messages using a protocol such as OSPF or IS-IS (for routers within the same Autonomous System) or BGP4 (for routers in neighboring domains). Route convergence can take from several minutes to tens of minutes to reach a consistent view of network topology after a fault because of routing advertisement oscillations (route flaps) during BGP's complicated path selection process. In a large network, the storm of routing protocol messages and processing overhead can bring a network to its knees. Router vendors are seeking to speed reconvergence with faster route processors and minor routing enhancements (such as route flap “damping”[14] to reduce route oscillation) but at the same time, networks (and therefore route tables) are continually becoming larger.

Once the routing table is synchronized, the router can calculate the forwarding table (also known as the Forwarding Information Base or FIB) for its line cards by calculating the best path to each destination. Packet forwarding can then resume. In reality, the FIB may be recalculated many times during reconvergence and partial routing states may be re-advertised. This entire process can take from five to 15 minutes[15] or much longer if there is significant route flapping or if manual intervention is required. Clearly, this is inadequate for a carrier-class router. “Five-9s” availability (99.999%) requires five minutes or less of downtime per year!

Non-stop Forwarding

A few vendors[16] have taken a leap forward by maintaining the router's forwarding table and link states and continuing packet forwarding during route processor failure and restart. During control module failover, layer-2 link state information is maintained for ATM, Frame Relay and Ethernet links, and the router continues to forward packets over routes that were available on the last-known state of the network. Failover is nearly instant and there should be little or no packet loss.

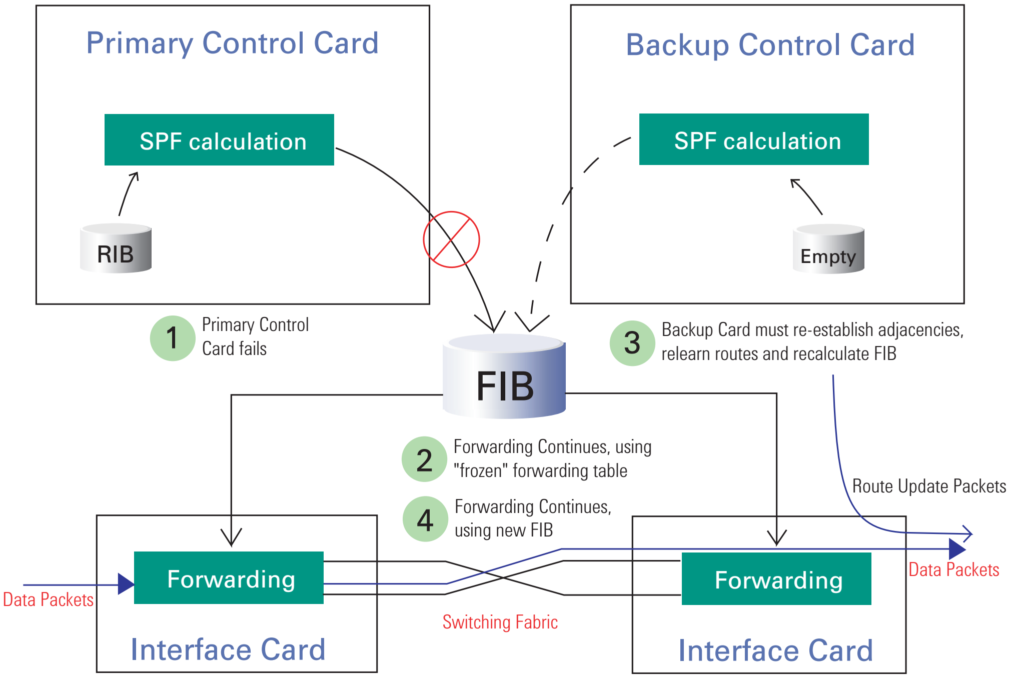


Figure 6: Non-stop Forwarding

Regardless of other resiliency mechanisms used, Non-stop Forwarding offers a big advantage because it prevents or reduces packet loss during reconvergence. The failed router must still relearn the network topology and recalculate its routing and forwarding tables, but while this is occurring, it continues to forward packets according to its existing forwarding table.

Critics of Non-stop Forwarding call it "headless forwarding" and claim that during failover, other network changes or failures could cause the failing router to ignore route update messages. The router would continue to forward packets according to its "stale" forwarding table, potentially resulting in forwarding loops, routing "black holes" and misdirected or discarded packets. Whether this is a likely or serious scenario depends on the frequency and magnitude of network route changes and remains to be tested.

Graceful Restart

There is an immense flurry of activity within the IETF to standardize protocol extensions to BGP[17], OSPF[18], IS-IS[19] and MPLS[20] that build upon "plain" Non-stop Forwarding and address some of its limitations. Known by names such as "Graceful Restart", these enhancements enable a router to stay on the forwarding path even as its routing software restarts. The restarting router continues to use its last-known, "frozen" forwarding table. The resiliency capabilities of several router vendors[21] are based on such extensions.

Without Graceful Restart, the neighbors of a restarting router remove the restarting router from their forwarding paths and reconverge on alternative paths, potentially causing a storm of route update messages and route flapping throughout the network. However, if the restarting router and all its immediate neighbors use the Graceful Restart procedures, the restarting router can rediscover its neighbors and relearn its routing table from them without triggering them to start reconvergence.

Graceful Restart relies on the cooperation of neighbor routers, which must implement the same protocol extensions as the restarting router. If just one neighbor does not know these extensions, then the fallback is to use the "regular" (slow) restart procedure.

Therefore, service providers using routers from multiple vendors or interfacing with other provider networks will be very keen to test router interoperability and provider interworking.

Route reconvergence is accelerated during Graceful Restart because route updates are communicated in a single block to the restarting router without interruption. There are no repeated recalculations of the FIB; the restarting router calculates the new FIB only once, after all updates have been communicated and acknowledged. Another advantage is that route updates are localized or isolated within the restarting router and its immediate neighbors. This prevents route flapping and network-wide instability.

Graceful Restart is not a silver bullet. If the network topology changes (for example, if there is another network failure) during the restart procedure, the restarting router may not be able to change its forwarding table quickly enough (or at all during restart), causing its entries to become “stale”. This could result in routing loops, network instability and lost packets. To avoid this, the Hitless Restart[22] proposal (an alternative to the OSPF draft cited above) suggests automatically reverting to “regular” restart procedures if such a topology change is detected before the restarting router is ready. This may be preferable but is obviously still not ideal. Because both scenarios are undesirable and could affect availability, service providers will want to test the likely impact on their networks.

Based on results from an experimental OSPF testbed, the recently proposed[23] IBB (“I'll Be Back”) enhancements to Graceful / Hitless Restart would solve the dilemma of how to react to topology updates during restart. When a “cooperating neighbor” of the restarting router receives a route update message, the neighbor first checks whether the update would impact any routes that traverse the restarting router. If the update would cause a routing loop or black hole, the neighbor establishes a new route around the restarting router. New protocol messages are proposed that would enable this. Because topology changes occur often in a large network, the IBB capability is potentially very useful for ensuring high network availability. If the processing overheads are indeed as small as the authors suggest, it is hoped that router vendors will implement this proposal within their carrier-class routers.

Carrier routers may need Graceful Restart for multiple routing protocols (BGP, OSPF and IS-IS) and for MPLS. It has been suggested that there could be undesirable interactions or dependencies between different routing protocols as they attempt to restart. Specifically, if the Interior Gateway Protocol (OSPF or IS-IS) is unstable or flaps during restart, it may trigger BGP to recommence its restart or to fallback to a “regular” restart. Service providers may wish to simulate and test this in their own evaluation laboratories.

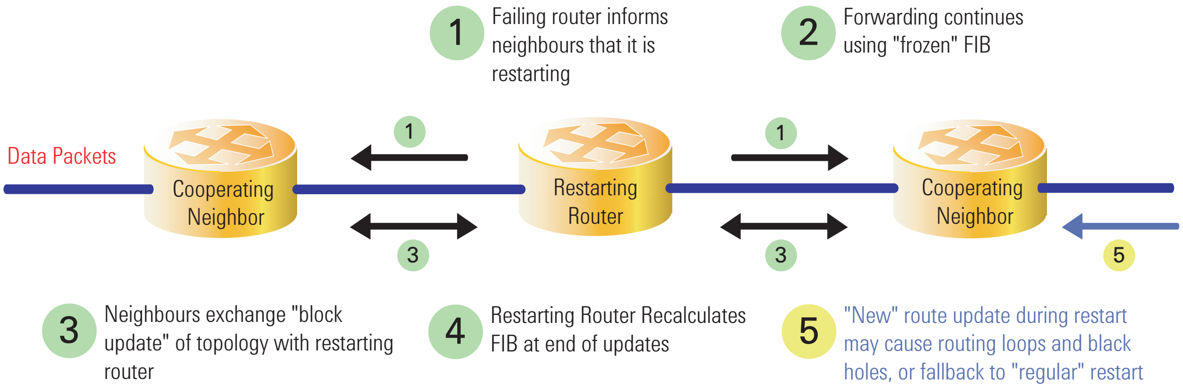

1 Failing router informs neighbours that it is restarting
2 Forwarding continues using "frozen" FIB
Data Packets
Cooperating Neighbor
1
Restarting Router
1
Cooperating Neighbor
3
3
5
3 Neighbours exchange "block update" of topology with restarting router
4 Restarting Router Recalculates FIB at end of updates
5 "New" route update during restart may cause routing loops and black holes, or fallback to "regular" restart

Figure 7: Graceful Restart

Non-stop Routing

At least two vendors[24] have announced routers with fully redundant “pre-booted” backup control cards, and other vendors are following. The backup maintains a complete copy of routing state information carried by the primary control card, including the routing table and routing sessions.

During failover, the router continues communication with its neighbors and persists in maintaining up-to-date routing and forwarding tables. There is little or no disruption to routing protocol interactions with the network. There is no need for a lengthy route reconvergence, no “flooding” of the network with update messages, and no packet loss. Router vendors claim[25] that this eliminates the need for duplicate hardware configurations — that is, only one router is needed in each POP.

Non-stop Routing enables new routes to be learned immediately following failover, avoiding the possibility of stale forwarding table entries, potential routing loops and "black holes". Control module restarts (following a failure or a software upgrade) are claimed to be a leading cause of router downtime, so Non-stop Routing should significantly reduce network outages.

Some implementations of Non-stop routing use mirrored routing cards, wherein the backup card runs the same processes as the primary card in lockstep — similar to telephony switches. The advantage of this approach is simplicity and instant or near-instant failover with practically zero possibility of routing loops or packet loss. However, if the primary card fails due to a software bug, the backup card may suffer exactly the same fate.

In contrast, the control cards of other vendors are "loosely coupled". The backup obtains updated state information from the primary card regularly (for example, once per second). Failover could take a little longer (perhaps a few seconds) but because the backup card does not run the same processes in lockstep with the primary card, fatal software bugs are less likely to be replicated. One could speculate that the backup card might fail if the software bug were somehow caused by a unique combination of routing state — for example, a very large and highly-meshed network topology — or by an errored or inappropriate routing message being sent repeatedly from a different vendor's router.

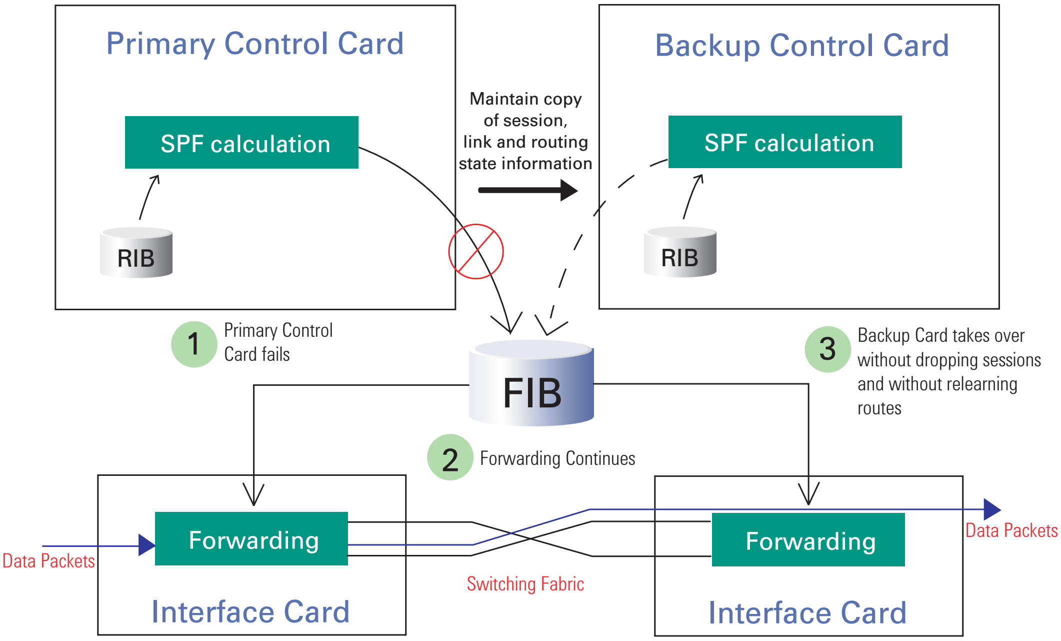


Figure 8: Non-stop Routing

At least one vendor is confidently backing its NSR technology with a Service Level Agreement[26] (SLA) guaranteeing 99.999% availability in a single router.

Another vendor is developing a router with a flexible architecture inspired by supercomputer microprocessing, with tasks distributed and shared amongst multiple processors. Recovery is claimed to be "hitless" in the event of processor failure.

Critics of Non-stop Routing (NSR) argue that the benefits of NSR are small compared to Graceful Restart, especially at the edge of the network where there are fewer routes to learn and where routes are generally more stable. Some vendors claim that the added costs of Non-stop Routing cannot be justified because routers are not yet sufficiently mature to avoid deployment of redundant routers at every POP. On the other hand, Graceful Restart is yet to be verified in a large, multi-vendor network with a realistic simulation of Internet-scale traffic, multiple routing protocols and MPLS. These claims and counter-claims are yet to be tested. In any case, these resiliency technologies are not incompatible and we may soon see carrier-class routers that offer both Graceful Restart and Non-stop Routing.

Hitless Upgrade

Two major service providers suffered network outages of up to 26 hours that were suspected to have been caused by faulty[27] or insufficiently-tested[28] software during a switch upgrade. Network upgrades are, in fact, the most common source[29] of outages because new defects are easily introduced to a network during software updates or when hardware replacement procedures are followed incorrectly. Many vendors are now touting the ability to upgrade router software while the device is in service, without downtime or dropped packets. This feature is often called "hitless upgrade" or "in-service upgrade".

Routers that implement Non-stop Routing have a distinct advantage here. To perform Hitless Upgrade, the software on the backup control card is first updated. Secondly, control is transferred to the backup card so that the software on the primary card can be updated. Finally, control can be transferred back to the primary card.

To upgrade a router that implements Non-stop Forwarding or Graceful / Hitless Restart, the routing engine is stopped and packets are forwarded using the frozen FIB while the software is updated. When the upgrade is complete, the control card is rebooted, the topology is relearned and the FIB is recalculated. Availability may be compromised if external network topology changes occur during this procedure.

Hitless Upgrade may be achieved in other devices[30] by using a modular architecture that isolates memory from failures.

Redundant Configurations

Regardless of the resiliency of individual routers, two or more routers can work together in redundant configurations to provide high availability. Multi-homing is a technique commonly used at the customer-provider boundary or between providers to safeguard against single link failures by supplying redundant links. For example, many corporate networks use two WAN links and most ISPs use at least two diverse connections to their own providers to ensure uninterrupted access to the Internet.

Measuring Routing Resiliency

Although the emergence of high-availability carrier-class routers is generally good news for service providers and their customers, it can be challenging to evaluate the plethora of new resiliency technologies and competing vendor claims. Each provider network is unique — different routing protocols are deployed, different topologies are already being used to provide redundancy, available services and SLA offerings vary, and each network also has unique legacy router interworking challenges. While every service provider will have its own test requirements, we present here a general strategy for evaluating and measuring some of the newer resiliency technologies that we have discussed. We will followup this article with a new series of test cases for the industry-standard Journal of Internet Test Methodologies[31], the reference used by equipment developers and service providers worldwide as the basis for their test plans.

Evaluating and Measuring Non-stop Forwarding

Verification of Non-stop Forwarding requires test equipment that tightly integrates IP packet generation and performance measurement with the simulation of large networks through emulation of multiple routing protocols.

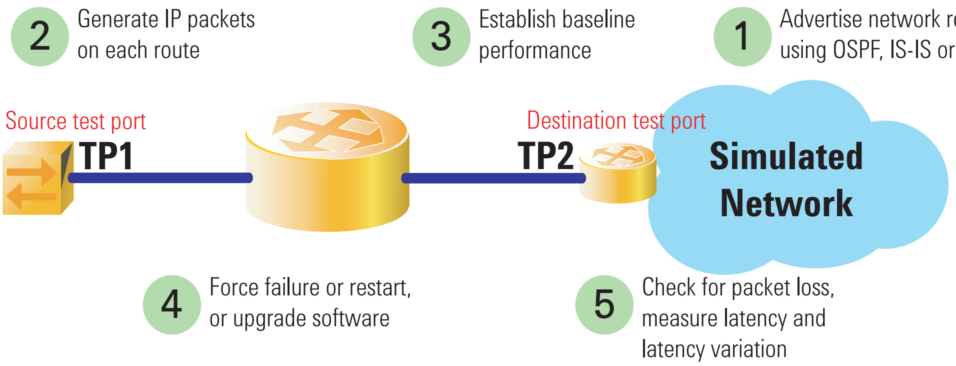

2 Generate IP packets on each route
3 Establish baseline performance
1 Advertise network routes using OSPF, IS-IS or BGP4
Source test port
TP1
Destination test port
TP2
Simulated Network
4 Force failure or restart, or upgrade software
5 Check for packet loss, measure latency and latency variation

Figure 9: Testing Non-stop Forwarding

To verify a router's ability to maintain packet forwarding during failover, first advertise a set of routes (using OSPF, BGP4 or IS-IS) from your test equipment, send traffic on those routes through the System Under Test (SUT), establish that traffic is being forwarded with zero or negligible loss, and baseline the latency and latency variation. Failover can then be initiated by extracting a primary control card or by forcing the routing engine or routing processes to restart. Continue to test forwarding performance until recovery is complete. Measure the packet loss and time-to-recovery of packet forwarding; in theory, there should be no packet loss. Compare the latency and latency variation measured during the test against baseline performance to determine the potential impact on SLAs.

A similar method can be used to evaluate “hitless upgrade”. Instead of restarting or extracting a control card, perform a “hot” upgrade of the router software during the test and measure performance degradation during the upgrade.

Evaluating and Measuring Graceful Restart

Graceful Restart requires both the restarting router and its immediate neighbors to implement the Graceful Restart procedures and routing protocol extensions. Therefore, it is important to test a “complete system” — including both a restarting router and two or more cooperating neighbors (sometimes known as receiving routers). The functionality and performance of both the restarting router and its cooperating neighbors will affect the outcome of the test. The test equipment should be connected only to ports of the cooperating neighbors — it should not be connected directly to the restarting router.

For this reason, Graceful Restart can be successfully verified and measured using test equipment that does not emulate the Graceful Restart extensions. This is also helpful because (at the time of writing) Graceful Restart is still a series of IETF drafts that are under development and subject to change.

There are five tests of interest:

1. Functional Test: Verification of Graceful Restart's continuous forwarding capability
2. Topology Change Test: Characterization of the impact of external topology changes on Graceful Restart
3. Performance Test: Measurement of Restart duration
4. Live Upgrade Test: Run the same tests above but instead of restarting the control card, re-install the router software to simulate an “in-service” upgrade
5. Interoperability Test: Run the above tests using restarting and receiving routers from different vendors

These five tests can be collapsed down to two test scenarios, described below.

To test continuous forwarding, configure a test system with three routers connected in sequence and three test ports (see Figure 10). Router RR is used as the restarting router, while routers CN1 and CN2 are used as the cooperating neighbor routers. Test instrument ports TP1 and TP3 are connected to router CN1, while test port TP2 is connected to router CN2. TP1 is used as a source test port, whereas TP2 and TP3 are destination test ports.

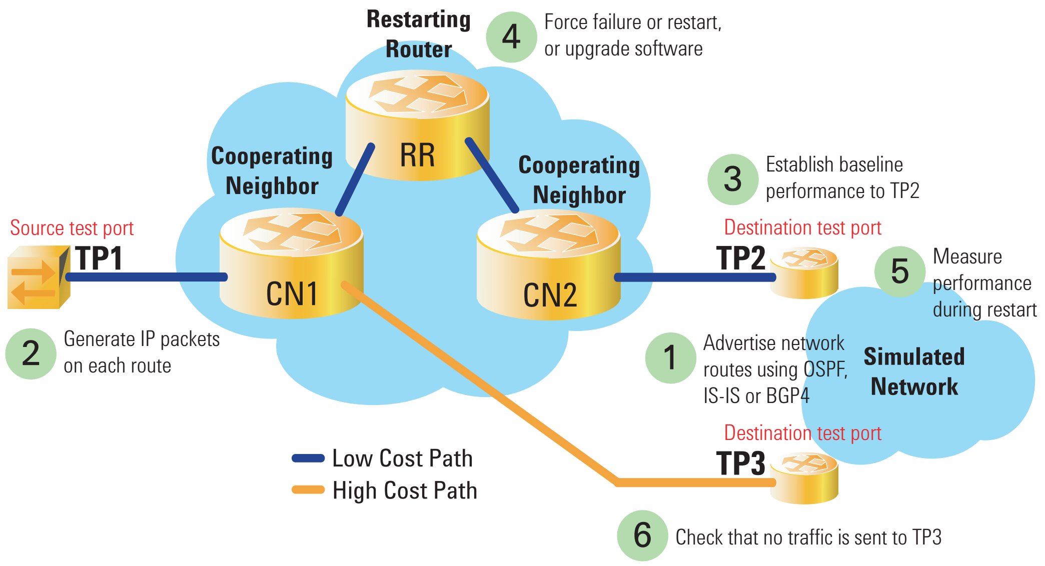

Restarting Router
4
Force failure or restart, or upgrade software
RR
Cooperating Neighbor
Cooperating Neighbor
3
Establish baseline performance to TP2
Source test port
TP1
CN1
CN2
Destination test port
TP2
5
Measure performance during restart
2
Generate IP packets on each route
1
Advertise network routes using OSPF, IS-IS or BGP4
Simulated Network
Destination test port
TP3
Low Cost Path
High Cost Path
6
Check that no traffic is sent to TP3

Figure 10: Testing Graceful Restart — Continuous Forwarding

Simulate a network behind destination ports TP2 and TP3 by advertising routes using OSPF, BGP4 or IS-IS. Setup link costs (weights) such that the default route to the simulated network is via the low cost path TP1-CN1-RR-CN2-TP2 and the alternative route TP1-CN1-TP3 offers a higher-cost path to the same simulated network.

As in the Non-stop Forwarding test, advertise a large number of routes through the SUT, generate test traffic from TP1, establish traffic forwarding, and baseline the packet performance (loss, latency and latency variation) to TP2. Force control card failure or restart to occur in router RR (or perform a live upgrade of the router software). Ensure continuity of forwarding to TP2 and measure IP performance both during the restart procedure and at the end of restart (when the FIB is recalculated). The performance of the cooperating routers (CN1 and CN2) also enters the equation since they are on the forwarding path.

When restart is complete, turn off traffic generation. As a final check, ensure that no traffic was rerouted at any time via the alternative high-cost path (i.e., to TP3).

A similar configuration is used to test the impact of an external topology change on Graceful Restart and to measure Restart duration (see Figure 11). However, this time a third "cooperating neighbor" router is required.

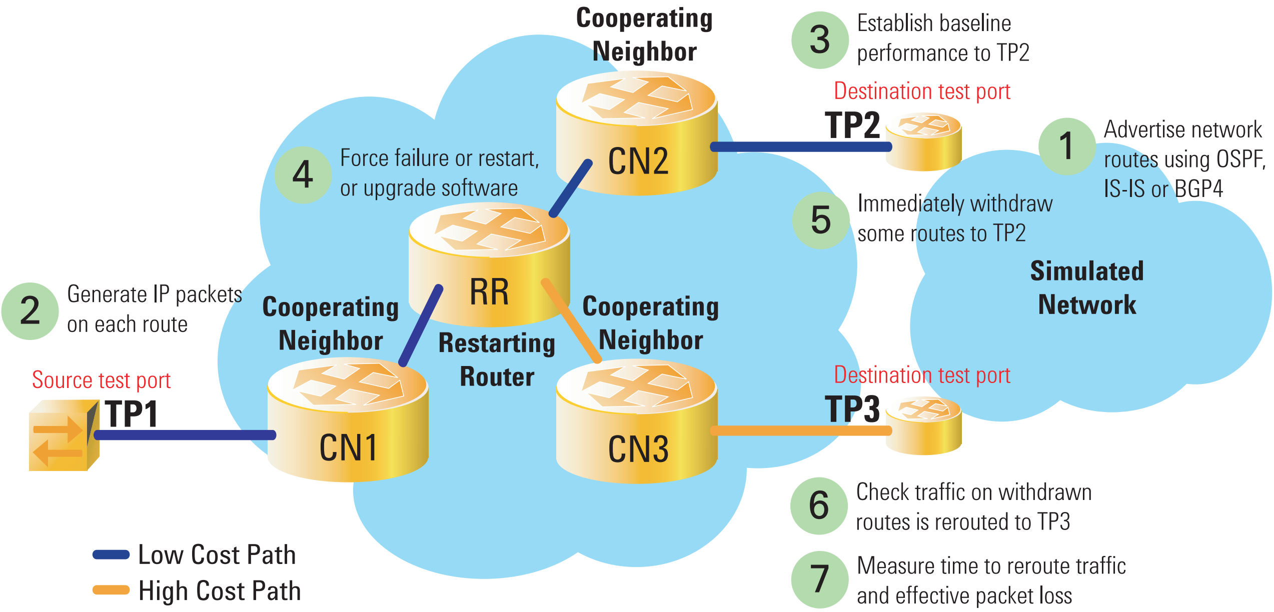

Cooperating Neighbor
CN2
3 Establish baseline performance to TP2
Destination test port
TP2
1 Advertise network routes using OSPF, IS-IS or BGP4
4 Force failure or restart, or upgrade software
5 Immediately withdraw some routes to TP2
Simulated Network
RR
Restarting Router
2 Generate IP packets on each route
Cooperating Neighbor
Cooperating Neighbor
Source test port
TP1
CN1
CN3
Destination test port
TP3
6 Check traffic on withdrawn routes is rerouted to TP3
7 Measure time to reroute traffic and effective packet loss
Low Cost Path
High Cost Path

Figure 11: Testing Graceful Restart - Delayed Reroute

Begin by following the same steps as for testing continuous forwarding: simulate a network behind test ports TP2 and TP3, advertise the high and low cost routes, generate IP test traffic from TP1 to TP2 and measure the baseline forwarding performance.

Following failure or upgrade of the restarting router, immediately withdraw some routes (on the low cost path) at TP2. While the system is busy "gracefully restarting", this topology change should have no effect on forwarded traffic. The restarting router will use its "stale" forwarding table to continue to forward all traffic to TP2. The restarting router will not yet have knowledge of the withdrawn routes (or if it does, it will not use this knowledge to recalculate its FIB until Graceful Restart is complete).

When restart is complete, RR will learn the topology change from CN2. It will then rebuild its FIB and re-advertise the topology change to its neighbors CN1 and CN3. Traffic on the withdrawn routes will now be forwarded by the restarting router to the high-cost path via TP3. Check that traffic on those routes is being forwarded to TP3 and compare the IP packet performance to the baseline.

Finally, to quantify the restart duration, note the time taken for the first packet to arrive at TP3 following commencement of restart. Calculate also the "effective" packet loss — that is, the packets that were still arriving at TP2 following the withdrawal of their routes.

Evaluating and Measuring Hitless Restart

As discussed earlier, the OSPF Hitless Restart procedure differs from the alternative Graceful Restart procedure in at least one important aspect: when a topology change occurs, Hitless Restart "aborts" and falls back to a "regular" restart rather than allowing the possibility of sustained packet loss during restart due to routing loops and black holes. We can use the same procedure as in Graceful Restart to test continuous forwarding. However, to test that a topology change causes Hitless Restart to abort and to measure Hitless Restart performance, we need two additional test scenarios.

The first Hitless Restart test scenario is almost identical to the last scenario described under Graceful Restart. The simulated network is configured behind test ports TP2 and TP3, routes are advertised, packet traffic is generated, performance is baselined, the router is restarted or upgraded, and some of the routes to TP2 are withdrawn (see Figure 12).

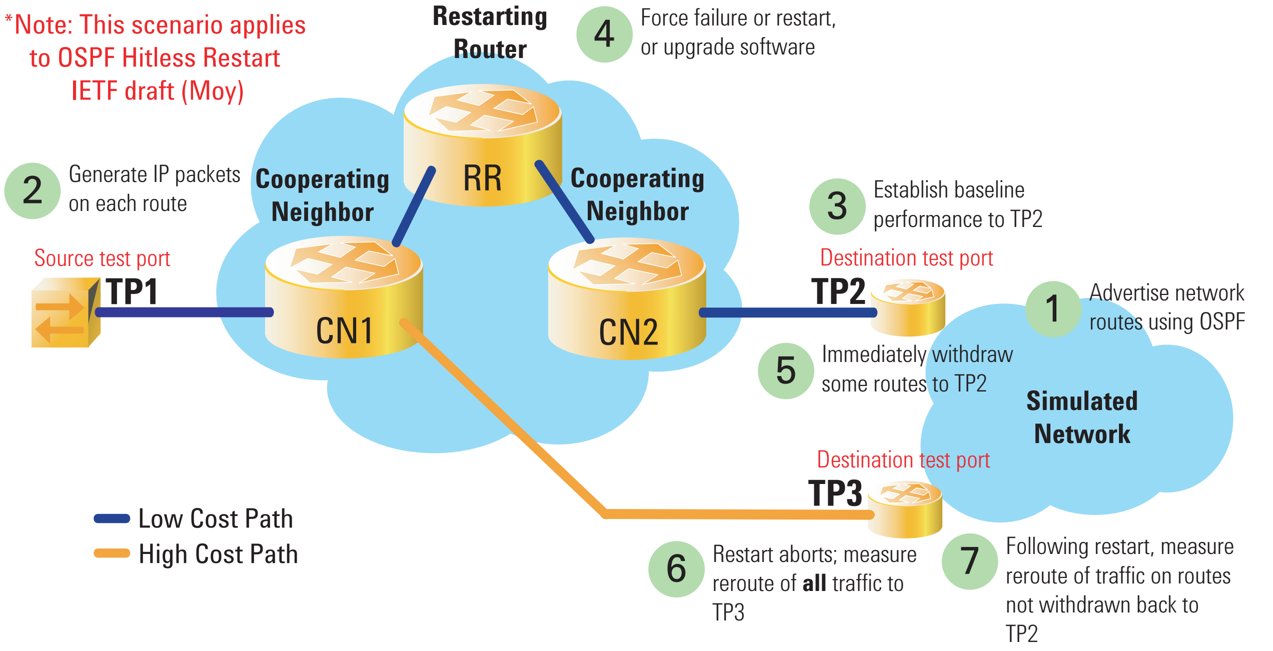

*Note: This scenario applies to OSPF Hitless Restart IETF draft (Moy)
Restarting Router
4 Force failure or restart, or upgrade software
2 Generate IP packets on each route
Cooperating Neighbor
RR
Cooperating Neighbor
3 Establish baseline performance to TP2
Source test port
TP1
CN1
CN2
Destination test port
TP2
1 Advertise network routes using OSPF
5 Immediately withdraw some routes to TP2
Simulated Network
Destination test port
TP3
Low Cost Path
High Cost Path
6 Restart aborts; measure reroute of **all** traffic to TP3
7 Following restart, measure reroute of traffic on routes not withdrawn back to TP2

Figure 12: Testing Hitless Restart - Abort on Topology Change

Withdrawal of these routes should cause Hitless Restart to abort. The three routers should fall back to a "regular" restart. The restarting router should discard its FIB and stop all traffic forwarding. Router CN1 should reroute traffic to the high cost path (to TP3). Check that all traffic is rerouted to TP3, measure the reconvergence time by noting the time taken for packets to begin arriving at TP3, and compare IP performance to the baseline.

Following "regular" restart, the restarting router will announce its presence and reconvergence will occur again. This time, traffic on all except the withdrawn routes will be routed back through the restarting router to TP2. Once again, check that the appropriate traffic has been rerouted to TP2, measure the reconvergence time by noting the time taken for packets to begin arriving at TP2, and compare IP performance to the baseline.

The above procedure measures regular restart / reconvergence duration. To measure Hitless Restart duration, an additional test scenario is required (see Figure 13).

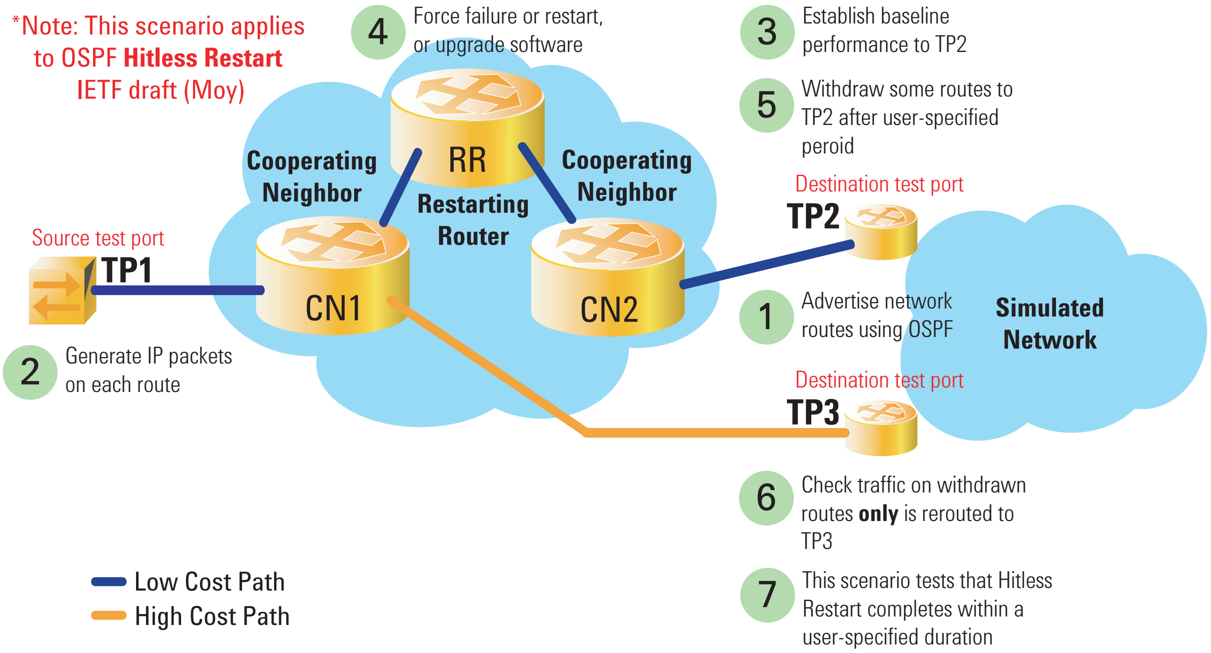

*Note: This scenario applies to OSPF **Hitless Restart** IETF draft (Moy)
4
Force failure or restart, or upgrade software
3
Establish baseline performance to TP2
5
Withdraw some routes to TP2 after user-specified peroid
Cooperating Neighbor
RR
Restarting Router
Cooperating Neighbor
Destination test port
TP2
Source test port
TP1
CN1
CN2
1
Advertise network routes using OSPF
Simulated Network
2
Generate IP packets on each route
Destination test port
TP3
6
Check traffic on withdrawn routes **only** is rerouted to TP3
Low Cost Path
High Cost Path
7
This scenario tests that Hitless Restart completes within a user-specified duration

Figure 13: Testing Hitless Restart - Restart Time

Once again, simulate the network, generate test traffic, baseline IP performance and restart or upgrade the router. This time, wait for a user-specified period before continuing. This period should be related to the desired maximum restart duration. During this period, Hitless Restart should complete.

At the end of the specified period, withdraw some of the routes to TP2. If Hitless Restart has indeed completed, only the traffic on the withdrawn routes will be rerouted to TP3. If Hitless Restart has not yet completed then the Hitless Restart procedure will abort, causing fallback to a "regular" restart, in which all traffic will be rerouted to TP3.

This procedure cannot directly measure the duration of Hitless Restart, but it can verify that Hitless Restart occurs within a user-specified duration. It does this by injecting a topology change, causing Hitless Restart to abort if it has not yet completed. This test scenario can be repeated with different "wait periods" to estimate Hitless Restart duration to a greater accuracy.

Evaluating and Measuring Non-stop Routing

Of all the resiliency technologies described in this paper, Non-stop Routing is arguably the most difficult to measure. If it works perfectly, the control card failover is almost instant and undetectable, and there is nothing to measure.

Configure the test in the same way as the Non-stop Forwarding test case, connecting three test instrument ports to the router. Simulate a network behind destination test ports TP2 and TP3 by advertising a large number of routes using OSPF, IS-IS or BGP4 from both test ports. Configure the routes through TP3 as higher-cost paths so that by default, all traffic will be routed to TP2. Load each route with IP test packets from TP1 and establish baseline packet performance to TP2 (see Figure 14).

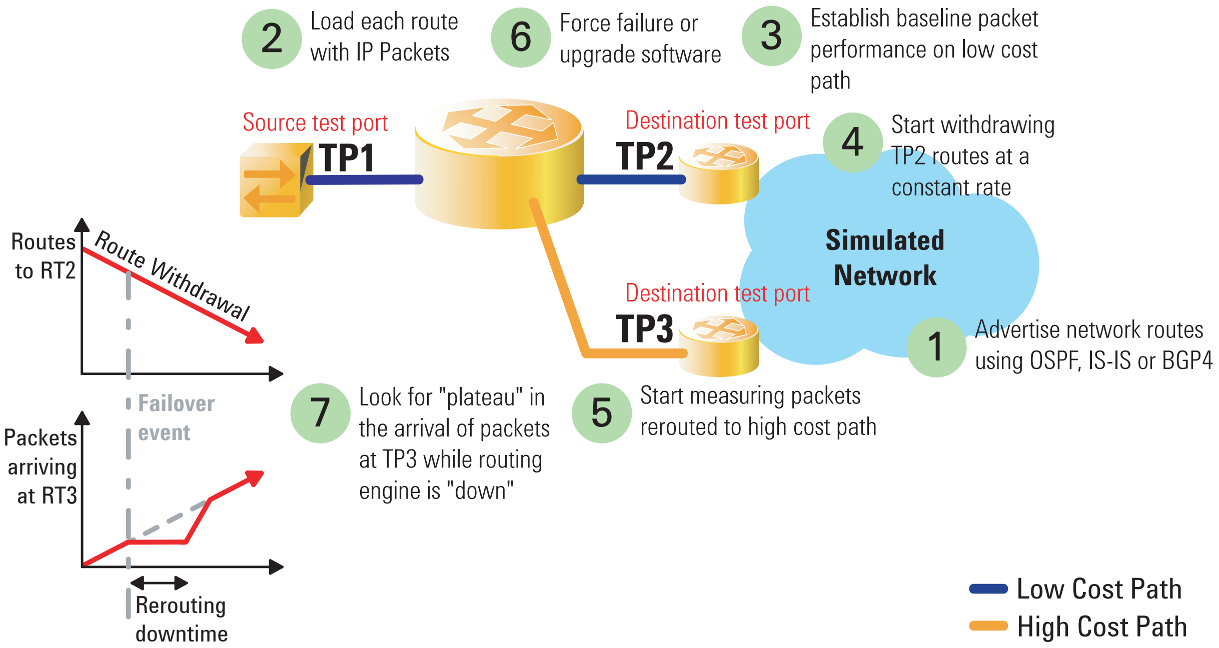

2 Load each route with IP Packets
6 Force failure or upgrade software
3 Establish baseline packet performance on low cost path
Source test port
TP1
Destination test port
TP2
4 Start withdrawing TP2 routes at a constant rate
Simulated Network
Destination test port
TP3
1 Advertise network routes using OSPF, IS-IS or BGP4
Routes to RT2
Route Withdrawal
Failover event
Packets arriving at RT3
7 Look for "plateau" in the arrival of packets at TP3 while routing engine is "down"
5 Start measuring packets rerouted to high cost path
Rerouting downtime
Low Cost Path
High Cost Path

Figure 14: Testing Non-stop Routing

The next parts of this test will require some automation within the test instrument. Start withdrawing routes from TP2 at a constant rate (see the first graph in the diagram above). As each route is withdrawn, after a short “rerouting delay”, the router will forward traffic to the alternative, higher-cost path (TP3). At the same time, measure packet performance at TP3. In particular, graph the bit rate or bandwidth of packets arriving at TP3 (see the second graph in Figure 14). You should begin to see a straight line, perhaps with some small “bumps” caused by variations in the rerouting speed.

Force control card failover by extracting the router's primary control card, restarting its routing processes or initiating software upgrade. Continue to withdraw routes from TP2 and continue to measure packet performance at both ports.

If the router's failover to the backup control card is almost instantaneous (sub-second) and seamless, rerouting performance will be unaffected and the graph of packet bandwidth will continue along a relatively straight line (see dashed line in the second graph above).

If the router takes more than a few seconds to failover and to establish routing processes on the backup control card, then a plateau will be observed in the second graph during the routing engine's “downtime” (see solid line in the second graph above). The duration of this plateau is a good estimate of the time taken for the router to failover to the backup control card. When routing processes are fully reestablished, traffic on the withdrawn routes should be quickly redirected to TP3 and the packet bandwidth at TP3 should “catch up” to the expected rate. Ensure that this “catch up” does occur — that is, ensure that all route withdrawals were processed and none were missed.

Although this test scenario was designed for testing Non-stop Routing, it can be used to test the impact of other resiliency technologies on routing performance.

Evaluating and Measuring MPLS Fast Reroute

Testing MPLS Fast Reroute is a complex procedure that requires a test instrument that tightly integrates simulation of network topologies (using OSPF or IS-IS), emulation of MPLS signaling protocols, and generation and performance measurement of labeled test traffic. For test scenarios such as this that require multiple protocols, an automated application is generally preferred.

To measure the time it takes to redirect traffic to a backup LSP tunnel after the primary LSP has been dropped, three router test ports are needed: one source port, a destination port for the primary LSP, and a second destination port for the backup LSP. Begin by simulating a common network topology on the two destination ports of the router under test using OSPF or IS-IS. Establish the LSP tunnels and generate traffic on the source port through the router. Physically sever the link between the SUT and the primary destination port to force failover from the primary to the backup LSP tunnel. Finally, measure reconvergence time and the time taken to reroute traffic to the backup destination port.

When MPLS reroutes traffic to a backup LSP tunnel, the backup LSP should respect the same Class of Service (CoS) priorities configured on the primary LSP. Configure the SUT and traffic source with several different streams and repeat the above test, confirming that the backup LSP maintains the relative stream priorities.

For a complete description of the MPLS Fast Reroute test scenario, see Journal Test Case 53 (JTC 053) of The Journal of Internet Test Methodologies[31]. The Journal includes detailed test configurations, procedures, variables and expected results.

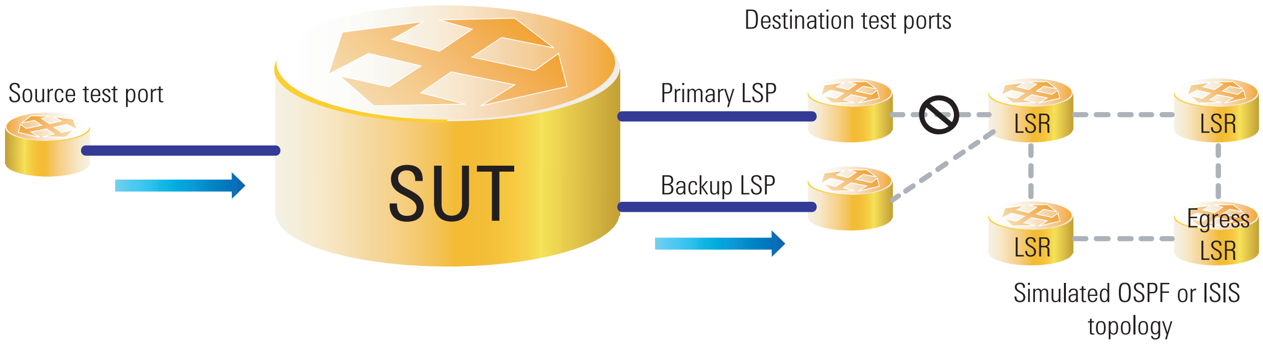


Figure 15: Verifying MPLS Fast Reroute

Conclusion

Router vendors recently introduced new resiliency technologies such as Non-stop Forwarding, Non-stop Routing, Graceful Restart and MPLS Fast Reroute to improve the reliability and availability of their carrier-class routers. The jury is still out on which technologies will dominate but it is likely that service providers will select a combination of several approaches at different protocol layers to achieve the IP network availability that they require to offer new services, reduce downtime and lower operational costs.

Service providers need to verify the claims of equipment vendors and evaluate these technologies for their own network applications. In this paper, we have compared and contrasted the approaches taken by different vendors and outlined several methodologies to verify and measure router resiliency. Service providers can use these test methodologies during the evaluation, benchmarking, acceptance testing and deployment of carrier-class core and edge routers. Similarly, equipment vendors can demonstrate the reliability and fault-tolerance capabilities of their latest routers to customers within their Proof-of-Concept laboratories.

We will illustrate and describe these Router Resiliency test methodologies in more detail and develop them into fully-fledged test cases in a future edition of The Journal of Internet Test Methodologies.

References

[1] Infonetics Research, "The Tier 1 Service Provider Opportunity, US/Canada 2001", Nov 2001; http://www.infonetics.com/pdf/press/nr_spo_us_t1_01.pdf

[2] BTexact Technologies, "Carrier requirements of core IP routers — 2002", Feb 2002; http://www.btexact.com/ideas/whitepapers?doc=42267

[3] Network Magazine, Nov 2001; http://www.networkmagazine.com/article/NMG20011102s0005 and http://img.cmpnet.com/networkmag2000/content/200111/non_anatomy_table.gif

[4] Yankee Group, "Core Routers: Challengers & Challenges to the Status Quo", Carrier Convergence Infrastructure Report, Vol 2, No. 11, Oct 2001

[5] USA Today, http://www/usatoday.com/; see also http://www.sciodata.com/home.asp

[6] "Experimental Study of Internet Stability and Wide-Area Backbone Failure", Craig Labovitz and Abha Ahuja, Merit Networks; data for this study was taken from MichNet (ISP) NOC Internet log from 33 backbone routers and hundreds of customer routers; see http://www.unix.ecs.umass.edu/~lgao/class/routing/lect11.ppt

[7] The Risks Digest, Volume 20 Issue 5, http://catless.ncl.ac.uk/Risks/20.05.html#subj4

[8] Outline of the IEEE 802.17 RPR Draft Standard, Version 0.3, June 2002, Resilient Packet Ring Alliance; http://www.rpralliance.org/articles/80217Outline.pdf

[9] "The PPP Multilink Protocol (MP)", IETF RFC 1990, http://www.ietf.org/rfc/rfc1990.txt

[10] "End-to-End Multilink Frame Relay Implementation Agreement", FRF.15, Frame Relay Forum, http://www.frforum.com/5000/Approved/FRF.15/frf15.pdf

[11] For some vendor examples of MPLS FRR implementations, see: Cisco FRR Feature Module http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120limit/120st/120st16/frr.htm#xtocid0; Juniper Solution Brief http://www.juniper.net/solutions/sol_prof/351001.html; "Lucent Technologies unveils TMX 880 MPLS core switch", Mar 12, 2002 http://www.lucent.com/press/0302/020312.nsc.html; Riverstone Webinar http://www.riverstonenet.com/webinar_archive/carrier_class/q_and_a.shtml

[12] See http://www.ilabs.interop.net/details?topic=MPLS

[13] Avici White Paper, "Building a Reliable and Scalable Internet", IEEE HPSR 20002, Chris Gunner, May 2002; http://www.ieice.org/hpsr2002/documents/archives/IEEE_HSPR_mtg_Final.pdf

[14] IETF RFC 2439 — BGP Route Flap Damping http://www.ietf.org/rfc/rfc2439.txt

[15] Communications Design, Apr 2001; http://www.commsdesign.com/

[16] For example, Cisco announced "Non Stop Forwarding" (NSF) and "Stateful Switchover" (SSO) in May 2002; see Cisco GRIP Press Release, May 14 2002 http://www.cisco.com/warp/public/732/Tech/grip/press.shtml; see also http://newsroom.cisco.com/dlls/innovators/Core_IP/banks_marthe_q_a.html

[17] "Graceful Restart Mechanism for BGP", draft-ietf-idr-restart-xx.txt

[18] Draft-nguyen-ospf-lls-xx.txt, draft-nguyen-ospf-oob-resync-xx.txt, draft-nguyen-ospf-restart-xx.txt

[19] Draft-ietf-isis-restart-xx.txt, draft-ash-ospf-isis-congestion-control-xx.txt

[20] "Graceful Restart Mechanism for LDP" draft-ietf-mpls-ldp-restart-xx.txt and "Graceful Restart Mechanism for BGP with MPLS", draft-ietf-mpls-bgp-mpls-restart-xx.txt

[21] See "Cisco Nonstop Forwarding", http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120limit/120s/120s22/nsf120s.htm; "The Trusted Edge: Edge Routers with Backbone", Juniper Solution Brief http://www.juniper.net/solutions/sol_prof/351003.html; and Riverstone Networks White Paper — "Toward a Hitless Network: BGP Graceful Restart" http://www.riverstonenet.com/technology/bgp_restart.shtml

[22] "Hitless Restart for OSPF", draft-ietf-ospf-hitless-restart-xx.txt and "Hitless Extended Restart for OSPF", draft-lindem-ospf-hitless-extended-restart-xx.txt

[23] "Avoiding Instability during Graceful Shutdown of OSPF", Aman Shaikh, Rohit Dube and Anujan Varma, UCSC, http://www.ieee-infocom.org/2002/papers/713.pdf

[24] See Alcatel press release, April 15 2002 http://www.cid.alcatel.com/doctypes/newsrelease/html/20020415.jhtml and Avici NSR white paper http://www.avici.com/technology/whitepapers/NSRTechnology.pdf

[25] Alcatel ACEIS media release; http://www.cid.alcatel.com/doctypes/leadstory/aceis/ACEIS_media.pdf

[26] Avici press release; http://63.111.106.66/press/2002/pr20020604.shtml

[27] The Risks Digest, Volume 19 Issue 72, http://catless.ncl.ac.uk/Risks/19.72.html#subj6.1; see also http://www.att.com/press/0498/980422.bsb.html

[28] “MCI-Worldcom Outage”, August 1999, http://www.mids.org/TQR/101/mid/mci.html

[29] See “Network Outage Hits ATT's ATM Users”, Computerworld, Feb 2001 http://www.itworld.com/Net/2318/CWSTO58077/

[30] For example, Equipe's Evail service switch; see “Start-up aims to end carrier switch-upgrade headaches”, Jun 2000, http://www.nwfusion.com/edge/news/2000/0621edgeevail.html

[31] The Journal of Internet Test Methodologies; http://www.agilent.com/comms/TheJournal

November 22, 2002
5988-8459EN

Agilent Technologies