導入ガイド

Cisco NSF/SSO（ノンストップ フォワーディング / ステートフル スイッチオーバー）導入ガイド

Cisco Nonstop Forwarding（NSF; ノンストップ フォワーディング）/Stateful Switchover（SSO; ステートフル スイッチオーバー）は、ハードウェアまたはソフトウェア障害による意図しないダウンタイムからシステムを保護し、ネットワーク サービスの継続的な提供を可能にします。NSF/SSO をネットワークの重要な部分に展開すれば、システムとサービスの可用性が向上するとともに、将来、In Service Software Upgrade（ISSU）で提供されるダウンタイムの計画化を目的とした機能を活用できるようになります。

重要なビジネス システムへのネットワーク アクセスを実現しようとしている企業においても、卓越したネットワーク サービスと接続性を顧客に提供することを追求しているネットワーク プロバイダーにおいても、コンポーネントの障害によって発生するダウンタイムを短縮することが業務には必須であると認識しています。シスコのお客様は、ビジネス要件と予算を考慮しながら、サービスが停止することのない冗長ネットワークの設計と運用の実現に向けて努力しています。

Cisco NSF/SSO は、ある種のネットワーク機能停止による影響を抑制するために開発された拡張機能から発展しました。Cisco NSF/SSO は、Route Processor Redundancy（RPR）および RPR Plus（RPR+）として知られる初期のテクノロジーを基盤としています。シャーシ内でハードウェアを冗長化（冗長 Route Processor [RP; ルート プロセッサ]）し、コントロール プレーンをデータ プレーンから分離することによって、たとえハードウェアまたはソフトウェアに障害が発生してそれが RP に支障を与えたとしても、パケット損失することなく継続的にパケットを転送できるようにします（独自のテスト結果[英語]は http://www.cisco.com/warp/public/732/Tech/grip/tech.shtml を参照してください）。

この文書は、Cisco NSF/SSO を展開することでネットワークサービスの可用性を高めようとしている設計スタッフと運用スタッフのための手引き書です。最初の項では、ネットワーク内で NSF/SSO を展開すべきポイントについて説明します。2 項と 3 項では、SSO と NSF の運用について検討します。4 項では、確実に展開を成功させるための実装手順について説明します。

注：この文書では、特に断りのないかぎり、「RP（ルート プロセッサ）」という用語は、ハードウェアの名称に関係なく、すべてのネットワーキング デバイス上のルート プロセッサ エンジンを指します。たとえば、Cisco 10000 シリーズ インターネット ルータでは、RP は Performance Routing Engine（PRE; パフォーマンス ルーティング エンジン）を指し、Cisco 12000 シリーズ ルータでは、RP は Gigabit Route Processor（GRP; ギガビット ルート プロセッサ）または Performance Route Processor（PRP; パフォーマンス ルート プロセッサ）を指します。また、Cisco Catalyst® 6500 シリーズ スイッチと Cisco 7600 シリーズ ルータではスーパーバイザという用語が使用され、Cisco 7500 シリーズ ルータの RP は Route Switch Processor（RSP; ルート スイッチ プロセッサ）を指します。

NSF/SSO の展開

Cisco NSF/SSO という言葉からは、すべてのネットワーク ノードに復元力の向上というメリットを提供するように思えます。しかし、実際には、この機能から最大の効果を得られるのはエッジ デバイスです。シングルポイント オブ フェイラーは、ネットワーク エッジの境界に存在する傾向があります。さらにサービス プロバイダーのような通信事業の場合、スケールメリットを基盤にしているので、そのエッジにはシングルポイント オブ フェイラーがより多く存在します。通常は、上位層とバックボーンのノード間のパスを冗長化することにより、単一ノードでの障害がサービスに影響を与えないようにします。したがって、上位層とバックボーンのノードには、シャーシ内で RP を冗長化させたり、ネットワーク復元力を装備したりはしません。代わりにそれらのノードは、代替パスへの高速なルーティング コンバージェンスによって可用性を向上させています。つまり、リンクまたはノードの障害をただちに検出し、トラフィックをすみやかに代替パスにルーティングします。Multiprotocol Label Switching（MPLS; マルチプロトコル ラベル スイッチング）Virtual Private Network（VPN; 仮想私設網）ネットワークは、トラフィック処理などの機能を組み込んでおり、コアでのリンクとノードの保護によって、迅速にルートを変更してパスの復元力を実現します。ルーティング プロトコルのコンバージェンスは、ネットワーク サービスの可用性に直接影響を与えますが、複

雑な問題なので、この文書では詳しく述べません。ルーティング プロトコル タイマーの操作と NSF/SSO に関する情報については、下記の URL にある高可用性に関する文書『**Cisco NSF and Timer Manipulation for Fast Convergence**』（英語）を参照してください。http://www.cisco.com/en/US/tech/tk869/tk769/technologies_white_paper09186a00801dce40.shtml

図 1　Cisco NSF/SSO の主な展開ポイント

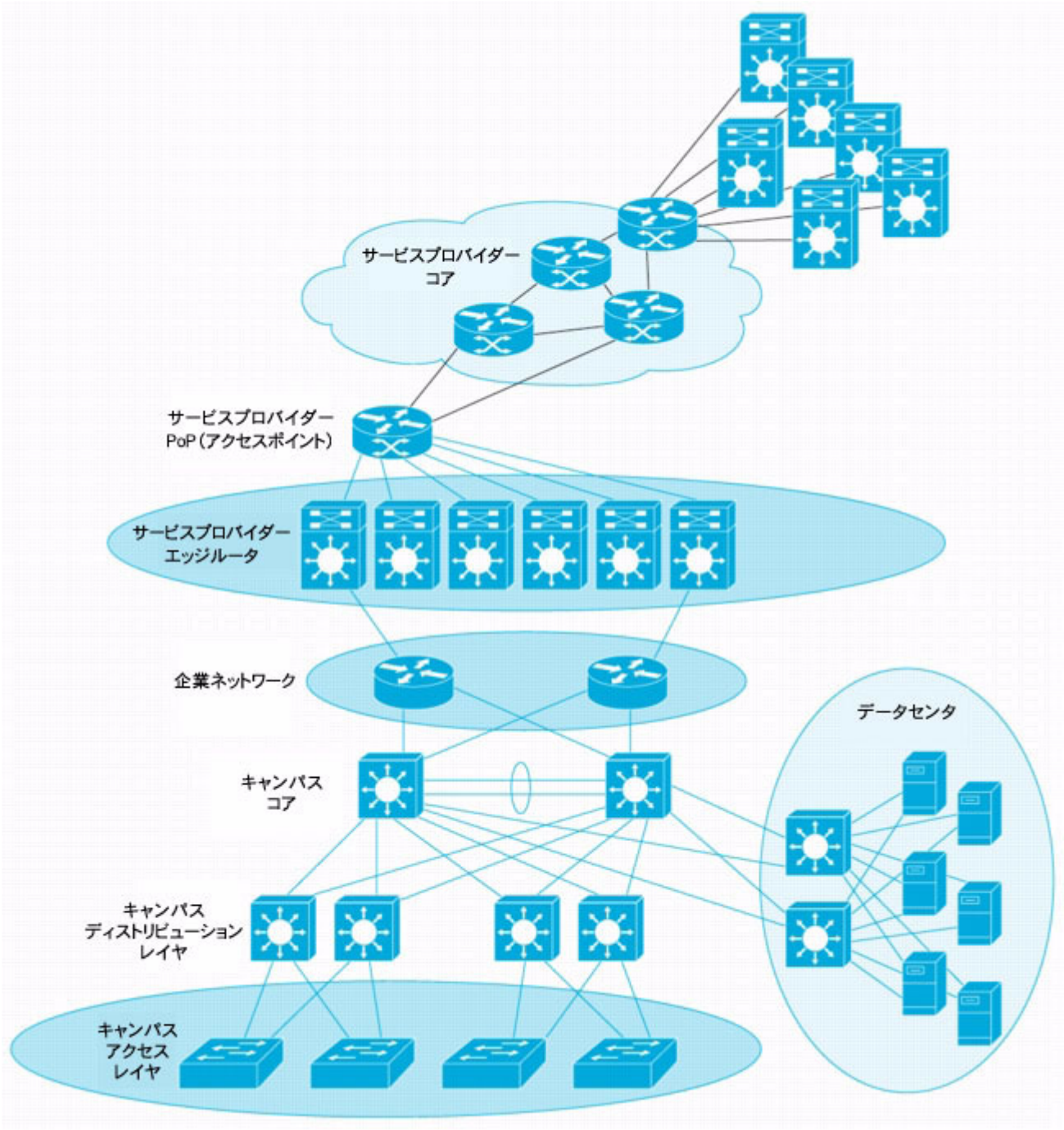


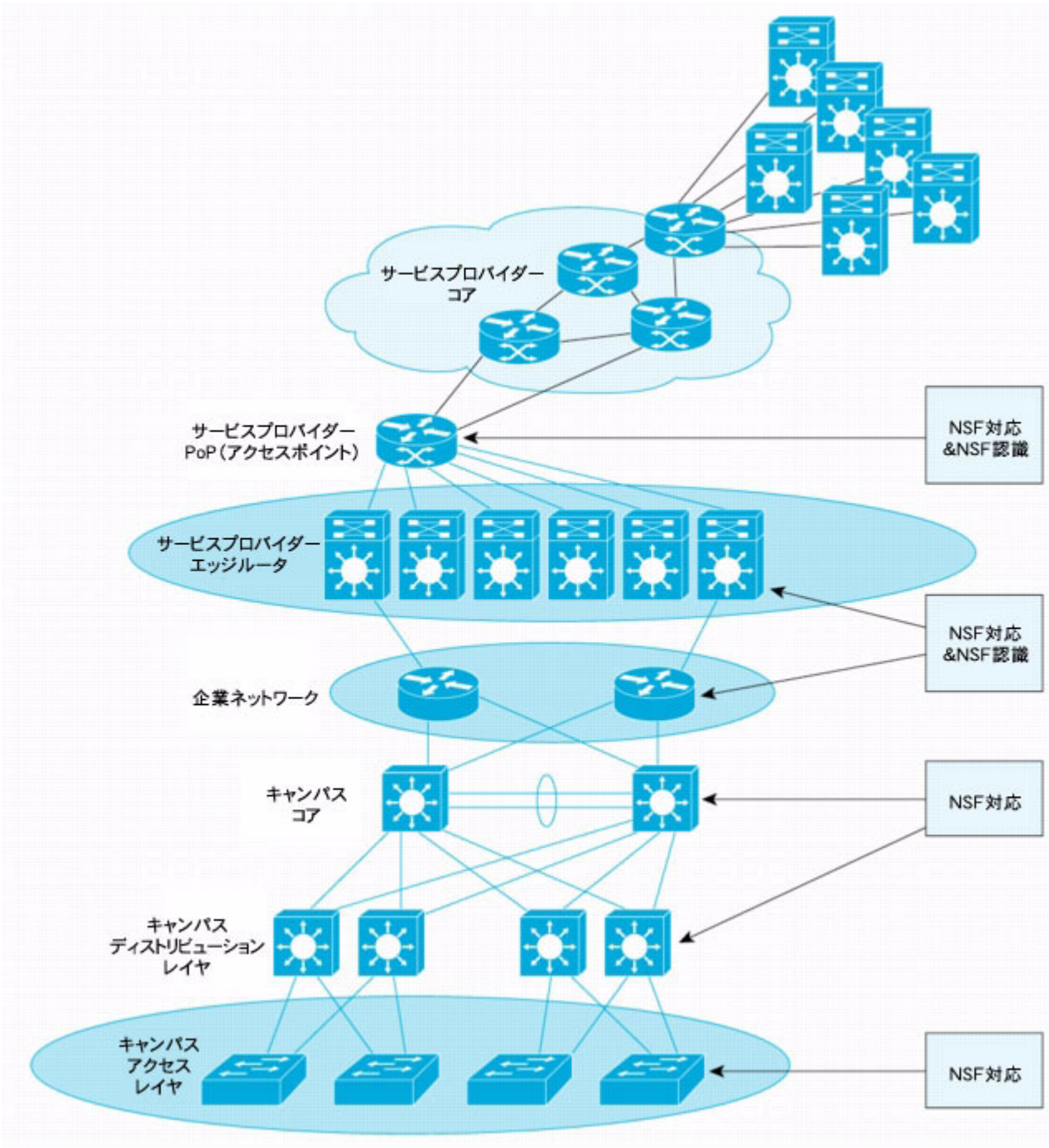
図 1 は、Cisco NSF/SSO を展開すべきポイントを示しています。色の付いた楕円で示しているように、NSF/SSO はサービス プロバイダー ネットワークのエッジに展開すると最も効果的です。NSF はサービス プロバイダーのエッジ ルータで、メンテナンスや何らかの障害などの理由により RP がオフラインになっても、お客様がその影響を受けないことを保証します。特に、お客様またはネットワークが、単一のエッジ ルータのみでサービス プロバイダーと相互接続している場合に、最大の効果を発揮します。こうしたケースでは、サービス プロバイダーのエッジ ルータがシングルポイント オブ フェイラーとなるので、この機能がないと、ノードに何らかの障害が発生した場合、そのパスを使用しているすべてのトラフィック フローが停止します。NSF を設定すると、シャーシ内の冗長 RP へスイッチオーバーする間もトラフィック フローは継続するため、サービスの向上と、ネットワーク寸断やルーティング プロトコル変動の抑制につながります。

多くのネットワークでは、トポロジー内の他の場所に設定しても効果が得られます。たとえば企業では、NSF/SSO をサービス プロバイダーとのエッジ境界に展開すると効果的です。これらのデバイスは、通常重要なネットワーク サービスを提供しており、再コンバージェンスによるパフォーマンス低下やパケットの損失は大きな問題となるからです。NSF/SSO は Cisco Catalyst 6500シリーズ スイッチで装備できるので、シングルポイント オブ フェイラーとなる接続が存在するデータ センタの重要なディストリビューション レイヤ デバイスや、キャンパス アクセス レイヤに展開できます。詳細については、「キャンパス ネットワークの NSF/SSO」を参照してください。

Cisco NSF 機能では、隣接ノードが 1 つの役割を果たします（図 2 を参照）。RP のスイッチオーバー中でもパケット転送を継続できる場合、そのノードは **NSF 対応**ノードです。NSF/SSO の展開から最大の効果を得るには、隣接またはルーティング プロトコル ピア ノードが **NSF を認識**できる必要があります。実装には必ず必要というわけではありませんが、再起動するノードでパケット転送を継続できることをルーティング ピアが認識し、またスイッチオーバー後にルーティング テーブルの整合性の復元と確認を支援しなければ、限られた効果しか得られません。これについては、ルーティング プロトコルごとに NSF 運用の詳細を説明する際に取り上げます。

シスコの NSF と SSO は、組み合わせて展開するように設計されています。NSF は SSO に基づいて、リンクとインターフェイスがスイッチオーバー中でも動作し、また下位レイヤのプロトコルの状態が維持されるようにします。ただし NSF を別個に設定したり、また SSO を NSF なしで有効にしたりすることは可能です。

図 2　NSF 認識デバイスと NSF 対応デバイスの連携



キャンパス ネットワークの NSF/SSO

キャンパス ネットワークは通常、高度な冗長性と広域な帯域幅を備えた設計となっています。キャンパス内では、どのリンクまたはコンポーネントで障害が発生しても、二重の等コスト パスと高速コンバージェンスによってトラフィックは代替パスをとることが可能です。ただし、接続の維持、パケット損失の低減、および特定のネットワーク サービスを提供するノードを経由するパス フローの一貫性を確保するうえで、NSF/SSO が効果を発揮する場所があります。

図 3　NSF 認識デバイスと NSF 対応デバイスの連携

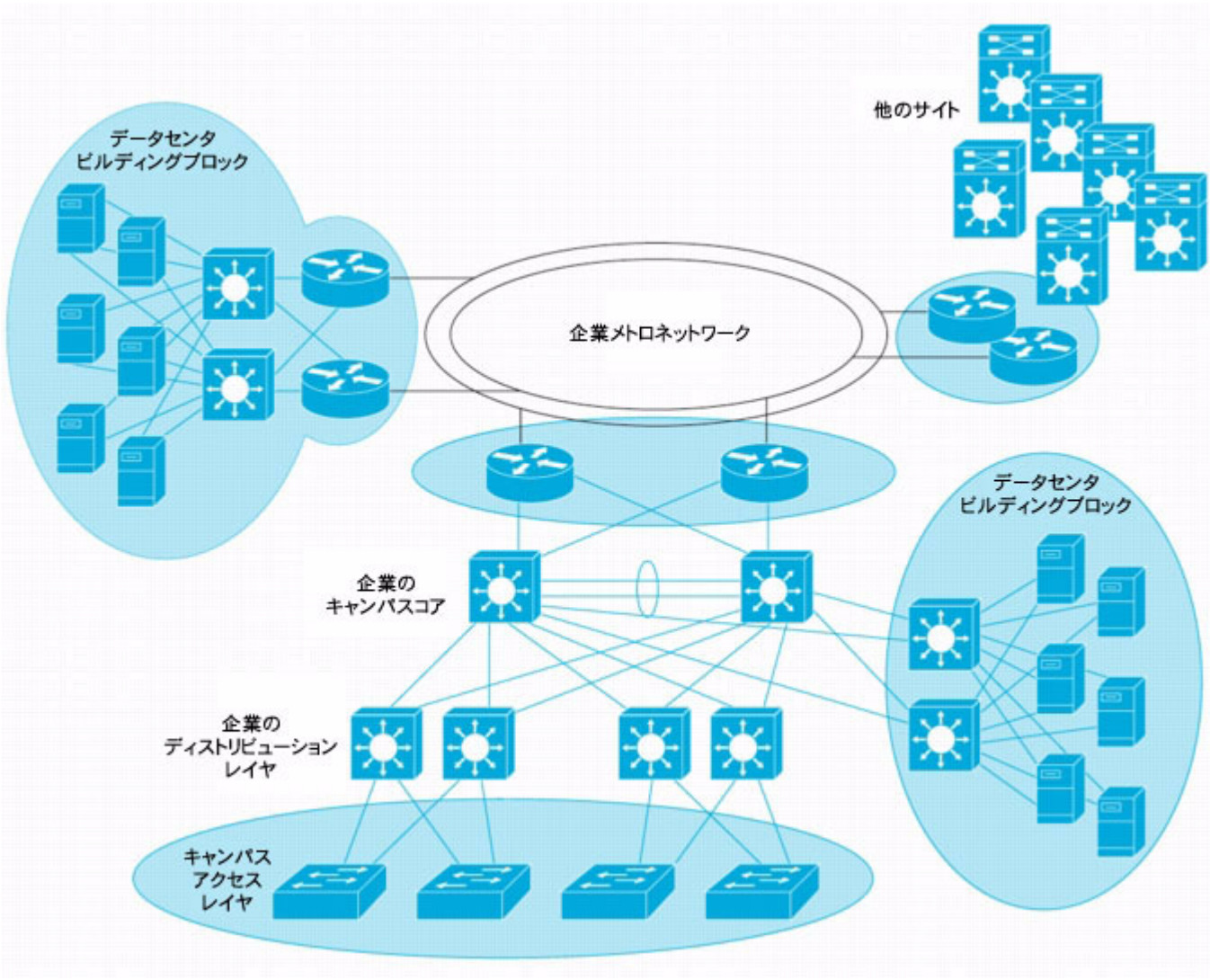


図 3 では、色付きの円で囲まれた部分が、NSF/SSO が最大の効果を発揮すると予想される場所です。

検討すべき最初の場所は、アクセス レイヤです。大規模企業では多くの場合、共通の機器とモジュールを使用することによって設計を簡略化し、運用の一貫性を確保することで、必要な予備品を最小限に抑えながら可用性を向上させます。Cisco Catalyst 6500 シリーズまたは 4500 シリーズ スイッチでは、エンド ステーションと IP テレフォニーにワイヤリング クローゼット接続が提供されるので、SSO はスーパーバイザによる障害、またはソフトウェア問題によるサービスの停止からの保護を実現します。アクセス レイヤは通常、レイヤ 2 サービスを提供し、冗長スイッチでディストリビューション レイヤが構成されます。レイヤ 2 アクセス レイヤでは、NSF なしで展開した SSO で効果が得られます。一部の企業は、アクセス レイヤでレイヤ 3 ルーティングを展開しています。その場合は、NSF/SSO を使用できます。

検討すべきもう 1 つの場所は、キャンパス メトロポリタン ネットワーク エッジです。多くの企業ではキャンパスは拡張され、そこでは複数の建物が相互接続されています。Metropolitan-Area Network（MAN; メトロポリタン エリア ネットワーク）は、2 台のルータまたはスイッチで各建物またはサイトを相互接続して構成される場合があります。メトロポリタン エリア サービス

はサービス プロバイダーによって提供され、ダーク ファイバ経由で相互接続されるか、企業所有のファイバ パスで構成されます。いずれの場合も、各サイトがメトロポリタン ネットワークに接続するキャンパス エッジの重要性は非常に高まるので、NSF/SSO が効果的です。

最後に、フロントエンドのデータ センタ、サーバ、コンピューティング クラスタ、およびメインフレームに使用される機器についても、NSF/SSO は効果的です。ここで特に有利なことは、ハードウェアおよびソフトウェア ベースの IP サービス機器またはブレード（ファイアウォール、コンテンツ管理システム、ロードバランシング システムなど）を通過するトラフィック パスが保全されることです。

図 4　データ センタの NSF/SSO

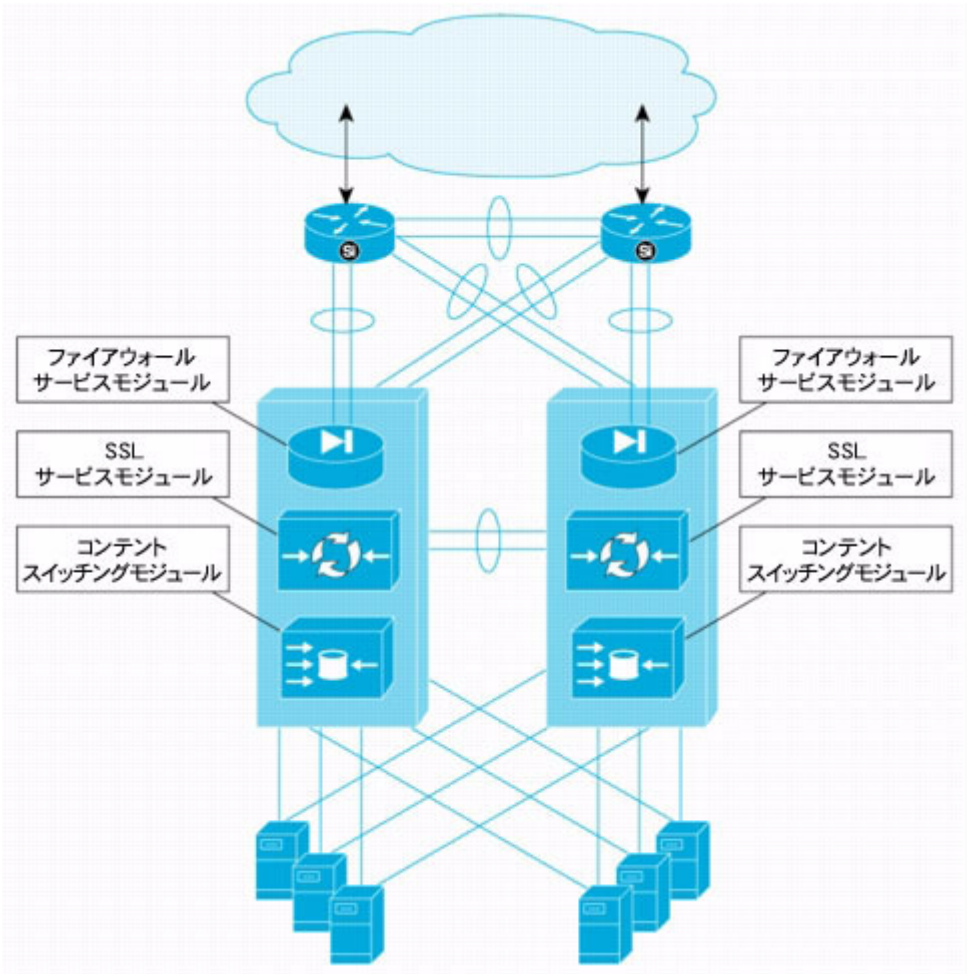


図 4 は、データ センタの設計を示しています。この図は、2 台の Cisco Catalyst 6500 シリーズ スイッチに統合型サービス モジュールを展開する例になっています。具体的には、ファイアウォール サービス モジュール、SSL サービス モジュール、およびコンテント スイッチング サービス モジュールを使用して、接続先サーバのアプリケーションに向かうトラフィックに重要なサービスを提供しています。

この環境では、片方の Cisco Catalyst 6500 シリーズ スイッチのスーパーバイザに障害が発生しても、トラフィックが同じパスで伝送され続けます。NSF/SSO のサポートによって、障害とネットワークの再コンバージェンスの影響は最小限に抑えられ、トラフィックの損失量が抑制されるとともに、Mean Time To Repair（MTTR; 平均復旧時間）も短縮します。サービス モジュール、シャーシ電源、またはシャーシ全体の障害に影響する重大な障害からの保護は、並列で動作する冗長スイッチで引き続き提供されます。

SSO の運用上の考慮事項

シスコの SSO は、他の 2 つの Cisco IOS® ソフトウェア インフラストラクチャ サブシステム、***Redundancy Facility*** と ***Checkpoint Facility*** に依存しています。PPP、High-Level Data Link Control（HDLC; ハイレベル データリンク制御）、フレーム リレーといった個々のプロトコルを制御するソフトウェアは、Checkpoint Facility と Redundancy Facility を使用して、リンクの状態とレイヤ 2 プロトコルの詳細がスタンバイ RP に複製されるようにします。これによって、RP のスイッチオーバー中にリンクの動作が維持されます。

以前の冗長モード（RPR など）では、こうした品質を提示できませんでした。RPR モードでは、スタンバイ RP は電源投入時に Cisco IOS ソフトウェア イメージをロードし、スタンバイ モードで初期化されます。スタートアップ コンフィギュレーションはスタンバイ RP に同期化されますが、その後の変更は同期化されません。スイッチオーバーが発生すると、スタンバイ RP はアクティブ RP として再初期化され、すべてのライン カードをリロードしてシステムを再起動します。ライン カードがすべてリロードされるため、隣接ルータではほとんどのタイプのポイントツーポイント接続で物理リンクの障害が検出されます。RPR+ モードでは、スタンバイ RP は完全に初期化され、設定されます。これによって、RPR+ ではアクティブ RP に障害が発生したり、手動でスイッチオーバーを実行したりした場合のスイッチオーバー時間を大幅に短縮できます。スタートアップ コンフィギュレーションとランニング コンフィギュレーションは、つねに両方ともアクティブ RP とスタンバイ RP で同期され、ライン カードがスイッチオーバー中にリセットされることはありません。インターフェイスは移行中も動作しているため、隣接ルータで物理リンクのフラップ（リンクの一時的なダウンとアップ）が検出されることもありません。ただし、ライン カード、プロトコル、およびアプリケーションの状態情報は同期化されないため、一部のレイヤ 2 プロトコルでは障害が発生します。冗長性モードを SSO に設定すると、ライン カード、プロトコル、およびアプリケーションの状態情報が同期化されて冗長 RP は「ホット」スタンバイとなり、いつでもただちに移行できるようになります。

現在、SSO を使用して同期化を実行するには、両方の RP で同じレベルのソフトウェア リリースが動作している必要があります。開発中の In Service Software Upgrade（ISSU）が利用できるようになると、この制約はなくなり、NSF/SSO を活用することによって、サービスに影響を与えずにソフトウェア アップグレードを実行できるようになります。

SSO の運用上の主な効果と利点は、RP がプライマリ RP からホット スタンバイ RP に切り替わるときに、隣接デバイスでリンク障害が検出されないことです。これは、RP のスイッチオーバーにのみ当てはまります。シャーシ全体が電源を失ったり障害を起こしたりした場合、またはライン カードに障害が発生した場合、リンクは障害を起こし、ピアでそのイベントが検出されます。もちろん、これはリンク障害を検出できるポイントツーポイントのギガビット イーサネット インターフェイスや Packet over SONET（POS）インターフェイスなどの場合です。NSF が有効になっていても、物理リンクの障害はピアによって検出可能で、NSF 認識は無効になります。

SSO のプロトコルサポート

SSO でサポートされているライン プロトコルとアプリケーションは、SSO を認識できます。Cisco IOS ソフトウェアの機能またはプロトコルが SSO を認識できるのは、Redundancy Facility と Checkpoint Facility が提供する機能を使用することにより、RP のスイッチオーバーを通じて部分的または完全に安定した動作が維持される場合です。SSO 認識プロトコルおよびアプリケーション（PPP、フレームリレー、Asynchronous Transfer Mode [ATM; 非同期転送モード]、SNMP [簡易ネットワーク管理プロトコル] など）の状態情報は、アクティブからスタンバイへと同期化され、それらのプロトコルとアプリケーションのステートフル スイッチオーバーが実現されます。

SSO を認識できないプロトコルとアプリケーションでダイナミックに生成された状態は、スイッチオーバーが発生すると消失するので、再初期化と再起動が必要です。これらのプロトコルとアプリケーションでは、状態情報が確立または再構築されるまでにある程度のパケット損失が発生する場合があります。

2004 年 10 月の時点で、SSO は PPP、Multilink PPP（MLPPP; マルチリンク ポイントツーポイント プロトコル）、HDLC、フレーム リレー、ATM、およびイーサネットをサポートしています。スイッチング製品には、表 1 に記載されている機能およびプロトコルのサポートも含まれます。

表 1　SSO のスイッチング機能サポート

リンク ネゴシエーション	VLAN Trunking Protocol（VTP; VLAN トランキング プロトコル）	Dynamic Trunking Protocol（DTP; ダイナミック トランキング プロトコル）
VLAN（仮想 LAN）	802.1Q	Port Aggregation Protocol（PAgP; ポート集約プロトコル）
VLAN トランク	レイヤ 2 プロトコル トンネリング	MAC 移動通知
Spanning Tree Protocol（STP; スパニング ツリー プロトコル）	802.1Q トンネリング	フロー制御とトラフィック ストーム制御
Address Resolution Protocol（ARP; アドレス解決プロトコル）	ブリッジ グループ	音声 VLAN とインライン パワー
Cisco Discovery Protocol（CDP）	ポート セキュリティ	802.1X
Switched Port Analyzer（SPAN; スイッチド ポート アナライザ）/Remote Switched Port Analyzer（RSPAN; リモート スイッチド ポート アナライザ）	Unidirectional Link Detection（UDLD; 単一方向リンク検出）プロトコル	Link Aggregation Control Protocol（802.3ad—LACP）
Internet Group Management Protocol（IGMP）スヌーピング		

製品ではコンフィギュレーションと状態情報も維持され、レイヤ 4 でのトランスペアレントなフェールオーバーが可能です。これには、Quality of Service（QoS; サービス品質）、セキュリティ機能、および Access Control List（ACL; アクセス制御リスト）の維持が含まれます。

特定のプロトコルごとの状態の同期化、制限事項、およびコンフィギュレーションの詳細については、該当するシスコのマニュアルを参照してください。

NSF の運用上の考慮事項

Cisco NSF はレイヤ 3 のルーティング冗長性機能と考えることができます。NSF では、フォワーディング プレーンからのコントロール プレーンの分離を活用します。コントロール プレーンはルーティング プロトコルのインテリジェンスであり、フォワーディング プレーンは利用可能な場合、ハードウェア アクセラレーションを使用してパケット交換を行います。NSF は Cisco Express Forwarding（CEF）と緊密に連携しています。Cisco 12000 および Cisco 7600 シリーズ ルータや Cisco Catalyst 6500 シリーズ スイッチなどの分散ルーティング ハードウェアは、CEF の情報を Forwarding Information Base（FIB; フォワーディング情報ベース）の形式でライン カードにダウンロードします。こうして RP のスイッチオーバー中でも、ライン カードは保持しているルーティング情報を使用してトラフィックの転送を継続できます。

また、NSF は Checkpoint Facility と Redundancy Facility を使用して、CEF の状態情報をスタンバイ RP に複製します。ホット スタンバイ RP に切り替わって「アクティブ」になると、NSF 対応および NSF 設定済みルーティング プロトコルによってネイバーとの隣接関係が再設定され、ルーティング情報が交換されます。ルーティング情報の交換後、Routing Information Base（RIB; ルーティング情報ベース）は FIB によって検証され、さらに必要に応じてアップデートされて、ルーティング情報の正確さとピアとの同期化が保証されます。

ルーティング プロトコルの隣接関係はプライマリ RP がダウンすると失われ、スタンバイ RP がアクティブになったあとで再確立されることに注意してください。さらにその後、ルーティング プロトコル情報はピアと交換されます。これを実行し、ピアまたは隣接ルータから、スイッチオーバー中のルータへのトラフィック転送を確実に継続させるために、ルーティング プロトコルの拡張機能が使用されます。

運用および展開の点からみると、上記の実現には隣接ルータでルーティング プロトコルの拡張機能がサポートされていることが必要です。ルーティング プロトコルの拡張機能によって、ネイバーはピアがパケット転送を継続できることをあらかじめ認識できます。一方、隣接関係が短時間途絶するために、ルーティング プロトコル情報の送信を要求する場合があることも、あらかじめ認識できます。スイッチオーバー中でも転送を継続できるルータは、**NSF 対応**ルータです。ルーティング プロトコルの拡張機能で、再起動するルータへのトラフィック転送を継続する機能をサポートするデバイスは、**NSF 認識**デバイスです。シスコ デバイスで NSF 対応のものは、NSF 認識デバイスでもあります。一部のソフトウェア バージョンとシスコ製品には、NSF 認識はサポートしていても NSF 対応ではないものがあります。

NSF のプロトコル サポート

前述のとおり、Cisco NSF では、スイッチオーバー中のコントロール プレーンが常時アクティブに維持されるわけではありません。代わりに、フォワーディング プレーンでは既知のルートが使用され、ルーティング プロトコル情報はスイッチオーバー後に復旧することになります。シスコのネットワーキング デバイスでは、パケット転送は CEF によって提供されます。CEF は FIB を保持しており、スイッチオーバーの時点で最新だった FIB 情報を使用して、スイッチオーバー中もパケット転送を継続します。パケット転送の継続が可能なので、スイッチオーバー中のダウンタイムはなくなります。

Cisco NSF では、Border Gateway Protocol（BGP）、Intermediate System-to-Intermediate System（IS-IS）、Open Shortest Path First（OSPF）、および EIGRP の各ルーティング プロトコルがサポートされています。Cisco NSF は、MPLS 関連のプロトコルもサポートしています（製品とリリースの入手方法については、該当する文書を参照してください）。スイッチオーバー中、各プロトコルでは CEF によってパケット転送が継続され、一方ルーティング プロトコルによって RIB が再構築されます。

Cisco NSF の現在の実装は、次の要件を想定して設計されています。

- シスコのお客様のニーズを満たすスケーラビリティを提供する。
- 多数のシスコ製品にわたって展開できる。
- 複数の障害シナリオにわたってネットワークの完全性を維持する。

シスコはネットワーク コミュニティおよび IETF と連携し、広く使用されているルーティング プロトコルに対する複数の機能拡張を推進することで、効果的なソリューションを生み出しています。プロトコルの拡張機能の基礎となる規格とドラフトについては、「関連する規格とドラフト」にまとめてあります。

次の項では、プロトコルの拡張機能と、サポートされている各ルーティング プロトコルの実装について説明します。

BGP NSF

シスコの BGP Nonstop Forwarding（NSF、別名グレースフル リスタート）のサポートは、IETF の規格案に記述されている実装仕様に準拠しています。この実装によると、パケット転送の継続を実現するには、以下の条件が満たされている必要があります。

- NSF 対応ルータとピア ルータは、それぞれが BGP グレースフル リスタートをサポートする。
- ピア ルータは NSF 対応ルータが使用できなくなっても、即座にそれを明示しない。

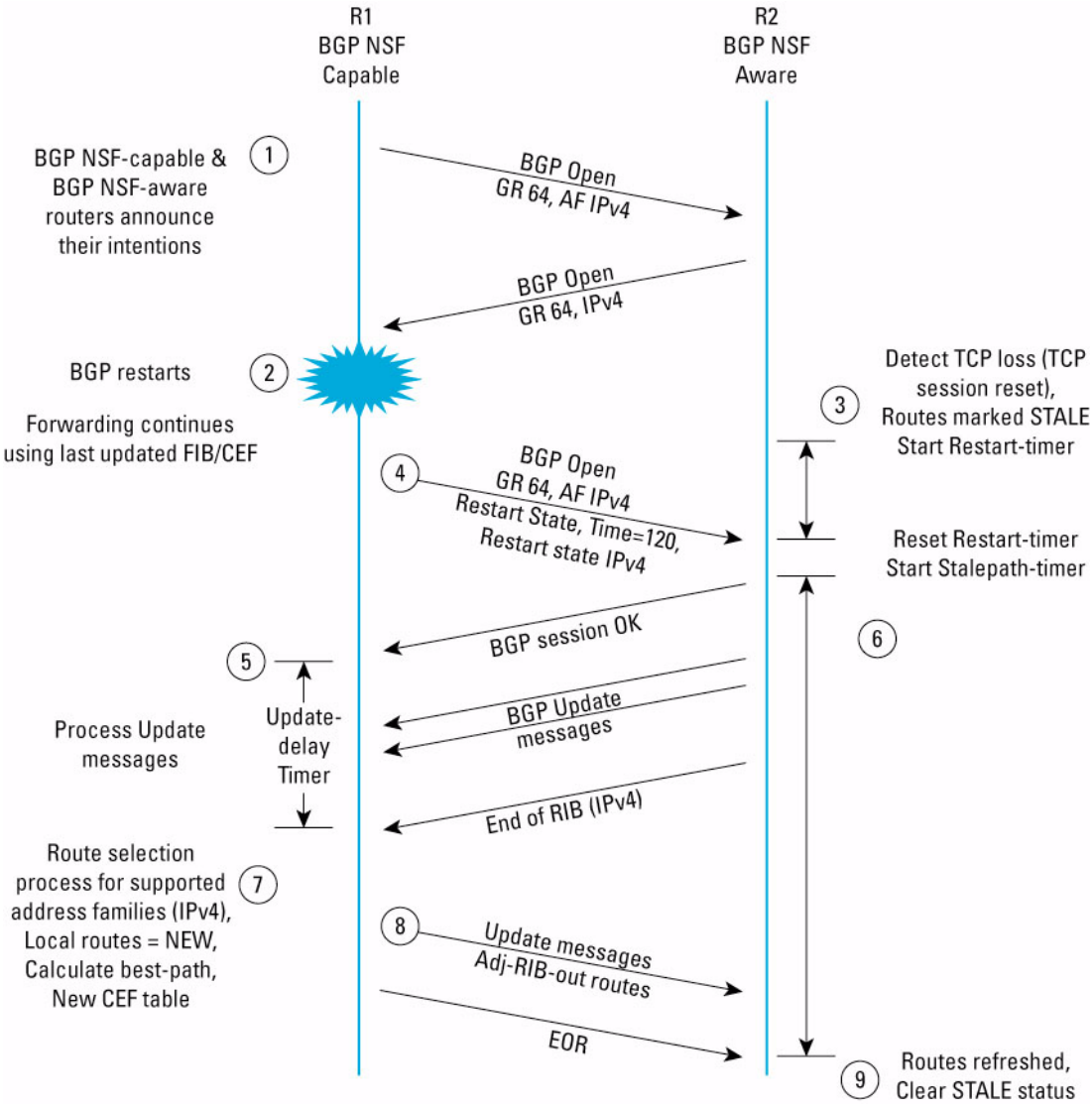
- ピア ルータは NSF 対応ルータの状態変化をいずれのピアにも伝達しない。これによって、当該ルータの突然の障害に伴うパフォーマンスへの悪影響が、ネットワーク全体に及ぶことを回避する。
- ピア ルータは BGP アップデートを送信して、再起動する NSF 対応ルータの BGP RIB 再取得を支援する。
- ピア ルータは、End of RIB マーカーの送信によって、最初のルーティング アップデートの完了を通知する。
- ピア ルータは一時的に（再起動する NSF 対応ルータがルーティング情報を再取得するまでの間）、再起動ルータに関連するあらゆるルートを「stale」とマーキングする。ただし、パケット転送のためにそれらのルートの使用は継続する。

最初の BGP 接続が確立されると、プロトコルの調整が始まります。NSF 対応ルータとそのピアは、セッションを確立する最初の BGP OPEN 中に、新しい BGP Capability（#64）を交換することによって、BGP グレースフル リスタート メカニズムに対応していることを示します。

ルータは、NSF 対応であるかどうかに関係なく、Capability 64 を送信することに注意してください。Capability 64 だけでは、再起動可能かどうかはわかりません。この信号が示すのは、当該ルータが IETF のドラフトで規定されている BGP の拡張機能を実装しているということです。そのため、BGP グレースフル リスタート対応で設定されている Cisco 7200 シリーズ ルータは、デュアル RP をサポートしていないために BGP を再起動できない場合でも、Capability 64 をピアにアドバタイズします。

さらに、NSF 対応ルータは、一連の Address Family Identifier（AFI）と Subsequent Address Family Identifier（SAFI）を提供し、それらについて BGP リスタートの前後でフォワーディング ステートを維持する機能を備えています。AFI と SAFI は、BGP で情報を伝送できる各種のプロトコルを示します。これには、IPv4、IPv6、MPLS、およびユニキャスト / マルチキャスト ルーティングなどのプロトコル サポートが含まれます。

図 5　BGP プロトコル拡張機能



BGP グレースフル リスタート プロトコル拡張機能の手順

この項では、RP のスイッチオーバー中に動作する BGP グレースフル リスタートの手順の一例を示します。図 5 は、BGP プロトコル拡張機能の手順を示しています。ここで、R1 は再起動ルータ、R2 はピア（受信ルータ）です。目的は、トラフィックのルートを変更して再起動ルータを迂回させる NSF 対応ルータのピアを使用せずに、BGP セッションを再開することです。

1. ルータ R1 の BGP プロセスが開始され、ルータ R2 とのピアリング関係が確立されます。ルータ R1 は「OPEN」メッセージを R2 に送信します。OPEN メッセージには、「グレースフル リスタート」の Capability（コード 64）、IPv4 のアドレス ファミリー、およびユニキャストの SAFI が含まれています。R2 は「グレースフル リスタート」をサポートしているので、独自の「OPEN」メッセージ（GR=64 と AF=IPv4 を含む）によって確認応答を送信します。

2. RP のスイッチオーバーが発生し、ルータ R1 の BGP プロセスが、新たにアクティブになった RP で再開します。R1 はこの RP 上に RIB を持っていないため、それをピア ルータから再取得する必要があります。R1 は、最後にアップデートされた FIB と CEF テーブルを使用して、ピア ルータ（R2）宛て（または経由）の IP パケット転送を継続します。

3. 受信ルータ（R2）は、再起動ルータとの間の TCP セッションが消失したことを検出すると、再起動ルータから学習したルートをただちに stale とマーキングします。R2 で stale とマーキングされるのは、R1 から学習したルートだけです。R2 にほかにもピアがある場合、そのピアから学習したルートは UP ステートのままです。またルータ R2 では、再起動ルータ用に「リスタート タイマー」が初期化されます。このタイマーのデフォルト設定は 120 秒です。リスタート タイマーの長さは、受信ルータが再起動ルータからの OPEN メッセージを待ち受ける時間です。受信ルータは、指定された「リスタート タイム」内に再起動ルータから OPEN メッセージを受信しなければ、stale ルートをすべて削除します。R2 で R1 の OPEN メッセージが受信されると、リスタート タイマーはリセットされます。この間に、ルータ R1 とルータ R2 は最後にアップデートされた CEF テーブルを使用して、トラフィック転送を継続します。

4. R1 の BGP プロセスが初期化されました。R1 は次に、R2 との BGP セッションの再確立を試みます。まず新しい TCP セッションを確立し、次に「OPEN」メッセージを送信します（リスタート ステート ビットの設定は、Restart Time = n、Forwarding State = IPv4）。リスタート タイムはデフォルトでは 120 秒で、設定可能です。R2 はこの「OPEN」メッセージを受信すると、自身のリスタート タイマーをリセットし、「ステイルパス タイマー」を開始します。「ステイルパス タイマー」のデフォルトは 360 秒で、これも設定可能です。

5. 両ルータがセッションの再確立に成功します。この時点で、R1 の「OPEN」メッセージ内の Forwarding State が IPv4 で設定されていないことを R2 が認識した場合（通常、Forwarding State は IPv4 で設定されます）、R2 はただちに再起動ルータから学習したすべての stale ルートを削除し、そのルーティング データベースを再計算します。

6. R2 は R1 に対して UPDATE メッセージの送信を開始します。このメッセージには IP プレフィクス情報が含まれており、R1 はそれに応じてメッセージを処理します。R1 はアップデート遅延タイマーを開始し、全ての NSF ピアから「End of RIB（EOR）」を受信するまで、最大 120 秒待ちます。R1 で BGP ルート選択プロセスが開始されるのは、すべてのピアから EOR 指示を受信したあと（または BGP アップデート遅延タイマーが切れたあと）です。ルート選択プロセスが完了すると、新しいルーティング情報データベースが利用可能になり、CEF の情報もそれに応じてアップデートされます。

7. R1 がすべてのピアから EOR を受信すると、BGP ルート選択プロセスが開始されます。

8. このプロセスが完了すると、R1 は R2 に対してプレフィクス情報付き UPDATE メッセージの送信を開始します。R1 は EOR 指示を R2 に送信することで、このプロセスを完了します。これにより、次は R2 でルート選択プロセスが開始されます。

9. R2 は EOR を待つ間、「ステイルパス タイム」の監視も行います。このタイマーが切れると、stale ルートはすべて削除され、「通常の」BGP プロセスが有効になります。R2 でルート選択プロセスが完了すると、BGP のすべての stale エントリが新しい情報でリフレッシュされるか、または BGP の RIB と FIB から削除されます。これでネットワークのコンバージェンスが完了します。

BGP NSF の展開の例

BGP ネットワークの設計と展開には、さまざまなバリエーションがあります。話を簡単にするため、ルータの機能に関して BGP 設計を検討します。ネットワーク トポロジー内の配置に応じて、特定のルータが達成すべきことは何でしょうか。BGP ネットワーク内のルータには、基本的に次の 3 つのタイプがあります。

- **AS 間ルータ**は、eBGP と iBGP を組み合わせて実行し、さまざまな Autonomous System（AS; 自律システム）を接続します。これには、エンタープライズ カスタマーをサービス プロバイダーのネットワークに接続するエッジ ルータ、サービス プロバイダーの AS 同士を接続するインターネット ピアリング ポイント、および BGP 連合のサブ AS の境界に存在するエッジ ルータなど、多くのバリエーションがあります（RFC 3065 を参照）。ただしこれらの各ルータの機能は、Cisco NSF の点では同等です。
- **AS 内ルータ**は、個々の AS のディストリビューション レイヤまたはコアに存在します。これらのルータは iBGP のみを実行し、同一 AS 内のルータとしかやり取りしません。AS 外部に関して保持する情報は、すべて AS 間ルータによって伝達されます。

- **ルート リフレクタ**は、BGP ルーティング情報の集約ポイントおよび配信ポイントとして機能します。AS 内ルータは、ルート リフレクタに BGP ルーティング情報をレポートし、ルート リフレクタから情報を受信します。ルート リフレクタによって、すべての iBGP ピアをフルメッシュにする制約がなくなるため、BGP ネットワークのスケーラビリティは向上します。ルート リフレクタの最も一般的な展開シナリオは、次の 2 つです。
  - **中央集中型**ルート リフレクタ : BGP ネットワークのコアに存在し、AS 内にある他のすべてのルータからほぼ等距離です。AS 内の各ルータは、このルート リフレクタとの BGP セッションを確立します。多くの場合、この構成では冗長ルート リフレクタが配置されます。
  - **分散型**ルート リフレクタ : AS 内のルータの一部は管理上グループ化され、ローカルのルート リフレクタが配置されます。各ルータはこのルート リフレクタに対して BGP セッションを確立します。これらのルート リフレクタは次に別の領域で、他のルート リフレクタと BGP セッションを確立するか、またはコアにある他のルート リフレクタおよび AS 内ルータとメッシュ接続を確立します。このタイプの構成の一般的な例は、サービス プロバイダーが各 POP にローカルのルート リフレクタを配置する場合です。

AS 間の例

図 6 は、複数の異なる AS にピアが配置されている eBGP の展開を示しています。この図には、可能な設計がいくつか示されています。ルータ R1 とルータ R2 は AS100 に属します。ピアリング ポイントには、RR1 と RR2 の 2 台のルート リフレクタが配置されています。可能な設計の 1 つは、接続先の AS（AS200）にある 2 つの異なるルータに対して、2 つのリンクと 2 つの eBGP セッションを使用するものです。別の設計では eBGP マルチホップを使用して、単一のルータに対し 2 つのリンクを使用します（図の AS300 との接続を参照）。さらにもう 1 つの可能性は、AS400 との接続のような単一の接続です。AS400 は、AS300 を経由する別のパスを備えていることに注意してください。

図からは、AS100 にピアリングしている一部のルータが、NSF 認識ではない場合があるということもわかります。前述のように、NSF/SSO が最大の効果を発揮するのは、ピア ルータが NSF 認識ルータである場合です。ただし理解を深めるために、ピアが NSF を認識しない場合のトラフィック フローの動作についても説明します。

注 : NSF 対応ルータは NSF 認識ルータでもあります。

検討のために、R2 がスイッチオーバーを実行するケースを取り上げます。

まず、AS100 と AS400 の間の動作を検討します。AS400 は AS100 と単一のルータ（R6）経由で接続されていますが、このルータは BGP NSF 認識ルータなので、スイッチオーバー中でも R2 へのトラフィック転送を継続します。さらに R6 は、R2 との接続損失をどのピアにも通知しません。また R2 の上流のルータも、R2 を経由して AS400 に向かうパケットの転送を継続します。NSF/SSO は完全に意図したとおり機能します。RP のスイッチオーバー中も転送は継続し、ルーティング プロトコルの混乱もまったく起こりません。

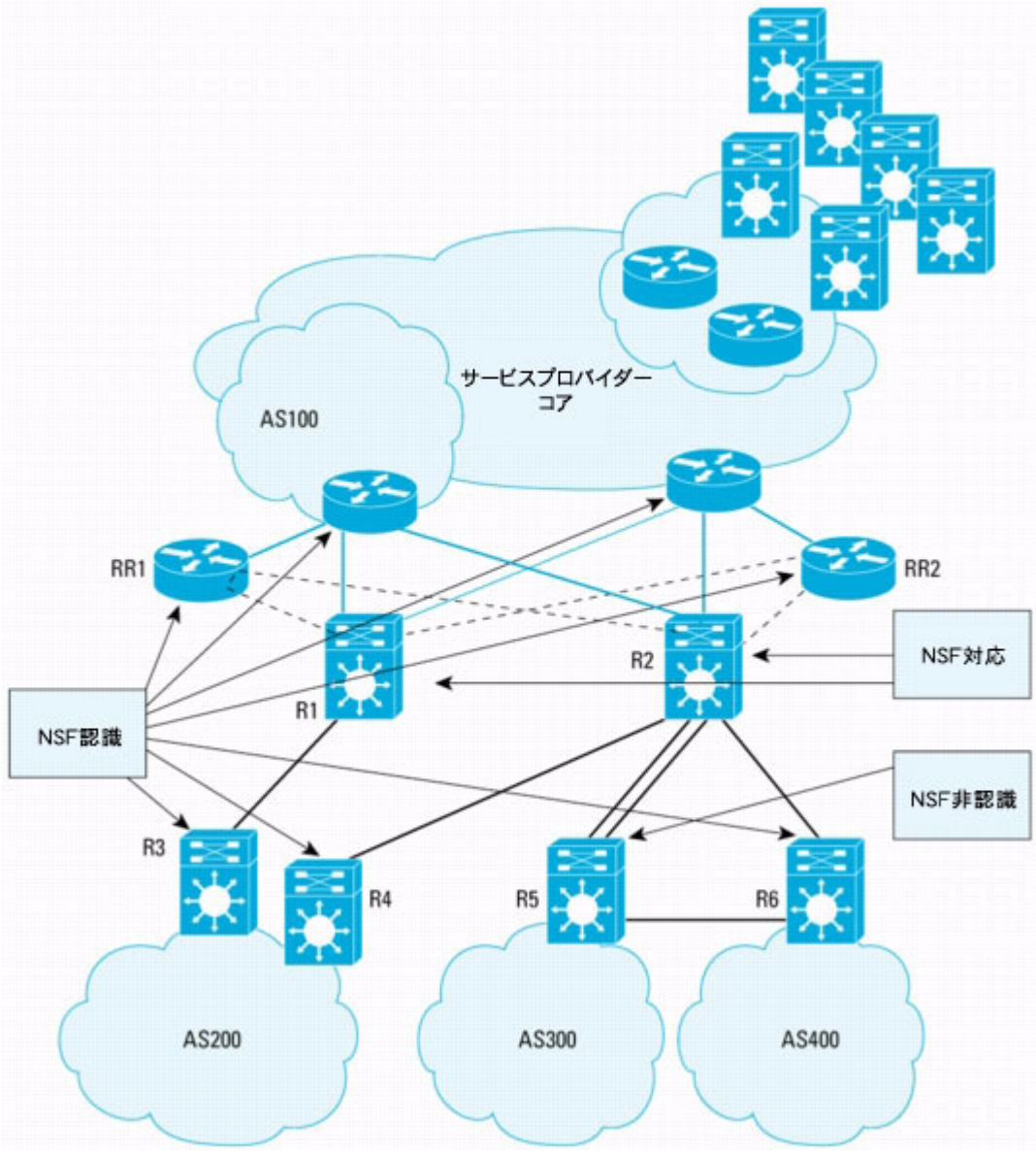
同じことは、AS100 と AS200 の間のトラフィック フローにも当てはまります。ここでは 2 つの異なるルータに対して、2 つの接続が管理ドメイン間で使用されています。R3 と R4 は、両方とも BGP NSF 認識ルータです。BGP で使用されている TCP セッションが R2 上の RP のスイッチオーバーで消失しても、同様に問題なく処理されます。スイッチオーバー中のトラフィックは、BGP が選択した最適パスで継続的に転送されます。

次に、AS300 に出入りするトラフィックを検討します。R5 は BGP NSF 認識ルータではありません。おそらく R5 では、BGP NSF 認識のサポートを提供した最初のバージョン（Cisco IOS ソフトウェア リリース 12.0(22)S）よりも前のソフトウェアが実行されています。R2 でスイッチオーバーが実行されると、TCP/BGP セッションの障害が R5 で検出されます。次に R5 は、トラフィックをルーティングして障害を迂回させようとします。その結果 AS100 宛てのトラフィックは、AS400 の R6 を経由するルートをとりますが、R2 は NSF/SSO 対応で設定されているため、R5 に対して AS300 に向かうトラフィックの転送を継続します。これは、NSF 認識ピアと NSF 非認識ピアが混在するときに発生する可能性のある、非対称ルーティングの例です。非対称ルーティングは望ましくない状態であり、ある程度のパケット損失が発生することもありますが、それでも R2 の再初期化に伴うネットワークの混乱よりは望ましい状態です。

R5 が R6 に接続されていない場合を考えます。R2 は以前に R5 から学習したルートをクリアしません。R2 は、最後にアップデートされた CEF テーブルを使用して、R5 への IP パケットの転送を継続する必要があります。R5 は NSF 非認識ルータなので、R2 との BGP セッションを喪失し、BGP セッションを最初から初期化します。R2 は AS300 宛てのパケット転送を R5 経由で継続しますが、そのトラフィックにはリターン パスがありません。R5 が R2 との間で再コンバージェンスを完了するまでに、パケット損失が発生します。

このルールには例外があります。R5 がネクスト ホップとして R2 を指定するデフォルトのスタティック ルートを持ち、また BGP のみを使用していたとすると、そのルートは R5 によってアドバタイズされ、R2 の BGP テーブルにエントリされます。この場合、R5 のそのルートは R2 で維持され、また R5 に必要なのはデフォルトのルートだけなので、パケット損失は発生しません。

図 6　BGP の AS 間展開の例



BGP と IGP の関係

このシナリオでは、展開上重要な考慮事項があります。このトポロジーでは、AS100 からネクスト ホップに到達できるようにするために、Interior Gateway Protocol（IGP; 内部ゲートウェイ プロトコル）（OSPF または IS-IS）を運用するのが非常に一般的です。BGP と選択された IGP プロトコルは相互に依存します。最適パスを計算する際、BGP は特定の宛先プレフィクスをアドバタイズするルータの IP アドレスを認識しています。ただし、そのアドバタイズしているルータに到達するネクスト ホップを決定するには、IGP からの情報が必要です。

BGP グレースフル リスタートでは BGP のコンバージェンスのタイミングを変更できるので、BGP で最適パスの選択が実行されるときに、IGP のコンバージェンスが未完了という状態になる可能性があります。したがって、アドバタイズしているルータへのパスが IGP によって計算されていないため、宛先プレフィクスが BGP にいくらか存在しても、CEF テーブルには追加できません。これによりパケット損失が発生することもあるので、BGP グレースフル リスタートに加えて、IS-IS または OSPF 用の NSF を設定することを強く推奨します。

ルート リフレクタとの相互作用

このトポロジーでは、R1、R2、RR1、RR2、およびコア対面ルータは NSF を認識します。ここでは、ルート リフレクタがコントロール プレーンとして展開され、iBGP のフルメッシュ構成の要件を軽減しています。そのためルート リフレクタは転送パス内にはありませんが、ルート リフレクタ クライアントとして、R1、R2、および他のルート リフレクタと iBGP のピアリング アレンジメントを構成しています。このトポロジーでは、IGP NSF（OSPF または IS-IS のいずれか）の実装が想定されています。BGP NSF とルート リフレクタに関しては、次のことを考慮する必要があります。

- R2 は BGP を再開する際、既存の CEF テーブルと FIB に基づいて、コア対面ルータ宛て（または経由）のパケット転送を継続する。
- その間、R2 が保持するピアリング アレンジメントはルート リフレクタとのアレンジメントだけである。コア対面ルータとの直接のピアリングはない。
- ルート リフレクタは NSF を認識できるので、R2 が BGP を再開したことは隠蔽される。ルート リフレクタは、コア対面ルータや他のルート リフレクタ ピアに対して、情報の伝達を停止する。ネットワーク内の他のルータは、R2 を経由してトラフィック転送を継続する。

ルート リフレクタが実際に NSF 対応で、BGP プロセスを再開する別のバリエーションを検討します。ルート リフレクタが BGP を再開すると、すべてのクライアントはルート リフレクタによって反映されたルーティング情報を維持します。バックアップ ルート リフレクタに切り替えるクライアントはありません。

NSF 対応ルート リフレクタを使用するときは、特別に考慮すべき点がいくつかあります。まず考慮することは、ルート リフレクタが持つ BGP ピアの数と BGP データの総量が、おそらく AS 内の他のルータよりも多くなるということです。このため、スイッチオーバー中の最適パス選択の完了にかかる時間が長くなる場合があります。第 2 に、ネットワーク設計者は、ネットワークでパケット転送の連続性とルーティングの安定性を確保するという要件と、コンバージェンスの完了までにルーティングが大幅に変化する可能性との間で、折り合いをつける必要があります。RP のスイッチオーバー中、Cisco NSF は BGP のルーティング情報ではなく、CEF テーブルを使用してパケットを転送することに注意してください。

Cisco NSF をルート リフレクタで使用する場合、もう 1 つ設定の調整が必要になることがあります。再コンバージェンスのプロセスが全体で 360 秒を超えると予想される場合は、**bgp graceful-restart stalepath-time 360** コマンドのデフォルト値をルート リフレクタのすべてのピアで調整する必要があります。ステイルパス タイムの値は、コンバージェンス時間の予想値（秒）に 30 ～ 60 秒のバッファ ゾーンを加算した値に等しくなるように調整します。これにより、コンバージェンス時間がネットワークの状態変化に基づいて変動しても対応できます。

BGP グレースフル リスタートをルート リフレクタで使用すべきかどうかの決定は複雑であり、ネットワークの運用に大きく左右されます。ネットワーク設計者は、この決定にあたって重要なファクタを比較検討する必要があります。次の疑問に答えることが必要です。

- ほかに可用性戦略はないか。バックアップ ルート リフレクタを使用した場合、フェールオーバー時間は許容可能か。
- 再起動するルート リフレクタで再コンバージェンスが実行され、そのピア ルータが新しい情報に基づいた転送を開始できるようになるまでの時間はどれくらいか。
- ルート リフレクタの再コンバージェンス中、ほかに BGP ルーティングで大幅な変更が発生する可能性はないか。

これらの疑問は、ルート リフレクタで Cisco NSF/SSO を使用するかどうかを決定する際に問題となりますが、Cisco NSF/SSO を展開する場所と方法を決定する際に役立つ一般的な疑問でもあります。

特定のネットワーク展開で、別のケースとトポロジーが可能な場合もあります。したがって、NSF/SSO を導入するときは、実際にネットワークでアクティブ化する前に、すべてのケースについてその効果を分析することが重要です。

BGP NSF の設定

設計と展開オプションが決まれば、設定は非常に簡単です。

BGP NSF（グレースフル リスタート）は、ルータの bgp グローバル コンフィギュレーション コマンドを使用して設定します。

```
Router(config-route)# [no] bgp graceful-restart
```

```
Router(config-route)# [no] bgp graceful-restart restart-time n
```

```
Router(config-route)# [no] bgp update-delay n
```

```
Router(config-route)# [no] bgp graceful-restart stalepath-time n
```

bgp graceful-restart コマンドは、Cisco NSF 対応ルータ、およびグレースフル リスタートに係わるすべての NSF 認識ピアで実行する必要があります。グレースフル リスタートはデフォルトでは無効なので、NSF 対応ルータとすべてのピア ルータで正しく設定することが必要です。

bgp graceful-restart restart-time n コマンドは、再起動ルータの障害が検出されたあと、ピアが TCP セッションの再接続と、新しい BGP OPEN メッセージを待ち受ける最大時間を指定します。TCP および BGP セッションが再確立されないうちにこのタイマーが切れた場合、BGP セッションは障害を起こしたとみなされ、通常の BGP 復旧手順が有効になります。再起動時間のデフォルト値は 120 秒です。

bgp update-delay n コマンドは、Cisco NSF 対応ルータで実行できます。このコマンドは、最初のピアが再接続されたあとのタイム インターバルを指定します。再起動ルータはこのインターバル中に、すべての BGP アップデートと END OF RECORD（EOR）マーカーを、設定されているすべてのピアから受け取ります。**n** のデフォルト値は 120 秒で、常に秒単位で指定します。再起動ルータが多数のピアを持ち、それぞれが多数のアップデートを送信するときは、この値をデフォルト値よりも大きくすることが必要な場合もあります。

bgp graceful-restart stalepath-time n コマンドは、再起動ルータの NSF 認識ピアで実行できます。このタイマーは、再起動ルータとの BGP セッションを再確立したあと、ピアが転送に stale ルートを使用できる最大時間を設定します。デフォルト値は 360 秒です。これはコンバージェンスを完了するには十分な時間ですが、大規模なネットワークでは、この値を大きくすることが必要な場合もあります。

OSPF NSF

BGP の場合と同様に、OSPF NSF の目的は、RP のスイッチオーバー発生時にグレースフル リスタートを実行することです。グレースフル リスタートは、ルーティングへの影響が最小限に抑えられ、パケット転送が中断しない方法で実行する必要があります。

OSPF はリンク ステート ルーティング プロトコルの 1 つで、同一ルーティング エリア内のすべてのルータが、一貫したルーティング トポロジーのビューを保持するようになります。たとえば、ルーティング トポロジーに変更があった場合は、Link State Advertisement（LSA; リンク ステート アドバタイズ）が OSPF エリア全体にフラッディングされます。これにより、エリア内のすべてのルータで Shortest Path First（SPF）計算が実行され、ルーティング テーブルが更新され、FIB テーブルの再読み込みが行われます。

再コンバージェンス中はネットワークが不安定になり、悪影響が生じることがあります。RP のスイッチオーバーは復旧処理であり、ルーティング トポロジーの変更ではありません。ルーティング トポロジーは以前の状態に復帰させる必要があるからです。再起動ルータが LSA フラッディングとネイバー隣接関係のフラップを引き起こさずにルーティング情報を再学習できれば、ルーティングの不安定化を回避できます。

OSPF ルーティング プロトコルでこの目標を達成するには、主に次の 2 つの課題に対処する必要があります。

- スイッチオーバー発生時にネイバー隣接関係を維持し、不要な LSA フラッディングを回避する。
- 新たにアクティブになった RP の Link State Database（LSDB; リンク ステート データベース）を、隣接ネイバーと再び同期させる。

ネイバー隣接関係の維持

OSPF がデュアル RP を備えた NSF ルータで有効になっている場合、ルーティング プロセスはアクティブ RP でのみ実行されます。スタンバイ RP には、OSPF 関連のルーティング情報、LSDB、およびネイバー データ構造は保持されていません。スイッチオーバーが発生すると、ネイバー関係は再確立する必要があります。

OSPF Hello プロトコルを使用することで、ネイバー関係の確立と維持、およびネイバー間の双方向通信の確認ができます。ルータが受信するネイバーの Hello パケット内にそのルータがリストされていれば、双方向通信が成立していることがわかります。

スイッチオーバーが発生すると、再起動ルータは Hello パケットを送信することによって、ネイバー隣接関係の再確立を試みます。新たにアクティブになった RP にはネイバーの状態情報が存在しないため、この Hello パケットのネイバー リストには、ネイバー情報がまったく含まれません。そのあとプロトコルに変更がなければ、この Hello パケットを受信するネイバーは双方向チェックに失敗し、再起動ルータとの既存のネイバー隣接関係をリセットします。隣接ルータは同時に、隣接関係の変更を反映するためにアップデート LSA をフラッディングするので、ルーティングの混乱が引き起こされます。

シスコは、プロトコル機能拡張を OSPF に導入することによって、この問題を解決しました。シスコの実装は、IETF の 3 つのドラフトで提案されている方式に準拠しています（「関連する規格とドラフト」を参照）。ネイバー隣接関係のフラップを防止するため、シスコの OSPF NSF の実装では、Hello プロトコルに新しいビットとして Restart Signal を導入しています。Hello パケットに Restart Signal ビットが設定されることにより、そのルータで RP のスイッチオーバーが実行されていることがわかります。ネイバーはこの Hello パケットを受信すると、OSPF NSF 手順に従って、双方向接続チェックを無視します。

Restart Signal ビットは、Hello パケットの Link Local Signaling（LLS）データ ブロックにある Extended Options Type Length Value（EO-TLV）に格納されています。LLS データ ブロックが Hello パケットに存在することは、IETF のドラフトで導入された L ビットによって示されます。L ビットは OSPF の Options フィールドに設定されます。ビットの値は 0x10 です。

NSF の実行中、Restart Signal ビットが設定された Hello パケットは 2 秒間隔で送信されます。これは、スイッチオーバー後のコンバージェンス時間を短縮するためです。この Restart Signal ビットが設定された 2 秒間隔の Hello パケットは、「Fast Hello」と呼ばれています。Restart Signal ビットは、ネイバー隣接関係が復旧するとクリアされます。

LSDB の再同期化

OSPF NSF では OSPF の状態情報がスタンバイ RP に保持されないため、新たにアクティブになった RP は、LSDB をネイバーと同期化する必要があります。

OSPF プロトコルは現在 RFC 2328 で定義されており、次の 2 つの方法で LSDB を同期化できます。

- ネイバー隣接関係の確立中に LSDB を初期化する
- ネイバー隣接関係の確立後、およびトポロジーの変更発生時に、フラッディング メカニズムを利用して LSDB を同期化する

これらの方法は、RP のスイッチオーバーの場合はいずれも実行不可能です。第 1 の方法が不可能なのは、LSA フラッディング回避のためには、RP のスイッチオーバー中もネイバー隣接関係を維持する必要があるからです。第 2 の同期方法は、変更のみが再同期化される差分式のため不十分です。この差分式の LSDB 同期化では、FIB 内のルートをすべて検証することができません。スイッチオーバー後はすべてのルートを検証し、トポロジー全体の完全性を維持することが不可欠です。

Cisco OSPF NSF では、Out of Band（OOB）LSDB 再同期化を使用することによって、この問題に対処しています。OOB 再同期化メカニズムは IETF のドラフトで定義されており、ネイバー隣接関係が確立されたあとに LSDB を完全に再同期化できます。

この OOB 再同期化機能を通知するため、新しいビットである LSDB Resynchronization（LR）ビットが定義されています。LR ビットは LLS データ ブロック内の EO-TLV に設定されます。このデータ ブロックは、すべての Hello パケットと Database Description（DBD）パケットに含まれています。

LR ビットに加えて、DBD パケットには新たに R ビットも導入されています。R ビットは、OOB 再同期化手順がアクティブになっていることを示すために使用されます。この R ビットは、DBD パケットの Options フィールドのフラグに設定されます。

LR ビットを導入すると、OSPF NSF ルータは、OSPF ネイバーが NSF 手順をサポートできるかどうかを識別できます。OSPF の動作中に LR ビットの設定された Hello パケットをネイバーから受信すると、そのネイバーが NSF 認識で、NSF 手順を実行できることがわかります。R ビットを導入すると、ルータは通常の LSDB 同期化または OOB 再同期化のどちらが実行されているかを識別できます。

OOB 再同期化メカニズムを使用した LSDB 同期化プロセスは、すべての隣接ネイバー間で実行されるわけではありません。これは RFC 2328 で定義されている従来の LSDB 同期化と同じ方法により、ルータ間で実行されます。たとえばブロードキャスト ネットワークでは、再起動ルータが Designated Router（DR; 指定ルータ）または Backup DR（BDR; バックアップ指定ルータ）でない場合は、指定ルータとの間でのみ OOB 再同期化が行われます。再起動ルータが NSF 認識ネイバーとポイントツーポイントで接続されている場合は、そのネイバーと OOB 再同期化が行われます。

注：NSF 非認識ルータが検出されると、OSPF NSF 対応ルータはセグメントでの NSF 処理を無効にします。デフォルトでは、他のセグメントで NSF 処理が継続されます。（OSPF）nsf [enforce global] CLI（コマンドライン インターフェイス）オプションが設定された場合、NSF 処理はすべてのセグメントで終了します。また、共通セグメント上の 2 台のルータが同時に NSF の実行を試みた場合、NSF 処理は両方のルータで終了します。

OSPF NSF プロトコル拡張機能の手順

図 7 は、R1 再起動の直後に、デュアル RP の NSF 対応ルータと NSF 認識ルータとの間で OSPF NSF が実行される手順を示しています。

1. 再起動ルータ（R1）は、FIB 内のルートを「stale」とマーキングします。また、NSF リスタート タイマーを開始します。このタイマーは、DR/BDR 選択と OOB 再同期化のトリガーとなります。

2. R1 は、RS ビットの設定された Fast Hello パケットをマルチキャストし、OSPF NSF 手順が開始されたことを通知します。LR ビットも設定されます。ネイバー情報はスイッチオーバー後まで維持されないため、この Hello パケットのネイバー リストは空白です。NSF 対応ネイバーと NSF 認識ネイバーでは、NSF プロセスのステータスに関係なく、Hello パケットに必ず LR ビットが設定されることに注意してください。

3. R2 は RS ビットの設定された Hello パケットを R1 から受信し、R1 で NSF 再起動手順が実行されていることを認識します。そのため双方向チェックは無視されます。一方 R1 では、ネイバーの Finite State Machine（FSM; 有限状態マシン）が Full ステートで維持されます。Resync-Timeout というタイマーがこの時点で開始されます。このタイマーによって、RS ビットの設定された Hello パケットが最初に受信されてから OOB 再同期化が開始されるまでの遅延が制限されます。

注 : OOB 再同期化タイマーは、dead-interval タイマーかデフォルトの 40 秒のいずれか大きい方の値に設定されます。たとえば、dead-interval タイマーが 40 秒未満の値に設定されている場合、OOB 再同期化タイマーは 40 秒のままです。逆に、dead-interval タイマーが（個々のネットワーク構成に固有の何らかの理由で）40 秒を超える値に引き上げられると、OOB 再同期化タイマーも同じ値に設定されます。これは自動的に行われるため、ルータに特別な設定は不要です。OOB 再同期化タイマーは、CLI コマンド **ip ospf resync-timeout seconds** によって明示的に設定できます。このコマンドは、必要に応じて再起動ルータの NSF 認識ピアで有効にできます。コマンドの有効化はインターフェイスごとに行います。詳細については、CSCdz80936 を参照してください。

4. R2はユニキャストHelloパケットをR1に返信します。R2は通常のHelloタイマーを待たずに、ただちにHelloパケットに応答します。注：R2 からの Hello パケットには、RS ビットは設定されません。

5. R1 は R2 から Fast Hello を受信すると、ネイバー隣接関係のステートを 2-way に移行させますが、NSF 側からはステートは Full とみなされます。

6. R1 は NSF リスタート タイマーが切れるまで待ちます。これは Hello インターバルの設定値（デフォルトは 20 秒）の 2 倍です。このタイマーが切れると、DR/BDR 選択と OOB LSDB 再同期化が開始されます。NSF 非認識ルータがセグメントに存在する可能性があるため、この「待ち時間」によって再起動ルータがすべてのネイバーの状態を学習できるようになっています。また、RS ビットはクリアされます。DR/BDR 選択のあと、R1 はネイバー隣接関係のステートを EXSTART に移行させます。

注 :（OSPF）nsf [enforce global] CLI オプションが設定されている場合は、Hello パケットが LR ビットなしでピアから受信されると、ただちに OSPF NSF が無効になり、DR/BDR 選択が進行します。

7. R1 は R2 に対して、R ビットの設定された DBD パケットの送信を開始します。

8. R2 は、R ビットの設定された DBD パケットを R1 から受信すると、ネイバーFSM を EXSTART に移行させ、LSDB 同期化を開始します。R2 は再同期化タイマーをキャンセルします。

9. R1 と R2 は、RFC 2328 に記述されている通常の LSDB 同期化と同じ方法で、LSDB 同期化を実行します。R1 は LSDB 同期化プロセス中に自動作成された LSA を受信しても、すぐにその LSA を削除することはせず、それを保存し、「stale」とマーキングします。

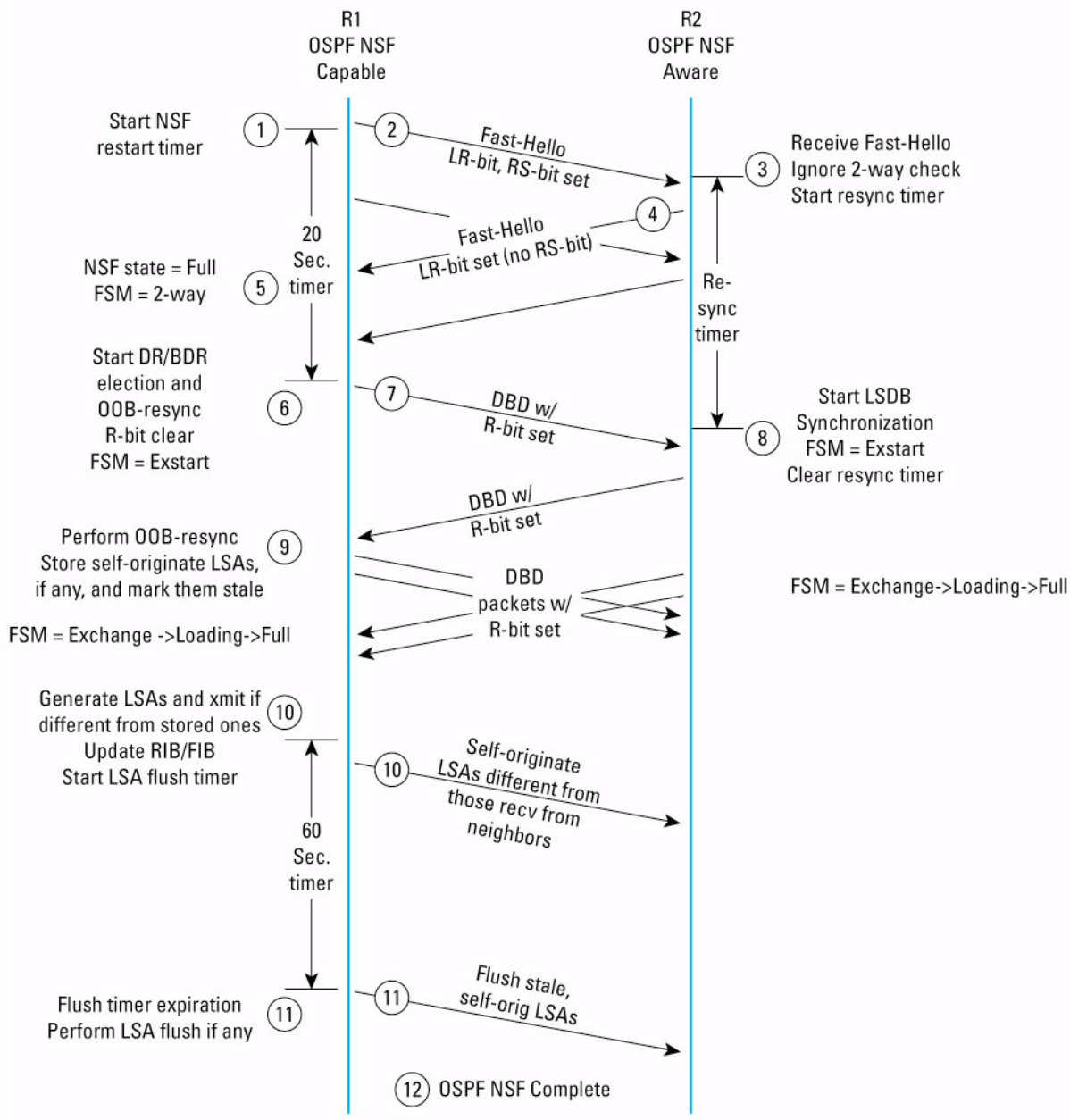
10. OOB 再同期化がこの段階で完了します。R1 でルータ LSA とネットワーク LSA の作成が開始されます。R1 がその LSA をネイバーに送信するのは、以前にネイバーから学習したものと異なる場合だけです。同じものであれば、R1 はその LSA の「stale」ステータスをクリアするだけです。またこの段階で、R1 は RIB と FIB のアップデートも開始します。

注 : ここでは内部的な RIB コンバージェンス信号の受信が想定されているため、R1 は LSA フラッシュ タイマーを開始します。RIB のコンバージェンスは、OSPF でコンバージェンスが行われる場合だけでなく、すべての NSF ルーティング プロトコルに基づいています（NSF 再起動を行うプロトコルが OSPF だけではない場合）。これによって、他のプロトコルが OSPF に再配信されることで自動作成される外部の LSA が、すぐに削除されることを防止します。

11. R1 で LSA フラッシュ タイマーが切れたことが検出されます（LSA フラッシュ タイマーのデフォルト値は 60 秒）。データベースに残っている「stale」フラグの設定された LSA がすべて削除されます。

12. OSPF NSF が完了します。

図 7　OSPF NSF の手順



OSPF NSF の展開

OSPF NSF は、次の場所に展開することを推奨します。

- シングルポイント オブ フェイラーとなるルータ。
- RP のスイッチオーバーが発生した場合に、ネットワーク不安定化の原因となるルータ。
- OSPF NSF 対応ルータのネイバー。OSPF NSF 手順には再起動ルータとそのネイバーの両方が関与するので、これらのネイバーは NSF を認識できる必要がある。これは必須ではないが、NSF/SSO で最大の効果を得るためには必要。

NSF 非認識ネイバーが存在しても NSF の利点を活用することはある程度可能なので、NSF は漸次的に展開できます。再起動ルータは、セグメント内のネイバーが NSF を認識できないことを検出すると、デフォルトでは、そのセグメントの NSF 手順を終了させるだけです。ほかのセグメントの NSF 手順は継続されます。

混在環境では、RP のスイッチオーバー中、および NSF 手順が完了するまでに、非対称ルーティングが発生することがあります。ルーティングは NSF 手順が完了すると対称になります。

以下は、NSF が動作しているときのトラフィック フローについての説明です。トラフィック フローを次の 3 つの段階で図示します。

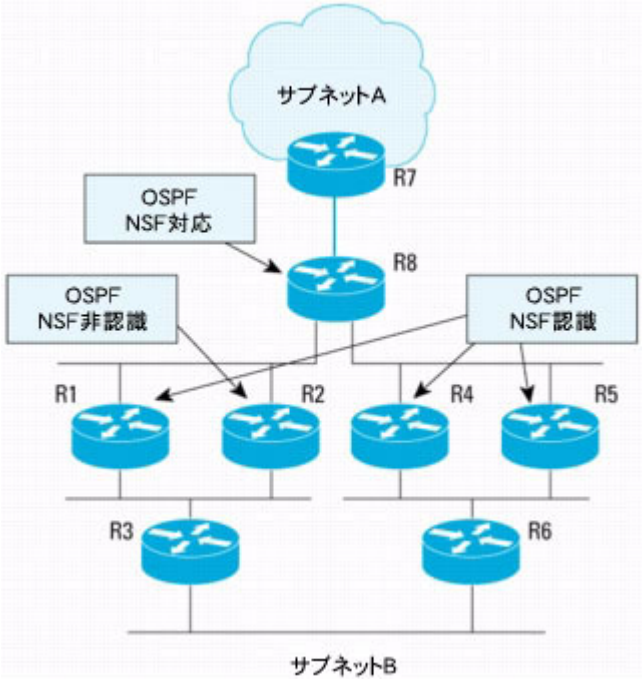
1. RP のスイッチオーバー開始前
2. RP のスイッチオーバー発生時、および NSF 実行中
3. NSF プロセスの完了後

理解しやすいように、OSPF NSF 再起動ルータとして 1 台のエッジ ルータに焦点を合わせ、そのネイバー ルータの 1 つが NSF 非認識ルータであるとします。

OSPF NSF 非認識ネイバーが存在するときのトラフィック フロー

図 8 では、R8 が NSF 再起動ルータです。そのネイバーのうち 4 台（R1、R4、R5、R7）は NSF 認識ルータです。R2 は NSF 非認識ルータです。この設計は、NSF 非認識ルータが存在するときのトラフィック フローの結果を示すという目的に合わせ、意図的に構成したものです。またリンクはすべて等コストで、「enforce global」コンフィギュレーションオプションは無効になっているものとします。「enforce global」オプションが設定されていると、NSF 非認識ネイバーが検出された場合に、ネットワークの全セグメントですべての OSPF NSF 手順が強制的に終了されます。

図 8　NSF 非認識ネイバーが存在する場合の OSPF NSF の例



サブネット A からサブネット B へのトラフィック フロー

- R8 の RP のスイッチオーバー開始前：

– R8 はトラフィック フローを R1、R2、R4、および R5 にロードバランシングします。次に、そのトラフィック フローは R3 と R6 へ送られ、さらにサブネット B へと送られます。これを図示したものが図 9 です。

- R8 の RP のスイッチオーバー発生時、および NSF 実行中：
  – R8 は R1、R2、R4、R5 に対してトラフィック転送を継続します（図 9）。

注：R2 は NSF 認識ルータではありませんが転送先に含まれます。これは、ネイバーが NSF を認識できるかどうかにかかわらず、再起動ルータのライン カード上の FIB がスイッチオーバー中は変化しないためです。

- R8 のスイッチオーバー完了後：
  – トラフィック フローはスイッチオーバー前と同様に、図 9 に示されているパスと同じパスで伝送されます。

図 9　スイッチオーバー前のサブネット A からサブネット B へのトラフィック フロー

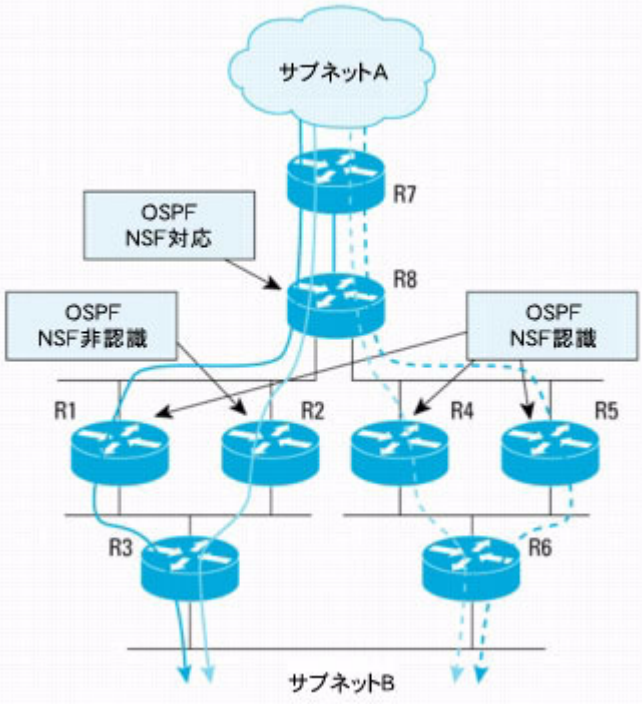
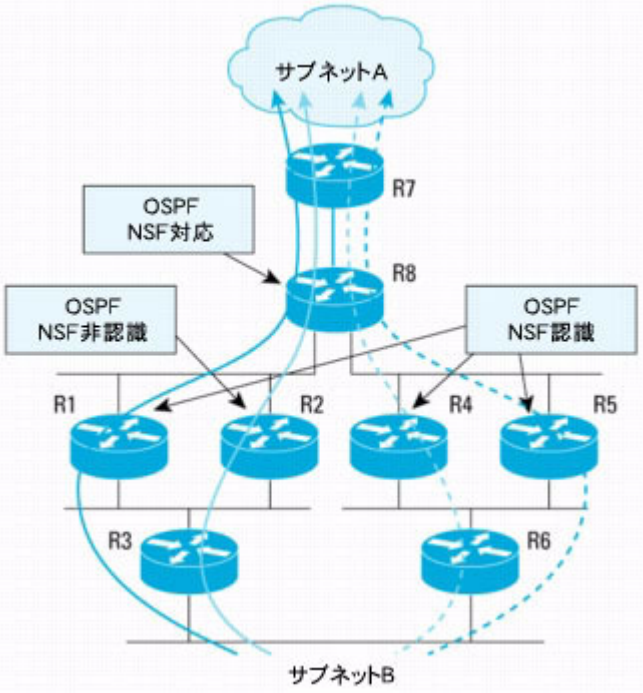


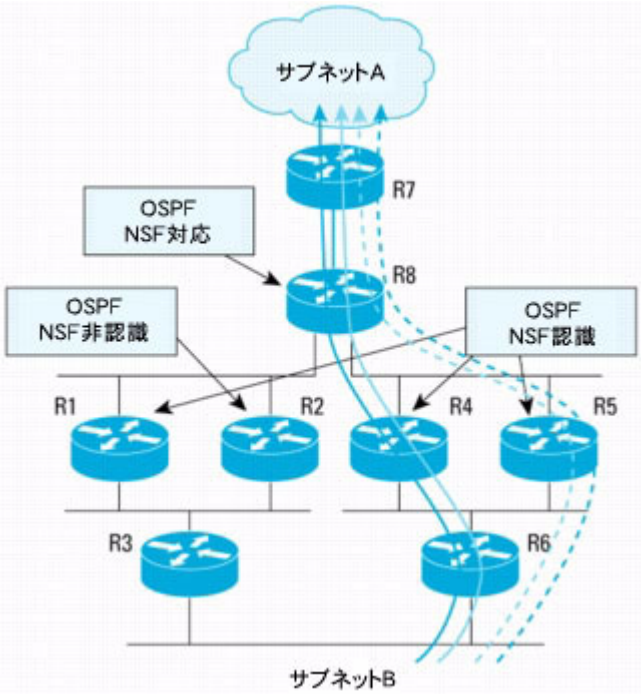
図 10　スイッチオーバー前のサブネット B からサブネット A へのトラフィック フロー



サブネット B からサブネット A へのトラフィック フロー

- R8 の RP のスイッチオーバー開始前：
  - サブネット B からのトラフィックは、R3 と R6 の両方へ送られるものとします。R3 は、サブネット A 宛てのトラフィックを、R1 と R2 にロードバランシングします。同様に、R6 はサブネット A 宛ての受信トラフィックを R4 と R5 にロードバランシングします。スイッチオーバー前のトラフィック フローは図 10 に示されています。
- R8 の RP のスイッチオーバー発生時、および NSF 実行中：
  - トラフィックは R6 にのみ伝送され、さらに R4 と R5 にロードバランシングされます（図 11）。
  - フローは R1 と R2 のパスには流れません。これは、R2 が NSF を認識しないためです。そのため R8 は、R1 を R2 に接続しているセグメントで NSF 手順を終了させます。
  - これによって、R8 と R1 および R2 の間の OSPF 隣接関係にフラップが発生します。R1 と R2 は、トポロジーをアップデートするために LSA を R3 にフラッディングします。その結果、R8 を経由するルートが R1、R2、および R3 で削除され、R6 に既知のパスだけが残ります。
  - ここで、サブネット B からのトラフィックは R6 へ向かうようになります。サブネット B からサブネット A に向かうトラフィックは、図 11 に示すように、ネットワークの右側だけを流れます。
- R8 のスイッチオーバー完了後：
  - トラフィック フローは、図 9 と図 10 に示された（スイッチオーバー前と同じ）もとのパスに戻ります。

図 11 R2 が NSF を認識しないために R1 と R2 を迂回するトラフィック

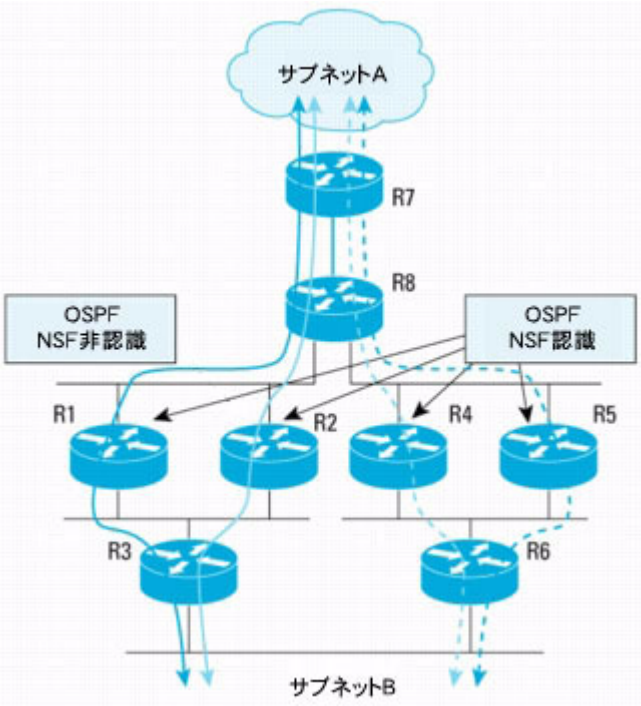


NSF プロセス中にトポロジー変更が発生した場合のトラフィック フロー

NSF 手順の進行中にトポロジー変更が発生することはほとんどありません。発生した場合も、NSF プロセスは継続します。この項では、NSF 手順の進行中にトポロジー変更が発生した場合のトラフィックへの影響について説明します。ここまでに図示したものと同じネットワークトポロジーを例として使用しますが、ここでは R2 は NSF 認識ルータとします。

図 12 は、R8 でスイッチオーバーが発生したために、NSF 手順が進行しているときのトラフィック フローを示しています。

図 12　スイッチオーバー前とスイッチオーバー中のトラフィック フロー



ここでは R6 のリンクが障害を起こしたために、トポロジー変更が発生するものとします。この場合は R6 で LSA が作成され、OSPF エリアにフラッディングされます。

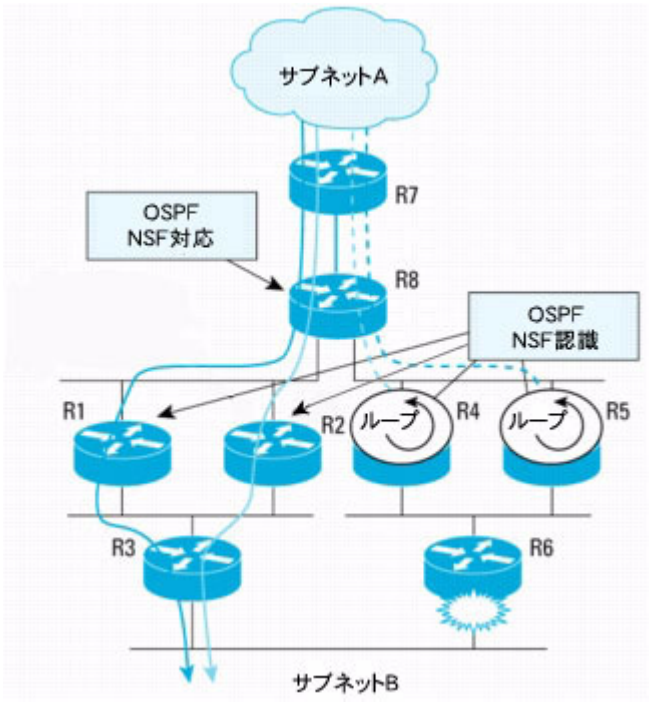
LSA を受信した R4 と R5 は、R6 を経由するサブネット B へのパスが存在しなくなったことを認識します。R4 と R5 はパスを再計算し、サブネット B には R8 経由で到達できると判断します。そのため、R8 がサブネット B に到達するためのネクストホップとして選定されます。これによって、一時的にルーティング ループが発生します（図 13）。

このルーティング ループの発生は短時間です。NSF 手順はまだ進行中であり、OOB 再同期化手順が R8 と R4、または R5 の間で完了すると、R8 は R6 を経由するサブネット B へのパスが利用できなくなったことを認識するので、ルーティング ループは解消します。

注 : ルーティング ループは、NSF プロセス中にトポロジー変更があると必ず発生するわけではありません。ネットワーク トポロジー、変更のタイプ（スタブ ネットワークのフラップではルーティング ループは発生しません）、および変更のタイミングに左右されます。

NSF 手順が有効なときにトポロジー変更が発生した場合のもう 1 つの例は、NSF を終了させることです。NSF プロセスをトポロジー変更中に終了させると、Cisco NSF のメリットは完全になくなります。NSF がトポロジー変更中に終了された場合、4 つのフローはいずれもサブネット B に到達しません。トラフィックの消失です。NSF/SSO が実装されていなくても、ネットワークの再コンバージェンスが完了するまでは、RP のスイッチオーバーによってトラフィック損失が発生するのが一般的です。あらゆる面を考えると、こうした結果を招くよりも NSF/SSO がもたらすメリットの方が大きくなります。

図 13 NSF 手順中にルーティング変更によって生じる一時的なループ



OSPF NSF の設定

OSPF で NSF の動作を設定するには、ルータの OSPF コンフィギュレーション モードで nsf コマンドを使用します。

```
router(config)# router ospf 100
router(config-router)# nsf
```

注：ルータを NSF 認識ルータに設定する必要はありません。ルータで、NSF 手順をサポートできる Cisco IOS ソフトウェア リリースが実行されていれば、設定なしで NSF を認識します。

OSPF NSF 非認識ルータが検出された場合に、オプションでルータ全体の OSPF NSF プロセスを終了させるには、「enforce global」キーワードを設定します。

```
router(config)# router ospf 1
router(config-router)# nsf enforce global
```

IS-IS NSF

IS-IS NSF の目的は、RP のスイッチオーバーが発生したときにグレースフル リスタートを実行することです。グレースフル リスタートは、ルーティングへの影響が最小限に抑えられ、パケット転送が中断しない方法で実行する必要があります。

IS-IS は OSPF と同様、リンク ステート ルーティング プロトコルの 1 つです。したがって、同じルーティング エリア内のルータは、すべて一貫したルーティング トポロジーのビューを保持する必要があります。たとえば、ルーティング トポロジーに変更があった場合は、Link State Protocol（LSP）データ ユニットが IS-IS エリア全体にフラッディングします。その結果、エリア内のすべてのルータで SPF アルゴリズムが実行され、RIB が更新され、FIB の再読み込みが行われます。

ネットワークが再コンバージェンス中に不安定になり、パケットの配信に悪影響を与えることもあります。RP のスイッチオーバーは復旧処理と捉えることはできますが、ルーティング トポロジーの変更とは捉えられません。ルーティング トポロジーはスイッチオーバー後、以前の状態に復帰します。再起動ルータが、LSP フラッディングとネイバー隣接関係のフラップを引き起こさずにルーティング情報を再学習または維持できれば、ルーティングの不安定化は回避できます。

OSPF の場合と同様、IS-IS ルーティング プロトコルでこの目標を達成するには、主に次の 2 つの課題に対処する必要があります。

- スイッチオーバー発生時にネイバー隣接関係を維持し、不要な LSP フラッディングを回避する。
- 新たにアクティブになった RP の LSDB を、隣接ネイバーと再び同期させる。

この問題に対処するには、2 つのソリューションがあります。1 つはシスコ固有のステートフル ルーティング ソリューションで、もう 1 つは OSPF と BGP で使用される前述の方法によく似ています。Cisco IOS ソフトウェア固有のソリューションでは、チェックポイント機能を使用して、スタンバイ RP の IS-IS 隣接関係とデータベースの状態をバックアップします。2 つめのソリューションは IETF の成果に基づき、IS-IS Hello PDU 内の新しい TLV を使用します。したがって 2 つめの方法では、支援ネイバーが機能する必要があります。

シスコのステートフル ソリューション

シスコのステートフル ルーティング ソリューションを使用すると、隣接関係と LSP 情報がすべてスタンバイ RP に保存（チェックポイント化）されます。スイッチオーバー後、新たにアクティブになった RP はチェックポイント化されたデータを使用して隣接関係を維持するので、ルーティング テーブルを迅速に再構築できます。

このシスコ固有のソリューションは、前述の 2 つの問題（隣接関係の再取得と LSDB の再同期化）に革新的かつ独特な方法で対処します。

隣接関係の維持

IS-IS プロトコルでは、隣接関係は Hello メッセージの定期的な伝送によって維持されます。Intermediate System（IS）で隣接関係のホールディング タイム中に Hello を受信できないと、その隣接関係は廃棄されます。適切な状態情報が Hello に含まれていない場合（Hello に、受信 IS の [ポイントツーポイント リンクの] システム ID または [LAN セグメントの]MAC アドレスがリストされていない場合）も、隣接関係は再初期化されます。したがって NSF メカニズムでは、タイムアウトのためにネイバーで隣接関係が廃棄されないように、すばやく再起動を行う必要があります。さらに、NSF プロセスでは適切な情報が Hello に含まれ、ネイバーで再起動が認識されることを防止できるように、状態を維持する必要があります。

シスコのソリューションは、適切な状態情報をチェックポイント化し、再起動後にそれを使用することでこれらの課題を克服し、ネイバーで隣接関係が廃棄されることを防止します。メカニズムは、ポイントツーポイントと LAN の両方の隣接関係に対応して設計されています。

LSP データベースの同期化

データベースの同期化は、再初期化プロセスのもう 1 つの部分です。IS-IS プロトコルでは、LSDB を隣接ルータと同期化するメカニズムを利用できます。通常の状況ではリブートによって、隣接関係の再初期化とそれに続く LSP データベースの同期化が引き起こされます。隣接関係の再初期化はシスコの IS-IS ステートフル ソリューションによって抑制されるので、トポロジー変更を起こさずにルータの LSP データベースを同期化するために、特定のメカニズムが使用されます。

この場合もメカニズムは、ポイントツーポイントと LAN の両方のインターフェイスに対応するように開発されています。

IETF ソリューション

IETF ソリューションが定義するメカニズムでは、ルータが再起動してネイバーでダウン ステートを繰り返さずに隣接関係を再確立できることを、再起動ルータからネイバーに通知します。一方、データベース同期化が正しく開始されることも通知します。前述のシスコのステートフル IS-IS ルーティング ソリューションと異なり、IETF ソリューションはステートレスです。このソリューションでは、LSP データベースの内容と隣接関係の情報はチェックポイント化されません。

IETF ソリューションでは、再起動ルータの再起動を隠蔽しません。再起動ルータは再起動したことを明示し、LSP データベースの内容をネイバーから取得できるようにします。ネイバーはこれを認識し、再起動ルータと連携します。

隣接関係の再取得

隣接関係の再取得は、再初期化の最初のステップです。再起動ルータは、隣接関係が再取得されるので、ネイバーでは隣接関係再初期化の必要がないことをネイバーに対して明示的に通知します。これは、新しい「再起動」オプション（TLV）を Hello PDU に含めることで実現されます。この TLV の存在によって送信者が新しい再起動機能をサポートしていることがわかります。またこの TLV には、再起動中の情報伝達に使用するフラグが付いています。この機能をサポートしているルータから送信される Hello メッセージには、この TLV が含まれます。この TLV には、2 つのフラグが含まれています。「Restart Request」を示す RR と「Restart Acknowledgement」を伝達する RA、そして「Remaining Time」で、これは許容できる復旧時間を再起動ルータに通知します。

再起動ルータの隣接ルータは、「再起動」TLV に RR ビットが設定された Hello メッセージを受信すると、再起動ルータとの隣接関係を「Up」ステートのまま維持し、この再起動への確認応答として RA ビットを設定した Hello メッセージを送信します。

複数のレベル

ルータは、特定のインターフェイスでレベル 1 とレベル 2 の両方として動作している場合、上記の動作をレベルごとに実行します。

- LAN インターフェイス：ルータはレベル 1 とレベル 2 の両方の Hello を送受信し、Complete Sequence Number PDU（CSNP）同期化をレベルごとに実行します。
- ポイントツーポイント インターフェイス：Hello メッセージは（両レベルのサポートを示すものが）1 つだけ必要です。ただし CSNP 同期化はレベルごとに実行されます。

LSP データベースの同期化

ルータは再起動すると、ネイバーごとに保持されている LSP 状態を反映した CSNP をそれぞれのインターフェイスで受信できます。CSNP が正しく受信されるまでは、RR ビットの設定された「再起動」Hello が再送信されて、CSNP の着信を保証します。この LSP がすべて受信されると、同期化が完了します。

LSP の作成とフラッディング

隣接関係がすべて再初期化されると、ルータでは利用可能な隣接関係情報がすべて再取得されたものとみなされ、IS-IS で独自の LSP を作成できるようになります。ローカライズされた再起動を実現するためには、再起動前のルータの状態を反映する十分な情報が取得されるまで、このルータの LSP を作成および伝達しないことが重要です。また、このローカル LSP の再作成フェーズより前に、ローカル ルータの LSP の古くなったコピーが受信されることもあります。

通常の場合、作成を終えているルータから受信された LSP のコピーは削除する必要があります。ただし再起動ルータの場合は、作成される必要のない新しい LSP が受信されることもあります（レベル 1 の SPF が実行されて、レベル 2 に伝達の必要のあるプレフィクスが検出された場合）。「余計な」LSP を削除すると、他のすべてのルータに影響が及び、その FIB は混乱します。NSF 認識ルータでは、プロトコルと IS-IS レベル間の再分配がすべて実行されるまで、この「余計な」LSP は無視されます。「余計な」LSP は、同期化ポイントに達してから削除されます。

同様に、レベル間情報の再分配は、このルータの LSP が他のノードにフラッディングする前に再開されます。レベル 1 またはレベル 2 の LSP の送信は、他のレベルの SPF が実行され、伝播する必要のあるレベル間情報がこの LSP に含まれていることが確認できるまで延期されます。

注：SPF 計算の「最初の反復」中に情報が RIB に追加されなくても、FIB にエントリが保持されているため、これらの宛先へのトラフィックは廃棄されません。IS-IS 以外のルーティング プロトコル情報の再分配は、最後の NSF IS-IS LSP の作成前に、RIB でアップデートされる適切なルーティング情報に依存することがあります。

SPF 計算

LSP データベースが再同期化されると、リンク ステート データベースは最新になります。SPF 計算が実行され、再初期化されたすべての情報が RIB と FIB に伝播されます。このプロセスでリフレッシュされなかったルートはすべて古くなっているので、ブラック ホールとルーティング ループを抑制するためのホールド タイム経過後に削除されます。

IS-IS NSF プロトコル拡張機能の手順

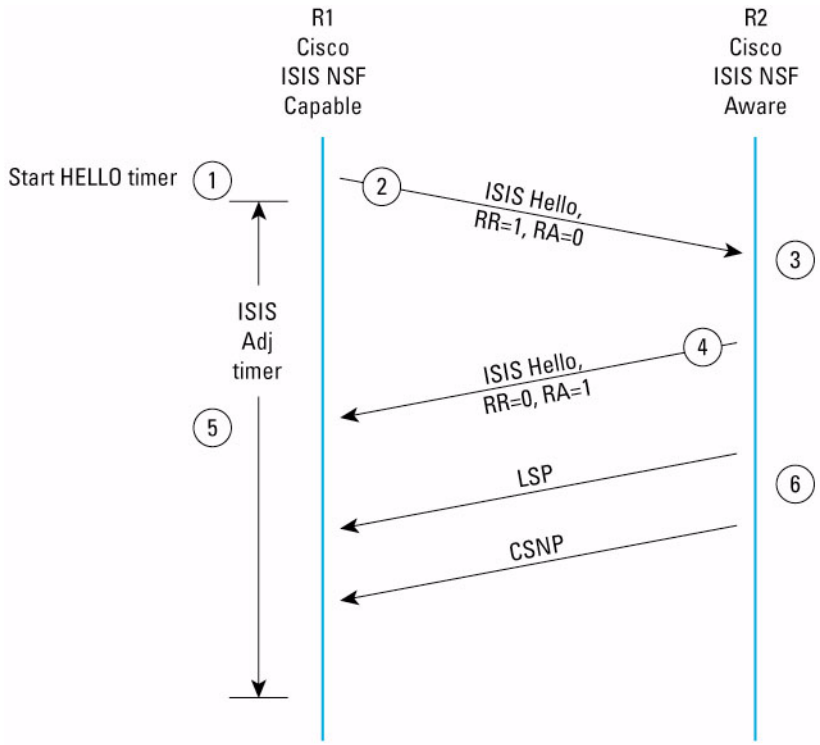
この項では、IETF の実装に対応した IS-IS NSF について説明します。

シスコの IS-IS NSF の IETF 実装

次のシーケンスは、IETF の実装に準拠して IS-IS 手順を記述したものです。図 14 を参照してください。

1. R1 が再起動します。
2. R1 は Hello メッセージを送信します。このメッセージには、RR ビットが設定され、RA ビットがクリアされた TLV 211 が含まれており、R1 が再起動したことが示されます。
3. R2 は R1 の Hello メッセージを受信します。
4. R2 は IS-IS NSF を認識できるので、RR ビットがクリアされ、RA ビットが設定された TLV 211 を含む Hello メッセージで応答します。これによって、前に R1 から受信した Hello を R2 が確認したことが示されます。
5. R1 は R2 から Hello メッセージを受信します。
6. インターフェイスがポイントツーポイント インターフェイスである場合、または R2 が（R1 を除いて）、IS-IS Hello（IIH）に再起動 TLV を含むルータの中で（送信元 MAC アドレスに基づくプライオリティも含めて）最高のルータ プライオリティを持っている場合、R2 は CSNP の完全セットを送信します。この CSNP と上記の 4 で送信された Hello メッセージの両方が受信されると、隣接関係タイマーはキャンセルされます。隣接関係タイマーが切れた場合は、R1 は RR ビットの設定された Hello メッセージを再送信します。

図 14 IETF の実装対応の IS-IS NSF 手順



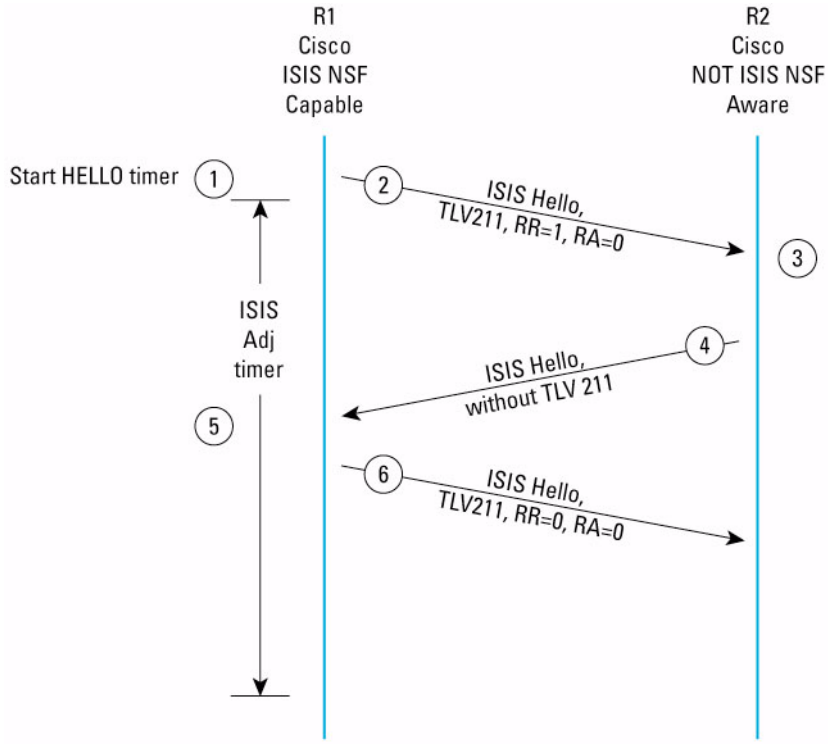
IS-IS NSF 非認識ピアとの手順例

下の説明は、IETF IS-IS NSF 機能が有効、ただしピアは NSF 非認識ルータである場合です。図 15 のダイヤグラムを参照してください。

1. R1 が再起動します。
2. R1 は Hello メッセージを送信します。このメッセージには、RR ビットが設定され、RA ビットがクリアされた TLV 211 が含まれており、R1 が再起動したことが示されます。
3. R2 は R1 から Hello メッセージを受信します。
4. R2 は NSF を認識できないため、TLV 211 なしの Hello メッセージで応答します。隣接関係は廃棄されます。
5. R1 は TLV 211 なしの Hello メッセージを受信します。
6. R2 との隣接関係を再初期化します。R1 は、TLV 211 の RR および RA ビットをクリアして Hello メッセージを送信します。目的は、必ずしも隣接関係を再初期化することではなく（R2 では再初期化済みであるため）、「通常の」隣接関係取得プロセスを実行することです。

注：NSF 非認識ルータが検出されると、IETF IS-IS NSF 対応ルータは、他のすべてのセグメントで NSF 処理を無効にします。

図 15 IS-IS NSF 非認識ピアとの IETF IS-IS NSF 手順



IS-IS NSF の展開

IS-IS NSF の展開で推奨される対象ポイントは次の 2 つです。

- シングルポイント オブ フェイラーとなるルータ
- RP のスイッチオーバー発生時に、望ましくないネットワークの不安定化を引き起こすルータ

Cisco IS-IS NSF は、隣接ルータの NSF 機能にかかわらず、同レベルの効果で機能する利点を備えています。IETF バージョンを展開する場合は、IS-IS NSF 手順に再起動ルータとそのネイバーの両方が関与するため、IS-IS NSF 対応ルータのネイバーは NSF 認識ルータである必要があります。

タイマー調整に関する考慮事項

隣接関係を廃棄する標準的なタイムアウト時間は、ポイントツーポイント リンクでは 30 秒、LAN では 10 秒です。この時間内に Hello の伝送が再開できれば、ネイバーはその隣接関係を廃棄しません。したがって、Hello のホールド タイムは、隣接関係が時間切れになる前にプロセスを再開できるだけの十分な長さに設定する必要があります。

「スムーズに」再起動するという目標は、リンク（およびその後のトポロジー）の変更にすばやく反応するという目標と矛盾することになりますが、ホールド タイムの設定値を長くするだけでは、スムーズな再起動は保証できません。Hello タイマーが変動すると、Hello とホールド タイムの時間切れはホールド インターバル全体に均一に分散します。これはどの瞬間にも、多数の隣接関係が時間切れ寸前になっていることを意味します。時間切れ寸前の隣接関係がすべて失われないようにするには、Hello の乗数を 1 より大きくするのが唯一妥当な方法です。これは一般的に行われていることですが、NSF では絶対要件です。Hello の乗数を 2 または 3 にすると、（少なくとも Hello が失われていない隣接関係では、）再起動プロセスで復旧のための Hello インターバルは最大になります。インターフェイスの数が多いときは、NSF ルータの IS-IS 再起動時間を決定する必要があります。

IS-IS NSF の設定

IS-IS で NSF 動作を設定するには、ルータの IS-IS コンフィギュレーション モードで **nsf** コマンドを使用します。デフォルトでは NSF 再起動はオフですが、ルータにはデフォルトで IETF TLV が含まれています。動作モード（シスコまたは IETF）はこの段階で選択されます。

```
router(config)# router isis
```

```
router(config-router)# nsf [cisco/ietf]
```

次のコマンドは、2 つの再起動間のインターバルを（0 ～ 1440 分の範囲で）指定します。ルータのアクティブおよびスタンバイ RP がこの時間より長く稼働していないと、IS-IS NSF はキャンセルされます。デフォルト値は 5 分です。

```
router(config)# router isis
```

```
router(config-router)# nsf interval 600
```

次のコマンドは、IS-IS 隣接関係を持つインターフェイスが再起動完了前にすべて動作状態になるように、NSF 再起動の待ち時間を（1 ～ 60 秒の範囲で）設定します。デフォルト値は 10 秒です。

```
router(config)# router isis
```

```
router(config-router)# nsf interface wait 20
```

次の IETF モード専用コマンドは、overload ビットの設定された独自の LSP が作成されてフラッディングされる前に、LSP データベースが同期化されるのを NSF が待つ時間を（秒単位で）設定します。

```
router(config)# router isis
```

```
router(config-router)# nsf t3 manual 60
```

「adjacency」キーワードを使用すると、この上記の t3 時間は、スイッチオーバー前にネイバーにアドバタイズされる隣接関係のホールド タイムから決定されます。

```
router(config)# router isis
```

```
router(config-router)# nsf t3 adjacency
```

EIGRP NSF

Enhanced Interior Gateway Routing Protocol（EIGRP）は、さまざまなトポロジーとメディアに適した IGP です。EIGRP は拡張ディスタンス ベクタ ルーティングプロトコルで、Diffused Update Algorithm（DUAL）に基づいて、ネットワーク内の宛先への最短パスを計算します。設計の優れたネットワークでは、EIGRP のスケーラビリティが発揮されるので、きわめて高速のコンバージェンスが実現され、またオーバーヘッドトラフィックは最小限に抑えられます。これまでに説明した各種のルーティング プロトコルと同じく、EIGRP ルーティング プロトコルと NSF が相互作用する目的は、RP のスイッチオーバーが発生したときに、ルーティングへの影響が最小限に抑えられ、パケット転送が混乱しないような形で、グレースフルリスタートを実行することです。

ネイバー隣接関係の維持

他のプロトコルと同様、NSF を実現するには、スイッチオーバー中に再起動ルータのピアが再起動ルータへのパケット転送を継続する必要があります。したがって、ピアでネイバー隣接関係がリセットされないようにすることが必要です。

ネイバーによる隣接関係のリセットを防止するため、再起動ルータは新しい Restart（RS）ビットを EIGRP パケットのヘッダーに設定し、再起動を示すことで、スイッチオーバー中でもサービス提供ができることをピアに通知します。EIGRP NSF を設定すると、Hello パケットと NSF 再起動中に送信される最初の INIT アップデート パケットに、RS ビットが設定されます。RS

ビットを Hello パケットに設定することによって、再起動ルータはスイッチオーバーを迅速にネイバーに通知できます。またこれによって NSF 認識ピアは、通常の隣接関係検出と起動の方法を使用するのではなく、NSF 拡張機能に従う必要があることを認識します。

NSF 非認識ネイバーは新しい RS ビットを無視します。このネイバーは INIT アップデート パケットを受信するか、またはホールド タイマーが切れると、隣接関係をリセットします。

Hello パケットまたは INIT パケットによって再起動の通知を受信すると、ネイバーは再起動ピアをピア リストに書き込み、再起動ルータとの隣接関係を維持します。隣接ルータは、再起動ルータに関する状態変更を自身のネイバーにはまったく伝達しません。代わりに、再起動ルータを経由するルートを stale とマーキングし、再起動ルータへのパケット転送を継続します。これによって、ルータの障害に伴うネットワーク パフォーマンスへの悪影響を回避できます。

EIGRP NSF プロトコル拡張機能の手順

図 16 は、NSF 対応ルータでスイッチオーバーが実行されているときの、NSF 対応 EIGRP ルータと NSF 認識 EIGRP ピア間のプロトコル交換を示しています。

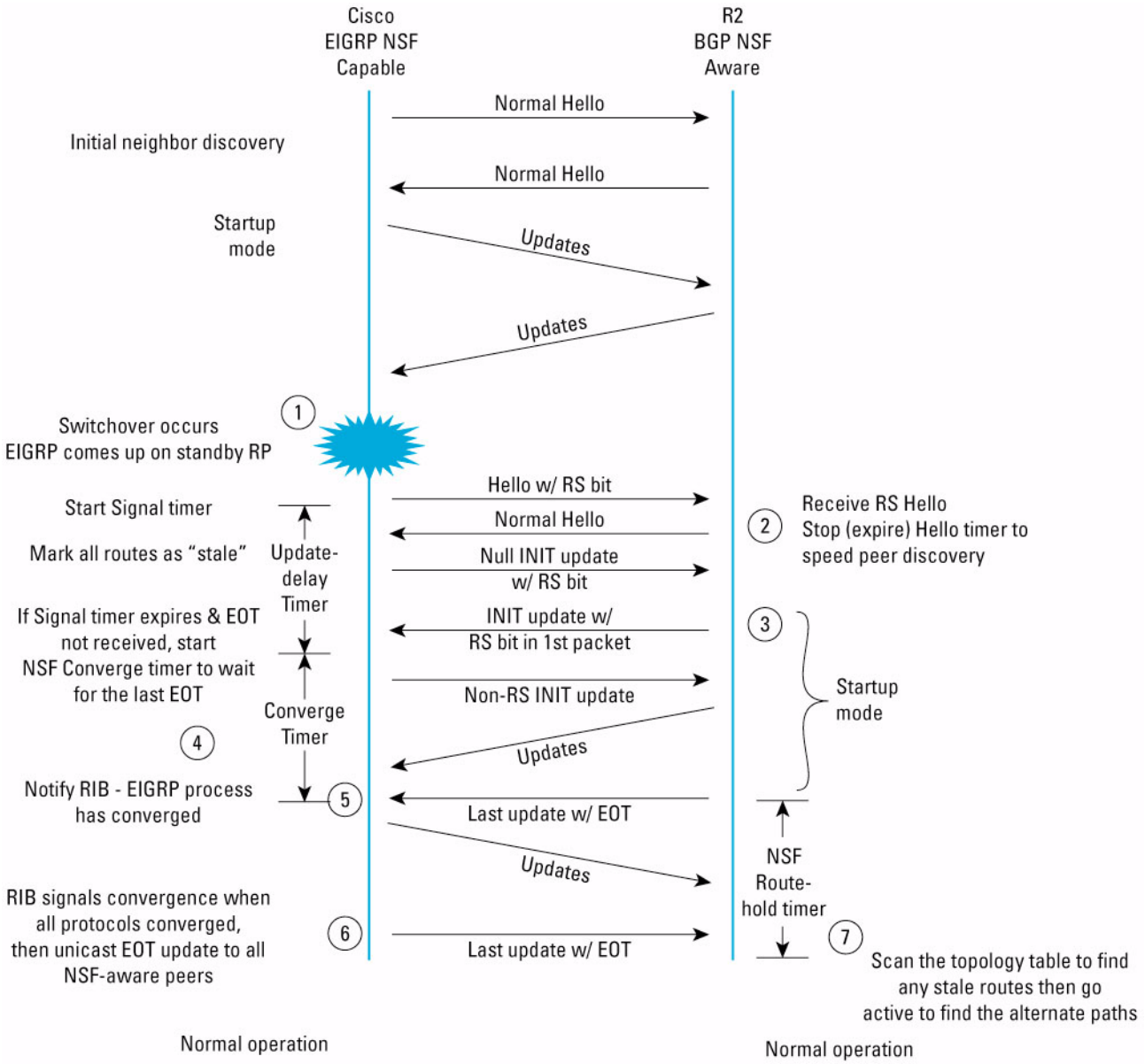
隣接関係が最初に形成されるとき、RS ビットは使用されません。そのため、ピアが EIGRP NSF 手順をサポートできるかどうかを、CLI コマンドから事前に判断することは不可能です。EIGRP NSF に対する各ルータのサポートを判別するには、それらのルータにアクセスするか、または Cisco IOS ソフトウェアのバージョンを確認する必要があります。

1. スイッチオーバーが発生し、スタンバイ RP がアクティブになると、EIGRP は RS ビットの設定された Hello を生成します。これによって再起動が実行されたことがピアに通知されます。
2. ここではピアは EIGRP NSF 認識ルータなので、RS ビットを認識し、再起動ルータに関してそのフォワーディング ステートを維持します。つまりこのピアは、隣接関係をリセットせず、何ごともなかったかのように再起動ルータを経由したパケット転送を継続します。

   再起動ルータがまだピアを再検出していない場合、NSF 認識ルータが INIT パケットの前に Hello パケットを受信することがあります。再起動ルータのピア再検出プロセスを早めるため、NSF 認識ルータはただちにより短い Hello タイム インターバルで Hello パケットを返信します。
3. 次に NSF 認識ネイバーは、最初のアップデート パケットに RS ビットを設定して、そのトポロジー テーブルを再起動ルータに送信します。これによって、そのネイバーが NSF を認識でき、再起動ルータを支援できることを通知します。NSF 認識ネイバーではスイッチオーバーまたは再起動が実行されなかったため、Hello パケットに RS ビットは設定されません。

注 : ルータは NSF を認識できても再起動手順は実行しません。この状態は、再起動ネイバーがリロードされ、コールド スタートから稼働状態になるときに発生します。

図 16 EIGRP NSF の手順



4. 再起動ルータとピアはルーティング アップデートを交換し、NSF 対応ルータはデータベースを再構築します。再起動ルータは、トポロジー テーブル アップデートで End of Table（EOT）マーカーを受信すると、プロセスが完了したことを認識します。NSF 認識ルータはそれぞれが最後のアップデート パケットで End of Table マーカーを送信し、テーブル内容の終了を示す必要があります。

   End of Table マーカー方式に加えて、EIGRP ではタイマー（NSF コンバージ タイマー）が使用され、すべての End of Table マーカーを受信するまでの最大待ち時間が設定されます。

5. 再起動ルータは、EOT マーカーをすべてのピアから受信すると、ただちに通常どおりアップデートを送信し、RIB を伝達します。再起動ルータでネイバーからの EOT マーカーがすべて受信されるか、または NSF コンバージ タイマーが切れると、EIGRP によって Diffusing Update Algorithm（DUAL）計算が実行され、ループのない最適ルートがトポロジー データベースの宛先ごとに選択されます。さらに RIB にコンバージェンスが通知されます。

6. その後、RIB はすべてのプロトコルからコンバージェンス信号を受信すると、EIGRP に対して RIB のコンバージェンスを通知します。RIB のコンバージェンスが完了すると、再起動ルータは再起動に係わった NSF 認識ピアに EOT アップデートを送信します。RIB のコンバージェンスの通知後、再起動ルータによって送信されるこの EOT アップデートには、パケットヘッダー内に EOT フラグのみが含まれ、トポロジー情報は含まれません。

7. NSF 認識ピアは、EOT マーカーを再起動ルータから受信すると、再起動ネイバーのコンバージェンスが完了した時間を認識します。次に、ピアはトポロジー テーブルを走査し、再起動ネイバーが送信元となっているルートを検索します。ピアはルートのタイムスタンプを再起動イベントのタイムスタンプと比較し、ルートがまだ利用できるかどうかを判別します。さらにピアはアクティブになり、再起動ルータを経由するルートで利用できなくなったものについて、代替パスを検出します。

この時点で NSF 拡張機能はすべて完了し、通常の EIGRP 処理が継続します。

EIGRP NSF の展開

再起動ルータが NSF を正しく実行するには、ピア ルータで EIGRP NSF 機能を備えた Cisco IOS ソフトウェア バージョンが動作していることも必要です。ピア ルータで EIGRP NSF 機能を持たない Cisco IOS ソフトウェア バージョンが動作している場合、スイッチオーバーの結果は隣接関係がリセットされるのと変わりません。これは非 NSF EIGRP ルータが、再起動ルータから INIT アップデート パケットを受信すると隣接関係をリセットするためです。

この EIGRP NSF の設計では、2 つの隣接ルータでスイッチオーバーまたは NSF 再起動が同時に実行されるケースはサポートしていません。両方のルータが同時に再起動した場合は、Hello パケットまたは INIT アップデート パケット内の RS ビットによって、一方に他方の再起動が通知されます。両方のルータで NSF 以外の通常の再起動が実行され、それらのピア関係は NSF 以外の方法で再確立されます。

EIGRP NSF では、次のことを認識する 3 つの新しいタイマーが追加されています。

- シグナル タイマー — 各 EIGRP プロセスは、スイッチオーバー イベントを通知されるとシグナル タイマーを開始します。RS ビットの設定された Hello がこの時間内に送信されます。
- コンバージ タイマー — コンバージ タイマーは、シグナル タイマーの時間内に未受信のスタートアップ アップデートがあった場合、最後の EOT アップデートを待ち受けるために使用します。EIGRP プロセスでネイバーが検出されない場合、またはシグナル タイマーの時間内にすべてのスタートアップ アップデートをネイバーから受信した場合、コンバージ タイマーは開始されません。
- ルートホールド タイマー — NSF 認識ピアは、再起動ルータからの EOT を待ち受けるためにルートホールド タイマーを開始します。このタイマーが時間切れになると、ピアは待ち状態を解除してトポロジー テーブルの走査を開始し、さらに再起動ルータによってアップデートされていないルートでアクティブになります。再起動時間が長すぎる場合は、ピアでの代替パスの検出を早めてトラフィックの消失を回避できるように、ルートホールド タイマーを調整（短縮）します。

EIGRP NSF の設定

EIGRP NSF はデフォルトでは無効です。NSF は、次のコマンドで有効または無効にします。

```
router eigrp <AS-number>

[no] nsf
```

タイマーは、次のコマンドで指定できます。

```
router eigrp <AS-number>

   [no] timers nsf signal <seconds>

   [no] timers nsf converge <seconds>

   [no] timers nsf route-hold <seconds>
```

MPLS ネットワークの高可用性

Cisco IOS ソフトウェア リリース 12.2(25)S では、MPLS 環境に対応した HA が導入されています。MPLS High Availability（MPLS-HA）のサポートの主な対象は、MPLS ネットワークへのアクセスを提供する Service Provider Edge（PE; プロバイダー エッジ）デバイスです。これらのデバイスは、MPLS コアに基づかない純粋な IP ネットワークのエッジ ルータと同様に、MPLS VPN サービスのお客様にとってシングルポイント オブ フェイラーとなることがよくあります。

この文書の作成時点では、MPLS-HA 機能のサポートは Cisco 7500 シリーズ ルータでのみリリースされています。他の製品についても順次サポートされる予定です。Cisco 7500 シリーズ製品の MPLS-HA では、MPLS レイヤ 3 VPN 用の NSF/SSO、および Label Distribution Protocol（LDP）NSF（グレースフル リスタート）のサポートが可能となっています。これまでに説明した他のプロトコルと同様、LDP の実装では、隣接ピア ルータでの LDP グレースフル リスタート（NSF 認識）が必要になります。LDP GR 認識は、Cisco IOS ソフトウェア リリース 12.0(29)S 以上が稼働する Cisco 12000 シリーズ製品で利用できます。

MPLS の完全サポートを提供していない Cisco IOS ソフトウェアが稼働しているネットワークでは、RP のスイッチオーバー中に MPLS トラフィックのパケット損失が発生します。ただしこの場合でも、SSO に維持されるリンク レイヤの状態によってすばやく回復できるため、NSF/SSO を有効にすることには、ある程度の利点があります。試験の測定結果によると、多くのリンク タイプでは、MPLS インターフェイスを備えたルータでスイッチオーバーが発生した場合のトラフィック損失は、そのルータで RPR+ のスイッチオーバーが実行された場合とほぼ同じです。ただし一部のリンク タイプでは、トラフィックの損失が減少します。転送を継続するには LDP プロセスでラベルの完全な再起動と学習が必要ですが、再起動の速度は NSF/SSO を有効にすることで向上します。

MPLS-HA が Cisco 12000 シリーズやその他の製品で利用可能になると、MPLS レイヤ 3 VPN、さらに MPLS レイヤ 2 VPN で同じようにパケット損失をゼロにすることが可能になります。

MPLS-HA 機能

現在 Cisco IOS ソフトウェア リリース 12.2(25) では、次の MPLS 機能で RP スイッチオーバー後のデータ転送が継続できます。

- MPLS Virtual Private Network（VPN; 仮想私設網）：これにより、ルータは BGP NSF グレースフル リスタート メカニズムを利用して、VPN プレフィクス情報を失わずにサービスの混乱から回復できます。現在 BGP グレースフル リスタートでは VPNv4 VRF がサポートされているため、BGP グレースフル リスタートを実行するルータは、再起動が発生しても VPN プレフィクス情報を維持できます。
- MPLS LDP：MPLS LDP では、SSO および NSF（グレースフル リスタート）を利用することで、RP が MPLS フォワーディング ステートを失うことなくコントロール プレーン サービスの LDP コンポーネントの混乱から回復できます。LDP グレースフル リスタートは、直接接続されていないピア（ターゲット セッション）だけでなく、直接接続されているピア間の LDP セッションでも機能します。
- Any Transport over MPLS（AToM）：AToM では、SSO、NSF、およびグレースフル リスタートを利用することで、RP が MPLS フォワーディング ステートを失うことなくコントロール プレーン サービスの LDP コンポーネントの混乱から回復できるようになります。
- また、IETF バージョン 8 アップグレードによる MPLS VPN MIB、MPLS LDP MIB の SSO サポートも備えられています。

MPLS-HA 共存機能

この文書の作成時点では、次の MPLS 機能は HA 用に有効になっていません。そのためスイッチオーバー後の状態情報は維持されませんが、NSF/SSO および MPLS-HA とは共存します。

- MPLS トラフィック処理
- MPLS QoS アプリケーション
- IPv6 over MPLS
- MPLS Label Switching Router（LSR; ラベル スイッチング ルータ）MIB
- MPLS TE MIB

- インターフェイス MIB への MPLS 機能拡張

詳細については、この文書の最後にある「参考文献」を参照してください。

MPLS-HA の前提条件

MPLS-HA は、これまでに動作を説明した NSF/SSO の基本機能を基にしています。前提条件は次のようにまとめられます。

- BGP NSF メカニズムを有効にする必要があります。BGP グレースフル リスタートによって、ルータは NSF モードで VPNv4 プレフィクスの MPLS 転送エントリを作成できます。転送エントリは再起動中も維持されます。また BGP は、プレフィクスおよび対応するラベル情報を保存し、再起動後その情報を回復します。
- コア ネットワークの LDP に対する NSF サポート
- コアで使用される IGP（OSPF または IS-IS）に対する NSF サポート
- PE および Customer Edge（CE; カスタマー エッジ）ルータ間のルーティング プロトコルに対する NSF サポート

MPLS-HA の動作

BGP は、プレフィクスにローカル ラベルを割り当てると、そのローカル ラベル バインディングをバックアップ RP でチェックポイント化します。チェックポイント機能は、状態情報をアクティブ RP からバックアップ RP にコピーします。これによって、バックアップ RP は最新情報とまったく同じコピーを保持するようになります。アクティブ RP に障害が発生しても、サービスを中断することなくバックアップ RP に切り替えることができます。チェックポイント化は、アクティブ RP がすべてのローカル ラベル バインディングを、バックアップ RP にコピーするバルク同期化を行うときに開始されます。その後アクティブ RP は、ラベルの割り当てまたは解放のときに、個々のプレフィクス ラベル バインディングをダイナミックにチェックポイント化します。これによって、BGP の再コンバージェンス前でもラベル付きパケットの転送を続けることができます。

BGP グレースフル リスタート機能を持つルータが接続を失うと、再起動ルータは次のように動作します。

1. ルータは他のルータとBGPセッションを確立し、同じようにグレースフル リスタート機能を持つ他のルータからBGPルートを再学習します。再起動ルータは隣接ルータからアップデートを受信するのを待ちます。隣接ルータが End of RIB マーカーを送信してアップデートの送信完了を通知すると、再起動ルータは自身のアップデートの送信を開始します。

2. 再起動ルータはチェックポイント データベースにアクセスして、各プレフィクスに割り当てられたラベルを検出します。ラベルが検出されると、再起動ルータはそれを隣接ルータにアドバタイズします。ラベルが検出されない場合は、新しいラベルを割り当て、それをアドバタイズします。

3. 再起動ルータは stale エントリのタイマーが切れてから、stale プレフィクスをすべて削除します。

BGP グレースフル リスタート機能を持つピア ルータは、再起動ルータを検出すると、次の動作を行います。

1. ピア ルータはすべてのルーティング アップデートを再起動ルータに送信します。アップデートの送信が完了すると、ピア ルータは End of RIB マーカーを再起動ルータに送信します。

2. ピア ルータは再起動ルータから学習した BGP ルートを、すぐには BGP ルーティング テーブルから削除しません。ピア ルータは再起動ルータからプレフィクスを学習していき、新しいプレフィクスとラベル情報が古い情報と一致すれば、stale ルートをリフレッシュします。

VPN NSF 用に設定されていないルータが、VPN NSF を備えたルータとの BGP セッションの確立を試みた場合、2 台のルータは通常の BGP セッションを確立しますが、VPN NSF は実行できません。

LDP グレースフル リスタート（NSF）

LDP NSF（LDP グレースフル リスタート [GR]）は、RP のスイッチオーバーが片方の LSR で発生するなどして 2 台の LSR 間の LDP 通信がいったん失われ、その後復旧する場合に、それらの LSR で LDP とフォワーディング ステートを保護するために使用できるメカニズムです。LDP GR によって、中断された LDP 通信が障害から復旧するまでの間も、以前に学習したラベルを使用するトラフィックでノンストップの MPLS 転送が可能になります。

この実装によって、LDP コンポーネント（コントロール プレーン）の再起動やネイバーとの LDP 通信の一時的な中断から保護されます。LDP コンポーネントの再起動（LDP 再起動）が発生すると、すべての LDP ネイバーとの LDP 通信が中断し、それらのネイバーから学習した LDP 状態が失われます。LDP コンポーネントは再起動せず、ネイバーとの通信が失われただけ（LDP セッション リセット）の場合は、そのネイバーから学習した LDP 状態は、関連するフォワーディング ステートも含めて維持されます。

LDP GR は、この文書で説明している他の NSF プロトコルと同様に動作します。LDP GR では、LDP 通信の障害から回復するために、LSR が次のように動作することが必要です。

- LDP 通信の障害が検出されると、関連するフォワーディング ステートを stale とマーキングし、維持する。
- stale 状態を使用して転送を継続する。
- LDP 通信が再確立されると、stale のフォワーディング ステートを復旧し、リフレッシュする。
- stale のフォワーディング ステートが、要求される時間内に「リフレッシュ」されない場合は、それを削除する。

LDP GR プロトコル拡張機能

LDP を使用してラベル マッピング情報を交換する 2 台の LSR は、LDP ピアと呼ばれています。LDP ピアの 1 つが LDP GR 対応で、ピアが少なくとも LDP GR を認識できる場合は、MPLS-HA が可能です。LDP は、ラベル スイッチング パスを経由してピア間に「LDP セッション」を確立することで機能します。単一の LDP セッションによって、各ピアは他のピアのラベル マッピングを学習できます。

図 17 は、2 台の LSR 間の LDP GR メッセージ フローを示しています。

LDP では、Hello メッセージが交換され、LDP メッセージを伝送するために TCP セッションが確立されると規定しています。

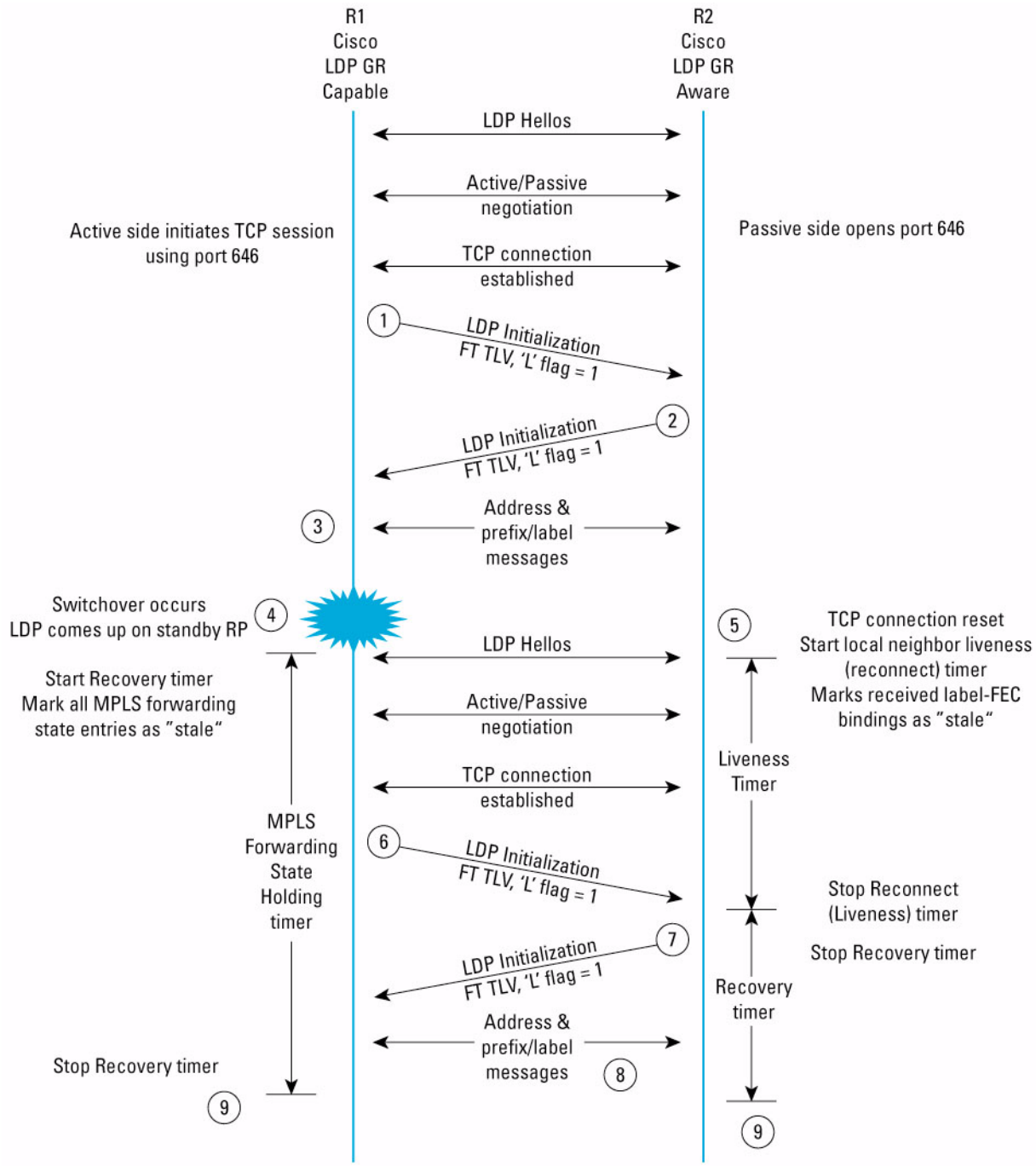
1 & 2. LSR は、LDP 初期化メッセージの中にオプション パラメータとして Fault Tolerant（FT）Session TLV を含めることによって、LDP グレースフル リスタートのサポートが可能であることを示します。L（Learn from Network）フラグは、LDP GR 手順が使用されることを示します。オプションの FT Session TLV は、下位互換性を確保できるように定義されています。TLV の「U ビット」が設定されると、受信側は LDP GR をサポートしていない場合に TLV を自動的に削除します。その場合、LDP セッションは確立しますが、GR は実行されません。

LDP GR 関連で、FT Session TLV に存在するタイマー フィールドは次の 2 つです。

- **再接続タイムアウト：**TLV の送信側が受信側に要求する、LDP 通信障害検出後の待ち時間（ミリ秒）です。待ち状態の間、受信側では、送信側と受信側との間のリンクを通過する（確立済み）LSP の MPLS フォワーディング ステートが維持されます。FT 再接続タイムアウトは、TLV の送信側のコントロール プレーンが再起動できるように、十分長くする必要があります。特にその LDP コンポーネントが、送信側でネイバーと LDP メッセージを交換できるような状態になることが必要です。このタイマーのデフォルト値は 120 秒です。
  FT 再接続タイムアウトを 0 に設定すると、TLV の送信側では再起動の前後にそのフォワーディング ステートは維持されませんが、RFC 3478 の 3.3 項「Restart of LDP communication with a neighbor LSR」で定義されている手順はサポートすることが示されます。

- **復旧時間：**再起動する LSR での復旧時間とは、再起動の前後に維持していた MPLS フォワーディング ステートを、LSR が維持しようとする時間（ミリ秒）です。これは、LSR が FT Session TLV を含む初期化メッセージを再起動後に送信する瞬間から始まります。

図 17 LDP グレースフル リスタート拡張機能



この時間を 0 にすると、再起動の前後に MPLS フォワーディング ステートは維持されていないことが示されます。

3. 再起動の前にラベル情報が交換されます。

4. RPのハードウェアまたはソフトウェア障害によってスイッチオーバーが発生すると、ルータの再起動するLDPコンポーネントは、ピアに対して新しい TCP セッションを確立します。LDP GR 対応ルータは、MPLS フォワーディング ステート ホールディング タイマー（forwarding-holding）という内部タイマーを開始し、MPLS フォワーディング ステートのエントリをすべて「stale」とマーキングします。ルータは LDP 再起動モードになります。この forwarding-holding タイマーのデフォルト値は 600 秒です。

5. LDP GR 認識ピアである R2 は、障害が検出されて再起動すると、liveness タイマーというタイマーを初期化します。このタイマーは、ピアの LSR が stale のラベル /FEC バインディングを維持する時間を示します。このタイマーは、再起動ルータによって以前にアドバタイズされた FT 再接続タイムアウト、および Neighbor Liveness タイマーというローカル タイマーよりも小さな値に設定されます。

   この時間内に LSR がネイバーとの LDP セッションを確立しなければ、stale バインディングはすべて削除されます。

6. 再起動 LSR は、FT Session TLV で送信される復旧時間を、MPLS フォワーディング ステート ホールディング タイマーの現在の値に設定します。このタイマーは、フォワーディング ステートがその後維持される時間を示します。

7. LDP セッションが確立します。再接続タイマーが切れる前にネイバーとの LDP セッションが再確立した場合は、再接続タイマーが停止し、復旧タイマーが開始します。

8. LSR はアドレス プレフィクスとラベル メッセージを交換します。

9. ピア ルータが再アドバタイズしていない stale の送信ラベル バインディングを削除するために、LDP コンポーネントはもう 1 つのタイマー（max-recovery）を使用します。このタイマーは初期化メッセージが送信されると開始します。デフォルト値は 120 秒です。このタイマーが切れると、LDP は stale とマーキングされ、タイマーの時間切れに関連する再起動ネイバーから学習した LIB 内の送信ラベル バインディングを、すべて削除します。ラベル バインディングが削除されると、関連するフォワーディング ステートはすべて削除されます。

ピア側の処理

LDP ピアは、既存の LDP セッションの障害によって、進行中の LDP の復旧を認識します。LDP セッションの障害は、次の場合に検出されると考えられます。

- セッションの最後の Hello 隣接関係が失われる。
- セッションのキープアライブ タイマーが時間切れになる。
- 接続が失われたことが TCP レイヤから通知される。
- Shutdown または Notification メッセージの受信によって通常のクローズが起動される。

LDP GR 認識ルータは LDP セッション リセット モードになります。LDP GR 認識ルータは、再起動ルータとの LDP セッションを再確立する時間を設定するためにタイマーを開始します。ルータが LDP セッションの再確立を待ち受ける時間は、ピアから FT TLV で受信される再接続時間、または neighbor-liveness タイマーの値よりも短くなります。タイマーが切れる前に LDP セッションが確立された場合、ルータはそのネイバーと関連する stale ラベル バインディングを削除します。neighbor-liveness タイマーのデフォルト値は 120 秒です。

再起動手順中は既存のラベル バインディングが使用されます。送信ラベル バインディングは、LDP セッションの再確立後、再起動ルータによってアドバタイズされます。LDP がアドバタイズされたプレフィクスのバインディングを保持している場合は、新しいバインディングが既存の LDP 手順を使用して学習されます。ラベル バインディングが保持されている場合、LDP は新しいバインディングが学習されたときに、保持されているバインディングの「stale」マークを消去します。同じラベルがアドバタイズされた場合は、動作は必要ありません。新しいラベルがアドバタイズされると、LDP はそのイベントを処理するために既存の手順を使用して、ラベル バインディングをアップデートする必要があります。

MPLS-HA の設定

MPLS-HA のサポートを有効にするには、まずルータを SSO モードにする必要があります。次に、LDP グレースフル リスタートを有効にすることが必要です。NSF で最大の効果が得られるように、IGP プロトコルと PE-CE プロトコルも有効にする必要があります。次に、LDP の具体的な設定について説明します。

LDP グレースフル リスタートの設定

LDP GR は、次のグローバル コマンドによって有効化します。

```
Router(config)# mpls ldp graceful-restart
```

LDP GR タイマーの値は、関連する次のコマンドで指定します。

```
Router(config)# mpls ldp graceful-restart timers <timer> <value>
```

このコマンドでは、次のタイマーを指定できます。

- forwarding-holding：コントロール プレーンの再起動後に、MPLS フォワーディング ステートを維持する時間を指定します。
- neighbor-liveness：ルータが LDP セッションの再確立を待ち受ける時間を指定します。
- max-recovery：LDP セッションの再確立後に、ルータが stale のラベル /FEC バインディングを保持する時間を指定します。

LDP グレースフル リスタートの状態は、次のコマンドを使用して決定できます。

```
Router# show mpls ldp graceful-restart

LDP Graceful Restart is enabled

Neighbor Liveness Timer: 5 seconds

Max Recovery Time: 200 seconds

Down Neighbor Database (0 records):

Graceful Restart-enabled Sessions:

VRF default:

    Peer LDP Ident: 18.18.18.18:0, State: estab

    Peer LDP Ident: 17.17.17.17:0, State: estab
```

NSF/SSO のソフトウェアとハードウェアのサポート

Cisco NSF/SSO のサポートは、Cisco IOS ソフトウェア 12.0(22)S で初めて登場しました。その後、他のリリースにも拡張され、Cisco 7500 シリーズ ルータ用の Cisco IOS ソフトウェア リリース 12.2(25)S では MPLS-HA のサポートが導入されました。NSF/SSO は現在、幅広いシスコ製品、RP、およびライン カード ハードウェアで利用できます。また NSF 認識も、さまざまなシスコ製品ファミリーの複数のリリースで一般的な機能となっています。

ハードウェアの制限

ルータには SSO をサポートするために、互換性のある RP とライン カードが搭載されている必要があります。さらに、次のタイプの RP を混在させるときは注意が必要です。

- Cisco 12000 シリーズ ルータ：GRP および GRP-B RP を併せて使用できます。このルータで PRP を使用する場合は、もう 1 つ PRP を用意してペアにする必要があります。
- Cisco 10000 シリーズ インターネット ルータ：2 つの PRE-1 または PRE-2 を使用する必要があります。このルータのもとの PRE は、Cisco NSF/SSO をサポートしていません。MPLS-HA は、Cisco IOS ソフトウェア リリース 12.2S で PRE-2 用にサポートされます。

- Cisco 7500：RSP-2 と RSP-4 を組み合わせて使用できます。また、RSP-8 と RSP-16 も組み合わせて使用できます。ただし、RSP-8 または RSP-16 を RSP-2 または RSP-4 と混在させることはできません。
- Cisco Catalyst 6500 シリーズおよび Cisco 7600 シリーズでは、Supervisor Engine 2 と Supervisor Engine 720 がサポートされています。同種のスーパーバイザを使用する必要があります。
- RP ハードウェアの物理的特性が異なる場合、一部の製品では警告が表示されることもあります（メモリ容量が異なる場合など）。その相違が許容されるものである場合、システムは SSO モードに入ります。ネットワーク設計者は、スイッチオーバー発生時に物理的特性の相違がパフォーマンスに影響を与えないことを確認する必要があります。

さまざまなライン カードで Cisco NSF/SSO がサポートされています。パフォーマンスを最適化するには、ルータ シャーシ内のすべてのカードで Cisco SSO がサポートされている必要があります。各プラットフォームで現在サポートされているライン カードの一覧については、シスコの最新マニュアルで確認してください。Cisco SSO でサポートされていない特定のライン カードの場合、そのライン カードは RPR+ モードで動作します。RP のスイッチオーバー時には、カード上の分散転送情報が消去されます。これにより、スイッチオーバーが発生した場合、そのカードを経由して到達できる宛先へのトラフィックに損失が発生します。他のライン カードは、スイッチオーバー中も転送を継続します。

各種ライン カードまたはモジュール組み合わせのサポートに関する詳細は、CCO 上の該当するマニュアルを参照するか、シスコの担当者にお問い合わせください。

NSF/SSO の実装手順

この項では、NSF/SSO の展開時に踏むべき実際の実装手順について概説します。

ピアの組み合わせの検討

実装前の最初のステップはネットワークを再検討し、手持ちのさまざまなピアの組み合わせを分析することです。シスコ製品だけで構成されたネットワークを利用しているお客様は、展開が最も簡単です。マルチベンダー環境を利用しているお客様は、グレースフル リスタート規格とプロトコル拡張機能のサポート レベルを確認する必要があります。いずれの場合も、HA 認識のサポートでは、さまざまな組み合わせを利用できます。

次に、実装の戦略を考えます。推奨する方法は、コアから実施することです。上位層のコア対面ルータを、必要な NSF 認識をサポートする Cisco IOS ソフトウェアのレベルまでアップグレードします。さらにサイトごと、または特定のロケーションで作業を行い、NSF/SSO をエッジ デバイスに実装します。

ルート リフレクタを使用する場合は、このルータが BGP NSF を認識するようにします。これが完了すると、ネットワーク エッジ境界で NSF/SSO の有効化を開始できます。

OSPF NSF 認識はデフォルトでオンですが、BGP NSF 認識は設定の必要があります。またピアをリセットして、BGP NSF がその特定のピアについて有効になるようにする必要があります。OSPF NSF が有効で、BGP NSF が無効の状態は望ましくないので、実装の観点では避ける必要があります。

サービス プロバイダーでの展開例

設定の準備

サービス プロバイダーの POP またはサイト内で NSF/SSO を展開するための最初の準備ステップは、サイト内のすべてのルータの機能別一覧を作成することです。たとえば、カスタマー アクセス ルータ、アグリゲーション レイヤ ルータ、ルート リフレクタ、コア ルータなどです。

NSF/SSO の展開で予想される適切なレベルのソフトウェアが、すべてのルータで実行されていることを再確認します。

ピアリングの設定をルータごとに再確認し、実際の設定が実行されるときにすべての OSPF または IS-IS および BGP ピアが含まれるようにします。

特定のネットワークで実行されるステップを簡単にまとめると、次のようになります（この例では、IGP を OSPF とし、ルート リフレクタを使用すると仮定しています）。

1. ルート リフレクタで BGP NSF を設定し、それらが NSF を認識するようにします。
2. コア対面ルータの OSPF NSF 機能を再確認します。
3. アグリゲーション ルータで OSPF NSF と BGP NSF を設定します（RR のピアのリセットは、この時点、またはアクセス ルータの設定後に行うことができます）。
4. カスタマー アクセス ルータ（エッジ ルータ）で OSPF NSF と BGP NSF を設定します。
5. BGP ピアをリセットし、すべての NSF 対応ピアが機能のネゴシエーションを行うようにします。
6. すべてのデュアル RP ルータを SSO 用に設定します。
7. カスタマーの NSF 機能を再確認します。BGP が SP エッジとカスタマー ネットワークの間で使用され、コードが NSF をサポートしている場合は、カスタマー ルータで BGP NSF を設定し、BGP セッションをリセットします。

以下に、NSF/SSO を実装する詳細なコンフィギュレーション コマンドを例示します。

段階的な実装手順

この項では展開の例、および NSF/SSO の実装に使用する一連の CLI コンフィギュレーション コマンドを示します。

Cisco 7500 シリーズ ルータと Cisco 12000 シリーズ ルータについては、特記すべき運用上の注意事項がいくつかあります。

Cisco 7500 シリーズ ルータ運用上の注意事項

SSO スイッチオーバーが発生すると、新しいスレーブ（以前のアクティブ）は、ブート シーケンスを開始する前に 5 分間だけ ROMMON にとどまります。これは設計に従った動作です。Cisco 7500 上のスレーブは、（Cisco 12000 と異なり）マスターの支援がなければ起動できません。マスターは RP を部分的に起動するためにサブセット イメージをバスで提供します。この 5 分間は、サブセット イメージをスレーブにロードするという、プロセッサに大きな負荷がかかるタスクからの影響を最小限に抑えながら、新たにアクティブになった RP が通常の復旧動作を行えるように設計されています。

スレーブがオンラインになると、トラフィックはスレーブがバスに再接続するときにもう一度中断します。このトラフィックの中断は、スイッチオーバーが発生した際の最初のトラフィック損失以下になります。

Cisco 12000 シリーズ ルータ運用上の注意事項

スイッチオーバーが発生すると、新しいスレーブ（以前のアクティブ）はただちにブート プロセスを開始します。プロセッサは個別に起動できるため、アクティブ プロセッサの負荷が下がるまで待つ必要はありません。

スレーブがオンラインになるときにトラフィック損失は発生しません。

設定

設定の最初のステップは、最上位レベルのすべてのルート リフレクタで、NSF 認識のために BGP グレースフル リスタートを有効にすることです。IGP も NSF を認識できる必要があります。OSPF 認識はデフォルトでオンなので、設定の必要はありません。ルート リフレクタで BGP グレースフル リスタート設定の動作を行っても、トラフィックへの影響はまったくありません。次のコマンドを使用します。

```
Router(config)#router bgp <as number>

Router(config-router)#bgp graceful-restart
```

次に、このサイトのアグリゲーション レイヤ ルータで OSPF NSF を有効にします（関連する Cisco IOS ソフトウェア ベースのデバイスが存在する場合）。

```
Router(config)#router ospf <process id>

Router(config-router)#nsf
```

同じコマンドを使用して、サイト内のすべてのカスタマー アクセス レイヤ ルータまたはエッジ ルータで、BGP および OSPF グレースフル リスタートを有効にします。

```
Router(config)#router ospf <process id>

Router(config-router)#nsf

Router(config)#router bgp <as number>

Router(config-router)#bgp graceful-restart
```

この時点で、ルータは「望ましくない」設定状態の 1 つになります。OSPF NSF は稼働していますが、BGP ピアがリセットされていないため、BGP NSF は未稼働です。NSF/SSO の設定手順ではこうした状態は回避できないので、次のステップをすみやかに実行し、SSO/NSF を完全に稼働状態にします。

次に、BGP ピアをリセットする必要があります。通常サービス プロバイダーの POP の設計では、冗長アグリゲーション ルータが配置されます。この場合、**clear ip bgp ＊**を一方のアグリゲーション レイヤ ルータで実行してピアが再確立するのを待ち、次にもう一方のアグリゲーション レイヤ ルータで **clear ip bgp ＊**を実行するのが最も簡単です。これによって、設定した最初のサイトの BGP RR ピアがすべてキャッチされます。そのあと他のサイトを設定していくときは、**clear ip bgp ＊**を使用してリセット済みサイトの BGP RR ピアを再度リセットする必要はありません。BGP ピアをそれぞれ任意にリセットすることは可能ですが、すべての BGP ピアがグレースフル リスタート機能を備えるようにすることが重要です。

次に、デュアル プロセッサを備えたアグリゲーション レイヤ ルータとアクセス レイヤ ルータを、すべて Redundancy Mode SSO で設定します。

Cisco 7500 シリーズ ルータでは、ハードウェア モジュール コマンドを設定することが必要です。

```
router#conf t

router(config)#hardware-module slot 6 image disk0:<image-name>

router(config)#hardware-module slot 7 image disk0:<image-name>
```

注：スロットは、ルータが Cisco 7507 と Cisco 7513 のどちらであるかによって異なります。

両タイプのルータに、次のコマンドを使用して SSO を設定します。

```
router#conf t

router(config)#redundancy

router(config-red)#mode sso
```

SSO を設定すると、スレーブは自動的にリセットされます。Cisco 12000 シリーズ ルータのリセット中、トラフィック損失は発生しません。Cisco 7500 シリーズ ルータでは、スレーブをリセットすると、スレーブがオンラインに復帰するときに短時間だけトラフィック損失が発生します。

最後に、NSF/SSO の設定を各ルータで確認する必要があります。

OSPF では、次のコマンドで NSF が有効になっていることを確認します。

```
router>sh ip ospf | inc Non-Stop
```

Non-Stop Forwarding enabled

BGP では、各ピアを次の行で確認する必要があります。

```
Graceful Restart Capability: advertised and received
```

機能はアドバタイズされ、かつ受信される必要があります。受信されない場合は、ピアの相手側で BGP グレースフル リスタートが設定されていないか、またはそのピアがリセットされていないかのいずれかです。

最後に **sh redundancy** コマンドで、ルータが SSO モードで動作していることを確認します。

```
router>sh red
Redundant System Information:
------------------------------
       Available system uptime = 12 minutes
Switchovers system experienced = 0
              Standby failures = 0
        Last switchover reason = none
                  Hardware Mode = Duplex
    Configured Redundancy Mode = sso
     Operating Redundancy Mode = sso
              Maintenance Mode = Disabled
                Communications = Up
```

まとめと利点

ここ数年の間に、ネットワークの可用性は、サービス プロバイダーと企業の双方にとってますます重要な問題となってきました。シスコは、Cisco IOS ソフトウェアの HA インフラストラクチャと Cisco NSF/SSO、およびグレースフル リスタート用のさまざまなルーティング プロトコル拡張機能など、高可用性ネットワーキングに向けた包括的戦略を実践しています。ユーザとベンダーがこうした拡張機能による展開の経験を蓄積するにしたがって、プロトコルそのものとネットワークの展開方法に、さらに改良が加えられます。

シスコは SSO を、Cisco IOS ソフトウェアのインフラストラクチャ機能として実装しています。Cisco SSO は、実際にはレイヤ 2 接続の維持以上の機能を提供しており、すべてのサポート プラットフォームとインフラストラクチャの状態の管理も行っています。レイヤ 2 接続の維持は、Cisco SSO の提供サービスの中で最も目につきやすいものであるにすぎません。

まとめとして、シスコはお客様のニーズに対応する革新的機能を提供し続けます。シスコの HA 戦略は単純です。つまり、ダウンタイムの原因となるあらゆる可能性に対処し、MTBF を拡大して MTTR を短縮するための特徴、機能、ベスト プラクティス設計の推奨事項、および運用手順を提供することです。Cisco NSF/SSO は現在、既存のハードウェアに展開することが可能です。Cisco NSF/SSO は、ネットワーク ルーティング プロトコルの再コンバージェンス、およびそれに伴うトラフィック バーストと CPU 負荷を最小限に抑えます。また、ネットワーク利用の計画性を強化し、信頼性に対するユーザの認知度も向上させます。

関連する規格とドラフト

RP のスイッチオーバー発生時に転送を継続させるメカニズムは、完全には規格化されていません。関連する文書を以下に挙げます。

シスコの BGP の実装は、draft-ietf-idr-restart-nn.txt に記されている仕様に準拠しています。この文書の作成時点での最新バージョンは draft-ietf-idr-restart-10 で、これは規格案です。

シスコの OSPF の実装は、次の IETF のドラフトに記されている仕様に準拠しています。

- draft-nguyen-ospf-lls-04 2004-01-07 IESG で処理中 — 検索画面 DRAFT TRACKER から参照してください。
- draft-nguyen-ospf-oob-resync-04 2004-01-07 IESG で処理中 — 検索画面 DRAFT TRACKER から参照してください。
- draft-nguyen-ospf-restart-04 2004-01-07 IESG で処理中 — 検索画面 DRAFT TRACKER から参照してください。

OSPF Hitless Restart の現在の標準は、RFC 3623『Graceful OSPF Restart』です。現在（この文書の作成時）シスコの実装は、RFC 3623 と相互運用できません。

IS-IS（IETF オプション）の Cisco NSF の実装は、RFC 3847『Restart Signaling for Intermediate System to Intermediate System（IS-IS）』に記されている仕様に準拠しています。シスコ固有のステートフル実装もサポートされており、設定が可能です。

シスコの LDP の実装は、RFC 3478『Graceful Restart Mechanism for Label Distribution Protocol』に記されている仕様に準拠しています。

用語集

Autonomous System（AS; 自律システム）	技術的には、共通の管理制御を受けているルータのグループ。実用上は、BGP コンフィギュレーションで一般的に設定された AS 番号を共有するルータのグループ
Cisco Express Forwarding（CEF）	FIB をさらに最適化したもので、非常に高速の IP パケット交換が可能。Distributed Cisco Express Forwarding（dCEF）は、ライン カード上で動作する CEF の一種
コンバージェンス	ネットワーク上のすべてのルータで、ピア ルータからのルーティング情報の受信と処理がすべて完了すること
eBGP	異なる AS に属するルータ間の BGP ピアリング接続
Forwarding Information Base（FIB; フォワーディング情報ベース）	RIB を検査して特定の宛先 IP アドレスに対して唯一の最適パスを選択することで形成される、最適化されたルーティング テーブル。ロード シェアリングまたはロード バランシングが有効の場合は、複数の最適パスが選択されることもある
iBGP	同一の AS に属するルータ間の BGP ピアリング接続
Interior Gateway Protocol（IGP; 内部ゲートウェイ プロトコル）	AS 内で動作し、ネクスト ホップの接続情報を提供するプロトコル（OSPF、IS-IS、EIGRP が標準的）
Nonstop Forwarding（NSF; ノンストップ フォワーディング）	RP のスイッチオーバー中にバックグラウンドでフォワーディング ステートを維持し、ルーティング プロトコルの再コンバージェンスを行うルータの機能。シスコは BGP、OSPF、および IS-IS の機能を拡張しているので、それらの拡張機能を一括して Cisco NSF という
NSF 対応ルータ	NSF が実装され、RP に障害が発生してもパケット転送を継続できるルータ。注：NSF 対応ルータはいずれもが NSF 認識ルータである。ただし NSF 認識ルータが NSF 対応ルータである必要はない
NSF 認識ルータ	ルーティング プロトコルに必要な変更を加えて、NSF 対応ネイバーを支援できるルータ
NSF 非認識ルータ	NSF 認識ルータではないルータ
再起動ルータ	RP のスイッチオーバーが実行されているルータ
Routing Information Base（RIB; ルーティング情報ベース）	1 台のルータに関するすべてのルーティング情報を集約したもの。さまざまな宛先 IP アドレスに対する複数の参照が含まれている場合もある
ルート選択プロセス	Border Gateway Protocol（BGP）が、ピアから取得した利用可能な情報をすべて使用して、特定の宛先への最適ルートを選択するプロセス。別名最適パス選択
Stateful Switchover（SSO; ステートフル スイッチオーバー）	プラットフォーム、インフラストラクチャ、およびレイヤ 2 接続に関する情報をデュアル RP 間で共有するプロセス。SSO では、RP のスイッチオーバーの前後でレイヤ 2 接続を維持することも可能

参考文献

MPLS ハイ アベイラビリティ — 概要（英語）：

http://www.cisco.com/en/US/products/sw/iosswrel/ps1838/products_feature_guide09186a008029b23d.html

MPLS VPN：SSO/NSF のサポート（英語）：

http://www.cisco.com/en/US/products/sw/iosswrel/ps1838/products_feature_guide09186a008029b289.html

MPLS LDP：SSO/NSF のサポートとグレースフル リスタート（英語）：

http://www.cisco.com/en/US/products/sw/iosswrel/ps1838/products_feature_guide09186a008029b285.html

MPLS LSR MIB（英語）：

http://www.cisco.com/en/US/products/sw/iosswrel/ps1838/products_feature_guide09186a008029b23c.html

MPLS LDP MIB バージョン 8 アップグレード（英語）：

http://www.cisco.com/en/US/products/sw/iosswrel/ps1838/products_feature_guide09186a00801b1bdc.html

CISCO SYSTEMS

シスコシステムズ株式会社

URL: http://www.cisco.com/jp/

問合せ URL: http://www.cisco.com/jp/go/contactcenter/

〒 107-0052 東京都港区赤坂 2-14-27 国際新赤坂ビル東館

TEL: 03-6670-2992

電話でのお問合せは、 以下の時間帯で受付けております。

平日 10:00 ～ 12:00 および 13:00 ～ 17:00

お問合せ先

05.04