Influence des bonnes pratiques sur les incidents BGP

François Contat[1], Sarah Nataf[2], and Guillaume Valadon[1]
francois.contat(@)ssi.gouv.fr
sarah.nataf(@)orange.com
guillaume.valadon(@)ssi.gouv.fr

[1] ANSSI 51 bd. de La Tour-Maubourg 75700 Paris 07 SP
[2] France Télécom Orange 1 av. Nelson Mandela 94110 Arcueil

Mots-clés: BGP, incidents, bonnes pratiques, configuration

Résumé Partant du constat que le protocole BGP n'utilise pas de mécanismes de sécurité forts, cet article vise à présenter les incidents BGP rencontrés par les opérateurs ainsi que les contre-mesures mises en place (filtres, annonces plus spécifiques, ou bien encore surveillance des préfixes gérés).

La contribution majeure de cette soumission repose dans la description de ces incidents et dans le détail de leur gestion. Nous proposons ainsi une approche à la fois théorique et pratique pour appréhender les problématiques liées à la sécurité du protocole BGP. À l'occasion des dix ans du SSTIC, cette soumission est également l'occasion de faire le point sur les incidents qui ont été les plus médiatisés depuis 2002.

Introduction

Le protocole de routage BGP, défini dans la RFC 4271 [25], est utilisé par tous les acteurs de l'Internet pour s'interconnecter et échanger des routes. Les interconnexions qu'il permet de réaliser constituent la structure de l'Internet. Par conséquent, BGP est un protocole dont le fonctionnement influence directement celui de l'Internet.

Partant du constat que BGP n'utilise pas de mécanismes de sécurité forts, cet article présente tout d'abord les problèmes auxquels ce protocole est confronté. Les contre-mesures et bonnes pratiques associées sont ensuite détaillées et illustrées à l'aide d'exemples de configuration de routeurs BGP. À l'occasion des dix ans du SSTIC, les incidents les plus médiatisés de ces dix dernières années sont également détaillés.

En plus de l'état de l'art sur les incidents et leurs contre-mesures, la contribution majeure de cet article repose sur un retour d'expérience

de la surveillance des annonces BGP sur un réseau opérationnel. Nous proposons ainsi une approche théorique et pratique pour appréhender les problématiques liées à la sécurité du protocole BGP.

Cet article est organisé comme suit. Dans la première section, une description du protocole BGP est donnée. Les deuxième et troisième sections décrivent les protections que l'on peut mettre en place sur les interconnexions BGP ainsi que les approches pour limiter les effets de messages BGP malformés. La quatrième section présente les incidents BGP les plus médiatisés depuis 2002. Enfin, le bilan de la surveillance des annonces BGP sur un réseau opérationnel est présenté.

1 Présentation du protocole BGP

L'Internet repose sur des réseaux de diverses natures administrés indépendamment par un grand nombre d'opérateurs. Chacun d'eux gère des blocs d'adresses IP qu'il peut diviser en préfixes de plus petites tailles, pour ses besoins ou ceux de ses clients. Ces réseaux indépendants sont interconnectés grâce au protocole BGP [25], dont l'objectif est d'échanger des préfixes. Dans la terminologie BGP, un réseau est appelé *Autonomous System* (AS) et identifié par un numéro d'AS (AS Number). Une entreprise utilisant le protocole BGP aura généralement un numéro d'AS unique dans l'Internet.

1.1 Structure des interconnexions BGP

Il existe deux types de sessions BGP qui correspondent à deux usages distincts du protocole. Dans la figure 1, les sessions eBGP (pour *external BGP*) permettent aux trois AS de partager leurs routes. C'est ce type d'interconnexion qui est utilisé pour connecter les différents opérateurs de l'Internet.

Au sein de chaque AS, les routeurs BGP internes sont interconnectés en iBGP (pour *internal BGP*). Ce type de session assure, dans le réseau interne, les échanges de l'ensemble des routes apprises par le biais de l'eBGP ainsi que des routes spécifiques aux besoins internes de l'AS.

La figure 2 représente différentes interconnexions eBGP établies entre différents AS. En pratique, une interconnexion BGP s'effectue entre deux AS. Dans cet exemple, le routeur de l'AS 64496 annonce les préfixes qu'il gère au routeur de l'AS 64497 (appelé pair de l'AS 64496 ou voisin eBGP),

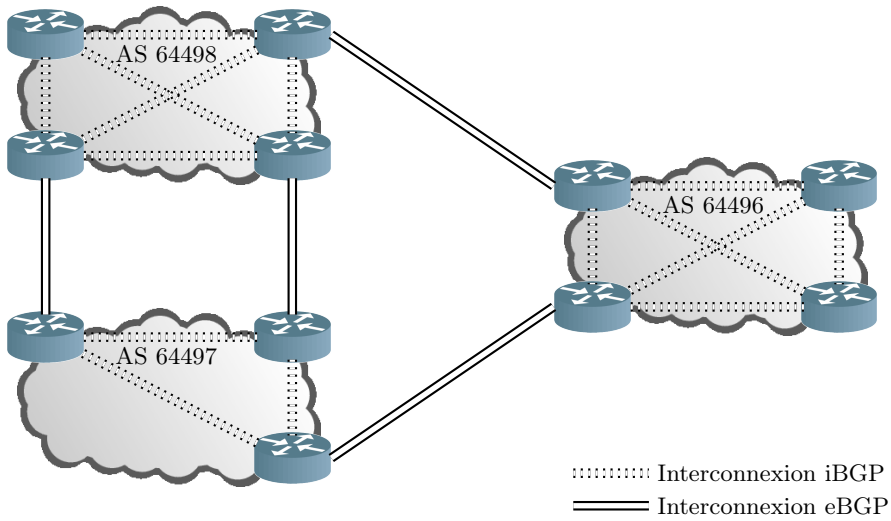


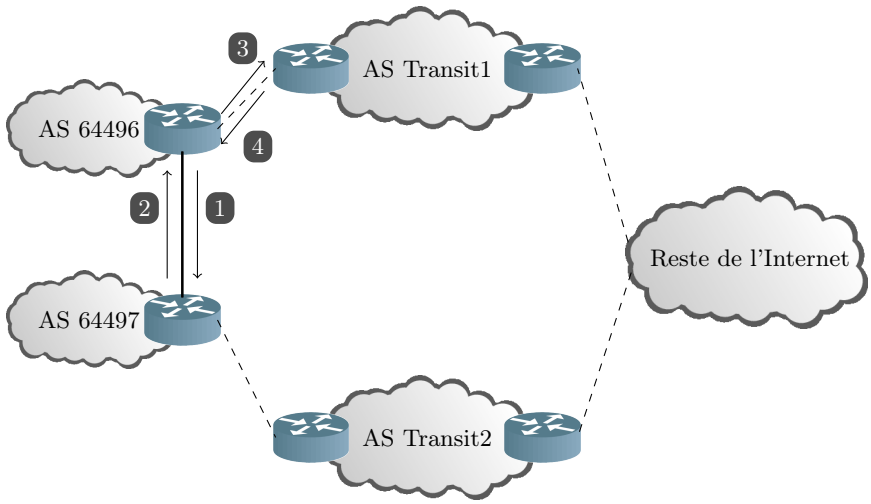
Figure 1. Exemple d'interconnexions eBGP et iBGP.

et réciproquement. Une annonce indique à son pair BGP qu'il sait acheminer le trafic à destination des préfixes listés. Les interconnexions se découpent en deux catégories :

1. Le peering : c'est un accord réciproque où chaque interlocuteur annonce exclusivement à l'autre les préfixes qu'il gère et ceux de ses clients. Sur la figure 2, les AS 64496 et 64497 sont en relation de peering.
2. Le transit : accord entre un client et un fournisseur de transit. Le client annonce les préfixes qu'il gère, tandis que l'opérateur de transit lui annonce le reste des préfixes constituant l'Internet. Sur la figure 2, les interconnexions de transit sont représentées en pointillés.

Une route est composée d'un préfixe destination ainsi que d'attributs, dont le chemin d'AS (en anglais, *AS PATH*). Dans la figure 1, l'AS 64496 connaît deux chemins d'AS différents pour joindre les préfixes de l'AS 64497 : `64498 64497` [3] (via une interconnexion de transit) et `64497` (via l'interconnexion de peering). Pour l'AS 64496, les chemins d'AS associés au préfixe 203.0.113.0/24 annoncé par l'AS 64497 représente la liste des AS qui seront successivement traversé par un paquet émis par une machine de l'AS 64496 à destination d'une machine appartenant au préfixe 203.0.113.0/24.

3. l'AS gérant le préfixe se situe à droite du chemin d'AS.



1 AS 64496 annonce à AS 64497 ses préfixes
2 AS 64497 annonce à AS 64496 ses préfixes
3 AS 64496 annonce à AS Transit1 ses préfixes
4 AS Transit1 annonce à AS 64496 les préfixes de l'Internet

Figure 2. Interconnexions d'AS en BGP.

1.2 Les messages BGP

Le protocole BGP utilise le port TCP 179. En règle générale, une interconnexion est effectuée entre deux routeurs directement connectés, et demeure stable dans le temps. En pratique, la mise en place d'une interconnexion BGP entre deux pairs nécessite, sur chaque routeur, les éléments de configuration suivants :

1. le numéro d'AS local,
2. le numéro d'AS distant,
3. l'adresse IP du pair.

A l'aide de la syntaxe de l'IOS de Cisco, la figure 3 présente la création d'une interconnexion BGP entre l'AS local (64498), le pair (192.0.2.1) de l'AS distant (64496) sur un routeur de l'AS 64498.

```
router bgp 64498
  neighbor 192.0.2.99 remote-as 64496
```

Listing 1.1. Exemple d'interconnexion BGP.

Le protocole BGP est constitué de quatre messages servant à l'établissement de session, son maintien, à l'échange de données et à la fermeture de session. Voici leur fonctionnement détaillé :

1. **OPEN** : ce message est envoyé par un routeur cherchant à communiquer avec son homologue. À la réception de ce message, ce dernier émettra également un message OPEN pour valider l'établissement de la session BGP. Ce message permet à un routeur d'annoncer son numéro d'AS, son identifiant, ainsi que les capacités supportées (comme le support d'IPv6).
2. **UPDATE** : ce message contient la liste des préfixes qu'un routeur souhaite annoncer (ou dont il souhaite supprimer l'annonce) à son pair ainsi que les différents attributs associés comme le chemin d'AS, le *next hop*, ou l'origine de la route (eBGP ou iBGP).
3. **KEEPALIVE** : message visant à maintenir la session. Par défaut, un routeur Cisco envoie un message KEEPALIVE toutes les 60 secondes. La session est rompue si aucun message KEEPALIVE n'a été reçu durant 180 secondes.
4. **NOTIFICATION** : message envoyé par un pair BGP pour signaler à son interlocuteur qu'une erreur a eu lieu. La session est immédiatement coupée suite à l'émission de ce message.

1.3 Les incidents touchant le protocole BGP

Différents problèmes peuvent toucher le protocole BGP :

- **La disparition de préfixes :** un préfixe peut disparaître de l'Internet si son émetteur supprime l'annonce de ce préfixe auprès de ses interlocuteurs BGP. La conséquence directe de cette disparition est la non disponibilité des réseaux concernés. Ce genre d'incident est généralement imputable à une erreur humaine.
- **L'usurpation de préfixes** [4] **:** étant donné qu'il n'y a aucune authentification des annonces de préfixes, il est possible d'annoncer à son fournisseur un préfixe appartenant à un autre réseau. Les conséquences de ce genre d'incidents peuvent être plus ou moins graves selon l'annonce qui est faite. Le réseau victime peut ainsi devenir injoignable pour tout ou partie de l'Internet. Ce type d'incidents peut également entraîner une redirection du trafic destiné au réseau victime vers le réseau ayant usurpé les préfixes.

4. en anglais, *hijack*.

– **Un bug logiciel sur un équipement :** les routeurs BGP peuvent faire l'objet d'erreur d'implantation. Ces bugs sont susceptibles d'avoir différents niveaux d'incidence sur le réseau Internet, allant jusqu'à la coupure de la session BGP et l'arrêt de l'utilisation du lien touché. Du fait de l'impact qu'ils peuvent avoir sur le réseau mondial, les constructeurs sont extrêmement réactifs quant à la recherche de bugs et la mise à disposition de correctifs.
– **Une attaque en déni de service** visant les routeurs BGP. Ce type d'attaque n'est pas lié au protocole BGP mais elle peut, si elle est menée via un réseau à grande capacité de débit, sous la charge de traitement qu'elle représente pour l'équipement BGP ciblé, induire une mise hors service de ce dernier. Ce type d'incident n'a pas été constaté pour l'instant mais doit être pris en compte.

2 Le protocole BGP et la sécurité

Cette section décrit des configurations de routeurs permettant de protéger les interconnexions BGP et filtrer les annonces reçues. Pour illustrer les exemples, la syntaxe de l'IOS de Cisco a été utilisée. Des configurations similaires peuvent être réalisées sur les routeurs d'autres constructeurs.

2.1 Sécurité des sessions BGP

Une session BGP s'établit en clair et se base sur des informations dont l'usurpation est possible dans certains contextes (comme le numéro d'AS et l'adresse IP). Mis à part de rares exceptions, une session eBGP est mise en place entre deux routeurs directement connectés. Afin d'écarter toute injection de données via un paquet forgé en amont du réseau, la plupart des constructeurs de routeurs implémentent le *TTL security*. Étant donné qu'un paquet BGP ne passe normalement à travers aucun routeur, son TTL[5] est seulement décrémenté par le routeur le recevant. L'exemple 4 permet donc de rejeter silencieusement tout paquet transportant des messages BGP dont le TTL n'est pas de 254.

```
router bgp 64496
  neighbor 192.0.2.1 remote-as 64497
  neighbor 192.0.2.1 ttl-security hops 1
```

Listing 1.2. Exemple de configuration du *TTL security*.

5. pour *Time To Live*.

Un mécanisme d'authentification, appelé TCP MD5 [14] basé sur un secret partagé est disponible sur la plupart des implantations du protocole BGP, comme le montre l'exemple 4. Lors de la spécification de cette extension, la prédiction des numéros de séquence TCP des routeurs était une menace importante [23]. Un attaquant ayant la possibilité d'usurper un routeur et de deviner le numéro de séquence TCP pouvait alors injecter des messages UPDATE pour annoncer ou supprimer des routes, ou des messages NOTIFICATION pour stopper la session. Avec ce mécanisme, chaque segment TCP émis est protégé en intégrité à l'aide de la fonction de hachage MD5 et du secret partagé. Cela permet au récepteur de vérifier que le segment TCP a bien été émis par un routeur connaissant le secret.

```
router bgp 64496
  neighbor 192.0.2.1 remote-as 64497
  neighbor 192.0.2.1 password m07d3p4s5
```

Listing 1.3. Exemple de configuration du mécanisme *TCP MD5*.

Le protocole de routage BGP s'établit à travers une relation de confiance entre les opérateurs. Actuellement, il n'existe pas de mécanisme de signature permettant de vérifier si un message UPDATE annonçant des préfixes est légitime. Un routeur BGP est toutefois capable d'appliquer des filtres sur les routes annoncées et/ou reçues, sur le même principe qu'une règle de pare-feu. Un transitaire peut par exemple n'accepter que les routes appartenant réellement à ses clients, empêchant ainsi les usurpations de préfixes. Afin d'automatiser la gestion et l'utilisation de ces filtres, la RFC 2622 [6] a défini le RPSL [6] dont le but est de définir auprès de son IRR [7] l'ensemble de la politique de routage de son AS. Un exemple de filtres de préfixes est donné dans la section 2.3.

2.2 Détails sur les effets du max-prefix sur les sessions eBGP

Lorsqu'une session eBGP est établie, sans aucun filtre, chacun des pairs peut annoncer tous les préfixes qu'il souhaite sans restriction. En pratique, un AS possédant seulement quatre préfixes n'a pas de raison d'en annoncer plus.

La configuration d'un nombre maximum de préfixes échangés par session eBGP est donc le mécanisme principal configuré sur toutes les sessions externes pour isoler rapidement des débordements. Selon les options de configuration, ce mécanisme déclenche automatiquement l'envoi d'un

6. pour *Routing Policy Specification Language*.
7. pour *Internet Routing Registry*.

message NOTIFICATION, puis la rupture de session BGP lorsque le pair annonce un nombre de préfixes supérieur au seuil fixé. En général, le seuil est atteint suite à une erreur humaine, c'est-à-dire une configuration incorrecte entraînant la réannonce d'une table de routage complète vers le pair. La fermeture de la session BGP protège le pair à la fois sur le plan de contrôle [8] et le plan de transfert [9] (congestion évitée par reroutage du trafic via d'autres pairs BGP encore fonctionnels).

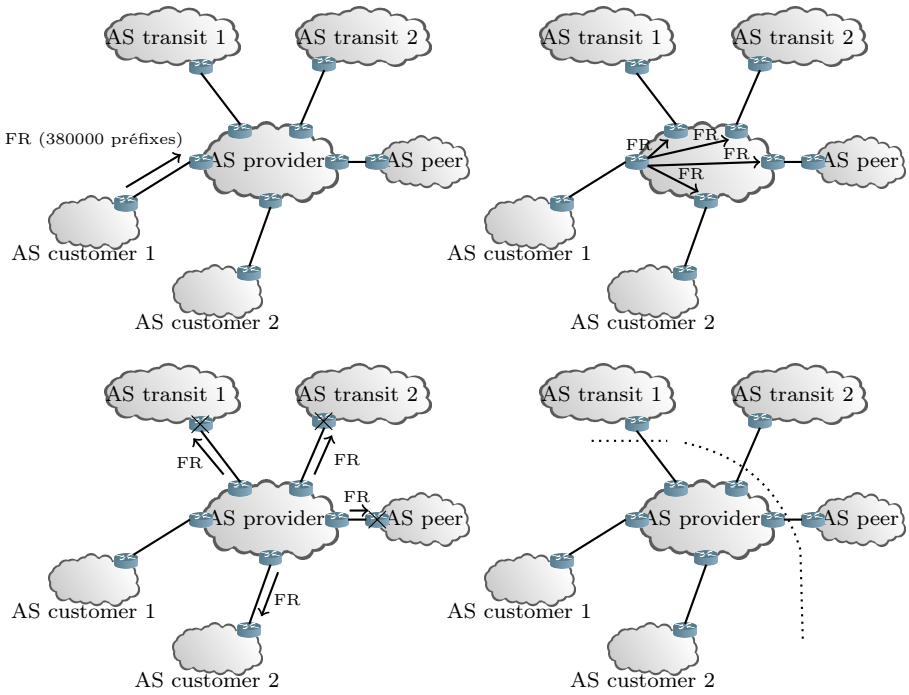


Figure 3. Illustration de l'effet de max-prefix.

Cas concret (figure 3) : L'interconnexion entre les deux AS (ASCustomer - ASProvider) est de type client-transitaire. Le client réannonce par erreur la table de routage Internet (figure 3, en haut à gauche). Sans la configuration de *max-prefix*, le transitaire (appelé ASProvider dans la figure) accepte les 380 000 préfixes constituant l'Internet (FR dans la figure

8. partie d'un routeur traitant le protocole BGP et le calcul des tables de routage.
9. partie du routeur décidant de la route et de l'interface de sortie d'un paquet.

3). Sans filtres sur les préfixes clients, et puisque les routes des clients sont en général préférées aux routes des pairs et des transitaires, le routeur déroule l'algorithme de sélection des meilleurs chemins sur ces routes et les élit comme *BEST*.

Le routeur réannonce alors via iBGP ces 380 000 routes (figure 3, en haut à droite) à l'ensemble des routeurs de l'ASProvider avec des attributs préférentiels. Ces routeurs déroulent l'algorithme de sélection des meilleures routes BGP et les élisent également comme *BEST*[10]. Ils les réannoncent alors à l'ensemble de leurs pairs eBGP : pairs, clients et transitaires (figure 3, en bas à gauche).

Heureusement, tous ses pairs ont configuré une valeur de *max-prefix* sur les sessions eBGP vers l'ASProvider, et coupent au fur et à mesure les sessions eBGP vers l'ASProvider, ce qui empêche la propagation de ces routes erronées (figure 3, en bas à droite). Le réseau ASProvider est peu à peu isolé de l'Internet ; il lui faut à corriger les erreurs de configuration pour rétablir de nouvelles sessions eBGP cohérentes et retrouver sa connectivité initiale.

```
router bgp 64496
 neighbor 192.0.2.1 remote-as 64497
 neighbor 192.0.2.1 maximum-prefix 20 50 restart 5
```

Listing 1.4. Exemple de configuration de la règle *max-prefix*.

Dans l'exemple 4, le routeur de l'AS 64497 portant l'adresse IP 192.0.2.1 ne peut envoyer plus de 20 préfixes sur la session eBGP décrite. Le deuxième argument (50) renseigne à partir de quel seuil, exprimé en pourcentage, une notification doit être envoyée par le biais d'un *trap* SNMP. Enfin, la mention *restart 5* indique que la session eBGP fera l'objet d'une nouvelle tentative d'établissement toutes les cinq minutes. Si le nombre de préfixes annoncés par le pair dépasse toujours le seuil fixé, alors la session sera à nouveau coupée, cinq minutes s'écouleront avant une nouvelle tentative de rétablissement de la session.

2.3 Filtres sur le plan de contrôle

Filtres sur les AS origines, chemins d'AS, et listes de préfixes Ces mesures sont les plus efficaces pour lutter contre les attaques de type

10. dès cet instant, le réseau ASProvider écoule l'ensemble de son trafic Internet via l'interconnexion avec le client, ce qui génère vraisemblablement une congestion.

usurpation de préfixes ou réannonce de tables de routage, mais elles sont très complexes à maintenir dans le temps pour une entité de type opérateur de transit et échangeant des routes avec de nombreux pairs.

```
ip as-path access-list 1 permit ^64497(_64497)*$

route-map filtrage-as permit 10
 match as-path 1

router bgp 64496
 neighbor 192.0.2.1 remote-as 64497
 neighbor 192.0.2.1 route-map filtrage-as in
 bgp maxas-limit 50
```

Listing 1.5. Exemple de configuration de filtre sur AS l'origine.

L'exemple 10 présente un exemple de filtre basé sur l'AS origine. En premier lieu, il convient de déclarer un filtre, ou *access-list*, sur un chemin d'AS. L'*access-list 1* sur l'as-path précise que l'AS d'origine doit être l'AS 64497, sur le principe des expressions rationnelles. Il peut être répété ou non : en allongeant artificiellement un chemin d'AS, l'AS origine affaiblit cette route pour faire de l'ingénierie de trafic par exemple et signaler un chemin de secours.

L'application de l'access-list se réalise via une *route-map*, nommée dans cet exemple «filtrage-as», qui permet d'appliquer des modifications aux préfixes reçus (diminution de la priorité, changement de *next-hop*, etc.). La *route-map* est ensuite renseignée dans la configuration, au niveau de la déclaration BGP de l'interconnexion avec le pair. Il est spécifié dans cet exemple que la route-map s'applique sur les préfixes reçus en entrée (mot-clé «in»), donc de la part de l'AS 64497. De plus, la ligne *bgp maxas-limit 50* indique que le nombre maximum d'AS dans les chemins d'AS ne devra pas excéder 50. Toute route BGP ne correspondant pas aux critères ci-dessus sera immédiatement refusée par le routeur.

```
ip prefix-list prefixes-acceptes seq 5 permit 198.18.0.0/15 le 24

route-map filtrage-prefixe permit 10
 match ip address prefix-list prefixes-acceptes

router bgp 64496
 neighbor 192.0.2.1 remote-as 64497
 neighbor 192.0.2.1 route-map filtrage-prefixe in
```

Listing 1.6. Exemple de configuration sur des prefixes.

Afin de filtrer des préfixes, la configuration de l'exemple 10 peut être appliquée. Il faut tout d'abord déclarer une *prefix-list*. Ici la *prefix-list*

«prefixes-acceptes» renseigne tous les préfixes qui seront acceptés par cette règle. Cette *prefix-list* acceptera tout préfixe compris ou égaux à 198.18.0.0/15 et dont le masque de sous réseau sera égal ou inférieur à 24. La *prefix-list* est ensuite appliquée à une *route-map* nommée «filtrage-prefixe». La mise en place de ce filtrage se réalise au niveau de l'interconnexion BGP avec l'AS 64497, en exprimant de manière explicite que la *route-map* «filtrage-prefixe» s'applique à tous les préfixes reçus par cette interconnexion BGP avec le pair 192.0.2.1.

Filtres sur les propriétés de certains attributs (e.g. longueur de chemins d'AS, masques des routes) Ces filtres ont notamment pour objectif d'empêcher la propagation de routes malformées. L'exemple 10 montre la mise en place d'un filtre validant la longueur de l'AS_PATH. En général, la valeur configurée est très large, de l'ordre de 50 ou 75, même si un chemin d'AS est composé de moins d'une dizaine d'AS différents au maximum. Dans les annonces BGP échangées sur Intenet en mars 2012, le chemin d'AS le plus long est de 32. En moyenne, les chemins d'AS sont compris entre 3 et 4.

3 Impacts des messages malformés sur les implantations

3.1 Messages UPDATES malformés

Les piles BGP peuvent avoir des réactions très diverses sur la réception d'un message de type UPDATE BGP mal géré par un routeur. Le paquet peut être détecté comme ayant un attribut malformé, la route n'est alors pas prise en compte pour la sélection des meilleurs chemins et aucune conséquence supplémentaire n'est observée, pas même de génération de *traps* SNMP pour remonter l'erreur.

Dans certains cas, le traitement de l'UPDATE peut, par ordre de gravité, provoquer : un arrêt inopiné du processus BGP, l'arrêt du processus de routage complet du routeur, voire le redémarrage complet du routeur. La route comportant un attribut mal géré peut également être élue comme *BEST* par l'algorithme de sélection, et être réannoncée avec son attribut inchangé à ses pairs eBGP et iBGP en accord avec les politiques de routage, propageant les risques de pannes aux autres nœuds du réseau.

Certains routeurs peuvent être sensibles à des séquences d'UPDATEs BGP, auquel cas un UPDATE unique, ou une séquence correcte d'UPDATEs, ne génèrent pas de panne rendant le diagnostic difficile. Ce fut

ainsi le cas de l'incident du 7 novembre 2011 détaillé dans la section suivante. Sur les cas de panne les plus simples, la mise en place de filtres a été suffisante pour résorber les pannes. Dans le cas des UPDATEs avec des chemins d'AS longs, par chance le paquet est rejeté par le filtre avant traitement, ce qui permet d'éviter les redémarrages du routeur ou des sessions BGP.

En revanche, l'envoi de notifications sur détection de l'attribut malformé est dépendant de l'implantation du constructeur : il est possible que les routeurs forcent l'arrêt de la session BGP, isolant le réseau de l'opérateur des autres réseaux. En effet, les routeurs suivant scrupuleusement les normes doivent remettre à zéro la session BGP sur réception d'un UPDATE contenant un attribut malformé [25]. Or certaines piles peuvent ne pas respecter cette assertion, ou encore ne pas considérer un attribut comme malformé à cause des zones d'ombre des normes ou d'une erreur d'implantation.

Les graphes de la figure 4 montrent l'impact sur le plan de transfert de trois liens avec un réseau tiers, dont les routeurs ont été victimes d'un défaut d'implantation. Le 27 août 2010, sur réception de routes avec un attribut 99 (détaillé dans la section suivante), ces pairs BGP ont coupé leurs sessions, isolant l'AS pendant plus d'une trentaine de minutes. Des remontées temporaires de sessions ont été observées mais une nouvelle réception de l'attribut 99 les faisait automatiquement cesser, amplifiant l'instabilité. On voit clairement que la totalité du trafic n'a pas été reprise sur les autres liens.

Cette gestion simple et robuste des pannes se révèle finalement catastrophique, car elle génère isolement et instabilité sur l'Internet. Notons que dès que la session redémarre, il est fort possible que les mêmes routes soient traitées et renvoyées : mêmes causes, mêmes effets. Devant la disparité des traitements et la multiplication des incidents, un effort de normalisation de la gestion des messages d'erreurs BGP a été finalisé en 2011-2012 [24,18]. D'autres efforts se concentrent sur l'accélération de la récupération d'information BGP après la fermeture d'une session [19,10].

3.2 Protection des services de l'opérateur des attributs malformés

Une session BGP est multi-protocolaire : lors de la phase d'établissement, les messages OPEN signalent au pair, via des capacités, les diffé-

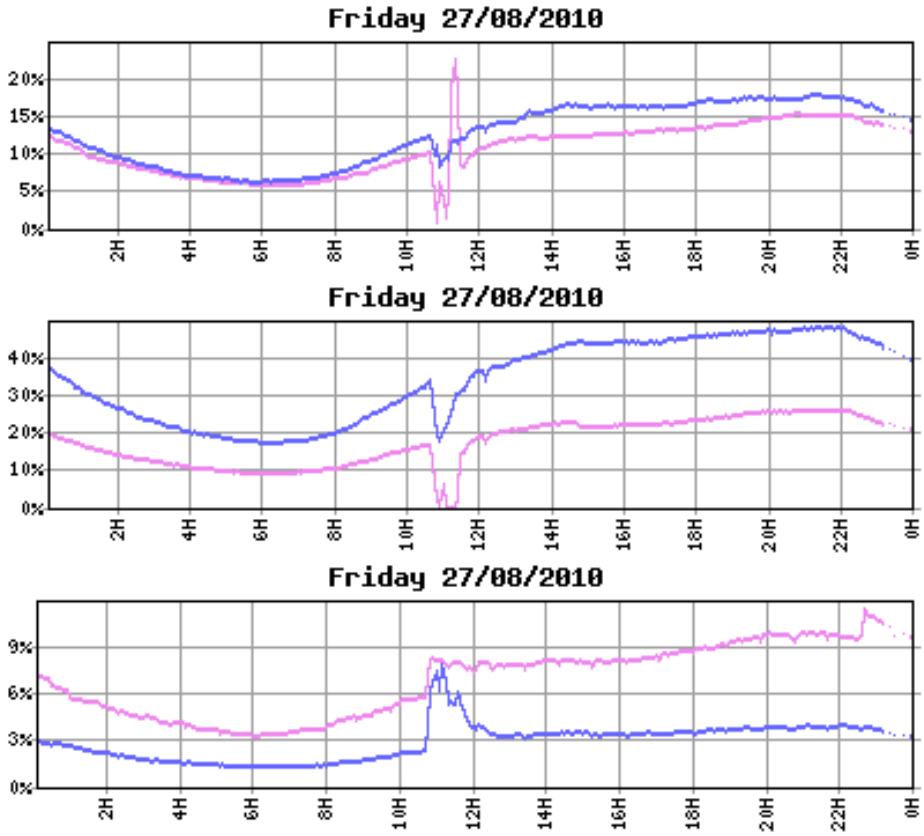


Figure 4. Impact du bug attribut 99 sur les liens de peering.

rentes familles et sous-familles d'adresses concernées par les échanges de routes (par exemple IPv4 unicast, IPv6 unicast, VPN IPv4 unicast, IPv4 multicast). Pour limiter les propagations de pannes aux autres nœuds ainsi que les contagions des pannes Internet aux services internes (VPN MPLS, IPTV, VoIP interne), l'architecture BGP doit être définie avec soin. Il est ainsi possible d'appliquer les bonnes pratiques suivantes :

- éviter le multiplexage des échanges de routes de familles et sous-familles d'adresses différentes sur la même session BGP (e.g. IPv4 et IPv6 unicast pour les routes Internet d'une part, VPN unicast et multicast d'autre part pour éviter la perte des routes VPN lors de fermeture de sessions BGP due à des routes IPv4 malformées) ;
- dédier par exemple des équipements aux échanges BGP Internet sur la périphérie du réseau et libérer le cœur de BGP avec une commutation MPLS ; les routeurs de bordure peuvent même n'avoir que des routes par défaut vers des points d'évasion Internet, ce qui leur évite de gérer des routes potentiellement néfastes.
- avoir une stratégie équipements fondée sur plusieurs constructeurs, en espérant que les implantations des différents constructeurs n'aient pas le même défaut. Bien sûr, dans un monde idéal, il faudrait également que les pairs eBGP du côté des transitaires IP restent opérationnels sur réception des UPDATEs malformés, ce qui est en pratique impossible à vérifier.

4 Les incidents BGP depuis 10 ans

Depuis la première édition du SSTIC, un certain nombre d'incidents BGP ont été médiatisés. Ils sont principalement de deux types : bugs d'implantation logicielle et usurpation de préfixes. La mémoire collective de l'Internet permet de retrouver la trace des incidents ayant eu ces dix dernières années, et dont les impacts ont été discutés publiquement : la liste n'est donc pas exhaustive.

2004 En mai, l'hébergeur malaisien DataOne a usurpé deux préfixes de Yahoo [21]. Le 24 décembre, le fournisseur d'accès turque TTNET a réannoncé l'ensemble des routes de l'Internet provoquant ainsi des problèmes de routage globaux [26]. Au cours de ces deux incidents, le trafic à destination des préfixes usurpés a été redirigé vers les AS à l'origine de ces problèmes.

2005 Le 7 mai, le transitaire Cogent a annoncé un préfixe de Google [28]. En raison de cette annonce erronée, les services offerts par Google ont été partiellement indisponibles entre 15 et 60 minutes suivant les AS touchés.

2006 Le 22 janvier, la société américaine Con Edison a annoncé des préfixes à la place de ses clients et d'autres AS sans relation avec eux [27]. Le trafic des AS touchés a été redirigé vers le réseau de Con Edison.

2008 Le 24 février, sur ordre du gouvernement pakistanais, le fournisseur Pakistan Telecom a annoncé à son transitaire, PCCW, des préfixes plus spécifiques que ceux annoncés par YouTube [22]. Il est important de noter que lorsqu'une table de routage contient un réseau et un sous-réseau plus spécifique, i.e. dont le masque est plus long, les routes correspondantes aux préfixes les plus spécifiques l'emportent. Le transitaire n'ayant pas mis en place de filtre sur les annonces de Pakistan Telecom, une partie du trafic de YouTube a ainsi été redirigé chez Pakistan Telecom.

Le 11 novembre, le fournisseur brésilien CTBC a réanonncé l'intégralité de la table de routage de l'Internet a ses pairs BGP [16,9]. Cet incident est original du point de vue des impacts : les systèmes de filtrage et de détection d'usurpation ont correctement fonctionné et seul le trafic des clients de CTBC a été perturbé.

2009 Le 16 février, une annonce de préfixe émise depuis un opérateur tchèque, SuproNet, à l'aide d'un routeur MikroTik a provoqué des arrêts massifs de routeurs Cisco [12,13]. Étant donné la part de marché de Cisco, cela a engendré une grosse perturbation des annonces de préfixes sur tout l'Internet. Le routeur MikroTik, à l'origine de l'incident, a annoncé des préfixes avec des chemins d'AS très longs. Au fur et à mesure de la propagation de ces annonces dans l'Internet, ces chemins ont rapidement atteint une taille de 255 AS déclenchant un bug alors inconnu dans des routeurs Cisco éloignés de SuproNet, et la coupure des sessions BGP.

2010 Le 8 avril, l'opérateur China Telecom a annoncé 50 000 préfixes ne lui appartenant pas [17]. Une partie du trafic à destination de ces préfixes a donc été redirigé vers China Telecom.

Le 27 août, des tests préliminaires d'une implantation de secure BGP a engendré une perturbation de l'Internet pendant 30 minutes environ [11]. Dans le cadre de cette expérience, un nouvel attribut BGP, de type 99, a été annoncé. Inconnu des implantations BGP déployés, cet attribut

aurait dû être transféré par les routeurs BGP sans aucune conséquence. Cependant des routeurs Cisco utilisant IOS XR [8] ont corrompu l'attribut avant de le retransmettre à leurs pairs [29]. À la réception de cet attribut corrompu, ces pairs ont donc stoppé leurs sessions BGP, provoquant ainsi la perturbation. Cet incident est publiquement connu sous le nom «bug attribut 99».

2011 Le premier décembre, des routeurs Redback ont subitement coupé leurs sessions avec certains de leurs pairs BGP. Une analyse [15] des paquets reçus a montré qu'il s'agissait de messages BGP de type UPDATE comprenant des numéros d'AS nuls. Le traitement d'un numéro AS nul n'étant pas correctement détaillé dans les spécifications du protocole BGP, les concepteurs des routeurs Redback ont considéré que ce comportement était anormal, et par conséquent décidé de stopper les sessions. Suite à cet incident, un document a été produit par l'IETF [20] préconisant de journaliser la réception de numéro d'AS nul sans couper la session BGP.

Le 7 novembre, un autre incident a touché certains routeurs Juniper. Nous le décrivons en détails dans la section 4.1.

2012 Le 23 février, le fournisseur d'accès australien Telstra a été isolé d'Internet pendant une demi-heure environ [5]. L'un de ses clients, Dodo, lui a annoncé l'intégralité des routes de l'Internet. Aucun filtre n'étant présent sur le routeur BGP de Telstra, celui-ci a appris ces routes puis les a réannoncées en iBGP. Telstra a donc redistribué la table de routage de l'Internet apprise par son client à ses propres pairs. Le mécanisme de protection appelé *max-prefix*, détaillé dans la section 2.2, mis en place par les pairs de Telstra a bloqué ces annonces provoquant ainsi l'isolation de Telstra. Cet incident est similaire à l'incident ayant touché CTBC en 2008.

4.1 Retour sur l'incident Juniper du 7 novembre 2011

L'incident ayant touché des routeurs Juniper le lundi 7 novembre [7] demeure le plus spectaculaire de l'année 2011 car il a concerné des transitaires français et internationaux importants. Durant plusieurs dizaines de minutes, ces transitaires n'ont pas été en mesure d'acheminer correctement le trafic de leurs clients.

Cet incident a été fortement médiatisé, cependant les origines exactes du problème ne sont pas connues et l'on ne sait pas le reproduire. Les

routeurs affectés par l'incident sont les Juniper de type MX avec le chipset Trio sous Junos 10.x sortis avant août 2011. Le déclencheur serait une séquence de messages BGP légitimes de type UPDATE qui provoquent une erreur dans le module MPC [11], et par conséquent le redémarrage des cartes de ligne. Pour un pair BGP, la coupure des liens durant l'incident est perçue comme une disparition de son homologue : le chemin entre les deux routeurs n'est plus actif, et le routeur coupe la session BGP.

Le 8 août 2011, Juniper a publié un avis de sécurité non public présentant ce problème et les mises à jour le corrigeant. Le constructeur avait par ailleurs conseillé à ses clients de mettre à jour leurs routeurs. Cet incident était décrit comme ayant peu de risque de se produire car la réception du message BGP de type UPDATE n'est pas la seule cause du problème. L'avis indique ainsi que seule une combinaison complexe de messages BGP précédant cet UPDATE engendre une erreur dans le module MPC qui redémarre seul et sans intervention d'un administrateur.

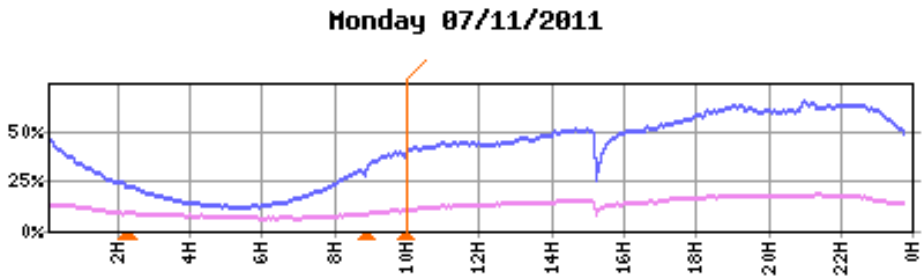


Figure 5. Diminution de trafic vers un transitaire. L'axe des ordonnées correspond à la capacité du lien : 100% étant le maximum.

Les archives BGP publiques collectées à Londres et à Amsterdam permettent d'identifier plusieurs phénomènes intéressants concernant deux transitaires ayant publiquement annoncé avoir été victimes de cet incident.

1. Le nombre total de préfixes vus à Londres et à Amsterdam est resté constant durant l'incident. Un nombre réduit de pairs a cependant supprimé des annonces de préfixes. Malgré l'arrêt de certains sessions,

11. en anglais, Modular Port Concentrators.

la structure de l'Internet et le protocole BGP ont permis de maintenir les routes ;

2. une diminution de trafic a toutefois été perçue sur les liens avec les transitaires touchés, en conséquence des changements d'annonces de préfixes. La figure 5 met en évidence cette diminution en entrée (ligne du haut) et en sortie (ligne du bas) d'un lien.

Les constatations liées à cet incident permettent de mettre en évidence le problème de dépendance vis-à-vis d'un équipement unique à la fois pour les opérateurs et les transitaires.

5 Bilan de trois années de surveillance des annonces BGP sur un réseau opérationnel

La surveillance des annonces BGP liées à l'AS 3215 a commencé fin 2008. Cette section vise à présenter un bilan de la surveillance mise en place à travers différents outils : BGPmon [1], Cyclops [2], le *Routing Information Service* [4], et l'*Internet Alert Registry* [3]. Les critères surveillés sont :

1. les blocs d'adresses sous responsabilité de l'AS 3215 : toute annonce de préfixes identiques ou de préfixes plus spécifiques par d'autres AS dits «AS origines» est suspecte ;
2. l'AS reconnu comme transitaire pour les préfixes de l'AS 3215 (AS dit *upstream*, second numéro d'AS constaté dans les chemins d'AS) ;
3. les annonces ayant pour AS origine l'AS 3215 (nouveau préfixe apparu sur les sondes).

Entre 2009 et 2011, seuls quelques incidents avec des impacts variables ont été constatés. Divers cas sont présentés dans cette section.

En août 2009, un AS tiers a annoncé un préfixe /24 appartenant à un bloc d'adresses sous la responsabilité de l'AS 3215. Ce bloc n'est habituellement pas annoncé sur l'Internet : ces préfixes ne sont pas susceptibles de recevoir du trafic depuis d'autres réseaux externes. Il n'est donc pas possible de détourner du trafic légitime depuis l'Internet. En outre, ce préfixe est désagrégé et inclus dans les tables de routage des protocoles internes (IGP [12]). Pour un préfixe réseau donné, lorsqu'une table de routage contient des routes issues de plusieurs protocoles différents comme

12. pour *Interior Gateway Protocol*.

par exemple eBGP et l'IGP, un mécanisme de préférence de route intervient : les routes apprises via les protocoles internes sont en général privilégiées sur les routes apprises via eBGP. Ainsi, les routes internes étant préférées aux routes reçues de l'extérieur, le trafic interne n'a pas non plus été perturbé. Aucun impact sur les services n'a été constaté.

L'opérateur tiers a été contacté. Aucune réponse n'a été reçue mais les annonces illégitimes ont stoppé après environ 48h.

Le second incident s'est produit en juin 2010, lorsqu'un opérateur a annoncé par erreur un préfixe /22 faisant parti d'un bloc /16 de l'AS 3215. Cette plage d'adresses est utilisée pour le service Internet résidentiel.

```
======================================================================
Possible Prefix Hijack
======================================================================
Your prefix:          92.142.0.0/16:
Update time:          2010-06-03 11:15 (UTC)
Detected by #peers:   56
Detected prefix:      92.142.8.0/22
----------------------------------------------------------------------
```

Listing 1.7. Alerte d'usurpation de préfixe fournie par le service BGPmon [1].

Grâce aux alertes remontées par les outils de surveillance BGP, figure 9, la réaction des équipes d'exploitation a été très rapide :

- d'une part, pour annoncer des préfixes plus spécifiques (92.142.8.0/23 et 92.142.10.0/23) afin de rediriger le trafic à destination de ces blocs vers l'AS 3215 ;
- d'autre part, pour contacter le NOC de l'AS ayant usurpé le préfixe afin de signaler et résoudre le problème d'annonce.

Suite à cet échange avec le NOC, l'opérateur tiers a en quelques minutes corrigé la faute de frappe lors de la mise en place d'une nouvelle annonce (erreur humaine). Aucun problème chez les clients lié à cet incident n'a été remonté.

D'autres alarmes ont mis en évidence des erreurs de configuration :

- Un AS a réannoncé un sous-ensemble des préfixes de l'AS 3215 avec son propre AS en origine (erreur de type «usurpation de préfixe»). Cela n'a pas porté à conséquence côté service car le trafic était ensuite correctement réacheminé vers les destinations finales.
- Un AS a réannoncé l'ensemble des préfixes 3215 avec son propre AS en origine : dans ce cas on peut rapidement vérifier que l'opérateur a en réalité tenté de réannoncer l'ensemble de la table de routage de

l'Internet. Ces incidents sont en général de courte durée (occurrences également les 08/04/2010, 14/01/2011, 21/10/2011) car le réseau fautif est isolé au fur et à mesure des débordements d'annonces comme décrit dans la section 2.2.

- Un préfixe plus spécifique et habituellement non annoncé aux pairs BGP a été «leaké» lors d'une maintenance sur un routeur de l'AS 3215 (erreur de type «usurpation d'un nouveau préfixe») ; les alertes remontées par les outils de surveillance BGP ont accéléré la détection des erreurs de configuration et donc leur correction.
- Un AS, voisin de l'AS 3215, a réannoncé l'ensemble des préfixes de l'AS 3215 ainsi que d'autres destinations vers ses transiteurs. Il s'est alors mis en position de transitaire pour les blocs de l'AS 3215 pour tout l'Internet au lieu de se limiter à ses pairs. La surveillance des alertes a mis en évidence cet événement : erreur de type «changement de chemin d'AS». Notons que ce type d'erreur est toujours de courte durée car l'AS fautif devient d'une part rapidement congestionné, et d'autre part isolé par la tombée successive des sessions eBGP avec ses voisins (seuil du «max-prefix» atteint).

Depuis 2011 en revanche, la fréquence des alertes remontées pour les possibles usurpations de préfixes a considérablement augmenté. En effet, la plage 2.0.0.0/12 a été progressivement mise en production sur le réseau. Ce bloc d'adresses IP était normalement vierge de toute utilisation, mais avant même le déploiement, et par de simples recherches dans l'historique de routage de ce préfixe, il est devenu clair que les exploitants rencontreraient de nombreux contretemps.

D'une part, un important trafic a été reçu à destination de ces plages d'adresses dès la mise en place du routage, alors qu'aucun service n'utilisait effectivement ces adresses. Ces pics sont toujours constatés, avec des sources et des trafics variés, même sur des adresses non affectées à des machines finales. Le graphe 6 présente par exemple 15 jours de trafic (en nombre de paquets par seconde) vers l'adresse 2.3.4.5.

D'autre part, comme le montre la figure 7, de nombreux enregistrement DNS pointaient vers l'adresse 2.2.2.2 alors quelle n'a jamais été allouée par un RIR [13].

En une année, il y a ainsi eu 18 occurrences d'incidents de type usurpation de préfixes pour lesquels un AS tiers a illégitimement annoncé un préfixe plus spécifique du 2/12. Des annonces de «/32», préfixes normale-

13. pour *Regional Internet Registry*. Organisme en charge de l'attribution des blocs d'adresses IP aux opérateurs.

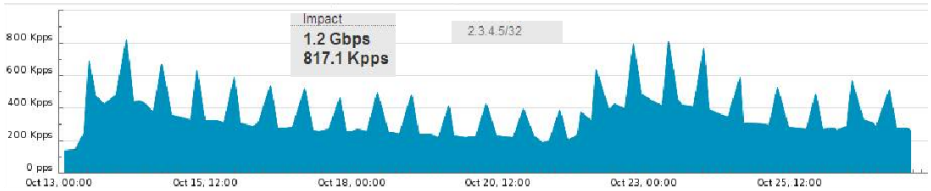


Figure 6. Trafic vers l'adresse 2.3.4.5.

Domaines	Type	IP
bestdigitalmultimedia.com	A	2.2.2.2
bestmultimediasoftware.net	A	2.2.2.2
besttubeshoses.com	A	2.2.2.2
bjtv.net	A	2.2.2.2
caitiaobu.com	A	2.2.2.2
cgtao8.com	A	2.2.2.2
chinalpa.com	A	2.2.2.2
chinesetop.com	A	2.2.2.2
chuanjia.net	A	2.2.2.2
enomhostingsample.com	A	2.2.2.2
google18.com	A	2.2.2.2
googmap.com	A	2.2.2.2
jinxiangnk.com	A	2.2.2.2
jmtee.com	A	2.2.2.2
laoshugen.net	A	2.2.2.2

Figure 7. Exemples de domaines pointant toujours vers l'adresse 2.2.2.2.

ment non propagés sur l'Internet mais détectés par certaines sondes, ont notamment été observées. Actuellement, l'adresse 2.2.2.2 semble toujours utilisée par des tiers, comme le montre la figure 7.

```
Megaweb-player.in, boss-tubes.com, 8dian.cn,
centerall4media.net, helpdatacentre.com and at least 200 other
hosts point to 2.2.2.2.  Gaer4g8ae48g4ae.biz and 212.89.208.in-addr.
    arpa use
2.2.2.2 as a mail server

It is blacklisted in four lists.
```

Listing 1.8. Blacklisting de l'adresse 2.2.2.2.

Rapidement, plusieurs mesures ont été prises par l'AS 3215 :

- au niveau BGP : contacter les AS diffusant de manière illégitime des blocs inclus dans le 2/12 pour faire cesser ces annonces qui durent en général de quelques minutes à quelques heures : les RIR sont systématiquement en copie des mails envoyés, à noter que le taux de réponse via mail est quasi-nul ;
- au niveau des services : mettre en place des mesures opérationnelles visant à interdire en interne l'utilisation d'une liste d'adresses (2.1.1.1, 2.2.2.2, etc) de manière à ce qu'aucun client ni service ne soit affecté par un trafic indésirable.

Comme indiqué dans cette section, la surveillance BGP est un élément nécessaire pour accélérer la réactivité en cas d'incident et la connaissance du fonctionnement du réseau. En parallèle, des travaux sont en cours à l'IETF pour normaliser une infrastructure de routage sécurisée ayant pour objectif de certifier les préfixes annoncés par les pairs eBGP, et ainsi sécuriser l'infrastructure de routage sur l'Internet. Il s'agit de RPKI (Resource Public Key Infrastructure). Cette IGC repose sur l'usage, par les LIR [14], de certificats X.509, ainsi que d'objets signés attestant que les annonces de routes sont légitimes sur les préfixes opérés par un AS (ROAs ou Route Origination Authorizations). On peut raisonnablement penser que le temps d'adoption d'une telle technologie sera très long :

- à la fin 2011, la plupart des OS des routeurs BGP sur l'Internet ne supportaient pas les extensions RPKI ;
- les opérateurs sont réticents à déployer des technologies qui surchargent le plan de contrôle des routeurs BGP (à la fois en mémoire mais surtout en CPU) et qui seraient susceptibles de ralentir les processus de convergence de bout en bout sur des événements réseaux.

14. pour *Local Internet Registry*. Opérateur qui attribue des préfixes à ses clients.

Pour éviter des dénis de service, il est logiquement préconisé de ne jamais bloquer les annonces BGP dont la signature est refusée mais de remonter une simple alerte, ce pour se prémunir des erreurs d'enregistrement et privilégier la continuité de service. Toutefois, même sans automatisation du blocage, RPKI pourrait être un outil puissant pour lever des alertes.

Conclusion

Cet article nous a donné l'occasion de présenter le protocole BGP et de fournir des exemples concernant les différentes techniques utilisées pour le protéger. Après avoir présenté les incidents BGP survenus depuis la première édition du SSTIC, nous avons exposé des détails sur des incidents rencontrés par un grand opérateur français et décrit la manière dont ils ont été traités.

Le protocole BGP ne comporte pas de mécanismes cryptographiques forts permettant de signer les annonces de préfixes, cependant les mécanismes mis en œuvre par les opérateurs visent à détecter les erreurs au plus tôt afin de les corriger rapidement.

Sur la plupart des cas opérationnels détaillés dans cette section, les problèmes ont été résolus en quelques minutes, soit le temps de propagation des nouvelles annonces. Aucun problème chez les clients lié à cet incident n'a été remonté, aucun AS tiers n'a maintenu ces routes.

Parmi toutes les menaces visant le protocole BGP, l'usurpation de préfixes et les problèmes d'implantation demeurent les plus importantes car elles provoquent de graves perturbations dans l'acheminement du trafic sur Internet. L'utilisation des bonnes pratiques de déploiement et la connaissance du fonctionnement des implantations permettent d'en limiter la portée.

Références

1. BGPmon.net, a BGP monitoring and analyzer tool. <https://bgpmon.net/>.
2. Cyclops. <http://cyclops.cs.ucla.edu/>.
3. Internet alert registry. <http://www.cs.unm.edu/~karlinjf/IAR/>.
4. Routing Information Service (RIS). <http://ripe.net/ris/>.
5. How the Internet in Australia went down under. BGPmon.net blog, 2012.

6. C. Alaettinoglu, C. Villamizar, E. Gerich, D. Kessens, D. Meyer, T. Bates, D. Karrenberg, and M. Terpstra. Routing Policy Specification Language (RPSL). RFC 2622 (Proposed Standard), June 1999. Updated by RFC 4012.
7. CERTA-2011-AVI-619. Vulnérabilité dans Juniper, 2011. <http://www.certa.ssi.gouv.fr/site/CERTA-2011-AVI-619/CERTA-2011-AVI-619.html>.
8. Cisco. Cisco IOS XR Software Border Gateway Protocol Vulnerability, 2010. <http://www.cisco.com/en/US/products/csa/cisco-sa-20100827-bgp.html>.
9. D. McPherson. When Hijacking the Internet... ARBOR SERT, 2008. <http://ddos.arbornetworks.com/2008/11/when-hijacking-the-internet/>.
10. E. Chen and al. Notification Message support for BGP Graceful Restart. Internet Draft, Internet Engineering Task Force, 2011.
11. E. Romijn. RIPE NCC and Duke University BGP Experiment. RIPE Labs, 2010. <https://labs.ripe.net/Members/erik/ripe-ncc-and-duke-university-bgp-experiment/>.
12. E. Zmijewski. Longer is not always better. renesys|blog, 2009. <http://www.renesys.com/blog/2009/02/longer-is-not-better.shtml>.
13. E. Zmijewski. Reckless Driving on the Internet. renesys|blog, 2009. <http://www.renesys.com/blog/2009/02/the-flap-heard-around-the-worl.shtml>.
14. A. Heffernan. Protection of BGP Sessions via the TCP MD5 Signature Option. RFC 2385 (Proposed Standard), August 1998. Obsoleted by RFC 5925.
15. I. Ybema. bgp update destroying transit on redback routers ? NANOG mailing list, 2011. <http://seclists.org/nanog/2011/Dec/9>.
16. J. Cowie. Brazil Leak : If a tree falls in the rainforest.... renesys|blog, 2008. <http://www.renesys.com/blog/2008/11/brazil-leak-if-a-tree-falls-in.shtml>.
17. J. Cowie. China's 18-Minute Mystery. renesys|blog, 2010. <http://www.renesys.com/blog/2010/11/chinas-18-minute-mystery.shtml>.
18. J. Scudder and E. Chen and P. Mohapatra and K. Patel. Revised Error Handling for BGP UPDATE Messages, 2011.
19. K. Patel and E. Chen and R. Fernando and J. Scudder. Accelerated Routing Convergence for BGP Graceful Restart. Internet Draft, Internet Engineering Task Force, 2011.
20. W. Kumari, R. Bush, and H. Schiller. Codification of AS 0 processing. Internet Draft, Internet Engineering Task Force, 2011.
21. L. Benkis. Practical BGP Security : Architecture, Techniques and Tools. Technical report, Renesys, 2005.
22. M. Brown. Pakistan hijacks YouTube. renesys|blog, 2008. <http://www.renesys.com/blog/2008/02/pakistan_hijacks_youtube_1.shtml>.
23. M. Zalewski. Strange Attractors and TCP/IP Sequence Number Analysis : Cisco IOS, 2001. <http://lcamtuf.coredump.cx/oldtcp/tcpseq.html#ios>.
24. R. Shaklir. Operational Requirements for Enhanced Error Handling Behaviour in BGP-4. Internet Draft, Internet Engineering Task Force, 2011.
25. Y. Rekhter, T. Li, and S. Hares. A Border Gateway Protocol 4 (BGP-4). RFC 4271 (Draft Standard), January 2006. Updated by RFC 6286.
26. T. Underwood. Internet-Wide Catastrophe-Last Year. renesys|blog, 2005. <http://www.renesys.com/blog/2005/12/internetwide_nearcatastrophela.shtml>.

27. T. Underwood. Cond-Ed Steals the 'Net. renesys|blog, 2006. <http://www.renesys.com/blog/2006/01/coned-steals-the-net.shtml>.
28. T. Wan and P. van Oorschot. Analysis of BGP Prefix Origins During Google's May 2005 Outage. 2006.
29. Tassos. Decoding the RIPE BGP experiment, 2010. <http://ccie-in-3-months.blogspot.com/2010/08/decoding-ripe-experiment.html>.