L'Art d'Anticiper les Changements IGP pour Acheminer Optimalement la Patate en Transit

Jean-Romain Luttringer, Quentin Bramas, Cristel Pelsser, Pascal Mérindol

L'Art d'Anticiper les Changements IGP pour Acheminer Optimalement la Patate en Transit

J-R. Luttringer[1], Q. Bramas[1], C. Pelsser[1] et P. Mérindol[1]

[1] *Université de Strasbourg, ICube*

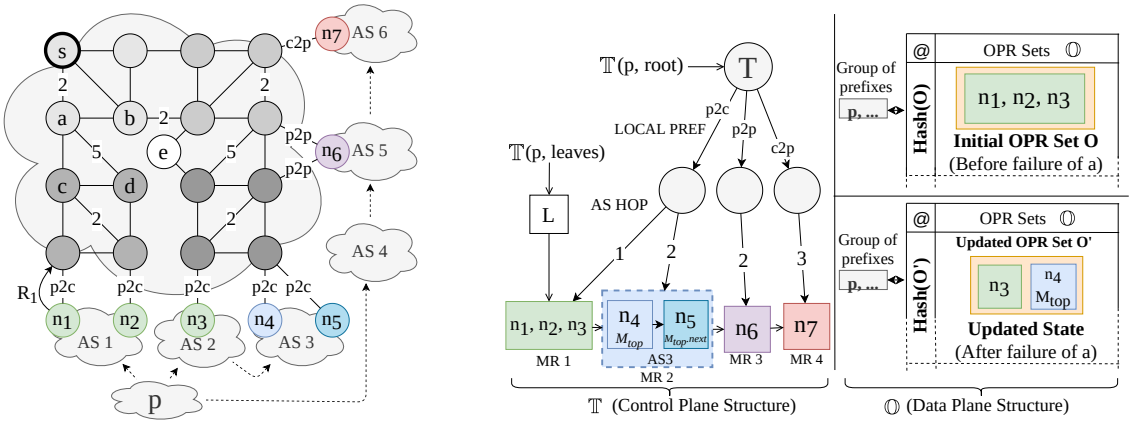
Le routage des données en transit dans les systèmes autonomes (AS) d'Internet se fait selon le paradigme de la patate chaude. Les meilleures routes inter-domaines (BGP) sont sélectionnées grâce à un ordre lexicographique dont l'une des règles stipule de choisir la meilleure distance intra-domaine (IGP) parmi les meilleures routes existantes (ordonnées selon les critères précédents, par ex., préférence économique et nombre de sauts d'AS). Cette pratique est appelée patate chaude car les AS qui l'appliquent évacuent ainsi efficacement le trafic en transit. Cette dépendance de BGP vis à vis de l'IGP implique que BGP doit re-converger après chaque événement interne se produisant dans l'AS (et ce processus est particulièrement lent car traité naïvement). Avec **OPTIC, Optimal Protection Technique for Inter/intra-domain Convergence**, l'objectif de notre travail est de ramener ce temps de convergence à une durée marginale dans la plupart des cas. Pour cela, OPTIC crée et manipule efficacement des ensembles de passerelles BGP contenant les meilleures routes BGP antérieures et postérieures à tout changement IGP. Ces ensembles sont partagés par groupe de préfixes ayant des passerelles identiques. Ainsi, leur mise à jour, construction et utilisation s'opèrent à la granularité du groupe et non du préfixe. Non seulement OPTIC garantit un re-routage rapide vers la meilleure passerelle en cas de changement interne mais assure aussi efficacement sa propre re-convergence face à tous les types de changements : il met à jour ses nouveaux ensembles protecteurs (pour la nouvelle route post-convergence) face à tous les évènements futurs avec un coût inférieur ou égal à celui de BGP pour la gestion de la panne précédente !

Mots-clefs : IGP, BGP, convergence, résilience aux pannes

1 Introduction

Une des briques essentielles à Internet est l'acheminement des paquets en transit à travers les domaines qui le composent. Dans un domaine (ou système autonome, AS), les routeurs commutent les paquets en fonction du préfixe IP auquel la destination du paquet appartient. Lorsque le préfixe n'appartient pas à l'AS où le paquet est commuté, on dira qu'il s'agit de trafic en transit. Ce trafic est acheminé via un mécanisme de routage appelé patate chaude : l'opérateur de l'AS cherche à se débarrasser du paquet le plus vite possible selon la route interne la plus courte.

Cette route est obtenue en appliquant le processus de décision BGP dont l'une des dernières règles stipule que la meilleure route est celle qui minimise la distance dans l'AS — meilleure route parmi celles dont les critères strictement inter-domaines (préférence économique et longueur en saut d'AS en particulier) sont égaux. En d'autres termes, la sélection des routes BGP dépend des changements de l'IGP. Malheureusement, BGP converge lentement, même de manière interne avec iBGP. Pour augmenter la visibilité des routes à l'intérieur du domaine et diminuer le temps de convergence, les opérateurs peuvent utiliser *add-Path* [UFP+16] mais la mise à jour du meilleur prochain saut reste relativement lente car il existe beaucoup de préfixes BGP à traiter (plus de 800K) : en cas de panne, beaucoup d'entrées doivent être mises à jour. PIC (Prefix Independent Convergence) [FMB+11] permet, grâce à une table de routage hiérarchique, de grouper les préfixes BGP pour réduire le nombre d'entrées et de basculer rapidement, en cas de changement IGP, vers une route de secours pré-calculée pour chaque groupe. Cependant la route de secours n'est pas nécessairement la route optimale (post-convergence BGP). De plus, PIC suppose la bi-connexité du réseau comme hypothèse garantie. Comme l'alternative de secours n'est ni nécessairement active ni optimale, PIC se repose encore et toujours sur BGP pour finalement re-converger à son tour. Or les groupes de PIC ne sont d'aucune utilité pour cela : la re-convergence BGP est à nouveau nécessaire pour l'ensemble des préfixes.



(a) Exemple illustrant l'opportunité d'utiliser plusieurs passerelles aux caractéristiques différentes.

(b) Représentation des routes apprises dans le plan de contrôle et de données.

FIGURE 1: Des routes vers les prochains sauts de bordure : représentation et structuration.

Cet aspect est critique : comment anticiper *et* garantir l'optimalité face à *tous* les événements sans les considérer un à un ? La réponse est plus simple qu'il n'y parait et est expliquée en détail dans la partie 2. Un ensemble de passerelles possédant les mêmes attributs inter-domaines, et suffisant pour assurer la 2-connectivité, contient nécessairement la meilleure passerelle après un changement IGP. Nous exploitons cette propriété pour construire facilement des ensembles de passerelles protecteurs après *n'importe quel* changement IGP. Ces ensembles, partagés en mémoire par groupe de préfixes, permettent de basculer très vite vers la nouvelle meilleure route BGP pour tous les préfixes d'un groupe.

Néanmoins, ces ensembles doivent potentiellement être mis à jour après un événement. Cette opération se doit donc également d'être efficace. Pour mettre en évidence la faisabilité d'OPTIC, et avant même de montrer que le nombre et la taille de ces ensembles sont limités en théorie comme en pratique (partie 4), nous allons expliquer comment facilement maintenir ces ensembles à jour lors de changement BGP ou IGP (partie 3). En d'autres termes, nous allons montrer qu'OPTIC est capable de maintenir efficacement ses groupes de préfixes ayant des ensembles de passerelles en commun pour toujours avoir un coup d'avance sur BGP. Alors que la convergence vers la nouvelle passerelle est quasi instantanée, la mise à jour des groupes pour garantir la protection face à un prochain événement à un coût au pire égal à celui de BGP.

2 Garantir une protection optimale face à tout évènement IGP

Dans cette partie et dans la suite de l'article, nous ferons l'hypothèse que la visibilité des routes est suffisante grâce à une solution comme AddPath et une architecture iBGP très simple, par exemple avec un seul route reflector. Des hypothèses plus réalistes sont abordées dans [LBPM21], et OPTIC peut s'y adapter. Sur l'exemple donné en figure 1a, nous montrons les limites d'une solution comme PIC. Avec cette fonctionnalité de re-routage, le routeur s mémoriserait uniquement les routes via n_1 (optimal) et n_2 (secours si n_1 tombe en panne) vers le préfixe p. Après un événement IGP, PIC se contente de restaurer rapidement la connectivité vers p via la meilleure route mémorisée encore active. Or, si le lien $a-c$ tombe en panne, la route via n_3 devient optimale. PIC se contente de restaurer la connectivité vers n_1, qui est toujours joignable, et offre donc une route sous-optimale vers p jusqu'à ce que BGP converge. Pire encore, si le routeur a tombe en panne, PIC n'a pas de route de secours, car il fait l'hypothèse d'un réseau 2-connexe sans vérification préalable ni ajustement : s est alors déconnectée de p jusqu'à la re-convergence. **PIC n'apporte ni une protection complète, ni un re-reroutage optimal, et n'assure donc pas seul la re-convergence de BGP.**

Notre objectif est d'assurer une re-convergence optimale et immédiate avec une méthode efficace. Notre solution consiste à trouver, pour chaque préfixe p, un ensemble de passerelles garanti de contenir la nouvelle meilleure route après n'importe quel événement IGP. Pour calculer cet ensemble, (i) nous groupons

les passerelles ayant les mêmes attributs inter-domaine (cad sans considérer la distance IGP) et (ii) nous empilons ces ensembles (en commençant par les passerelles avec les meilleurs attributs inter-domaine) jusqu'à ce que les passerelles de l'union de ces ensembles offrent deux chemins disjoints vers le préfixe p. L'union ainsi formée, qu'on appelle ensemble protecteur, contient donc suffisamment de passerelles pour tolérer tout changement IGP (chemins disjoints vers p). Comme les routes ont été considérées suivant leurs attributs inter-domaine, ces dernières resteront les meilleures routes après n'importe quel changement IGP car ces changements ne modifient pas les attributs inter-domaine des routes. Après un événement IGP, pour basculer sur la nouvelle route optimale de manière quasi-immédiate, il suffit donc de sélectionner dans l'ensemble protecteur la passerelle avec le poids IGP le plus faible. Les préfixes ayant un ensemble protecteur identique le partagent en mémoire : ainsi, une unique mise à jour de l'ensemble protecteur bénéficie à tous les préfixes.

Sur la Fig. 1a, l'ensemble des passerelles possédant les meilleurs attributs inter-domaine $\{n_1,n_2,n_3\}$ suffit à offrir deux chemins disjoints vers p. Si le nœud a est supprimé, la nouvelle meilleure passerelle est celle possédant le plus petit poids IGP : n_3. Cependant, après la panne de a, cet ensemble composé uniquement de n_3 n'offre plus deux chemins disjoints vers p : il est nécessaire de rajouter l'ensemble des passerelles possédant les deuxièmes meilleurs attributs inter-domaine afin de re-créer un ensemble protecteur.

3 Toujours un coup d'avance, ou la gestion efficace des groupes

La principale difficulté à surmonter est de garantir une reconstruction efficace des groupes après un changement IGP : la bascule vers la meilleure passerelle étant déjà réalisée, comment préparer efficacement les prochains groupes protecteurs pour anticiper n'importe quel futur événement IGP ? Sur la Fig. 1, on peut observer que le plan de contrôle d'OPTIC est construit sur base d'un arbre de préfixes $\mathbb{T}$ dont les feuilles $\mathbb{L}$ sont triées selon les attributs inter-domaine. Les premières (meilleures) feuilles permettent de construire les ensembles protecteurs, transférés dans le plan de données $\mathbb{O}$. C'est sur ces ensembles réduits que pointent les groupes de préfixes, une fonction de hachage étant appliquée aux contenus des ensembles afin de les identifier de manière unique. Ainsi, dans le plan de données, un groupe de préfixes est associé à un ensemble de passerelles contenant la meilleure route courante et la meilleure route après tout changement IGP.

Après un changement IGP OPTIC opère une bascule vers la meilleure passerelle en appliquant un simple minimum dans chaque ensemble. Le changement IGP est donc pris en compte de manière optimale et quasi-instantanée. Ensuite, OPTIC se met à jour afin d'être prêt pour le *prochain* changement IGP. Pour chaque groupe de préfixe, si les passerelles de l'ensemble protecteur associé offrent toujours deux chemins disjoints vers p, ce-dernier reste inchangé. En revanche, si cette propriété n'est plus vérifiée, le groupe est mis à jour via les informations contenues dans $\mathbb{L}$. Cette mise à jour s'opère préfixe par préfixe dans le groupe concerné (car ils ne partagent pas nécessairement le même $\mathbb{L}$). **OPTIC opérant à la granularité des groupes de préfixes, son coût de mise à jour pour anticiper le prochain événement IGP est inférieur ou au pire égal au coût de BGP pour réagir à l'événement courant.** Sur la Fig. 1a, seule la panne du nœud a provoque une modification du groupe (pour le préfixe p). Son effet est visible dans la Fig. 1b : la passerelle n_4 doit être ajoutée car n_1 et n_2 sont inaccessibles. Dans tous les autres cas, l'ensemble IGP arrondi n_1, n_2, n_3 est suffisant pour assurer la protection, même après une panne.

Après une annonce BGP Quand une route BGP est apprise/modifiée, il suffit de l'insérer (resp. la modifier) dans $\mathbb{T}$ et d'appliquer les éventuels changements associés dans $\mathbb{O}$ si la route modifie effectivement l'ensemble protecteur. Le coût de cette mise à jour d'OPTIC est équivalent à celui de BGP.

4 Analyse du nombre de groupes : un plan de données compact

Nous fournissons ici [†] un modèle d'analyse assez défavorable car ne prenant pas en compte les préférences régionales limitant le nombre d'annonces à considérer dans la réalité. Soit un AS avec B passerelles biconnectées annonçant P préfixes au total. Chaque préfixe p est annoncé par un sous ensemble $b \leq B$ de passerelles, choisies aléatoirement selon une loi uniforme. Pour chaque préfixe, l'étalement de la politique

†. Les résultats présentés sont issus d'une version optimisée d'OPTIC expliquée ici : `https://optic-icube.github.io/`

TABLE 1: Nombre de groupes distincts ($|\mathbb{O}|$) selon plusieurs configurations.

Type of AS	# gateways per class	# prefix per class	# distinct OPR sets	OPR sets median size	Lower bound
Stub	$(10;20;0)$	$(700K;100K;0K)$	3945	4	235
Tier 3	$(10;50;100)$	$(500K;200K;100K)$	46010	3	6219
Tier 2	$(5;500;2000)$	$(500K;200K;100K)$	263219	2	197194
Tier 1	$(0;50;5000)$	$(0K;600K;200K)$	232180	2	199633

BGP est représenté par un entier entre 1 et ps choisi aléatoirement selon une loi uniforme. Chaque sous ensemble de taille $n \leq b$ a une probabilité d'existence p_n ou p'_n suivant la manière dont il est construit. Notre modèle calcule le nombre $|\mathbb{O}| = |\mathbb{O}_{B,P,ps}|$ d'ensembles uniques en fonction de B, P, et ps. La quantité $|\mathbb{O}_{B,P,ps}|$ est donc le nombre d'ensembles de passerelles distinctes, i.e. le nombre de groupes de préfixes.

Suivant sa configuration interne, un ensemble de taille n ($2 \leq n \leq b$) est dans $\mathbb{O}_{B,P,ps}$ avec une probabilité $\mathbb{P}_{B,P,ps,n}$ ou $\mathbb{P}'_{B,P,ps,n}$. Comme il existe respectivement $\binom{B}{n}$ ou $B\binom{B-1}{n-1}$ de ces ensembles, nous en déduisons :

$$|\mathbb{O}_{B,P,ps}| = \sum_{n=2}^{b} \binom{B}{n} \mathbb{P}_{B,P,ps,n} \; + \; B\binom{B-1}{n-1} \mathbb{P}'_{B,P,ps,n} \tag{1}$$

$$\mathbb{P}_{B,ps,P,n} = 1 - \left(1 - \binom{B}{n}^{-1}\right)^{p_n P} \quad , \quad \mathbb{P}'_{B,ps,P,n} = 1 - \left(1 - \left(B\binom{B-1}{n-1}\right)^{-1}\right)^{p'_n P} \tag{2}$$

$$p_n = \sum_{i=1}^{ps} \binom{b}{n} \frac{1}{ps^n} \left(1 - \frac{i}{ps}\right)^{b-n} \quad , \quad p'_n = \sum_{i=1}^{ps} b\binom{b-1}{n-1} \frac{1}{ps^{n-1}} \frac{i-1}{ps} \left(1 - \frac{i}{ps}\right)^{b-n} \tag{3}$$

Le tableau 1 décrit le nombre de groupes obtenus en raffinant l'analyse. Les AS étant triés en fonction de la dispersion de leurs préférences locales en trois classes (fournisseurs, paires, clients) et de leur structure topologique — nombre de passerelles et préfixes appris par classes de voisins. **Pour la majorité des AS, tels que les réseaux Stubs et les réseaux de transit avec un nombre de passerelles inférieur à la centaine, le nombre de groupes $|\mathbb{O}|$ est très réduit.** Pour les gros réseaux de transit, le nombre de groupe distincts $|\mathbb{O}|$ est fortement dépendant de la décomposition en sous classes : pour les grands Tier 1, $|\mathbb{O}|$ est relativement élevé mais OPTIC est de toute façon proche de la borne minimale pour la protection. Globalement, comme le nombre de groupes est (bien) plus faible que le nombre d'entrées BGP, OPTIC est à même de répondre rapidement à la plupart des changements IGP très efficacement.

5 Conclusion

OPTIC découple l'IGP de BGP à l'aide de groupes de préfixes BGP peu nombreux, petits et stables. Chaque groupe pointe vers un ensemble de routes protecteur commun offrant deux chemins disjoints vers le préfixe et contenant les meilleures routes pre- et post-convergence pour n'importe quel événement IGP. Afin de protéger le trafic de transit en cas de nouveau changement, c'est à dire pour anticiper la *prochaine* panne avec de nouveaux ensembles protecteurs, le coût de la mise à jour des groupes d'OPTIC est limité pour être inférieur – ou au pire égal bien que généralement très inférieur – au temps pris par BGP pour se "remettre de la panne précédente" ! Lors de changements internes ou de pannes de bordure, OPTIC a un coup d'avance sur BGP pour un coût moindre : la bascule vers la route post-convergence est quasi-immédiate et les prochains groupes protecteurs reconstruits efficacement si nécessaire.

Références

[FMB+11] Clarence Filsfils, Pradosh Mohapatra, John Bettink, Pranav Dharwadkar, Peter De Vriendt, Yuri Tsier, Virginie Van Den Schrieck, Olivier Bonaventure, and Pierre Francois. Bgp prefix independent convergence (pic) technical report. *Cisco, Tech. Rep, Tech. Rep*, 2011.

[LBPM21] Jean-Romain Luttringer, Quentin Bramas, Cristel Pelsser, and Pascal Mérindol. A fast-convergence routing of the hot-potato, 2021. `https://arxiv.org/abs/2101.09002`. To appear in INFOCOM'21.

[UFP+16] Jim Uttaro, Pierre Francois, Keyur Patel, Jeffrey Haas, Adam Simpson, and Roberto Fragassi. Best Practices for Advertisement of Multiple Paths in IBGP. Internet-Draft draft-ietf-idr-add-paths-guidelines-08, IETF Secretariat, April 2016.