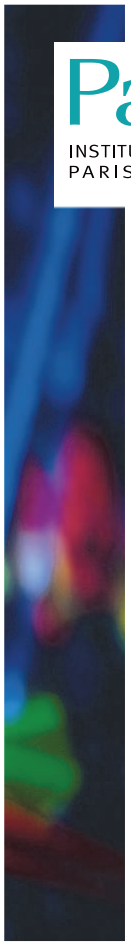
Enabling inter-domain path diversity

Xavier Misseri



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0057

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Informatique et Réseaux »

présentée et soutenue publiquement par

Xavier MISSERI

le 10 octobre 2013

Vers une utilisation de la diversité de chemins dans l'Internet

Enabling inter-domain path diversity

Directeur de thèse :**Jean-Louis ROUGIER,**
Telecom Paristech, France
Co-encadrement de la thèse :**Olivier BONAVENTURE,**
Université catholique de Louvain, Belgique

Jury

M. Chadi BARAKAT, Chargé de recherche, INRIA	Rapporteur
M. Peter REICHL, Professor, University of Vienna	Rapporteur
M. Jordi DOMINGO-PASCUAL, Professor, Universitat Politècnica de Catalunya	Examinateur
M. Maurice GAGNAIRE, Professeur, TELECOM ParisTech	Examinateur
M. Olaf M. MAENNEL, Assistant professor, Loughborough University	Examinateur

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Remerciement

Je tiens tout d'abord à remercier très chaleureusement mon directeur de thèse, Jean-Louis Rougier, de m'avoir donné l'opportunité de réaliser cette thèse. Je le remercie pour son encadrement et son soutien ainsi que pour la confiance qu'il m'a accordé. Je tiens également à exprimer ma gratitude à Chadi Barakat, Peter Reichl, Olivier Bonaventure, Jordi Domingo-Pascual, Maurice Gagnaire et Olaf M. Maennel d'avoir bien voulu me donner de leur temps pour faire partie de ce jury. Je suis reconnaissant à l'égard de Chadi Barakat et de Peter Reichl d'avoir accepté de juger mon travail de thèse en étant rapporteurs. Je remercie fortement Olivier Bonaventure pour ses conseils et aussi de m'avoir accueilli dans son équipe à Louvain-la-Neuve.

Je remercie fortement les personnes avec qui j'ai eu le plaisir de travailler : Ivan Gojmerac, Damien Saucez, Stefano Moretti, Virginie Van Den Schrieck, Lamine Lamali et Hélia Pouyllau.

Je remercie également mes collègues de Telecom Paristech (et d'ailleurs), anciens et actuels, qui ont permis une atmosphère de travail très agréable. Un grand merci à Claudio, Raluca, Stefano Iellamo, Céline, Lamine, Claude, Antonino, Luigi, Mattia, Mathis, Sameh, Ahlem, Ghida, Davide Tamaro, Davide Cuda, Marc, Giuseppe, Isabel, Konstantinos, Mario, Felipe, Stefano Secci, Hélia, Dorice, Lilianna, Paolo, Sylvio, Maciej, Walter, Michele, Silvia, Rino, Andrea... et toute l'équipe d'INFRES, pour leur soutien et les très bons moments que j'ai passé à leurs côtés. Ces années de thèse m'ont permis de faire de très belles rencontres.

A ma compagne Cécile, des milliards de mercis pour son soutien, sa compréhension et sa patience... j'espère pouvoir te rendre ça un jour :-) d'ici peu... Je te serai toujours reconnaissant pour ta présence, ton aide et ton énorme capacité à me supporter... Un grand merci à ma famille d'avoir toujours été présent pour moi, de m'avoir encouragé et de m'avoir toujours soutenu dans les choix que j'ai pu faire. A mes parents, à mon frère et à mes petits neveu et nièce (trop adorables...), je vous remercie énormément d'être là ! ! Je remercie aussi grandement ma belle-famille, qui a toujours été là, pour me soutenir et m'aider.

Un très grand merci aussi à tous mes amis de longue date, Pierre-Louis, Julien, Olivier... de m'avoir soutenu et aussi supporté... et de m'avoir changé les idées quand j'en avais besoin.

Abstract

In the present Internet, autonomous systems are inter-connected in such a way that several paths may exist from one source to a destination. Nevertheless, as inter-domain routing is based on BGP-4, which selects a single path per destination prefix, it prevents carriers and end-users to use the vast inherent path diversity. The addition of multi-path capabilities to the Internet has long been advocated for both robustness, quality of service and traffic engineering purposes.

In this thesis we consider a new service where carriers offer additional routes to their customers (w.r.t. to the BGP default route) as a free or value-added service. These alternate routes can be used by customers to optimize their communications, by bypassing some congested points in the Internet (e.g. a “tussled” peering points), to help them to meet their traffic engineering objectives (better delays etc.) or just for robustness purposes (e.g, shift to a disjoint alternate route if needed).

First we propose a simple architecture that allows a network service provider to benefit from the diversity it currently receives. Then we extend this architecture in order to make the propagation of the Internet path diversity possible, not only to direct neighbors but also to their neighbors and so on. We take advantage of this advance to relax the route selection processes of autonomous systems in order to make them be able to set up new routing paradigms.

Nevertheless announcing additional paths can lead to scalability issues, so each carrier could receive more paths than what it could manage. We quantify this issue and we underline easy adaptations and small path filterings which make the number of paths drop to a manageable amount.

Last but not least we set up an auction-type route allocation framework, which gives to network service providers the opportunities first to propagate to their neighbors only the paths the said neighbors are interested in and second to leverage a new routing selection paradigm based on commercial agreements and negotiations.

Résumé

Internet est constitué de systèmes autonomes inter-connectés de telle manière que plusieurs chemins existent afin d'aller d'une source à une destination. Néanmoins, l'actuel protocole de routage inter-domaine (BGP), qui ne sélectionne qu'une seule route utilisable, empêche les opérateurs et les utilisateurs finaux de bénéficier la vaste diversité de chemin inhérente à Internet. La mise en marche de cette diversité de chemin a depuis longtemps été reconnue comme une avancée afin d'obtenir des améliorations de robustesse, de qualité de service et d'ingénierie de trafic.

Nous considérons, dans cette thèse, un nouveau service par lequel les opérateurs de télécommunications offrent des routes supplémentaires à leurs clients (en plus de la route par défaut fournie par BGP) comme un service gratuit ou à valeur ajoutée. Ces routes supplémentaires peuvent être utilisées par des clients afin d'optimiser leurs communications, en outrepassant des points de congestion d'Internet, ou les aider à atteindre leurs objectifs d'ingénierie de trafic (meilleurs délais etc.) ou dans un but de robustesse (par exemple en basculant sur un chemin disjoint en cas de panne).

Nous proposons d'abord une architecture simple permettant à un opérateur de télécommunication de bénéficier de la diversité de chemin qu'il reçoit déjà. Nous étendons ensuite cette architecture afin de rendre possible la propagation de cette diversité de chemin, non seulement aux voisins directs mais aussi, de proche en proche, aux autres domaines. Nous profitons de cette occasion pour relaxer la sélection des routes des différents domaines afin de leur permettre de mettre en place de nouveaux paradigmes de routage.

Néanmoins, annoncer des chemins additionnels peut entrainer des problèmes de passage à l'échelle car chaque opérateur peut potentiellement recevoir plus de chemins que ce qu'il peut gérer. Nous quantifions ce problème et mettons en avant des modifications et filtrages simples permettant de réduire ce nombre à un niveau acceptable.

En dernier, nous proposons un processus, inspiré des ventes aux enchères, permettant aux opérateurs de propager aux domaines voisins seulement les chemins qui intéressent les dits voisins. De plus, ce processus permet de mettre en avant un nouveau paradigme de propagation de routes, basé sur des négociations et accords commerciaux.

Synthèse en français

Introduction

Internet est né dans les années 60 en tant que réseau académique et de recherche [1, 2]. Aujourd'hui 34% de la population mondiale y est connecté [3] ce qui laisse présager une croissance encore importante.

Internet est souvent considéré comme une invention dont les retombés sont aussi importantes que l'écriture ou l'impression [4]. C'est une révolution qui induit une ré-organisation importante de la société, sur de multiples domaines telles que l'économie, la religion ou la politique [5].

Internet fait maintenant partie de notre vie journalière, est considéré comme un succès et semble déjà être arrivé à maturité. Cependant, son évolution continue et de nouvelles utilisation émergent (par exemple, la télé-chirurgie [6] ou la communication inter-cérébrale [7]). La présente thèse se place dans ce contexte, où nous ne sommes pas encore parfaitement conscient du potentiel d'Internet et à quel point il change notre vie.

Nous abordons ici un changement profond. Alors que l'Internet a toujours été contraint de n'utiliser qu'un seul chemin entre une source et une destination cette thèse tente de mettre en marche l'utilisation des multiples chemins, intrinsèquement disponibles mais actuellement non utilisables.

BGP [8] (Border Gateway Protocol) est le protocole de routage actuellement utilisé afin de propager les informations d'accessibilité des préfixes de l'Internet. BGP est contraint par son processus de de sélection de route qui ne sélectionne, in fine, qu'une seule route. Quelques travaux on permit d'assouplir quelque peu cette contrainte (cf. Multipath BGP [9, 10]) mais le nombre de routes choisies et la maîtrise de ce choix sont très limités.

La figure 1 illustre la diversité potentielle de chemins pour aller d'une source à une destination. BGP ne sélectionne qu'un chemin (le chemin rouge) alors que d'autres chemins existent.

Utiliser la diversité de chemins disponible peut apporter plusieurs bénéfices.

Tout d'abord, les flux ayant la même destination convergent actuellement vers le même chemin, causant ainsi des points de congestion [11]. Le premier bénéfice est alors de pouvoir éviter/contourner les points de congestion.

De plus, la multitude d'applications existantes [12] nécessite des chemins de caractéristiques différentes (délai, bande passante, perte de paquets... [13]), afin d'améliorer le ressenti utilisateur. L'utilisation de plusieurs chemins pourrait per-

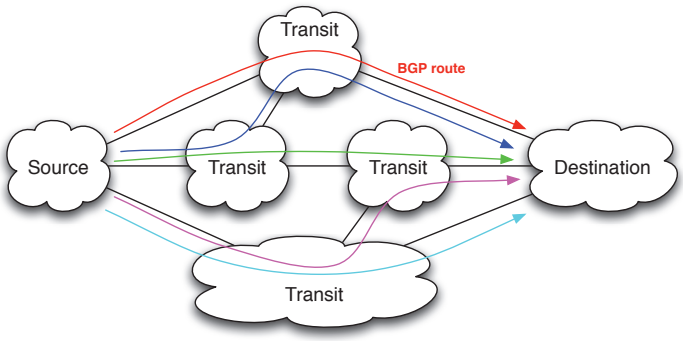


FIGURE 1 – Chemins potentiels

mettre la sélection de chemins adaptés à chaque utilisation/application.

Enfin, les fournisseur d'accès à internet et les fournisseurs de transit ont des besoins d'ingénierie de trafic importants et l'utilisation de plusieurs chemins s'avère être un outil important pour ce type de besoins. Le partage de charge serait alors plus aisé et obtenir des chemins disjoints permettrait alors de basculer plus rapidement sur un chemin valide en cas de panne.

Fournir une solution afin de débloquer la diversité de chemins sous-jacente n'est pas chose aisée. La présente thèse se base sur les contributions [14] et [15] afin d'établir une liste de pré-requis qu'une solution doit réunir afin de pouvoir atteindre son but. On retrouve deux types de pré-requis.

Les premiers sont de type technique et permettent de propager l'information de routage et d'utiliser les différents chemins propagés. Une architecture de multi-chemins inter-domaines se doit alors de modifier le plan contrôle et le plan de donnée actuellement utilisés. A cela s'ajoute qu'un ISP peut ne pas adopter le multi-chemins et se contenter de l'internet tel qu'il est actuellement conçu. L'architecture doit alors être adoptable de manière incrémentale, c.a.d., qu'un ISP peut en bénéficier même si d'autres ISPs n'ont pas adopté l'architecture. Ceci inclus le fait que la solution proposée doit être compatible avec l'Internet existant, c.a.d., le routage BGP actuel.

Les seconds pré-requis sont de type économiques. La solution proposée doit fournir un retour sur investissement évident. Bien qu'il soit difficile de savoir de quelles manières les ISPs peuvent monétiser de tels services, une manière d'atteindre ce but est de proposer une architecture nécessitant un investissement raisonnable. Pour cela les technologies utilisées doivent être existantes, ce qui diminue drastiquement les coûts de migrations. De plus, l'architecture ne doit pas être restreinte à un service spécifique. En effet, les ISPs potentiellement intéressés par le multi-chemins peuvent chacun proposer des services à valeur ajoutée différents. Enfin, la solution doit respecter les paradigmes de l'Internet. Par exemple l'Internet doit rester décentraliser et les ISPs doivent avoir un pouvoir équivalent à ce qu'ils ont qu'actuellement sur leur propre réseau. De plus, la solution ne doit pas né-

cessiter de partage d'information commercialement sensible entre opérateurs (par exemple, les informations concernant les accords commerciaux entre domaines).

Un effort substantiel a déjà été fourni afin de permettre l'utilisation de la diversité de chemins de l'Internet. Néanmoins aucun des travaux proposés ne réuni les prérequis précédemment énoncés.

Certaines propositions ne se concentrent que sur la propagation de plusieurs chemins [16] tandis que d'autres ne se focalisent que sur l'utilisation de chemins déjà reçus [9, 10, 17, 18] D'autres travaux focalisent sur des solution adaptées exclusivement à la convergence en cas de panne [19, 20, 21, 22], ce qui constitue une utilisation limitée de la diversité. Enfin, d'autres contributions [23, 24, 25, 20, 26], plus complètes, restent à l'état de propositions. Ces architectures ne reposent pas sur des technologies actuellement disponibles et certaines propositions nécessitent de plus une migration architectural brutale, ce qui n'est pas envisageable.

Le présent résumé de thèse est organisé comme suit.

Nous proposons d'abord une première architecture permettant l'utilisation locale de la diversité de chemins qu'un opérateur reçoit grâce à ses connexions eBGP. Grâce aux données fournies par Route Views [27], nous montrons que la diversité reçue est déjà conséquente et que l'instabilité induite est limitée.

Ensuite, nous généralisons l'architecture. La diversité accumulée par un opérateur peut maintenant être propagée à ses voisins qui peuvent faire de même avec leurs voisins. La diversité est donc propagée aux autres domaines ayant adopté l'architecture. A l'aide d'une simulation, nous évaluons la diversité de chemins disjoints et concluons qu'avec la règle de préférence pour les clients ("prefer customer rule"), communément utilisée dans l'Internet, un opérateur peut ne pas être en mesure de joindre 50% des domaines avec deux chemins disjoints. Nous proposons alors une assouplissement du processus de sélection des routes, comprenant un critère assurant la stabilité de l'Internet, ce qui permet d'obtenir une diversité de chemins disjoints proche la la limite théorique.

La quantité de chemins que l'assouplissement permet d'obtenir est si importante qu'elle peut entrainer des problèmes de passage à l'échelle au niveau de la table de routage des routeurs. Nous explorons donc le cas où toutes les routes sont propagées de proche en proche, sans filtrage. Nous quantifions le problème et montrons qu'en apportant quelques petites adaptations, le nombre d'entrées à insérer dans la table de routage des routeurs devient gérable.

Enfin, nous proposons un processus dans lequel les fournisseurs de transit peuvent vendre l'accès aux routes à leurs voisins. Le système proposé est inspiré des enchères combinatoires et réuni certaines propriétés intéressantes. Le vendeur et les acheteurs potentiels ont tous un intérêt à participer au jeux d'allocation. De plus, dans certaines conditions, les acheteurs potentiels dévoilent la vrai valeur qu'ils évaluent pour l'ensemble des biens qu'ils veulent acheter. Enfin, le temps de calcul de ce système est polynomial, d'un point de vue du nombre d'acheteurs potentiels et du nombre de biens à vendre, ce qui le rend adoptable dans un contexte inter-domaine.

Utiliser la diversité de chemins localement reçue

Dans cette section, nous nous concentrons sur la diversité de chemins qu'un opérateur peut proposer à ses clients. Cette diversité est actuellement reçue par l'opérateur grâce à ces connections externes (eBGP).

Architecture

L'architecture que nous proposons dans cette section est une première étape vers la mise en marche de la diversité de chemins dans l'Internet. Cependant, elle a comme avantage d'être facilement utilisable et de ne pas nécessiter de coopération avec les autres opérateurs.

La figure 2 illustre notre proposition.

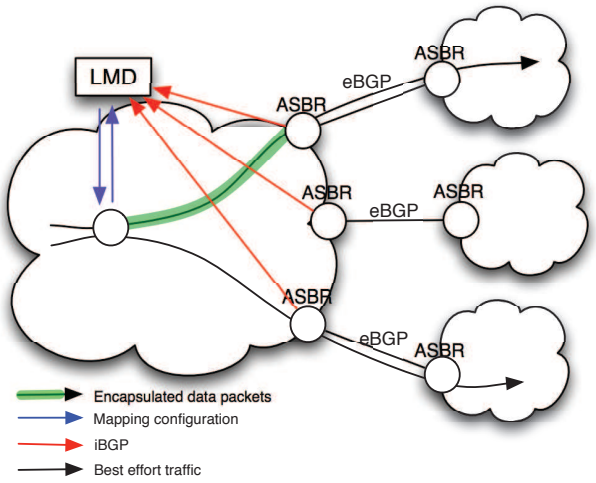


FIGURE 2 – Utilisation interne de l'encapsulation et du mapping system fourni par LISP

L'architecture repose sur le concept Map-and-Encaps. D'un point de vue plan contrôle, la partie de "mapping" permet de stocker la diversité de chemins reçue par eBGP dans le LMD (Local Mapping Distributor). Chaque chemin est associé à un point de sortie du domaine. D'un point de vue plan de donnée, la partie encapsulation permet de forcer les paquets à suivre le chemin qui a précédemment été choisit par le LMD. Nous basons l'architecture sur le protocole LISP [28], qui propose un type spécifique d'encapsulation (basé sur de l'IP dans de l'IP).

Plus précisément, chaque annonce eBGP reçue est directement stockée dans le LMD avec l'identité de l'ASBR recevant l'annonce. Quand un paquet de données arrive sur un router, ce dernier demande au LMD comment il doit le traiter. Le LMD lui donne alors l'ordre d'encapsuler le paquet et de placer l'adresse de l'ASBR de sortie dans le champ de destination de l'entête d'encapsulation. Le router s'exécute et envoie le paquet sur le chemin permettant d'atteindre l'ASBR de sortie. Tous les

routeurs intermédiaires routent le paquets en ne prenant en compte que l'entête d'encapsulation. Une fois arrivé à l'ASBR, le paquet est dé-capsulé et le paquet original est envoyé vers l'ASBR d'entrée du domaine suivant.

Ce type d'architecture peut être utilisée pour les besoins internes d'un opérateur (cf. figure 2). Néanmoins, nous pensons que les opérateurs voudront partager leurs diversités avec leurs clients stubs (cf. figure 3) afin de proposer de nouveaux services et ainsi obtenir de nouveaux profits.

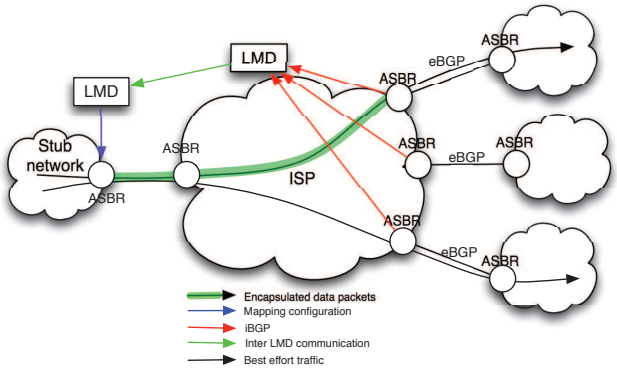


FIGURE 3 – Diversité de chemins propagée aux réseaux stubs

Dans ce cas, l'encapsulation est effectuée par le client stub et la dé-capsulation est effectuée par l'opérateur, au niveau de l'ASBR de sortie.

Au niveau des deux LMDs, chacun des acteurs (le client ou l'opérateur) peut effectuer un filtrage des routes qui seront utilisées. Tout d'abord, l'opérateur voudra peut être ne proposer qu'un sous ensemble de la diversité qu'il a, en fonction du contrat qu'il a établi avec son client. Enfin, le client peut ne prendre en compte qu'une partie des routes proposées par l'opérateur, suivant sa politique de routage (stabilité, performance...).

Evaluation

Utiliser la diversité de routage peut s'avérer dangereux dans le sens où les routes propagées de part et d'autres de l'Internet peuvent être instables.

Nous utilisons, dans cette section, des données publiques fournies par Route Views [27] et CAIDA [29] afin d'évaluer une borne inférieur à la diversité que différents opérateurs peuvent partager avec leurs clients. Nos évaluations focalisent sur les ASes 3356, 13030, 4436 et 2914.

Comme chaque domaine peut filtrer les routes à sa guise au sein de son LMD, nous simulons différents processus de sélection. Ces processus sont basés sur les attributs BGP car ce sont les seules informations que Route Views fournit. Voici les processus de sélection utilisés :

ALL : sélectionne toutes les routes quelles que soient leurs métriques.

LP : ne sélectionne que les routes ayant la meilleure local préférence.

ASPL : ne sélectionne que les routes ayant le plus court chemin d'ASes.

LP+ASPL (∼Multipath BGP) : ne sélection que les routes ayant la plus grande locale préférence et, parmi les routes résultantes, celles ayant le plus petit chemin d'ASes. Nous approximons cette sélection à multipath BGP.

BGP : sélectionne les routes en fonction du processus de sélection BGP.

La figure 4 montre les résultats de l'évaluation pour l'ASes 3356 et pour tous les processus de sélection de routes.

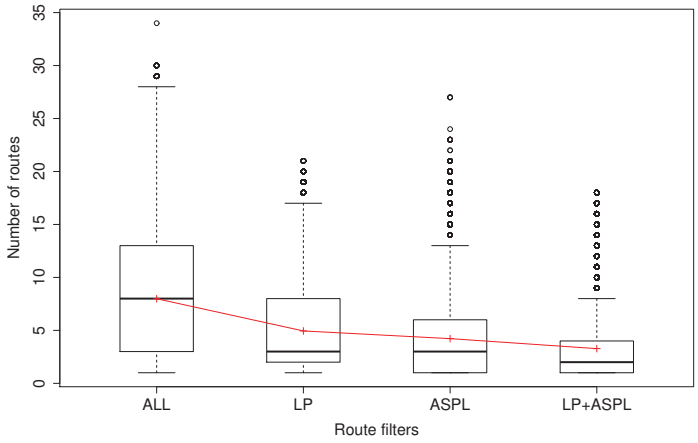


FIGURE 4 – 3356 : diversité de chemins

Dans cette évaluation, les préfixes ont des diversités de chemins très diverses. Certains obtiennent plus de 30 routes (cf. sélection *ALL*) tandis que d'autres n'obtiennent qu'une seule route. La diversité utilisable dépend de manière importante du filtrage utilisé.

Nous pouvons conclure que la diversité de chemins actuellement reçue par les opérateurs est existante et non négligeable. Il est donc fortement probable qu'elle puisse être la source de nouveaux services proposés à leurs clients. Il est d'autant plus important de préciser que ces résultats sont obtenus alors qu'une quantité importante d'annonces BGP n'ont pas été analysées car les données Routes Views et CAIDA ne sont pas exhaustives.

Les mêmes évaluations ont été effectuées sur les ASes 13030, 4436 et 2914 et des résultats similaires ont été mis en avant.

Néanmoins, obtenir de la diversité de routage peut s'avérer dangereux. En effet, chaque chemin pris en compte par le processus de sélection est accompagné par l'instabilité qui lui est associé. Nous avons effectué une deuxième série d'évaluations permettant d'évaluer la quantité de churn et de la comparer avec les bénéfices fournis par l'obtention de cette diversité. Pour cela, nous avons effectué les

même évaluations que précédemment mais en prenant en compte l'historique des changements de métriques, les effacements... de chaque chemin.

La figure 5 présente la corrélation entre l'instabilité et la diversité de routage obtenue pour l'AS 3356. Une extrapolation linéaire accompagne chaque série de points, afin d'en faciliter la lecture nous avons aussi ajouté une extrapolation linéaire de l'instabilité fournie par BGP dans un but de comparaison.

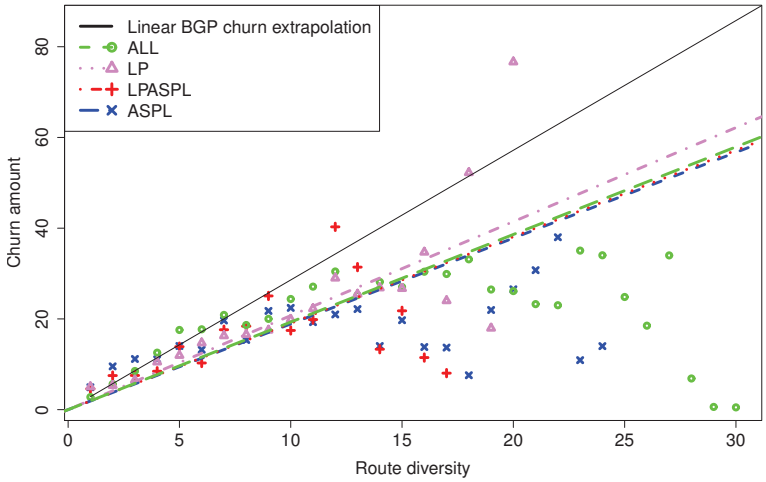


FIGURE 5 – 3356 : Instabilité en fonction de la diversité

Nous pouvons voir que, quelque soit le processus de sélection utilisé, le nombre moyen de churn par route est inférieur au nombre de churn de la route choisie par BGP. Nos obtenons la même conclusion avec les ASes 13030, 4436 et 2914.

Mise en marche de la diversité globale

Architecture

Nous présentons ici l'architecture IDRD (Inter-Domain Route Diversity) qui permet l'utilisation de la diversité de chemins intrinsèquement présents dans l'Internet. L'architecture est basée sur le paradigme de "map-et-encap". LISP [28] est ici donné comme exemple de protocole "map-et-encap" permettant de l'implémenter. Néanmoins, d'autres protocoles (par exemple MPLS) peuvent être utilisés.

Nous souhaitons clarifier le fait qu'IDRD n'a pas pour but de remplacer BGP. Nous considérons cette solution comme un ajout à l'Internet actuel, dans le sens où BGP garde son rôle actuel.

La figure 6 fournit une vision globale de l'architecture.

Au niveau du plan de données, nous choisissons de marquer la séparation entre l'intra-domaine et l'inter-domaine. Cela permet entre autre de laisser à chaque

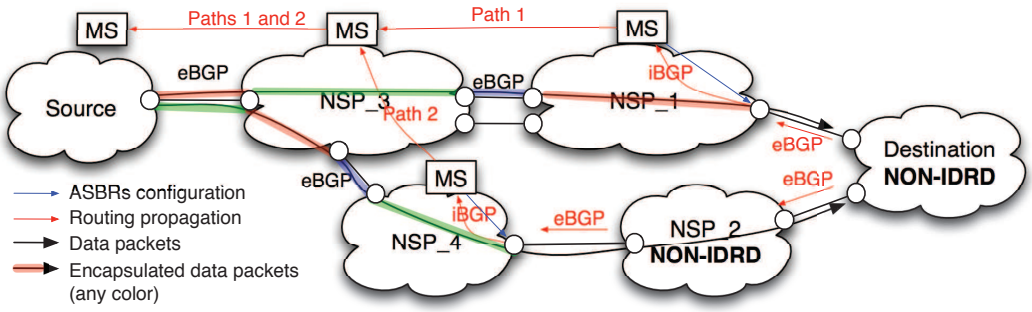


FIGURE 6 – Architecture IDRD.

domaine le choix de la technologie qu'il utilise, c.a.d. un domaine peut choisir une encapsulation IP alors que son voisin choisit MPLS.

Au niveau intra-domaine, les paquets sont encapsulés par l'ASBR (AS Border Router) d'entrée (point (1) sur la figure 7) de chaque domaine dans le but de forcer son chemin jusqu'à l'ASBR de sortie (point (3)) qui aura été préalablement choisit par le MS. Le même processus d'encapsulation intervient à l'inter-domaine.

Puisque plusieurs chemins sont disponibles afin d'atteindre une destination donnée, l'adresse IP de destination n'est pas une information suffisante afin de router le paquet. Nous ajoutons alors au sein du paquet, en plus de l'adresse IP de destination, un identifiant servant à spécifier le chemin qui doit être suivi. Nous l'appelons le Path-ID (path identifier).

Sur le plan contrôle, chaque domaine ayant adopté IDRD est muni d'un Mapping System (MS), qui est connecté aux MSes des domaines voisins. Ce réseau overlay, formé par les MSes, sert à propager la diversité de routage et les métriques associées. L'information de routage, composée au moins par le préfixe, le chemin d'ASes et les éléments de configuration (Path-IDs...), sont propagés d'un MS à ceux des domaines voisins. Le protocole d'échange de données de routage est laissé à l'appréciation des opérateurs. Néanmoins, BGP add-path [16] peut être utilisé pour la communication entre les Mapping Systems de chaque opérateurs. D'autres paradigmes de propagation de routes peuvent être utilisés. Par exemple, une négociation pour la vente de routes requiert alors des métriques/champs/messages additionnels (c.a.d., proposition, préférence, acceptation...), tels que proposés dans [30].

Comme un domaine peut potentiellement être connecté à un voisin n'ayant pas adopté IDRD, il reçoit des annonces eBGP. Ces annonces peuvent être insérées dans le MS et devenir une source de diversité prise en compte par IDRD. Le premier domaine à déployer IDRD sera alors en mesure d'insérer dans le MS les routes eBGP reçues de ses voisins (cf. point (B)) et ainsi profiter de la diversité de routage qu'il reçoit déjà (tel que décrit dans la section précédente).

Une fois les routes propagées d'un domaine à un autre, le MS sélectionne les

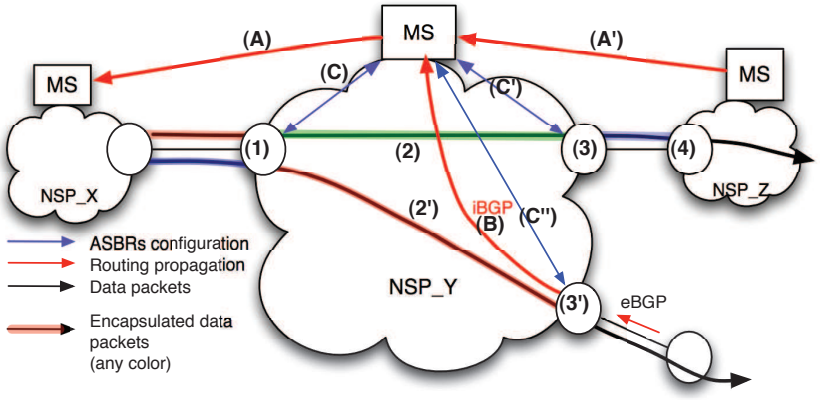


FIGURE 7 – Zoom sur un domaine.

routes qu'il souhaite utiliser. Il les configure ensuite dans les ASBRs d'entrée et de sortie de son réseau (points (C), (C') and (C") sur la figure 7). Il est possible d'envoyer les configurations dès que la route est sélectionnée, ce qui implique un importante table de routage, ou d'attendre qu'un ASBR fasse la demande explicite d'une route lorsque des paquets arrivent sur l'ASBR d'entrée, ce qui implique une latence supplémentaire. LISP [28] contient déjà le protocole de communication entre les MSes et les routeurs.

Assouplissement des politiques de sélections de routes

Dans cette section, nous assouplissons le processus de décision du routage inter-domaine et proposons une règle de stabilité assurant que l'ensemble du système reste stable.

Besoin de stabilité

Motivation pour l'assouplissement de la sélection de chemins

BGP est communément configuré pour respecter les règles "Valley Free" et "Prefer Customer" [31].

La règle "Valley Free" spécifie qu'un NSP n'exporte à ses peers et à ses providers que des routes provenant de ses clients. Cette règle a encore du sens dans un contexte multi-chemins car un domaine ne voudra sans doute pas fournir de service de transit entre ses voisins non clients.

La règle "Prefer Customer" spécifie que, dans le cas où un NSP recevrait des routes provenant de clients, de peers et de providers, celles provenant des clients ont la priorité sur les autres et que seule une route cliente peut alors être sélectionnée. Alors que cette règle a du sens dans le cadre de l'Internet actuel (c.a.d.,

dans le cas où un seul chemin est sélectionné), il est très probable qu'un fournisseur de transit veuille sélectionner simultanément des routes clientes et des routes peers/providers dans le cadre d'un internet multi-chemins. L'assouplissement de cette règle peut donc devenir intéressant.

La figure 8 illustre l'augmentation de diversité apportée par l'assouplissement de cette règle. AS_1 reçoit les annonces de routage permettant de joindre l'AS destination par l'intermédiaire d'un client (AS_4), d'un peer (AS_2) et d'un provider (AS_3). Avec la règle de préférence du client, AS_1 ne peut propager à AS_0 que le chemin annoncé par le client (en rouge) alors que d'autres chemins existent. Assouplir cette règle permettrait l'utilisation des autres chemins.

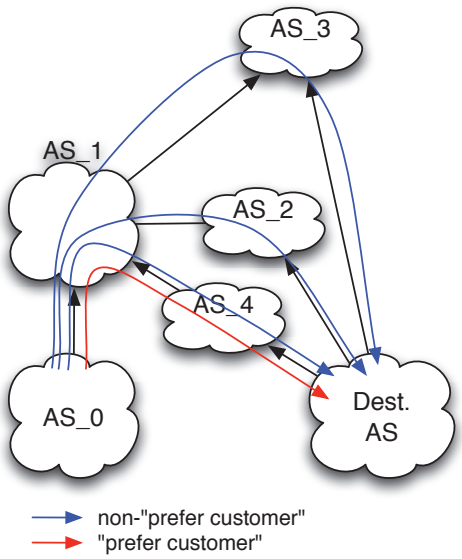


FIGURE 8 – diversités potentielles dans les deux cas : "Prefer customer" et non-"prefer customer".

Notations

Chaque domaine d est connecté à un ensemble de clients, de peers et de providers, respectivement notés comme $C(d)$, $Pe(d)$ et $Pr(d)$. Le domaine d reçoit :

- Un ensemble de routes venant de $C(d)$ (c.a.d., la diversité cliente) : $R_{d,C}$ (où C représente l'ensemble des clients).
- Un ensemble de routes venant de $Pe(d)$ and $Pr(d)$ (c.a.d., la diversité peer/provider) : $R_{d,P}$ (où P représente l'ensemble des *peers et providers*).

d peut exporter un ensemble de routes, à un voisin n, différent de celui exporté à un autre voisin m, même si les deux entretiennent les même relations commerciales avec d (c.a.d., peer, client ou provider).

Pour chaque voisin n, le domaine d utilise un process de sélection de routes λ_d^n afin de sélectionner, parmi les deux ensembles de routes reçues, celles qui seront exportées vers n : $E_d^n = E_{d,C}^n \cup E_{d,P}^n = \lambda_d^n(R_{d,C} \cup R_{d,P})$ où E_d^n représente l'ensemble des routes sélectionnées par d et exportées vers n, $E_{d,C}^n \subseteq R_{d,C}$ l'ensemble des routes clientes sélectionnées et $E_{d,P}^n \subseteq R_{d,P}$ l'ensemble des routes peer/provider sélectionnées.

Il est important de noter que nous ne fournissons pas, ici, de formulation explicite d'un processus de sélection de routes λ. Nous fournissons un assouplissement conséquent laissant aux NSPs le pouvoir de développer de nouveaux processus basés sur leurs propres contraintes.

Condition de stabilité IDRD

L'assouplissement trop important des processus de sélection peut entrainer des oscillations des annonces de routage à l'échelle de l'Internet [32]. Nous proposons alors un critère de stabilité – en plus des règles "Valley Free" [31] – permettant de s'assurer de la stabilité de l'internet dans sa globalité.

Les règles "Valley Free" s'expriment, à l'aide des notations précédemment détaillées, de la façon suivante :

- Chaque AS d envoie un ensemble de routes clientes/providers/peers ($E_{d,C}^x \cup E_{d,P}^x$) à un client x.
- Chaque AS d envoie seulement un ensemble de routes clientes ($E_{d,C}^x$) à un voisin peer ou provider x. $\lambda_d^x(R_{d,P}) = \emptyset$, donc $E_d^x = E_{d,C}^x = \lambda_d^x(R_{d,C})$

Sous l'hypothèse qu'il existe une hiérarchie entre opérateurs et qu'aucun domaine n'est provider d'un de ses providers, le critère de stabilité suivant, associé avec les règles de "Valley Free", assure la stabilité de l'Internet dans sa globalité.

> Critère de stabilité IDRD : Les routes reçues de peers ou de providers ne doivent avoir aucun impact sur la selection des routes clientes envoyées vers un peer ou un provider.
> Plus formellement, $\forall x \in P(d)$, nous avons :
> – $E_d^x = E_{d,C}^x \cup E_{d,P}^x = \lambda_d^x(R_{d,C}, R_{d,P})$
> – avec $\forall R'_{d,P}$ et $R''_{d,P}$:
> – $\lambda_d^x(\mathbf{R_{d,C}} \cup R'_{d,P}) = \mathbf{E_{d,C}^x}$
> – $\lambda_d^x(\mathbf{R_{d,C}} \cup R''_{d,P}) = \mathbf{E_{d,C}^x}$

Par souci de clarté, ce critère de stabilité peut également être formulé d'une manière un peu moins concentré. Le critère souligne que l'ensemble des routes clientes sélectionnées, qui doivent être envoyés à un voisin peer ou provider, doit être indépendant de l'ensemble des routes provenant de peers ou de providers. Il est important de noter que la sélection des routes clientes en vue d'être propagées aux voisins clients peut dépendre des routes reçues des peers/providers. Le critère de stabilité décrit plus haut équivaut à la combinaison de deux processus de sélection distincts :

- Un pour les voisins peer/provider : $E_{d,C\cup P}^{x\in P} = E_{d,C}^{x\in P} = \lambda_{d,C}^{x\in P}(\cancel{\boldsymbol{R_{d,P}}}, R_{d,C})$

– Un pour les voisins clients : $E^{x\in C}_{d,C\cup P} = E^{x\in C}_{d,C} \cup E^{x\in C}_{d,P} = \lambda^{x\in C}_{d,C}(\boldsymbol{R_{d,P}}, R_{d,C}) \cup \lambda^{x\in C}_{d,P}(R_{d,P}, R_{d,C})$

Nous souhaitons clarifier le fait que le fait de propager plusieurs routes n'est pas une source d'instabilité. L'instabilité vient de l'assouplissement du processus de sélection des routes. Le tableau 1 illustre les différents cas en met en avant

	"Prefer client"	New stability criterion	Full relaxation
Single path	**BGP : Stable**	Equivalent to "Prefer Client" : Stable	Unstable
Multi-path	Multi-path BGP : Stable	**IDRD : Stable**	Unstable

TABLE 1 – Identification de la cause d'instabilités.

lesquels peuvent entraîner de l'instabilité.

Evaluation

Le provisionnement des routes disjointes est, à court terme, l'une des utilisations les plus importantes de la propagation de la diversité de routage. Dans ce contexte, nous évaluons le nombre de chemins disjoints qu'un domaine peut utiliser pour atteindre tous les autres ASes. Comme nous ne pouvons nous baser sur des données de routage classique (par exemple, Route Views) en raison de leur dépendance au processus de sélection de BGP., nous avons mis en place une méthodologie d'évaluation dédiée.

Diversité disjointe maximale potentielle

Nous noterons C_s et C_d le nombre de domaines voisins, respectivement pour l'AS source s et l'AS destination d, et $P^{dj}_{s->d}$ le nombre de chemins disjoints disponibles pour que s puisse atteindre d. $P^{dj}_{s->d}$ subit la contrainte suivante : $max(P^{dj}_{s->d}) \leq min(C_s, C_d)$. En effet, le nombre de chemins disjoints entre deux domaines ne peut pas être plus élevé que le nombre de voisins du domaine source ou du domaine destination.

En raison de cette contrainte, nous évaluons d'abord le nombre maximum de chemins disjoints pouvant être utilisé pour joindre chaque AS. CAIDA [29] fournit la liste des relations commerciales entre les paires de domaines voisins [1] sur la base duquel nous pouvons évaluer le nombre maximum de chemins disjoints pouvant être utilisés pour joindre un domaine donné, qui est intrinsèquement limité par le nombre de ses voisins. En ne tenant compte que d'ASes multi-homés (c.a.d., 23526 ASes au total), on obtient la courbe noire représentée dans les figures 10 et 9. Comme nous prenons uniquement en compte les domaines multi-homés,

1. Datant du 16 Janvier 2011.

chaque domaine a une accessibilité maximale disjointe qui est strictement supérieur à 1.

Nous sélectionnons ensuite des ASes voulant joindre le reste de l'internet via des chemins disjoints. Nous recréons le graphe de l'Internet à l'aide des données fournies par CAIDA et mesurons, pour chaque destination, combien de chemins disjoints sont utilisables.

Résultats

Les figures 10 et 9 montrent la diversité disjointes utilisables dans les deux cas de respect de la règle "prefer customer" et de respect de la règle de stabilité précédemment énoncée. La courbe noire représente nombre maximum de chemins disjoints. Par exemple, une importante proportion des ASes (environ 50%) sont connectés à deux domaines tandis que les autres sont connectés à 3 voisins ou plus.

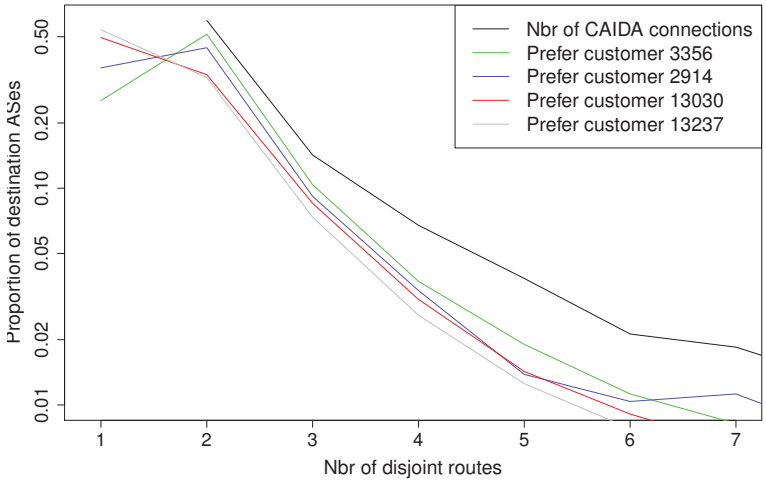


FIGURE 9 – Chemins disjoints : règle "Prefer customer".

Le premier résultat est que la règle "prefer customer" tronque de manière conséquente la diversité disjointe. Toutes les courbes des domaines analysés sur la figure 9 sont nettement en dessous du la diversité potentielle. Cependant, la diversité disjointe fournie par la nouvelle règle de stabilité (cf. Figure 10) est très proche de la diversité disjointe potentielle.

Avec la règle "prefer customer", le nombre d'ASes non accessibles par des chemins disjoints varie beaucoup en fonction de l'AS_{Source}. Par exemple, l'AS 3356 (rang CAIDA : 1) ne peut pas accéder à 25% des destinations par des chemins disjoints tandis que ce chiffre monte à 50% pour l'AS 13237. Utiliser la règle de

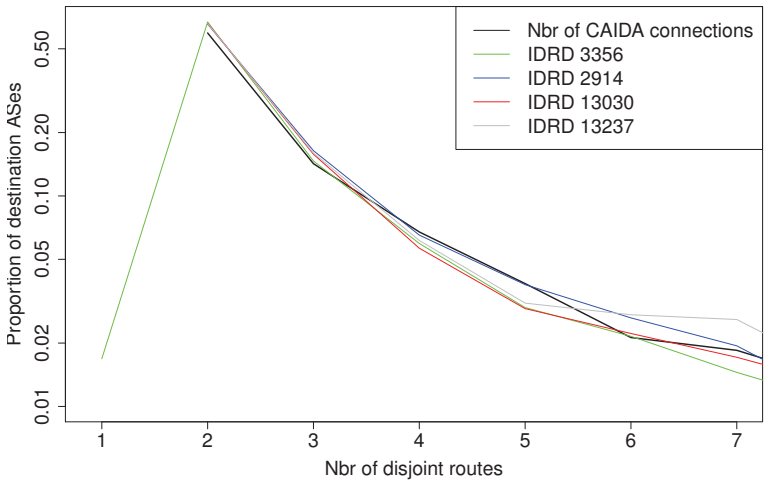


FIGURE 10 – Chemins disjoints : règle de stabilité IDRD.

stabilité d'IDRD permet à presque tous les domaines d'être accessibles par des chemins disjoints.

Nous focalisons ensuite sur les destinations n'étant pas accessibles par des chemins disjoints puisque l'obtention de chemins disjoints constitue probablement la première utilisation d'une architecture tel qu'IDRD. La figure 11 présente l'évolution du nombre de destination non accessibles par des chemins disjoints dans les deux cas, "prefer customer" et IDRD.

Cette figure met en avant que la règle "prefer customer" ne propose pas le même niveau de chemins disjoints que l'on soit tier one, tiers two ou plus petit. En effet, le nombre de destination non accessible par des chemins disjoints, déjà important pour les tiers-ones (environ 25%), augmente fortement pour les domaines dont le rang CAIDA est inférieur à 20 (c.a.d., plus de 99% des domaines). L'assouplissement de la sélection des routes et le critère de stabilité IDRD est quasiment insensible au rang CAIDA de l'AS source. Cela signifie que trouver des chemins disjoints peut représenter une motivation pour adopter IDRD, pour tous les domaines, petits ou grand.

Analyse de passage à l'échelle des schémas d'identifications de chemins

Besoin d'identification

L'architecture proposée spécifie deux points clefs. Le premier est l'utilisation de l'encapsulation afin de forcer les paquets à suivre un chemin qui n'a pas été

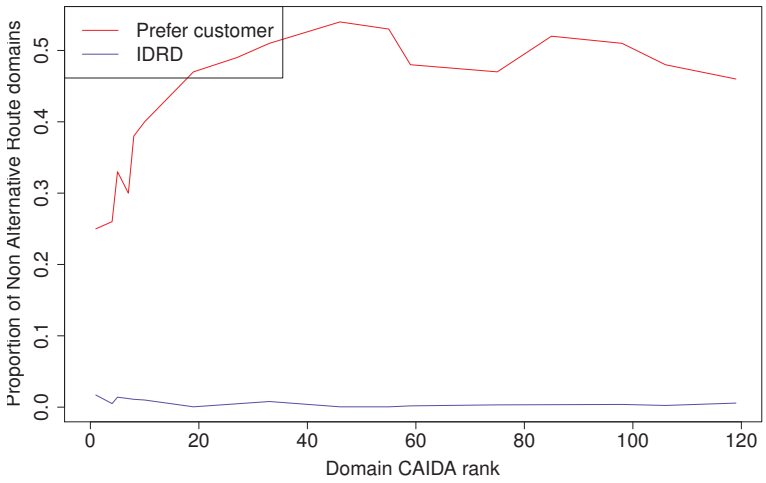


FIGURE 11 – Proportion d'ASes destinations sans chemins disjoints.

spécifié par le protocole de routage sous-jacent. Le second point est l'utilisation d'un identifiant de chemin, transporté au sein même de chaque paquet, permettant de sélectionner, à chaque étape, le chemin sur lequel le paquet doit être envoyé. Cette identification de chemin peut prendre la forme d'un label MPLS, d'un LISP instance-ID...

La signification de ces labels doit être connues au niveau des routeurs cœurs de l'Internet, afin d'effectuer la bonne décision de routage. De manière grossière, la multiplication du nombre de chemins servant à atteindre une destination devrait multiplier le nombre d'entrée dans la table de routage des routeurs. Néanmoins, un schéma d'identification efficace peut potentiellement permettre de ne pas augmenter de manière trop importante leurs tailles.

Où le problème de passage à l'échelle peut-il apparaitre

Au niveau du point d'identification des flux (point A)

Au niveau du routeur A (cf. figure 12) est effectuée la transition entre le domaine IP et le domaine où les paquets sont encapsulés. Ce routeur doit donc connaitre l'association entre les caractéristiques des paquets IPs et les chemins qu'ils doivent suivre. Par analogie avec MPLS [33], nous appelons cette association la FEC (c.a.d., Forwarding Equivalent Class). Le nombre d'entrées dans la FEC dépend du nombre de préfixes IP et du nombre de chemins par préfixes.

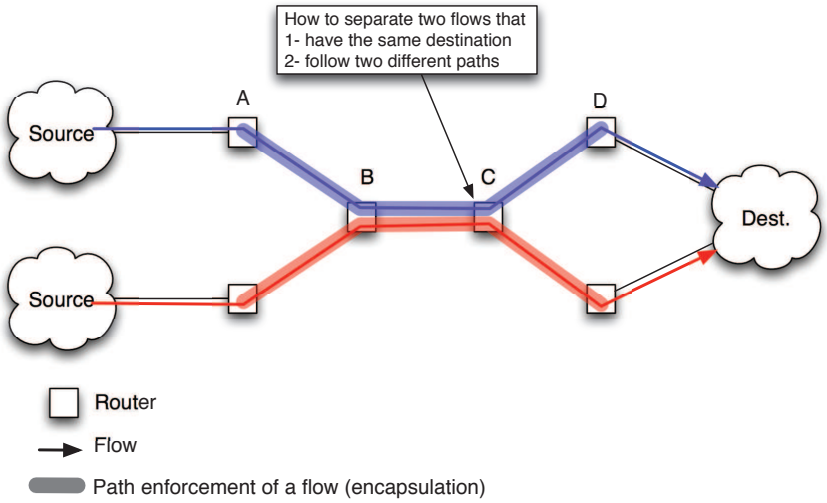


FIGURE 12 – Points d'identification.

Au niveau des routeurs coeurs (points B and C)

Au niveau des points B et C, seuls les chemins à suivre sont identifiés. En effet, deux flux suivant le même chemin peuvent être identifiés de manière similaire.

Au niveau du routeur de sortie du tunnel (point D)

Au niveau du routeur D, la décision de routage n'est effectuée qu'en fonction de l'adresse IP de destination. Il n'y a donc pas de problème de passage à l'échelle dû à l'adoption du multi-chemin à ce niveau là.

Analyses de passage à l'échelle des différents schémas d'identification de chemins

Nous effectuons ici une analyse de passage à l'échelle de l'identification des chemins dans le cas où aucun filtrage de route n'est effectué par les domaines (c.a.d., tous les domaines propages toutes les routes dont ils ont connaissance).

Stub-to-Stub Path-ID :

Dans ce schéma d'identification, les paquets sont encapsulés dès la sortie du domaine source et contiennent dès lors le path-ID identifiant le chemin à suivre. Ces paquets ne sont décapsulés qu'à l'arrivée au domaine destination (cf. figure 13).

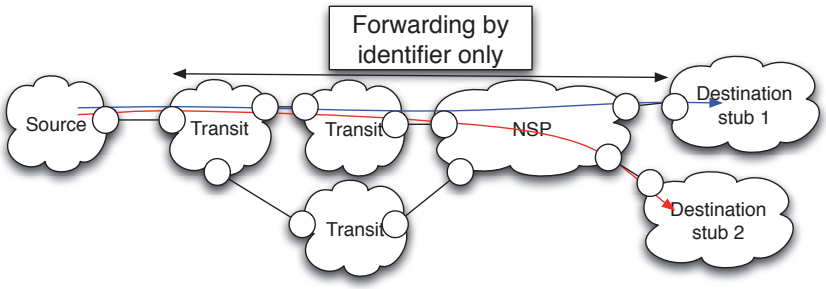


FIGURE 13 – Stub-to-Stub path-id.

Stub-to-Transit Path-ID :

Dans ce schéma d'identification, le Path-ID n'est utilisé que jusqu'à l'avant dernier domaine du chemin d'AS (cf. figure 14). Défini ainsi, le Path-ID identifie la suite de domaines transit traversés par le flux. Une fois que le paquet arrive au dernier transit, celui ci est dé-capsulé et la décision de routage est effectuée grâce à l'adresse IP de destination du paquet originel.

Le nombre total de domaine de transit étant largement inférieur au nombre total de domaines, cela a potentiellement un impact positif sur le nombre d'entrées dans la table de routage des routeurs cœur.

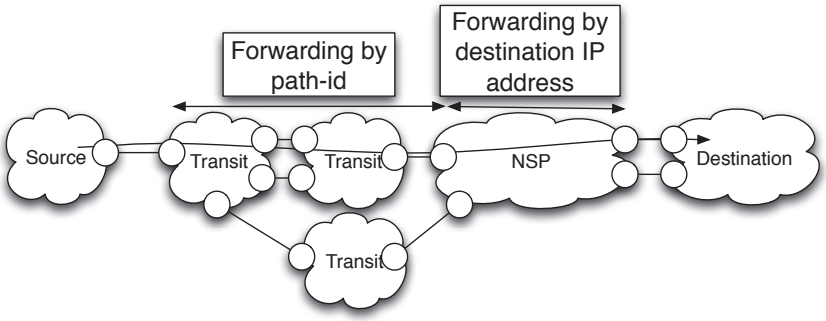


FIGURE 14 – Stub-to-Transit path-ID use.

Evaluation "pire cas" : impact sur les potentiels goulets d'étranglements

Les deux schémas d'identification peuvent avoir un impact sur la taille de la table de routage des routeurs cœurs et sur la taille de la FEC des routeurs de

bordure, entre le domaine purement IP et le domaine d'encapsulation IDRD (c.a.d., point A sur la figure 12).

En adoptant le point de vue de plusieurs domaines, nous évaluons dans cette section le nombre de chemins qu'un domaine peut recevoir et insère dans sa table de routage si aucun filtrage de route n'est appliqué.

La figure 15 montre l'évaluation du nombre de Path-IDs qu'un domaine devrait prendre en compte dans le cas de l'utilisation de tous les chemins potentiels. L'évaluation porte sur les deux schémas d'identification des chemins et aussi bien sur le passage à l'échelle de la FEC que sur celle de la table de routage.

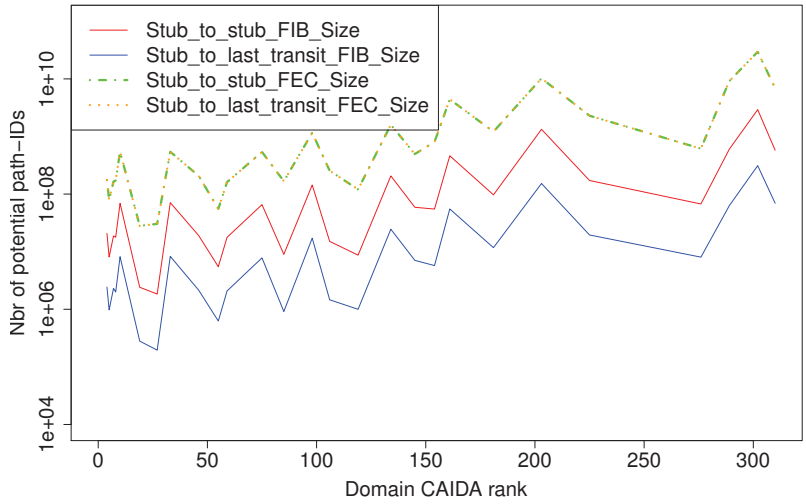


FIGURE 15 – Nombre de path-IDs nécessaire en fonction du rang CAIDA.

Le nombre potentiel d'entrée dans la table de routage varie soit de 10^6 à 10^9, soit de 10^5 to 10^8, suivant le schéma d'identification des chemins. La taille de la table de routage varie beaucoup d'un AS à l'autre car le nombre de chemin potentiel dépend fortement du nombre de voisins de l'AS considéré.

Le schéma le plus avantageux est le stub-to-transit où un path-ID est utilisé pour encodé le chemin d'ASes transits, sans prendre en compte les stubs, dont l'identification est contenue dans les adresses IP source et destination du paquet d'origine. De plus, le nombre d'entrées varie en fonction du rang CAIDA de l'AS analysé. Un Tier-one reçoit moins de chemin que les plus petits domaines. Ceci s'explique par l'application des règles de "Valley Free" car les tiers-one, n'ayant pas de provider, ne reçoivent que les annonces clientes de leurs voisin, ce qui limite leur nombre.

De même que pour la FIB, la taille de la FEC est importante (entre 10^7 et 10^{10} entrées) pour les deux types de schéma d'identification de chemin.

Ces résultats ne sont pas encourageant car les nombres d'entrées dans la FEC

et dans la FIB sont très important en regard de la capacité des équipements actuels (c.a.d., ordre de grandeur 10^6 entrées). De plus, le nombre d'entrées augmente sensiblement pour les ASes de faible taille (cf. rang CAIDA). ce qui rend le multi-chemins très difficilement adoptable.

Filtrer ?

Comme souligné dans la section précédente, propager tous les chemins de l'internet entrainerait d'importants problèmes de passage à l'échelle aussi bien au niveau de la FEC que de la FIB. Nous explorons ici des hypothèses raisonnables permettant de filtrer les annonces de routage et ainsi diminuer sensiblement le nombre d'entrées dans ces tables sans retirer au routage multi-chemins ses avantages.

Hypothèse à propos de la FEC

Tout d'abord nous posons comme hypothèse que les réseaux stubs ne veulent envoyer du trafic qu'à une proportion limitée de préfixes IPs. Pour cela, nus basons notre raisonnement sur les mesures de Mikians et al. [34] qui mettent en avant que seulement 61 000 préfixes destination sont largement utilisés (c.a.d., joints par un nombre conséquents de domaines – c.a.d., plus de 3 000 domaines). Ces préfixes sont ceux ayant la plus forte probabilité d'être joint par de multiples chemins. Nous pouvons donc poser comme hypothèse que le nombre de préfixes qu'un domaine stub veut joindre est borné par ce nombre, faisant ainsi décroitre le nombre maximum de préfixes de la FEC à 61 000.

De plus, nous pouvons poser comme hypothèse qu'un domaine stub n'a besoin que d'un nombre limité de chemins (une dizaine) afin d'accéder à un préfixe destination. Ceci nous amène à une FEC composée d'environ 610 000 entrées qui est un nombre envisageable avec les équipements actuels.

Hypothèse à propos de la FIB

Les chemins pouvant servir à contacter une destination ont un nombre d'ASes très différents. En effet, certaines destination peuvent être desservies par un chemin de longueur 3 et un autre chemin de longueur 13. Afin de réduire le nombre d'entrée dans la FIB, nous souhaitons limiter la différence entre le nombre d'ASes du plus petit chemin et les chemins sélectionnés.

Nous proposons donc de filtrer les chemins reçus et propagés. Un domaine reçoit un ensemble de chemins P_d pour atteindre une destination d et sélectionne un chemin $p \in P_d$ si et seulement si :

$$AS\_path\_length(p) \leq min_{x \in P_d}(AS\_path\_length(x)) + \delta$$

δ est le nombre d'ASes supplémentaires, comparé au chemin le plus court, autorisés pour qu'un chemin soit sélectionné.

Il est raisonnable de penser que les NSPs fourniront des chemins associés à des services. Alors que diminuer le nombre d'ASes sur un chemin a généralement un impact positif sur les caractéristiques du chemin (meilleur délai...) cela peut rendre la recherche de chemin disjoint plus difficile. En effet, par exemple, ne prendre que les chemins ayant le nombre minimum d'ASes ($\delta = 0$) peut empêcher un NSP de fournir des chemins disjoints, ce qui est considéré comme un service important. Il y a donc un compromis à trouver afin de réduire le nombre de chemin (c.a.d., réduire δ) sans empêcher de trouver des chemins disjoints.

Pour cela, nous avons effectué une évaluation afin de trouver la valeur de δ tel qu'il est possible de trouver des chemins disjoints pour toutes les destinations (résultats dans le tableau 2).

		Difference between 2 disjoint paths				
		-2 ASes	-1 AS	0 AS	1 AS	2 ASes
AS	rank	Percentage of destination ASes				
3257	5	2.1%	15%	59%	22%	1.9%
2914	7	2.0%	13%	60%	22%	1.7%
174	8	2.1%	15%	59%	21%	2.0%
3303	19	1.8%	13%	59%	24%	1.4%
13030	27	1.8%	13%	53%	29%	2.3%
6762	33	1.3%	10%	59%	27%	1.9%
4589	55	1.7%	13%	58%	25%	2.0%
4436	98	3.8%	22%	49%	23%	1.9%
8426	106	3.2%	19%	55%	21%	1.8%

TABLE 2 – Différence de longueur entre le chemin BGP ($\delta = 0$) et le plus petit chemin disjoint.

Le tableau 2 nous permet de conclure que $\delta = 2$ permet d'atteindre une partie significative de l'internet par des chemins disjoints et nous prendrons cette valeur pour la prochaine évaluation.

Evaluation de la taille de la FIB

Nous évaluons dans cette section la taille de la FIB en prenant en compte la valeur de δ précédemment calculée.

La figure 16 présente l'évaluation avec $\delta = 2$. Il est clairement montré que le nombre d'entrée dans la FIB décroit d'un ou de deux ordre de grandeurs. En utilisant le schéma d'identification des chemins Stub-to-Transit, le nombre maximum d'entrées à insérer dans la FIB est alors de $4 \cdot 10^6$, ce qui est juste un ordre de grandeur supérieur à ce que les routeurs actuels doivent faire face. Ceci semble alors un nombre raisonnable.

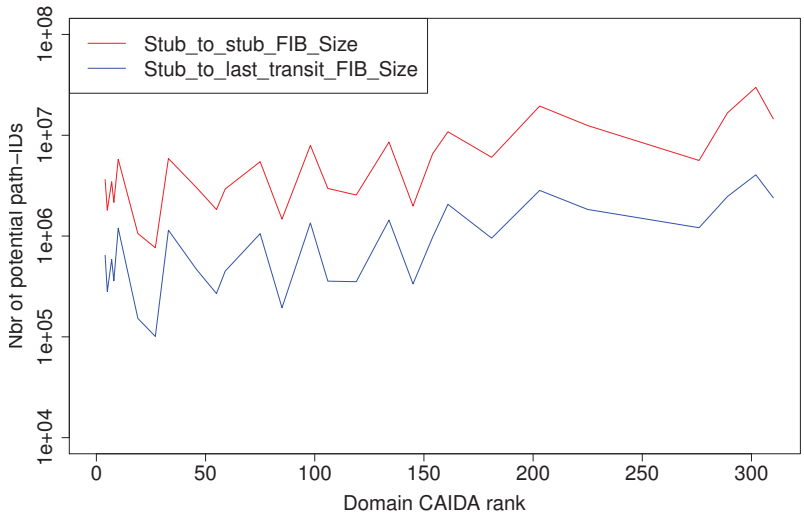


FIGURE 16 – Nombre de path-IDs en fonction du rang CAIDA ($\delta = 2$).

Vente de routes inter-domaine par un processus inspiré des enchères

Contexte

Dans la section précédente, nous avons vu que ne pas filtrer la propagation des routes entraîne un important problème de passage à l'échelle. Nous mettons aussi en avant que de petites adaptations et quelques filtrages permettent de diminuer drastiquement ce problème afin de le rendre contrôlable.

Néanmoins, les propositions de la section précédentes ne reposent que sur du filtrage sans négociation avec les voisins. Kwong et al. [35] suggèrent l'utilisation de mécanismes de marché afin de limiter l'augmentation du nombre de routes de l'Internet et nous adoptons, dans cette section, la même perspective.

Dans le contexte de cette section, le NSP souhaitant propager ses routes les filtres en fonction des desiderata de ses voisins (cf. figure 17). Filtrer les routes en fonction des besoins des voisins permet de mettre en place une nouvelle approche commerciale de la propagation de routes car chaque voisin peut potentiellement payer afin d'obtenir les routes qu'il souhaite.

Néanmoins, une bonne connaissance des besoins de chaque voisin est nécessaire, ce qui semble difficile d'obtenir. Il est donc nécessaire de mettre en place un processus d'association afin de sélectionner, pour chaque voisin, l'ensemble de routes à lui envoyer. De plus, le calcul du prix de chaque route est difficile car chaque route a des caractéristiques uniques et le prix dépend de l'intérêt que les voisins ont de cette route.

Le but du présent travail est donc de répondre à ces deux questions : 1/ A quels voisins envoyer quel ensemble de routes ? 2/ Quel prix devra payer chaque voisin ?

Le processus d'allocation proposé permet à un NSP d'offrir des routes additionnelles à ses voisins afin qu'ils puissent obtenir la qualité qu'ils souhaitent ou effectue de l'ingénierie de trafic avancée. Néanmoins, les chemins fournis peuvent traverser des ASes ne participant pas au processus d'allocation et le NSP n'a aucun contrôle sur ces ASes et sur les caractéristiques de ces routes. Ce processus alloue donc un privilège d'utilisation des routes et aucune réservation de ressource (QoS) n'est effectuée.

La propagation des routes aux voisins suit le processus suivant :

– Etape 1 : **Publication des chemins** : Le NSP (le vendeur) envoie à ses voisins les informations associées à l'ensemble des routes (e.g., chemin d'ASes, caractéristiques du chemin, résultats de mesures [36]...) et les détails du processus d'allocation (par exemple le type of mécanisme...).
– Step 2 : **Les enchères** : Chaque voisin envoie un ensemble d'enchères, chacune comprenant un ensemble de routes et le prix que le voisin est prêt à payer cet ensemble.
– Step 3 : **Association entre voisins et chemins et calcul des prix** : Après avoir reçu les enchères, le vendeur est en mesure d'associer pour chaque voisin, un ensemble de routes à lui propager et de calculer le prix qu'il doit payer.
– Step 4 : **Configuration du chemin et paiement** : Le vendeur alloue les routes en les configurant dans les routeurs intermédiaires, de telle manière que chaque gagnant est en mesure d'utiliser les routes qu'il a gagné et en paie le prix associé.

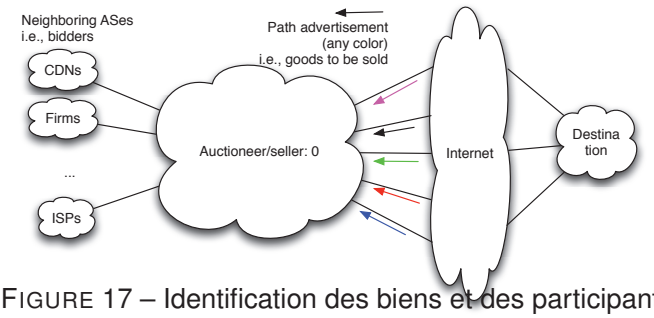


FIGURE 17 – Identification des biens et des participants

D'un point de vue du contexte du routage inter-domaine, le processus d'allocation se doit de prendre en compte certaines contraintes. Tout d'abord à la vue de l'évolution du prix de la bande passante aux points de peering (une chute de deux ordres de grandeur [37]), nous posons l'hypothèse que le vendeur de routes est capable de prendre en charge tout le trafic que ses voisins voudrait lui envoyer et est donc en surcapacité, au regard de la demande de bande passante.

De la même manière qu'actuellement, une route peut être allouée à plusieurs voisins. Ainsi, sans contrainte de bande passante, une route peut être fournies à

tous les voisins. Nous considérons donc qu'une route est un bien infiniment duplicable.

Ainsi, les caractéristiques du routage inter-domaine sont les suivantes :

- les voisins peuvent concourir dans le but de gagner plusieurs routes. Le processus doit alors être combinatoire.
- Les routes peuvent être dupliquées indéfiniment.
- le processus d'allocation ne doit générer que peu de messages. Un processus effectué en un seul coup (c.a.d., non itératif) doit être préféré.
- Une route n'est allouée à un voisin qu'une seule fois. En effet, une route étant considérée ici comme une autorisation d'utilisation d'un chemin, recevoir plusieurs fois la même autorisation n'apporte rien de plus au voisin.

Propriétés requises du processus d'allocation

De plus, nous considérons les propriétés suivantes comme essentielles à l'adoption d'un tel processus par les NSPs.

- **Implementable** : Avec ses 450 000 prefixes IPv4, le routage inter-domaine doit déjà traiter une quantité importante d'informations. Le processus d'allocation doit alors pouvoir passer à l'échelle et être implémentable.
- **"Truthful"** : Un mécanisme est "Truthful" si les joueurs (ici les acheteurs potentiels) disent la vérité, dans le prix qu'il sont capable de payer pour obtenir chaque route. Cette propriété permet au vendeur de connaitre les types de routes qui sont réellement intéressants, afin de mettre en place les futures allocations, et empêche les potentiels acheteurs de manipuler les montants qu'ils annoncent afin de modifier le prix qu'ils devront payer.
- **Maximiser le revenu du vendeur** : Le processus doit fournir un revenu substantiel au vendeur.
- **Ne pas connaître l'intérêt de chaque voisin** : L'évaluation des prix de chaque route est nouveau dans le monde du routage. Le vendeur ne connait donc pas à combien chaque route peut être évaluée par les acheteurs potentiels.

Ces propriétés sont étudiées dans le cadre de la théorie des enchères. le processus d'allocation que nous proposons est largement inspirée de cette théorie, tout en restant compatible avec les contraintes fournies par le contexte inter-domaine.

Processus d'allocation

Conventionnellement, l'élection des gagnants d'une enchère est effectuée en maximisant le bien être social. Ceci n'est pas possible dans notre cas car le nombre de bien à vendre est potentiellement illimité.

Il est intéressant de remarquer que si le vendeur ne sélectionne aucun gagnant, son revenu est de zéro. De même, si le vendeur sélectionne tout le monde et fournie toutes les routes demandées, sont revenu est aussi nul car chaque participant, se sachant par avance gagnant, n'annoncera qu'une valeur nulle pour chaque route. Il doit alors exister un revenu maximum que le vendeur pourrait gagner. C'est

ce revenu que nous visons et la sélection des gagnant du processus d'allocation sont sélectionnées tel que le revenu du vendeur est maximisé. Le maximum est illustré figure 18.

Nous adoptons ici une approche itérative afin de trouver le point de revenue maximum, tel qu'illustré dans l'algorithme 1.

Algorithm 1 Calcul de la courbe de revenu

$Revenue\_max = 0$;
$Winners\_max = 0$;
ForAll $Winning\_bids \subset bids$ **do**
$Revenue = compute\_revenue(Winning\_bids)$
if $Revenue > Revenue\_max$ **then**
$Revenue\_max = Revenue$
$Winners\_max = Winning\_bids$
end if
EndFor

Un bid (c.a.d., une enchère) est composé d'un ensemble de routes et d'une valeur que l'acheteur potentiel serait prêt à payer s'il est élu vainqueur. Afin d'éviter de parcourir l'ensemble des sous ensembles des enchères soumises (les "bids") nous classons les bids par valeur moyenne.

Nous utilisons alors la définition suivante.

Definition 1. *Une élection du gagnant par la moyenne est une élection des gagnants telle que si un joueur i gagne grâce à un de ses bids, chaque joueur j dont la moyenne d'un de ses bids et supérieure au bid gagnant de i gagne aussi.*

Ce type d'élection est intéressant car aucun bid perdant ou aucun ensemble de bids perdants ne peut être utilisé pour remettre en cause l'élection d'un gagnant.

Theorem 1. *Quand les biens sont infiniment duplicables, élection du gagnant par la moyenne est à l'épreuve des coalitions de bids perdants.*

La fgure 18 représente le revenue que pourrait gagner le vendeur en fixant la valeur moyenne séparant les bids gagnants des bids perdants. Le point proposant le plus de revenu est le point $\mathfrak{M}$. Chaque bid dont la valeur moyenne est supérieure à $\mathfrak{M}$ est un bid gagnant. Dans le cas où un joueur aurait fourni plusieurs bids gagnant, celui ci gagne les routes contenues dans le bid de plus forte valeur.

Exemple d'une fonction de paiement simple

Dans cette section, nous analysons les caractéristiques d'une fonction de paiement simple basée sur le bid du premier perdant. En dépit de son apparente simplicité, cette fonction de paiement fourni aux joueurs la motivation pour dire la vérité quand chaque joueur ne fourni qu'un seul bid.

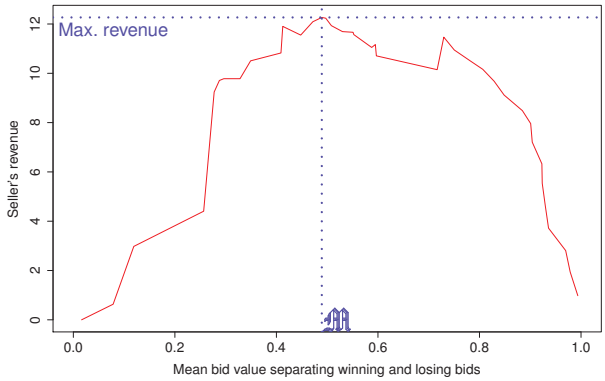


FIGURE 18 – Example d'évolution du revenu du vendeur.

Présentation de la fonction de paiement

Definition 2. *Un paiement **first-loser-mean-bid** est une fonction de paiement où les gagnants paient chacun des biens gagnés la valeur moyenne du bid le plus haut fourni par les perdants :*

$$\begin{aligned} \textit{pour chaque } g \subseteq G,\ p(g) &= |g| \times \mathfrak{M} \\ \textit{et } \mathfrak{M} &= max\{\overline{v_i} | i \in L\} \end{aligned}$$

Voici un petit exemple d'application du processus d'allocation avec 4 acheteurs potentiels (c.a.d., 1, 2, 3 and 4) souhaitant acheter les 2 biens a et b. Le tableau 3 présente un exemple de bids, qui sont ensuite triés par valeur moyenne dans le tableau 4 et traités afin de calculer le revenue du vendeur et l'allocation des biens à chaque valeur dans le tableau 5.

En ayant trié les bids par valeur moyenne, le vendeur est maintenant capable de scanner toutes les combinaisons de gagnants en parcourant toutes les valeurs moyennes de la plus haute à la plus basse (cf. tableau 4). A chaque étape est calculé le revenu du vendeur, à l'aide de la fonction de paiement **first-loser-mean-bid**, dans le but de sélectionner l'allocation apportant le maximum de revenu.

A la première étape, le vendeur spécifie que seuls les participants ayant misés plus de 7, comme valeur moyenne, sont sélectionnés en tant que gagnant (cf. première colonne dans le tableau 5). Le participant (1) gagne donc le bien b et le prix de chaque bien gagné est de 6 (c.a.d. la valeur moyenne du plus haut bid des perdants), ce qui fourni un revenu de 6 au vendeur.

A la seconde étape, le vendeur sélectionne les bids dont la valeur moyenne est supérieur à 6. Le prix est alors de 5 par bien et seuls les participant (1) et (2) gagnent. Le revenu du vendeur est alors de 10.

Le revenu est calculé à toutes les étapes et le vendeur sélectionne celle lui fournissant le plus de revenu (c.a.d. l'étape 3, en rouge), avec un revenu de 15).

	(1)	(2)	(3)	(4)
a	-	6	5	2
b	7	-	5	-
ab	11	9	6	3

TABLE 3 – Bids

$\Rightarrow$

Mean	Bidder	Goods
7	(1)	b
6	(2)	a
5.5	(1)	ab
5	(3)	$a \oplus b$
4.5	(2)	ab
3	(3)	ab
2	(4)	a
1.5	(4)	ab

TABLE 4 – Bids ordonnés en fonction de leurs valeurs moyennes

$\Downarrow$

	Minimum winning mean bid							
	7	6	5.5	5	4.5	3	2	1.5
(1)	b	b	ab	ab	ab	ab	ab	ab
(2)	$\emptyset$	a	a	a	ab	ab	ab	ab
(3)	$\emptyset$	$\emptyset$	$\emptyset$	$a \oplus b$	$a \oplus b$	ab	ab	ab
(4)	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	a	ab
Route price	6	5	5	2	2	2	0	0
Revenue	6	10	15	8	10	12	0	0

TABLE 5 – Calcul du revenu ($\oplus$ représente l'opération XOR)

Propriétés interessantes

Nous prouvons que ce processus d'allocation réuni certaines propriétés intéressantes, quand il est utilisés dans certaines circonstances.

Véracité

Tout d'abord, dans le cas où chaque participant ne fourni qu'une seule mise, nous prouvons que chacun d'entre eux a une incitation à donner une valeur de mise (c.a.d. une enchère) qui reflète parfaitement le montant qu'il est capable de payer. Cette propriété est importante car elle permet au vendeur d'éviter que les participants ne tentent de manipuler l'allocation des routes (et le prix payé) en modifiant la valeur de la mise.

Forme la grande coalition

Ensuite, l'utilisation de cette fonction de paiement forme la grande coalition. C'est à dire que chaque participant potentiel ne perd rien à participer au processus d'allocation. Et, plus important, le vendeur est enclin à accepter tous les nouveaux participants au processus d'allocation. Cette propriété résulte d'un contexte spécifique. En effet, le contexte de routage inter-domaine dans lequel nous nous plaçons permet aux participant d'échanger de l'utilité (c.a.d. l'utilisation de la route allouée) avec le vendeur et réciproquement. Néanmoins, les participants ne sont pas en mesure d'échanger cette utilité entre eux car, une fois une route allouée à

un participant, il lui est impossible de la partager ou d'en transmettre la propriété à un autre participant.

Faible complexité

Enfin, le processus, dans sa globalité, ne nécessite que peu de calculs. En effet, la complexité est polynomial aussi bien en fonction du nombre de participants que du nombre de biens à vendre. Nous avons implémenter ce processus et avons simulé des allocations pour un grand nombre de routes à vendre (c.a.d. de 16 à 128 routes) et un grand nombre de participants (c.a.d. de 20 à 3 000 participants) et l'allocation a été traitée en moins d'une seconde dans le pire des cas. Cette faible complexité est intéressante dans un contexte de routage inter-domaine, où les sélections de route doivent être effectuées très rapidement.

Acronyms

ACM: Association for Computing Machinery
ADSL: Asymmetric Digital Subscriber Line
AS: Autonomous System
ASBR: Autonomous System Border Router
ASPL: AS Path Length
BGP: Border Gateway Protocol
CAIDA: The Cooperative Association for Internet Data Analysis
CDN: Content Delivery Network
COMNET: COMputer NETworks
DFZ: Default Free Zone
DNS: Domain Name Server
DSCP: Differentiated Services Code Point
eBGP: external Border Gateway Protocol
ED: École Doctorale
EDITE: École Doctorale Informatique, Télécommunication et Électronique
ENST: École Nationale Supérieure des Télécommunications
FD: Flap Dampening
FEC: Forwarding Equivalence Class
FIB: Forwarding Information Base
iBGP: internal Border Gateway Protocol
IDRD: Inter-Domain Route Diversity
IEEE: Institute of Electrical and Electronics Engineers
IETF: Internet Engineering Task Force
IGP: Interior Gateway Protocol
INRIA: Institut National de Recherche en Informatique et en Automatique
IOS: Internetwork Operating System
IP: Internet Protocol

ISP: Internet Service Provider
LCAF: LISP Canonical Address Format
LISP: Locator/ID Separation Protocol
LMD: Local Mapping Distributor
LP: Local Preference
MED: Multi-Exit Discriminator
MIRO: Multi-path Interdomain ROuting
MP-TCP: MultiPath Transmission Control Protocol
MPLS: MultiProtocol Label Switching
MS: Mapping System
NHLFE: Next Hop Label Forwarding Entry
NIRA: A New Internet Routing Architecture
NS-BGP: Neighbor Specific Border Gateway Protocol
NSP: Network Service Provider
NTT: Nippon Telegraph and Telephone Corporation
PCE: Path Computation Element
RCN: Root Cause Notification
RCP: Routing Control Platform
RFC: Request For Comments
RIB: Routing Information Base
STAMP: SelecTive Announcement Multi-Process routing protocol
TCP: Transmission Control Protocol
TE: Traffic Engineering
TIPP: TheTopology Information Propagation Protocol
UBPS: Upper Bound Payment Space
VCG: Vickrey Clarke Groves
VF: Valley Free
VPN: Virtual Private Network
VRF: Virtual Routing and Forwarding
WDP: Winner Determination Problem
XOR: eXclusive OR
YAMR: Yet Another Multipath Routing

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 The Internet

The Internet was born in the late 60's as a research/academic network [1, 2]. The first network service providers emerged in the late 80's and began to provide commercial services in the 90's. Since then, 2.4 billions people have been connected through the Internet [3]. Although it may appear as an important figure, it represents 34% of the worldwide population, which envisions an important growth potential.

Despite the Internet may be considered as already making part of the current society, its influence goes beyond the imagination. Indeed it allows a rapid and wide propagation of information and is often considered as a revolution comparable to writing or to letter press [4]. This revolution leads to a re-organisation of the society with consequences that reach a wide variety of fields, such as economy, religion or politic [5].

Whereas books, radio and television allow to share information in a single direction (from an emitter to a receiver), the Internet allows the use of bidirectional communications. This places each individual at the same level of any other institution. Furthermore, while elder technologies were specific for certain restricted types of information (e.g., text and image for books and sound for radio), the Internet allows the propagation of a wide variety of information types. Video, text and sound are already carried over the Internet and new information types begin to emerge, such as movement (e.g., telesurgery [6]) or brain-to-brain transmission [7].

Today's Internet is part of the daily life of lots of people and is considered as a great success. Nevertheless it is already hard to identify its whole current potential. Moreover its evolution, driven by the already identified future utilizations, may provide even more changes and revolutions. The current thesis, which deals with the enabling of the multipath in the Internet, takes place in such context. Indeed Internet multipath has already been envisioned in the literature. Nevertheless it has not been adopted by the community yet and its impact is not fully understood.

1.1.1 The Current Single Route Internet

Internet is composed of an important number of domains/autonomous systems (around 40 000 [38]), which own and propagate address prefixes. Every Autonomous Systems (AS) belongs either to a network service provider (around 6 000 transit domains) or to a stub network (around 34 000 stub domains), could they be universities, CDNs (Content Delivery Network) or firms.

While a given domain d can naturally reach (i.e., send IP packets to) its own prefixes, reaching address prefixes of other domains necessitates that the individual domains are mutually interconnected. Moreover, as d can not be directly connected to every other domain in the Internet, it must make use of so-called transit domains for the transportation of IP packets from/to non-adjacent domains. This transit domain may also have a limited reachability and therefore may also rely on a bigger transit network.

The Figure 1.1 illustrates such relationships. Domains connected with an arrow $\rightarrow$ have a *transit relationship*, which means that, from a commercial point of view, the small domain pays its providers to benefit from its transit service. Networks generally try to avoid using transit services in order to minimize the cost of their communications. Therefore they connect directly their networks with the Network Service Providers (NSPs) of approximately the same size. Unlike with the providers, exchanging traffic between equivalent size networks is generally free and allows to bypass providers and their costs. This commercial relationship is named “shared cost” or “peering” relationship (domains connected via a line — in Figure 1.1). nevertheless being connected to every other domains of the same size is not possible as, by definition, small domains do not geographically span the world. Transit services provided by bigger NSPs are therefore required to reach far domains.

A hierarchy naturally emerges between small and big domains [39]. At the top of the hierarchy are the Tier 1s, which are commonly defined as the domains that do not have providers (around 10 providers are tier 1s [1] [29]). Each of these domains is connected to other Tier 1s via a “shared cost” relationship. Besides Tier 1 domains, “smaller” domains that have some providers and some peers (i.e, “shared cost” relationships) are named Tier 2s. And domains who always reach other networks through a provider are commonly named Tier 3s.

1.1.2 Current inter-domain routing protocol

BGP (Border Gateway protocol) is currently the inter-domain routing protocol. As illustrated in Figure 1.2 a router running BGP first receives routing information (aka BGP updates) from neighbors, could they be external or internal. These updates are filtered at the entrance (e.g., to avoid loop) and stored in the Routing Information Base (RIB) which serve as BGP update cache (i.e., information stored but not always used). The RIB may contain several routes toward a single destination. The router then compute the best route thanks to the BGP decision

1. this figure is approximative as AS relationships are not public. This information comes from inferences [40, 41].

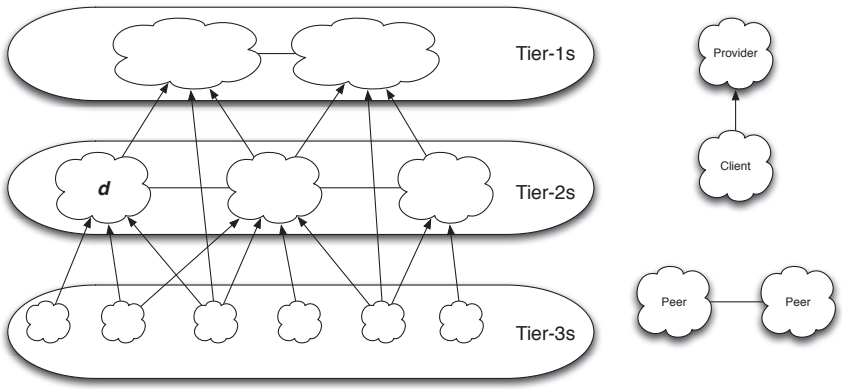


Figure 1.1: Internet hierarchy

process and put this route into the routing table of the router (called FIB: Forwarding Information Base). From that point, if some data packets are to be forwarded to this destination, only the route inserted in the routing table is used. Alternate routes cached in the RIB are not taken into account for packet forwarding. Then the routes contained in the routing table are potentially sent to neighbors.

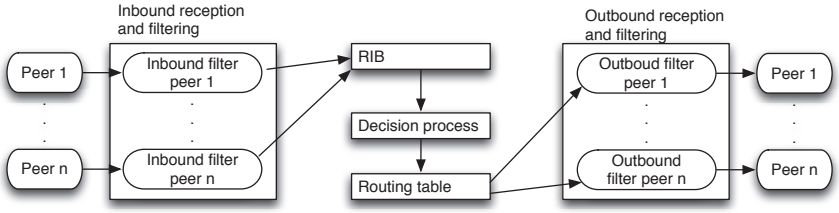


Figure 1.2: BGP process in a router

A domain is composed of several routers and a specific one may have simultaneously two different kinds of connection:

- Internal connections: in such a case the router is connected to other routers that belong to the same domain.
- External connections: in such a case, the router is connected to routers belonging to other domains. This router is named an ASBR (Autonomous System Border Router).

To handle these two types of technical relationships, BGP acts slightly differently in these two situations. We usually differentiate these cases by using different names, one for external BGP peering relationships (eBGP) and one for internal BGP ones (iBGP) (illustrated in Figure 1.3). A relationship is considered as external if the AS number of the neighbor differs from the local one. If the AS number is the same,

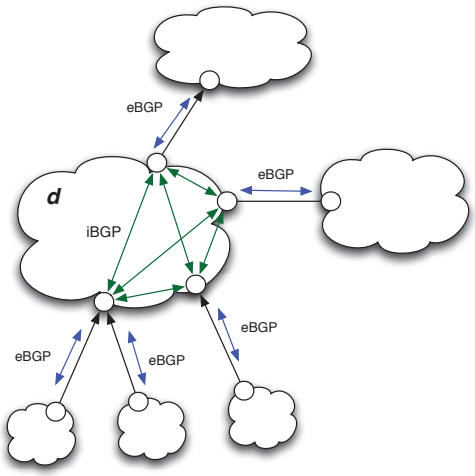


Figure 1.3: Internal and external BGP peerings

the relationship is internal to the domain. The BGP decision process, which selects the route which is to be inserted into the routing table, is declined with the following steps:

- Highest Local_Preference: the Local_Preference is a metric only propagated internally to a domain (i.e., via iBGP). It represents the preference the domain has for the received updates. It is used in general to rank neighbor domains according to their transit prices.
- Shortest AS path: Each BGP update contains the list of ASes/domains the path is composed of (i.e., the AS path). Choosing the shortest AS path update intends to minimize the distance to the destination.
- Lowest origin type: the origin type is an information included in each BGP update. It informs that the update as been created either at the real location of the prefix or from other BGP updates. The decision process prefers the updates generated locally to the network being announced.
- Lowest MED (multi-exit discriminator): the MED is a value communicated by the neighboring domains. In the case where two domains are connected by multiple links, one of the domains can ask (with the MED) to the other to send the packets through a specific link. The MED of several updates are compares if they come from the same neighbor.
- eBGP over iBGP: if several routes remain, BGP prefers external routes (eBGP) over internal routes (iBGP).
- Lowest IGP (Interior Gateway Protocol) metric: if several routes remain, BGP selects the one which the exit ASBR is the closest.
- Lowest Router ID: if several routes remain, BGP breaks the tie by selecting the route which next hop has the lowest router-id (the router-id of a router is generally on one of its IP addresses).

The “eBGP over iBGP” and “Lowest IGP metric” steps are considered as “hot potatoe” steps, meaning that the router chooses the route which allows for the autonomous system to get rid of the packets (i.e., its chooses the closest autonomous system’s exit).

It is important to notice that, whatever could be the routes that are compared by this process, only one can emerge. Indeed, if two routes have exactly the same attribute values, the last step (i.e., “Lowest Router ID”) selects one on a value that can not be the same (i.e., the IP address of the BGP next-hop) in both updates.

1.1.3 Internet policies

Taking into account the commercial relationships exposed in Section 1.1.1 (i.e., client, peer and provider), NSPs respect two fundamental best practices.

First a NSP may be able to reach a single prefix either via a client, a peer or a provider. In such a case, reaching this prefix via the provider is costly, whereas reaching it via a peer is free and reaching it via a client brings money[2]. Therefore the NSPs adopt the “**prefer customer**” rule which allows him to earn money. Technically speaking, this rule is configured thanks to the Local_Preference attribute [31] (i.e., first step of the BGP decision process explained in Section 1.1.2). When receiving a route from a client, the NSP put a higher Local_Preference than the ones of the routes coming from peers and providers. Then a router can compare peer/provider routes with client routes by just comparing the Local_Preference values. In addition to minimize the cost of transit, this rule is proven to ensure, among other rules, the whole Internet stability [31]. The configuration of a higher Local_Preference value to client neighbors is known as the Gao & Rexford condition [31].

Second, a NSP avoids providing the transit service to domains that are not its clients. It wants therefore to avoid all communication except if the sender or the receiver is a client. This filtering is called the “**Valley Free**” policy and is performed at the control plane level thanks to these two rules:

- A route coming from a client can be propagated to every other neighboring domain. This rule allows for all the domains, could they be clients, peers or providers, to reach the client of a NSP.
- A route coming from a peer or a provider can not be propagated to peers and providers. This filtering is performed by tagging routes coming from peers and providers with a specific community [42]. Then no route with this community is advertised to others peers/providers [43]. First it allows avoiding a NSP to propose transit between two providers/peers as none of these neighbors pay. Second, the propagation of the peer/provider routes to the clients provides them the reachability of the entire world.

The Valley Free rule does not prevent clients to reach the whole Internet as all peer/provider prefixes are propagated to clients and the client prefixes are propagated to clients/peers/providers.

2. Transit pricing between two NSPs is performed according to the amount of data transferred, as opposed to the flap-rate pricing generally applied at the end-user connexion.

Generally speaking, the "**prefer customer**" rule and the "**Valley Free**" policy are well applied [44, 45]. Indeed Network Service Providers have an incentive to use them. First when a customer route and a peer/provider route are available, as NSPs can only select and propagated one route, it is commercially interesting to select the customer route (i.e., and being compliant with the "**prefer customer**" rule) as the forwarded packets are sent through a link which makes the domain earn money. Second, in the case where a peer or a provider route is selected (i.e., no customer route is available), the domain has an incentive to not send it to peers and providers as doing so would make the domain pay twice - i.e., once for receiving data through a peer/provider link and once for sending the same data through another peer/provider link. This makes the domain compliant with the "**Valley Free**" policy.

1.1.4 Potential diversity

As domains are generally connected to several neighbors, several paths may be available to reach a single destination domain/network. First because each of the external connections may propose a path to reach the destination. Second because each neighbor may also receive several paths which could potentially be propagated. Figure 1.4 illustrates the potential path diversity a source domain may

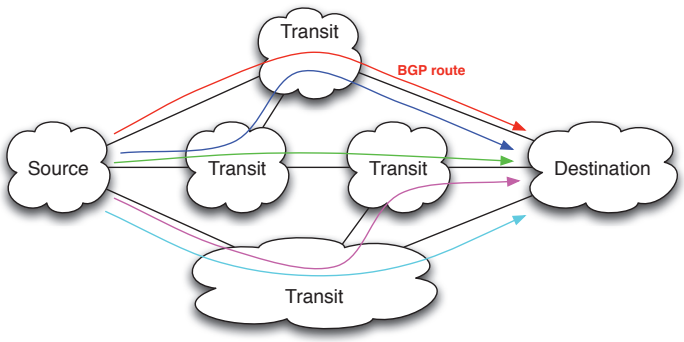


Figure 1.4: Inter-domain potential paths

benefit in order to reach a destination domain.

As underlined in Section 1.1.2, domains communicate with their neighbors using the Border Gateway Protocol (BGP) [46]. BGP fulfills the role of globally announcing prefix reachability between individual IP networks belonging to different administrative domains (i.e., ASes).

Mühlbauer et al. [47] underline that an important potential and un-exploited path diversity exists in the current Internet. Nevertheless the current inter-domain routing protocol BGP truncates all this potential by selecting only one route, potentially using an arbitrary tie break.

It is also interesting to notice that the BGP decision process steps hardly take into account the quality of the path. One may argue that the minimization of the AS path may lead to both minimize the propagation delay and the potential bottlenecks and points of failure. Nevertheless, as this step is in second position, it is not a predominant selection. Moreover, minimizing the AS path length is not equivalent to minimizing the length of the path. Indeed some domains may add a few number of routers whereas other domains may add an important number of routers in the path, leading to extra communication delays and potential bottlenecks. Last there are no obvious relation between AS path length and the path characteristics as other metrics may greatly impact the characteristic of a path (e.g., the bandwidth).

1.2 Motivations for Inter-Domain Multipath

The use of internet is very diverse. All the flows that cross it have not the same properties and do not need the same path characteristics. For instance on one hand Voice Over IP sends very few amount of data but requires very small propagation delays and very low jitter. On the other hand downloading a file may require the sending of an important quantity of data while not requiring a small propagation delay or jitter, compared to other uses. A lot of other application flows (e.g., mail, chat, streaming, P2P... [12]) require different profiles which are commonly expressed with criteria like delay, jitter, bandwidth and packet loss rate [13].

Even if there exists a lot of different application needs, a global over-dimensioning of network service provider capacities may fit all these needs. Nevertheless over-capacity is not generally adopted. Consequently a lot of congestion points exist in the internet, and half of them are located in inter-domain exchange points [11]. Bypassing these congestion points would probably make the path length increase, which increases the propagation delay, or go through links which price are higher. Path characteristics differ a lot and a path which is good for one application could be bad for another.

Beside the technical characteristics of flows and their adequacy with the paths, some specific flows are critical, in the sense that they should not drop for a long time, even if the underlying path fails. Such flows require at least a second usable path which is to be used if the first one fails (i.e., a disjoint path). Paths are currently not well protected by the convergence of BGP. For instance, path exploration can make the Internet converge in several minutes [48, 49], which is a too high convergence delay, even for the uses that do not specifically require high reliability (e.g., web browsing).

From a security point of view, some BGP weaknesses give the opportunity to malicious ASes to hijack prefixes (i.e., falsely originate prefixes) [50, 51]. It can either be used to impact payments of providers or to attract and analyse the traffic. Inter-domain multipath may also be used to detect prefix hijacking [52].

These needs could potentially be filled with the adoption of the "Routing as a service" paradigm proposed by Lakshminarayanan et al. [53]. They propose to exploit some alternate paths via deflecting the traffic through Internet checkpoints. The path which is to be followed by a specific stream becomes a resource that

is to be sold by NSPs to clients. This type of paradigm allows for both the use of paths adapted to the traffic characteristics and the unblocking of new income revenue for NSPs. Nevertheless this paradigm requires the availability and usability of different routes whereas, in the current Internet, only one path is selected to reach a destination.

From an NSP point of view, inter-domain traffic engineering (TE) is an incentive to develop inter-domain multipath. The simultaneous use of several paths is an opportunity to balance the load between end-to-end paths and to avoid bottlenecks, could they be close or far away. A lot of work have been done to allow intra-domain TE [54, 33, 55]. Nevertheless, from an inter-domain perspective, traffic engineering is limited as it only focuses on external links [56, 57] (e.g., for load balancing or to direct the traffic to a different neighbor). A perspective to unlock the traffic engineering capabilities is to centralize the routing plane, outside the routers, in order to get a global knowledge about the network. Some works [58, 59, 60, 61] have been proposed to separate the control plane from the data plane and expose the perspective given by such a separation. Nevertheless, they mostly focus on intra-domain traffic engineering.

1.3 Solution Requirements

In this section we define the required *design criteria* in order to produce an inter-domain multipath proposal which is realistically applicable in today's Internet, These *design criteria* are based on valuable insights into successful system design from [14] and [15].

1.3.1 Provided services

The deployment of a new Internet architecture is normally associated to substantial investments, and therefore it need to generate clear benefits and financial profit. Taking into account all the motivations underlined in section 1.2, we conclude that the adoption of an inter-domain multipath architecture must support the vast majority of new services, such as:

- Traffic engineering: The use of multiple paths should open new possibilities concerning load balancing, the selection of paths according to new criteria, etc.
- Path selection by the end users: For each application a user may select the best path among a set of paths offered by its provider(s).
- Simultaneous use of several paths, either by an NSP or by an end user: Both users and NSPs may choose to simultaneously utilize several paths either to balance the load or to perform recovery from failures.
- Extensive customization of route selection: In contrast to BGP, the new multipath architecture must allow for significant relaxations in the route selection process.

While we aim at supporting a plethora of new network services, all the specified functionalities are by no means mandatory. For instance, network service providers

may choose to manage the set of client routes strictly by themselves, i.e., without notifying the clients. In such a case, the clients benefit from path diversity without the need to take any actions on their own.

1.3.2 Control plane changes

In the current Internet, as a router is naturally connected to several neighbors, it already receives several paths to reach a single destination. Once these paths are received, they are inserted in the Routing Information Base (i.e., RIB) and the BGP decision process selects one path which is to be inserted in the Forwarding Information Base (i.e., FIB).

Therefore, it is necessary to **change the routing selection process** in order to select several paths to be used [3]. This change must be performed such the stability of the entire system is preserved [31].

Once the multiple paths are selected and inserted into the FIB, they need to be propagated to neighbors in order to be put into use for path diversity. The propagation of diverse routes increases substantially the usable diversity as the number of available paths is not anymore bounded by the number of neighboring domains.

1.3.3 Data plane changes

Once several paths are selected and inserted into the FIB, they should become usable by the user.A user must be able to choose the path he wants to follow.

In order to be able to perform this choice, **each path must be indexed into the FIB**, which enables multiple users to use different paths, each one using a different index. In order to specify the path which is to be followed, **data packets must transport the index of the path**. An extended/extra header must be used to transport the path index in each packet. An in-depth discussion about this indexing is available in Section 4.

1.3.4 Backward compatibility

The chosen solution/architecture must be compatible with BGP-4 as the current inter-domain routing protocol. Some applications necessitate special routes according to their technical or business requirements, while others operate well with the BGP route. Therefore, the plain BGP route must be included in the set of diverse routes in the inter-domain mutipath solution, as selected according to the BGP decision process [9].

1.3.5 Incremental deployability

The solution/architecture must be incrementally deployable. By this we mean that players in the Internet must be able to implement the new architecture on

3. Some relaxations of the BGP decision process already exist and we analyse them in Section 1.4.1.

their own, without having to synchronize with or wait for other actors. The direct consequences of this principle are:

- *Technical independence of actors.* A domain wanting to deploy the solution should not suffer from the choices made by its neighbors. Therefore, the implementation of the architecture within one domain must have no impact on neighbor domains which do not aim at deploying it, and vice versa.
- *Non-disruptiveness.* The adoption of the architecture must have no general impact on the way Internet routing is currently functioning. Therefore, if one domain migrates to a specific solution, other domains should not experience any changes with respect to the propagation of the (single) BGP best routes.
- *Intra-AS router inter-operability.* In order to be easily adoptable by a domain, the internal migration towards a new technology should be feasible in a gradual manner. Therefore, the migration must avoid excessive synchronized intra-network manipulations and changes. More precisely, a migration process requiring a closely coordinated reconfiguration of all routers within a domain must be avoided. To this end, conventional network elements and multi-path network elements must inter-operate seamlessly.

1.3.6 Incentive and cost effectiveness

By “incentive” we mean that an AS must be readily willing to adopt the novel architecture. The first domain that adopts it must be able to benefit from it immediately. In the case of inter-domain path diversity, early adopters must benefit from the routing diversity they already receive thanks to multiple eBGP peerings with neighbor domains, which is normally truncated by the BGP selection. Furthermore, a non multi-path domain connected to multi-path domains must easily identify value in migrating to the new solution. Firstly, it should benefit from the diversity it already receives (as previously described for early adopters), and secondly, it should even further increase its utility by setting up muti-path neighboring relationships with other domains.

“Cost effectiveness” is closely related to the incentive. It means that, despite the adoption of a new technology represents a cost, it must allow for either a higher revenue increase or a higher cost saving.

1.3.7 Reliance on existing technologies

In order to be easily adoptable, the architecture should primarily make use of existing technologies. In this context, we identify the following subcases:

- *Technologies already in active use.* This category is comprised of all the elements which are already deployed in the operational network (e.g., routers, switches, etc.). The less these elements need to be replaced or modified, the easier will the novel architecture be adoptable, both for technical migration and for economic reasons.
- *Additional technologies.* New elements may be necessary to benefit from inter-domain path diversity (i.e., elements which are currently not used in operational networks). Although most novel elements will not necessarily

need to be based on existing technologies, in order to achieve timeliness and cost efficiency of deployment, the re-use of readily available building blocks is preferred.

Moreover, designing the new architecture in such a way that it does not mandate excessive use of completely new technologies and component types should facilitate the migration and also lower the cost of adoption, at least in terms of change management.

1.3.8 Respecting Internet paradigms

The Internet is, by nature, decentralized and each Autonomous System comprising the Internet has sovereignty over its network and the choices it makes about routing and forwarding. The Valley Free propagation of routes may serve as a good illustration of local policies that translate the commercial policies of NSPs into technical configurations. Therefore, providing an inter-domain multi-path solution which could potentially bypass the Valley Free filtering is an example of solution which is doomed to failure.

1.3.9 Respecting NSPs' commercial information

Network technologies, internal architectures, and inter-AS relationships can be considered as business secrets, and it is important to bear in mind that architectures which necessitate the exposure of business secrets will likely not be adopted by the relevant Internet actors. E.g., the negotiated relationships between various Autonomous Systems are normally unknown to third parties. And in spite of the fact that this information may be inferred [62, 41, 40], NSPs are still not willing to provide it on demand. Therefore, a multi-path solution which, for instance, requires the centralized listing of ASes relationships will hardly be adopted by the NSPs.

1.3.10 Scalability

The number of routes a router has to manage has been identified as a growing potential issue [63].

Although some works [64] claim that FIB scalability will not represent the major problem in Internet routing in the years to come, addressing this issue nevertheless remains in the focus of research and engineering efforts (cf. LISP [28]).

In any case, the insertion of the available path diversity into the FIB would certainly put an additional burden on the available FIB memory space, and therefore the adoption of any novel inter-domain multi-path architecture must avoid further aggravations of this (potential) issue.

1.4 Related Work

Some work has been performed in the field of inter-domain path diversity. Despite the following solutions propose interesting approaches and advances for putting the multiple viable paths into use, a simple and effective solution, which covers

all the previously exposed requirements, has not yet been formulated. Indeed all the requirements exposed in Section 1.3 are very hard to unify into one solution/architecture.

For instance, some works only pursue a single goal. As a consequence they only focus on specific needs (e.g., failure recovery) that make proposals very expensive in terms of deployment efforts for such specific needs and hardly adaptable for other purposes. Other works are highly disruptive as they assume deep change in the routing paradigms and/or assume very close coordination between network service providers, which makes the adoption of such solutions very hard.

The solutions are analysed in the order of the number of compatibilities with the requirements presented in the previous section. For instance the first solutions only focus on either data plane or control plane changes and the last ones, more complete, focus on both data plane and control plane requirements.

1.4.1 Multipath BGP

In order to enable simultaneous use of multiple routes, the *Multipath BGP* extension [9] [10] slightly changes the BGP decision process.

Multipath BGP relaxes the decision process of BGP while remaining coherent with the choice made by the conventional BGP. The new decision process is composed by, at least, the selection on the local preference (i.e., highest one), on the AS path length (i.e., the shortest one) and on the IGP cost.

If several paths have the same values for these metrics, the router configures them into the forwarding table (i.e., the FIB) and balances the traffic in equal shares among these paths. Despite this multiple path use, *Multipath BGP* does not modify the propagation of routes. It only propagates the route chosen by the whole decision process (including the Tie Break) and the additional equivalent routes (i.e., that have the same metrics) are inserted in the FIB but not propagated.

This technique is already implemented by manufacturers into commercial routers and does not require any change in the network. Nevertheless the selection process of routes is static and there is no way, besides modifying the local pref or underlying IGP routing, to customize it. Moreover if several paths are chosen, the traffic is separated equally among the paths which is quite limited in term of traffic engineering. Last but not least, as the path diversity is not propagated, it is not possible to know upstream the amount of available diversity and to have details about it (i.e., AS paths...).

Even if this technique is very simple and already implemented, its current adoption is very limited. Augustin et al. [65] underlines that most of the already adopt load balancing techniques take place inside a single domain (i.e. IGP equal-cost multipath or MPLS traffic engineering) and that very few core networks enable BGP multipath capabilities in their routers.

1.4.2 Source selectable path diversity via routing deflections

Yang et al. propose in [17] a solution which allows the use of non-equal paths without creating loops. Each router selects several paths and the end-user is able

to select which path to be used thanks to a new field (i.e., a tag) between the IP and the TCP headers. This field is computed by each router to select the path and forward the traffic. The routers may select a set of routes that have not the same cost. Consequently, the authors propose two simple rules in order to avoid loop. Despite this solution is mainly focused on intra-domain diversity, Yang et al. briefly propose a way to influence the choice of the exit ASBR of the domain, thanks to this technique.

This proposal is very similar to Multipath BGP [9, 10] as several routes are inserted in the FIB, while only one is propagated. Contrary to Multipath BGP, this proposal allows for the end system to selects the route which is to be used. Nevertheless, as path diversity is not propagated, the only usable path diversity is the one received by the current routing protocols. Moreover the end user does not receive information about available paths and is not able to directly associate paths with tags. He must therefore explore the tag space to learn the possible diversity and must perform extra analyses (e.g., traceroute) to obtain information about the paths.

1.4.3 Path Splicing

Motiwala et al. propose Path Splicing [18] as a multipath proposal. This work focuses on the intra-domain but is extended to have an impact at the inter-domain level.

On the intra-domain side, Path Splicing rely on the information gathered by a link state routing protocol. By locally simulating perturbation in the topology, each router is locally capable to find alternate paths which property is almost the same as the one of the shortest path. It pre-computes backup topologies for arbitrary failure combinations by removing edges from the underlying topology or by setting high costs on some edges. Each backup topologies/routes are inserted into different forwarding tables. End systems insert a “shim” header in between the TCP and IP headers in order to control the path taken by the packets in the network by indicating, for each hop, which forwarding table to be used to forward the packet.

Despite no diversity is explicitly propagated, Path Splicing is able to use more paths than what is allowed by the already mentioned solutions [9, 10, 17]. Indeed as it relies on a link state routing protocol, path diversity may be inferred by the local analysis of the network graph.

On the inter-domain level, the same idea is presented. The shim header is used to select the exit ASBR of each intermediate domain. Therefore an entry ASBR looks at the extra header to deduce the exit point that is to be used. This entry ASBR must be aware of several exit ASBRs for a given prefix. Therefore the ASBRs must be full meshed in order to receive a diversity and this proposal could hardly be used in conjunction with a route reflector [66]. Moreover, as for the intra-domain use, this diversity is not propagated to neighbors. Contrary to the intra-domain use, in which each router is aware of the whole topology (thanks to the link state routing protocol), the inter-domain use does not propose any solution to obtain any information about the path an end-host may potentially use. It is

therefore obliged to perform extra data plane analyses (e.g., traceroute) to obtain information about the potential paths.

1.4.4 BGP Add-Path

BGP Add-Path [16] is presented as a BGP extension that allows the announcement of several paths per prefix. Each path is identified in the BGP message by a path identifier. Contrary to the Multiprotocol Extensions for BGP-4 [67], BGP Add-Path propagates routes that belong to the same routing instance. Therefore a router receiving several paths from the same neighbor (with different path identifiers) will compare them thanks to the BGP decision process as if there where no path identifier. The goal of this identifier is to well separate paths in there propagation and reception: two advertisements coming from the same neighbor and having two different identifiers advertise two routes, which will be both inserted into the RIB of the receiver. On the contrary, two advertisements that have the same identifier advertise the same route and only the latest advertised route will be inserted into the RIB. Only one path is selected to be put into the routing table. The extra paths are considered as extra paths and are only put into use (i.e., put into the FIB) if the primary path fails.

Such an approach may potentially be generalized in order to fully propagate the available path diversity. Nevertheless this proposal only addresses the issue of propagating multiple paths and does not propose any change in the way the path diversity is computed/selected by routers. Consequently, there is no method to select and use simultaneously several paths, which restrains its use to failure recovery. Nevertheless, even if BGP add-path does not propose such possibilities, it is a well advanced IETF work and requires few changes onto the existing BGP protocol. It can therefore be useful for the control plane of our solution.

1.4.5 D-BGP and B-BGP

Wang and al. [19] propose D-BGP and B-BGP as two interdomain routing proposals that propagate path diversity. The goal of these proposals is also to speed-up the recovery of BGP.

D-BGP extends BGP to make it propagate the most disjoint alternate path in conjunction to the best path. The two paths are propagated in the same routing advertisement and path diversity is compared thanks to the current BGP decision process. On the data plane level, no modification is proposed and only one path is put into use. An alternate route is used only when a failure occurs but is nevertheless kept into the RIB of the router. In case of link failure, Root Cause Notification (RCN) [68, 49] is used to localize the failure and then quickly deduce which alternate routes are appropriate (i.e., routes that do not include the failed link) to be used.

Unfortunately, RCN propagates some information (i.e., router level information) that NSPs may not want to reveal to there neighbors. Therefore Wang and al. also propose B-BGP as an extension of D-BGP that addresses this problem by

using bloom filters, then allowing the identification of several failed paths without identifying the precise failure.

This proposal seems easy to implement and backward compatible. Nevertheless the scope of utilization is limited to path recovery. It is therefore not possible to put several paths into simultaneous use.

1.4.6 Reliable interdomain routing through multiple complementary routing processes

Liao and al. [21, 69] also propose a solution to decrease the convergence time in case of failure and therefore avoid transient instabilities (such as path exploration). The proposal is based on the running of several BGP processes that compute complementary routes. This work focuses on a new protocol, named STAMP, which propagates the disjoint diversity.

The STAMP protocol makes each router running two BGP processes. The paths of the two processes are mostly disjoint paths. Once a failure is detected, packets are forwarded thanks to the route of the alternate routing process.

In order to obtain the most disjoint path, the authors propose a specific way of propagation, which results in the construction of a two disjoint path in the downhill section of the path (i.e., path from top level domains to the destination domain). STAMP only deals with the downhill section of a path as the authors prove that transient loops can only happen in this part of the path.

In order to achieve this, the BGP protocol must be slightly changed to introduce a “Lock” attribute, which is used to specify the default path, and all the providers must support several BGP processes. This proposal is very close to the inbound traffic engineering problem as a network specifies, with the “lock” attribute, the best path the inbound flow should follow. Nevertheless there is no way to be sure domains will follow this guideline.

This proposal can be easily implemented and adopted by network service providers. Nevertheless, this work only addresses the path recovery problem and do not allow for the simultaneous use of several paths. Furthermore, all the ASes of the path from the source to the destination must have adopted this proposal in order to obtain its benefits. Therefore early adopters hardly obtain benefits.

1.4.7 R-BGP: Staying Connected In a Connected World

Kushman et al. propose R-BGP [22] as a BGP extension that provides most disjoint routes. A domain receiving a route, from a neighbor, propagates to the same neighbor, in return, the most disjoint route, by comparison to the received route. As a result, the neighbor is aware of (at least) two mostly disjoint routes. R-BGP does not need modification on the data plane as extra paths are only used for fail-over purpose. In case of failure, Root Cause Information (i.e., a modified version of the Root Cause Notification [68, 49]) is utilize to specify the AS which contains the link failure. This allows the routers to know where the failure occurs and to precisely select the routes to be withdrawn. With such a process, the authors prove that no transient loop can occur.

On the forwarding plane, packets are forwarded onto different routing tables (i.e., one for the main routes and one for the backup routes), as for VPNs. Kushman et al. underline that it is theoretically possible to simultaneously use these two paths. Nevertheless no technical detailed are provided.

Despite R-BGP is presented as an extension of BGP, the authors do not specify how BGP must be changed. Indeed, this proposal requires that the route sent to a neighbor depends on the route it received from it, which leads to a Neighbor Specific BGP approach [70, 71] (i.e., selecting the route to be propagate according to the identity of the neighbor). Such an approach leads to important architectural changes but there is no architecture proposal that allows its adoption. An early adopter hardly benefit from the adoption of R-BGP. Indeed, this proposal relies on the paths sent back by neighbors, which must also adopt it.

1.4.8 MIRO: Multi-path interdomain routing

Xu et al. propose MIRO [23] as a complete multipath inter-domain architecture. It proposes to store the received inter-domain path diversity in a local device, separated from routers (e.g. with a RCP [59, 60]). This diversity can then potentially be shared with other domains in order to make them benefit from the local diversity.

At the data plane level, packets following an alternate route are encapsulated to enforce their path till the domain proposing the alternate path. In order to separate them from the conventional traffic, packets carry a flow identifier which is used at the forwarding plane to select the path to be used. Two packets having the same destination address and two different flow identifiers may therefore be forwarded to different paths.

This proposal seems to be the one which is compliant with the highest number of requirements exposed in Section 1.3. Indeed, this proposal addresses both data and control plane changes and is not focused on a single goal (e.g., path recovery...). Moreover it is backward compatible as it lets BGP running underneath and seems to be incrementally deployable. MIRO is close to our architecture. Nevertheless this proposal describes the concepts with a very high perspective and some of the key points, which we aim at addressing in the present dissertation, remain open issues.

For instance the separation between flows is performed thanks to an identifier inserted into the encapsulation header. Nevertheless the global organization scheme of flow identifiers remains an open issue whereas it can be an early bottleneck of the architecture.

Concerning message exchange, MIRO specifies that routes are *pulled* (i.e., sent on demand) and not *pushed* (i.e., sent without demand, as for BGP) in order to prevent from an explosion of the FIB size. The pulling of routes is a great change in the inter-domain routing context. While it considerably reduces the scalability issue of the number of routes an AS may learn, it prevents the domains from being aware of the potential diversity it may benefit.

1.4.9 NIRA: A New Inter-Domain Routing Architecture

Yang et al. propose in [25] a new routing architecture for Internet. Even if multi-path is not the core of the proposal, the architecture proposes a way to enable the use of the underlying path diversity.

As a path is generally composed by an uphill part (i.e., recursively from a NSP to one of its providers) and a downhill part (i.e., recursively from a provider to one of its clients) till the destination, this work proposes to make the sending host choose separately these two parts. Each top provider receives a block of addresses that it re-allocates to its clients and so on. Each domain receives a number of prefix that is exponential with regard to the number of upper provider layers it has above it (i.e., provider, provider of provider...). Finally the architecture allocates to each host several IP addresses, each one associated to a sequence of direct or indirect providers. One IP address identifies a path from the host to a top level provider. This path could be used as the downhill or the uphill part of a path. The sender chooses both the uphill part and downhill part (i.e., uphill from the point of view of the receiver) and uses the associated IP addresses as source and destination addresses in its packet header. In the uphill part, packet are forwarded thanks to the sender address and in the downhill part, packets are forwarded thanks to the destination address.

At the control plane level, a new inter-domain routing protocol (TIPP) is proposed to propagate topology information while BGP could be kept in the core network (i.e., top level providers). A client is able to directly choose the uphill part of the path but needs to rely on a name-to-route lookup service (comparable to a DNS) to obtain the potential downhill paths and the corresponding IP addresses.

This work underlines a new routing paradigm, which allows for an important path diversity utilization. Nevertheless this approach is very disruptive. At the data plane level, the forwarding process must be changed in order to forward traffic according to the sender IP address, which impacts every single internet router. Moreover the addressing plan is highly correlated to the internet hierarchy (i.e., which domain is a provider and which one is a client). There is a great concern in term of hierarchy dynamic as a Tier 2 network may become a Tier 1, which would make it change its address allocation and the ones of its clients.

1.4.10 YAMR: Yet Another Multipath Routing

YAMR allows domains to construct a set of alternate paths that could avoid the failure of each link of the default path, providing a way to quickly switch from the default path to an alternative one in case of failure. Once a domain selects a set of routes, it propagates this diversity to its neighbors. The default path is the same as the one selected by the BGP decision process. On the control plane level, routes are differentiated thanks to a new label field in the routing messages. On the data plane level, packets bring a new 32-bit field that is used to choose the path the packet is forwarded to.

In order to reduce the churn, a router which detects a failure does not propagate the withdrawal of the impacted route if an alternative path is known. The traffic

which should be sent through the failed link is forwarded across a deflection path close to the previous failed path. Packets are then forwarded to the alternate path without the sender being aware of it. This withdrawal retention makes the router hide a path failure.

[20, 26] describe YAMR as a protocol and focus on the algorithm used to find link disjoint paths. Data plane and control plane changes are addressed. Nevertheless, the proposition only deals with path recovery issue. Furthermore it seems that a significant number of domains must adopt YAMR before early adopters benefit from it. Finally, It does not address the issue of implementation and how this protocol can be adopted/deployed in the current Internet.

1.4.11 Pathlet routing

Godfrey et al. propose Pathlet routing [24]. Each domain contains one or several $vnodes$ (i.e., virtual nodes). A pathlet is defined as a set of consecutive vnodes and an end-to-end path is composed with a list of concatenated pathlets. Pathlet information are advertised to neighboring domains via a path vector routing protocol and only policy filtering is performed (i.e., there is no selection of a best path, contrary to BGP). On the data plane level, the senders choose/construct end-to-end paths by concatenating pathlets they received the advertisement. Packet forwarding is comparable with strictly source routing [72] as a new field, included in the packet header, is used to store the sequence of pathlet and specify the whole path. Each $vnode$ forwards the packets according to the most external pathlet identifier. Once it reaches the end of the pathlet, the identifier is erased and the forwarding is performed based on the next one - as for MPLS label stacking [33].

This proposal has the advantage to reduce a lot the forwarding table size of routers. Moreover Pathlet routing is compatible with several requirements (e.g, data and control plane changes, agnostic about the potential services...)

Nevertheless, some other crucial requirements are absent. Indeed, the sender needs to be aware of all the potential paths in order to make its choice. Therefore, despite the decrease of FIB entries into the core routers, the equipment at the frontier between the IP network and the pathlet Internet needs to know every single path (i.e., pathlet list) the end user could use. This leads to a scalability issue at the edge of Internet, which remains opened.

Furthermore Pathlet routing is quite disruptive as the inter-domain routing protocol and forwarding capabilities of routers must be completely modified to advertise pathlets and to forward packets according to pathlets' identifiers.

1.5 Contributions of this thesis

The goal of this thesis is to show how feasible are the propagation and the use of multiple paths. The thesis is organized as follows.

In the current chapter (i.e., Chapter 1), we first highlight in Section 1.1 the required information to understand the work presented in the next chapters. The required knowledge are from the routing field. Nevertheless, it is declined in both

technical (i.e., BGP) and commercial (i.e., Valley Free, Prefer Client) considerations. Indeed these two aspects are the two sides of inter-domain routing and should not be separated. We also underline, in Section 1.2, how the enabling of the inter-domain multi-path can leverage some new uses and perspectives as well for NSPs and for end users. Nevertheless, using the Internet path diversity is not easy. We underline in Section 1.3 a list of requirements that a solution must respect in order to be adopted by NSPs. Having in mind these requirements, we present a complete state of the art about the Inter-domain path-diversity in Section 1.4. We show that none of the previous works unify all the requirements presented in Section 1.3. Indeed, while some solutions are meanly theoretical (i.e., very little has been done to propose ways of implementing the proposed solutions), some are very restrictive in term of usage (e.g., they focus on path recovery) or propose a highly disruptive approach.

In Chapter 2, we propose a first architecture that allows for the local use of path diversity. Indeed we propose that NSPs cumulate the diversity that is currently received by eBGP and share it with their stub clients. This proposal is simple, already configurable in current commercial routers and is backward compatible. From a global Internet point of view, this proposal may seem limited, in term of diversity. Nevertheless, we show, in Section 2.5, thanks to an extensive evaluation (performed with real BGP data) that the diversity currently received by a NSP is already important. Therefore sharing it with the stub clients gives NSPs the opportunity to propose value-added services which does not require any coordination with neighboring NSPs. The presented local path diversity use is a first step to a global path diversity enabling. We also analyse this first step from the point of view of the stability of paths, which can be potentially used. We show that the way the NSP selects its paths has an important impact on the stability and that enabling path diversity may be an opportunity to better select the paths according to their stability.

In Chapter 3 we first generalize in Section 3.2 the architecture described in Chapter 2 in order to allow NSPs to share their path diversities among them. The architecture we propose is compliant with the requirements exposed in Section 1.3. We point out, through an evaluation covering the whole Internet graph in Section 3.4, that the “Prefer Customer” rule (described in Section 1.1.3) is restrictive in term of diversity. Indeed, it prevents NSPs to simultaneously obtain disjoint paths for reaching from 20% to 50% (depending on the source domain) of destination domains, which would be required to perform fast failover [4]. Therefore we propose the relaxation of the “Prefer Customer” rule. As this rule also allows for the Internet to remain stable, we propose a new criterion, in Section 3.3, that allows for both the stability of the global Internet and for the relaxation of the “Prefer Customer” rule. Our evaluation of the potential diversity, according to this new stability criterion, shows that the potential path diversity between a source domain and a destination domain is so large that the maximum number of disjoint paths between these domains is only limited by the numbers of their external connections.

4. We consider fast failover as the first incentive to adopt inter-domain path diversity and such a limitation would lead to the non-adoption of our proposal.

As previously underlined in Chapter 3, the path diversity allowed by the proposed stability criterion is very high and may thus lead to scalability issue. Solving this potential scalability issue is possible by filtering the received path diversity. A NSP can filter the routes:

- either by selecting the routes on its own. In this case, each NSP filters the received diversity according to the information which is propagated inside each route advertisement. This type of filtering is most likely to happen at the beginning of the adoption of the inter-domain diversity propagation, when NSPs do not have precise negociation framework with neighbors.
- either by negotiating with neighbors to obtain the interesting routes. Thanks to negociation between neighboring domains, each one of them may know which path diversity each one of its neighbors is interested in. In such a case, routes are not filtered after reception but rather filtered before being sent to neighbors.

The Chapter 4 analyses the case where NSPs propagate the full path diversity to their neighbors. Each one of them is responsible of its own selection. We consider this approach as a worst case evaluation. First we identify in Section 4.2.5 where the scalability issues of such an approach can arise. We show that, despite what is commonly admitted, the scalability issue can not only appear in the core routers but also at the edge of the Internet (i.e., at the boundary between stub networks and NSPs). We then analyse, in Section 4.3, how the path diversity can be organized and propose schemes of identifying the paths, leading to different evaluation about their scalability. We quantify the scalability issue of propagating the whole diversity of the Internet and show that some small changes may help to reduce this scalability issue.

We assume that this routing table size increase will be monetized by NSPs as they will probably not be willing to implement new routes at their expenses. Therefore in Chapter 5 we expose a techno-economic approach to sell and propagate interesting routes in order to leverage value added services related to route propagation. Beside the revenue income of the NSPs, it also has for consequences the reduction of the scalability issue underlined in Chapter 4, as only routes that are bought are inserted into the FIB. We set-up an auction-type framework where a domain sells routes to its neighbors and each neighbor compete by bidding an amount it is able to pay to win the routes it is interested in. As a route is infinitely duplicable by the seller, the framework can not fit into conventional combinatorial auctions and we propose a new allocation mechanism. We set up a new way to determine which bidders win the routes (cf. Section 5.4) and the prices they have to pay (cf. Section 5.6). We manage to set-up a framework which is truthfull when each bidder submit one bid, meaning that bidders bid the true valuations of the routes they compete for, and which form the grand coalition, meaning that the seller has an incentive to accept every new bidder and that every potential bidder has an incentive to participate to the mechanism. Moreover we show, in Section 5.7, that the framework computation time is polynomial regarding the number of bidders and the number of routes.

Last but not least, we draw in Chapter 6 the conclusion of the current work.

While we summarize the main results of the current document, we also underline how the presented work can be deepened with further works.

Chapter 2

Enabling the locally received path diversity

2.1 Introduction

In this chapter, we focus on the diversity a Network Service Provider can provide to its customers. The path diversity that is currently received by a NSP thanks to its number of external connection is essential for traffic engineering (TE) and robustness [73]. Unfortunately, end clients (i.e., stub networks) cannot benefit from this diversity as their providers only announce them one route per prefix. As already underlined in Section 1.1.2, this limitation is the result of the BGP decision process that determines one single best route per prefix (potentially using arbitrary tie-break rules).

Bypassing this BGP limitation and thus offering several routes would increase the routing flexibility, make the Internet more robust to failures, and improve performance [70, 74]. It would also offer a richer and more flexible set of high-level policies than currently available with BGP such as choosing routes based on their stability or performance.

The architecture we propose in this chapter is the first step of the use of the Internet path diversity. We design it in order to be compliant with most of the requirements exposed in Section 1.3. Indeed, whereas this proposal does not allow for the global propagation of path diversity (i.e., it allows for the use of the locally received paths), it has the great advantage to be simple, backward compatible and to propose the use of an already interesting amount of paths.

Deflection is a way to enforce the paths followed by packets [75] and benefit from Internet multipath routing. The architecture we propose deflects the traffic thanks to the mapping-and-encapsulation paradigm to take advantage of the path diversity currently truncated by the BGP decision process. On the one hand, encapsulation is used to dynamically construct tunnels that enable traffic deflection. On the other hand, mappings are used in the control-plane to manage the deflection. We use the available eBGP paths to construct mappings, where each BGP prefix is assigned a list of potential NSP egress points. We leverage the Locator/Id Separation Protocol (LISP) [28, 76] to construct our architecture but this

choice does not preclude the use of any other mapping-and-encapsulation solution (e.g., MPLS [33]). The choice of LISP is motivated by the fact that it provides an advanced mapping-and-encapsulation protocol that supports incremental deployment as well as a flexible control-plane. Moreover, LISP runs on top of IP which makes it deployable in virtually every network which is not the case of MPLS.

In this Section, we first show how to exploit diversity in an AS without interacting with its neighboring transit domains. We then show how Internet Service Providers (ISPs) can build value-added services (see Section 2.3) by offering diversity to their customers in order to help them selecting stable and/or fast paths. We evaluate the benefits of our approach with simulation on two large ISPs. We first study route diversity with different possible routing policies (i.e., which routes are offered to customers). We then determine the churn impact of these policies on the control-plane (i.e., the dynamics of mapping changes) and the data-plane (i.e., the dynamics of path changes).

This Chapter is organized as follows. We propose in Section 2.2 a mapping-and-encapsulation architecture leveraging the LISP protocol and allows a network to use the path diversity of its ISP. In Section 2.3 we present two distinct use-cases that benefit from our architecture and present five route selection process policies that can be implemented with our architecture. We then evaluate our proposition with trace driven simulations. We first describe a precise methodology to succeed such evaluation in Section 2.4 and extensively describe our results in Section 2.5. The evaluation shows that our architecture offers an important gain in term of diversity compared to BGP with a limited and controllable overhead. Finally, Section 2.6 starts a general discussion around our architecture and Section 2.7 concludes this work.

2.1.1 Locator/ID Separation Protocol (LISP)

LISP has been initially designed to make the Internet more scalable by separating core prefixes from end site prefixes [28, 77]. LISP has today a wider scope of applications and is considered in several domains such as Traffic Engineering or migration issues (see, e.g., [78]).

LISP is a Map-and-Encap mechanism whose data-plane consists of an encapsulation protocol and the control-plane consists of a *mapping system*. Several mapping systems have been proposed and are considered at the IETF [28]. This encapsulating scheme can be used to forward a packet through a deflection point and thus enforcing its path. The mapping system stores and distributes (push or pull) the *mappings* to the LISP encapsulating routers. A mapping links IP prefixes with a list of IP addresses that can be used as deflection points. When a LISP router receives a packet, it encapsulates it with a LISP header whose outer destination corresponds to the mapping value (the deflection point). The encapsulated packet is then forwarded until it reaches the LISP router identified by the outer LISP header destination address. This router decapsulates the LISP packet and forwards the inner packet to the end host identified by the inner packet destination address.

Each locator is assigned a priority and a weight. On the one hand, the priority determines the eligible mappings (i.e. the deflection points). On the other hand, the weight determines how the load must be balanced among the deflection points with the highest priority.

2.2 Architecture

This section proposes a solution to reveal path diversity in the control-plane and to ensure the use of this diversity in the data-plane.

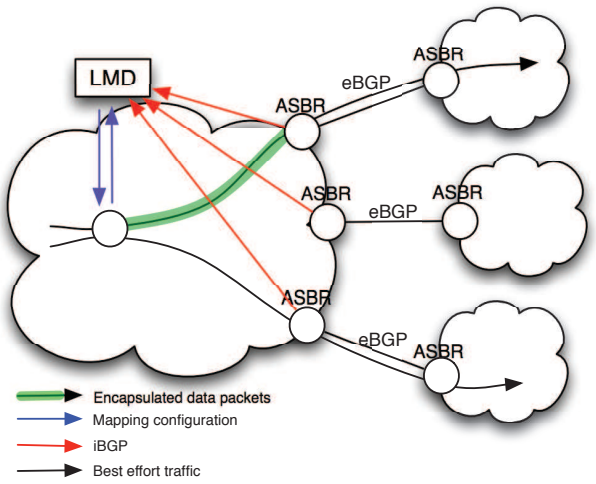


Figure 2.1: Intra-AS use of LISP encapsulation and mapping system

2.2.1 Description of the architecture

Our architecture offers to stubs the flexible capability of controlling the network exit point for their traffic. Depending on the use-case (see Section 2.3), stub networks can either choose their own exit points or choose the ones of their providers. The architecture to achieve this goal must respect the three following requirements:

$(R1)$ **local route diversity management** : the architecture must allow operators to implement their own route diversity policies,

$(R2)$ **path enforcement** : the architecture must provide a mean to enforce path until the chosen exit point of the domain (independently from BGP best routes),

$(R3)$ **incremental and local deployment** : the architecture must be practically, locally, and incrementally deployable. In other words, BGP should not have to be changed and an AS must have immediate benefits adopting the architecture even if there is no global deployment.

To this aim, our architecture relies on the Mapping-and-Encapsulation principle. On the one hand, the mapping part of our solution ensures that every BGP prefix

is mapped to a list of exit Autonomous System Boundary Routers (ASBR). On the other hand, the encapsulation part (i.e., tunneling) is used to enforce the path through the chosen exit ASBR. Such an exit choice may not be that of the BGP decision process. Fig. 2.1 summarizes our architecture.

Several solutions may be used for building the mapping (at the control-plane) [79, 80] and the encapsulation (at the data-plane) [81, 33, 28] parts. To the best of our knowledge, among these solutions, the LISP architecture [28, 82] is the only that provides at the same time mapping and encapsulation, and that meets all our requirements. LISP offers a complete mapping system which allows to manage and announce routing diversity to encapsulating routers ($R1$). It also provides an encapsulation scheme ($R2$) on top of IP ($R3$) for ease of deployment. Finally, LISP requires no changes to end-systems or to most routers, fulfilling the aim of an incrementally deployable protocol. LISP is still under development at the IETF but it is already a mature architecture and is available on the latest Cisco IOS releases.

2.2.2 Control-plane

At the control-plane level, our architecture relies on a centralized and extended mapping system that we call Local Mapping Distributor (LMD). In addition to storing and distributing mappings, the LMD first has to generate mappings from the diverse Internet routes. Every mapping associates a BGP prefix to the list of addresses of the exit routers that can be used to send packets to that particular prefix. Mappings are built from all the eBGP advertisements received by the AS. More precisely, each ASBR maintains an iBGP session with the LMD and propagates the routes it receives from its neighboring ASes. To benefit from all the potential path diversity and bypass current BGP policies, it would be very useful for ASBRs to activate BGP add-path [16] or BGP Best External [83].

As the chosen routes may no longer be congruent with BGP routing, some difficulties may arise at the level of the exit ASBR. When packets are decapsulated, a (BGP) routing decision may be taken at this point with the potential risks of: (i) forwarding the packet to another exit ASBR (the one providing the default BGP best route) and (ii) sending the packet to another neighboring domain that peers with the selected exit ASBR. To tackle this issue, each ASBR can be configured with several VRFs, one for each neighboring router it is connected to.

For every eBGP update message received, the ASBR sends it to the LMD without modifying the BGP next hop. The ASBR is therefore aware of both the exit ASBR, which is the BGP peer sending the update to the LMD, and the entry ASBR of the next AS, which is contained in the next-hop field of the BGP update. Each neighbor ASBR (belonging to a neighboring domain) is associated by configuration, in both the LMD and the exit ASBR, with a LISP instance-ID. Normally, the LISP instance-ID is used for identify the specific VRF that has to process a LISP packet. In our case, the VRF is directly associated with a given entry ASBR of the neighboring domain. Each data packet arriving at the exit ASBR, and containing an instance-ID, is forwarded according to the routing table from the VRF identified by the instance-ID. In other words, the instance-ID is used to distinguish routes in

order to use the appropriate exit links.

The LMD uses the received information and its local policies to construct the subset of ASBRs that can be used to reach the BGP prefix. Each such ASBR is assigned a priority indicating the one to use preferably. The way the subset of ASBRs is computed (i.e., the mapping function) and ranked is open. For example, one can choose to use a fixed number of exit ASBRs (e.g., 2), based on price whereas others can perform more complex selection process to insert, for instance, only disjoint routes that have proven reliability or stability. The use of performance evaluation tools may be used to rank paths based on real measurements [36, 84]. Examples of mapping function strategies are proposed and evaluated in Section 2.4.

The mapping configuration may be sent to the encapsulating routers either on-demand or directly. With the first approach, the encapsulating router send a request to the LMD when a flow must be forwarded to a path which configuration is not yet in the router. Packets are sent toward the BGP path during the LMD answering delay. If the mapping configurations are sent directly, the LMD configures the encapsulating routers without any demand from the routers. A deeper discussion about these two approaches is available in Section 2.6.1

2.2.3 Traffic forwarding

Devices encapsulating packets can be deployed in any part of the network. For instance, the encapsulation can be performed by a load-balancer deployed at the exit of a stub network could it be a multi-homed network or a single-homed one (e.g., small office, home network. . .).

When an encapsulating router receives a packet, it analyses the packet (e.g., destination IP address, DSCP field...) and determines, thanks to the mapping records, the IP address of the exit ASBR to use to deflect the packet and the instance-ID associated with the VRF of the exit ASBR that identifies the entry ASBR of the non-LISP NSP. Traffic is then encapsulated with a LISP header, which destination IP address is the exit ASBR as well as the instance-ID associated with the exit ASBR's VRF. Several entries may be used for the same type of traffic. Every such entry is assigned a priority and traffic is balanced between ASBRs of same lowest value priority. Nevertheless, the different packets of a stream will go through the same ASBR thanks to the hash based routing locator selection algorithm described in the LISP specification [28]. The ASBR that receives the packet decapsulates it and forwards it to the neighbor ASBR associated with the LISP instance-ID.

2.3 Deployment Use Cases

This section proposes a general scheme to reveal and use path diversity within an autonomous system. In this section, we propose two deployment use cases and five route selection policies that our architecture enables.

2.3.1 Stub's local diversity usage

Our architecture can be used by multi-homed stub networks, i.e., networks connected to several Internet service providers, to take advantage of the diversity offered by their different providers as illustrated in Figure 2.1. Our architecture allows to tune outbound traffic engineering, without requiring any form of coordination with the ISPs. As compared to existing techniques [56], the solution may offer a wider choice of route selection policies, such as differentiated route selection for different egress routers or more flexible use of load balancing on multiple paths.

2.3.2 Stub's provider diversity usage

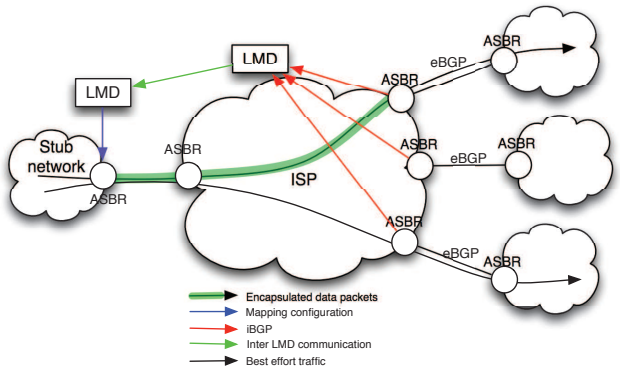


Figure 2.2: Path diversity provided by the ISP and transmitted to the stub network

The most novel and promising applications arise when the solution is implemented within an ISP. In this scenario the stub network and its provider settle for a mapping function. The ISP is responsible for collecting eBGP routes and uses the mapping function to construct mappings on behalf of its customer. As depicted in Figure 2.2, the ISP's path diversity is summarized in mappings and distributed to the stub via the LMD. For every outgoing packet, the stub network uses the mappings to determine the most appropriate ISP exit point to use and sends the packet, encapsulated, to this exit point. When the ISP's exit point receives such a packet, it decapsulates it and forwards the packet according to its forwarding table.

The ISP manages mappings for its customers, based on some pre-negotiated objectives (e.g., performance, stability) implemented by the mapping function. Some large stub networks with complex traffic engineering requirements may, on the contrary, prefer to implement their own route selection policies by managing their own Local Mapping Distributor (as represented in Fig. 2.2). Nevertheless, the needed data to make the architecture work are the prefixes and their associated ASBR exit points. All other information (e.g., AS path) are not necessary but may be useful to perform path selection.

Some paths can be more expensive than others hence the ISP might not reveal

1. Ignore routes having an unreachable BGP nexthop
2. Prefer routes having the highest local-pref
3. Prefer routes having the shortest AS-Path
4. Prefer routes having the smallest MED
5. Prefer routes learned via eBGP sessions over routes learned via iBGP sessions
6. Prefer routes having the closest next-hop
7. Tie breaking rules: prefer routes learned from the router with lowest router id

Figure 2.3: Vanilla BGP decision process

its full diversity to customers. However, in Section 2.4, we show that the diversity remaining after typical filtering policies is still of interest for large providers. A customer can thus benefit from this diversity to improve its robustness or traffic engineering (e.g., [73, 85, 86]). The ISP can easily deploy the service as it relies on existing building blocks (essentially LISP). We also show in the next section that the overhead (in terms of route changes or churn) is limited even while using the diversity of routes.

Our flexible architecture can be a real differentiation factor for ISPs and also an opportunity to develop added-value services, such as the route selection and prioritization service or the LMD interconnection service. For instance, the mapping system allows different levels of route diversity to be offered to different customers, based on their subscription. In particular, special peerings, shortcuts (e.g., with performance and/or protection guarantees) may be reserved for specific customers or traffic.

It is worth noticing that our architecture is not limited to stub ASes and can be used in the stub part of ASes. The distinction is important as transit ASes can have parts of their network that are not used for transit (e.g., the part of the network that connects residential ADSL customers). Thus, an ISP can announce its diversity to its customers, as long as customers use this diversity only in their stub parts. Moreover, an AS that is multiconnected can receive diversity from its different ISPs and implement its own LMD to deflect its outgoing traffic among the diverse routes of its ISPs. Nevertheless, our solution is restricted to stub networks and cannot be used in transit as it could cause loops and routing instabilities due to violation of Gao-Rexford rules [31].

2.3.3 Route Selection Process Policies

EBGP feeds provides the maximum available diversity usable by the network. The network selects the EBGP routes that fulfill policies implemented in the mapping function. Path selection policies implemented by mapping functions can be constructed in many several ways and are not restricted to the BGP decision process that we remind in Figure 2.3.

In the following, we propose different path selection process policies and evaluate them in Section 2.4. For the sake of illustration, we use a path selection process inspired by the BGP best path selection process because a lot of practically interesting aspects are taken into account such as commercial relationship (with the Local Preference), path length (with the AS path length), and loop avoidance. As the BGP decision process provides, at the end, only one route, we relax it to allow the selection of multiple paths. Nevertheless, the potential of the architecture is not limited to such decision processes and more advanced route selection may be performed (e.g., based on packet loss, contract with customers. . .).

ALL : selects all the routes independently of their metrics. This filtering policy is identified in the document as **ALL**.

LP : selects routes that have the highest local preference. Local preference is the metric reflecting, most of the time, the cost of the peering/transit link. Selecting the highest local preference is equivalent to minimizing the cost of the inter-AS forwarding. This filtering policy is identified in the document as **LP**.

ASPL : selects routes that have the shortest AS path length. AS path length is the a technical distance metric to the destination that ensures the absence of routing loops. Selecting the path only through this criterion is meant to roughly search for better technical quality regardless of the price. This filtering policy is identified in the document as **ASPL**.

LP+ASPL ($\sim$Multipath BGP): selects routes that have the highest local preference and the shortest AS path length. This path selection allows us to select shortest paths while minimizing the transit cost. This filtering policy is identified in the document as **LP+ASPL**. For the purpose of comparison, the decision process performed by Multipath BGP [1] can be approximated with this decision process. The MED filtering is missing and leads to a maximization of the Multipath BGP diversity. Nevertheless we think that the lack of MED values has a small impact as these values are not always taken into account by ISPs.

BGP : selects routes that have the highest local preference, the shortest AS path length and the lowest router ID. This path selection emulates the BGP decision process. We use it for the sake of completeness only. This filtering policy is identified in the document as **BGP**.

The remainder of this section evaluates LMD and the five new selection process presented above.

2.4 Evaluation Methodology

In this section we propose a methodology to evaluate our solution to increase the usable Internet routing diversity. Section 2.5 details the results obtained using

1. Multipath BGP takes routes that have the same highest Local Preference then the same shortest AS path length and the same best MED.

this methodology. Despite the diversity is finally profitable for stub networks, we focus our analysis on the diversity that a transit network can propagate to its stub clients (as described in Figure 2.2) in order to establish the relationship between diversity and churn. Even though we only explain it later, Table 2.1 shows that our approach is general as it highlights that ASes potentially have a large number of stub ASes connected to them.

Thanks to our methodology, it is possible to complete a valid evaluation even though it is hardly possible for researchers to obtain perfectly accurate BGP information. That explains why we have to reconstruct BGP information from partial knowledge. To do so, we assume that the two most common Internet routing policies (i.e., Valley-Free and Prefer Customer [31]) are respected by transit ASes. Despite one can measure some current routing advertisements that do not respect these policies [87], we consider that these advertisements are incorrect but still propagated in the Internet because of router misconfigurations [88, 89].

2.4.1 Getting eBGP routes

We used one week of eBGP data provided by Route Views [27] (from march 7th to march 13th, 2011) to simulate the BGP routes received by several ASes (each AS is an independent evaluation). We choose to concentrate on the ASes for which we were able to reconstruct the maximum diversity from Route Views data as explained below. In order to asses the results we obtain on this one week evaluation, we performed the same evaluation, on the same ASes, during another week (i.e., from may 1st to may 7th, 2011). The results are analysed for the Week 1 (i.e., march 2011) and aggregated results of the Week 2 (i.e., may 2011) are provided all along, for completeness.

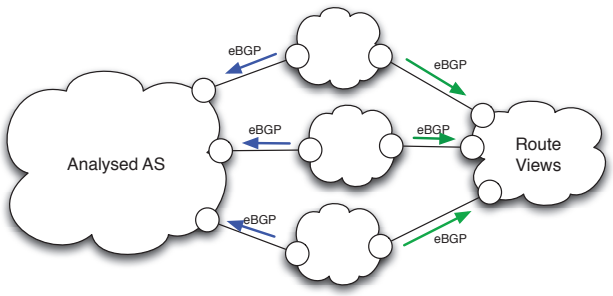


Figure 2.4: Route Views connectivity

Route Views project collects routing information from several ISPs in location spread around the world [27] with some Route Views collecting routers connected with eBGP multi-hop to several neighboring ASes. For each prefix, every BGP event that causes the best route to be changed at the provider routers results in a BGP update message sent to the Route Views routers which BGP RIB-INs are regularly archived and made publicly available. The provided RIB-INs contain all the received routes (i.e., no BGP best path selection process). In fine, we get a

routing table with full path diversity.

We assume that Route Views' neighbors advertise their whole BGP routing table to Route Views routers. Hence, every route that the neighbor sent to one of its neighbors (blue arrows in Figure 2.4) is also sent to Route Views (green arrows).

As a matter of fact, Route Views eBGP data have some limitations (see Section 2.4.2). First it does not provide local preference values so we must infer them using, for instance, the CAIDA relationship Database [29] (see Section 2.4.2). Second domains connected to Route Views send all updates with no filtering whereas they would normally send only a subset to another provider (according to the Valley Free conditions [31]). We must then filter the BGP updates to only accept the ones that are compatible with the Valley Free conditions (see Section 2.4.2). Mühlbauer et al. [90] underline the wide utilization of the valley-free filtering.

2.4.2 Limitations of Route Views data

As Route Views is only used for route collection and thus has no data plane, the BGP view it provides is biased. For example, Route Views only provides a few MED values and no IGP costs which means that it is impossible to determine the MED as the topologies are unknown from our dataset. In this section, we describe the techniques we applied on the Route Views data to minimize the impact of the bias.

Local preference estimation

As the Route Views infrastructure only uses eBGP, there is no local preference propagated through the updates. Local preferences are inserted by the receiving AS and are generally used to promote the exit of an AS according to the commercial agreement with the connected neighbor. At a peering link, for instance, where bandwidth is free, there will be a higher local preference than at transit link (where consumed bandwidth use must be paid). In the Route Views infrastructure, there is no reason to insert a local preference at BGP update reception because there is no commercial relationship between the domains and Route Views routers and they do not provide packet forwarding. Consequently we had to insert local preferences for every route, according to the AS we evaluate the diversity.

By knowing the relationships between the analyzed AS number and the ASes connected to Route Views, it becomes possible to insert realistic classes of local preferences (i.e., one class for customers, one for providers, and one for peers). Commercial relationships between ASes are partially known and provided by CAIDA [29]. We computed the relationship knowledge for all the ASes listed in the CAIDA relationship list and extracted the ones having the most well known relationship with ASes connected to Route Views. We focused on the AS numbers 3356, 2914, 13030 and 4436.

In the table 2.1 are listed the figures associated to the chosen ASes. For instance, among all domains connected to AS 13030, 49 are connected to at least one of the Route Views routers. Among these 49 domains, 2 are providers,

which represent 66% of 13030's providers (according to the CAIDA relationship database), 24 are peers (22%) and 23 are clients (2.4%).

ASn	3356	2914	13030	4436
Name Name	Level 3 Communications	NTT America	Init7 Global Backbone	nLayer Communications
AS rank (from CAIDA)	1	6	27	34
Available providers	0 (100%)	7 (78%)	2 (66%)	4 (50%)
Available peers	14 (73%)	14 (17%)	24 (22%)	33 (77%)
Available clients	43(1.7%)	11 (2.0%)	23(2.4%)	1 (1.5%)
Stub clients	1 656	259	383	28

Table 2.1: Information on the analyzed ASes

It is interesting to note that, in general, the number of stub clients, to which the NSP's diversity could be propagated, is an important part of the clients of the Network Service Providers. For instance, the 1 656 stub clients of Level 3 represent 64% of the total number of clients.

No Valley Free compatibility for the received BGP updates

Each connection between two ASes can be classified depending on the type of relationship they commercially negotiated. Without loss of generality, we only consider the two most prominent relationship types: peering and transit. A transit relationship means that one of the two ASes, the provider, provides wider IP reachability than the one the other AS, the client, proposes to it. The client pays for the amount of data sent to its provider. A peering relationship means that both providers exchange the same amount of traffic. An AS will only advertise its client subnets and its own ones to its providers and its peers, in order to limit as much as possible the bandwidth use with its peers and providers, but will advertise all its routes to its clients. These rules are known as the Valley Free (VF) [31].

As there is no commercial relationship between the Route Views routers and the Ases connected to them, the ISPs advertise all the best routes they learn to Route Views. In the case of Route Views taking the place of a real ISP, the amount of received BGP update would differ depending on the commercial relationship with the neighbor with respect to the Valley Free conditions. These conditions are not always respected but seem to be a reasonable assumption. Therefore the received updates need to be filtered. The Valley Free conditions allow an AS to:

1. send all the routes in its routing table to a client,
2. send only routes of local and client prefixes to a peer,
3. send only routes of local and client prefixes to a provider.

Non Valley Free updates detection Each update received by a Route Views router contains an AS path `AS1-AS2-AS3...` We denote as AS0 the AS we want to evaluate the diversity and the stability. First we only accept updates coming

from ASes (e.g., AS1) that have a well known relationship with AS0 (according to CAIDA [29]). Second we filter the accepted updates to be compliant with the Valley Free conditions. To achieve that, we must know the commercial relationship between AS1 and AS2. For instance, if AS1 is a client of AS0, it would not send it updates coming from AS2 if AS2 is a peer of AS1.

BGP update erasing A VF compatible route can be replaced by a non VF compatible routing update coming from the same AS. By just filtering the update, we would occult an implicit withdrawal of the VF compatible route. For that reason, all non VF compatible routes are replaced by withdrawals that are computed like the other withdrawals. This technique overestimates the number of withdrawals and hence gives an upper bound on the number of withdrawals in our evaluation.

Withdrawal process The filtering of BGP withdrawals can not be performed with the previously described process as BGP withdrawals do not contain the AS path. Therefore, a withdrawal is taken into account only if a route for the same prefix associated with the same next hop remains into the RIB.

2.4.3 Evaluation process

The evaluation process of a single AS and a selected selection policy can be divided in two processes.

First we construct the routing table the AS should receive by applying the modified selection process. We take as input the whole routing information that has been cumulated by Route Views at the beginning of the evaluation period. Each routing entry is computed in order to insert the local preference inferred according to the inter-domain commercial relationship database provided by CAIDA[29] (cf. Section 2.4.2). Then each entry is filtered in order to only accept the ones that are compliant with the Valley Free conditions (cf. Section 2.4.2). Last the evaluation process applies to the remaining entries one of the route selection policies (i.e., one of the selection processes detailed in Section 2.3.3). The output is considered as the FIB (i.e., in our case the LISP mapping locator list) of the analysed AS, which would be used to forward the packets if the architecture were used by the said AS.

Then we compute the instability of the FIB. We compute each single BGP update Route Views provides, during the evaluation period, to evaluate the instability of the FIB . It is important to remember that instability is commonly considered as a major issue of the Internet. Routers can be impacted by the intensive computation needed to treat the BGP updates which might result in wide-scale instability, prefix reachability issues, data forwarding loops, and packet losses [91, 92]. Consequently, the amount of diversity that can be announced is a tradeoff between diversity needs and instability (i.e., churn) as with a new route comes the related instability. To evaluate the evolution of churn and path diversity with the local use of LISP and redistribution of the BGP routes into the LMD, we define the path diversity as the number of routes that can be used for a given prefix and the instability as the number of route changes, among the LMD entries, that have a direct impact

on the data-plane. Whereas some instability consequences can be handled by the computation capabilities of routers, this instability measurement is independent from routers capabilities.

In our case, each BGP advertisement we take into account is firstly computed to insert the inferred local preference and filtered in order to only accept the ones that are compiant with the Valley Free conditions. Then we compute the impact of each routing update on the forwarding table in order to update it and conclude, for each update, if it impacts the data plane. In our evaluation, routing updates are considered as instabilities if they have one of the following impacts:

- a mapping locator is erased. A mapping locator may be erased due to policy (e.g., stability, metric change) or BGP withdrawal. The associated traffic is then redirected through another domain with potential packet loss and desequencing.
- a new mapping locator arises because of local policy (e.g., stability or BGP metric change) or a new BGP route. Load balancing implementations are usually using computation based on destination IP address of packets and the number of available paths. The result of a new mapping locator is that the encapsulating routers will re-order the flows among the different exit points leading to potential packet loss and desequencing.

2.5 Results

We present in this section the results of our evaluation. We have to highlight that the underlined diversity and instability are the ones received by transit networks and potentially propagated to their stub clients. We evaluate the gain in term of diversity for the different route selection process policies proposed in Section 2.3.3 and the cost in term of churn (i.e., stability) on the control plane. We first show the path diversity a NSP may enjoy in Section 2.5.1. Then we analyse the factors that impact the stability in Section 2.5.2.

2.5.1 Examples of allowed path diversity

Our evaluation focuses on four ASes: 3356, 2914, 13030, and 4436. While results for AS 3356 and AS 13030 are presented in details, we only provide, for conciseness, summarized results for ASes 2914 and 4436.

We can see in the Figure 2.5 and Figure 2.7 that both ASes (i.e., 3356 and 13030) can benefit from a high route diversity. The amount of available paths is greatly correlated with the path selection policy [90].

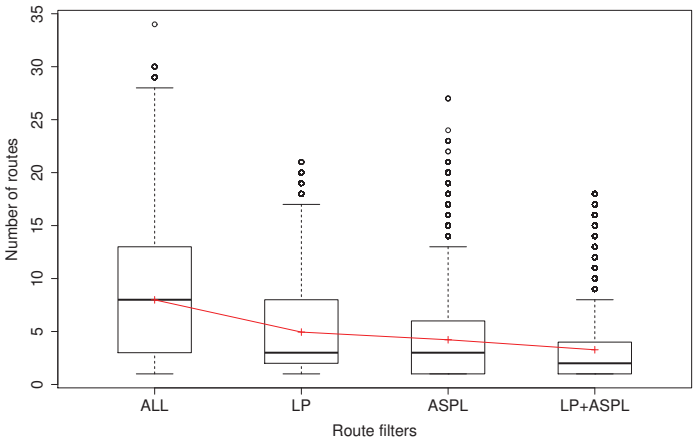


Figure 2.5: 3356: path diversity

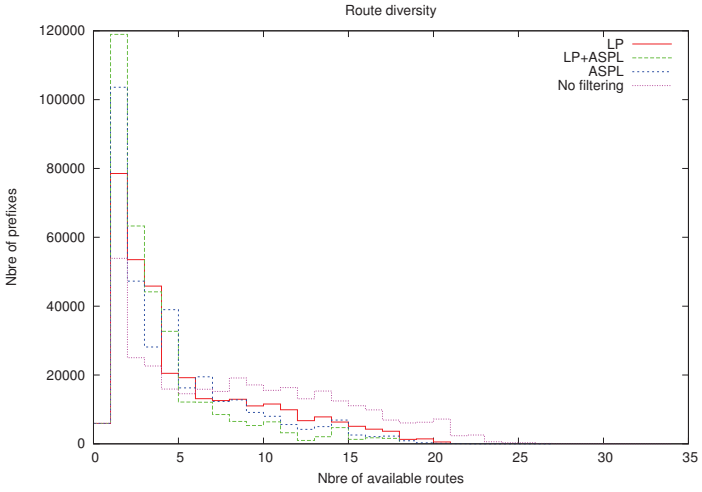


Figure 2.6: 3356: diversity distribution

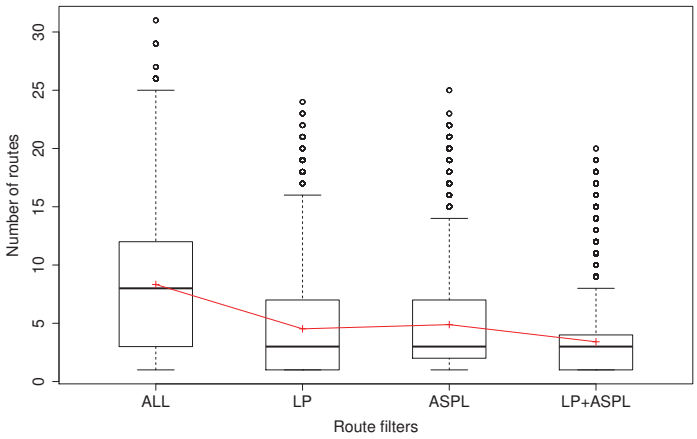


Figure 2.7: 13030: path diversity

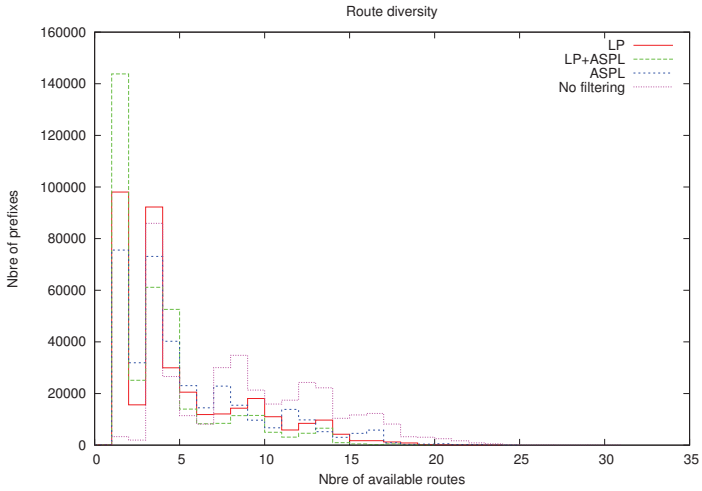


Figure 2.8: 13030: diversity distribution

In the LP selection case, as almost all the prefixes are known and advertised at least by one client it leads AS 3356 to only accept diversity coming from clients (i.e., the highest local preference). AS 3356 is actually known to have a *customer cone*[2] higher than 90% and therefore receives updates from clients for more than 90% of the prefixes. This filtering can make prefixes with a lot of potential diversity decrease to very little diversity because one or two clients propagate it. In our

2. Customer cone: for an AS, the customer cone is the ratio between the prefixes belonging to its recursive clients and the number of overall prefixes in the world (see http://as-rank.caida.org/?mode0=as-intro#customer-cone).

evaluation, a high proportion of the prefixes (about 200,000 prefixes) decreases to a diversity less than 5 routes.

On the contrary, an overall diversity of 1 or 2 routes (see ALL) is seldom at AS 13030 as it is connected to 3 provider links and hence, a large proportion of prefixes have three routes. These are prefixes which reachability is only provided by the provider links that advertise nearly all prefixes of the Internet. Other prefixes are reachable through more than 3 paths. LP filtering has no effect on that peak as the reachability is assured by providers only. As for AS 3356, prefixes advertised by at least one client lose their diversity with the LP selection. As opposed to the ALL diversity graph, a high number of prefixes (almost 100,000) are reachable by only one route.

Results using the ASPL filtering are very different. For the case of AS 3356, ASPL filtering truncates diversity more than LP filtering. The median values of LP and ASPL filtering are similar but much more prefixes have a diversity of 1 for ASPL filtering. This makes the mean value of the diversity lower for ASPL than for LP. On the contrary, ASPL filtering truncates less the diversity than LP filtering at AS 13030.

ASPL and LP filtering are not totally correlated. Given a prefix that is advertised both by clients and peers, the updates announced by clients propose the shortest ASPL with a probability of 62% for AS 3356 and 16% for AS 13030. Therefore LP+ASPL filtering combines the effects of the two filtering techniques and proposes a low level of diversity. First it takes into account routes coming from the best LP (i.e., the clients for most of the prefixes) and then takes only 62% for AS 3356 and 16% for AS 13030 of these diversities due to path length.

Therefore Multipath BGP, which can be approximated by the LP+ASPL decision process, truncates a lot the received diversity for both AS 3356 and AS 13030. Furthermore, Multipath BGP can not be relaxed and there is not way to use non equal routes for prefixes which have very low diversity.

We report prefix diversity mean and quartile values for the two weeks of the evaluation for the four ASes in Tables 2.2, 2.3, 2.4 and 2.5. It is worth noting that diversity levels are not significantly different between the two snapshots separated by a period of two months, reinforcing so the generality of our observations.

ASn	3356						13030					
	min	25%	50%	75%	max	mean	min	25%	50%	75%	max	mean
ALL	1	3	8	13	34	8.12	1	3	8	12	31	8.33
LP	1	2	3	8	21	5.03	1	1	3	7	24	4.52
ASPL	1	1	3	6	27	4.29	1	2	3	7	25	4.88
LP+ASPL	1	1	2	4	18	3.33	1	1	3	4	20	3.41

Table 2.2: ASes 3356 and 13030 - Week 1 diversity results

ASn	3356						13030					
	min	25%	50%	75%	max	mean	min	25%	50%	75%	max	mean
ALL	1	3	8	13	31	8.38	1	3	8	12	31	8.61
LP	1	2	3	8	22	5.29	1	1	3	7	25	4.90
ASPL	1	1	3	6	25	4.43	1	2	3	7	26	5.01
LP+ASPL	1	1	2	4	19	3.51	1	1	3	5	22	3.65

Table 2.3: ASes 3356 and 13030 - Week 2 diversity results

ASn	2914						4436					
	min	25%	50%	75%	max	mean	min	25%	50%	75%	max	mean
ALL	1	16	18	20	42	18.13	1	11	12	15	29	13.05
LP	1	2	3	6	15	5.16	1	4	7	11	18	7.00
ASPL	1	2	6	12	29	7.13	1	4	7	9	27	7.00
LP+ASPL	1	1	2	3	15	2.84	1	2	4	7	16	4.54

Table 2.4: ASes 2914 and 4436 - Week 1 diversity results

ASn	2914						4436					
	min	25%	50%	75%	max	mean	min	25%	50%	75%	max	mean
ALL	1	16	17	20	40	18.05	1	11	12	16	31	13.29
LP	1	2	3	6	15	5.24	1	3	7	11	20	7.07
ASPL	1	2	6	11	29	7.10	1	4	7	9	29	7.10
LP+ASPL	1	1	2	3	15	2.89	1	2	5	7	18	4.66

Table 2.5: ASes 2914 and 4436 - Week 2 diversity results

2.5.2 Churn: the cost of path diversity

Flap Dampening

The evaluation of the next section (i.e., Section 2.5.2) compares the evolution of the churn with the augmentation of the routing diversity offered by the policies described in Section 2.3.3. As underlined in Section 2.4, we define the churn as the routing updates that have a direct impact on the data plane.

Before estimating the cost of path diversity in term of churn, we first have to determine if route flap dampening [93] can still be applied in a diversity context. Route flap dampening is a mechanism to reduce signaling messages in BGP. To do so, a BGP router computes a penalty value for every prefix in its routing table. The penalty increases with the rate of updates received for the prefix and the update is not propagated if the penalty exceeds a threshold, until the penalty comes back to a more acceptable value. Flap dampening reduces churn, at the cost of a longer convergence time. Without the use of flap dampening, a single route may alternatively be propagated and withdrawn with a high frequency (i.e., several times per minutes). Such a flapping may impact either the BGP route of a route selected according to another selection policy and may therefore bias the diversity cost evaluation of the next section.

By performing flap dampening, **we aim at avoiding that a small amount of flapping routes tilt the evaluation either in favour of BGP or in favour of a relaxed decision policy.**

In our case, it is therefore interesting to study whether a flap dampening pre-processing of routes impacts the diversity. If route flap dampening does not alter the route diversity, it can be safely applied to reduce the global churn on a system that aims at increase the usable diversity.

We simulated flap dampening with the default parameters recommended by Cisco [3] on our dataset and observe that route flap dampening does not significantly impact the diversity but greatly reduces the global churn. Figure 2.9 shows the impact of flap dampening on the diversity and Figure 2.10 shows the impact of flap dampening on the churn, for AS 13030. Both figures aggregate the information with boxplots representing the 5th, 25th, 50th, 75th, and 95th percentiles as well as the mean and outliers. The results obtained with route flap dampening are represented with the labels prefixed with `FD-`. When Figure 2.10 shows no visible difference on the diversity, Figure 2.10 shows that all the churn is significantly reduced with route flap dampening where the mean and the third quarter values notably decrease. More precisely, flap dampening pre-processing decreases the routing diversity by between 2.3% and 3.0% depending on the policy whereas it makes the churn drop by between 26% and 43%. It is interesting to note that the flap dampening process reduces the churn regardless of the selection processes (LP, BGP...).

We observe the same benefit for every AS we analyze (cf. Table 2.6). Therefore we apply the flap dampening pre-filtering for the rest of the evaluation.

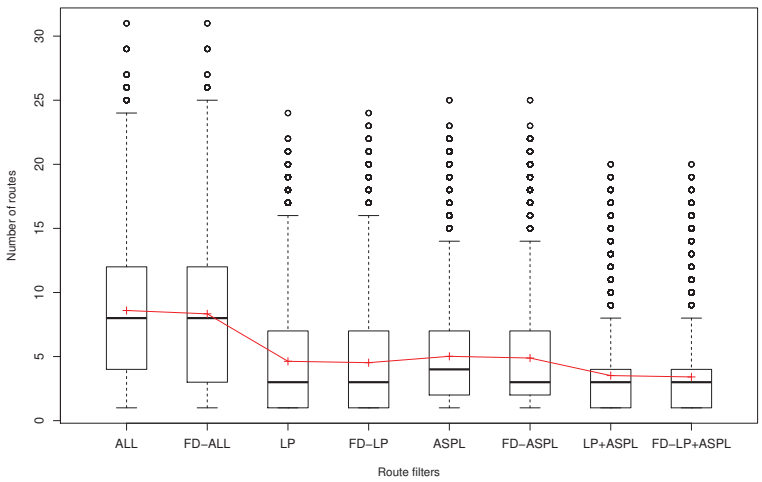


Figure 2.9: Path diversity comparison with and without flap dampening

3. see `http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a00800c95bb.shtml`

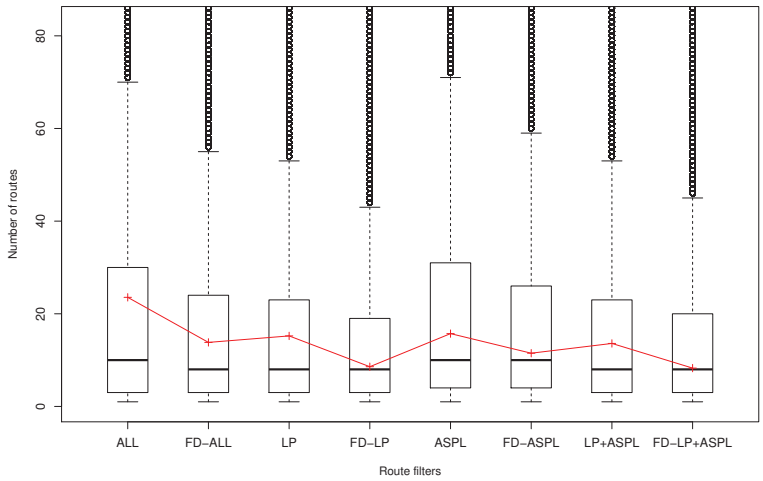


Figure 2.10: Churn comparison with and without flap dampening

	Decrease of diversity				Decrease of churn amount			
ASn	3356	13030	2914	4436	3356	13030	2914	4436
ALL	3.5%	3%	0.7%	0.73%	33.5%	41%	19%	18%
LP	3.0%	2.3%	0.4%	0.22%	34%	44%	23%	21%
ASPL	3.6%	2.7%	0.7%	0.62%	29%	27%	5.2%	6.4%
LP+ASPL	3.1%	3%	1.0%	0.17%	29%	39%	12%	8.7%
BGP	1.8%	0.05%	0%	0%	24%	35%	20%	15%

Table 2.6: Flap dampening: Churn and diversity decrease

Tradeoff between path diversity and stability

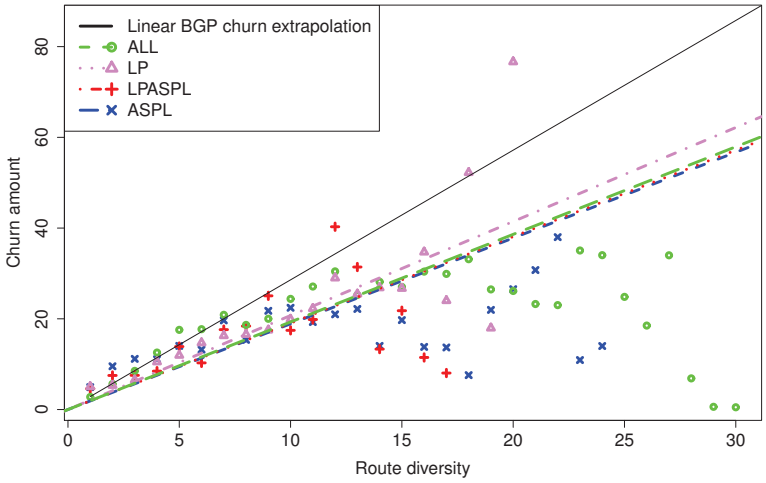


Figure 2.11: 3356: Instability Vs Diversity

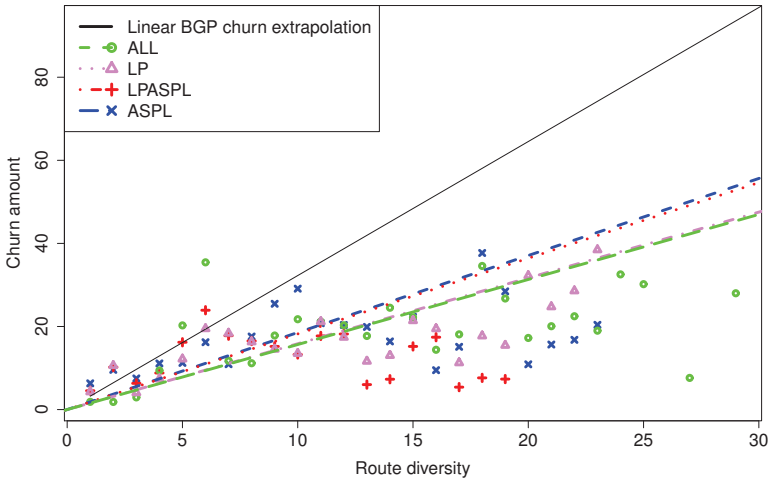


Figure 2.12: 13030: Instability Vs Diversity

High diversity is a consequence of a relaxation of the route selection process. However, this relaxation may have two contradictory consequences. On the one hand, it decreases the impact of some updates as the change of BGP metrics may not impact the result of the relaxed selection. For example, a change in the AS path that changes the BGP selection would not change the `LP` selection. On the other

hand, a routing update that may not change the BGP best route could change the selection of the path diversity. For instance a change in the AS path length would not change the BGP best route if this route comes from the only neighbor with the highest local preference whereas it could change the routes selected by ASPL filtering. From this contradiction comes the necessity to analyze the churn (i.e., the instability) obtained with every policy and its relation with the usable diversity. Figure 2.11 and Figure 2.12 present the relation between instability and diversity for AS 3356 and AS 13030. A linear extrapolation of the BGP churn is used for comparison in both graphs. They contain the graphs of churn according to the diversity for each path selection. Linear regressions have also been calculated for each case for the sake of readability and comparison (with the linear extrapolation of the churn of BGP).

On the one hand, the different cases of AS 3356 are close to that of the BGP extrapolation, but with smallest amount of churn. Some flapping prefixes (less than 0.03%) introduced a high churn and flap dampening reduced their instability. As these flaps affect prefixes with high diversity, AS 3356 keeps on getting a lot from the adoption of a flap damping algorithm.

On the other hand, AS 13030 shows lower instability than the BGP extrapolation even if the instability is higher for some low diversity prefixes. The reason why AS 13030 is more stable than AS 3356 is that it is connected to several providers. Indeed, providers rarely send withdrawals to their clients as they almost always have reachability for all the prefixes.

Tables 2.7 and 2.8 report the mean number of churn per prefix for the two distinct weeks of the evaluation and highlights that the same occurs for the ASes 2914 and 4436. We can also see that the BGP decision process usually selects routes with higher churn than other selection processes. For each selection process, the tables also provide, in brackets, the percentage of churn decrease, compared to the one provided by BGP.

ASn	3356	2914	13030	4436
ALL	1.93 (32.3%)	0.96 (73.5%)	1.56 (51.6%)	1.00 (71.0%)
LP	2.07 (27.4%)	0.86 (76.2%)	1.58 (50.9%)	0.88 (74.5%)
ASPL	1.89 (33.7%)	2.09 (42.2%)	1.85 (42.5%)	2.41 (30.1%)
LP+ASPL	1.90 (33.3%)	1.88 (48.1%)	1.82 (43.5%)	2.79 (19.1%)
BGP	2.85	3.62	3.22	3.45

Table 2.7: Week 1: Mean churn per path

ASn	3356	2914	13030	4436
ALL	1.42 (43.4%)	0.74 (75.7%)	1.12 (58.3%)	0.85 (73.5%)
LP	1.31 (47.8%)	0.89 (70.8%)	1.18 (56.1%)	0.88 (72.6%)
ASPL	1.39 (44.6%)	1.49 (51.1%)	1.39 (48.3%)	1.56 (51.4%)
LP+ASPL	1.33 (47.0%)	1.85 (39.3%)	1.43 (46.8%)	1.64 (48.9%)
BGP	2.51	3.05	2.69	3.21

Table 2.8: Week 2: Mean churn per path

2.6 Discussion

Section 2.5 estimates the diversity gain that can be achieved with our architecture and its related cost in term of churn. The dataset we use leads to some approximations that we discuss in this section in order to assess the accuracy of the evaluation and draw general conclusions.

2.6.1 Impact of architectural choices

As underlined in Section 2.2, the route propagation between the LMD and the encapsulating router can either be pulled on-demand (i.e., the encapsulating router requests a route when needed) or directly pushed by the LMD (like BGP, which propagates routes without demand)

These two approaches have different advantages and drawbacks we detail below. In consequence, the choice between these two paradigms is let to the network services provider and its clients, according to their requirements.

On-demand mapping insertion

When mappings are obtained on-demand, whenever an encapsulating router receives a packet for which no mapping matches in its LISP cache, the router requests the LMD the mapping corresponding to the packet. Such a process has three consequences.

- First, as shown in [94] it allows to reduce the memory needed on the encapsulating routers as only the used paths are put into there LISP caches. The reason behind this memory gains is that routers, on any given period of time, only contact a small fraction of the IP space.
- Second, it induces a latency when the first packet, which is to be put onto a specific path, arrives. Indeed this specific path is not readily available and some mapping resolution time is necessary. However, during that period there is no packet loss as packets can be forwarded onto the legacy BGP path.
- Third, as routes change with time, mappings pulled by the encapsulating router cache may be outdated. To avoid using outdated mappings, the LMD can follow a state-full approach to remember the routers that are using the mapping and inform them of any change. An alternative to avoid maintaining state at the LMD is to use the mapping versioning feature of LISP [95] that allows decapsulating routers to verify whether the mapping used to encapsulate a packet was outdated and if this is the case, trigger the encapsulating router to update its mapping.

Therefore, at the price of a path establishment latency, the whole diversity and the associated churn are managed only by the LMD whereas encapsulating routers only manage the routes (and the associated churn) they really use. This approach suits well if the stub network uses rarely alternate path. Nevertheless, due to the latency, it is not suitable in the case where the stub network wants to quickly switch to an alternate path in after a failure of the primary path.

Proactive mapping insertion

When mappings are installed proactively, each time the LMD creates or updates a mapping because of a routing update it sends the mapping to the encapsulating routers encompassed by the LMD. Proactive insertion has four consequences.

- First, any new flows can readily use paths as all the usable paths are already inserted into the LISP cache. Therefore no extra latency is added by the system in order to use the alternate paths.
- Second, mappings in the LISP cache are always up-to-date as the system inherits from the BGP reactivity as BGP updates are treated and installed directly in the encapsulating router caches.
- Third, as all mappings are installed in advance, the LISP cache size can become important whenever a few alternate paths are utilized. This point is the main scalability issue of this approach as the cache size is directly proportional to the number of prefixes in the Internet, and the number of routes per prefix clients are willing to obtain.
- Fourth, the encapsulating routers receive and treat every single routing update whereas some updates impact routes that are not currently uses. This leads to an important computation requirement for end-site routers. The churn evaluation of the current section gives the amount of churn encapsulating routers must be able to manage in this case.

When mappings are installed proactively, there is no latency in the usage of the most appropriate paths. Nevertheless the whole path diversity and the associated churn have to be managed by each encapsulating routers. Therefore, this approach is adapted for clients with strong performance requirements or fast recovery needs.

2.6.2 Potential error due to the lack of BGP feeds

All existing ASes are not connected to the Route Views architecture. Therefore it were not possible to obtain BGP feeds from all the BGP neighbors of 3356 or 13030. We analyzed feeds coming from 50 neighbors among more than 2500 neighbors (19 peers and 2,600 clients or so) of 3356 and 31 neighbors among more than 1,000 neighbors (3 providers, 100 neighbors and roughly 900 peers) of the 13030 (figures given by CAIDA). So the route diversity and the instability are under-estimated. As the diversity has to be compared to the BGP one route selection, the real case (i.e., the real diversity) is better and our analysis gives a lower-bound of the reality. As the instability increase has to be compared to the diversity increase, we believe that it keeps on evolving with the same behavior (same or higher increase of the diversity than the instability: see Section 2.5.2) as we assume that the analyzed BGP feed set is a representative sample of the real connection amount of both 3356 and 13030. Route Views does not provide a complete vision of the Internet but the impact of this incompleteness is limited for the ASes 3356 and 13030 [96].

2.6.3 Potential error due to the lack of BGP relationship knowledge

We filtered BGP updates in order to only accept updates compatible with valley free conditions. Due to lack knowledge about inter ASes relationship, some updates have not been taken into account even if they were potentially valley free compliant. We are mainly interested by the ratio between diversity and instability. This lack of knowledge may lead to a deviation of this ratio compared to the one with the same data but with full AS relationship knowledge. This part proposes to evaluate the ratio of the diversity over the instability.

Table 2.9 presents the number of updates that have been analyzed. It is separated into 5 columns that are:

ALL : all the analyzed BGP updates,

VF : BGP updates that are compatible with the Valley Free conditions. Only these updates have been used to calculate the diversity and the instability,

non VF : BGP updates that are not compatible with the Valley Free conditions,

ND (not determined) : BGP updates for which we are not able (due to our lack of knowledge about AS relationships) to establish whether they are compatible or not with the Valley Free conditions,

Error : estimation of the number of Valley Free compliant updates among the ND updates. We can define this as the error of the data treatment. The error is linearly calculated by extrapolating on the ND updates the proportion of VF and the non VF updates already detected.
$ERROR = ND * VF/(VF + non\,VF)$

	ALL	VF	non VF	ND	Error
3356	114M	16.9M	89.9M	7.2M	1.14M(6.7%)
2914	116.2M	37.1M	62.6M	16.5M	6.14M(16.6%)
13030	120M	18.4M	81M	20.6M	3.15M(17.1%)
4436	84.9M	31.0M	46.8M	7.07M	2.81M(9.1%)

Table 2.9: Update numbers classed by category

Such an error leads to a deviation (in term of number of interesting routing updates) of 6.7% for 3356 and 17% for 13030, for instance. The consequence of this suppression is that diversity and instability have been under-estimated for both networks. The worst cases would be that the impact of the ignored updates is to give a high instability increase for a rather small diversity increase. Even in such a case, such an increase would not significantly change the conclusions of the evaluation in Section 2.5.2. In order to illustrate that point, the percentage error presented in the Table 2.9 can be compared with the decrease of churn highlight in Tables 2.7 and 2.8. Indeed, Tables 2.7 and 2.8 provide percentages of instability decrease (i.e., between 19.1% and 76.2%) that are more important than the percentage or error provided in Table 2.9 (i.e., between 6.7% and 17.1%) which leads to conclude

that the error due to the lack of relationship knowledge do not significantly modify the results.

2.6.4 Outcomes

Diversity and instability are present and are not homogeneous among the prefixes. We highlighted that some prefixes suffer from a lack of diversity (one or two routes) whereas other prefixes have a lot more routes than really necessary (more than 10 routes).

This shows that naive mechanisms are not adapted and that one would benefit from more flexible decision policies which is precisely what we aim to offer with our architecture. For instance, a restrictive selection may be used to filter the prefixes with high diversity (potentially taking into account new parameters such as route stability or long term performance measurements) whereas a relaxation of the policy may be of interest to increase the diversity for some prefixes with low diversity.

Some prefixes may bring a lot of instability and stability analysis may be used to select routes that are more stable. As we have observed, these unstable prefixes, for the most, benefit from a high diversity. Hence this stability filtering will not make them lose their whole diversity (unless the instability takes place in the destination network). Consequently, an advanced selection process is of interest to wisely select routes.

It is important to highlight that, contrary to BGP, the instability provided by our use of diversity is not propagated to the neighbors as the diversity is used locally in the stub part of the network and the instability impacts only the local choices as we only propagate the BGP best route to neighbors.

2.7 Conclusion

This chapter presents an architecture aimed at better exploiting the path diversity that is inherently available in the Internet. The proposal is based on the LISP Map-and-Encap mechanisms in order to overcome current BGP limitations. It allows an ISP to offer its path diversity (at least partially) to its customers for the sake of traffic engineering and robustness purposes. We describe the architecture and potential applications. The solution opens new perspectives in the definition of a wide range of possible route selection policies, based for instance on economical, performance, or stability criteria.

In order to assess the potential benefits of the proposal, we conducted an evaluation based on simple route selection policies (i.e., which routes are offered by the ISP to its customers). First the route diversity that diverse transit ASes may offer is studied, depending on the different policies. We also focused on the routing stability criterion by analysing the quantity of routing updates that cause changes in data forwarding. Our evaluation shows, at least for the studied networks, that the increase in diversity comes with a controllable and acceptable overhead. Furthermore, it was shown that taking into account the stability of routes is promising as

it does not significantly alter path diversity while being essential for the scalability and robustness of our proposal.

Chapter 3

IDRD: Enabling Inter-Domain Route Diversity

3.1 Introduction

In this Chapter, we aim at generalizing the approach of the previous chapter. Instead of propagating the path diversity only to stub clients, we propose a way to propagate it to transit neighbors and make them able to use it simultaneously.

As previously underlined, the Border Gateway Protocol (BGP) [8] fulfills the role of globally disseminating the routing information between individual IP networks belonging to different administrative domains (i.e., ASes).In this chapter we address three closely related problems attributable to BGP and its best practices.

Firstly, BGP was originally designed to select and propagate a single route (i.e., *path*) for each IP prefix. Such an approach automatically impedes the use of the AS path diversity which is inherently present in the Internet graph (cf. [97]).

Secondly, BGP does have some Traffic Engineering (TE) potential [98]. However, BGP route selection, which is based on the very strict *BGP decision process* [1], limits the traffic engineering capabilities in their scope and applications. Our objective is to relax BGP route choice policies in order to enable a richer set of route selection policies (for TE, value-added services, etc.)

Thirdly, the "prefer customer" condition (a.k.a. Gao & Rexford condition [31]), which assures the stability of the current single route inter-domain routing in the Internet, truncates the use of the potential routing diversity (as shown in our evaluation in Section 3.4). As this condition may prevent domains from proposing advanced services to clients, the relaxation of this rule becomes highly attractive.

We address the described problem along two complementary lines of thought. On the one hand, we propose *Inter-Domain Route Diversity* (IDRD), as an architecture which allows ISPs to propagate multiple routes per prefix towards their customers, providers and peers, while fully respecting the fundamental BGP export policies (a.k.a. Valley Free export policies) within the current inter-AS business

1. The BGP decision process corresponds to a list of successive, static rules based on comparisons of global and local path attributes (e.g., *LOCAL_PREF*, *AS Path* length, *MED - Multi-Exit Discriminator*, and next-hop IGP cost).

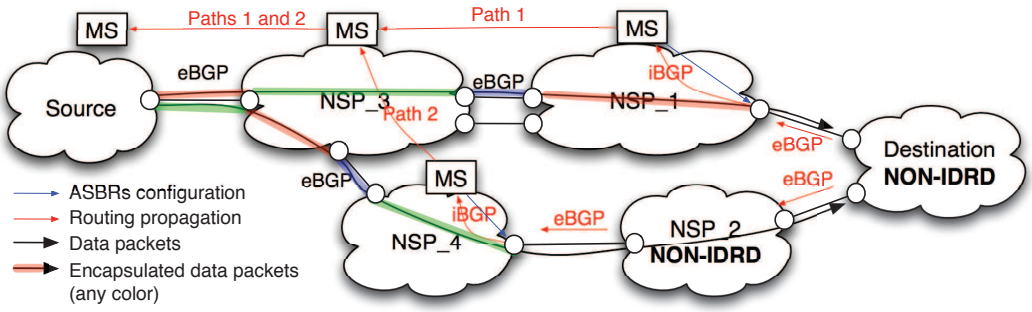


Figure 3.1: IDRD topology.

relationships. On the other hand, we recognize that the *prefer client routes* condition [31] is too restrictive with respect to the degree of diversity it allows for (cf. evaluation in Section 3.4). Therefore, we somewhat relax the selection process in the context of multi-path routing, and allow for the absence of the strict client route preference under certain conditions in order to simultaneously propagate and use both client and peer/provider routes.

The rest of this chapter is structured as follows. First, we present in Section 3.2 the architecture of IDRD starting from our main design objectives, after which we provide an overview of the IDRD data and control planes. In order to increase the potential usable routing diversity, in Section 3.3 we propose a relaxation of route selection criteria, including a modification of the “prefer customer” rule [31], accompanied with a comprehensive proof of stability for the relaxed criterion. In Section 3.4, we highlight the impact of the policy relaxation on the path disjointness potential in the current Internet graph based on real world data from CAIDA, after which we summarize this chapter with an outline of our main conclusions in Section 3.5.

3.2 Architecture for route diversity

In this section, we present Inter-Domain Route Diversity (IDRD), which enables the use of the inherent topological diversity present in today’s Internet. In order to enable the propagation (control plane) and the use (data plane) of diversity, we base our architecture on the *map-and-encap* paradigm. We rely on the Locator/Identifier Separation Protocol (LISP) as an example implementation, which is a map-and-encap protocol under ongoing development at the IETF [28]. Figure 3.1 provides an overview of the architecture. Each domain manages a Mapping System which is connected to the ones of its neighbors. In this way, a control plane in parallel to BGP is formed, which is used to propagate multiple paths between the ASes. In the data plane, packets are encapsulated at the entry router (AS Border

Router – ASBR) of each domain in order to enforce its path until the chosen exit ASBR. Encapsulation is also used to enforce the path between individual domains, whereby the required parameters are provided by the local Mapping System (cf. [99, 100] for examples of mapping systems associated with LISP). The Mapping System (MS) is local to a domain but can either be internal or external with respect to the encapsulating routers, and structurally it can either follow a centralized or a decentralized logic.

The detailed architecture we present is based on LISP, but we stress that each domain may choose an individual solution that suits its network best. E.g., an alternative architecture may be based on MPLS [33] for the encapsulation scheme, and on an extended version of the Path Computation Element [101] (PCE) for the management of path diversity. In the following, we present a LISP-based architecture, which may however easily be mapped to an MPLS based architecture.

Before presenting the IDRD architecture in detail, we first wish to provide clarity on the character of this solution, i.e., to stress that IDRD by no means aims at replacing BGP. Instead, we see IDRD as an add-on to the present Internet, which can be deployed by some domains in addition to BGP.

In order to illustrate our approach and to facilitate orientation with respect to the technical concepts introduced in the next subsections, in Figure 3.2 we provide an example overview of the global IDRD architecture. Interfaces and components that are directly associated with IDRD are named. For instance, the interfaces (A) is one that links two mapping systems and the component (1) is the entry ASBR of the domain (according to the forwarding plane).

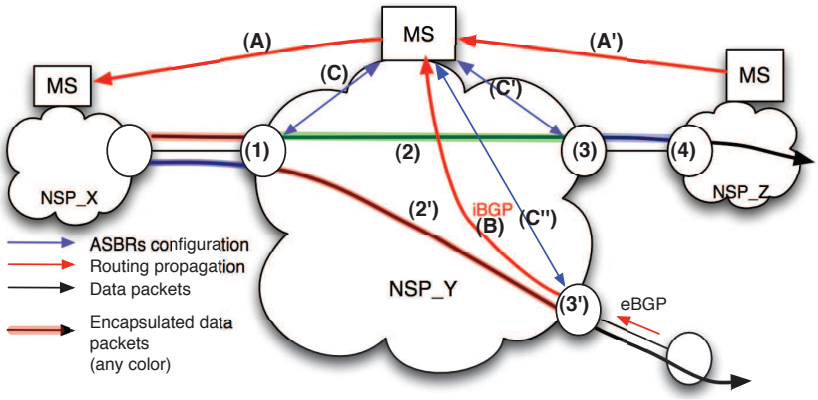


Figure 3.2: Zoom-in on one domain.

The Section is organized as follows. Sections 3.2.1, 3.2.2 and 3.2.3 describe the architecture by providing the protocols, which could be used for its adoption. Then Section 3.2.4 provides a functional representation of the architecture and develop the lifetimes of both packets and routing updates.

3.2.1 IDRD Data Plane

Due to the availability of several paths per destination prefix, the destination IP address no longer provides sufficient information for making forwarding decisions. To this end, in addition to the destination IP prefix, an identifier is used for specifying the path which is to be used, which we call the *path identifier* (i.e., path-ID). An in-depth discussion about the *path identifier*, its scalability as well as its Internet-wide organization is available in Chapter 4. This chapter also addresses the issue of FIB scalability, due to its close dependence upon the choice of path identification scheme. In order to enable the usage of IDRD's alternative paths in the present Internet, we propose to apply packet encapsulation in a similar scheme as in [75]. LISP provides both the encapsulation scheme and an *instance-ID* [28] which can be respectively used as path enforcement encapsulation and path-ID. Moreover, LISP is an IETF RFC and it has already been implemented in some commercial routers [2].

We choose to make a separation between intra-AS and inter-AS packet treatment according to the Internet's current structure as this allows for individual per-NSP choices of intra-domain technologies, i.e., either in favor of native IP-based network provisioning or in favor of MPLS-based schemes. Therefore, in order to put into use the end-to-end paths, IDRD must address their enforcements at two different levels of hierarchy.

Intra-domain path enforcement: A packet that arrives at a domain entrance (at Point (1) in Figure 3.2) triggers an ASBR request to the mapping system [3] about which egress ASBR the packet must be forwarded to. Subsequently, the Mapping System (MS) specifies the egress ASBR according to the packet's *path identifier* (i.e., LISP instance-ID), after which the packet gets encapsulated by the ingress ASBR and forwarded (Point (2)) to the correct exit ASBR. Finally, the packet gets de-capsulated once it arrives at the exit ASBR (Point (3)).

Inter-domain path enforcement: When arriving at the egress ASBR (Point (3) in Figure 3.2), the de-capsulated packet again gets encapsulated in order to enforce its path towards the next domain's ASBR (Point (4)). The next hop (destination address of the LISP extra header) is thereby chosen by the MS based on the IP destination address and the *path identifier*, which is contained in the intra-AS encapsulation header. Both the encapsulating IP destination address and path identifier are changed at the exit ASBR according to the LISP re-encapsulation process provided in [102].

Whenever our architecture is put into use by a NSP, there is a boundary between the non-encapsulated IP network of a stub domain (or an NSP that did not adopt IDRD) and the LISP network of its IDRD neighbor (e.g., between the source AS and NSP_3 in Figure 3.1). At that boundary, the encapsulating ASBR receives the IP packets and must LISP-encapsulate them and insert both the correct IP address of the next-hop ASBR and the Instance-ID (i.e., the Path-ID) associated with the path which is to be followed. This IP-packet to LISP-encapsulated-packet

2. Cf. http://lisp.cisco.com/

3. The mapping information could be either pushed or pulled. Cf. Section 3.2.2 for more details about it.

transition is performed thanks to the LISP Canonical Address Format [103], which allows to associate IP fields with the encapsulation values (Instance-ID, etc.).

Figure 3.3: LISP encapsulation headers (cf. [28]).

Figure 3.3 provides a representation of the LISP encapsulation header [28]. There are two headers. The first one (i.e., noted IH for intra-header) is the original one, for the end-host to end-host communication. The second is the LISP extra header which is composed by an IP outer header (noted OH), a UDP header and a LISP specific one. Four IP addresses are used. First the two original addresses - i.e., the end hosts sender and receiver - are stored in the intra header and are never modified. Then the extra header contains the addresses of both the encapsulating element (as source address) and the de-capsulating element (as destination address). Beside these addresses, the Instance-ID, which is to be used as a Path-ID, is present in a LISP specific header. It is to be noted that the LISP encapsulation is based on an IP-in-IP encapsulation. Therefore every single router that does not understand the LISP protocol (e.g., an intra AS router) is able to forward the packet till the de-capsulating router (i.e., an exit ASBR) by analysing the outer header as if it were a conventional IP packet.

Figure 3.3 provides an IPv4-in-IPv4 encapsulation representation. Nevertheless, the LISP protocol also supports IPv4-in-IPv6, IPv6-in-IPv6 and IPv6-in-IPv4 encapsulations. Please refer to [28] to obtain the details about other fields of the LISP encapsulation.

3.2.2 IDRD Control Plane

In the IDRD architecture, the information on path diversity is stored in the individual domains' Mapping Systems (MSs), and thus the propagation of path diversity between two IDRD domains is also performed at the MS level (Point (A) and Point (A') in Figure 3.2). In the most simple case, in which path diversity is *pushed* between the domains, the inter-MS protocol is very close to BGP as it is also based on the *path vector* principle, propagating the AS path in order to prevent loops.

As IDRD domains may be connected to neighbors that have not adopted IDRD, eBGP routing information coming from these neighbors can be redistributed into the MS (Point (B)). In other words, conventional BGP is a source of diversity. Moreover, IDRD does not replace BGP as the default inter-domain routing protocol, as IDRD domains keep on advertising BGP routes. This property allows IDRD to fulfill the "Backward Compatibility" and "Incremental Deployability" design criteria. Therefore, even early adopters that do not have any IDRD neighbors, benefit from the adoption of the architecture. Indeed, early adopters can benefit from the route diversity they receive from different eBGP neighbors and choose wisely among the different exit ASBRs depending on the path to be followed. Such early adoption and its benefits are described in Chapter 2 .

In the following parts, we focus on the architecture in the context of its inter-domain adoption.

Route propagation

BGP to MS route redistribution: This case is only relevant if one of the neighbors is non-IDRD compatible. In such a case, the adjacent IDRD domain receives eBGP messages at the inter-domain peering points. Pure iBGP can subsequently be used by the IDRD domain for sending these BGP messages from its ASBRs to the IDRD Mapping System (at Point (B) in Figure 3.2). In order to be technically able to send several paths to the Mapping System, ASBRs may use the *BGP add-path* extension [16]. In such a case, each route is identified in the BGP update thanks to an identifier. The couple (Next_Hop, ADD_path_Identifier) is sufficient for the Mapping System to uniquely identify each one of the routes it receives from all the ASBRs.

Inter-MS routing propagation: The propagation of multiple paths between IDRD-enabled neighbor ASes is performed at the MS-level, i.e., their MSes communicate directly (Points (A) and (A') in Figure 3.2). This diversity information contains at least the BGP metrics and may also contain additional information, e.g., price, capacity, etc. Overall, the choice of inter-MS protocol which is to be used depends directly on the type of relationship between the adjacent domains. The most simple type is certainly a push-based paradigm similar to the one already in use by BGP. In such a case, the use of BGP add-path [16] can readily be applied between the MSes. The information that needs to be propagated is composed of: (i) the advertised prefixes, (ii) a set of routes per prefix, along with the associated BGP metrics, and (iii) an entry ASBR per path, along with an entry path-ID. Other route propagation paradigms could also be considered, e.g., based on negotiation. In that case, a

negotiation framework, including proposal/acceptance/path preferences/etc. messages, should be made use of, as proposed in [30].

Mapping System entries

In its data base, the Mapping System must keep a minimum set of information in order to allow for the system's functioning. It must store the association between the sets of incoming flows and the paths they must be forwarded to. Colours are used to highlight the association between incoming flow information and outgoing flow information.

Domain incoming flows are defined by the following set of information, which is stored in the MS and partially configured in the **entry ASBR**:
- An end-destination IP prefix,
- An entry ASBR,
- An incoming path-ID,

Each *domain incoming flow* is further associated to its internal path defined by:
- A set of outgoing path-IDs,
- Each one associated with a next hop ASBR (i.e., an exit ASBR of the domain).

Each *domain incoming flow*, once forwarded toward an exit ASBR, becomes a domain outgoing flow.

Domain outgoing flows are defined by the following set of information, which is also stored in the MS but configured in the **exit ASBR**:
- An end-destination IP prefix (which must be the same as in the **incoming paths**),
- The IP address of the exit ASBR (which must be coherent with the next hop ASBR of the **incoming paths**),
- An incoming path-ID (which must be coherent with the outgoing path-IDs of the **incoming paths**).

The *domain outgoing flows* are further associated to the external/outgoing path defined by:
- A set of outgoing path-IDs,
- A next hop ASBR (i.e., an entry ASBR of the next AS) per outgoing path-ID.

It is important to note that the IP destination of a flow is not mandatory in order to forward it toward the correct path. Once a packet is encapsulated and associated with a Path-ID, the IP destination address may not be required anymore as all the forwarding information may be inserted into the Path-ID. Indeed, in Chapter 4 we analyse the scalability of Path-ID organisations and recommend to forward the packets according to the Path-IDs only.

Other information could be stored to perform advanced path selection as poten-

tial new routing paradigms may emerge. To this end, the following information may be useful; the price of the path or of the exit link, information on stability, bandwidth availability, delays, etc.

ASBR configuration

Once the diversity is propagated between ASes, the Autonomous System Border Routers (ASBRs) must be made aware of the corresponding mapping information. The information about **domain incoming flows** (cf. Section 3.2.2) is configured in an entry ASBR whereas the information about **domain outgoing flows** is configured in an exit ASBR.

The MS can either directly push the mapping information or await mapping requests from neighbor ASBRs (Points (C), (C') and (C") in Figure 3.2). While the push approach allows for the propagation of the messages to the ASBRs as soon as the MS receives them, at the same time it increases the size of the forwarding table with entries that may not be used at the data plane level. On the other hand, the pull approach introduces latency in forwarding the first packets on a path (i.e., during the mapping request), but it only inserts the required routing entries into the ASBR forwarding table. The choice between push and pull approaches in the configuration of routers is local to each AS and may therefore differ according to the policy of each domain.

Concerning the pull approach, [28] already proposes processes and messages for the LISP protocol that allow a router to request a mapping entry from the mapping server. This pull configuration protocol is already in use in some commercial routers. And as far the push approach is concerned, configurations can be performed via the *Netconf* protocol [104], which is also available on commercial routers.

We stress that not all the ASBRs of a domain must necessarily be able to deal with the use of routing diversity - instead, a domain may prefer to have only a small number of routers address this issue. In such a case, only a subset of ASBRs need to be compliant with the IDRD architecture.

3.2.3 An MPLS-based architecture

It is easy to adapt the previously described LISP-based architecture in order to obtain an MPLS equivalent, in which the intra-domain and/or inter-domain encapsulation are based on label switching. To this end, the whole data-plane description in Section 3.2.1 should be adapted: Indeed, the IP address of the next hop router in this case becomes an MPLS label and a second MPLS label is used as a Path-ID. Thereby, the mapping entries of the MS do not change except that the IP addresses of the next-hops and Path-IDs are assigned MPLS labels.

The inter-MS path propagation depends on the type of encapsulation that the domain has negotiated with its neighbors. Indeed, if inter-AS communications are encapsulated with MPLS, both the Path-ID and the entry ASBR are MPLS labels, whereas no changes are needed if the inter-AS communication is LISP-based.

An end-to-end path may consist of arbitrary combinations of LISP- and/or MPLS-domains, making label format translation mandatory in order to ensure end-to-end system coherency. This translation is performed at each ASBR connecting domains of different technologies. Thereby, either LISP to MPLS translation is performed by using the MPLS FEC-to-NHLFE table [105], or MPLS to LISP translation is carried out based on the LISP mapping [28, 103].

3.2.4 A functional representation of the architecture

Figure 3.4 gives a global block diagram overview of the IDRD control and data planes, agnostic regarding the chosen technology (i.e., MPLS or LISP). The IDRD control plane is coloured in **blue** whereas the IDRD data plane is coloured in **green**. In parallel, the conventional data plane and control plane are respectively coloured in **black** and **red**. We focus, in Figure 3.4, on the architecture of a single network service provider “A”, which is connected to some neighbors that have already adopted IDRD and to one neighbor that did not adopt it. Domain A owns ASBRs, that can forward IDRD traffics and conventional traffics, and a Mapping System, which is interfaced with the ASBRs.

The domain we stand for (AS A) is represented with two ASBRs (ASBR_1 and ASBR_2) and its mapping system. ASBR_4, which belongs to another AS, is not compatible with IDRD and can belong to either an IDRD domain [4] or a non-IDRD domain. In all the cases, it can only forward conventional IP packets and propagate conventional BGP updates.

Each ASBR contains both the control plane and the data plane. For space reasons, neighboring domains are not fully represented. As the neighboring ASes B and C are assumed to have adopted the IDRD architecture, their respective mapping systems are also represented. The routing information propagation between the MSs is represented (cf. “Inter-MS propagation" boxes) and underlines the minimum required information to be propagated in order to make the architecture work. For instance, the prefix X is propagated from C to A and from A to B. The routing information contains 2 routes, each one defined by the AS path and the next hop (i.e., the address of the next ASBR and the associated Path-ID). If selected by the mapping system, this information is configured in the different routers (cf. “Configuration” boxes).

Several routers may lie between two ASBRs. We do not represent these routers as they do not interact with the IDRD processes. Indeed, as the IDRD packets are encapsulated with conventional headers (i.e., IP or MPLS) these routers forward IDRD packets as if they were conventional packets.

4. An IDRD domain may have adopted the architecture in a sub-part of its network.

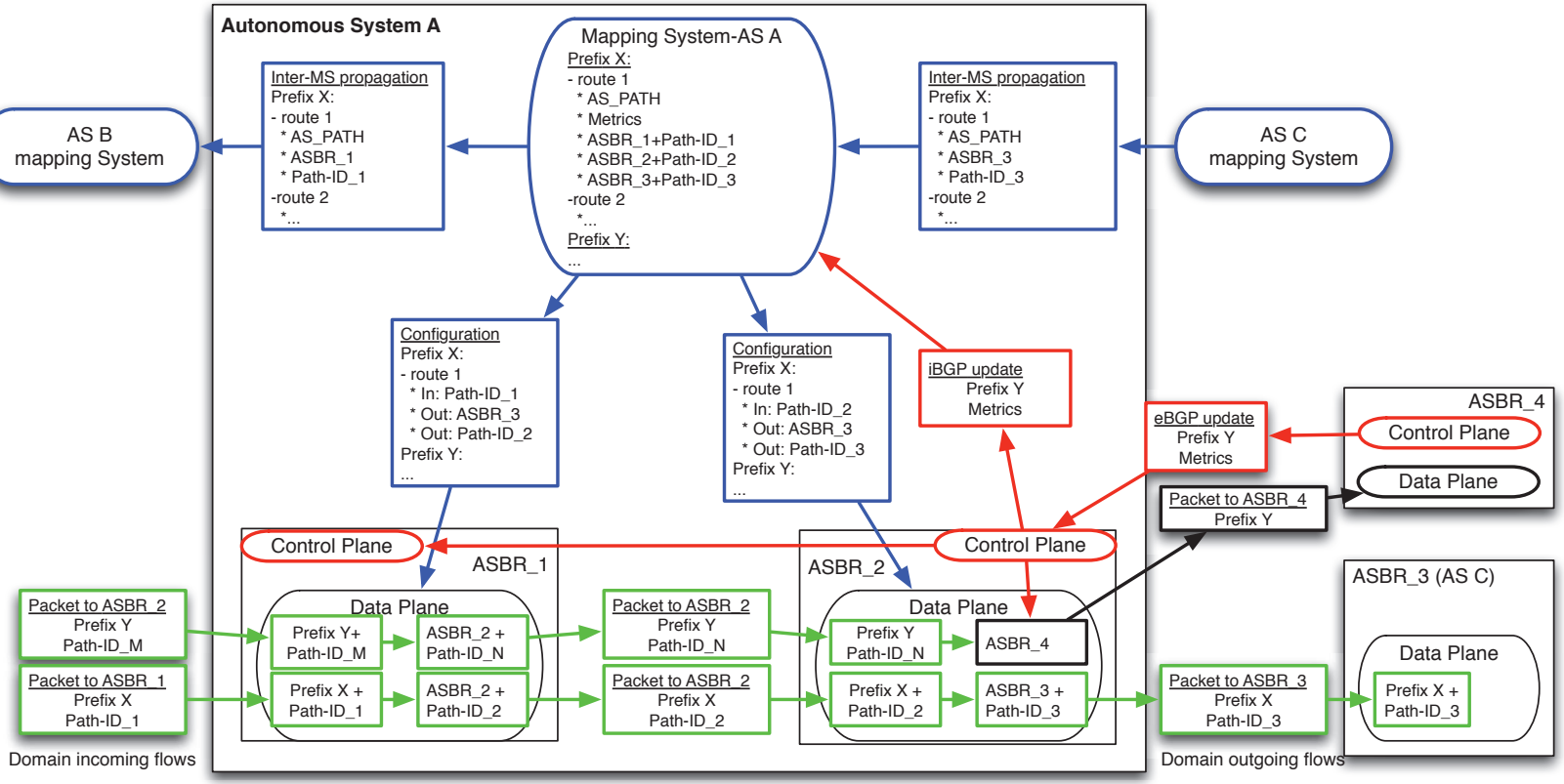


Figure 3.4: IDRD block diagram.

Life time of an IDRD update

We focus here on the IDRD control plane, which is in blue in Figure 3.4.

The mapping system of the domain A is connected to the mapping systems of IDRD neighbors (i.e., ASes B and C). Let us focus on an IDRD routing update coming from a neighbor. The MS of AS C sends a routing information to the MS of AS A. This update contains at least the AS path, in order to avoid loops, and the required information that will be used to forward the traffic. Once the MS of AS A receives the update:

- it filters updates that cause loops, as it is currently performed for BGP updates.
- then it filters again the update according to its routing selection process, which represents the policy of the domain. The Section 3.3, deals with the relaxation of this selection policy, in order to be able to perform more flexible path selections than the current BGP ones. It is interesting to note that we do not specify selection policies. We consider that the freedom to choose a policy is the keystone to leverage new internet paradigms. Therefore one can imagine simple policies, as studied in Chapters 2 and 4, or advanced policies, which need inter-domain negotiations, as the one studied in Chapter 5
- If the routing update is selected to be used, the mapping system forwards the routing update to the neighbors that are compliant with its export policy. Indeed IDRD allows for a selective export of routes and AS A may be willing to export a route only to a selected subset of neighbors. For instance, Valley Free is a selective export policy.

Once the routing update is propagated to his neighbors, AS A must configure his forwarding elements in order to make them correctly forward the traffic. Generally speaking, the mapping system needs to send to each involved forwarding elements the required information in order to:

- recognize the incoming traffic. Indeed a router will receive IDRD traffic and will be able to correctly process it if it has a matching entry with the carried prefix and Path-ID, at least. Without the knowledge of these information, the router will only be able to forward the packet toward the BGP route or the default route.
- forward the traffic to the correct IDRD path. Indeed once the traffic is well recognized, the router needs to know which IDRD forwarding element (i.e., its IP address or MPLS label) it must forward the traffic to and the new Path-ID value that is to be added to the packets.

If we focus on ASBR_1 of Figure 3.4, the mapping system sends to it the traffic recognition information (i.e., Prefix X and Path-ID_1) and the associated forwarding information (i.e., ASBR_3 and Path-ID_2). If ASBR_1 receives a packet carrying an IP address contained in the prefix X and the path-ID_1, it will decapsulate the packet, re-encapsulate it, add Path-ID_2 to the encapsulation and forward it to ASBR_3.

It is to be noted that the configuration of the ASBRs by the mapping system directly impacts the data plane, as it directly insert into the forwarding table the entries to be used.

Life time of an incoming BGP update

We focus here on the BGP control plane, which is in red in Figure 3.4.

Let us focus on ASBR_2. It receives from an external neighbor (i.e., ASBR_4) a conventional BGP update, containing a prefix and a set of metrics. First we have to highlight that this BGP route is propagated according to the current BGP practices. For instance if ASBR_2 selects the received routing update according to the BGP selection policy, it propagates it to other BGP routers of the domain (as it is currently the case). Therefore ASBR_1 receives from ASBR_2 the update, which allows for the conventional propagation of the BGP control plane inside AS A.

Beside the BGP propagation, once ASBR_2 receives the BGP updates, it sends it unchanged to the mapping system. The MS is therefore aware of the prefix, the BGP metrics and the identifier of the exit ASBR (i.e., ASBR_2). The MS is able to filter the update according to the local policy. If accepted, the BGP update is transformed into an IDRD update and inserted into the MS data base. This IDRD entry contains at least the prefix, the AS path, the identity of the exit ASBR and the Path-ID to be used.

Once the information is saved, the mapping system configures the forwarding elements and potentially propagates the IDRD routing update to neighboring ASes' mapping systems as described in the previous section.

It is important to note that the insertion of BGP routes in the mapping system has the same reactivity as the BGP propagation. Indeed BGP updates and withdrawals are sent to the MS as soon as ASBR_2 receives them, leading to an immediate repercussion of the underlying routing changes on the IDRD control plane.

Life time of a packet

We focus here on the IDRD data plane, which is in green in Figure 3.4.

Two cases arise. First we focus on the traffic which destination is prefix X. An incoming flow arrives at ASBR_1 from an external element. The incoming packets are already encapsulated and contain the required information to identify the path which is to be followed - i.e., a destination prefix (i.e., X) and a Path-ID value (i.e., Path-ID_1). ASBR_1 extracts from its forwarding data base the forwarding instruction. In this case, the incoming couple X+Path-ID_1 is associated with the outgoing couple ASBR_2+Path-ID_2. Therefore ASBR_1 must send the packets to ASBR_2 and use the Path-ID_2. Then it de-capsulates the incoming packets and encapsulates them, put the identifier of ASBR_2 and the Path-ID_2 in the extra header and forward them toward the path to ASBR_2. When the packets arrive to ASBR_2, the same process is used and the packets are forwarded to ASBR_3 with Path-ID_3. It is important to notice that, all along the path, the inner IP header, which contains the final destination prefix X, is not modified. Indeed the modifications only occur at the extra header level.

Then we focus on the incoming flow which is destined to prefix Y. ASBR_1 forwards the packets according to the previously described process. Arriving at ASBR_2, the packets are processed differently. Indeed the outgoing path is not

an IDRD path in the sense that packets are forwarded to a neighboring ASBR that does not deal with IDRD. Therefore the forwarding data base of ASBR_2 associates the incoming packets' couple X+Path-ID_N with a conventional BGP next hop (i.e., ASBR_4). ASBR_2 de-capsulates the packets and forward them directly to ASBR_4.

3.3 Relaxation of ISP policies

Current BGP operates at the granularity of individual ASes and it defines a path as a sequence of domains. Following this logic, IDRD defines route diversity as a set of paths (for a common prefix), whereby each path is defined as a sequence of domains.

In the next paragraphs, we explore ways for circumventing the actual constraints on the selection of routes, including the *prefer client* rule, on inter-AS level path diversity. We first examine the negative effects that disregarding the *prefer client* criterion has on inter-domain path stability by using an instructive example. Subsequently, we propose a stability criterion which provides ISPs with a vastly increased number of candidate routes, while still safeguarding the stability of inter-domain paths.

3.3.1 Need for stability

Motivation for relaxing the selection of paths

BGP, as the *de facto* inter-domain routing protocol, is mostly configured such that it reflects the Valley Free and *prefer customer* policies [31]. Concretely, Valley Free export policies reflect the business relationships of domains with the following rules: (a) ASes export only their selected customer routes (i.e., routes received from customer domains) to their peers and providers, (b) ASes export all of their selected routes to their customers, irrespectively of whether they come from clients, peers or providers. These policies are meaningful in both single-route and multi-route propagation contexts as peering ASes are not willing to propose transit between non-client neighbors (cf. Section 1.1.3).

Additionally, and as already mentioned in the section 1.1.3, the so-called *prefer customer* import rule states that a client route towards a particular prefix must always take priority in use and propagation over routes towards the same prefix which were learnt from peer or provider ASes. Gao and Rexford prove that this import rule ensures the stability of the single path BGP inter-domain routing [31]. The *prefer customer* rule makes sense in a single-route propagation context as, if a domain has to select only one route, it would naturally select and propagate the one coming from its clients. Nevertheless, this rule does not make much sense in the context of multiple-route propagation. Indeed, once a domain has propagated its clients' routes to its neighbors (and thus fulfilled its provider role), it may find interesting to also propagate its peers' and providers' routes at the same time. Consequently, relaxing this condition may become advantageous. Fig. 3.5 illustrates the potential of using both client and peer/provider routes: AS_1 receives

routes to the destination AS from a client (AS_4), a peer (AS_2) and a provider (AS_3). With the "prefer customer" rule, AS_1 can only propagate its client routes (i.e., routes coming from AS_4) to its client AS_0, whereas relaxing this rule would enable it to propagate all of its routes to AS_0.

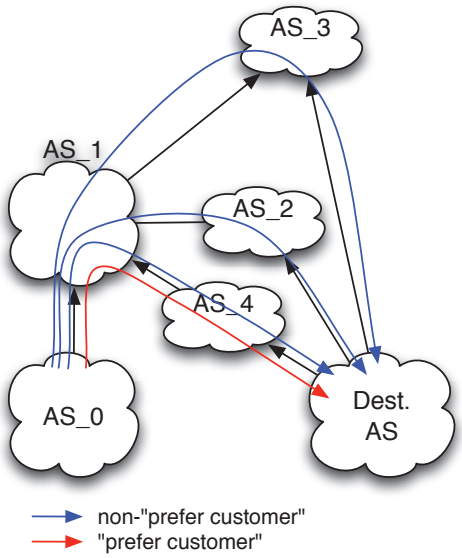


Figure 3.5: "Prefer customer" and non-"prefer customer" potential diversities.

While BGP currently operates sufficiently well, putting into use the vast diversity which is present in the Internet graph would of course yield a number of benefits (cf. [106] and Section 1.2). For instance, having available a set of multiple disjoint paths can be used for increasing resiliency [107]. The evaluation of path disjointness (cf. Section 3.4) demonstrates that disjoint paths can be identified for reaching almost all domains in today's Internet. This use case is especially interesting as it greatly increases network resiliency and capacity without significantly increasing router FIB size. Nevertheless we show in Section 3.4 that the "prefer customer" rule truncates a lot the usable path diversity and that path disjointness suffers from this truncation. Without the availability of such a simple service, NSPs will hardly adopt IDRD. The relaxation of the path selection, including the "prefer customer" rule, is therefore a requirement for the adoption of a multipath solution.

Policy relaxation related work

Wang et al. introduce in [61, 108] some ways to make the inter-domain path selection more flexible. Even if this work does not take into account multipath inter-domain routing, it emphasizes well the need for flexible decision processes. They underline in the same work the "Neighbor Specific" concept, which consists in the propagation of a different route to each neighboring domain, as one of the key

points for the relaxation of inter-domain policies.

Some mathematical works study Neighbor Specific (NS) BGP [85, 109] and have proposed a wider relaxation of the inter-domain routing. Indeed, in addition to the propagation of different routes to different neighbors, these works address the issue of relaxing the "prefer customer" rule, which ensure the global stability of the Internet [31], and prove that, in NS-BGP, this rule can be relaxed without harming the global Internet stability.

Nevertheless, these works deal with single path inter-domain routing while we adopt in this thesis an inter-domain multi-path approach.

All these works require by construction a multipath routing within an AS, as different routes will be offered and thus used by different neighbors. However, a **single** route is offered to each individual neighbor, whereas we consider, in this thesis, the possibility of exchanging a variety of paths between carriers, which requires specific **multipath** route export policies.

The proposed "prefer customer" relaxation is proven to be stable thanks to a proof given by Wang et al. in [85]. Nevertheless, this proof requires that the routes must be **strictly ordered** by the decision process in order to select the best route to be propagated to a neighbor. This constraint is not respected in our multi-path propagation as several paths are propagated to a neighbor and have therefore either the same rank or are not comparable. It makes the proof of [85] hard to adapt to our multipath context.

We consider that our inter-domain multipath propagation approach and the NS approach are orthogonally independent as domains may selectively adopt one without the other. This is the reason why we prove, in this section, the stability of the inter-domain routing relaxation in a worst case scenario - i.e. taking into account an inter-domain multipath and Neighbor Specific scenario.

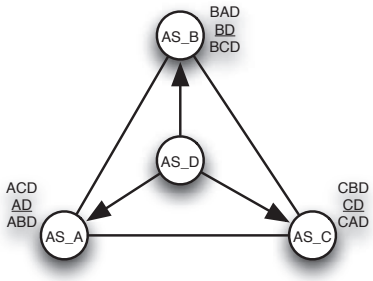
Instability example

Fig. 3.6 provides a simple but stunning example for the instabilities which might occur if the "prefer customer" rule were not respected. For the sake of clarity, we focus on a single route example, although such an instability may also appear in a multi-path context. AS_A, AS_B and AS_C are peers, and AS_D is the client of the other three ASes. Each AS uses a local decision process and ranks the candidate paths for route choice by priority. In our example, AS_A ranks the paths as ACD, AD, ABD. Nontheless, it can only advertise the path AD to its peer neighbors due to the Valley Free conditions.

The table in Fig. 3.6 presents the stepwise changes of path selection in each AS. Thereby, the selections which have changed since the last step are highlighted in red, the selections which are being propagated to neighbors are underlined, and paths which are being unselected (withdrawn) are ~~crossed out~~. A selected route is propagated to all the neighbors with respect to the Valley Free conditions.

According to the previously listed priority list, if AS_A receives the path CD from C (as in Line 3 of Fig. 3.6), it selects the path ACD and withdraws the path AD. And concurrently, AS_A sends the withdrawal of AD to its neighbors. However, it does

not propagate the newly selected path ACD to its peers (due to adherence to the Valley Free export conditions [31]).



Step	AS_A	AS_B	AS_C
1	AD	BD	CD
2	ACD (AD)	BD	CD
3	ACD	BD	CD
4	ACD	BD	CBD (CD)
5	AD	BD	CBD
6	AD	BAD (BD)	CBD
7	AD	BAD	CD
8	ACD (AD)	BAD CD	
9	ACD	BD	CD

Figure 3.6: Gadget example.

We can see from the table of Fig. 3.6 that Steps 3 and 9 are identical, which implies that the system will enter into oscillations (cf. also [32] for further examples of instabilities in inter-domain routing). It is interesting to note that this example does not oscillate by applying the "prefer customer" rule. Indeed, AS_A's decision process would have to be changed to prefer the path AD, then avoiding the oscillation.

In a multipath environment, the "prefer customer" rule is not mandatory (as explained in Section 3.3.1), but some caution still needs to be taken to avoid instabilities. The next section underlines the sufficient conditions to reliably avoid oscillations while relaxing the "prefer customer" rule in IDRD, followed by a comprehensive mathematical proof of IDRD's stability.

Notations

A domain d is connected to sets of clients, peers and providers, respectively defined as $C(d)$, $Pe(d)$ and $Pr(d)$. Domain d receives:

- A set of routes coming from $C(d)$ (i.e., client diversity): $R_{d,C}$ (where d denotes a *domain* and C denotes the whole set of its *clients*).
- A set of routes coming from $Pe(d)$ and $Pr(d)$ (i.e., peer/provider diversity): $R_{d,P}$ (where P denotes the whole set of its *peers and providers*).

d may export a different set of routes to each neighbor n than the one exported to another neighbor m even if both have the same relationship with d (i.e., peer, client or provider). This kind of routing is called "Neighbor Specific" [85]. For each neighbor n, domain d uses a decision process λ_d^n to select routes, among the two sets of route candidates, that will be exported to n: $E_d^n = E_{d,C}^n \cup E_{d,P}^n = \lambda_d^n(R_{d,C} \cup R_{d,P})$ where E_d^n denotes the set of routes selected by the domain d and exported to n, $E_{d,C}^n \subseteq R_{d,C}$ stands for the set of selected client routes and $E_{d,P}^n \subseteq R_{d,P}$ denotes the set of selected peer/provider routes, which are exported to n.

It is common that different routers of a specific domain export different sets of routes to the same neighbor. In this stability analyse, a given path, being advertised

by two routers, is considered as two paths. Therefore we only consider domain to domain route export policies, where a route propagated through different ASBRs is considered as several routes.

It is important to note that our formalization and the stability proof in Sect. 3.3.3 are compatible to both domains that consider the NS approach and domains that do not (or partially) consider it. In the non-NS approach, the set of routes propagated to different peers/providers is identical (i.e. $\forall x, y \in P$, $E^x_{d,P} = E^y_{d,P}$), as is the set of routes propagated to the different client neighbors (i.e. $\forall x, y \in C$, $E^x_{d,C} = E^y_{d,C}$).

While in this work we will not provide exact formulations of the route decision process λ, we do stress that the importance of relaxing the decision process both in order to increase the potential path diversity and to allow for customized decision processes. To this end, we can imagine a broad spectrum of approaches, ranging from simple selection criteria (e.g., selecting only disjoint routes) to more sophisticated multi-criteria schemes optimizing other objective functions.

In the next section, we provide an example for instability which can result from a too generous relaxation of path constraints. Subsequently, we provide a novel criterion in Section 3.3.1 which relaxes the “prefer customer” path selection rule, while reliably avoiding instabilities.

IDRD stability conditions

In the above example we have highlighted that the relaxation of path selection constraints can lead to oscillations. The potential oscillations stem from the fact that disregarding the “prefer client” criterion may lead to a cessation of client routes’ advertisement, meaning that provider-ASes discontinue their stable provisioning of routes for their clients’ prefixes and thus fall short of their role. Therefore, in this part we present a criterion which – in addition to the Valley Free conditions [31] (cf. Section 3.3.1) – relaxes the “prefer customer” condition while providing a new way to protect and ensure the reachability of customer prefixes. In terms of IDRD, the well known Valley Free conditions translate to:

- Each AS d sends a set of selected client/provider/peer routes ($E^x_{d,C} \cup E^x_{d,P}$) to a client x.
- Each AS d sends only a set of selected client routes ($E^x_{d,C}$) to a peer or provider neighbor x. $\lambda^x_d(R_{d,P}) = \emptyset$, so $E^x_d = E^x_{d,C} = \lambda^x_d(R_{d,C})$

We assume in the present analyse that a customer-provider hierarchy exists between domains, as it is currently the case in the Internet [29]. Therefore no customer provider cycle exists, meaning that if x is a (direct or indirect) customer of y, therefore y can not be a (direct or indirect) customer of x. In such a hierarchy, tiers-one domains are domains that do not have any providers.

In order to ensure the global stability of IDRD, the following stability criterion introduces a strong requirement in addition to the two Valley Free conditions stated above.

IDRD Stability Criterion: Routes received from peers and providers must have no impact on the selection of client routes sent to peer or provider neighbors.
More formally, $\forall x \in P(d)$, we have:
– $E_d^x = E_{d,C}^x = \lambda_d^x(R_{d,C} \cup R_{d,P})$
– with $\forall R'_{d,P}$ and $R''_{d,P}$:
  – $\lambda_d^x(\mathbf{R_{d,C}} \cup R'_{d,P}) = \mathbf{E_{d,C}^x}$
  – $\lambda_d^x(\mathbf{R_{d,C}} \cup R''_{d,P}) = \mathbf{E_{d,C}^x}$

For the sake of clarity, this stability criterion may also be formulated in a somewhat less condensed manner. Essentially, the criterion underlines that the set of selected client routes, which are to be sent to peers or/and providers, must be independent of the set of received peer and provider routes. It is important to note that the selection of client routes which are to be propagated to client neighbors may depend on received peer/provider routes. The stated stability criterion is equivalent to combining two separate selection processes:
– One for peer/provider neighbors: $E_{d,C\cup P}^{x\in P} = E_{d,C}^{x\in P} = \lambda_{d,C}^{x\in P}(\cancel{\boldsymbol{R_{d,P}}} \cup R_{d,C})$
– One for client neighbors: $E_{d,C\cup P}^{x\in C} = E_{d,C}^{x\in C} \cup E_{d,P}^{x\in C} = \lambda_{d,C}^{x\in C}(\boldsymbol{R_{d,P}} \cup R_{d,C}) \cup \lambda_{d,P}^{x\in C}(R_{d,P} \cup R_{d,C})$

We wish to underline that inter-domain *multi-path* is not the source of instability, but rather that inter-domain instability stems from the full relaxation of route selection rules. Table 3.1 illustrates the different cases (e.g., multi-path versus single path) and highlights when instabilities may occur. The "prefer client" rule well ensures the stability of inter-domain routing, even in a multi-path environment (cf. Multipath BGP). However, as evaluated in Section 3.4, the amount of path diversity is very limited in a "prefer client" environment. And while the full relaxation of this rule may lead to instability (cf. last column of Table 3.1), our proposed IDRD stability criterion ensures global route stability in both single- and multi-path environments while providing for a substantial amount of path diversity.

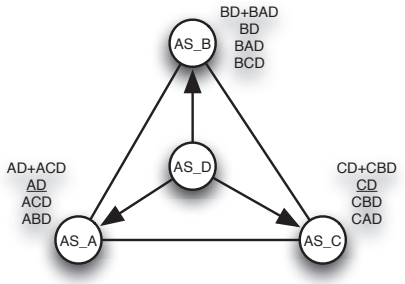
	"Prefer client"	New stability criterion	Full relaxation
Single path	**BGP: Stable**	Equivalent to "Prefer Client": Stable	Unstable
Multi-path	Multi-path BGP: Stable	**IDRD: Stable**	Unstable

Table 3.1: Identification of reasons for instability.

We can analyze the impact of this criterion on the oscillation example given in Section 3.3.1. In Step 4, AS_C receives the route BD from its peer AS_B and subsequently withdraws its client route CD propagated to AS_A and AS_B. The same occurs in Steps 6 and 8. However, this is strongly prohibited by the stated stability criterion, which mandates that the ASes select and propagate client routes,

to peers and providers, independently of routes received from peers and providers.

By taking into account the defined stability criterion, the system analyzed in 3.3.1 converges as shown in Fig. 3.7. The ranking of the paths is also modified to reflect the stability criterion.



Step	AS_A	AS_B	AS_C
1	AD	BD	CD
2	ACD AD	BD	CD
3	ACD AD	BD	CBD CD
4	ACD AD	BAD BD	CBD CD

Figure 3.7: Evolution of the routing decisions whith stability criterion.

3.3.2 Stability verification models and definitions

In order to prove the stability of IDRD, we first have to define the required verification tools. To this end, we follow the terminology and proof structure from the proof of stability for the Gao & Rexford conditions in [31].

We say that *activating* an AS applies the export policies of its neighboring domains. In other words, activating an AS causes its neighbors send it their routing messages once, after which the decision process is performed in order to select routing information that would be exported to neighbors.

The *activation sequence* tool is defined as a recursion of the activation over a series of domains. Using this tool, the propagation of a prefix on the path from domain d (which originates the prefix) to a domain r (which finally receives the routing messages) is performed by triggering sequentially the activation of all intermediate domains from d to r. During an activation sequence S, an AS d is given a set of activation numbers $A_S(d)$ where $A_S(d)(n)$ denotes the n^{th} activation number of d during the activation sequence S. For example, $A_S(d_1)(1) > A_S(d_2)(2)$ means that the activation sequence S activates d_2 at least twice before activating d_1 for the first time. We define $A_S(x) = \emptyset$ if x is not activated during the activation sequence S.

We define as *fair activation sequence* an infinite activation sequence in which every domain is activated an infinite amount of times. It is important to underline that if A_S is a fair activation sequence, $\forall d_1$ and d_2, two domains, and $\forall x$, an activation number of d_1, $\exists y$, an activation number of d_2, such that $A_S(d_2)(y) > A_S(d_1)(x)$ – i.e., at any step of a fair activation sequence, every domain will be activated at least once in a future step.

Further, we define the *state of the system* as the result of routing selection processes in the system. In the case of the Internet, the state of the system is the union of the selection process results in all of its domains. And in the case of a single domain, the state of the system is the set of selection process results within the domain itself.

In the rest of the Section, we call “IDRD system” the inter-domain routing system, where domains have adopted the IDRD architecture and are therefore able to propagate and select several paths.

Definition 3. *A state s is stable if it remains invariant under any activation sequence.*

Definition 4. *An IDRD system is safe if it always converges to a stable state from any initial state and within a finite number of activations.*

Theorem 2. *An IDRD system that respects the Valley Free conditions and the IDRD stability criterion is safe. Moreover, its stable state is unique if each decision process is deterministic.*

3.3.3 Proof of stability

Lemma 1. *An IDRD system which respects the stability criterion and the Valley Free conditions has a stable state. This stable state is unique if each decision process is deterministic.*

Proof

As the propagation/selection of route diversity for any specific prefix is independent from other prefixes, we only focus on the propagation of a single prefix that is originated in domain d.

In order to prove the present lemma, we use an activation sequence which contains two phases. Each phase can be considered as an activation sequence.

Phase 1 (P_1): We first *activate* the providers of domain d , and then their providers recursively , until activating the tier-ones [5]. Each domain is activated only once. Therefore, all direct or indirect providers of d (noted $Pr(d)$) are activated in Phase 1 in an order such that:

– if $Y \in Pr(d)$ and $X \in Pr(Y)$ then $A_{P_1}(X)(1) > A_{P_1}(Y)(1)$

– if $Z \notin Pr(d)$ then $A_{P_1}(Z) = \emptyset$

Phase 2 (P_2): After having completed Phase 1, we sequentially *activate* all the domains. We begin with all tier-one domains and activate a domain only when all of its providers have already been activated in Phase 2. Therefore, all ASes are activated in Phase 2 in an order such that if $Y \in C(X)$ then $A_{P_2}(X)(1) < A_{P_2}(Y)(1)$.

Phase 1 (P_1)

We make use of an induction proof in order to underline that if domains are activated according to the right activation sequence order (defined before), the routing diversity they export to their peers and providers is stable.

We consider the domain d as the base case (that is $A_{P_1}(d)(1) = 1$). d is then considered to be in a stable state. For the induction step, we activate the domain N. We assume that $R_{N,C(N)} = \cup_{Y \in C(N)} R_{N,Y} = \cup_{Y \in C(N)} E_Y^N$ is stable from, at least,

5. A tier-one is a domain that does not have any provider.

the previous step of the induction. Therefore $\forall R_{N,P(N)}$ and $\forall X \in Pr(N) \cup Pe(N)$, $E_N^X = \lambda_N^X(R_{N,C(N)})$ is stable thanks to the stability criterion.

As a result, once the first activation sequence has been completed, all (direct or indirect) providers of the destination have received a stable set of diverse client routes and export to their peers and providers a stable set of routes.

Phase 2 (P_2)

During the second activation sequence, we also use induction and we prove that the state reached by the domains at the end of this phase is stable. As the base case, all the tier-one domains (T_1) are first activated and we prove that the routes they export to their clients (i.e., $\cup_{Y \in C(T_1)} E_{T_1}^Y$) is stable as well. $\forall T_1$, $\forall Y \in C(T_1)$ we have:

$$E_{T_1}^Y = \lambda_{T_1}^Y(R_{T_1,C(T_1)} \cup R_{T_1,Pe(T_1)})$$

with

$$\begin{aligned} R_{T_1,Pe(T_1)} &= \cup_{X \in Pe(T_1)} E_X^{T_1} \\ &= \cup_{X \in Pe(T_1)} \lambda_X^{T_1}(R_{X,C(X)}) \end{aligned}$$

$\forall X \in Pe(T_1)$, $R_{X,C(X)}$ and $R_{T_1,C(T_1)}$ are stable since the Phase P_1. Therefore $E_{T_1}^Y$ is stable as well, $\forall Y \in C(T_1)$.

For the inductive step, we focus on the domain N and assume that all its providers send him a stable path diversity (i.e., $\cup_{X \in Pr(N)} E_X^N$ is stable). We prove that the diversity N exports to its clients (i.e., $\cup_{Y \in C(N)} E_N^Y$) is stable. $\forall Y \in C(N)$ we have:

$$E_N^Y = \lambda_N^Y(R_{N,C(N)} \cup R_{N,Pe(N)} \cup R_{N,Pr(N)})$$

with

$$\begin{aligned} R_{N,Pe(N)} &= \cup_{X \in Pe(N)} E_X^N \\ &= \cup_{X \in Pe(N)} \lambda_X^N(R_{X,C(X)}) \end{aligned}$$

and with

$$R_{N,Pr(N)} = \cup_{X \in Pr(N)} E_X^N$$

$\forall X \in Pe(N)$, $R_{X,C(X)}$ and $R_{N,C(N)}$ are stable since the end of the Phase P_1 and $\forall X \in Pr(N)$, E_X^N is stable since the previous iteration of the induction proof. E_N^Y is therefore stable as well, $\forall Y \in C(N)$.

At the end of the second activation sequence (i.e., after the completion of Phase P_1 and Phase P_2), all the domains receive and export stable client, peer and provider routes. All the domains are therefore in a stable state.

It must be noted that, as IDRD routing propagation is asynchronous, the stable state is not unique unless all the decision processes are deterministic. Therefore if the decision processes λ are deterministic, the domains' stable states reached at the end of Phase P_2 are the only ones that can be reached under any activation sequence.

□

Lemma 2. *An IDRD system that respects the stability criterion and the Valley Free conditions converges to its stable state for any initial state and under any fair activation sequence.*

As for the previous proof, we focus on the propagation of a single prefix, originated from domain d.

Proof

Given any fair activation sequence, we devise our proof by subsequently focusing on individual domains in the same order as the activation sequences of Lemma 1's proof (both phases). While the fair activation sequence may activate a lot of domains several times, we only focus on a single domain until its activation. We then focus on the next domain until its activation and so on., shifting our focus in the order foreseen in Phase P_1 and then Phase P_2. The activation sequence of this step is therefore separated in two phases: first, Phase P_1' where we shift our focus in the order of Phase P_1 and second, Phase P_2' where we shift our focus in the order of Phase P_2.

Phase 1 (P_1')

As for the Lemma 1, $A_{P_1'}(d)(1) = 1$.

For the inductive step, we focus on a domain N that has been activated in the first phase of the first lemma's proof (i.e., $A_{P_1}(N)(1) = n$). We prove that, if the first $n-1$ ASes of Phase P_1 currently export a stable set of diverse client routes to their providers, N selects and exports a stable routing diversity to its peers and providers. We assume that, after the activation number x of the phase P_1', the set of client route diversity received by N (i.e., $R_{N,C(N)}$) is stable. Because of the definition of a fair activation sequence, $\exists y, a \in \mathbb{N}^*$ such that $A_{P_1'}(N)(a) = y > x$. We can say that $\forall P \in Pr(N) \cup Pe(N)$:

$$E_N^P = \lambda_N^P(R_{N,C(N)}) = \lambda_N^P(\cup_{Y \in C(N)} E_Y^N)$$

With $\cup_{Y \in C(N)} E_Y^N$ stable given the previous iteration of the induction proof. Therefore the routing information that N exports to its peers and providers (i.e., E_N^P) is stable.

At the end of Phase P_1', each AS receives a stable set of diverse client routes and the sets of routes it exports to its peers and providers are stable as well.

Phase 2 (P_2')

Because of the definition of a fair activation sequence, $\exists y \in \mathbb{N}^*$ such that $\forall T_1$:

$$A_{P_2'}(T_1)(1) = y$$

and

$$\forall Y \in C(T_1)$$
$$E_{T_1}^Y = \lambda_{T_1}^Y(R_{T_1,C(T_1)} \cup R_{T_1,Pe(T_1)})$$

with

$$\begin{aligned} R_{T_1,Pe(T_1)} &= \cup_{X \in Pe(T_1)} E_X^{T_1} \\ &= \cup_{X \in Pe(T_1)} \lambda_X^{T_1}(R_{X,C(X)}) \end{aligned}$$

$\forall X \in Pe(T_1)$, $R_{X,C(X)}$ and $R_{T_1,C(T_1)}$ are stable since the end of the phase P_1'. $E_{T_1}^Y$ is therefore stable as well, $\forall Y \in C(T_1)$.

For the inductive step, we focus on domain N, which is the n^{th} activated domain in Lemma 1's Phase P_2. We assume that the routing diversity that N receives from its providers (i.e., $\cup_{Y \in Pr(N)} E_Y^N$) is stable since, at least, the previous step of the induction (i.e., the activation number z). Because of the definition of a fair activation sequence, $\exists a, y \in \mathbb{N}^*$ such that $A_{P_2'}(N)(a) = y > z$.

By its activation, N receives routes from neighbors and selects, for each client, a set among the received diversity to export it. $\forall Y \in C(N)$ we have:

$$E_N^Y = \lambda_N^Y(R_{N,C(N)} \cup R_{N,Pe(N)} \cup R_{N,Pr(N)})$$

with

$$\begin{aligned} R_{N,Pe(N)} &= \cup_{X \in Pe(N)} E_X^N \\ &= \cup_{X \in Pe(N)} \lambda_X^N(R_{X,C(X)}) \end{aligned}$$

and with

$$R_{N,Pr(N)} = \cup_{X \in Pr(N)} E_X^N$$

$\forall X \in Pe(N)$, $R_{X,C(X)}$ and $R_{N,C(N)}$ are stable since the end of Phase P_1' and $\forall Y \in Pr(N)$, E_Y^N are stable since the previous iteration of the induction proof. E_N^Z is therefore stable as well, $\forall Z \in C(N)$. At the end of the induction proof, all domains receive stable client, peer and provider routing diversities and they also export stable routing diversities. We conclude that all the domains reach their stable states at the end of Phase P_2', thus proving Lemma 2.

□

Lemma 3. *An IDRD system that respects the stability criterion and the Valley Free conditions always reaches a stable state within a finite number of activations.*

Proof

Given any fair activation sequence, each domain is activated an infinite number of times. We focus on a domain a. Domain a is separated from the domain d by at most a finite number of ASes (with no routing loops thanks to the path vector paradigm). Each intermediate domain is activated several times during the fair activation sequence. $\forall P$, a path from a to d, there exists an activation number A_p at which a receives this path. At activation number $A = max_p(A_p)$, a is aware of all the route/path diversity that intermediary ASes have selected and propagated. Therefore a receives the final set of diverse routes (potentially filtered by intermediary ASes) via IDRD in a finite number of activations, after which it reaches its own

stable state by applying its decision process. The same applies to all domains, such that the stable state of the global system is always reached under a finite number of activations.

□

3.4 Evaluation

3.4.1 Evaluation process

The provisioning of disjoint routes is one of the short term most prominent uses of routing diversity propagation. In this context, we evaluate the number of disjoint paths a domain can use to reach all the other ASes, whereby we underline that we address *potential paths*[6] which go beyond those currently propagated by BGP. To this end, we have established a dedicated evaluation methodology, as we cannot rely on data from the existing routing data retrieval projects (e.g., Route Views) due to their truncation inherited by the current use of the BGP decision process.

Maximum potential disjoint reachability

We denote by C_s and C_d the number of BGP neighboring domains, respectively for the source AS and the destination AS, and $P^{dj}_{s->d}$ the number of available disjoint paths for s to reach d. $P^{dj}_{s->d}$ has the following constraint: $max(P^{dj}_{s->d}) \leq min(C_s, C_d)$. Indeed the number of disjoint paths between two domains can not be higher than the number of neighbors of the source domain and the one of the destination domain.

Due to this constraint, we first evaluate the maximum number of disjoint paths each AS is reachable with. CAIDA [29] provides a list of relationships between pairs of neighbors[7] based on which we can evaluate the maximum disjoint reachability for each domain, i.e., the maximum number of disjoint paths for each domain, which is inherently limited by the number of neighboring domains. By taking into account only multihomed ASes (i.e., 23526 ASes in total), we obtain the black curve represented in Figures 3.12 and 3.11. As we only take into account multihomed domains, each domain has a maximum disjoint reachability that is strictly higher than 1.

Path disjointness

In the following, we briefly describe the process we use for obtaining the number of available disjoint paths between ASes. We consider a subset of domains, one by one, and compute disjoint paths between the considered one and all other ASes. The process itself consists of three major steps:

6. The notion *potential paths* is attributed to the fact that paths may be filtered by intermediate domains.

7. Dating from 16 January 2011.

1/ Generation of all the possible paths that IDRD (i.e., with the new stability criteria provided in Section 3.3) and the “prefer customer” rule may provide to a source AS. Figure 3.5 (in Section 3.3.1) shows the difference between the IDRD approach and the “prefer customer” approach.

As the performed evaluation addresses the problem of path disjointness, we do not take into account artificially longer paths. We denote as an artificially longer path a path where a part of the AS path can be replaced by a single inter-domain link. We can see in Figure 3.8 that the AS_4 can be replaced by the link that goes directly from AS_1 to AS_2. Therefore taking into account AS_4 leads to an artificially longer path and does not provide a new alternative to find a disjoint path.

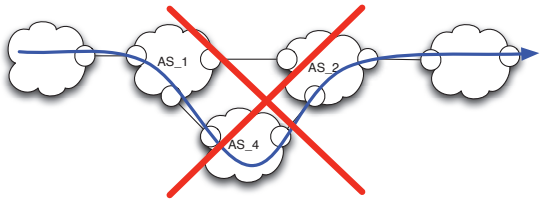


Figure 3.8: Artificially longer path.

Nevertheless, we take into account longer path that can not be reduced by replacing an AS with an inter-domain link (as in Figure 3.9). Such a longer path may be useful as the ASes that make the new path longer (i.e. AS_4 and AS_5 in Figure 3.9) could make the path interesting by avoiding other ASes (i.e. AS_2 in Figure 3.9)

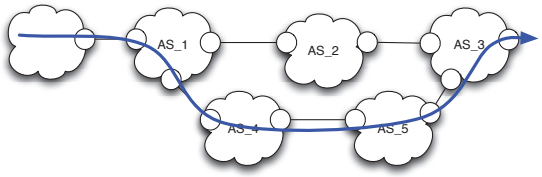


Figure 3.9: Non artificially longer path.

2/ Disjointness analysis: Each generated path is compared to the other paths (for a given prefix) to highlight which pairs of paths are disjoint, according to their AS paths.

3/ By using a graph representation, the maximum clique problem helps us to compute the maximum number of disjoint paths an AS may be reached with. Figure 3.10 provides such a representation. Each path is represented by a vertex, while an edge connects two vertices only if the represented paths are disjoint. Therefore, a clique is a subset of the graph in which each vertex represents a path that is

disjoint from all other paths represented by the vertices of the clique. The largest clique in the Figure 3.10 is $\{P1, P2, P3, P4\}$ and its size is 4.

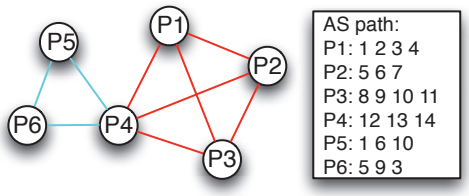


Figure 3.10: Example of cliques in a graph.

The computation of the maximal clique size is exponential with regard to the number of vertices (i.e., $O(3^{n/3})$ where n stands for the number of vertices – paths). Due to this computational complexity, we have performed the evaluation on 10% of the destination ASes (i.e., 2137 multi-homed domains) [8]. These ASes have been selected randomly.

IDRD & "prefer customer" comparison

For the sake of comparison, we highligth the difference between IDRD and the "prefer customer" rule by performing the evaluation on both cases.

If a customer route exists, the "prefer customer" rule implies that no peer or provider routes are usable.

Figure 3.5 (Section 3.3.1) provides an illustration to underline the difference between both cases. AS_1 is the provider of both AS_0 and AS_4, the peer of AS_2 and the client of AS_3. It must be noted that AS_0 may either be a stub or a transit network. With the "prefer customer" rule, AS_1 uses only its client diversity whereas IDRD allows AS_1 to use all its routes, irrespectively of whether they come from clients, peers or providers, and to propagate them to AS_0. We evaluate, in the next section, the differences between both approaches in finding disjoint routes.

Interestingly, the number of additional paths produced by the proposed relaxation (i.e., IDRD) is so large that we had to limit the allowed path length in order to limit the total number of paths to be evaluated. Therefore, we only considered (for each prefix) those paths which are not longer than the best AS path length plus 2. This limitation of the number of IDRD paths has therefore two consequences. Firstly, it reduces the potential diversity, such that the conclusions of the evaluation are at most negatively biased towards IDRD. Secondly, all the paths that are taken into account have a length that is comparable to the shortest path (no more than 2 ASes in addition to the shortest AS path).

8. Other methods could be used to achieve this evaluation. Nevertheless [110] proves that computing a maximum number of vertex-disjoint paths in acyclic graphs is NP-hard.

3.4.2 Results

Figures 3.12 and 3.11 show the disjoint path diversity that IDRD and the "prefer customer" rule, respectively, can provide to ASes. In black are the connection graphs (the same curve for both Figures 3.12 and 3.11) representing the external connectivity of domains (e.g., the number of eBGP neighbor relationships provided by CAIDA) which stands for the maximum disjoint reachability. A large proportion of the ASes are connected to two domains (i.e., more than 50%). The other ASes are connected to 3 or more domains and may potentially be reachable via 3 or more fully disjoint routes.

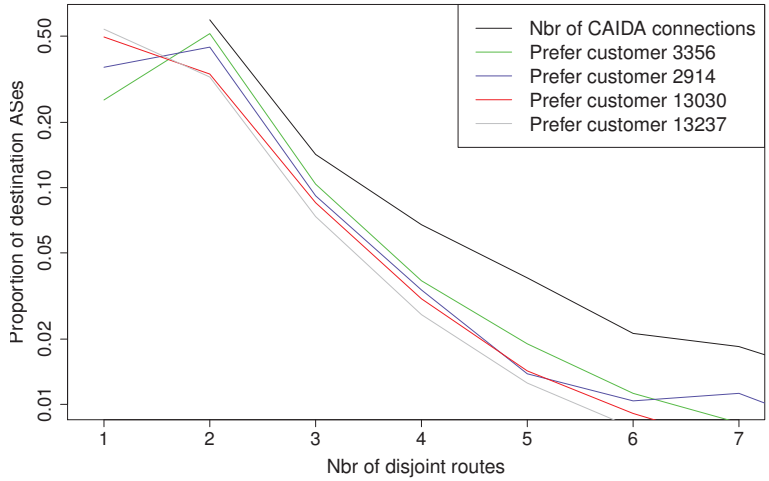


Figure 3.11: "Prefer customer" rule potential disjoint paths.

The first result that can be underlined is that the "prefer customer" rule truncates the potential disjoint diversity. The curves of all the analyzed domains in Figure 3.11 are below the one of the destination AS connectivities, which corresponds to a loss of diversity. In contrast, the diversity of IDRD (cf. Figure 3.12) is very close to the underlying potential diversity. We can see that the AS 3356 has a little percentage of destination domains (i.e., 1.6%) that are not reachable via disjoint paths. This is a consequence of the limitation of the number of paths we have taken into account for the evaluation of IDRD (cf. Section 3.4.1).

With the "prefer customer" rule, the number of "no disjoint path" ASes varies a lot depending on the AS_{source}. For instance, AS 3356 (rank CAIDA: 1) has about 25% of "no disjoint path" ASes whereas the AS 13237 (rank CAIDA: 46) has more than 50% of the destination ASes not reachable via disjoint paths. Using IDRD allows almost all domains (except for between 0% and 1.7%) to be reachable via multiple disjoint paths, and this figure is pretty much stable irrespectively of the analyzed AS.

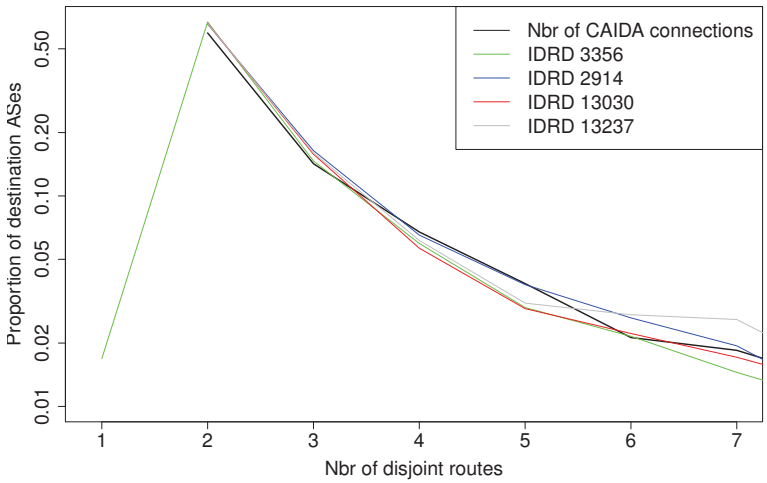


Figure 3.12: IDRD potential disjoint paths (relaxed route selection).

Some ASes in the IDRD graph (e.g., AS 3356) display a higher number of two-disjoint-paths destination domains than the proportion of domains having exactly two neighboring domains in the CAIDA connections graph (black curve). This is due to the fact that these extra destination domains have more than two BGP neighbors, but their disjoint path reachability is reduced to two because of the Valley Free export conditions of intermediate domains. Therefore, the increase in the number of two disjoint path domains is obtained at the cost of a decrease of more than two connections domains.

For the sake of presentation, we only presented full distributions for a small number of ASes. In the following, we focus on the number of "no disjoint path" destination ASes since path disjointness is probably the most interesting service which can be proposed by an architecture like IDRD.

We perform the analysis on 17 domains. Figure 3.13 presents the evolution of the number of "no disjoint path" destination ASes with respect to the CAIDA rank [9] of the ASes for both IDRD and the "prefer customer" rule.

This graph underlines that the "prefer customer" rule does not provide the same level of diversity for tier-ones, tier-twos or even lower tiers. Indeed, the number of "no disjoint path" destination ASes is quite already important for the tier-ones (i.e., 25% or so destination domains), and furthermore this figure almost doubles for ISPs that have a CAIDA rank lower than 20 (i.e., more than 99% of the ISPs). On the contrary, the relaxation proposed in IDRD is almost insensitive to the rank of the AS source. This means that finding disjoint paths could be an incentive to adopt IDRD, both for small and for large ISPs.

9. http://as-rank.caida.org

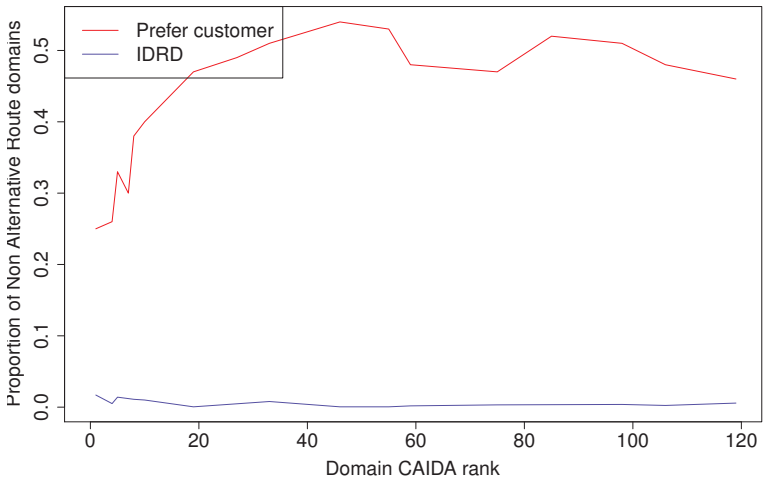


Figure 3.13: Proportion of “no disjoint path” destination ASes.

3.5 Conclusions

Today’s Internet displays vast potential path diversity due to the high degree of connectedness between the individual Autonomous Systems (ASes). However, BGP-4 as the single inter-domain routing protocol in the Internet does not allow for the utilization of multiple paths per destination prefix.

Starting with a motivation for AS-level path diversity, in this chapter we propose Inter-Domain Route Diversity (IDRD) as a novel, easy-to-deploy architecture which allows for the use of multiple inter-domain paths, while maintaining full backwards compatibility with the current Internet due to its reliance on existing and compatible technologies (i.e., LISP [28]). Indeed both current routing plane and data plane remain in parallel with IDRD, which ensure full incremental adoption.

One of the fundamental advancements of IDRD lies in the relaxation of BGP’s “prefer customer” rule, which currently impedes the use of the underlying diversity. To this end, we specify a novel route selection criterion which assures the stability of the global control plane when utilizing and propagating multiple paths per IP prefix. The new stability criteria allows a service provider to propagate and use several path, could they be client, peer of provider paths.

We have compared the IDRD relaxation to the “prefer customer” selection rule and we demonstrate that our relaxed route selection greatly increases the offered path diversity. First it allows NSPs to receive disjoint routes for a high proportion of destination, which is a required condition to perorm efficient fast recovery. Second it opens the potential for many useful customer applications and routing market, as studied in Chapter 5.

Chapter 4

Wide path diversity propagation: scalability analysis of path identification schemes

4.1 Introduction

With the current de facto inter-domain routing protocol (i.e., BGP), only one path is put into use and the benefits of path diversity, inherently present in the Internet, are never harvested. We proposed in the previous chapters ways to propagate and use this diversity. On the control plane level, our work relies on the propagation of several routes per prefix on an higher layer, keeping BGP running underneath. On the data plane level, they rely on encapsulation in order to enforce the traffic along a specified path. In order to segregate flows following different paths, we proposed in Section 3.2 the use of an additional identifier carried in packets and processed by the routers of the path. Therefore two flows having the same destination address may be forwarded differently according to the identifiers (i.e., the Path-IDs) they carry.

Enabling Internet path diversity potentially faces both scalability issues (i.e., insertion of additional routes into the Forwarding Information Base of routers) and policy violation issues (e.g., using non Valley Free routes) - cf. Section 4.2.3. The way in which identifiers are globally organised has a direct impact and may either solve or worsen both issues.

We propose in this chapter to present proposals of global organisations of identifiers and to evaluate their scalability. Advanced filtering may be performed to lower the scalability issue (as proposed in Chapter 5) but we aim here at studying the worst case scenario (i.e., when all routes are propagated) to study the different Path-ID schemes. The chapter is organized as follows: We present the context of the path identification in Section 4.2, including the issues that such an approach can encounter. In Section 4.3, we present some examples of global path identifications and evaluate their scalability in Section 4.4. Finally, we present in Section 4.5 some simple assumptions/modifications that allow to make a global path iden-

tification without encountering scalability issue.

4.2 Path identification:

4.2.1 The need for identification

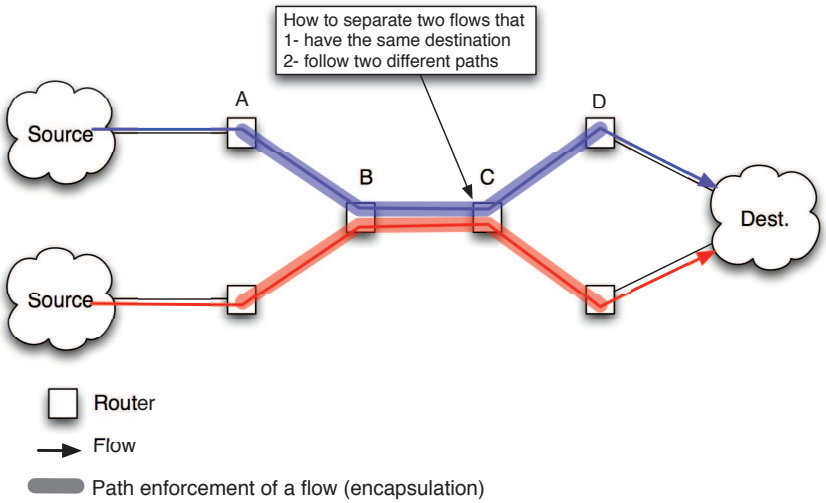


Figure 4.1: Points of identification.

The proposed architectures specify that several flows, addressed towards the same destination and identified by the destination address, may follow different paths. At least one of the paths which are to be used is not the path chosen by the underlying routing protocol.

Our approach relies on some key points of the architectures described in the Chapter 3.

Packets encapsulation

Packets are encapsulated to enforce their path till a deflection point/router, which lies on the chosen/alternative path. Original packets are not modified and an alternative path may be composed of several successive defection points/routers. From that consideration, the deflection routers must be capable to recognize the different flows in order to make them follow the correct path (cf. Figure 4.1). Nevertheless, the destination addresses of the original packets are not anymore sufficient to separate the individual flows following a path as the different paths potentially have the same destination prefix but nevertheless need to follow different paths.

Path identification

In addition to the inner destination address, an identifier must be used for specifying the route which is to be used. We name that identifier a Path Identifier (Path-ID) and each encapsulated packet carries its Path-ID. For scalability reasons, the value of this identifier must be local to each hop. However, at the same time, path-ID values must be aligned all along a path in order to assure end-to-end path coherence. In practical terms, intermediate routers must be able to swap incoming identifiers with the appropriate outgoing path identifiers.

Already existing labels or identifiers can be used as Path Identifiers. For instance, MPLS encapsulation path enforcement can use MPLS label [111] (20 bits per incoming/outgoing router interface) as Path-ID whereas IP encapsulation path enforcement can use the IPv6 Flow Label [112] (20 bits) or the 64 lowest-order bits of /64 IPv6 addresses and the LISP Instance ID for the LISP path enforcement [28] (24 bits). All these possibilities must only impact the extra header. The user packet (including the original/inner header) must not be modified.

4.2.2 Identifying what and where ?

Flow identification

In order to identify the appropriate path, the system has to identify the flows and deduce the path to be used. This identification takes place at the point A in Figure 4.1. This point is located at the frontier between the two following areas:

- The part of the network where packets are forwarded thanks to the routing protocol only (according to the destination IP address).
- The part of the network where packet path is enforced thanks to encapsulation. In this area, the flows are not always following the path provided by the underlying routing protocol.

At point A, the system identifies the flows according to the original IP header only. The result of this identification is the enforcement of the path of the flow and the ad-junction of new information in the packet (i.e., the path identifier)

Path identification

Path identification is performed at the points where the path needs to be enforced (e.g., points B and C). Contrary to the point A, the path identification at B and C does not re-identify the flows. Path identification only identifies the path that flows are forwarded to.

No identification

At the exit of the path enforcement area, flows can enter in an area where no path enforcement is performed (i.e., after the router D). Therefore, no flow or path identification is needed. Packets are forwarded according to the destination IP address, as it is currently performed in the Internet.

4.2.3 Path identification challenges

We address the path identification issue on an inter-domain basis. Therefore, we define a path as an AS path, regardless of the router level. However, the routing decisions are still performed at the router level and the knowledge of appropriate labels is thus mandatory at this level.

An efficient identification scheme needs to address the following points/issues:

– Forwarding information base (i.e., FIB) scalability: The current FIB in the Internet default free zone (i.e., DFZ[1]) is already about 450 000 [38] lines and has been growing at a fast pace during the last decade. The chosen identification must not worsen the scalability issue at the FIB level.
– Policy compliancy: The identification scheme must not allow a domain to use a path that it is not allowed to use (e.g., which is not compliant with the policies of its neighbors). We do not discuss here about potential policies which could be adopted by domains.

4.2.4 Local ways of identification

Several identification ways may be adopted.

Identification based on the original IP header

In addition to the IP destination address, the original header contains other fields that could potentially be used to segregate traffic on a path from the traffic of another path. Each router on a chosen path must then know the association between the set of fields and the next hop. The size of the forwarding plane is then linear with the number of paths allowed per prefix. Performing this identification method may be possible at the edge of the Internet as stub networks may directly negotiate with their NSPs the semantics of the IP field values. Nevertheless, it will be hardly possible to implement this type of identification in the DFZ of the Internet.

Identification based on both identifier and original destination IP address

For a given destination prefix, this identification scheme permits to directly associate the identifier with the path the packet has to follow. This scheme allows theoretically to adopt a huge number of paths per prefix (1 Million for a 20 bit path identifier). It could be interesting that two destination prefixes reachable via a common path be aggregated into the forwarding table. Nevertheless such a prefix aggregation[2] can not be performed with this proposal.

Identification based on the identifier only

With this solution, a router receiving a packet will perform the forwarding function according to the Path-ID only. This solution may have some limitations. A 20

1. DFZ: the default free zone is the part of Internet where routers have no default route.
2. Prefix aggregation: aggregation of all the prefixes of a domain into a single forwarding entry, as proposed in Section 4.2.5.

bit field allows the router to identify, at most, 1 048 576 paths. The BGP routing table of a router in the default free zone already contains 450 000 [38] or so lines. A 20 bit identifier would then limit between 2 and 3 the number of different paths per prefix. Whereas 2 or 3 paths seem to be sufficient for fast recovery purpose, it becomes very limited if NSPs want to provide advanced services. Nevertheless, a Path-ID may be used to identify a path to a domain (or a set of domains), instead of a prefix. There is currently 60 000 or so domains. Therefore this solution would provide (on average) 17.46 paths per destination domain.

In the current chapter, we try to estimate, in a worst case approach, the number of entries that would be necessary to take into account the current potential diversity (cf. evaluations in Sec. 4.4 and 4.5). Beside this important issue, it may be interesting to take into account the future potential growth of this number of entries. The figures 4.2 and 4.3 respectively present the evolution of the number of ASes and the evolution of the number of propagated prefixes in the Internet (graphs taken from [38]). Whereas the number of prefixes evolves exponentially, the growth of the number of ASes is linear. Therefore identifying a destination as a domain instead of a prefix has the interesting property to depend on the linear growth of the number of ASes and not on the exponential one of the number of prefixes.

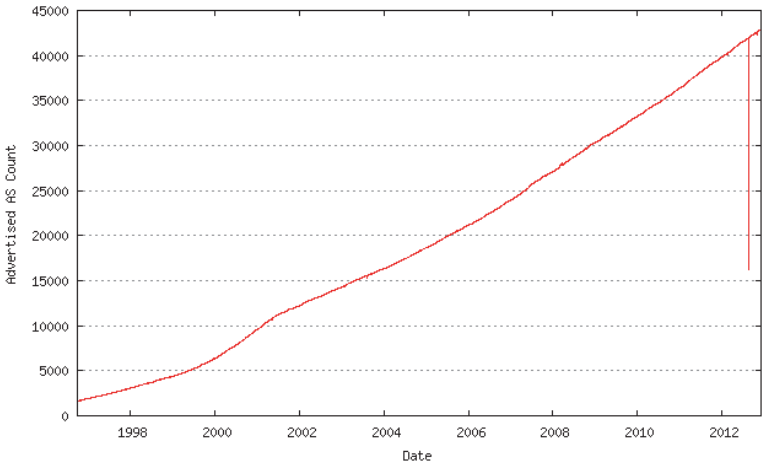


Figure 4.2: Evolution of the number of ASes (source: [38]).

4.2.5 Where can the scalability issue be addressed?

At the flow identification router (point A)

At the router A of Figure 4.1, the transition between basic IP forwarding and path enforcement encapsulation is performed. Therefore, this router knows the association between the fields of the original header (e.g., address, port, DSCP, etc.) and the path ID that will be used to forward the packets till the last domain. MPLS [33] uses the same type of association thanks to the Forwarding Equivalent

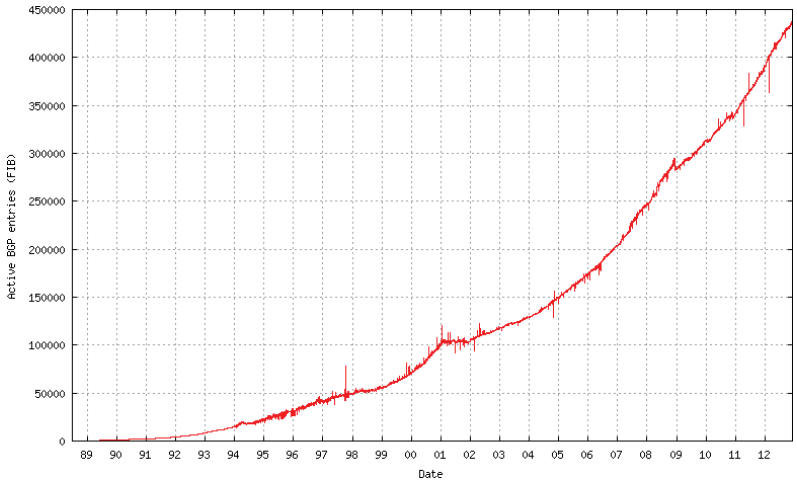


Figure 4.3: Evolution of the number of propagated prefixes (source: [38]).

Class (FEC). In our context, we use the same term for clearer descriptions and explanations.

The number of entries in the FEC depends on both the number of prefixes and the number of recorded paths per prefix. The Internet routing table currently contains 450 000 entries [38]. Allowing each entry to have X paths would make the number of routing entries increase by a factor of X, which is not acceptable, regarding the current size of the routing table.

By placing router A close to the source network, it is be possible to minimize this impact of issue. Indeed, a single source network hardly needs a lot of routes for all prefixes but would potentially need some alternate routes for a subset of prefixes, only. In that case, each domain can negotiate with its provider the characteristics of the entries of the FEC (i.e., amount of entries, types, etc.). On the contrary, placing the router A far from the source network would advance it to being shared between several source networks. Consequently, this would induce a substantial increase of the FEC due to the addition of the routing requirements of all the source networks.

At the path enforcement routers (points B and C)

At the points B and C, only path identification is performed.

Two contradictory effects impact the scalability of point B and C:

– Whereas point A may be placed in the source network, points B and C are far from it. They are therefore used by a great number of source domains, thereby aggregating their resource needs. It has for consequence to make the number of used path increase.
– While flow identification at point A is performed based on the IP destination header, the path identification can be performed, at points B and C, based on the Path-ID only. In such a case, some aggregation may be performed.

Several prefixes belonging to the same domain can be mapped to the same Path-ID at point A. Therefore, the number of Path-IDs points B and C need to manage would decrease.

At the path enforcement exit router (point D)

At the point D, a packet is forwarded thanks to its destination IP address, as it is currently performed. There is therefore no scalability issue at this point.

4.3 Analysis of specific path-ID schemes

In the following parts, we briefly analyse some identification scheme examples. These schemes are analysed according to the point of view of the scalability and path compliancy (defined in sections 4.2.3 and 4.2.5). No path filtering is performed to conduct a worst case scenario, therefore all the paths are propagated from a domain to its neighbors.

4.3.1 Stub-to-Stub Path-ID use:

In this scheme, the identification of the path is performed close to the source network. The forwarding of packets is then performed only according to the path-ID, included in the packet, till the destination network.

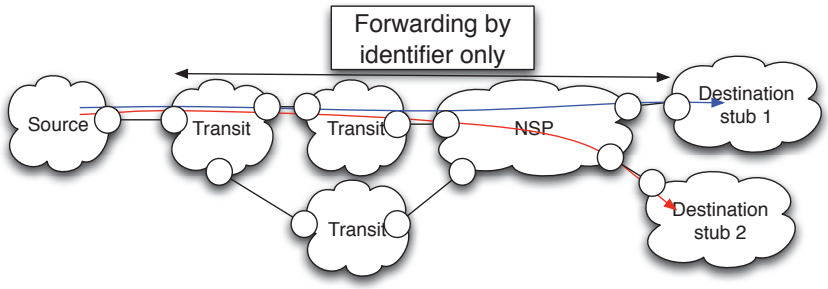


Figure 4.4: Stub-to-Stub path-id.

Let AS be the set of Internet domains, AS_i be the i^{th} domain of Internet and P_d^i the set of path from a domain d to AS_i. The number of potential paths received by d is then

$$\sum_{\{k|A_k \in AS; A_k \neq d\}} |P_d^k|$$

The FEC contains the association between a destination IP prefix and the diverse paths that can be used to reach that prefix. Therefore, a destination domain owning as set of X prefixes impacts the FEC by multiplying the number of paths to reach the domain by its number of prefixes X, at least when the aim is to stay compatible

with BGP which operates at the level of individual prefixes. Let Pr_i the set of prefixes originated from AS_i. The number of entry into the FEC in d is then the sum of the paths leading to each destination:

$$\sum_{\{k|A_k \in AS; A_k \neq d\}} |P_d^k| . |Pr_k| \tag{4.1}$$

Equation 4.1 shows that the number of entries in the FEC is directly proportional to both the number of prefixes originated from each destination domain and the number of paths potentially usable to reach each domain. It is important to underline that the number of prefixes advertised into the DFZ (currently 450 000) is increasing very quickly [38] and may have an important impact on the FEC.

Please note that even if a domain may technically be able to provide a particular path to its neighbors, it may still refrain from doing so if it considers this path not to be compliant with its policies. Technically, this can be simply implemented by not advertising the corresponding Path-ID.

4.3.2 Stub-to-Transit Path-ID use:

In this scheme, the path ID is only used to reach the penultimate AS of the AS path. Once packets arrive to this AS, it forwards them to the last domain according to their original IP addresses (cf. Figure 4.5). As we define it, the Path-ID identifies the set of domains the flow crosses. Once the packets arrive in the last transit domain, the end of the path can be unambiguously deduced from the original IP address. There is therefore no loss of path information and the diversity is the same as with the previous scheme (cf. Section 4.3.1).

Nevertheless, the number of Transit networks is far lower than the number of total ASes. Therefore, compared to the previous scheme, the number Path-IDs necessary to address all the potential paths decreases.

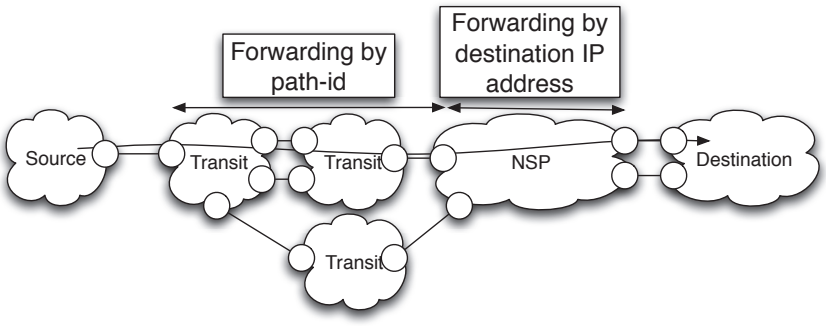


Figure 4.5: Stub-to-Transit path-ID use.

Let T be the set of Internet Transit domains. The number of potential paths

received by a domain d is then the sum of the paths leading to these transits:

$$\sum_{\{k|A_k \in T\}} |P_d^k| \tag{4.2}$$

It has been shown in [113] that more than 80% of Internet domains are non-transit domains (i.e., stub domains). Therefore, the current Path-ID scheme potentially makes the number of necessary Path-IDs drop. Indeed, the number of domains which needs to be addressed with the scheme drops from 100% (for the Stub-to-Stub Path-ID scheme) to 16% (for the current scheme).

The number of entries in the FEC depends both on the number of prefixes per domain (as for Eq. 4.1) Let $N(A_k)$ be the set of neighbors of A_k. Therefore, the number of entries in the FEC in d is:

$$\sum_{\{k|A_k \in T\}} \left[|P_d^k| . \sum_{\{i|A_i \in N(A_k)\}} |Pr_i| \right] \tag{4.3}$$

It can hardly be deduced from Eq. 4.3 whether the size of the FEC would reduce by adopting this scheme. We show in Section 4.4.1 that this scheme has got no impact on the size of the FEC in practice.

As for the previous scheme (i.e., Stub-to-Stub scheme), a domain which wants to prevent its neighbors from using one of its paths can do so by not propagating the associated Path-ID.

4.3.3 Source label stacking:

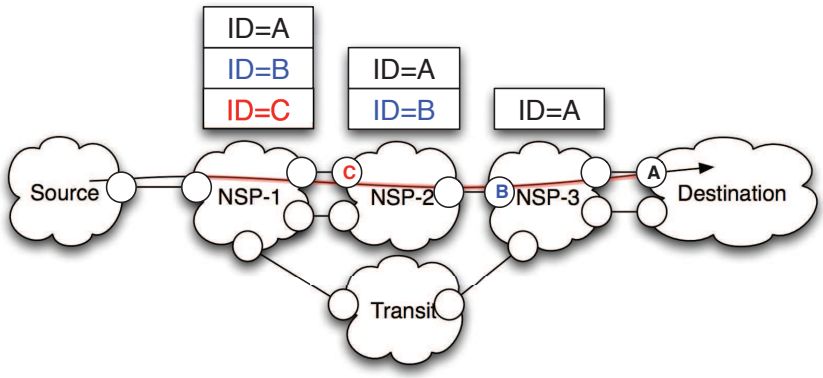


Figure 4.6: Source label stacking.

In this scheme, a packet is forwarded according to a local neighbor-specific label, whereas the Path-ID is a list of labels which identifies a series of domains lying on the path. With this technique, each AS would need a number of FIB entries that is the same than the number of its neighboring ASes. For instance, an AS

having 1 000 neighbors would only need 1 000 entries in its FIB. In order to assure the end to end path, each packet must carry a label associated with each domain of the AS path it will cross. This process is very close to the source routing [114] but applies at the inter-AS level. Therefore, no FIB issue is present. Note that the only difference between the current scheme and the Stub-to-Stub Path-ID scheme (cf. Sect 4.4) lies in the way the Path-ID is encoded: In the Stub-to-Stub case, the Path-ID has got global validity, while in the source label stacking scheme the individual Path-IDs have got only local/per-AS significance. The number of entries in the FEC is therefore exactly the same as the one of the Stub-to-Stub Path-ID scheme (cf. Eq. 4.1).

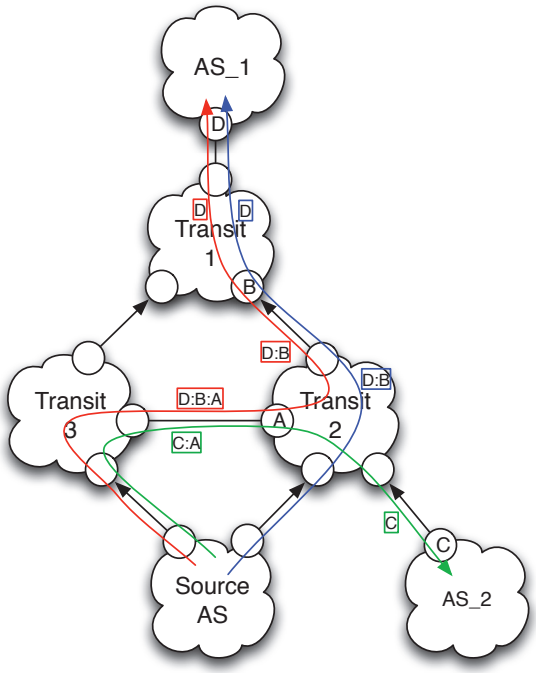


Figure 4.7: Valley Free violation with source identifier stacking.

In spite of the advantages that come along with the stacking solution, the underlying principle encounters a non-negligible issue. A user can potentially use paths which were not advertised to him by its neighbors, such that even non Valley Free paths can be utilized. An example of policy violation is shown is Figure 4.7. By knowing how to access *AS_2* and *AS_1*, the *Source AS* is able to use a non Valley Free path to reach *AS_1*.

Source AS receives policy compliant routes to reach *AS_2* (i.e., green path) and *AS_1* (i.e., blue path). Each path is identified by a series of labels, each one identifying the next transit AS. As *Source AS* is multi-homed, it receives labels, associated with paths, from each provider. Each label should only be used for traffic

sent to the provider providing the label. Nevertheless, nothing prevents *Source AS* to use the label provided by *Transit 2* for traffic sent through *Transit 3*. It is then able to send traffic to *AS_1*, using a non Valley Free path (i.e., red path). In order to prevent this type of violation, some sort of "deep label inspection"-filtering is necessary. This type of filtering would need to be performed at the entrance of each domain, which would of course be very costly (in terms of computational cost) and thus hardly applicable in reality. As this type of path identification is shown as non-policy compliant, we will not consider it in the following sections.

4.4 Worst case evaluation: Impact on the potential bottlenecks

The three previous global identification schemes may have an impact on both the Forwarding Information Base (FIB) of intermediary routers and the Forwarding Equivalent Class (FEC) of the router at the boundary between the IP world and the Path-ID world (i.e., point A in Figure 4.1). By adopting the perspective of different ASes, we evaluate in this section the number of paths a domain may receive if all the potential diversity is propagated (corresponding to the worst case scenario). Thanks to the AS relationships provided by CAIDA [29], we dumped all the potential diversity ASes may receive. We took into account the Valley Free conditions but did not consider the "Prefer Client" condition as we relaxed this condition in Chapter 3. It is important to note that disregarding the "Prefer Client" condition greatly increases the number of potential paths, which leads to a "worst case" evaluation.

4.4.1 Results

Figure 4.8 shows the evaluation of the number of Path-IDs that domains would need in order to be able to use all the potential paths. The evaluation does not take into account the "Source path-id stacking" as we showed it could be non-compliant with domain policies.

Impact on the forwarding table

The number of potential Path-IDs varies a lot from 10^5 to 10^9. The Stub-to-Stub scheme needs more Path-IDs than the Stub-to-transit, as previously assumed in Section 4.3.2. Nevertheless this difference is not so important as the ratio between both schemes is only about 9 or so, depending on the analysed AS.

The number of potential Path-IDs varies according to the CAIDA rank of the considered AS. A Tier 1 has less potential Path-IDs as its neighbors are clients or peers which only provide clients reachability which is compliant with Valley Free constraints. On the contrary, ASes that are low in the Internet hierarchy tend to have a lot more potential diversity/Path-IDs. This is due to the number of (direct or indirect) providers which provide **all** their potential paths/path-IDs.

In addition to the CAIDA rank, ASes having a lot of providers tend to have a lot of potential Path-IDs. For instance, the AS 6762, which has 11 providers (according

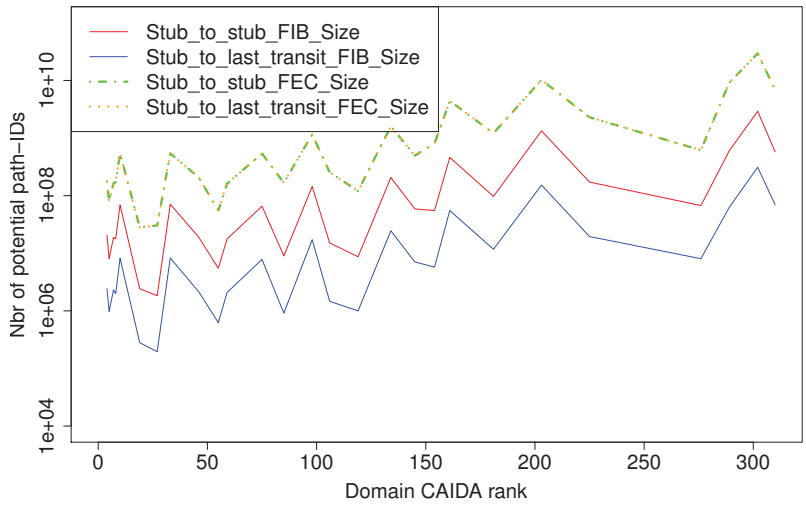


Figure 4.8: Number of necessary path-IDs according to the CAIDA ranking.

to CAIDA), and is ranked 33, has a potential number of Path-IDs of about $7 \cdot 10^7$. On the contrary, ASes 13237 and 4589, which have less providers (respectively 8 and 6) and have a higher ranking number (respectively 46 and 55) have numbers of potential Path-IDs of respectively $2 \cdot 10^7$ and $5 \cdot 10^6$.

Impact on the forwarding equivalent class

As for the FIB, the number of entries in the FEC increases a lot according to the relative position of the domain (i.e., high or low in the Internet hierarchy). The number of entries is the same for both identification schemes of sections 4.3.1 and 4.3.2. This number in correlated to the number of usable paths and not to the way we code these paths into Path-IDs. As we underlined before, the number of usable paths in both schemes is the same. Therefore, in our context Eq. 4.1 is equivalent to Eq. 4.3.

The number of entries varies from 10^7 to 10^{10} which is obviously not manageable.

4.4.2 Intermediate conclusion

From that evaluation, we can highlight two conclusions:

– Firstly, the number of entries in the routers (FIB and FEC) a domain may deal with is numerous. As expected, taking into account all the possible paths may not be possible and it seems necessary to filter the paths which are to be propagated to reduce the size of both the FIB and the FEC.

– Secondly, the scalability issue is more severe for ASes in lower rank. Indeed, we underline that small Tier-twos have more potential Path-IDs than Tier-ones.

4.5 What about filtering ?

As underlined in the previous section, propagating all the paths across the entire Internet and inserting them into the FIB and the FEC may lead to serious scalability issues. In this part, we propose to explore reasonable assumptions that could lead to a huge decrease of the entries that have to be inserted in both the FIB and the FEC.

4.5.1 FEC assumption

We assume that stub domains send traffic to a small subset of IP prefixes. Therefore, we propose to put the FEC as close as possible to the stub domains and configure only the entries that are used by clients (we assume that stub domains are clients of transit domains and name them “clients” without loss of generality). Mikians et al. [34] underline that only 61 000 destination prefixes are widely used – i.e., reached by a significant number of domains (i.e., more than 3 000 domains). These prefixes are the ones that have the highest probability to be reached via alternative routes. We can assume that the number of prefixes a stub domain reaches is bounded by this figure (i.e., 61 000). Therefore, it decreases reasonably from 450 000 to 61 000 the number or prefixes to be taken into account in the FEC.

Moreover, we can assume that the number of services an NSP delivers to its stub clients is limited and that it can hardly be higher than 10. The number of services that a domain wants to provide to its clients is limited to some quality services (delay, jitter, bandwidth) and resilience service (providing two disjoint paths). Therefore, providing alternative paths to reach significant prefixes (i.e., 61 000 or so) leads to have a FEC composed of about 610 000 entries which is of the same order of magnitude of the current routing table size.

It can be assumed that some domains may want to reach all the prefixes (e.g, CDNs such as Google, Akamai, etc.). Nevertheless, such domains are composed of several “regional” sites, close to their destinations. Each site reaches the destinations placed in its region and only takes into account the corresponding prefixes, therefore reducing the size of its FEC.

4.5.2 FIB assumption

Each alternative path can be associated with a service providers propose to their clients. The number of interesting paths (i.e., the ones which are supposed to be put into the FIB of backbones routers) drops if we take into account the services that ASes want to provide to their clients. We can reasonably assume that the quality of a path decreases with the length of the AS path. Therefore, we propose to filter the paths, to a destination d, a domain receives and select a path $p \in P_d$

(with P_d, the set of path to d) only if:

$$AS\_path\_length(p) \leq min_{x \in P_d}(AS\_path\_length(x)) + \delta$$

δ is the number of extra ASes a path can contain, compared to the shortest path. At the same time, alternative paths may require to have path with additional length, for the sake of resiliency (e.g., finding disjoint paths).

Thus there is a tradeoff to be found between quality and path resiliency and we perform next an evaluation of the value of δ which would allow for finding disjoint paths. We can develop more advanced path filtering algorithm (as in Chapter 5) that can be implemented in the architecture presented in Chapter 3. The present approach is just a "straight forward" pre-filtering, which could be performed before any advanced mechanism.

Setting the δ parameter

We believe that one of the first services that a domain may request from its provider is to obtain a path that is disjoint with the current BGP best route. Therefore we perform here an evaluation about the increase of path lengths associated with the adoption of a path that is disjoint with the current BGP best path. We then deduce a value of δ, which allows to obtain disjoint path for a high proportion of domains. In order to perform this evaluation, we extract the BGP routing tables of different domains from Route Views files [3]. Thanks to the AS relationship database provided by CAIDA [29], we then compute the disjoint paths from the BGP best route and select the shortest one. Our results are summarized in Table 4.1.

		Difference between 2 disjoint paths				
		-2 ASes	-1 AS	0 AS	1 AS	2 ASes
AS	rank	Percentage of destination ASes				
3257	5	2.1%	15%	59%	22%	1.9%
2914	7	2.0%	13%	60%	22%	1.7%
174	8	2.1%	15%	59%	21%	2.0%
3303	19	1.8%	13%	59%	24%	1.4%
13030	27	1.8%	13%	53%	29%	2.3%
6762	33	1.3%	10%	59%	27%	1.9%
4589	55	1.7%	13%	58%	25%	2.0%
4436	98	3.8%	22%	49%	23%	1.9%
8426	106	3.2%	19%	55%	21%	1.8%

Table 4.1: Impact of disjoint path on AS path length.

The table underlines that the use of a disjoint path has a small impact on the length of the path. For instance, AS 13030 can reach 53% of the destination ASes

3. The Route View files were extracted the same day as the CAIDA relationship database, i.e., on 16 January 2011

via a disjoint alternative path without any increase of the AS path length compared to the BGP best route. 29% of the destination ASes can be reached via a an alternative disjoint path one AS hop longer than the BGP best path, while some destination ASes (14.8%) can be reached via a disjoint alternative path that is shorter than the BGP best path. This last case happens when the BGP best route is not the shortest one, i.e., when the shortest path is advertised by a peer or a provider, whereas the route selected by BGP comes from a client.

From that Table 4.1, we can conclude that a great majority of the Internet can be reached via a disjoint path with an increase of the AS path of two or less ASes.

Therefore $\delta = 2$ is a good tradeoff for our evaluation and a route will be accepted if its AS path is not longer than the shortest one plus two.

Evaluation of FIB size

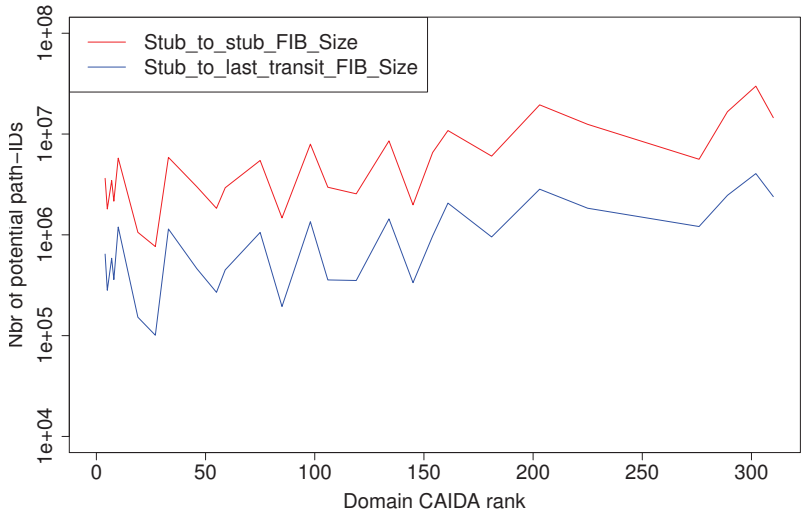


Figure 4.9: Number of necessary path-IDs according to the CAIDA ranking ($\delta = 2$).

Figure 4.9 presents the same evaluation that is presented in Section 4.4, but with $\delta = 2$. The evaluation of the size of the FEC is not performed as it has been addressed in the previous section (Section 4.5.1). The Figure 4.9 underlines that the number of entries in the FIB decreases roughly one or two orders of magnitude (depending on the AS). Using the Stub-to-Transit Path-ID scheme (described in Section 4.3.2) the maximum number of Path-IDs is of about $4 \cdot 10^6$, which is just one order of magnitude higher than what is currently handled by DFZ routers. It must be noticed that current FIBs in VPN routers of big providers have already reached this order of magnitude [115].

The domains having the most potential Path-IDs are the lowest domains (small Tier 2s). It can be opposed that these domains do not have infrastructures that are

adapted to such an amount of Path-IDs. Nevertheless, it must be noticed that such networks have small amounts of clients. They will therefore configure the Path-IDs that will be requested by clients, which should be low.

4.6 Conclusion

Today's Internet displays vast potential path diversity but BGP-4, as the single inter-domain routing protocol in the Internet, does not allow for the utilization of multiple paths per destination prefix. We proposed in Chapter 2 a way to enable the use of this diversity and we rely on the use of a path identifier, inserted in the packet. Nevertheless different manner of organizing these identifiers exist. As they can face scalability and policy issues, they could prevent the adoption of such proposals.

In this chapter, we proposed to explore some potential global path identification schemes. We have highlighted and quantified the worst case scalability issue such an organization may face. Furthermore we proposed some reasonable assumptions/filtering that would allow domains to bypass this issue without limiting the services they could provide to their clients.

We can conclude that choosing the correct Path-ID scheme (i.e., Stub-to-Transit), limiting the AS path (i.e., being shorter than the shortest one plus two) and associating Path-ID to destination ASes instead of prefixes, can make the FIB increase by only one order of magnitude compared to the current FIB, which is a manageable increase.

This analysis proposes an approach which is push based, meaning that an autonomous system propagates the routing information regardless of the willingness of its neighbors to use it. In the following chapter, we address the issue of negociating the routes which are to be propagated to the neighbors.

Chapter 5

Auction-type framework for selling inter-domain paths

5.1 Introduction

The Chapter 3 highlights an architecture, which allows for the propagation and the use of the Internet path diversity (cf. Fig. 1.4 for an illustration of the Internet path diversity). Nevertheless, enabling Internet path diversity faces huge scalability issues (i.e., insertion of additional routes into the Forwarding Information Base of routers) as the number of potential available paths can be very important. All this path diversity cannot be propagated from ASes to ASes as it would lead to the explosion of the Forwarding Information Base of routers. Therefore a selection, among the whole path diversity, must me performed in order to propagate a subset of this diversity. The previous chapter (i.e., Chapter 4) quantifies this diversity and underlines that it is possible, thanks to small adaptations, to make the number of paths decrease to the order of magnitude of 10^6, which seems manageable.

Nevertheless, the insertion of any additional route has an impact on the cost of the network. Network service providers will hardly implement these routes at their expenses and will therefore make their clients pay for this service. Nevertheless clients will pay if the NSP succeeds in proposing routes that are interesting, according to the clients' criteria.

Such an approach faces two important problems. First different paths have different characteristics (e.g., length, delay...) and each of these paths may be interesting for a neighbor but not for one another. The matching between paths and neighbors has never been studied as the current interdomain routing selects one path to be propagated to every neighboring domains. Second, network service providers aim at proposing path diversity as a value-added service, in particular to face the scalability issue mentioned above. Therefore the price establishment of a path is something very important which has not yet been studied. **We aim, in this chapter, at providing a simple framework that addresses these problems and fits the specific constraints of the inter-domain routing.**

Therefore we propose a route allocation framework, inspired by the auction theory, that aims at pricing routes and matching them to interested neighboring

domains. The allocation process outputs both route-to-neighbors matching and pricing and unifies good properties in order to motivate NSPs to adopt such an approach.

This chapter is organized as follows: In Section 5.2, we describe the context and the constraints of the inter-domain routing where the allocation process takes place. In Section 5.3 are presented the notations and the framework properties that are interesting in our context and the related work. Then the Section 5.4 presents the framework, including the matching between routes and neighbors. The Section 5.5 underlines a sufficient condition on the pricing function in order to form the grand coalition. Then we propose in Section 5.6 a simple pricing function which forms the grand coalition and is truthful dominant when each bidder submits a unique bid. We provide in Section 5.7 some evaluations of the framework. Last but not least, we conclude in Section 5.8 by summarizing the present chapter.

5.2 Background

5.2.1 Context

The architecture proposed in Chapter 3 allows for the propagation and the use of an important amount of paths. Internet routers currently have in their FIBs 450 000 entries [38] [1], which is already identified as a growing issue [63, 35]. Chapter 4 underlines that without any filtering, the amount of propagated paths could reach a non-manageable amount (i.e., order of magnitude of 10^{10}).

In order to address such an issue Kwong et al. [35] suggest the use of market mechanisms to limit the increase of the routing table size in a single path Internet. The proposed use of inter-domain multipath may worsen the already identified issue of the routing table size. In the current chapter we adopt the same perspective as [35] to avoid the explosion of the routing table of Internet routers.

In our context, each domain filters the routes according to the needs of his neighbors and makes neighbors pay according the routes they receive. Filtering according to the needs of each neighbor allows to leverage a new commercial approach of route propagation, as each neighboring domain may pay to receive its own specific route. Nevertheless a good knowledge about each neighbor is required in order to propagate the correct set of routes. This knowledge is very hard to obtain as it depends on the type of neighbors and reflects complex policies of neighboring domains, which may include commercially sensitive information. A matching process is therefore required to associate, for each neighbor, the best set of routes.

In the current chapter, a network service provider, which has already received several paths, aims at identifying the paths which deserve to be propagated to some of his **direct** neighbors (cf. Figure 5.1). The path allocation takes place between the NSP and his neighbors and no further domain is involved in the process.

Therefore the goal of the current chapter is to allow the NSP (the seller) to answer both following questions:

1. 450 000 routes or so in november 2012.

- Path to neighbor matching: to which neighbor do I send which route?
- Price computation: what price will each neighbor pay?

While the proposed framework differs from pure auctions (cf. Sections 5.2.2 and 5.3.5), its design is very close as both problems are well tackled by the auction theory respectively under the name of:

- Winner Determination Problem, which selects which bidder(s) win the item(s),
- Price Establishment, which compute the price(s) each winner must pay.

Therefore some comparisons between the proposed allocation framework and the auction theory are used all along the chapter to highlight some important aspects.

The propagation of paths to neighbors follows these five steps:

- Step 1: **coarse selection of path diversity**. There exists an important diversity and only a subset of it is interesting. We assume that the provider, which performs the path allocation, is capable to coarsely identify the interesting routes (e.g. filtering too long paths etc.), without being able to precisely match routes to neighbors yet.
- Step 2: **publication of paths**. The provider sends to his neighbors a set of path information (e.g., AS path, path characteristics, monitoring results [36]...) and the setting of the allocation process (e.g., type of mechanism, constraints...).
- Step 3: **Path bidding**. Each neighbor sends a set of bids, each one containing a set of routes and the corresponding price it is willing to pay for it.
- Step 4: **Path(s) matching and price(s) computation**. After receiving the bids, the provider is able to compute both the path to neighbor matching, which selects, for each neighbor, the set of paths he wins, and the prices neighbors must pay.
- Step 5: **Path configuration and payment**. Once matching and pricing are done, the provider configures the routing equipments such that each winner is able to use the paths he won, and propagates the full routing information to the winners. These route announcements and configurations can be seen as a grant to use the additional routes. In exchange to this grant, the neighbors pay the prices that have been computed by the mechanism.

Our contribution takes place at the steps 2, 3 and 4 where the provider put the allocation framework into use. It is important to note that the whole process takes place locally to the network service provider, which sells the routes (cf. Figure 5.1). Therefore this framework is incrementally adoptable and a domain can use this framework with some of his neighbors without waiting for other NSPs to participate to the mechanism.

The framework allows a carrier to offer additional routes to some selected customers in order to help them achieve their respective traffic engineering objectives. However, it is important to note that the end to end paths that are offered may pass through carriers that are not involved in such a mechanism. In other words, the carrier has usually no control on the route proposed (and their characteristics). Only a privilege to access these alternate routes is provided, with no quality of service (i.e., QoS) guarantees.

5.2.2 Inter-domain constraints

From an inter-domain point of view, the price of inter-domain peering bandwidth has drastically dropped during the last decade (a two order of magnitude decrease [37]). Therefore we assume that the network service provider that sells paths is able to forward the whole traffic demand of his clients, leading to the assumption that the NSP's capacities are over-dimensioned compared to the need of his neighbors.

Moreover, selling an inter-domain path is different from selling a conventional good in the sense that, after the transaction, the path does not belong to the buyer. Instead of really selling the path, the seller provides a grant to use the said path.

As it is currently the case, several buyers can be granted simultaneously to use a common path. Without any bandwidth constraint, all the neighbors could be granted to use the same path. Whereas the number of buyers is limited by the number of neighbors, we consider, for the rest of the chapter and without loss of generality, that an unlimited number of neighbors/bidders can win a single path.

Therefore inter-domain routes have the following properties:

- Neighbors may want to win several routes to reach the same destination. They must therefore be able to bid on bundle(s) of routes/goods.
- these routes can be **duplicated infinitely** by the seller in the sense that the seller can either propagate the routing information to only one neighbor or to all neighbors, which can be considered as a duplication of routes.
- if a neighbor wins a route, he can not sell it unchanged to one of his competitors. Indeed the characteristics of the route (e.g., delay, jitter...) change with each extra domain composing the path. For instance, a neighbor D wins the route containing the AS path A,B,C and wants to sell it to one of its neighbors. D is then able to sell the route containing the AS path A,B,C,D, which characteristics are not the same than the one with the AS path A,B,C.
- the inter-domain allocation process must generate low amount of messages. Therefore one round allocations should be preferred [2].
- a route is allocated to a neighbor at most once. Indeed, once a domain receives the routing information and is able to use the associated route, receiving again the same routing information does not give him any new grant or any additional capability.
- each neighbor is well identified. Indeed the seller and the potential buyers can each be identified by the shared physical inter-domain connections (e.g., BGP peering). Consequently there can not be any identity based manipulation. A neighbor can not steal another AS's identity, two neighbors can not merge in order to mutually bid and a neighbor can not bid under several identities.

2. The present work describes one single use of the process, which should be a one round process to minimize the communication overhead. Nevertheless, routes can be re-allocated on a regular basis (e.g., hourly, daily...) by using the same one round process. It could therefore elect different winners with different prices, if the interests of bidders changed since the previous allocation. The use of repeated allocation processes should be analysed from a repeated game perspective, which is outside the scope of the current work.

Routes are digital goods, as MP3s, and can therefore be duplicated infinitely (as previously underlined). Such goods are known in the literature either as digital goods or goods with unlimited supply (cf. related work in Section 5.3.5).

5.3 Framework required properties

5.3.1 Notations

We present in this section the notations, inspired from [116], for our work. A transit domain (i.e., the seller noted 0) proposes a set of routes G (G for goods) to all or part of his neighbors B (B for bidders). Without loss of generality, the framework should consider a set of paths G towards the same destination (destination AS or destination address prefix).

In the following part of the chapter and because of the analogy with the auction field, the NSP selling the routes is also named the "seller", routes are also named "goods" and neighbors are also named "bidders". Figure 5.1 identifies the seller, the bidders and the goods in the global environment.

For the sake of simplification (without loss of generality), we expose the case where the transit domain 0 proposes G to all his neighbors. Each neighbor $i \in B$ is a potential bidder and may therefore answer to the process in order to buy one or several route(s). Each bidder has a utility function $u_i(g)$, which associates each set of routes $g \subseteq G$ with a utility value. The utility function is private and represents the preference of bidder i for each set of routes g. Each bidder i provides a reported value $v_i(g)$, which represents the amount i is willing to pay in order to win the set g. The valuations are either public or at least known by the seller.

After receiving the bids from each bidder, the seller is able to elect the winners $W \subseteq B$ of the process. The seller is also capable to compute the price each winner must pay in order to obtain the set of routes he wins. Each bidder i is matched with a set of goods x_i. Bidders that do not win any route (i.e., $x_i = \emptyset$) are named losers L, with $L \cup W = B$. A bidder i who wins a set of routes $x_i = g \subseteq G$ pays a price $p_i(x_i = g) \leq v_i(x_i = g)$. His payoff is then $\pi_i = u_i(g) - p_i(g)$ whereas the payoff of the seller is the amount he wins - i.e., $\pi_0 = \sum_{i \in B} p_i(x_i)$ (with, $\forall g \subseteq G$, $u_0(g) = 0$).

5.3.2 Properties

The mechanism we design is intended to take place in both market and inter-domain routing contexts. Therefore we consider the following properties as being essential to the successful adoption of this framework:

– **Implementable**: With the current 450 000 IPv4 prefixes, inter-domain routing is already dealing with an important amount of information. The present allocation framework must occur in a context where each one of the said prefixes is associated with several routes (which is not the case with the current one route paradigm) and where several instances of the framework may run in parallel (e.g., one instance per prefix). Therefore the first requirement of the mechanism is to be scalable and implementable.

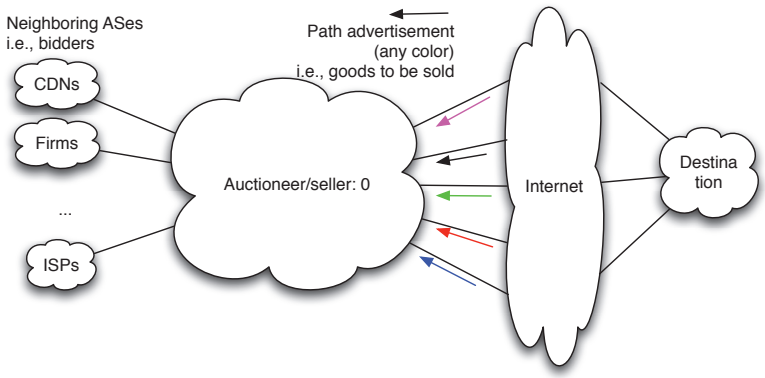


Figure 5.1: Goods and actors identification

- **Truthful**: A mechanism is truthful if bidding u_i is a dominant strategy for every bidder i (i.e., $v_i = u_i$). Such a property is interesting for the seller as it allows to assign the paths to those that really evaluates them more. Moreover it helps the seller to design future route allocations by learning the type of routes that are really interesting. Ultimately, it prevents bidders from manipulating bid values in order to influence the price they have to pay.
- **Form the grand coalition**: From a market perspective, it is very important that the seller has an incentive to accept every new bidder. Each external actor must be able to join the mechanism without its participation being refused by the seller. Moreover it is important that every potential bidder has an incentive to participate to the mechanism. From the point of view of cooperative game theory, if the framework accepts every new bidders and if every potential bidder is willing to participate, it forms the grand coalition, which includes the seller and every bidder.
- **Maximize the income of the seller**: We aim at providing to the seller a substantial income that will make him adopt this mechanism.
- **Deal with unknown valuation functions**: Pricing of alternate routes is a new possibility of the future multi-path Internet. Therefore a seller does not know yet how much each route could be priced. It is therefore impossible to fix prices and let buyers come if prices are interesting for them. Furthermore, we do not know yet the type of utility functions NSPs have for such a new service. We must therefore be agnostic about this information. Different approaches, which are cited in Section 5.3.5 may be used on a long run - i.e., when NSPs have sufficient knowledge about their clients needs. Nevertheless the unknown valuation function case is more adapted in this emerging market.

The three first properties (i.e., implementable, truthful and forming the grand coalition) may prevent the mechanism from closely approaching the maximum revenue of the seller. Nevertheless we consider these three properties essential to

the adoption of the framework. Therefore we aim at maximizing the revenue of the seller under the constraint that the three first requirements are respected.

The presented requirements are tackled by the auction theory (cf. related work in Section 5.3.5). This is the reason we chose to set up an auction-like allocation framework that we adapt in order to be compatible with the interdomain constraints presented in Section 5.2.2.

5.3.3 On the difficulty to both being truthful and form the grand coalition

In order to illustrate the fact that joining these two properties is difficult, we give as an example the case of auctions. Vickrey-Clarke-Groves (VCG) auctions have been well studied [117, 118, 119, 120] and are known to be truthful (i.e., bidders' dominant strategy is to tell the truth). Nevertheless Ausubel et al [121] underline that VCG auctions suffer from not enforcing the formation of the grand coalition. Indeed the seller may find an incentive to exclude some bidders in order to increase his revenue. In order to solve this issue, some works [116, 122, 123] propose the core-selecting auction, which ensures that the auction process leads to the grand coalition. Nevertheless such a type of auction is not truthful. It leads therefore to the necessity of finding a tradeoff between truthfulness and forming the grand coalition [122, 124]. Such works on conventional auctions (limited supply) do not apply in our context (unlimited supply). Nevertheless, it illustrates well the difficulties to both have the truthful property and form the grand coalition.

5.3.4 On the difficulty to both being truthful and maximize the seller's revenue

An ISP can expect to obtain a substantial revenue by selling his routes. The revenue the seller can obtain is bounded by the total amount bidders are able to pay - i.e., $max(revenue) = \sum_{i \in B} u_i$. It is to be noticed that this maximum revenue is hardly known by the seller. Indeed, the seller knows the bid values (i.e., v_i) which can be different from the utility values (i.e., u_i) if bidders lie (i.e., $v_i \neq u_i$).

As we underlined in section 5.2.2, routes are infinitely duplicable. Therefore the seller can potentially allocate a set of routes to every bidder without route depletion.

By aiming to obtain the maximum revenue, a common reflex would be to sell the routes the amount bidders are able to pay (i.e., a pay as you bid mechanism: $p_i = v_i$). Nevertheless, in our context, such a mechanism has two consequences. The first is that each bidder is elected as a winner and wins the good(s) he wants. Indeed, for each new bidder that wins a good, the seller increases his revenue. The maximum revenue is thus reached if all the bidders are elected as winners. Therefore competition between bidders totally disappears. The second consequence is the lack of truth, which encourage bidders to bid lower values. Indeed each bidder knows, before bidding, that he will win the set of goods he wants regardless of his bid(s). All bidders will thus decrease the price he will pay by reporting a valuation of zero, which will lead to a null seller's revenue, which is far from maximizing his revenue.

5.3.5 Related work

Pricing of items is a large and old field that has been well studied to sell physical goods. The emergence of computer science extended the field to the pricing of unlimited supply goods. Some works have been proposed to maximize the revenue of the seller in an unlimited supply context [125, 126, 127, 128, 129, 130]. While the computation of the maximum revenue price vector is NP-hard [125], contributions to the field generally propose price computation algorithms that approximate the maximum revenue of the seller. To the best of our knowledge, no contribution unify the properties presented in Section 5.3.2. [131] deals with the selling of one item while we deal with several items. Some other works do not ensure that bidders disclose their true utilities [125, 126, 130]. While [132] proposes a way to transform any non-truthful item pricing mechanism into a truthful mechanism, it degrades the efficiency of the revenue approximation and overall requires a minimum amount of buyers, which is not a reasonable assumption in our context. Some other works [128, 133] take into account historical records (e.g., Bayesian prior) or already know how much each buyer is willing to pay [127, 134, 129], which is also not the case in our context.

From a pure auction perspective, our framework gives the opportunity to sell sets of routes to neighbors, which could be compared to a combinatorial auction [3]. Combinatorial auctions have been well studied in the literature [135, 116]. Although they seem to fit our goal, we apply them in a context where **items to be sold can be duplicated** by the auctioneer (i.e., unlimited supply) and allocated simultaneously to different neighbors/bidders, which is not the case of these works.

Auctioning goods with unlimited supply is not common and some works have been done to address this problem. Goldberg et al. studied in [136] the selling of a single digital good and extended their work to multiple goods in [137]. Nevertheless, their approaches only take into account a single won good per winner and does not deal with combinatorial auctions. Other works also deal with digital auctions, but all focus on multiple identical items [138, 139, 140, 141]. We also aim at setting up our framework to make the bidders give their true valuations. A lot of work has been done in auction theory to try to set up truthful auctions which form the grand coalition [121, 122, 119, 116, 124, 118, 123]. Nevertheless all these works focus on a finite number of goods to be sold.

Other works [142, 143, 144] aim at providing frameworks that associate inter-domain routing with auctions or pricing. From a networking point of view, they propose to sell bandwidth along paths that everybody can use and fix transit prices according to pricing mechanisms, whereas we aim at selling the right to use specific paths. From a game theory point of view, they address these issues by using VCG approaches, which is not adapted in our infinite supply context [145].

From a pure networking point of view, an architecture has previously been proposed to sell inter-domain routes [146] and advocates for the use of auctions in this context, without providing any auction framework. While this architecture proposal is well adapted to federation of domains, it is highly centralized and thus not

3. Combinatorial auctions: auctions where bidders can bid simultaneously on several goods.

adapted to the Internet (distributed by nature). It requires a strong cooperation between carriers while we adopt an incremental perspective, where each NSP can adopt the framework on its own [4].

5.4 Framework

5.4.1 Computation process

Conventional matchings between goods and bidders (such as in auctions) are based on the maximization of the social welfare (i.e., $X(v) = argmax_x \sum_i v_i(x)$ [5] [119]). Such maximization selects in general a restricted number of winners as the number of available goods is lower than the number of goods bidders want to buy (i.e., goods are over-demanded). Nevertheless, in our context, goods/routes are infinitely duplicable by the seller and the maximization of the social welfare is equivalent to elect every bidders to be winners, which could lead to a seller's revenue of zero.

It must be noticed that accepting a new winner (or increase the set of winning bids) may face two contradictory effects:

- the new winner pays, which makes the payoff of the seller π_0 (i.e., his revenue) increase;
- depending on how prices are computed, the prices paid by winners may decrease, which makes the payoff of the seller π_0 decrease (as illustrated in Figure 5.2).

For instance, in multiple identical item auctions with limited supply, the VCG [6] price may reduce if you add another item. In such a case, adding items to make every bidder win makes the price and the seller's revenue drop to zero [119]. Therefore the revenue of the seller is null whether there is no winner (i.e., $|W| = 0$) or every bidder is elected to win (i.e., $W = B$). There must exist at least a set of winners W where $\pi_0 > 0$ (unless the selling process does not make sense). Therefore instead of selecting the winners by maximizing the social welfare, **we select the set of winners that maximizes the revenue of the seller**.

The revenue of the seller depends on both the set of winning bids and the prices winners pay. We adopt an iterative approach to elect the winners and the prices. We iteratively increase the set of winning bids in order to explore the bid space and compute, at each iteration, both the corresponding prices and the revenue of the seller. We can then deduce the seller's maximum revenue point and the associated winning bids and winners. This generic process is described in the Algorithm 2.

4. A NSP currently receives a path diversity thanks to its external BGP peerings (cf. Chapter 2). Therefore he can already sell this diversity to his clients without any cooperation with other NSPs.

5. It is interesting to note that, in order to be consistent, the maximization of the social welfare should take into account the real utility values $u_i(x)$ instead of the reported values $v_i(x)$. Nevertheless, such a maximization is generally performed by the seller or the auctioneer who only know the reported values and therefore base this maximization on these values.

6. Whereas VCG pricing computation is not applicable in our context, it is a good way to illustrate that increasing the set of winners may reduce prices.

Algorithm 2 Revenue curve computation

$Revenue\_max = 0$;
$Winners\_max = 0$;
ForAll $Winning\_bids \subseteq bids$ **do**
$Revenue = compute\_revenue(Winning\_bids)$
if $Revenue > Revenue\_max$ **then**
$Revenue\_max = Revenue$
$Winners\_max = Winning\_bids$
end if
EndFor

There is an important difference between the conventional auction approach [119] and the presented approach. The conventional approach selects first the winners (known as the Winner Determination Problem - i.e., WDP - which generally maximizes the sum of the valuations of winners) and uses the results as an input for computing the prices. In our approach, both winner selection and price computation are performed at the same step - i.e., during the revenue maximization process.

This algorithm computes the price for different sets of winning bids. A first approach for computing the maximum revenue of the seller is to compute, for all the combinations of winners, the prices and seller's revenue and keep the combination of winners providing the maximum revenue. Nevertheless this approach suffers from the huge number of potential winner sets. Therefore we use a ranking approach to add bids to the set of winning bids. A way to rank bidders must be found in order to easily identify, at each computation step of the algorithm, the bid which must be elected as extra winning bid of the next iteration.

5.4.2 How to rank bids fairly?

Some ranking have already been proposed in [147, 145], nevertheless their characteristics are proved in a context where a limited number of goods is available, which does not apply here.

In order to set up a fair winner determination process, a bid can be elected as a winning bid only if all the bidders that proposed more for the same set of routes also win.

Let b be a bidder and $v_b(g)$ its bid for the set of routes g (b may also submit other bids). Here are the requirements of a ranking. $v_b(g)$ can be elected as a winning bid:

- requirement 1: if all the higher ranked bids have already been elected as winning bids;
- requirement 2: and if no group of lower rank bids can propose more than $v_b(g)$ for the same set g.

The first requirement may be considered as a fair condition and the second one as a collusion-proof condition. It must be noted that a bidder may have several winning bids and/or several losing bids (i.e., for different sets of routes).

It is easy to rank bids for the same set of goods. Nevertheless, it is more

difficult to rank bids associated to different sets of goods as the sizes of the sets may be different and as the sets may not contain the same routes. In such a case, bid values can not be considered alone and set sizes must also be taken into account. Therefore we rank bids according to the mean valuation of the bids - i.e., $\overline{v_b(g)} = \frac{v_b(g)}{|g|}$. $\overline{v_b(g)}$ can be interpreted as the virtual value of each good/route contained in the bid.

Definition 5. *A **mean-bid winner determination** is a determination of the winners such that if a bidder i wins thanks to its bid v_i, each bidder j that has, at least, one higher mean-bid (i.e., $\overline{v_j} > \overline{v_i}$) also wins.*

By definition, the mean-bid winner determination is compatible with the first requirement of a bid ranking.

The Figure 5.2 presents what could be the output of Algorithm 2 on a single duplicable item. In this graphic example, 40 bidders bid in the range $[0, 1]$ and prices are computed according to a price function[7]. The curve traces the revenue that the seller could earn by fixing the mean-bid-value separating winners from losers. The one providing the highest revenue is the point $\mathfrak{M}$. Each bid which mean value is lower than $\mathfrak{M}$ is a losing bid. If a bidder has only losing bids, he does not win anything. **In the case he has several bids higher than $\mathfrak{M}$, he wins the set of goods which has the highest valuation** - i.e., $max_g\{v_i(g)|\overline{v_i(g)} > \mathfrak{M}\}$.

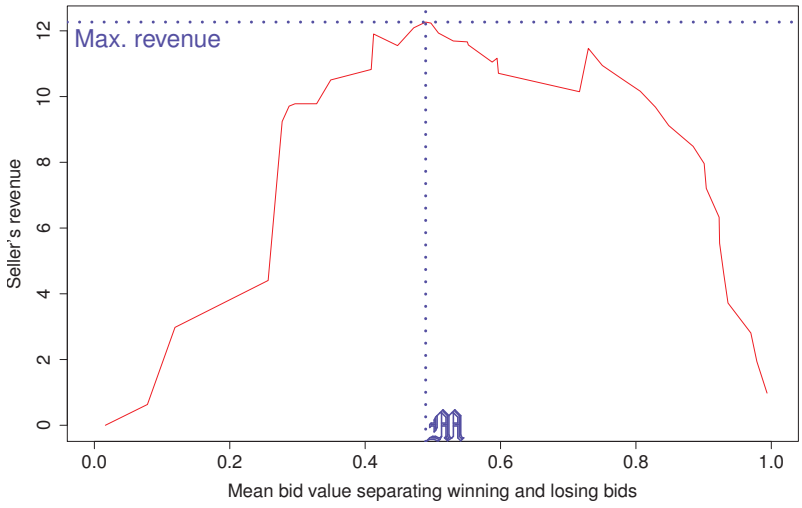


Figure 5.2: Example of the evolution of the seller's revenue.

Theorem 3. *When goods are infinitely duplicable by the seller, a mean-bid winner determination is loser collusion proof.*

7. Here, the first-loser-mean-bid payment function (detailed in Section 5.6) is used but other payment functions may be used.

Proof Theorem 3 advances that no losing bids or coalition of losing bids can propose to the seller more than what propose winners, for the same set of routes/goods. We rank all the bids by their mean values. For any winner bid v_1, we perform a reductio ad absurdum by imagining that a set of lower rank bids (i.e., $\{v_2, .., v_n\}$ with $\forall i \in \{2, .., n\}$, $\overline{v_i} < \overline{v_1}$) can propose a better common bid $v_{\{2,..,n\}}$ for exactly the same set of goods $g = g_1 = \cup_{i\in\{2,..,n\}} g_i$ (with g_i the set that bidder i bids on). We have to note that $\forall i, j \in \{2, .., n\}$, $g_i \cap g_j = \emptyset$ as a good is only present once in g and that goods are not duplicable by bidders (cf. Section. 5.2.2). If the coalition is successful, its valuation should be higher than the one of bidder 1 - i.e., $v_{\{2,..,n\}} = \sum_{i\in\{2,..,n\}} v_i > v_1$ and therefore $\overline{v_{\{2,..,n\}}} = \frac{\sum_{i\in\{2,..,n\}} v_i}{|g|} > \overline{v_1} = \frac{v_1}{|g|}$. Nevertheless we have the following, which yields a contradiction with the previous relation, and the proof follows:

$$\begin{aligned}
\overline{v_{\{2,..,n\}}} &= \frac{\sum_{i\in\{2,..,n\}} v_i}{|g|} \\
&= \frac{\sum_{i\in\{2,..,n\}} \frac{v_i \times |g_i|}{|g_i|}}{|g|} \\
&< \frac{v_1}{|g_1|} \times \frac{\sum_{i\in\{2,..,n\}} |g_i|}{|g|} \\
&= \frac{v_1}{|g_1|} = \overline{v_1}
\end{aligned}$$

□

The theorem 3 proves that the mean-bid winner determination is compatible with the second requirement of a bid ranking.

5.4.3 Consequences on bids

Performing a mean-bid winner determination has some consequences on bids. In order to illustrate such consequences, we give the following allocation example:

Table 5.1 presents an example of bids, which are then ranked by mean value in Table 5.2 and processed to compute both the seller's revenue[8] and the good allocation in Table 5.3. Some of the bids present in Table 5.1 may be considered as not rational. Nevertheless, this irrationality is only used in this section in order to underline the bidding rules. We assume in all the other sections that the players are rational. Having ranked the different bids by mean allows the seller to scan all the possible winner combinations by scanning it from high to low mean values (cf. Table 5.2). At each step, the revenue of the seller is computed (cf. revenue computation described in Section 5.5) in order to select the maximum revenue good allocation.

As a first step, the seller states that only bidders who bid more that 7, as mean value, are elected as winner (i.e., first colomn in Table 5.3). Therefore bidder (1) wins b.

8. This example illustrates the necessity to follow bidding rules, provided at the end of this section. Prices and revenue are here useless and are therefore not computed.

	(1)	(2)	(3)
a	6	4	5
b	7	4	5
ab	9	11	4

Table 5.1: Bids

$\Rightarrow$

Mean	Bidder	Goods
7	(1)	b
6	(1)	a
5.5	(2)	ab
5	(3)	$a \oplus b$
4.5	(1)	ab
4	(2)	$a \oplus b$
2	(3)	ab

Table 5.2: Bids ranked by mean values

$\Downarrow$

	Minimum winning mean bid						
	7	6	5.5	5	4.5	4	2
(1)	b	b	b	b	ab	ab	ab
(2)	$\emptyset$	$\emptyset$	ab	ab	ab	ab	ab
(3)	$\emptyset$	$\emptyset$	$\emptyset$	$a \oplus b$	$a \oplus b$	$a \oplus b$	$a \oplus b$
revenue	-	-	-	-	-	-	-

Table 5.3: Winner determination and good matching for every value of minimum winning mean bid

As a second step, the seller elects the more-than-6-mean-value bidders as winners and compute his potential revenue. In that case, even if two bids are considered, only bidder (1) wins as the bids went from him. We can see that the second bid (i.e., $(1) : a : 6$) is not taken into account. This is due to the fact that bidder (1) has won both routes a and b but is willing to pay only for one, which is the highest bid - i.e., $(1) : b : 7$. If a bidder bids for several sets that contain the same number of goods, only the highest one is to be taken into account by the mechanism. Therefore a bidder bids the same valuation for all the sets that have the same number of goods.

Arriving at the step where the minimum winning mean bid is 4, we can see that bidder (2) is able to win either $[(2) : ab : 11]$ or $[(2) : a \oplus b : 4]$[9]. The seller will naturally make (2) win ab as its valuation is higher than $a \oplus b$. A smaller set with a smaller mean value is never considered by the seller. Therefore a bidder that bids for two sets of different sizes must always give a higher mean-value to the smaller set, unless it will never be taken into account.

At last, arriving at the last step (i.e., minimum mean bid is 2), bidder (3) wins either $[(3) : a \oplus b : 5]$ or $[(3) : ab : 4]$. The seller chooses to allocate $a \oplus b$ to (3) as it maximizes the valuation of the winner. Therefore a bidder that bids for two sets of different size must always give a higher bid to the bigger set.

9. $\oplus$ for the conventional XOR operator.

Here are the conditions on the bids that a bidder must respect in order to propose bids that will be considered by the seller:

For all sets of goods A and B that bidder i bids:
– if $|A| = |B|$, then $\frac{v_i(B)}{|B|} = \overline{v_i(B)} = \overline{v_i(A)} = \frac{v_i(A)}{|A|}$
– if $|A| > |B|$, then $v_i(A) > v_i(B)$ and $\overline{v_i(B)} > \overline{v_i(A)}$

It must be noticed that the previous presented rules are not imposed by the seller. Bids that are not compliant with these rules are not taken into account by the auction framework. Therefore the bidders adopt these rules naturally in order to have their bids well considered by the seller.

Moreover, an important consequence of such rules is that if the seller proposes n goods/routes, each bidder proposes only n bids (each one providing one bidding value but potentially several route sets with the same size).

If bidders bid compliantly with the previously underlined bidding rules, a winning bid may correspond to several good sets (e.g., $[(3) : a \oplus b : 5]$ in the previous exemple). If bidder (3) wins the good $a \oplus b$, a choice must be made to allocate either a **or** b to (3) (but not a **and** b). The seller may ask that the bidder makes this choice before bidding therefore giving this last bidding rule, which is optional:

For all sets of goods A and B that bidder i bids, if $|A| = |B|$, then i bids exclusively for A or for B.

Contrary to the previously presented rule, the present rule is enforced by the seller and is not a consequence of the framework.

Even if this rule can be enforced, the seller may not adopt it and may accept that bidders bid several sets of the same size and the same valuation. In such a case, the seller will have to choose which set (e.g., $a \oplus b$ [10]) the winner obtains. This choice can be made according to specific objective functions of the seller (e.g., allocation of routes in such a way that they are propagated to an almost equal number of neighbors) which are outside of the scope of this work.

5.5 Forming the grand coalition

In this section, we highlight the complexity of using the core concept of cooperative game theory on the whole framework in order to form the grand coalition (i.e., Section 5.5.1). Then we shift to another perspective (i.e., Section 5.5.2) which permits a simpler formulation and ends by highlighting a sufficient condition which makes the mechanism form the grand coalition.

5.5.1 Core concept and auctions

As previously underlined in Section 5.4, we aim at designing an auction-like mechanism that leads to the formation of the grand coalition. The Core is a notion of cooperative game theory [118], which leads to the formation of the grand coalition.

10. Here the sets are singletons. Other negotiations may output several non singletons as $\{a, b\} \oplus \{e, f\}$.

Definition 6. *(**coalition game**) A coalition game with transferable utility is a pair* (N, w)*, where*

- N *is a finite set of players, indexed by* $i = \{0, ..., |N| - 1\}$*.*
- $w : 2^N \rightarrow \mathbb{R}$ *associates with each coalition* $S \subseteq N$ *a real-valued payoff* $w(S)$ *that the coalition's members can distribute among themselves. We assume that* $w(\emptyset) = 0$*.* w *is also called the characteristic function.*

Definition 7. *(**core**) A payoff vector* $\pi = \{\pi_0, ..., \pi_{N-1}\}$ *is in the core of a coalitional game* (N, w) *if and only if*

$$\forall S \subseteq N, \sum_{i \in S} \pi_i \geq w(S) \tag{5.1}$$

$$\text{with} \sum_{i \in N} \pi_i = w(N) \tag{5.2}$$

The Definition 7 underlines that, if a payoff vector π is in the core, a sub-group of players can not earn more by forming a sub-coalition.

From an auction perspective, let X be the set of possible allocations of goods and player 0 be the seller. A bidder $i \in N \backslash 0$, bidding v_i and paying a price p_i, has a payoff $\pi_i = v_i - p_i$ and the seller has the payoff $\pi_0 = \sum_{i \in N \backslash 0} p_i$. An auction is a Core Selecting Auction if it selects a payoff vector $\pi = \{\pi_0, ..., \pi_{N-1}\}$ which is compliant with the inequalities 5.1 and 5.2.

The value $w(S)$ of a coalition S is the maximum total value the players can share, according to the output of the Winner Determination Problem restricted to the coalition S. If the reader wants to read more about the core applied to limited supply auctions, we encourage him to read the following works [121, 116, 148, 123, 124, 122].

In order to compute a core allocation, the auctioneer must generate an important number of price constraints. Finding a core allocation or knowing if an allocation is in the core can be NP-Hard. According to the important number of bidders who could participate to the negotiation (i.e., order of magnitude of 100) and the quantity of goods to be sold (i.e., order of magnitude of 10), this approach is intense in term of computation requirements.

Day et al [149] propose an efficient way to compute core allocations. Their evaluations show that the mean computation time of a core allocation for a 1 000 bid auction is of approximately 600 seconds. Such an approach has been studied in the context of spectrum auctions, where each frequency is allocated during an important period of time (i.e., 10 to 25 years [150]). In the inter-domain context, where each computation may result in a short period allocation [11] and where several allocations may take place in parallel (i.e., for different destination prefixes), such an important computation time is not reasonable. In order to reduce the complexity of the approach, we propose another approach, in the next section, in order to form the grand coalition.

11. As already underlined, the auction may be repeated periodically by the seller.

5.5.2 Incentive to form the grand coalition

Simple formalisation

As we have seen, the computation of core prices of the whole game is combinatorial. In this section, we change our perspective. Instead of enforcing the grand coalition of the game by addressing the core computation at the level of the whole game, including the seller and all the bidders, we divide the game into smaller games.

First we have to highlight that this mechanism is a centralized one and that the seller has a special role. As underlined in Section 5.2.2, a bidder winning a good can not share it or sell it to another bidder. Furthermore, bidders are well identified and bidder merging or multiple bidding by a single bidder can not be performed. Last but not least, the winner determination is performed only according to the bids of the bidders and not on the availability of a good - i.e., a good won by a bidder does not prevent another bidder to win the same good (i.e., one copy). From a cooperative game theory perspective, these constraints result in the fact that the utility can only be transferred between the seller and the bidders (and not among bidders). Therefore we divide the whole game into smaller games (i.e., $A = \{A_1, ..., A_N\}$), each one with two players, the seller and one bidder $\{0; i\}$. The whole utility transfer between the players is therefore divided into bilateral utility transfers, each one between the seller and one bidder.

The grand coalition can be obtained if every bidder wins at participating to the allocation process and if the seller wins at accepting every bidder - i.e., being into the core of all the small games A_i. Each game A_i is defined as:

- two players - i.e., the seller 0 and one bidder i.
- a minimum mean-bid value $\mathfrak{M}$ from which the seller can identify the set of goods x_i that i wins - i.e., $x_i = argmax_g\{v_i(g)|\overline{v_i(g)} > \mathfrak{M}\}$. If $x_i = \emptyset$, bidder i is a loser.
- a price function $\mathfrak{F}(i)$, which provides the price p_i for the set of goods x_i that i wins.

The value of $\mathfrak{M}_i$ and the output of the price function $\mathfrak{F}(i)$ may depend on external information - i.e., the bids of other bidders (i.e., $-i$) - and may also depend on the bids of bidder i. We respectively denote as $\mathfrak{M}_S$ and $\mathfrak{F}_S$ the value of $\mathfrak{M}$ and the price function $\mathfrak{F}$ associated with the coalition $S \cup 0$ with $S \subseteq N \backslash 0$.

In order to lie in the core of this simple game, we rely on the definition given by the Eq. 5.1. The characteristic function of such a game is the total amount of revenue players can share, i.e., the sum of their utilities:

$$w(\{0 \cup i\}) = \pi_0^{\{0 \cup i\}} + \pi_i^{\{0 \cup i\}} = u_0^{S \backslash i}(S) + u_i(x_i) \tag{5.3}$$

where $u_i(x_i)$ is the utility value of bidder i for the set of goods x_i and $u_0^{S \backslash i}(S)$ is the amount of utility the seller brings into the game A_i. It is important to note that $u_0^{S \backslash i}(S)$ is the amount of utility won by the seller thanks to **other bidders** (i.e., $S \backslash i$), already taking into account the influence of bidder i on the values of $\mathfrak{M}_S$ and $\mathfrak{F}_S$.

At the end of the process, the seller and the bidder i have the following payoff:

$$\begin{aligned}\pi_0^{\{0\cup i\}} &= u_0^{S\setminus i}(S) + p_i = u_0^S(S)\\ \pi_i^{\{0\cup i\}} &= u_i(x_i) - p_i\end{aligned}$$

whereas their payoffs outside this small game A_i are:

$$\begin{aligned}\pi_0^{\{0\}} &= u_0^{S\setminus i}(S\setminus i)\\ \pi_i^{\{i\}} &= 0\end{aligned}$$

In order to be in the core of this small game, we must have, according to Eq. 5.1, the two following constraints which can easily been understood as the two following properties:

–

Definition 8. ***bidder incentive****: the bidder has an incentive to join the coalition, i.e., his payoff must not decrease by joining the seller:*

$$\pi_i^{\{0\cup i\}} \geq \pi_i^{\{i\}} = 0 \tag{5.4}$$

Therefore

$$p_i \leq u_i(x_i) \tag{5.5}$$

–

Definition 9. ***seller incentive****: the seller has an incentive to accept a new bidder, i.e., his payoff must not decrease with the addition of a new bidder:*

$$\pi_0^{\{0\cup i\}} \geq \pi_0^{\{0\}} \tag{5.6}$$

Therefore

$$u_0^{S\setminus i}(S) + p_i \geq u_0^{S\setminus i}(S\setminus i) \tag{5.7}$$

$$p_i \geq u_0^{S\setminus i}(S\setminus i) - u_0^{S\setminus i}(S) \tag{5.8}$$

The inequality 5.8 is very interesting. The right hand side represents the variation of the seller's payoff when it accepts the new bidder i, without yet taking into account what i pays. This inequality underlines that the price i must pay is higher than the negative impact i causes to the seller.

In our context, if the mechanism is compliant with the bidder incentive and seller incentive properties, it forms the grand coalition.

A sufficient condition to form the grand coalition

Let $\mathfrak{M}_{\mathfrak{F},N}$ be the mean-value maximizing the seller's revenue and be also the value separating winning bids from losing bids for the mechanism with bidders $N\setminus 0$ and the payment function $\mathfrak{F}$.

Let $\wp$ be the feasible payment space (i.e., $\wp = [0; UBPS]$) and $Bids$ be the set of bids. The **Upper Bound Payment Space** (**UBPS**) vector is defined as the set of maximum prices winners can pay - i.e., $p_{i\in B} \leq max(v_i|\overline{v_i} > \mathfrak{M}_{\mathfrak{F},N})$. The UBPS is derived from the core constraint given by the inequality 5.5.

Definition 10. *A payment function $\mathfrak{F} : \mathbb{R} \times Bids \rightarrow \wp$ computes, at each iteration of the process 2, the prices of the goods according to the minimum mean-bid-value $m \in \mathbb{R}$ and the set of bids $Bids$.*

Theorem 4. *In the present allocation framework (duplicable goods, mean bid-winner determination based on maximization of the seller's revenue), if bidders bid truthfully (i.e., $\forall g \in G$ and $\forall i \in B$, $u_i(g) = v_i(g)$) every non decreasing payment function $\mathfrak{F} : \mathbb{R} \times Bids \rightarrow \wp$ satisfies both bidder incentive and seller incentive properties.*

Proof Let r (revenue), W (winners) and L (losers) be the output of an allocation, with N the set of bidders, which prices are computed thanks to a non-decreasing function $\mathfrak{F}$. Let $\mathfrak{M}_{\mathfrak{F},N}$ be the mean-bid value maximizing the revenue of the seller (i.e., $r = \pi_0^{\{N\}}$), according to $\mathfrak{F}$. Winners are allocated the set of goods which both maximizes their valuations and have a mean-bid-value higher than $\mathfrak{M}_{\mathfrak{F},N}$ (i.e., $x_i = argmax_g\{v_i(g)|\overline{v_i(g)} > \mathfrak{M}_{\mathfrak{F},N}\}$).

It is interesting to notice that the Upper Bound Payment Space (**UBPS**) is the set of maximum prices winners can pay according to their bids. Nevertheless if bidders lie (e.g., if $v < u$), the **UBPS** can be lower than what they can really pay. In this proof we assume that bidders bid truthfully. Therefore the **UBPS** is the set of maximum prices winners can pay according to their utilities - i.e., $p_{i \in B} \leq max(v_i|\overline{v_i} > \mathfrak{M}_{\mathfrak{F},N}) = max(u_i|\overline{u_i} > \mathfrak{M}_{\mathfrak{F},N})$.

First if the price of a set of goods g (i.e., $p(g)$) is outside of $\wp = [0; UBPS]$, either the seller excludes the bidders who win g because their payments are negative (i.e., if $p(g) < 0$) or at least one winning bidder leaves the process because his payment is higher than what he can pay (i.e., if $p(g) > UBPS$). Therefore $p(g) \in \wp$ is a required condition to have all bidders participating to the process. Moreover, as $\wp = [0; UBPS]$, the payoff of each bidder can not be negative then **satisfying the bidder incentive property** for each bidder.

Second we prove that, if $\mathfrak{F} : \mathbb{R} \times Bids \rightarrow \wp$ is non decreasing in $Bids$ (either in values or in number of bids), the seller is not willing to exclude any bidder. Taking into account the allocation output previously described, we propose to the seller to include a new bidder i in the mechanism. Will the seller accept this new bidder ? (i.e., will the revenue of the seller increase ?). In order to answer to this question, we base our reasoning on the inequality 5.8, which can be transformed in $u_0^{\{N \cup i\}}(\{N\}) + p_i = u_0^{\{N \cup i\}}(\{N \cup i\}) \geq u_0^{\{N\}}(\{N\})$.
Therefore

$$\pi_0^{\{N \cup i\}} \geq \pi_0^{\{N\}} \tag{5.9}$$

which only underlines that the seller only accepts a new bidder i if his revenue increases. The impact of i on the revenue of the seller may differ according to the bids of i:

- either i is a winner (i.e., $\exists g \subseteq G$ such that $\overline{v_i(g)} \geq \mathfrak{M}_{\mathfrak{F},N}$). The seller allocates the set of goods g to the new winner (goods allocated to other winners do not change at point $\mathfrak{M}_{\mathfrak{F},N}$ in the seller revenue curve illustrated in Figure 5.2).

Two consequences arise: (1) the set of winning goods increases by the addition of g (all the others remaining identical), which makes the seller's revenue increase. (2) the set of bids (i.e., $Bids$) increases, by the addition of the new bidder i's bids. Thanks to the non-decreasing nature of the payment function $\mathfrak{F}$, the prices increase which makes also the seller's revenue increase.

– or i is a Loser (i.e., $\forall g \subseteq G$, $\overline{v_i(g)} < \mathfrak{M}_{\mathfrak{F},N}$). Therefore all his bids are losing bids, which makes the number of bids increase and therefore induces an increase of both the prices and the revenue of the seller - thanks to the non-decreasing nature of $\mathfrak{F}$.

We proved that the revenue, at point $\mathfrak{M}_{\mathfrak{F},N}$ in the revenue curve illustrated in Figure 5.2, increases by the addition of a new bidder. Nevertheless, the maximum revenue point may not be $\mathfrak{M}_{\mathfrak{F},N}$ anymore (i.e., $\mathfrak{M}_{\mathfrak{F},N\cup i} \neq \mathfrak{M}_{\mathfrak{F},N}$). Even if this point change may happen, the new maximum seller's revenue is higher than the one computed at $\mathfrak{M}_{\mathfrak{F},N}$, which **satisfies the seller incentive property**.

As the properties 8 and 9 are satisfied, the mechanism forms the grand coalition.

□

It is interesting to note that the pay-as-you bid (i.e., $p_i(g) = v_i(g)$) and constant price (i.e., $p_i(g) = k$) payment functions are non-decreasing functions of the bids and gives to the seller and the bidders an incentive to form the grand coalition.

5.6 An example of payment function

We analyse in this section the characteristics of a simple payment function, the first-loser-mean-bid payment. Despite the apparent simplicity of this payment function, it leads to the formation of the grand coalition and gives incentive to tell the truth.

5.6.1 Payment function presentation

Definition 11. *A **first-loser-mean-bid payment** is a payment function where the winners pay each of their won goods the highest mean-valuation of the losing bidders:*

$$\begin{aligned} \textit{for each } g \subseteq G,\ p(g) &= |g| \times \mathfrak{M} \\ \textit{and } \mathfrak{M} &= max\{\overline{v_i} | i \in L\} \end{aligned}$$

In such payment function, every good/route is paid the same price and does not take into account the identity of routes. The second equality of the definition 11 ensures the fairness of the price. It allows to avoid situations where the price is lower than a losing bid proposed by a winner (i.e., a bid comprised between $\mathfrak{M}$ and $max\{\overline{v_i} | i \in L\}$).

Theorem 5. *The **first-loser-mean-bid payment** is loser collusion proof.*

Proof The Theorem 5 states that no collusion of losers can propose more than a winner (or a set of winner) for the same set of routes. As the price to be paid (i.e., $\mathfrak{M}$) is higher that any mean-bid value of losing bidders, the proof is the same as the one of the Theorem 3.

□

5.6.2 Bidding concatenation

Let us imagine the current allocation process with 2 routes a and b and 4 bidders. Figure 5.3 shows the bids of the different bidders. The bids are ordered according to their mean values. We can see that, despite the bidder must specify which routes are interesting for him, this information is not required to process the matching and to compute the price. First by taking into account the bidding rules induced by the mean-bid winner determination (cf. Section 5.4), each bidder who bids for several sets of the same size will bid them via a common bidding value - i.e., submiting one single bid for several sets. Second, the first-loser-mean-bid payment function does not take into account the identities of routes to compute the price.

Consequently, in order to compute, for each bidder, how many routes he wins and what price he is about to pay, the seller only needs to know the price each bidder is willing to pay for each set size - i.e., regardless of the route identities.

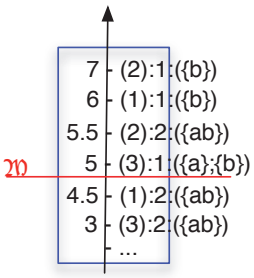


Figure 5.3: Required bidding information.

In Figure 5.3, the required information to compute the price is located in the blue square. For instance, if we focus on bidder (3), instead of bidding two bids for having one of the single routes (i.e., bidding $(3) : 1 : (a) \oplus (3) : 1 : (b)$), he concatenates the bids to express only one - i.e., $(3) : 1 : (\{a\}; \{b\})$, which must be considered as bidder (3) proposing to pay 5 for having either $\{a\}$ or $\{b\}$ (not both).

In combinatorial auctions, bidders can express $2^{|G|}$ different bids. In the present framework, bidders can only provide $|G|$ bids. The maximum number of bids to be analysed by the seller drops therefore from $|N| \times 2^{|G|}$ to $|N| \times |G|$.

Here is an overall example of the allocation process with the 4 bidders (i.e., bidders 1, 2, 3 and 4) competing for the 2 goods (i.e., a and b). The table 5.4 presents an example of bids, which are then ranked according to their mean values

in Table 5.5 and processed to compute both the seller's revenue and the good allocation in Table 5.6.

	(1)	(2)	(3)	(4)
a	6	-	5	2
b	-	7	5	-
ab	9	11	6	3

Table 5.4: Bids

$\Rightarrow$

Mean	Bidder	Goods
7	(2)	b
6	(1)	a
5.5	(2)	$a+b$
5	(3)	$a \oplus b$
4.5	(1)	$a+b$
3	(3)	$a+b$
2	(4)	a
1.5	(4)	$a+b$

Table 5.5: Bids ranked according to the mean

$\Downarrow$

	Minimum winning mean bid							
	7	6	5.5	5	4.5	3	2	1.5
(1)	$\emptyset$	a	a	a	ab	ab	ab	ab
(2)	b	b	ab	ab	ab	ab	ab	ab
(3)	$\emptyset$	$\emptyset$	$\emptyset$	$a \oplus b$	$a \oplus b$	ab	ab	ab
(4)	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	a	ab
Route price	6	5	5	2	2	2	0	0
Revenue	6	10	15	8	10	12	0	0

Table 5.6: Revenue computation

Having ranked the different bids according to the mean allows the seller to scan all the possible winner combinations by scanning it from high to low mean values (cf. Table 5.5). At each step, the revenue of the seller is computed (here, the first-loser-mean-bid payment, described in Section 5.6) in order to select the maximum revenue good allocation.

As a first step, the seller states that only bidders who bid more that 7, as mean bid value, are elected as winners (cf. first colomn in table 5.6). Therefore bidder (2) wins b and the price of each won route is 6 (i.e., the mean bid of the first loser), which gives to the seller a revenue of 6. As a second step, the seller elects the more-than-6-mean-value bidders as winners. The price which is to be paid is 5 per route as only bidders 3 and 4 remain losers and the maximum mean-bid value is 5. The revenue is therefore 10. Every step is computed and the seller selects the one maximizing its revenue (i.e., step 3, in red, with a revenue of 15).

5.6.3 Telling the truth

We prove show here an example of lying, which allows for a bidder to increase his payoff. Then we prove that if bidders submit only one bid, the allocation process

and the first-loser-mean bid payment is truthful. It must be noted that the seller can easily enforce bidders to bid only one value. Moreover one bid can cover several sets of routes. Indeed, as previously underlined in Section 5.4.3, one single bid value can be accompanied by several route sets. Nevertheless these sets must have the same size.

Example of lying

The decoupling between bid values of winners and prices they pay could be used to make bidders tell the truth. Indeed, with such a pricing scheme, winners have no influence on the final prices and can not increase their payoffs by trying to make the prices reduce.

In the current framework, such a decoupling does not ensure truthfulness. In order to highlight such an assertion, we rely on an example. Let us focus on the multi-bid version of the framework (i.e., bidders can submit several bids) as in the following example:

	(1)	(2)	(3)	(4)
a	8	-	7	1
b	8	5	7	1
ab	12	7	9	-

Table 5.7: Bids

$\Rightarrow$

Mean	Bidder	Goods
8	(1)	$a \oplus b$
7	(3)	$a \oplus b$
6	(1)	$a + b$
5	(2)	b
4.5	(3)	$a + b$
3.5	(2)	$a + b$
1	(4)	$a \oplus b$

Table 5.8: Bids ranked according to the mean

$\Downarrow$

	Minimum winning mean bid						
	8	7	6	5	4.5	3.5	1
(1)	$a \oplus b$	$a \oplus b$	ab	ab	ab	ab	ab
(2)	$\emptyset$	$\emptyset$	$\emptyset$	b	b	ab	ab
(3)	$\emptyset$	$a \oplus b$	$a \oplus b$	$a \oplus b$	ab	ab	ab
(4)	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$a \oplus b$
Route price	7	5	5	1	1	1	0
Revenue	7	10	15	4	5	6	0

Table 5.9: Revenue computation

In the example presented in Tables 5.7, 5.8 and 5.9, all bidders tell the truth, the maximum revenue point is reached for a price per route of 5 and the seller's revenue is 15. Please notice that the prices are computed thanks to the first-loser-mean-bid payment, described in Section 5.6, which decouples prices from the bids of winners.

Let us focus on bidder 1. He wins two routes for a whole price of 10 and his payoff is $\pi_1 = u_1(ab) - p(ab) = 12 - 10 = 2$. Bidder 1 may lie in order to increase his payoff. Instead of bidding 12 for both routes a and b, he bids 9, which remains compatible with the bidding rules explained in Section 5.4.3.

Tables 5.10, 5.11 and 5.12 present the result of such a lie.

	(1)	(2)	(3)	(4)
a	8	-	7	1
b	8	5	7	1
ab	**9**	7	9	-

Table 5.10: Bids

$\Rightarrow$

Mean	Bidder	Goods
8	(1)	$a \oplus b$
7	(3)	$a \oplus b$
5	(2)	b
4.5	(3)&(1)	$a + b$
3.5	(2)	$a + b$
1	(4)	$a \oplus b$

Table 5.11: Bids ranked according to the mean

$\Downarrow$

	Minimum winning mean bid					
	8	7	5	4.5	3.5	1
(1)	$a \oplus b$	$a \oplus b$	$a \oplus b$	ab	ab	ab
(2)	$\emptyset$	$\emptyset$	b	b	ab	ab
(3)	$\emptyset$	$a \oplus b$	$a \oplus b$	ab	ab	ab
(4)	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$a \oplus b$
Route price	7	5	1	1	1	0
Revenue	7	10	4	5	6	0

Table 5.12: Revenue computation

Instead of winning both routes, bidder 1 wins a unique route and pays a price 5. His payoff is now $\pi_1 = u_1(a \oplus b) - p(a \oplus b) = 8 - 5 = 3$.

Therefore bidder 1 is able to increase his payoff by modifying his bids and the decoupling of prices from bid values of winners is not a sufficient condition to ensure truthfulness in the present framework.

The one bid case

Here we study the case where bidders submit only one combinatorial bid and prove that telling the truth is a dominant strategy with the first-loser-mean-bid payment function.

Theorem 6. *In the present allocation framework (duplicable goods, mean bid-winner determination based on the maximization of the seller's revenue and first-loser-mean-bid payment), if each bidder bids only one bid, truth telling is a dominant strategy.*

Proof Here bidders submit only one bid (potentially containing several goods). A winner may lie in order to reduce the price he is about to pay and a loser may lie in order to become winner.

First, by modifying their bids, winning bidders are not capable to directly modify the price they pay as the price is computed thanks to bids of losing bidders.

Nevertheless a loser can make his losing bid become a winning bid by increasing the associated bid value. More formally, a bidder i lies and bids a valuation $v_i(g) > u_i(g)$ such that he is able to win the set of goods $g \subseteq G$ that he would have lost by telling the truth (i.e., when $v_i(g) = u_i(g)$). What price would bidder i pay? It is obvious that i is not willing to pay the price $p(g)^T > u_i(g)$ (i.e., where $p(g)^T$ is the value of $p(g)$ when i tells the truth). Nevertheless $p(g)$ could be reduced by the lie of i and take an interesting value $p(g)^L < u_i(g) < p(g)^T$ (i.e., L when i lies). Therefore the question is: can $p(g)$ decrease and take a value lower than $u_i(g)$ by increasing $v_i(g)$.

Figure 5.4, provided for illustration purpose, shows a zoom around the maximum revenue point and illustrates the elevation of the revenue curve because of the lie of a bidder. The black arrow shows the lie of bidder i (from a mean value $\overline{v_i(g)^T} = 0.41$ to a mean value $\overline{v_i(g)^L} = 0.52$ - cf. point 0 in Figure 5.4). Without the lie of the bidder, the revenue curve is represented by the red curve whereas the new revenue values, influenced by the lie, are represented by the blue diamonds.This bid modification makes some potential revenue points increase (cf. blue arrows that lead to points 1, 2, 3, 4 and 5), whereas some other points are not modified.

In a general case, we analyse here the increase of the revenue curve led by the lie. The following applies either if $\mathfrak{M}^T \in (u_i(g), v_i(g)]$ or not.

The lie (i.e., $v_i(g) > u_i(g)$) makes the seller's revenue curve change (as illustrated in Figure 5.4). It is important to notice that the modification of the seller's revenue curve only occurs in the value interval $(u_i(g), v_i(g)]$. Indeed $\forall m \notin (u_i(g), v_i(g)]$, the number of won goods is not changed and the revenue remains the same.

Then $\forall m \in (u_i(g), v_i(g)]$, the value of the revenue is increased by the price paid by i (i.e., $m \times |g|$). Therefore either the maximum revenue point remains the same (i.e., $\mathfrak{M}^T = \mathfrak{M}^L$) or $\exists m \in (u_i(g), v_i(g)]$ such that $r(m) \geq r(\mathfrak{M}^T)$. In both cases the new maximum revenue point $\mathfrak{M}^L \in (u_i(g), v_i(g)]$ provides a price that is higher than $u_i(g)$ leading to a negative payoff for i.

As a conclusion, truth telling gives to bidders their maximum payoff and is therefore a dominant strategy.

□

5.6.4 Formation of the grand coalition

Theorem 7. *In the present allocation framework (duplicable goods, mean bid-winner determination based on the maximization of the seller's revenue and first-loser-mean-bid payment), if each bidder bids only one bid, the allocation process leads to the grand coalition.*

Proof We first have to highlight that Theorem 6 proves that, if each bidder submits one bid, truthfullness is a dominant strategy. Then we rely on the theorem 4

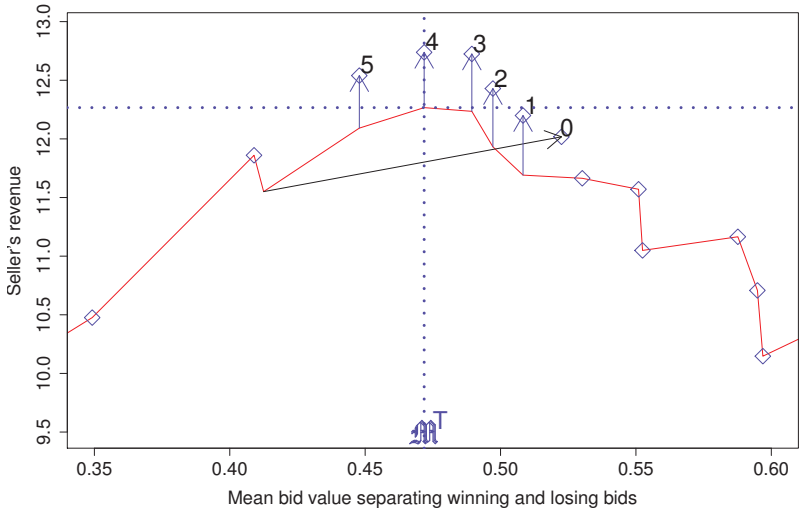


Figure 5.4: Revenue curve deformation.

and prove that a first-loser-mean-bid payment is a $Bids$ non decreasing function. We focus on **any** step of the revenue curve computation described in the algorithm 2, where the mean value separating potential winners from potential losers is m (potentially not the one giving the maximum revenue to the seller). At this point, the seller computes the price according to its payment function $\mathfrak{F}$. All the routes are priced the same amount $p(1) = m = max\{\overline{v_i}|i \in L\}$.

Is $\mathfrak{F}$ increasing by the addition of a new bidder ? When a new bidder $\{i\}$ is added to the negociation with the N players, it increases the set of bids $Bids$ with his own bids:

- his bids which mean values are higher than m have no impact on the price computed at point m.
- his bids which mean value are lower than the price $p(1)$ (i.e., the price computed before $\{i\}$ joins the process) have also no impact on the price computed at point m.

A first-loser-mean-bid payment is a $Bids$ non-decreasing function and therefore leads to the grand coalition when bidders bid one single bid each.

□

5.7 Evaluation

5.7.1 Computation complexity

The algorithm presented in Section 5.4 (i.e., Algorithm 2) is here modified in order to illustrate the impact of the first-loser-mean-bid payment function on the

overall process.

We have to underline that, with the first-loser-mean-bid payment, the price and the value separating the losers from the winners are both equal to the highest bid of the losers. Therefore, by taking place into a step of the Algorithm 2 and having already computed the revenue at this point m, the next revenue point which is to be computed is not anymore the one adding a new bid into the winning bid set but rather the one adding a new winner to the set of winners. Therefore the Algorithm 2 is changed to reflect that change:

Algorithm 3 Revenue curve computation

$Nbr\_winners = 0$;
$Revenue\_max = 0$;
$Nbr\_winner\_max = 0$;
$Rank\_by\_mean(Bids)$;
for $Nbr\_winners = 1 : Nbr\_bidders$ **do**
  $Revenue = compute\_revenue(Bids, Nbr\_winners)$
  if $Revenue > Revenue\_max$ **then**
    $Revenue\_max = Revenue$
    $Nbr\_winner\_max = Nbr\_winners$
  end if
end for

Therefore, once the ranking of bids is done, the number of revenue computation drops from $|Bids| = |B| \times |G|$ (i.e., the number of bids) to $|B|$ (i.e., the number of bidders). Each revenue computation is linear in the number of bids (i.e., $|Bids| = |B| \times |G|$) as it only looks at all the bids to select the ones that are elected as winning bids and compute the revenue with the price given by the first-loser-mean-bid payment function. The ranking of the bids does not change. It is an $O(Bids \times log(Bids))$ [151] operation. The overall process complexity, taking into account both bid ranking and revenue computation, is therefore in $O(|B|^2)$ and in $O(|G|log(|G|))$.

We implemented the allocation process and generated bids according to the bidding rules described in Section 5.4. The order of magnitude of possible bidders is in general of 100. Nevertheless some important ASes have a number of neighbors of the order of magnitude of 1 000 (e.g., Level 3 has about 3 700 neighbors [29]). We could assume that one allocation is “local” - i.e., meaning that the bidders are somewhat in a constrained perimeter (e.g., in Europe) leading to the decrease of this figure. Nevertheless we evaluate the allocation framework for numbers of bidders till 3 000 to reflect the worst cases.

We also constraint the number of routes to be of the order of magnitude of 10. Taking into account that we currently have only one single route per prefix with BGP, propagating 10 routes for a single prefix is, for a mid-term perspective, already very interesting. The following results have been computed on a single core 2.83GHz computer.

The Figure 5.5 shows the computation time of the allocation process according

to the number of players who participate. We did the same evaluation for different number of goods.

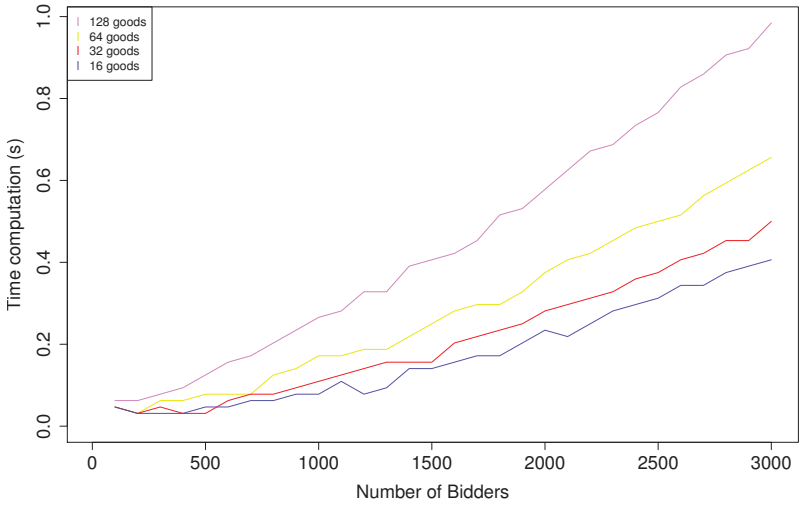


Figure 5.5: Time computation evolution regarding the Number of bidders.

The Figure 5.6 shows the same information but with the square-root of the time computation values. It shows that the allocation process is less than linear regarding the square of the number of bidders. It reinforces the square complexity of the whole allocation process, regarding the number of bidders, underlined previously in the current section.

The Figure 5.7 shows the computation time according to the number of goods, which are to be sold during the allocation. We also did the same evaluation for different number of bidders.

It is important to note that the computation of the allocation process does not last more than one second, even for the worst case (i.e.,128 routes and 3 000 bidders). Regarding the most probable cases (i.e., 10 routes and 100 bidders), computation power is not a limitation for performing such a mechanism.

5.7.2 Relation between maximum revenue point and number of goods

The way we generated the bids takes into account the bidding rules detailed in Section 5.4. For each bidder i, we use an induction approach to generate his bids. It must be noted that a given bidder can submit several bids, each one of them with a different number of goods:

– the one item bids of all the bidders are generated according to a half normal distribution with $\sigma^2 = 100$. The Figure 5.8 illustrates the distribution of the one item bids in an allocation with 1 000 bidders. Using this distribution simulates the fact that a large number of clients may be interested by an alternative

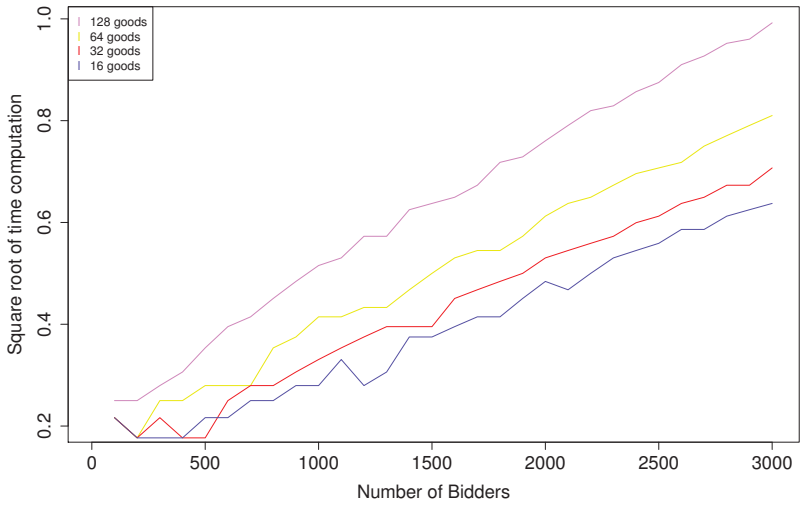


Figure 5.6: Time computation evolution (square roots) regarding the Number of bidders.

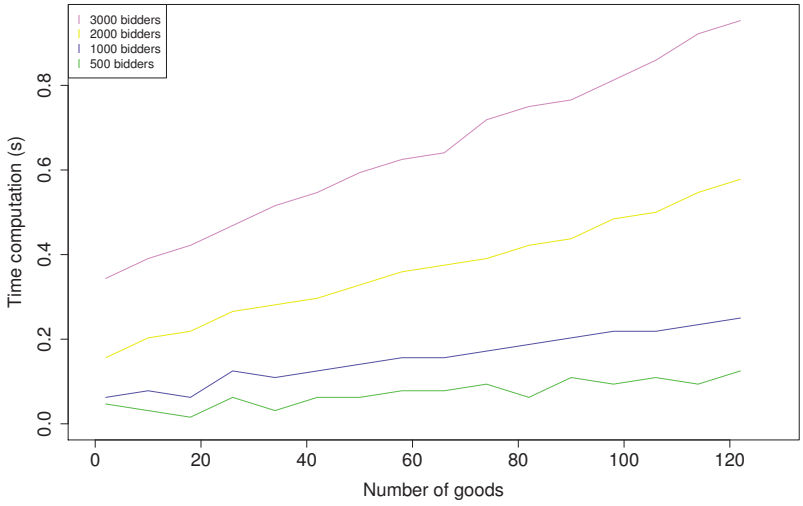


Figure 5.7: Time computation evolution regarding the Number of routes.

route but are not willing to pay a lot for it. On the contrary, a small amount of clients may be very interested in getting it and are therefore willing to pay a large price.

– then each x-route bid is generated randomly in the interval $[bid_i(x-1), bid_i(x-$

$1) \times \frac{x}{x-1}]$, with $bid_i(x-1)$ the bid i submits for $x-1$ goods. This interval is used in order to be compliant with the bidding rules detailed in Section 5.4.

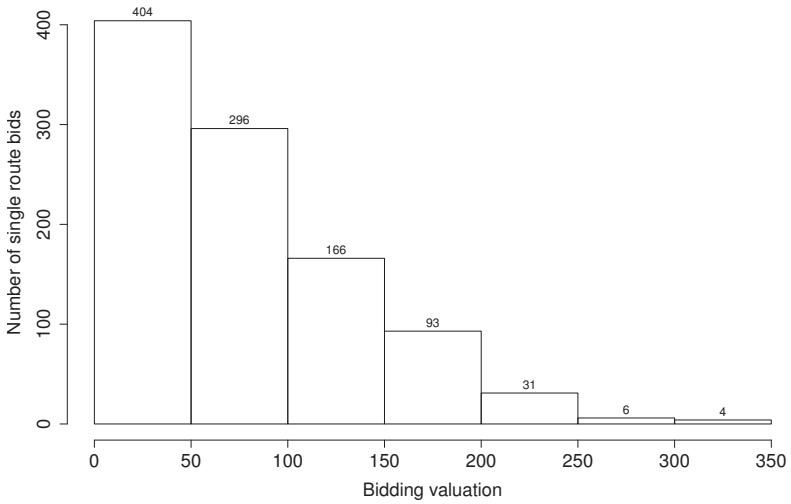


Figure 5.8: Half normal distribution of the single route bids.

The number of routes, which are to be proposed in the selling, have an impact on the output. Indeed, in some allocation examples, almost all the bidders can be elected as winners, whereas in other only a few proportion of them win.

The figure 5.9 shows such a phenomena. The process proposing only one route elects only half of the bidders as winners of the route - i.e., the revenue maximization point elects 477 winners over the 1 000 bidders. The increase in the number of goods makes the number of winners increase, till being close to electing every bidder. For instance, the process proposing 16 goods elects 728 winners over the 1 000 bidders.

Nevertheless, all the winners do not win the same amount of routes. The Figure 5.10 shows the number of winners according to the number of routes that are won for the process proposing 16 routes. It highlights that the winners win different numbers of routes. 237 winners win the whole set of routes (i.e, 16 routes) whereas 7 win 15 routes etc... and 165 winners win a single route each.

A lot of winners obtain all the routes and a lot of winners obtain only one route. This phenomena is a consequence of the bidding rules as once a bidder obtains several routes, he is likely to obtain easily the whole set of routes. For instance, if a bidder bids for 8 routes with a mean-bid-value of 50, the minimum mean-bid-value for a set of 9 routes is $50 \times \frac{8}{9} = 44.4$, which is very close. On the contrary, a bidder that bids for one route with a mean-bid-value of 50 can propose a bid for two routes down to $50 \times \frac{1}{2} = 25$. The highest the number of routes, the lower the maximum gap between two consecutive bids.

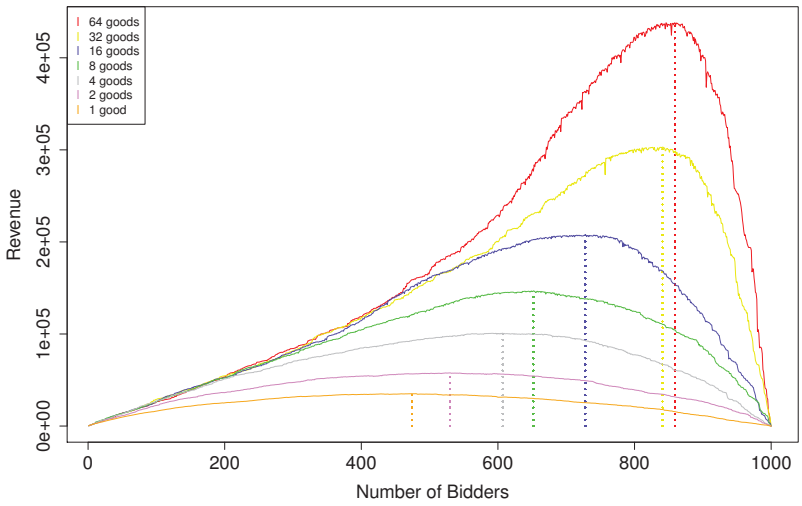


Figure 5.9: Maximum revenue point migration.

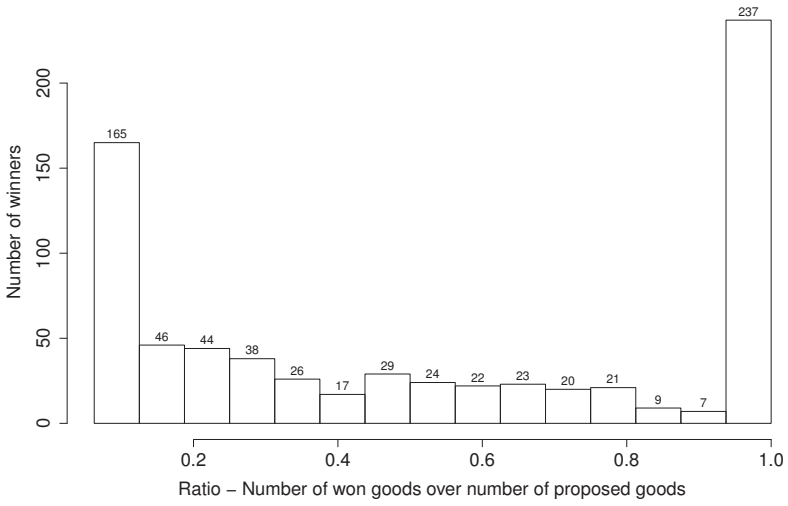


Figure 5.10: Ratio between the number of won goods over number of proposed goods (16 goods).

Even if a high proportion of winners obtain the whole package of routes, the payoff (i.e., $\pi_i = v_i(g) - p_i(g)$) they earn is far from being equivalent from one winner to the others. Indeed the Figure 5.11 shows the payoff distribution among

the winners for the same process as Figure 5.9. We can see that a high proportion of these winners (i.e., 701) get a small payoff (i.e., less than 200) whereas a few of them obtain a high payoff.

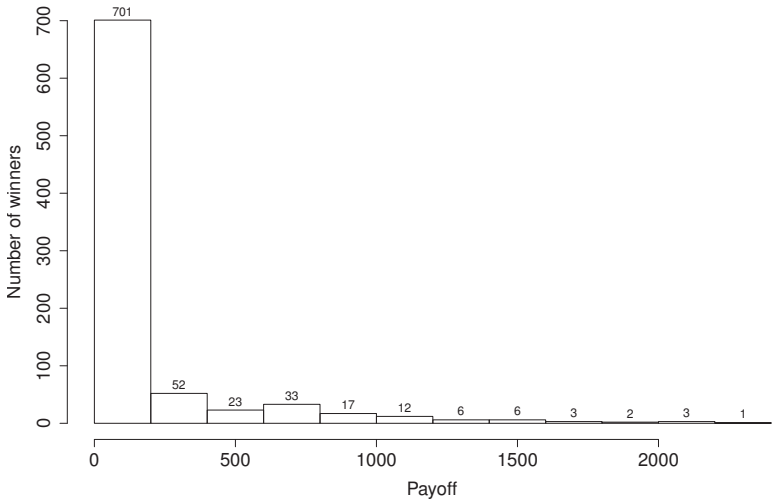


Figure 5.11: Distribution of the winners over the payoff.

5.7.3 Ratio between revenue and valuations

The goal of the current mechanism is, for the seller, to sell goods to the "correct" bidders and gain a substantial revenue. In order to be successful, the mechanism must provide a seller's revenue that is close to the potential payment of the winners.

In this section we evaluation the ration of the revenue of the seller over the bids. Two comparisons can be performed:

- One may evaluate the ratio between the revenue of the seller and the maximum price winners are willing to pay to obtain the goods they win (i.e., $\sum_{i \in W} v_i(x_i)$). This comparison makes sense as it evaluates, according to the whole set of won goods, if the goods have been sold to a price close to the amount they are valuated **by the winners**.
- The other way is to compare the revenue of the seller to the total amount of money bidders are able to pay (i.e., $\sum_{i \in B} max_{g \subseteq G} v_i(g)$). This comparison also makes sense as it underlines the amount of revenue the mechanism is able to extract and give to the seller according to the total amount **all the bidders** are able to pay.

Ratio between revenue and the sum of winners' valuation

The Figure 5.12 shows the ratio between the revenue of the seller and the valuations that winners could have paid to win their routes. We compute this ratio for different number of goods/routes and for different numbers of bidders.

We find that the seller receives as a revenue between 55% and 80% of the prices that winners could have paid (i.e., their valuations). The revenue that is not perceived by the seller is the difference between the valuations of bidders and the prices they pay - i.e., the winners' payoffs.

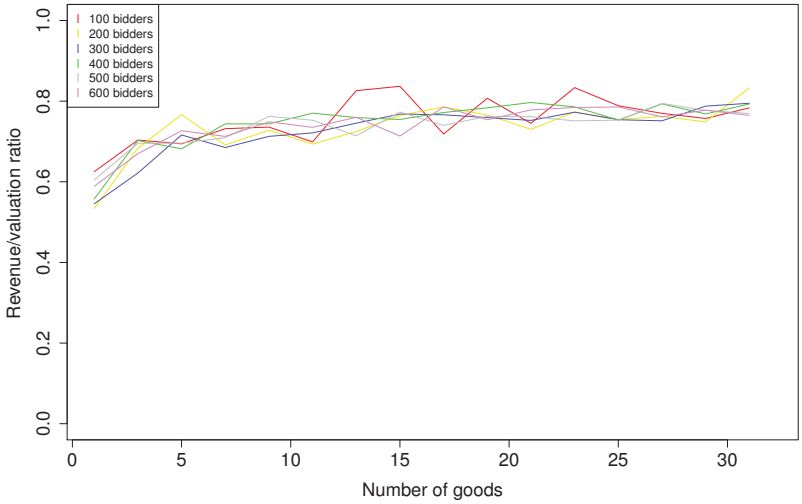


Figure 5.12: Revenue over valuation (**of winners only**) ratio.

The Figure 5.13 shows the same evaluation (i.e., revenue over valuation) for allocations which accept only one bid per bidder. This evaluation is important as we proved in Section 5.6 that truth telling is a dominant strategy for this type of mechanism. Therefore, whereas the seller is not sure that the valuations provided by the bidders in the evaluation of Figure 5.12 are true, the ones provided for the allocations simulated for the Figure 5.13 can be assumed to be equal to the real utility values.

It is interesting to notice that the ratio between revenue and valuation/utility is almost the same (i.e., approximately 60%) for different numbers of bidders and for different numbers of routes. It is also interesting to note that this revenue over valuation values are coherent with the ones of the one good processes of Figure 5.12 as the one good allocation, which leads to a one bid per bidder mechanism, provides a ration between revenue and valuation of around 60%.

The evaluations of Figure 5.13 give a less interesting ratio than Figure 5.12. It can be considered as the price the seller has to pay in order to enforce truthfulness - i.e., the price of truthfulness.

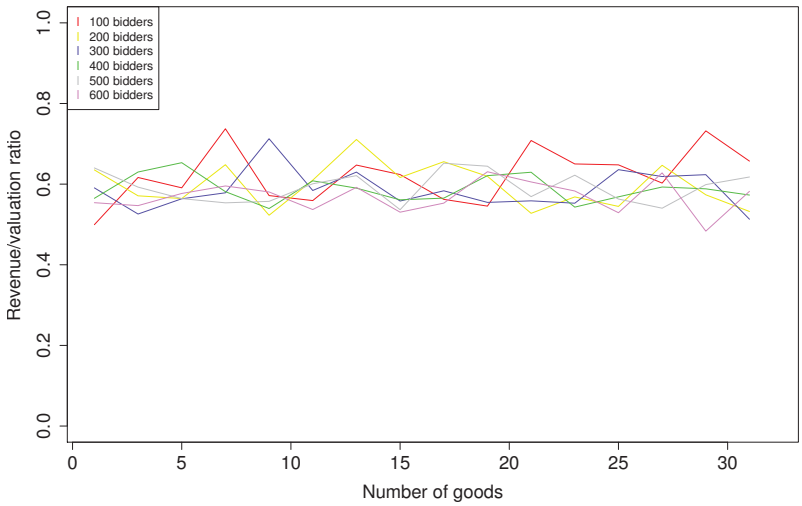


Figure 5.13: Revenue over valuation (**of winners only**) ratio - one bid allocation process.

Ratio between revenue and all bidders' valuation

Figures 5.14 and 5.15 show the results of the ratio between the revenue and the whole amount the totality of the bidders would be able to pay (i.e., $\sum_{i \in B} max_{g \in G} v_i(g)$).

As for the evaluation in the previous section, Figure 5.14 shows the ratio (for several allocation processes) in the case where each bidder submit several bids for several sets of goods. The ratio increases with the number of goods being sold, from approximately 45% to 60%.

The Figure 5.15 shows the same evaluation in the case of single bid allocations. In such a case truthfulness is ensured. The ratio remains approximately the same for the allocations for different numbers of goods and bidders (as for Figure 5.13). It is interesting to note that the ratio remain important (around 45%), which ensures the seller to obtain a substantial revenue.

5.8 Conclusion

Today's Internet displays vast potential path diversity, which is truncated by the inter-domain routing protocol BGP. The architecture proposed in Chapter 3 enables the propagation and the use of this path diversity. Nevertheless, the amount of available paths would be too important to be configured into the routers of the network service providers without any compensation. Moreover NSPs may consider alternate routes as a value-added service, which requires the selling of the correct route(s) to the correct buyer(s) and the establishment of a prices.

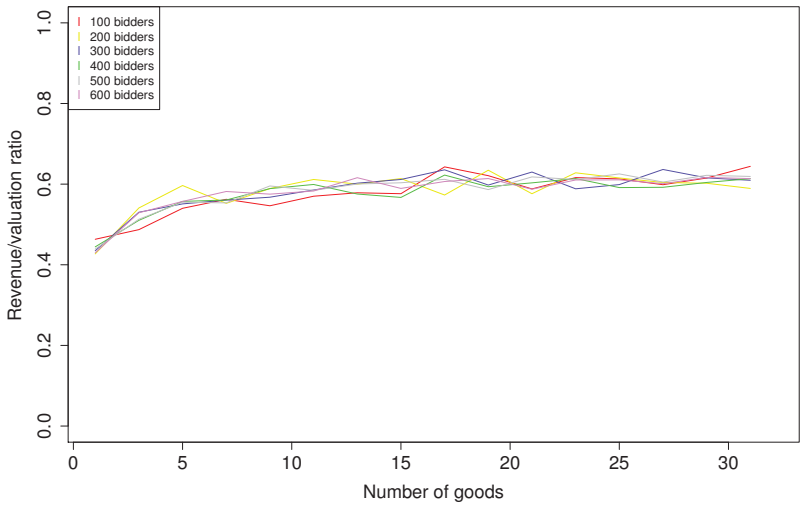


Figure 5.14: Revenue over valuation (**of all bidders**) ratio.

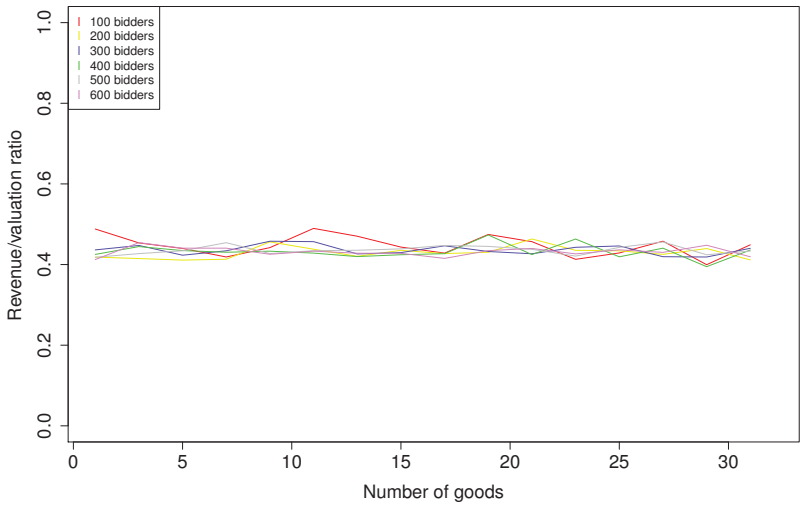


Figure 5.15: Revenue over valuation (**of all bidders**) ratio - one bid allocation processes.

We propose in this chapter an auction-like path allocation mechanism, organized by a network service provider to locally sell routes to its neighbors. Generally speaking, this allocation mechanism has for benefits to both propagate to each

neighbor only the paths it is interested in and to establish a price according to what neighbors can pay.

Furthermore, we propose a selection of the neighbors based on the maximization of the revenue of the seller and a pricing function that is dominant truthful when each bidder submit a single bid - i.e., each bidder has an incentive to bid the true value it evaluates the path or set of paths he bids for - and has the advantages to form the grand coalition - i.e., the seller gains at accepting every new participant to the auction and each bidder gains at participating.

Chapter 6

Conclusion

6.1 Summary

The goal of this thesis was to underline a way to enable the potential path diversity in the Internet.

The propagation and the use of the inter-domain path diversity can leverage some new uses and perspectives for both NSPs and end users (cf. Section 1.2). For instance it can help the users to find paths which characteristics (e.g., delay...) are close to the ones of the applications (e.g., VoIP...). It may also be used by NSPs to perform advanced traffic engineering or ensure the robustness of their networks (e.g., with fast failover).

Nevertheless, using the Internet path diversity is not easy as a solution must be compliment with an important number of requirements , provided in Section 1.3, such as the backward compatibility, the incremental deployability or the opening of potential increase in the revenue of NSPs. We present in Section 1.4 a state of the art about the Inter-domain path-diversity. It shows that the previous works are meanly high level, rarely address the incentives of NSPs to adopt them, are very restrictive in term of usage or are highly disruptive.

We propose in Chapter 2 a local architecture which allows for the use of the received path diversity. The described architecture is simple, backware compatible and relies on technologies that are already available in commercial routers. This proposal is the first step to the global adoption of path diversity. It allows an NSP to cumulate the diversity that it currently receives by eBGP and shares it with its stub networks. From a global Internet point of view, this proposal may seem limited. Nevertheless we show that the diversity currently received by a NSP is important and already allows for the proposal of new services (e.g., disjoint routes). Therefore sharing it with the stub clients gives NSPs the opportunity to propose diversity based value-added services which does not require any coordination with neighboring NSPs. We analyse the scalability from the point of view of the churn and show that the relative amount of churn decreases when several paths are selected. Moreover we show that the way the NSP selects its paths has an important impact on the amount of churn. Therefore enabling path diversity may be an opportunity to better select the paths according to their stability.

We propose in Chapter 3 a generalization of the already described architecture. This global architecture allows NSPs to share their path diversities among them. This architecture unifies all the requirements previously exposed - e.g., it is backward compatible, incrementally deployable, relies on existing technologies... As it is based on the LISP protocol, which is available in current routers, this architecture is readily usable. From a global adoption point of view, the route selection process needs to be relaxed in order to provide advanced services. Moreover we point out that the “Prefer Customer” rule is restrictive in term of diversity. Therefore we take advantage of the route selection relaxation to also relax the “prefer customer” rule. As this rule also allows for the Internet to remain stable, we propose a new criterion that allows for both the stability of the global Internet and for the relaxation of the “Prefer Customer” rule. Our evaluation of the potential diversity, according to this new stability criterion, shows that the potential path diversity between a source domain and a destination domain is very large.

The proposed stability criterion provides an important number of path to be taken into account and may thus lead to a scalability issue. Solving this potential scalability issue is possible by filtering the received path diversity. The Chapter 4 analyses the case where NSPs propagate their full path diversity to their neighbors, each one of them being responsible of its own selection. We consider this approach as a worst case evaluation and propose technical solutions to overcom this potential issue. We show that, despite what is commonly admitted, the scalability issue can not only appear in the core routers but also at the edge of the Internet (i.e., at the boundary between stub networks and NSPs). We quantify the scalability issue of propagating the whole diversity of the Internet. In a worst case situation, we reach a number of paths, that a router should put into its memory, of about 10^8, which is three order of magnitude higher than what is currently performed. We show that some small changes may help to handle this scalability issue. Indeed, by using destination AS routing instead of destination prefix routing, and by limiting the stretch of the AS path length for the same destination, the number of routes which needs to be put in the router memory drops to 10^6, which is just one order of magnitude higher than the current situation. 10^6 routes should be a manageable number of paths.

After the analyse of the worst case situation of the potential scalability issue, the Chapter 5 exposes a way to select the most interesting routes in order first to reduce the potential scalability issue and then to leverage some commercial route related value-added services. We set-up an auction framework where a domain sells routes to its neighbors and each neighbor compete by bidding an amount it is able to pay to win the routes. We focus on providing a solution that is implementable in the context of inter-domain routing, where the numbers of routes to be sold and buyers are important. We manage to set-up a framework which is truthful for one-bid auctions, meaning that bidders bid the true valuations of the routes they compete for, and which form the grand coalition, meaning that the seller has an incentive to accept every new bidder. Moreover we show that the auction framework computation time is polynomial regarding the number of bidders and the number of routes and that the price bidders pay is close to the amount they value the routes,

leading to a high seller's revenue.

6.2 Further work

We try, in this thesis, to tackle an important number of inter-domain path diversity difficulties. Nevertheless, other issues must be solved and some required steps must be succeeded to both harden the basis of these concepts and to motivate NSPs in its adoption.

6.2.1 Architecture test-lab

The architecture provided in the Chapter 4 has not yet been tested. Deploying this architecture in lab would aim at two goals. First some details may not be available yet in commercial routers. A test-bed conception of the whole system should make such problem arise.

Moreover, from an incentive point of view, we made our proposal compatible with the incentive that a solution must be compliant with in order to provoke interest from NSPs. A test-bed may also help in convincing such players, overall to underline how they can benefit from it even in the case of an early adopter.

6.2.2 Software Defined Networking

We detailed an IDRD architecture proposal in Chapter 3. This architecture is based on LISP, and its associated mapping system, for availability reasons (LISP is already implemented in some commercial routers).

Nevertheless the architecture could be implemented with the help of the Software Defined Networking [152, 153] (SDN) concept. Software-defined networking allows for the physical decoupling of the control plane from the data plane.

One SDN mechanism is OpenFlow [154] and some manufacturers already propose it in some new routers. OpenFlow allows an external element (e.g., a mappyng system) to directly change the forwarding entries of routers. MPLS has already been implemented in OpenFlow [155, 156] and IDRD could be implemented thanks to it. OpenFlow could be a good opportunity to implement the IDRD architecture for the lab.

6.2.3 Path auction extension

We proposed in Chapter 5 to sell paths thanks to auction negotiations. Nevertheless, we assume that the network is over-dimensioned, which may not be the case. Therefore the setting up of an auction process which takes into account bandwidth constraints and where bidders negotiate both prices and bandwidth amount to be used could be interesting. Negotiating two parameters (i.e., price and bandwidth) into an auction has already been studied [157]. Nevertheless, an infinity of values bid+bandwidth could be given as input (i.e., bid/valuation in regards to bandwidth). Either the whole bid/bandwidth function of each bidder is communicated to the seller, which is not feasible as the link between valuation and

bandwidth may be considered as a trade secret. Or the auction could be performed by the use of an iterative process where several rounds are required to approach the optimal values of both bandwidth and valuation. An iterative approach has the advantage that the value function is not propagated to the seller, nevertheless it increases a lot the communication requirement (i.e., number of messages), which may be incompatible in an inter-domain environment.

6.2.4 What about the end host ?

This thesis is mostly focused on NSPs as it mainly provides ways to propagate, use, select and sell routes in the Internet. Nevertheless these routes are about to be used by end users which should have the choice between them.

An important further work is the connection between end-users and the diverse paths a NSP may propose to him. By end-user we mean every end-host, could it be in a CDN or a university.

One perspective is the use of middleboxes, which choose the correct routes according to the characteristics (e.g., TCP header) of the packets which are to be forwarded.

Another approach is to use Multi-path TCP [158]. Indeed Multi-Path TCP has been developed principally to allow devices to simultaneously use their multiple interfaces. Nevertheless, several paths could be considered as interfaces and few changes on MP-TCP could make devices test and choose their paths. Current end-host do not understand routing protocols and the information and it is not possible to use them to provide path information. We must therefore find a way to make end-hosts aware of the path diversity without sending them routing updates. Propagating Path-IDs and path characteristics through DNS could be a solution, while letting MP-TCP test the different paths and choose the most suitable ones. This approach allows for the end user to directly express its own will, making the link between the quality of the service provided by his network service provider and the quality of experience he experiments [159].

6.2.5 CDNs and cloud

Communications between CDNs is an active field of research and standardization [160]. A lot of use cases for Content Delivery Network Interconnection have been highlighted in [161]. For instance failure recovery and quality of service improvement have already been expressed as important ones. Current path diversity CDNs are able to enjoy is the one they receive thanks to their multiple eBGP peerings. Such a diversity may be limited to reach their goals. The propagation of the Internet diversity may directly benefit to Content Service Providers as it could help them to meet they quality requirements. Such a use of path diversity has not yet been studied.

Bibliography

[1] S. Ruthfield, “The internet’s history and development: from wartime tool to fish-cam,” Crossroads, vol. 2, no. 1, pp. 2–4, Sep. 1995.

[2] H. Katie and L. Matthew, Where Wizards Stay Up Late: The Origins Of The Internet. Simon & Schuster, 1996.

[3] “Internet World Stats.” [Online]. Available: http://www.internetworldstats.com/

[4] M. Stephens, ““which communications revolution is it, anyway?”,” Journalism & Mass Communication Quarterly, vol. 75, no. 1, pp. 9–13, 1998. [Online]. Available: http://jmq.sagepub.com/content/75/1/9.abstract

[5] J. W. Carey, “Historical pragmatism and the internet,” New Media & Society, vol. 7, no. 4, pp. 443–455, 2005.

[6] R. Rayman, K. Croome, N. Galbraith, R. McClure, R. Morady, S. Peterson, S. Smith, V. Subotic, A. Van Wynsberghe, and S. Primak, “Long-distance robotic telesurgery: a feasibility study for care in remote environments,” The International Journal of Medical Robotics and Computer Assisted Surgery, vol. 2, no. 3, pp. 216–224, 2006.

[7] M. Pais-Vieira, M. Lebedev, C. Kunicki, J. Wang, and M. A. L. Nicolelis, “A brain-to-brain interface for real-time sharing of sensorimotor information.” Scientific reports, vol. 3, p. 1319, 2013.

[8] Y. Rekhter, T. Li, and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” RFC 4271 (Draft Standard), Internet Engineering Task Force, Jan. 2006, updated by RFCs 6286, 6608, 6793. [Online]. Available: http://www.ietf.org/rfc/rfc4271.txt

[9] “bgp best path selection algorithm [ip routing].” [Online]. Available: http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094431.shtml

[10] Juniper, “Configure BGP to Select Multiple BGP Paths.” [Online]. Available: http://www.juniper.net/techpubs/software/junos/junos53/swconfig53-ipv6/html/ipv6-bgp-config29.html

[11] A. Akella, S. Seshan, and A. Shaikh, “An empirical evaluation of wide-area internet bottlenecks,” in Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, ser. IMC ’03. New York, NY, USA: ACM, 2003, pp. 101–114.

[12] R. Serral-Gracià, E. Cerqueira, M. Curado, M. Yannuzzi, E. Monteiro, and X. Masip-Bruin, "An overview of quality of experience measurement challenges for video applications in ip networks," in Proceedings of the 8th international conference on Wired/Wireless Internet Communications, ser. WWIC'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 252–263.

[13] k. claffy, "Internet traffic characterization." Ph.D. dissertation, UC San Diego, 1994.

[14] J. Rexford and C. Dovrolis, "Future internet architecture: clean-slate versus evolutionary research," Commun. ACM, vol. 53, no. 9, pp. 36–40, Sep. 2010.

[15] D. D. Clark, J. Wroclawski, K. R. Sollins, and R. Braden, "Tussle in cyberspace: defining tomorrow's internet," IEEE/ACM Trans. Netw., vol. 13, no. 3, pp. 462–475, Jun. 2005.

[16] J. Scudder, A. Retana, D. Walton, and E. Chen, "Advertisement of Multiple Paths in BGP, IETF Internet Draft (Work in Progress)," September 2012.

[17] X. Yang and D. Wetherall, "Source selectable path diversity via routing deflections," in Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications, ser. SIGCOMM '06. ACM, 2006, pp. 159–170.

[18] M. Motiwala, M. Elmore, N. Feamster, and S. Vempala, "Path splicing," in Proceedings of the ACM SIGCOMM 2008 conference on Data communication, ser. SIGCOMM '08. New York, NY, USA: ACM, 2008, pp. 27–38.

[19] F. Wang and L. Gao, "Path diversity aware interdomain routing," in INFOCOM 2009, IEEE, 2009, pp. 307–315.

[20] I. Ganichev, B. Dai, P. B. Godfrey, and S. Shenker, "Yamr: yet another multipath routing protocol," SIGCOMM Comput. Commun. Rev., vol. 40, no. 5, pp. 13–19, Oct. 2010.

[21] Y. Liao, L. Gao, R. Guerin, and Z.-L. Zhang, "Reliable interdomain routing through multiple complementary routing processes," in Proceedings of the 2008 ACM CoNEXT Conference, ser. CoNEXT '08. New York, NY, USA: ACM, 2008, pp. 68:1–68:6.

[22] N. Kushman, S. Kandula, D. Katabi, and B. M. Maggs, "R-bgp: staying connected in a connected world," in Proceedings of the 4th USENIX conference on Networked systems design & implementation, ser. NSDI'07. Berkeley, CA, USA: USENIX Association, 2007, pp. 25–25.

[23] W. Xu and J. Rexford, "Miro: multi-path interdomain routing," SIGCOMM Comput. Commun. Rev., vol. 36, no. 4, pp. 171–182, Aug. 2006.

[24] P. B. Godfrey, I. Ganichev, S. Shenker, and I. Stoica, "Pathlet routing," SIGCOMM Comput. Commun. Rev., vol. 39, no. 4, pp. 111–122, 2009.

[25] X. Yang, D. Clark, and A. Berger, "Nira: A new inter-domain routing architecture," Networking, IEEE/ACM Transactions on, vol. 15, no. 4, pp. 775–788, Aug.

[26] I. A. Ganichev, D. BIn, P. B. Godfrey, and S. Shenker, "Yamr: Yet another multipath routing protocol," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-150, Oct 2009. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-150.html

[27] "University of Oregon Route Views Project." [Online]. Available: http://www.routeviews.org

[28] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis, "The Locator/ID Separation Protocol (LISP)," RFC 6830 (Experimental), Internet Engineering Task Force, Jan. 2013. [Online]. Available: http://www.ietf.org/rfc/rfc6830.txt

[29] "CAIDA AS rank." [Online]. Available: http://as-rank.caida.org/

[30] R. Mahajan, D. Wetherall, and T. Anderson, "Negotiation-based routing between neighboring isps," in Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation - Volume 2, ser. NSDI'05. Berkeley, CA, USA: USENIX Association, 2005, pp. 29–42.

[31] L. Gao and J. Rexford, "Stable internet routing without global coordination," SIGMETRICS Perform. Eval. Rev., vol. 28, no. 1, pp. 307–317, Jun. 2000.

[32] T. Griffin, F. Shepherd, and G. Wilfong, "The stable paths problem and interdomain routing," Networking, IEEE/ACM Transactions on, vol. 10, no. 2, pp. 232–243, 2002.

[33] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture," RFC 3031 (Proposed Standard), Internet Engineering Task Force, Jan. 2001, updated by RFCs 6178, 6790. [Online]. Available: http://www.ietf.org/rfc/rfc3031.txt

[34] J. Mikians, A. Dhamdhere, C. Dovrolis, P. Barlet-Ros, and J. Solé-Pareta, "Towards a Statistical Characterization of the Interdomain Traffic Matrix," in Networking 2012, May 2012, pp. 111–123.

[35] K.-W. Kwong and R. Guérin, "Controlling the growth of internet routing tables through market mechanisms," in Proceedings of the Re-Architecting the Internet Workshop, ser. ReARCH '10. New York, NY, USA: ACM, 2010, pp. 2:1–2:6.

[36] J. Seedorf and E. Burger, "Application-Layer Traffic Optimization (ALTO) Problem Statement," RFC 5693 (Informational), Internet Engineering Task Force, Oct. 2009. [Online]. Available: http://www.ietf.org/rfc/rfc5693.txt

[37] "DrPeering International." [Online]. Available: http://drpeering.net/

[38] G. Huston, "BGP Routing Table Analysis Report." [Online]. Available: bgp.potaroo.net

[39] R. Steenbergen, "A survey of interdomain routing policies," in North American Network Operators' Group (NANOG) 51, 2011.

[40] X. Dimitropoulos, D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, k. claffy, and G. Riley, "AS Relationships: Inference and Validation," ACM SIGCOMM Computer Communication Review (CCR), vol. 37, no. 1, pp. 29–40, Jan 2007.

[41] L. Subramanian, S. Agarwal, J. Rexford, and R. Katz, "Characterizing the internet hierarchy from multiple vantage points," in INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 2, 2002, pp. 618–627 vol.2.

[42] R. Chandra, P. Traina, and T. Li, "BGP Communities Attribute," RFC 1997 (Proposed Standard), Internet Engineering Task Force, Aug. 1996. [Online]. Available: http://www.ietf.org/rfc/rfc1997.txt

[43] V. Giotsas and S. Zhou, "Valley-free violation in internet routing - analysis based on bgp community data," in Communications (ICC), 2012 IEEE International Conference on, 2012, pp. 1193–1197.

[44] W. Mühlbauer, S. Uhlig, B. Fu, M. Meulle, and O. Maennel, "In search for an appropriate granularity to model routing policies," SIGCOMM Comput. Commun. Rev., vol. 37, no. 4, pp. 145–156, Aug. 2007.

[45] P. Gill, S. Goldberg, and M. Schapira, "A survey of interdomain routing policies," in North American Network Operators' Group (NANOG) 56, 2012.

[46] C. Perkins, P. Calhoun, and J. Bharatia, "Mobile IPv4 Challenge/Response Extensions (Revised)," RFC 4721 (Proposed Standard), Internet Engineering Task Force, Jan. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4721.txt

[47] W. Mühlbauer, A. Feldmann, O. Maennel, M. Roughan, and S. Uhlig, "Building an as-topology model that captures route diversity," in Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications, ser. SIGCOMM '06. New York, NY, USA: ACM, 2006.

[48] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed internet routing convergence," SIGCOMM Comput. Commun. Rev., vol. 30, no. 4, pp. 175–187, Aug. 2000.

[49] J. Chandrashekar, Z. Duan, Z.-L. Zhang, and J. Krasky, "Limiting path exploration in bgp," in INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE, vol. 4, March, pp. 2337–2348 vol. 4.

[50] P. Bangera and S. Gorinsky, "Impact of prefix hijacking on payments of providers," in Communication Systems and Networks (COMSNETS), 2011 Third International Conference on, 2011, pp. 1–10.

[51] F. Contat, S. Nataf, and G. Valadon, "Influence des bonnes pratiques sur les incidents bgp," 2012. [Online]. Available: https://www.sstic.org/2012/presentation/influence_des_bonnes_pratiques_sur_les_incidents_bgp/

[52] F. Wang, B. Dai, and J. Su, "How can multipath dissemination help to detect prefix hijacking?" in Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on, 2011, pp. 1–8.

[53] K. K. Lakshminarayanan, I. Stoica, S. Shenker, and J. Rexford, "Routing as a service," EECS Department, University of California,

Berkeley, Tech. Rep. UCB/EECS-2006-19, Feb 2006. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-19.html

[54] A. Feldmann and J. Rexford, "Ip network configuration for intradomain traffic engineering," IEEE Network Magazine, vol. 15, pp. 46–57, 2001.

[55] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels," RFC 3209 (Proposed Standard), Internet Engineering Task Force, Dec. 2001, updated by RFCs 3936, 4420, 4874, 5151, 5420, 5711, 6780, 6790. [Online]. Available: http://www.ietf.org/rfc/rfc3209.txt

[56] S. Uhlig and O. Bonaventure, "Designing bgp-based outbound traffic engineering techniques for stub ases," Comput. Commun. Rev, vol. 34, p. 2004, 2004.

[57] N. Feamster, J. Borkenhagen, and J. Rexford, "Guidelines for interdomain traffic engineering," SIGCOMM Comput. Commun. Rev., vol. 33, no. 5, pp. 19–30, Oct. 2003.

[58] Y. Wang and J. Rexford, "A modular rcp for flexible interdomain route control," in Proceedings of the 2006 ACM CoNEXT conference, ser. CoNEXT '06. New York, NY, USA: ACM, 2006, pp. 56:1–56:2.

[59] N. Feamster, H. Balakrishnan, J. Rexford, A. Shaikh, and J. van der Merwe, "The case for separating routing from routers," in Proceedings of the ACM SIGCOMM workshop on Future directions in network architecture, ser. FDNA '04. New York, NY, USA: ACM, 2004, pp. 5–12.

[60] M. Caesar, D. Caldwell, N. Feamster, J. Rexford, A. Shaikh, and J. van der Merwe, "Design and implementation of a routing control platform," in Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation - Volume 2, ser. NSDI'05. Berkeley, CA, USA: USENIX Association, 2005, pp. 15–28.

[61] Y. Wang, I. Avramopoulos, and J. Rexford, "Morpheus: making routing programmable," in Proceedings of the 2007 SIGCOMM workshop on Internet network management, ser. INM '07. New York, NY, USA: ACM, 2007, pp. 285–286.

[62] F. Wang and L. Gao, "On inferring and characterizing internet routing policies," in Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, ser. IMC '03. New York, NY, USA: ACM, 2003, pp. 15–26.

[63] D. Meyer, L. Zhang, and K. Fall, "Report from the IAB Workshop on Routing and Addressing," RFC 4984 (Informational), Internet Engineering Task Force, Sep. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4984.txt

[64] K. Fall, G. Iannaccone, S. Ratnasamy, and P. Brighten Godfrey, "Routing Tables: Is Smaller Really Much Better?" in Proceedings of HotNets 2009, 2009.

[65] B. Augustin, T. Friedman, and R. Teixeira, "Measuring multipath routing in the internet," Networking, IEEE/ACM Transactions on, vol. 19, no. 3, pp. 830–840, 2011.

[66] T. Bates, E. Chen, and R. Chandra, “BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP),” RFC 4456 (Draft Standard), Internet Engineering Task Force, Apr. 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4456.txt

[67] T. Bates, Y. Rekhter, R. Chandra, and D. Katz, “Multiprotocol Extensions for BGP-4,” RFC 2858 (Proposed Standard), Internet Engineering Task Force, Jun. 2000, obsoleted by RFC 4760. [Online]. Available: http://www.ietf.org/rfc/rfc2858.txt

[68] D. Pei, M. Azuma, D. Massey, and L. Zhang, “Bgp-rcn: improving bgp convergence through root cause notification,” Comput. Netw., vol. 48, no. 2, pp. 175–194, Jun. 2005.

[69] Y. Liao, L. Gao, R. Guerin, and Z.-L. Zhang, “Multi-process inter-domain routing,” ECE Department, UMass Amherst„ Tech. Rep. TR-08-CSE-09, 2008. [Online]. Available: http://rio.ecs.umass.edu/~yliao/pmwiki/uploads/Research/Publication/mpr-tech.pdf

[70] Y. Wang, M. Schapira, and J. Rexford, “Neighbor-specific bgp: more flexible routing policies while improving global stability,” Computing Research Repository, vol. abs/0906.3, pp. 217–228, 2009.

[71] A. Gurney and T. Griffin, “Neighbor-specific bgp: An algebraic exploration,” in Network Protocols (ICNP), 2010 18th IEEE International Conference on, Oct., pp. 103–112.

[72] J. Hui, J. Vasseur, D. Culler, and V. Manral, “An IPv6 Routing Header for Source Routes with the Routing Protocol for Low-Power and Lossy Networks (RPL),” RFC 6554 (Proposed Standard), Internet Engineering Task Force, Mar. 2012. [Online]. Available: http://www.ietf.org/rfc/rfc6554.txt

[73] M. Yannuzzi, X. Masip-Bruin, and O. Bonaventure, “Open issues in interdomain routing: a survey,” Network, IEEE, vol. 19, no. 6, pp. 49–56, 2005.

[74] V. Van den Schrieck, P. Francois, and O. Bonaventure, “Bgp add-paths: the scaling/performance tradeoffs,” IEEE J.Sel. A. Commun., vol. 28, no. 8, pp. 1299–1307, Oct. 2010.

[75] J. He and J. Rexford, “Toward internet-wide multipath routing,” Network, IEEE, vol. 22, no. 2, pp. 16 –21, 2008.

[76] D. Saucez, L. Iannone, O. Bonaventure, and D. Farinacci, “Designing a deployable internet: The locator/identifier separation protocol,” Internet Computing, IEEE, vol. 16, no. 6, pp. 14–21, 2012.

[77] B. Quoitin, L. Iannone, C. de Launois, and O. Bonaventure, “Evaluating the benefits of the locator/identifier separation,” in Proceedings of 2nd ACM/IEEE international workshop on Mobility in the evolving internet architecture, ser. MobiArch ’07. New York, NY, USA: ACM, 2007, pp. 5:1–5:6.

[78] D. Lee, “LISP Deployment at Facebook,” October 2010.

[79] P. Mockapetris, “Domain names - concepts and facilities,” RFC 1034 (INTERNET STANDARD), Internet Engineering Task Force, Nov. 1987, updated by RFCs 1101, 1183, 1348, 1876, 1982, 2065, 2181, 2308, 2535, 4033, 4034, 4035, 4343, 4035, 4592, 5936. [Online]. Available: http://www.ietf.org/rfc/rfc1034.txt

[80] J. Ahrenholz, “Host Identity Protocol Distributed Hash Table Interface,” RFC 6537 (Experimental), Internet Engineering Task Force, Feb. 2012. [Online]. Available: http://www.ietf.org/rfc/rfc6537.txt

[81] C. Perkins, “IP Encapsulation within IP,” RFC 2003 (Proposed Standard), Internet Engineering Task Force, Oct. 1996, updated by RFC 3168. [Online]. Available: http://www.ietf.org/rfc/rfc2003.txt

[82] V. Fuller and D. Farinacci, “Locator/ID Separation Protocol (LISP) Map-Server Interface,” RFC 6833 (Experimental), Internet Engineering Task Force, Jan. 2013. [Online]. Available: http://www.ietf.org/rfc/rfc6833.txt

[83] P. Marques, R. Fernando, E. Chen, P. Mohapatra, and H. Gredler, “Advertisement of the best external route in BGP, IETF Internet Draft (Work in Progress),” January 2012.

[84] D. Saucez, B. Donnet, L. Iannone, and O. Bonaventure, “Interdomain traffic engineering in a locator/identifier separation context,” in Internet Network Management Workshop, 2008. INM 2008. IEEE, 2008, pp. 1–6.

[85] Y. Wang, M. Schapira, and J. Rexford, “Neighbor-specific bgp: more flexible routing policies while improving global stability,” in Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems, ser. SIGMETRICS ’09. New York, NY, USA: ACM, 2009, pp. 217–228.

[86] V. Van den Schrieck, P. Francois, and O. Bonaventure, “BGP Add-Paths: The Scaling/Performance Tradeoffs,” IEEE Journal on Selected Areas in Communications, vol. 28, no. 8, pp. 1299–1307, Oct. 2010.

[87] M. Roughan, W. Willinger, O. Maennel, D. Perouli, and R. Bush, “10 lessons from 10 years of measuring and modeling the internet’s autonomous systems,” Selected Areas in Communications, IEEE Journal on, vol. 29, no. 9, pp. 1810–1821, 2011.

[88] S. Qiu, P. McDaniel, and F. Monrose, “Toward valley-free inter-domain routing,” in Communications, 2007. ICC ’07. IEEE International Conference on, 2007, pp. 2009–2016.

[89] F. Wang, J. Qiu, L. Gao, and J. Wang, “On understanding transient interdomain routing failures,” IEEE/ACM Trans. Netw., vol. 17, no. 3, pp. 740–751, Jun. 2009.

[90] W. Mühlbauer, S. Uhlig, B. Fu, M. Meulle, and O. Maennel, “In search for an appropriate granularity to model routing policies,” in Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications, ser. SIGCOMM ’07. New York, NY, USA: ACM, 2007, pp. 145–156.

[91] A. Elmokashfi, A. Kvalbein, and C. Dovrolis, “Bgp churn evolution: A perspective from the core,” Networking, IEEE/ACM Transactions on, vol. 20, no. 2, pp. 571–584, 2012.

[92] F. Wang, Z. M. Mao, J. Wang, L. Gao, and R. Bush, “A measurement study on the impact of routing events on end-to-end internet path performance,” SIGCOMM Comput. Commun. Rev., vol. 36, no. 4, pp. 375–386, Aug. 2006.

[93] C. Villamizar, R. Chandra, and R. Govindan, “BGP Route Flap Damping,” RFC 2439 (Proposed Standard), Internet Engineering Task Force, Nov. 1998. [Online]. Available: http://www.ietf.org/rfc/rfc2439.txt

[94] L. Iannone and O. Bonaventure, “On the cost of caching locator/id mappings,” in Proceedings of the 2007 ACM CoNEXT conference, ser. CoNEXT ’07. New York, NY, USA: ACM, 2007, pp. 7:1–7:12.

[95] L. Iannone, D. Saucez, and O. Bonaventure, “Locator/ID Separation Protocol (LISP) Map-Versioning,” RFC 6834 (Experimental), Internet Engineering Task Force, Jan. 2013. [Online]. Available: http://www.ietf.org/rfc/rfc6834.txt

[96] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang, “The (in)completeness of the observed internet as-level structure,” IEEE/ACM Trans. Netw., vol. 18, no. 1, pp. 109–122, Feb. 2010.

[97] S. Uhlig and S. Tandel, “Quantifying the bgp routes diversity inside a tier-1 network,” in Proceedings of the 5th international IFIP-TC6 conference on Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems, ser. NETWORKING’06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 1002–1013.

[98] B. Quoitin, C. Pelsser, L. Swinnen, O. Bonaventure, and S. Uhlig, “Interdomain traffic engineering with bgp,” Communications Magazine, IEEE, vol. 41, no. 5, pp. 122–128, 2003.

[99] L. Jakab, A. Cabellos-Aparicio, F. Coras, D. Saucez, and O. Bonaventure, “Lisp-tree: A dns hierarchy to support the lisp mapping system,” Selected Areas in Communications, IEEE Journal on, vol. 28, no. 8, pp. 1332–1343, 2010.

[100] V. Fuller, D. Farinacci, D. Meyer, and D. Lewis, “Locator/ID Separation Protocol Alternative Logical Topology (LISP+ALT),” RFC 6836 (Experimental), Internet Engineering Task Force, Jan. 2013. [Online]. Available: http://www.ietf.org/rfc/rfc6836.txt

[101] A. Farrel, J.-P. Vasseur, and J. Ash, “A Path Computation Element (PCE)-Based Architecture,” RFC 4655 (Informational), Internet Engineering Task Force, Aug. 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4655.txt

[102] D. Farinacci, P. Lahiri, and M. Kowal, “Lisp traffic engineering use-cases, ietf internet draft (work in progress),” January 2013.

[103] D. Farinacci, D. Meyer, and J. Snijders, “Lisp canonical address format (lcaf), ietf internet draft (work in progress),” March 2013.

[104] R. Enns, M. Bjorklund, J. Schoenwaelder, and A. Bierman, "Network Configuration Protocol (NETCONF)," RFC 6241 (Proposed Standard), Internet Engineering Task Force, Jun. 2011. [Online]. Available: http://www.ietf.org/rfc/rfc6241.txt

[105] T. Nadeau, C. Srinivasan, and A. Viswanathan, "Multiprotocol Label Switching (MPLS) Forwarding Equivalence Class To Next Hop Label Forwarding Entry (FEC-To-NHLFE) Management Information Base (MIB)," RFC 3814 (Proposed Standard), Internet Engineering Task Force, Jun. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3814.txt

[106] F. Valera, I. Van Beijnum, A. García-Martínez, and M. Bagnulo, Multi-path BGP: motivations and solutions. Cambridge University Press, 2011. [Online]. Available: http://e-archivo.uc3m.es/handle/10016/10324.

[107] M. Yannuzzi, X. Masip-Bruin, S. Sanchez, J. Domingo-Pascual, A. Oreda, and A. Sprintson, "On the challenges of establishing disjoint qos ip/mpls paths across multiple domains," Communications Magazine, IEEE, vol. 44, no. 12, pp. 60–66, 2006.

[108] Y. Wang, I. Avramopoulos, and J. Rexford, "Morpheus: Enabling Flexible Interdomain Routing Policies ," Princetown Technical Report, Tech. Rep., 2007. [Online]. Available: http://www.cs.princeton.edu/research/techreps/TR-802-07

[109] A. J. T. Gurney and T. G. Griffin, "Neighbor-specific bgp: An algebraic exploration," in Proceedings of the The 18th IEEE International Conference on Network Protocols, ser. ICNP '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 103–112.

[110] T. Erlebach, A. Hall, A. Panconesi, and D. Vukadinović, "Cuts and disjoint paths in the valley-free path model of internet bgp routing," in Combinatorial and Algorithmic Aspects of Networking. Springer Berlin / Heidelberg, 2005, vol. 3405, pp. 95–95.

[111] E. Rosen, D. Tappan, G. Fedorkow, Y. Rekhter, D. Farinacci, T. Li, and A. Conta, "MPLS Label Stack Encoding," RFC 3032 (Proposed Standard), Internet Engineering Task Force, Jan. 2001, updated by RFCs 3443, 4182, 5332, 3270, 5129, 5462, 5586. [Online]. Available: http://www.ietf.org/rfc/rfc3032.txt

[112] S. Amante, B. Carpenter, S. Jiang, and J. Rajahalme, "IPv6 Flow Label Specification," RFC 6437 (Proposed Standard), Internet Engineering Task Force, Nov. 2011. [Online]. Available: http://www.ietf.org/rfc/rfc6437.txt

[113] S. Secci, K. Liu, G. K. Rao, and B. Jabbari, "Resilient traffic engineering in a transit-edge separated internet routing." in ICC. IEEE, 2011, pp. 1–6.

[114] J. Postel, "Internet Protocol," RFC 791 (INTERNET STANDARD), Internet Engineering Task Force, Sep. 1981, updated by RFCs 1349, 2474. [Online]. Available: http://www.ietf.org/rfc/rfc791.txt

[115] Z. Houidi and M. Meulle, “A new vpn routing approach for large scale networks,” in Network Protocols (ICNP), 2010 18th IEEE International Conference on, 2010, pp. 124–133.

[116] R. Day and P. Milgrom, “Core-selecting package auctions,” International Journal of Game Theory, vol. 36, no. 3-4, pp. 393–407, Jul. 2007.

[117] L. M. Ausubel and P. Cramton, “Vickrey auctions with reserve pricing,” Tech. Rep., 2004.

[118] Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press, 2009, vol. 54, no. 1-4.

[119] P. Milgrom, Putting Auction Theory to Work. Cambridge University Press, 2004.

[120] P. Maille and B. Tuffin, “Why vcg auctions can hardly be applied to the pricing of inter-domain and ad hoc networks,” in Next Generation Internet Networks, 3rd EuroNGI Conference on, 2007, pp. 36–39.

[121] L. M. Ausubel and P. Milgrom, The Lovely but Lonely Vickrey Auction, P. Cramton, Y. Shoham, and R. Steinberg, Eds. MIT Press, 2006, vol. 94305, no. 03.

[122] B. Day and P. Milgrom, “Optimal Incentives in Core-Selecting Auctions,” in Handbook of Market Design, 2010.

[123] R. W. Day and S. Raghavan, “Computing core payments in combinatorial auctions,” SIGecom Exch., vol. 7, no. 1, pp. 22–24, Dec. 2007.

[124] A. Erdil and P. Klemperer, “A new payment rule for core-selecting package auctions,” Journal of the European Economic Association, vol. 8, no. 2-3, pp. 537–547, 2010.

[125] P. Briest and P. Krysta, “Single-minded unlimited supply pricing on sparse instances,” in In Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms, 2006, pp. 1093–1102.

[126] M.-F. Balcan and F. Constantin, “Sequential item pricing for unlimited supply,” in Proceedings of the 6th international conference on Internet and network economics, ser. WINE’10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 50–62.

[127] M.-F. Balcan and A. Blum, “Approximation algorithms and online mechanisms for item pricing,” in Proceedings of the 7th ACM conference on Electronic commerce, ser. EC ’06. New York, NY, USA: ACM, 2006, pp. 29–35.

[128] M.-F. Balcan, A. Blum, and Y. Mansour, “Item pricing for revenue maximization,” in Proceedings of the 9th ACM conference on Electronic commerce, ser. EC ’08. New York, NY, USA: ACM, 2008, pp. 50–59.

[129] G. Aggarwal, T. Feder, R. Motwani, and A. Zhu, “Algorithms for multi-product pricing,” in In Proceedings of the International Colloquium on Automata, Languages, and Programming, 2004, pp. 72–83.

[130] J. D. Hartline and V. Koltun, "Near-optimal pricing in near-linear time," in In 9th Workshop on Algorithms and Data Structures, 2005, pp. 422–431.

[131] J. D. Hartline and R. McGrew, "From optimal limited to unlimited supply auctions," in Proceedings of the 6th ACM conference on Electronic commerce, ser. EC '05. New York, NY, USA: ACM, 2005, pp. 175–182.

[132] M.-F. Balcan and A. Blum, "Mechanism design via machine learning," in Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science, ser. FOCS '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 605–614.

[133] I. Segal, "Optimal pricing mechanisms with unknown demand," American Economic Review, vol. 93, p. 2003, 2003.

[134] V. Guruswami, J. D. Hartline, A. R. Karlin, D. Kempe, C. Kenyon, and F. McSherry, "On profit-maximizing envy-free pricing," in Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms, ser. SODA '05. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2005, pp. 1164–1173.

[135] P. Cramton, Y. Shoham, and R. Steinberg, Combinatorial auctions. MIT Press, 2006.

[136] A. V. Goldberg, J. D. Hartline, and A. Wright, "Competitive auctions and digital goods," in Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms, ser. SODA '01. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2001.

[137] A. Goldberg, Jason, H. May, A. V. Goldberg, and J. D. Hartline, "Competitive auctions for multiple digital goods," in In Proc. 9th European Symposium on Algorithms, 2001.

[138] Z. Bar-Yossef, K. Hildrum, and F. Wu, "Incentive-compatible online auctions for digital goods," in Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms, ser. SODA '02. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2002, pp. 964–970.

[139] A. V. Goldberg and J. D. Hartline, "Envy-free auctions for digital goods," in Proceedings of the 4th ACM conference on Electronic commerce, ser. EC '03, 2003, pp. 29–35.

[140] A. V. Goldberg, J. D. Hartline, A. R. Karlin, and M. Saks, "A lower bound on the competitive ratio of truthful auctions," in In Proceedings 21st Symposium on Theoretical Aspects of Computer Science. Springer, 2004, pp. 644–655.

[141] A. V. Goldberg, J. D. Hartline, A. R. Karlin, M. Saks, and A. Wright, "Competitive auctions," Games and Economic Behavior, vol. 55, no. 2, 2006.

[142] P. Maille and B. Tuffin, "Multi-bid auctions for bandwidth allocation in communication networks," in In Proc. of IEEE INFOCOM, 2004.

[143] M. Dramitinos, G. D. Stamoulis, and C. Courcoubetis, "An auction mechanism for allocating the bandwidth of networks to their users," Computer Networks, vol. 51, no. 18, pp. 4979 – 4996, 2007.

[144] J. Feigenbaum, C. Papadimitriou, R. Sami, and S. Shenker, "A bgp-based mechanism for lowest-cost routing," Distrib. Comput., vol. 18, no. 1, pp. 61–72, Jul. 2005.

[145] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, Algorithmic Game Theory. Cambridge University Press, 2007.

[146] V. Valancius, N. Feamster, R. Johari, and V. Vazirani, "Mint: a market for internet transit," in Proceedings of the 2008 ACM CoNEXT Conference, ser. CoNEXT '08. New York, NY, USA: ACM, 2008.

[147] D. Lehmann, L. I. Oćallaghan, and Y. Shoham, "Truth revelation in approximately efficient combinatorial auctions," J. ACM, vol. 49, no. 5, pp. 577–602, Sep. 2002.

[148] P. Milgrom, "Incentives in core-selecting auctions," UCLA Department of Economics, Levine's Bibliography, Oct. 2006.

[149] R. Day and P. Cramton, "Quadratic core-selecting payment rules for combinatorial auctions," Tech. Rep., 2008.

[150] P. Cramton, "Spectrum auction design," Tech. Rep., 2009.

[151] C. H. Papadimitriou, "Computational complexity," in Encyclopedia of Computer Science. Chichester, UK: John Wiley and Sons Ltd., pp. 260–265.

[152] G. Goth, "Software-defined networking could shake up more than packets," Internet Computing, IEEE, vol. 15, no. 4, pp. 6–9, 2011.

[153] D. Staessens, S. Sharma, D. Colle, M. Pickavet, and P. Demeester, "Software defined networking: Meeting carrier grade requirements," in Local Metropolitan Area Networks (LANMAN), 2011 18th IEEE Workshop on, 2011, pp. 1–6.

[154] "OpenFlow." [Online]. Available: http://www.openflow.org/

[155] J. Kempf, S. Whyte, J. Ellithorpe, P. Kazemian, M. Haitjema, N. Beheshti, S. Stuart, and H. Green, "Openflow mpls and the open source label switched router," in Proceedings of the 23rd International Teletraffic Congress, ser. ITC '11. ITCP, 2011, pp. 8–14.

[156] A. R. Sharafat, S. Das, G. Parulkar, and N. McKeown, "Mpls-te and mpls vpns with openflow," SIGCOMM Comput. Commun. Rev., vol. 41, no. 4, pp. 452–453, Aug. 2011.

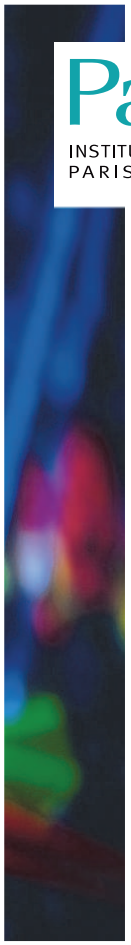
[157] E. Cantillon and M. Pesendorfer, "Combination bidding in multi-unit auctions," C.E.P.R. Discussion Papers, CEPR Discussion Papers 6083, Feb. 2007.

[158] A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, "Architectural Guidelines for Multipath TCP Development," RFC 6182 (Informational), Internet Engineering Task Force, Mar. 2011. [Online]. Available: http://www.ietf.org/rfc/rfc6182.txt

[159] P. Reichl, "From charging for quality of service to charging for quality of experience," annals of telecommunications - annales des télécommunications, vol. 65, no. 3-4, pp. 189–199, 2010.

[160] B. Niven-Jenkins, F. L. Faucheur, and N. Bitar, “Content Distribution Network Interconnection (CDNI) Problem Statement,” RFC 6707 (Informational), Internet Engineering Task Force, Sep. 2012. [Online]. Available: http://www.ietf.org/rfc/rfc6707.txt

[161] G. Bertrand, E. Stephan, T. Burbridge, P. Eardley, K. Ma, and G. Watson, “Use Cases for Content Delivery Network Interconnection,” RFC 6770 (Informational), Internet Engineering Task Force, Nov. 2012. [Online]. Available: http://www.ietf.org/rfc/rfc6770.txt



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

Vers une utilisation de la diversité de chemin dans l'Internet

Xavier MISSERI

RESUME :

Internet est constitué de systèmes autonomes inter-connectés de telle manière que plusieurs chemins existent afin d'aller d'une source à une destination. Néanmoins, l'actuel protocole de routage inter-domaine (BGP), qui ne sélectionne qu'une seule route utilisable, empêche les opérateurs et les utilisateurs finaux de bénéficier la vaste diversité de chemin inhérente à Internet. La mise en marche de cette diversité de chemin a depuis longtemps été reconnue comme une avancée afin d'obtenir des améliorations de robustesse, de qualité de service et d'ingénierie de trafic.

Nous considérons, dans cette thèse, un nouveau service par lequel les opérateurs de télécommunications offrent des routes supplémentaires à leurs clients (en plus de la route par défaut fournie par BGP) comme un service gratuit ou à valeur ajoutée. Ces routes supplémentaires peuvent être utilisées par des clients afin d'optimiser leurs communications, en outrepassant des points de congestion d'Internet, ou les aider à atteindre leurs objectifs d'ingénierie de trafic (meilleurs délais etc.) ou dans un but de robustesse (par exemple en basculant sur un chemin disjoint en cas de panne).

Nous proposons d'abord une architecture simple permettant à un opérateur de télécommunication de bénéficier de la diversité de chemin qu'il reçoit déjà. Nous étendons ensuite cette architecture afin de rendre possible la propagation de cette diversité de chemin, non seulement aux voisins directs mais aussi, de proche en proche, aux autres domaines. Nous profitons de cette occasion pour relaxer la sélection des routes des différents domaines afin de leur permettre de mettre en place de nouveaux paradigmes de routage.

Néanmoins, annoncer des chemins additionnels peut entrainer des problèmes de passage à l'échelle car chaque opérateur peut potentiellement recevoir plus de chemins que ce qu'il peut gérer. Nous quantifions ce problème et mettons en avant des modifications et filtrages simples permettant de réduire ce nombre à un niveau acceptable.

En dernier, nous proposons un processus, inspiré des ventes aux enchères, permettant aux opérateurs de propager aux domaines voisins seulement les chemins qui intéressent les dits voisins. De plus, ce processus permet de mettre en avant un nouveau paradigme de propagation de routes, basé sur des négociations et accords commerciaux.

THÈSE