OPTIC

an **O**ptimal **P**rotection **T**echnique for **I**ntra-domain **C**onvergence

Jean-Romain Luttringer

encadré par Pascal Mérindol

Master Réseaux Informatiques et Systèmes Embarqués
Université de Strasbourg

Laboratoire ICube

ICube - Sciences de l'ingénieur, de l'informatique et de l'imagerie

- Fondé en 2013 sous l'égide du CNRS, de l'Université de Strasbourg, de l'ENGEES et de l'INSA
- Nombreuses collaborations internationales
  - 4000 publications et 50 brevets
- 650 membres répartis dans 16 équipes formant 4 départements

Département de Recherche en Informatique

- Composé de 6 équipes
  - Imagerie, calcul parallèle, intelligence artificielle...

Équipe Réseaux

- Conception algorithmique/protocolaire d'architectures de communication (couches MAC & Réseau)
  - Internet des Objets : réseaux sans fils / équipements contraints
  - Réseaux de coeur : étude de protocoles, topologies, algorithmes, mesures, sécurité, routage

OPTIC

Conception d'une architecture de commutation pour la protection optimale et efficace de la composition BGP/IGP

Contexte Général - intra

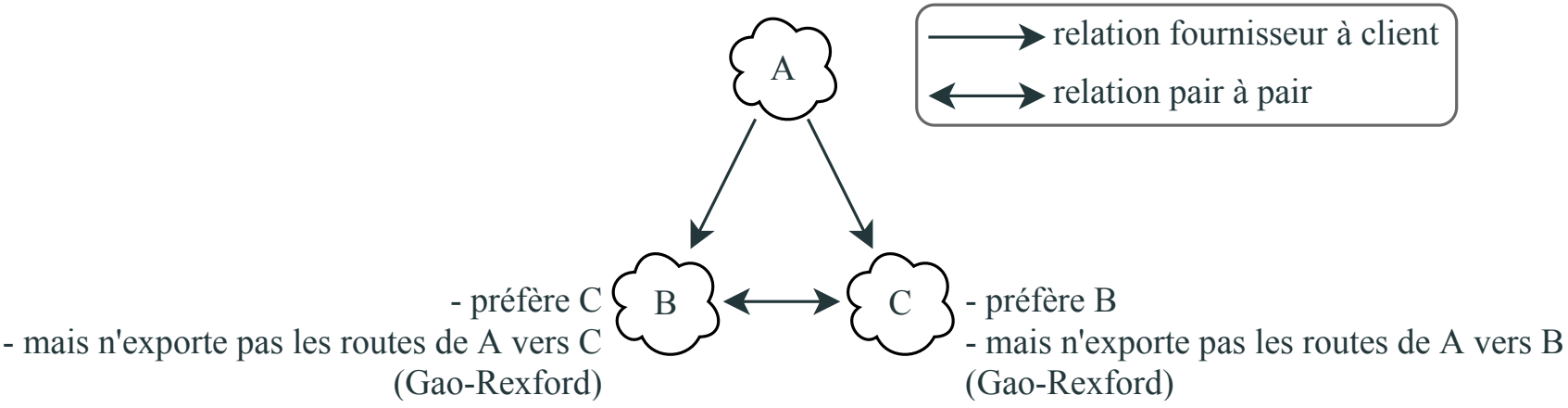
Acheminement de données dans un même domaine

- Domaines (AS) : ensemble de machines inter-connectées sous le contrôle d'une même entité administrative
- intra-AS : échange d'informations via protocole intra-domaine
- Calcul du meilleur chemin selon une algèbre simple, e.g., (min,+)
- Deux familles de protocoles
  - Protocole à vecteur de distance (e.g., RIP/IGRP)
    - Vue **locale** : connaissance du nombre de sauts vers la destination
    - Chemins calculés avec l'algorithme Bellman-Ford distribué
    - Problèmes de passage à l'échelle, convergence lente, anomalies...
  - Protocole à état des liens (e.g., OSPF/IS-IS/EIGRP)
    - Vue **globale** : connaissance quasi-complète de la topologie du réseau
    - Chemins calculés avec l'algorithme Dijkstra
    - Configuration plus complexe

Contexte Général - inter

Acheminement des données entre les domaines

- **Relations commerciales/hiérarchiques entre domaines (AS)** $\Rightarrow$ **nouveaux besoins** Informations filtrées, préférence locale
- **Échange d'informations via protocole à vecteur de chemins**



- **Un seul protocole utilisé en pratique :** Border Gateway Protocol (BGP)
  - Plus complexe, problèmes de convergence
  - Souple : laisse tout contrôle à l'opérateur

Contexte Général - dynamique

Acheminer les données dans des réseaux dynamiques

- **Dynamicité des réseaux IP : des reconfigurations, des pannes**
  - $\Rightarrow$ Re-calcul des chemins (convergence)
  - Temps de rétablissement de la connectivité borné/contraint

Solutions de re-routage rapides

Principalement basées sur des chemins de secours pré-calculés

- **Intra-domaine: convergence inférieure à la seconde**
  - Loop Free Alternate (LFA), U-turn, RLFA ...
- **Inter-domaine: convergence de l'ordre de la minute**
  - Convergence **lente**[1] (propagation, nombre de routes, décision complexe)
  - Indices du plan de données (R-BGP, Blink[2]) ou du plan de contrôle (Swift)

---

[1] Craig Labovitz et al. "Delayed Internet routing convergence". In: *ACM SIGCOMM Computer Communication Review* 30 (2000).

[2] Thomas Holterbach et al. "Blink : Fast Connectivity Recovery Entirely in the Data Plane". In: *Nsdi* (2019).

Contexte Général - composition

Des protocoles distincts mais en interaction: une panne interne peut affecter le trafic inter-domaine

Objectif : atténuer l'impact des pannes internes sur le trafic de transit

3 axes (**POE**) :

- **P**rotection : aucune panne interne ne doit affecter le trafic plus d'un certain temps (approx. temps de convergence intra-domaine)
- **O**ptimalité : le nouveau chemin est optimal (le meilleur après reconfiguration)
- **E**fficacité : temps de calcul viable

Comment gérer efficacement l'interaction de deux protocoles différents ?

Routage - intra vs inter

Routage intra-domaine (IGP)

- Simple et convergence relativement efficace
- Liens **pondérés**, partage des informations entre chaque noeud
- Choix du chemin **minimisant la somme des poids** de chaque liens
- **Cohérence** globale

Routage inter-domaine

- BGP = eBGP (échange externe) + iBGP (échange interne et comparaisons)
- Choix pouvant différer pour chaque noeud : **pas de cohérence globale**
- Classement local des routes via **filtres** et modification des ***attributs*** caractérisants les routes

Routage - intra dans inter

Routes classées suivant ordre lexicographique

1	Highest `local-pref` LP	(economical relationships)
2	Shortest `as-path` AS	
3	Lowest `origin`	(IGP over EGP over `unknown`)
4	Lowest MED	(cold potato routing)
5	Prefer routes learned via eBGP over iBGP	
6	Lowest IGP cost I	(hot potato routing)
7	Lowest `router-id` rid	(arbitrary tie-break)

Simplification du modèle sans perte de généralité: 3 attributs, deux catégories

Route r := attributs externes et attributs internes

$$\beta \circ \alpha$$

$$\{\underbrace{\text{client-provider, AS1-AS2}}_{\beta}, \text{EGP, None, eBGP, } \underbrace{15}_{\alpha}, \text{10.0.0.1}\}$$

Route - dépendance

Intra-domaine et inter-domaine : dépendance

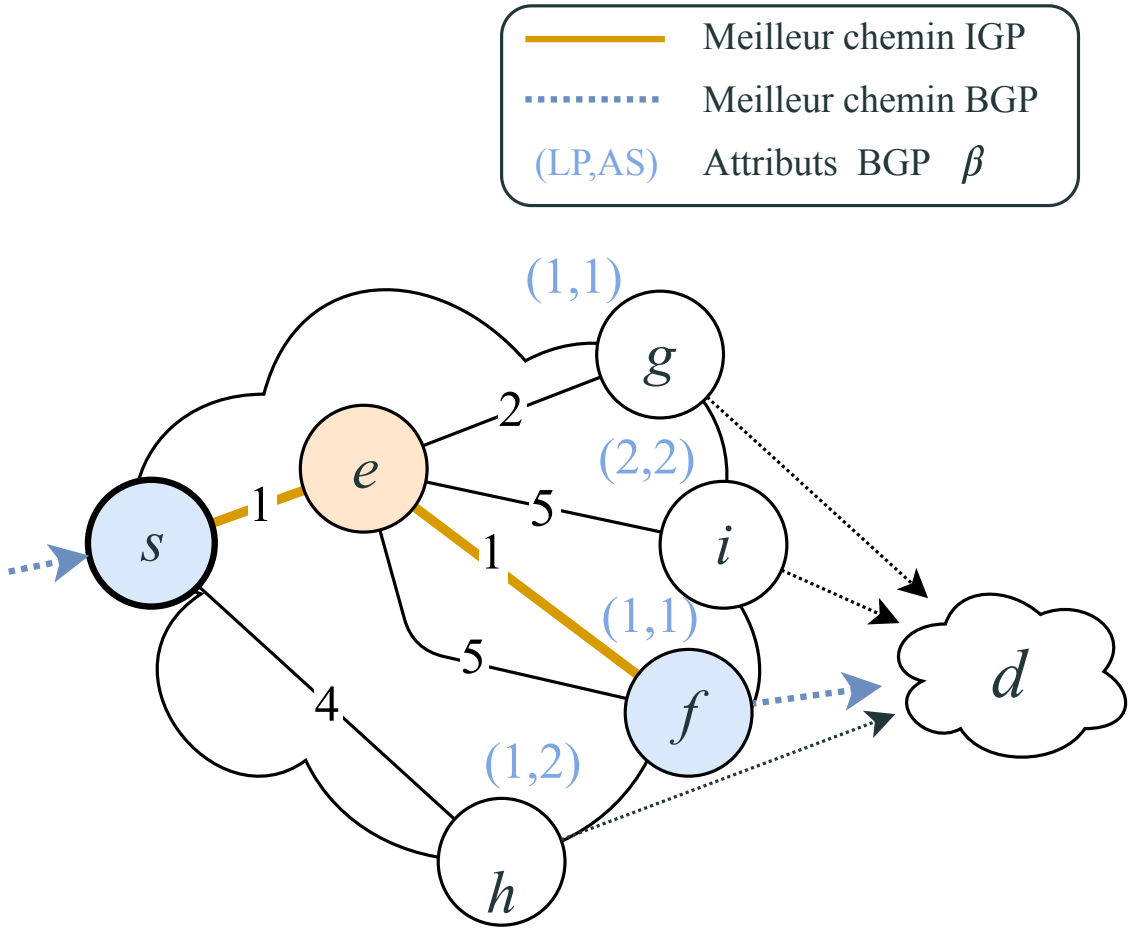
- **Entre les AS** : BGP
  Au sein d'un AS (entrée jusqu'à passerelle) : IGP
- **Le choix (iBGP) de la passerelle est influencé par les distances IGP**
  Si local-pref et as-path (β) égaux, départagées par distance IGP (α)

Si α modifié suite à une panne, le classement entre deux passerelles ayant le même β peut **évoluer**... mais ce n'est pas le cas pour un β différent !

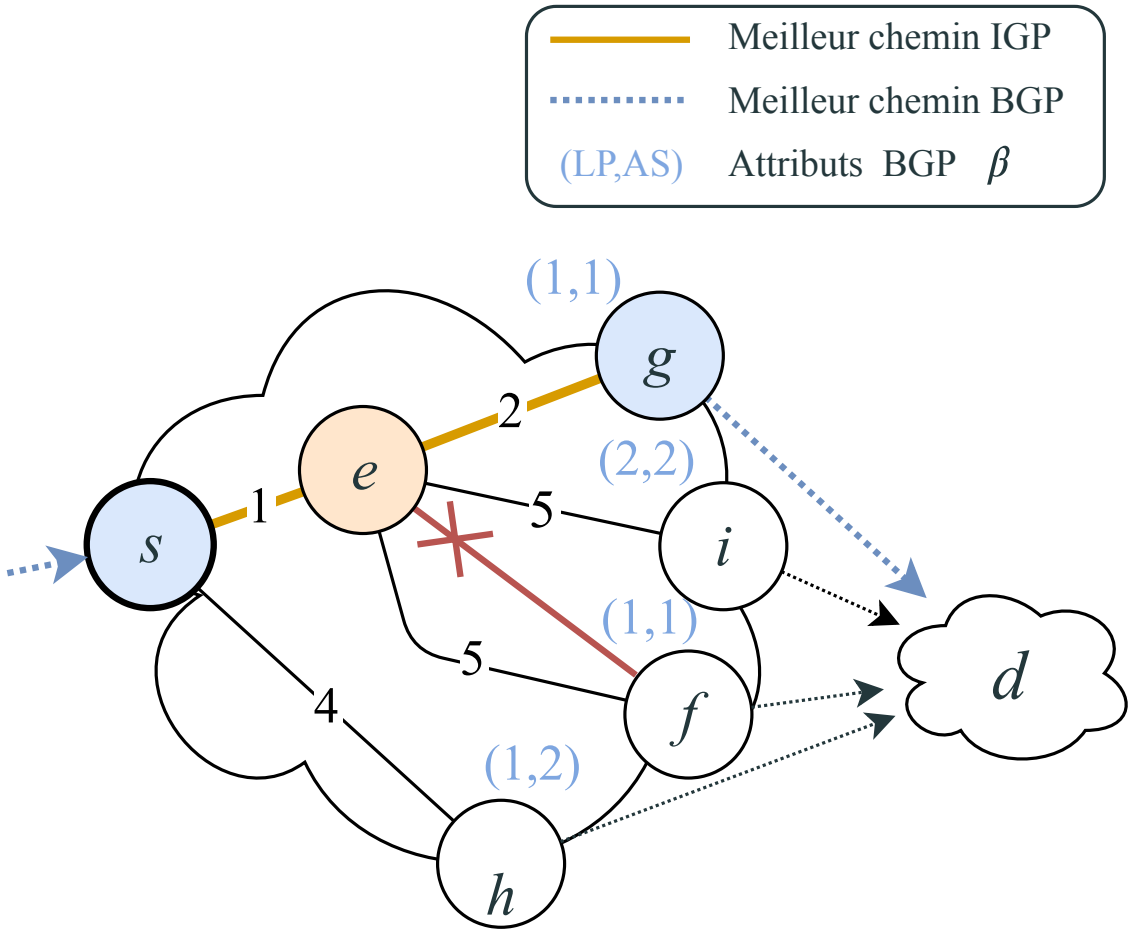
Routage - exemple

Higher is better

	f	$(1,1)$	$\circ$ 2
bad α	g	$(1,1)$	$\circ$ 3
	f	$(1,1)$	$\circ$ 6
bad β	h	$(1,2)$	$\circ$ 4
	i	$(2,2)$	$\circ$ 6



	g	$(1,1) \circ 3$
bad α	f	$(1,1) \circ 6$
bad β	h	$(1,2) \circ 4$
	i	$(2,2) \circ 6$



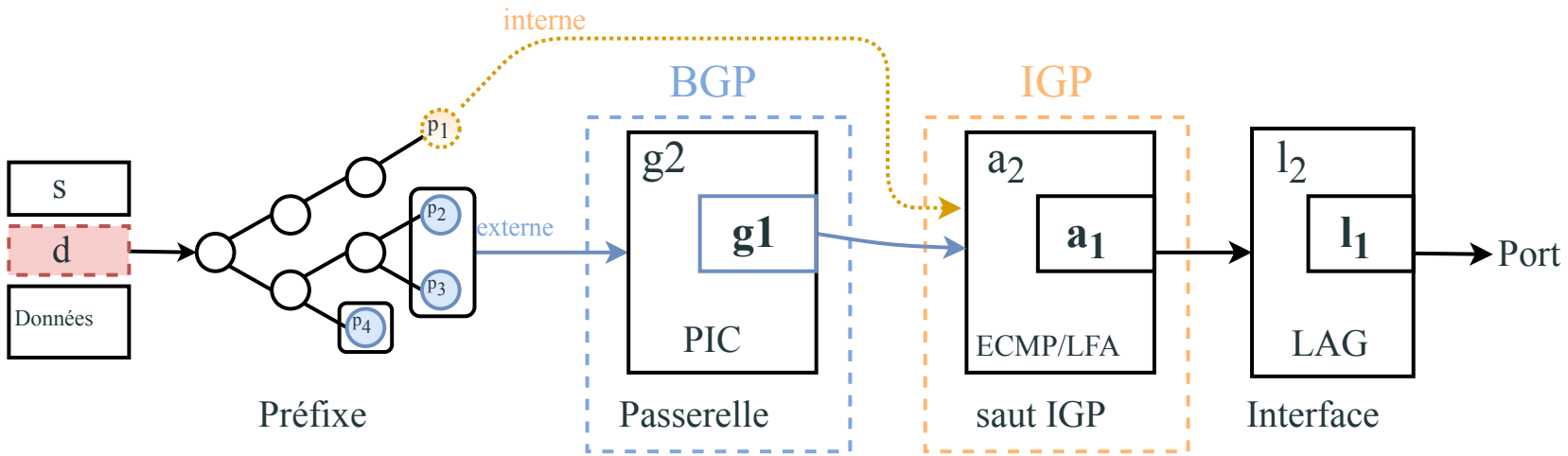
Panne intra-domaine modifiant la meilleure passerelle BGP

Routage - architecture

Vers POE malgré cette inter-dépendance ?

- Compromis existant : concessions sur P, O et E

PIC[3]: Architecture de commutation hiérarchique + BGP "doublé"

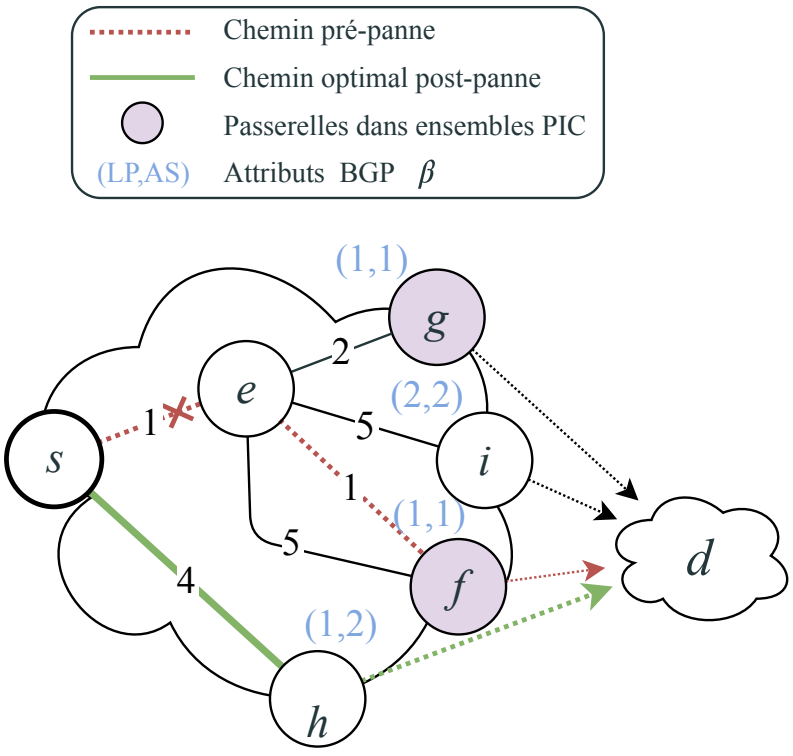


- Panne d'un routeur **interne** : mise à jour rapide grâce à **hiérarchie**
- Panne d'un routeur de **bordure** : BGP "**doublé**"

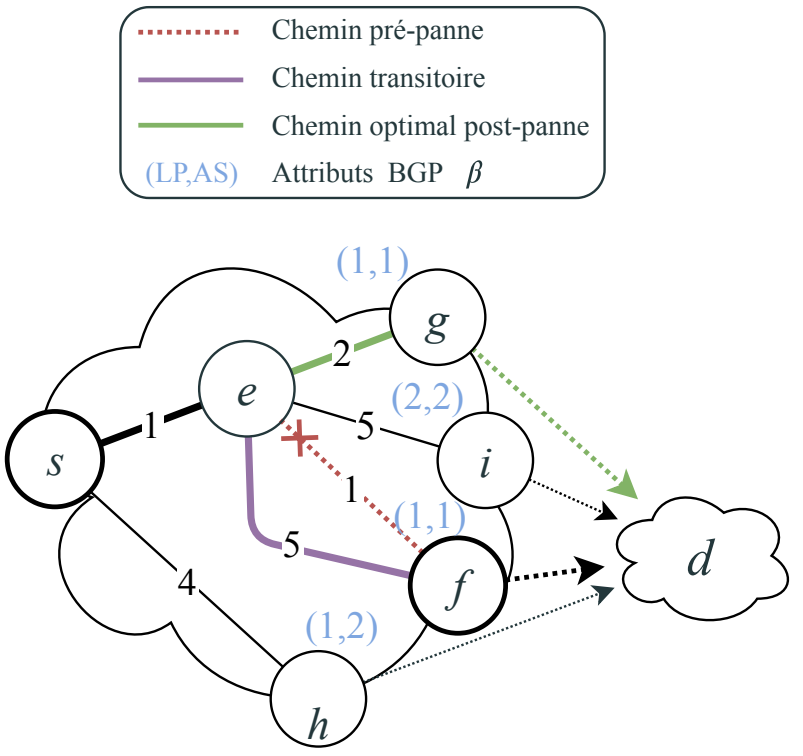
Protection et Optimalité **non garanties en pratique**

[3] Clarence Filsfils et al. "BGP prefix independent convergence (PIC) technical report". In: (2007).

Routage - Les deux défauts de PIC



Pas de protection :
aucune passerelle de secours pour la panne du lien $s \rightarrow e$ (ensemble de passerelles insuffisant)



Pas d'optimalité :
PIC ne prend pas en compte l'interaction IGP-BGP

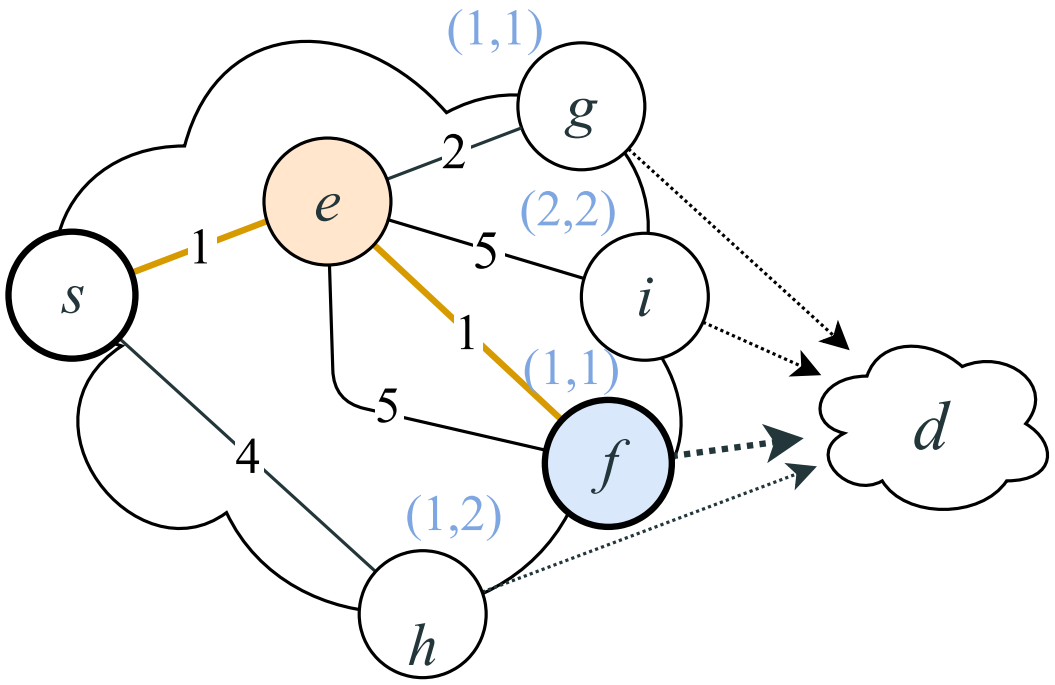
OPTIC - Contribution

Assurer la Protection et l'Optimalité

... sans oublier l'Efficacité

OPTIC - Contribution

f	$(1,1)$	2
g	$(1,1)$	3
h	$(1,2)$	4
i	$(2,2)$	6

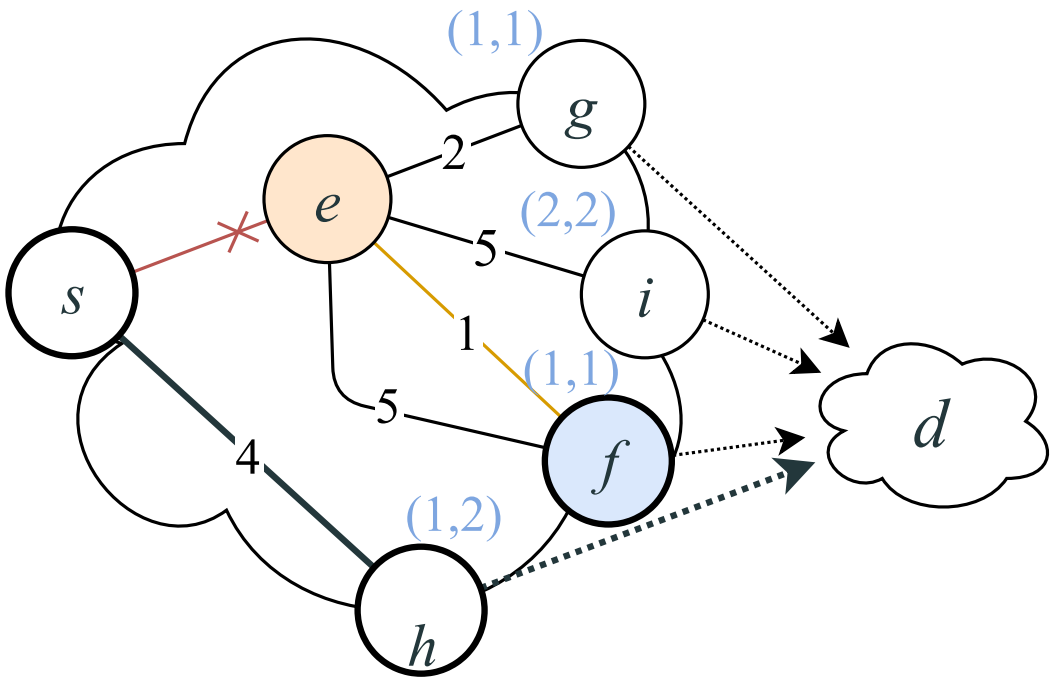


Comment construire un ensemble protecteur et optimal pour une panne ?

- Meilleure passerelle f : pannes sur le meilleur chemin interne ?

OPTIC - Contribution

			$s \rightarrow e$
f	(1,1)	2	∞
g	(1,1)	3	∞
h	(1,2)	4	4
i	(2,2)	6	∞

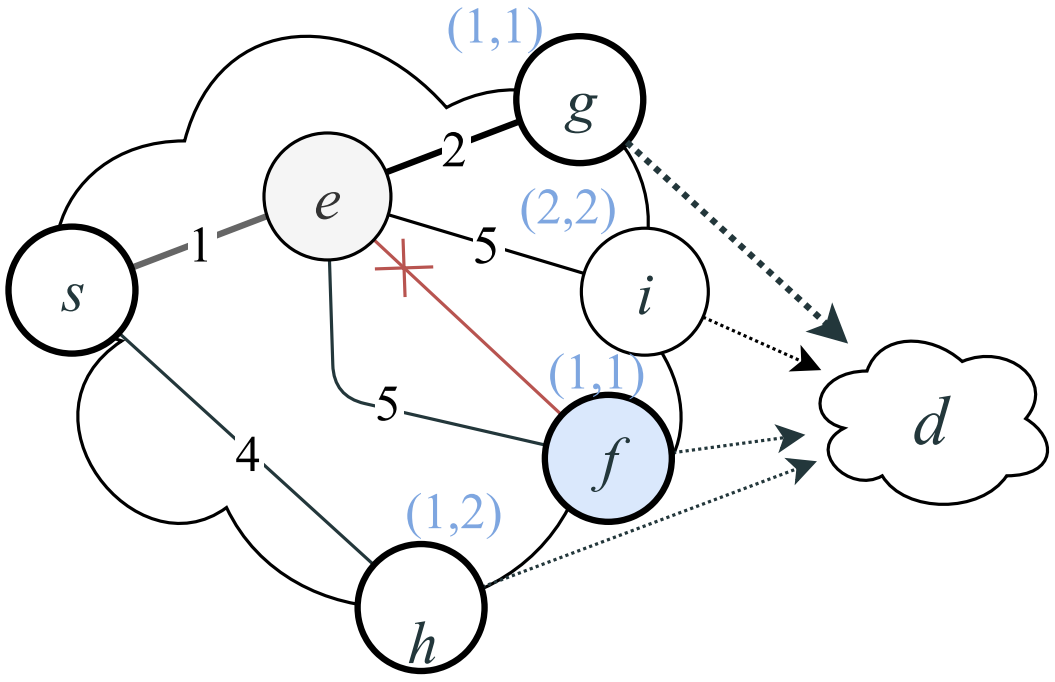


Comment construire un ensemble protecteur et optimal pour une panne ?

- Meilleure passerelle f : pannes sur le meilleur chemin interne ?
  - Si panne de $s \rightarrow e$: meilleure passerelle h

OPTIC - Contribution

			$s \to e$	$e \to f$
f	(1,1)	2	∞	6
g	(1,1)	3	∞	3
h	(1,2)	4	4	4
i	(2,2)	6	∞	6

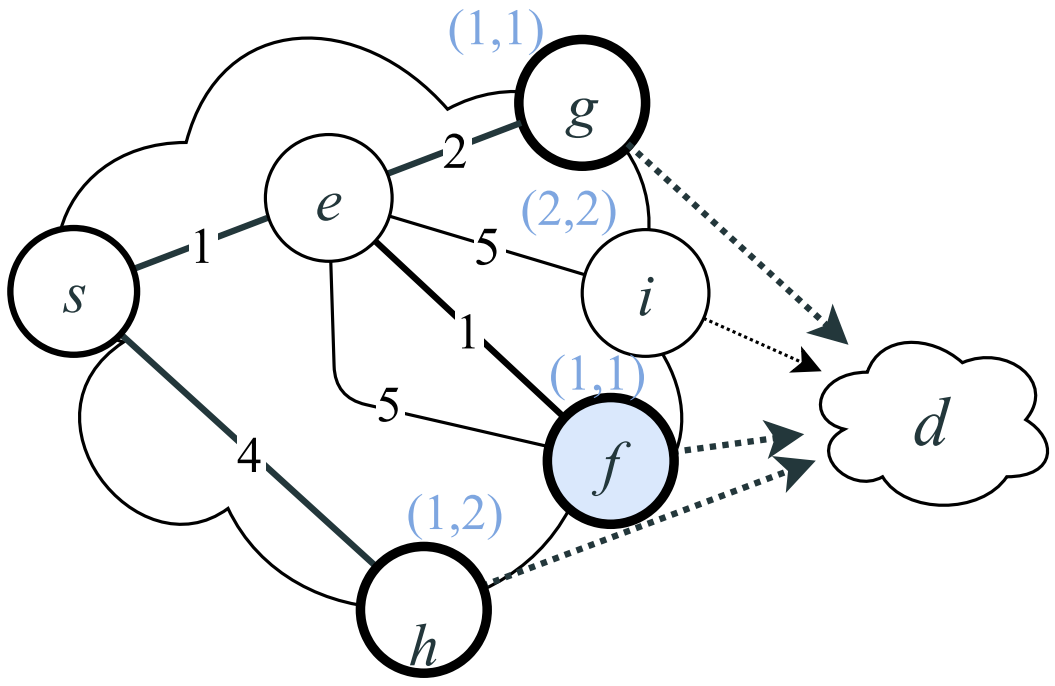


Comment construire un ensemble protecteur et optimal pour une panne ?

- Meilleure passerelle f : pannes sur le meilleur chemin interne ?
  - Si panne de $s \to e$: meilleure passerelle h
  - Si panne de $e \to f$: meilleure passerelle g
  - Sur cet exemple, e et f sont également protégés avec ces deux passerelles

OPTIC - Contribution

			$s \to e$	$e \to f$
f	(1,1)	2	∞	6
g	(1,1)	3	∞	3
h	(1,2)	4	4	4
i	(2,2)	6	∞	6



Comment construire un ensemble protecteur et optimal pour une panne ?

$\{f, g, h\}$ contient la nouvelle meilleure passerelle quelque soit la panne

$\{f, g, h\}$ est un ensemble Protégeant de manière Optimale contre 1 panne

OPTIC - Contribution

Notre première définition: ensembles 1-Optimal-Protecteur (1OP)

Soit b la meilleure passerelle vers un préfixe $\mathcal{P}$ et α la meilleure route intra-domaine vers b.

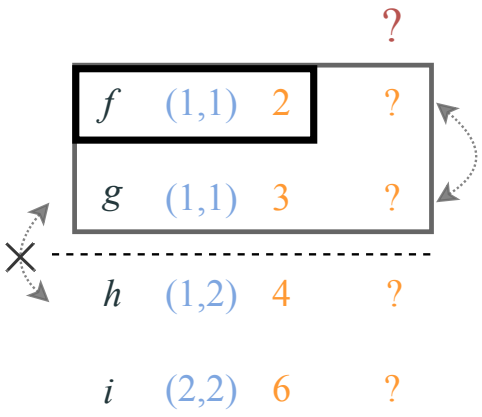
Un ensemble de passerelles $\mathcal{B}$ pour un préfixe $\mathcal{P}$ est ***1-optimal-protecteur*** (**1-OP**) si :

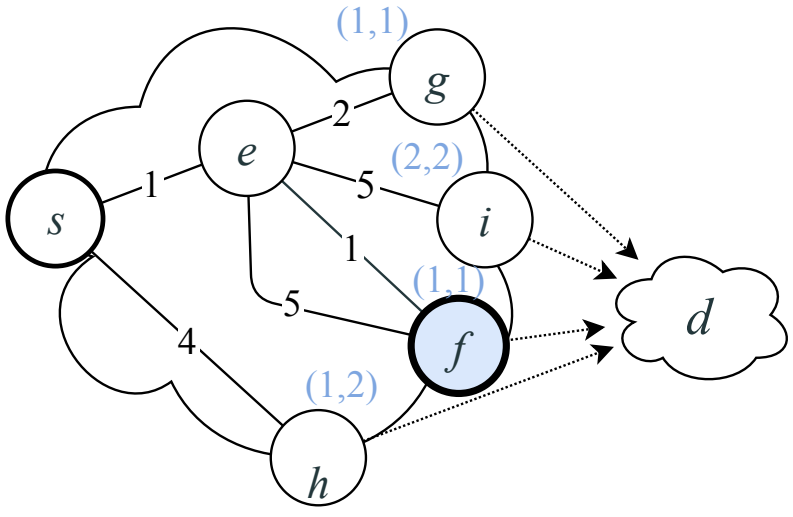
- b est dans $\mathcal{B}$
- Pour toute défaillance d'un composant de α (b incluse), la nouvelle meilleure passerelle b' vers $\mathcal{P}$ est dans $\mathcal{B}$

- Protection et Optimalité atteignables par construction mais gourmand en calcul
  - $\Rightarrow$ Efficacité ?
    Eviter le calcul d'un SPF par composant en panne

OPTIC - Contribution

Optimalité $\equiv$ Quelles passerelles peuvent devenir meilleures que f après une panne ?





Lemme: β-étanchéité

Le classement des routes, lorsque seul les attributs β sont considérés (local-pref et as-path) n'est pas affecté par une panne intra-domaine. (α (attribut interne) peut être modifié, mais pas β)

OPTIC - Contribution

Ensembles construits par bloc: une passerelle et toutes les passerelles pouvant être meilleure après une panne $\Rightarrow$ Toutes les passerelles ayant le même β

Définition: ensembles arrondis IGP

Soit $\mathcal{G}$ un ensemble arrondi IGP pour un préfixe p. Pour toutes r_1, r_2 vers p tel quel $r_1.\beta = r_2.\beta$, alors $r_1, r_2 \in \mathcal{G}$

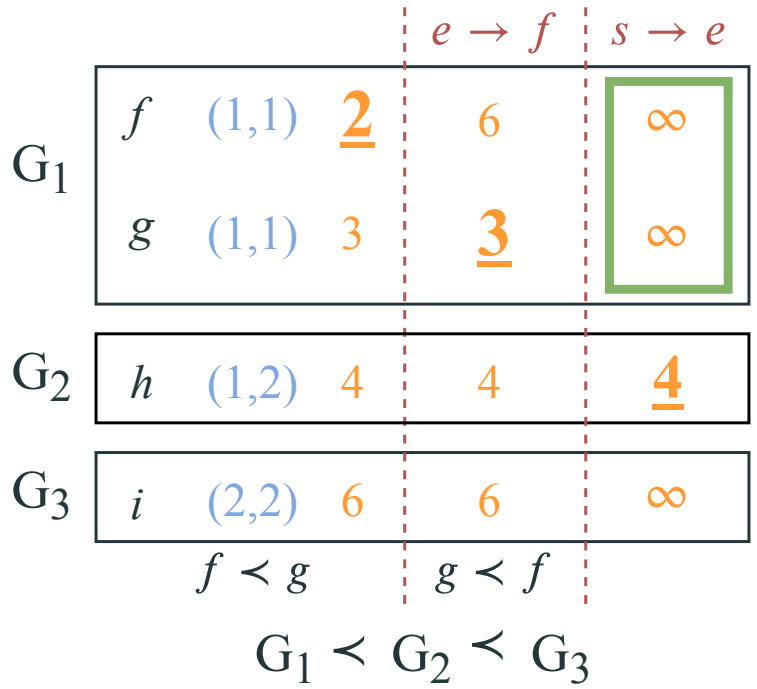
Mais un ensemble arrondi IGP ne suffit pas forcément à garantir la protection

Après une panne intra-domaine :

- L'ordre des routes au sein du même ensemble arrondi IGP peut changer

- Mais l'ordre des ensembles arrondis IGP ne change pas

Si nécessaire, parcours des ensembles arrondis IGP dans l'ordre



OPTIC - Contribution

⇒ Ajout de groupes arrondis IGP (O,E) jusqu'à la vérification de la Protection

Définition : Ensembles *P-rounded*

Un ensemble P-rounded set $\mathcal{O}$ est composé de l'union des P meilleurs groupes arrondis IGP, dans l'ordre, où P est le nombre de groupes arrondis IGP permettant d'assurer la protection de la destination d.

Formellement, $\mathcal{O}$ s'écrit $\bigcup_{i=1}^{P} \mathcal{O}_i$, où $\mathcal{O}_i$ est un ensemble arrondi IGP, et $\mathcal{O}_i \prec \mathcal{O}_{i+1}$

Propriété : Un ensemble P-rounded est 1-OP

Concepts formalisés et prouvés dans le mémoire.
Intuitivement : protection se démontre par construction
optimalité se démontre par le lemme de β-étanchéité

En pratique, des ensembles 2-rounded (2R) sont souvent suffisants
Plus précisément : Si le réseau est bi-connexe, des ensembles 2R sont 1OP.

OPTIC - Contribution

Un cas simplifié : OPTIC 2R ($=, \prec, \succ$: comparaison lexicographique de β)

```
update_best_route(r^new, 𝕊):
  𝒢, 𝒲 = 𝕊_g(dest), 𝕊_w(dest)
  r^best = min(𝒢); r^worst = max(𝒢)   #meilleure & pire route
  IF |𝒢| ≥ 2 AND r^new ≻ r^worst THEN 𝒲 = 𝒲 ∪ r^new   #mis de coté
  ELSE IF r^new ≺ r^best:   #nouvelle meilleure route
    IF r^worst = r^best THEN 𝒢 = r^new ∪ 𝒢
    ELSE: 𝒢' = 𝒢 ∪ r^new; 𝒢 = {r^new, r^best}; 𝒲 = 𝒲 ∪ 𝒢' \ 𝒢;
  ELSE IF r^worst = r^best THEN 𝒢 = 𝒢 ∪ r^new   #route equivalente
  ELSE IF r^new ≺ r^worst THEN 𝒢' = 𝒢 ∪ r^new; 𝒢 = {r^best, r^new};
    𝒲 = 𝒲 ∪ 𝒢' \ 𝒢;   #2eme meilleure
  ELSE IF r^new = r^worst THEN 𝒢 = 𝒢 ∪ r^new;   #égale aux 2emes meilleures routes
  𝕊_g(dest), 𝕊_w(dest) = 𝒢, 𝒲
```

Théorème (Prouvé)

Si le réseau sous-jacent est bi-connexe, OPTIC maintient des ensembles 2R

Protection & Optimalité : prouvées
Efficacité : à évaluer

Ensembles 1-OP partagés

Les destinations peuvent pointer vers le même ensemble 1-OP

Mise à jour de la nouvelle meilleure passerelle

- Parcours de chaque ensemble pour l'installer
- Mise à jour de chaque groupe ≡ mise à jour de chaque destination

Complexité liée au nombre d'ensembles 1-OP distincts (# groupes) et à leur taille

Analyse de la complexité de mise à jour -Théorie

- Une complexité théorique $\mathcal{K}$ liée au nombre $\mathcal{N}$ de groupes

$$\mathcal{N} = min(\mathcal{P}, 2^B - 1), \text{ avec } \begin{array}{l} \mathcal{P} \text{ le nombre de préfixes} \\ B \text{ le nombre de routeurs de bordure} \end{array}$$

- Égale au pire à la somme des tailles des combinaisons de passerelles

$$\text{si } \mathcal{N} = 2^B - 1, \mathcal{K} = \sum_{i=1}^{B} \binom{B}{i} \times i = \underline{B \times 2^{B-1}}$$

Petits ensembles 1-OP $\Rightarrow$ moins d'ensembles 1-OP

Petits ensembles $\Rightarrow$ probabilité accrue de partage $\Rightarrow$ moins de groupes

- En pratique, des facteurs limitant la taille des ensembles
  - Nombre de routes valides/annoncées limité par **filtres**
  - Ensemble arrondis IGP : routes ayant le même β
    Si beaucoup de valeurs de β possibles, ensembles réduits

Analyse de la complexité de mise à jour -Théorie

Soit ps (*policy spreading*) le nombre de valeurs possibles de β.

- Si β distribué parfaitement uniformément, $\frac{B}{ps}$ routes ayant β minimal
- Création de tous les ensembles distincts de taille 2 à $\frac{B}{ps}$ si grand nombre de destinations (d'où, si $\frac{B}{ps} \geq 2$) :

$$\sum_{i=2}^{\frac{B}{ps}} \binom{B}{i}$$

En pratique, beaucoup de facteurs

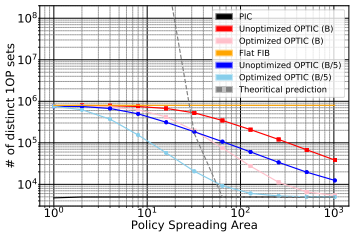
- Nombre de préfixes
- Nombre de routeurs annonçant les routes
- Distribution de β plus complexe et pas "parfaite"

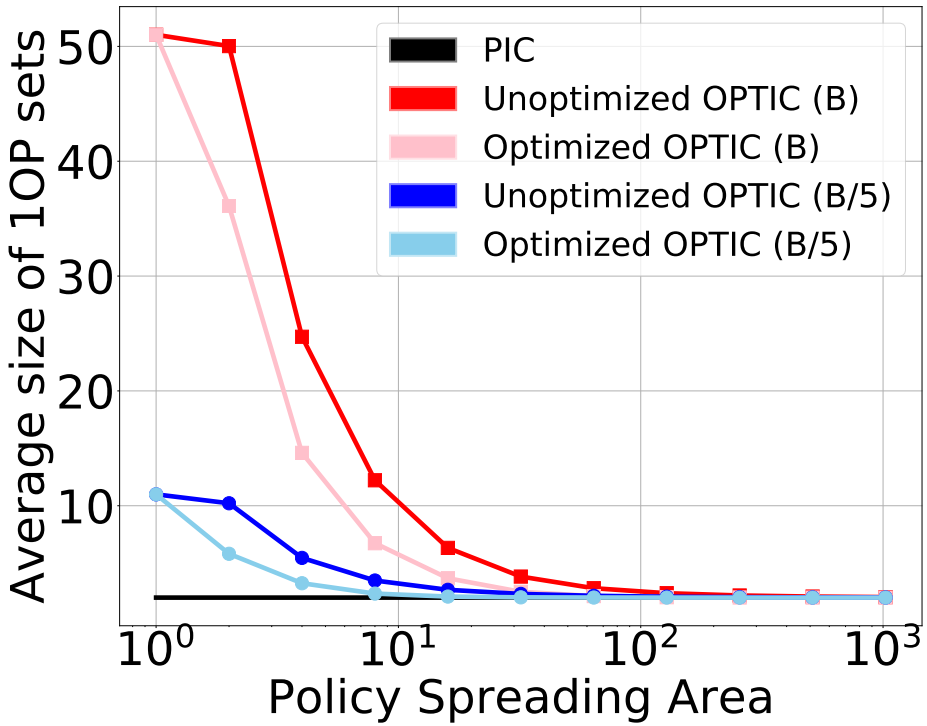
Simulateur python LenSim

- Explorer l'effet de différents paramètres sur la complexité
- Vérifier que l'ordre de grandeur du nombre de groupes OPTIC est viable

Analyse de complexité - Simulation

100 routeurs de bordures, 800000 préfixes, distributions uniformes

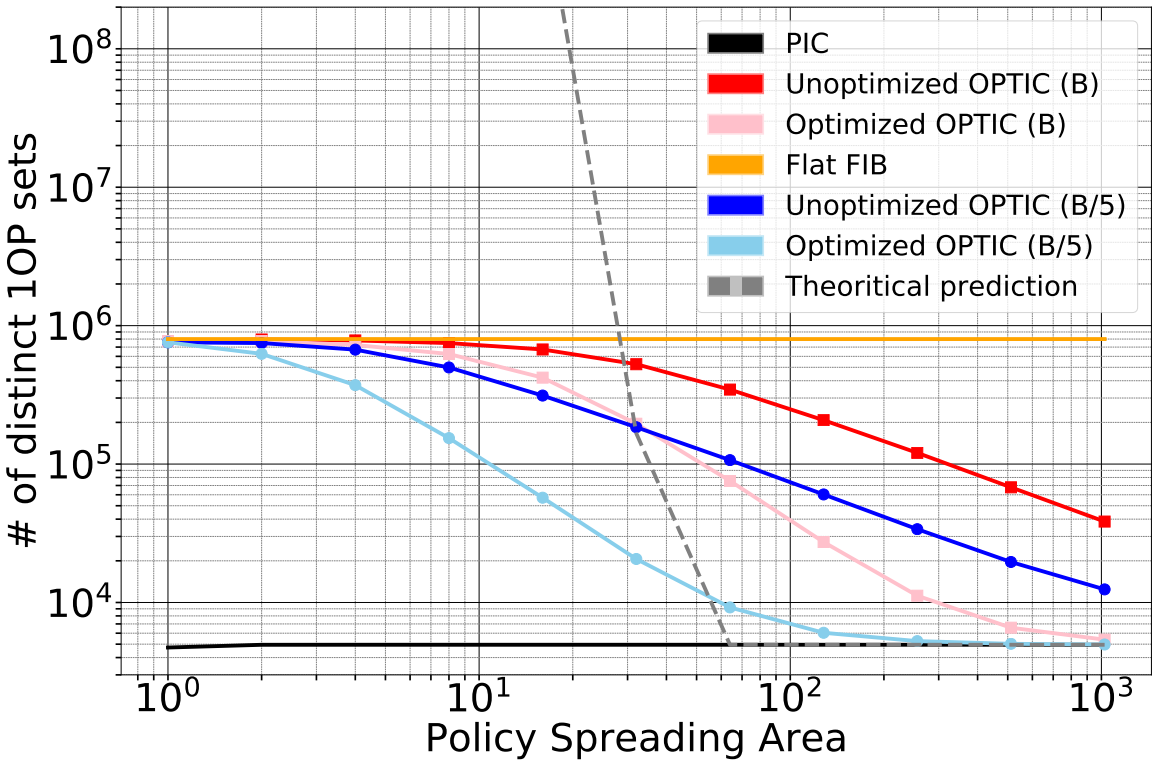


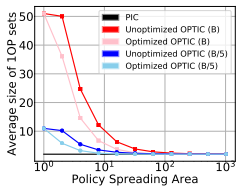


- Chute rapide de la taille des ensembles
- Taille minimale atteinte pour $ps = B$

Analyse de complexité - Simulation

100 routeurs de bordures, 800000 préfixes, distributions uniformes





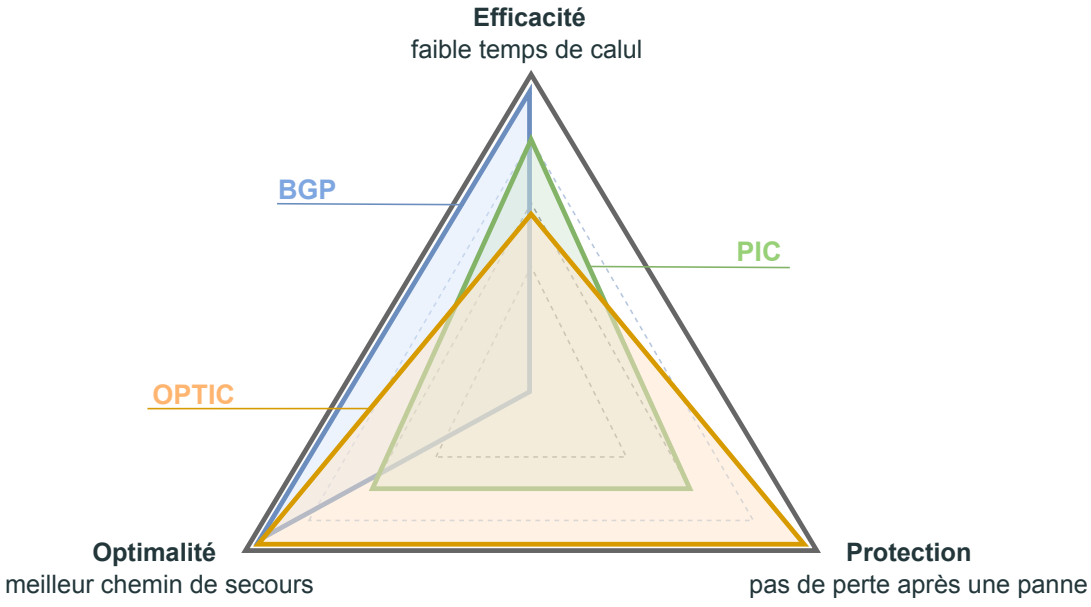
- Filtres & politiques ($\approx ps$) réduisent nettement la complexité
- Moins bien qu'espéré quand ps grand à cause du grand nombre de préfixes

Ne rejoint pas tout à fait PIC, mais **même ordre de grandeur**

Conclusion

OPTIC : Optimal Protection Technique for Intra-domain Convergence

- Développement d'un **cadre formel** pour la spécification d'OPTIC
- **Protection optimale** du trafic (**prouvée**)
- Une **efficacité** se rapprochant de l'existant (**évaluée**)



Perspectives

- D'autres ensembles 1-OP relachés et évaluations associées
- Amélioration du simulateur, données réelles, implémentation

OPTIC

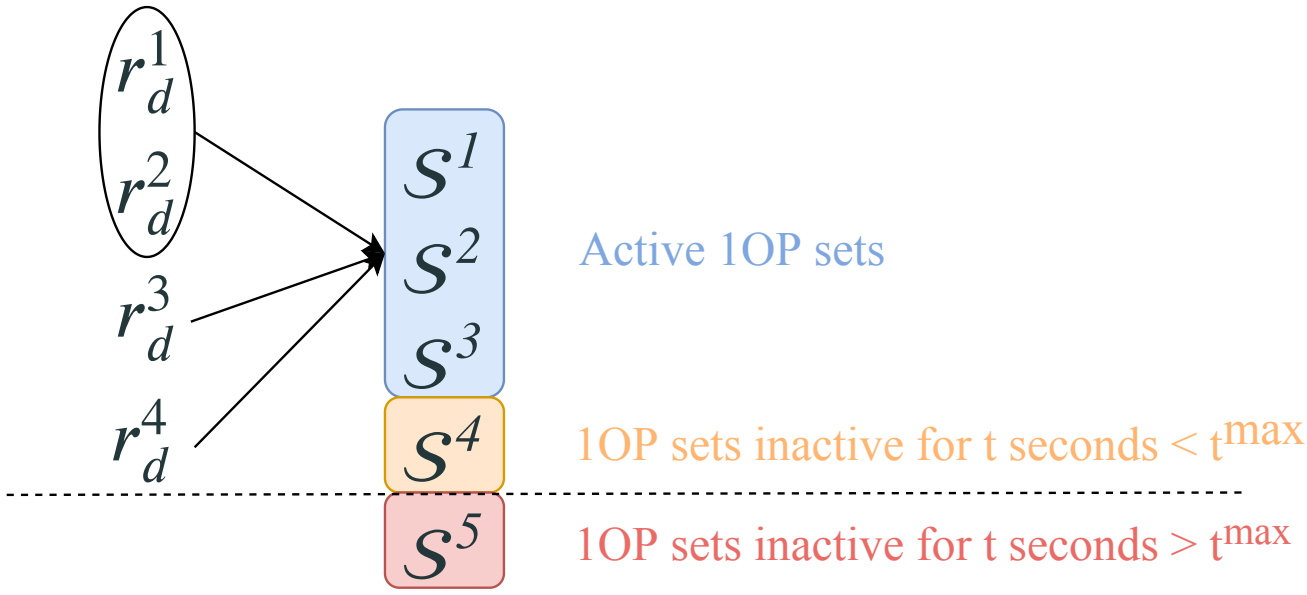
Optimal **P**rotection **T**echnique for **I**ntra-domain **C**onvergence

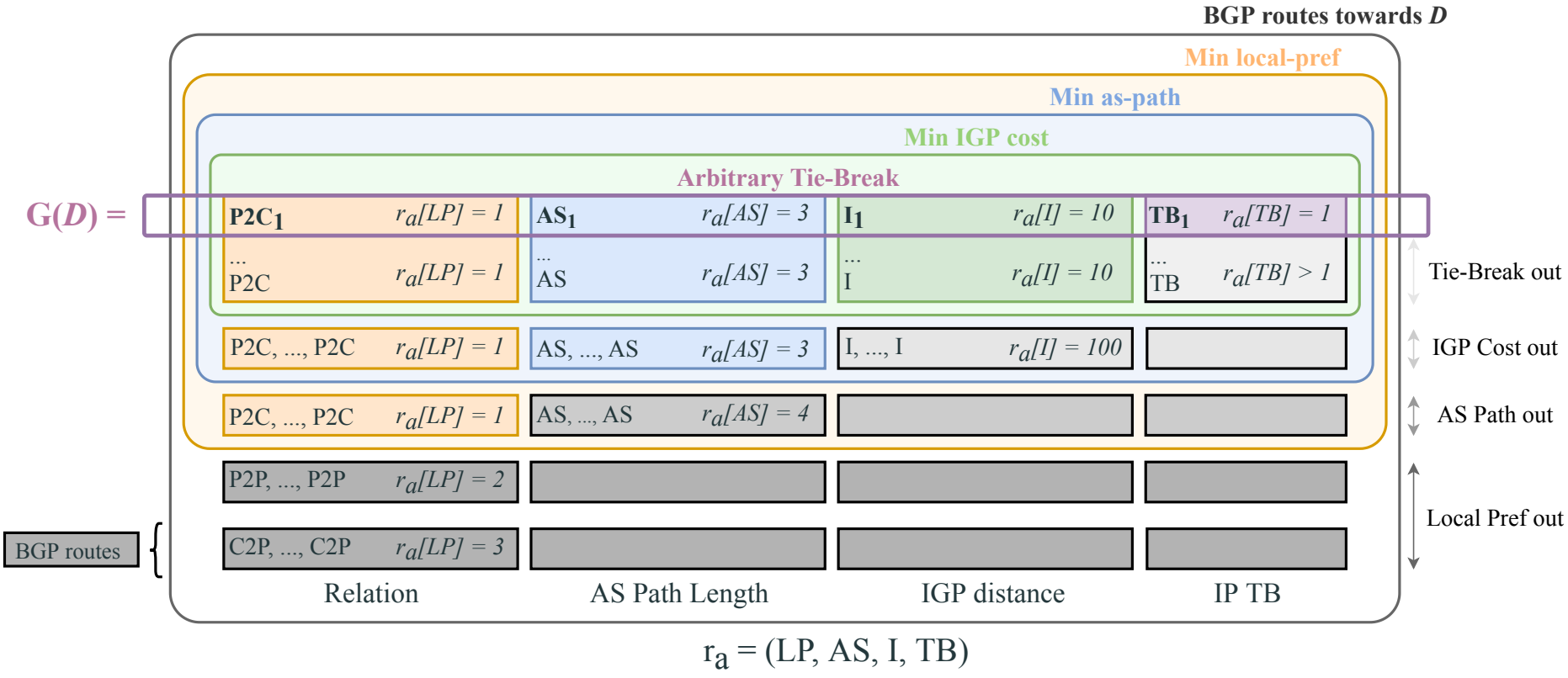
Jean-Romain Luttringer

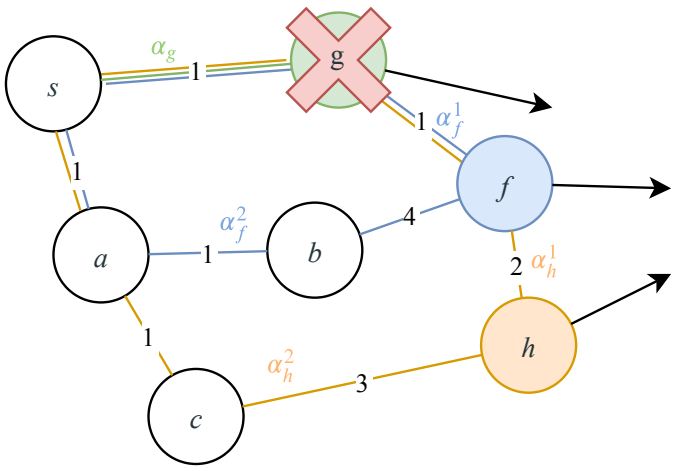
MERCI

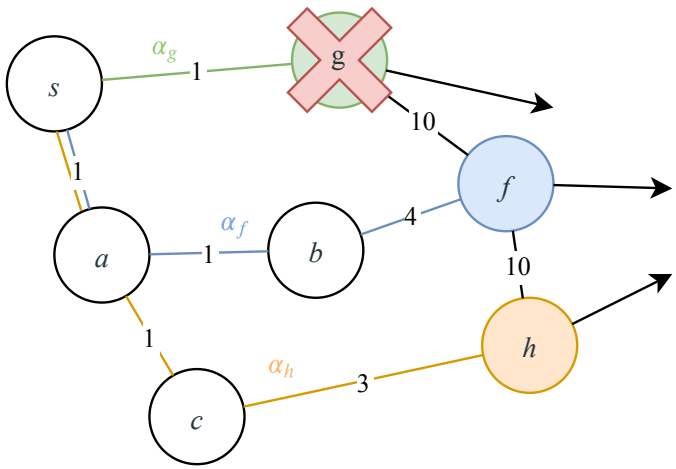
References

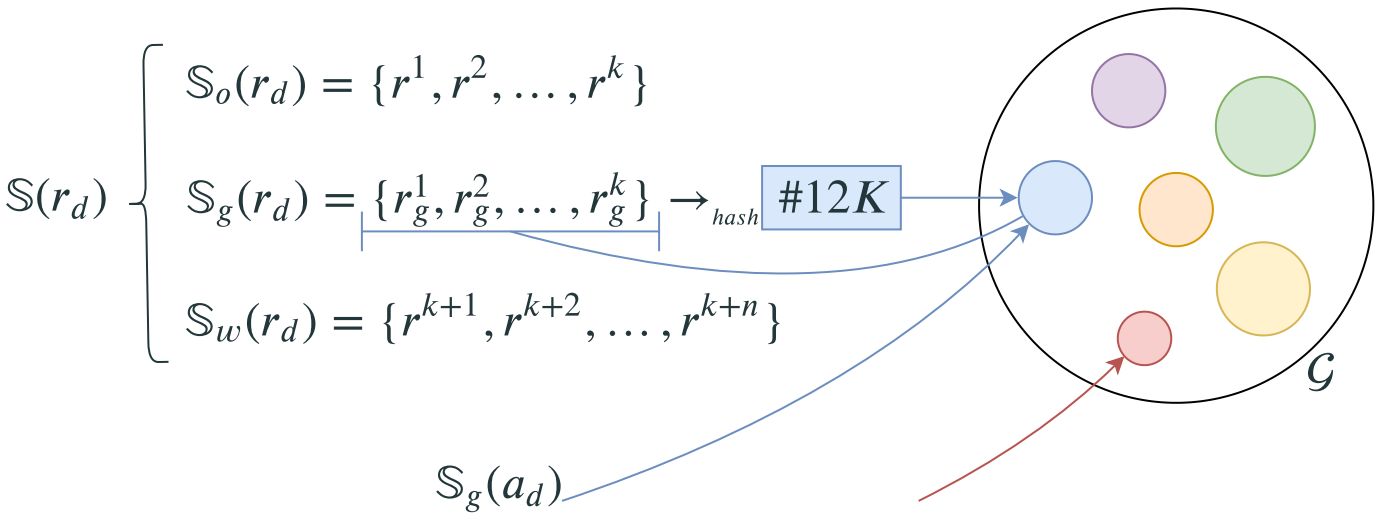
- Filsfils, Clarence et al. “BGP prefix independent convergence (PIC) technical report”. In: (2007).
- Holterbach, Thomas et al. “Blink : Fast Connectivity Recovery Entirely in the Data Plane”. In: *Nsdi* (2019).
- Labovitz, Craig et al. “Delayed Internet routing convergence”. In: *ACM SIGCOMM Computer Communication Review* 30 (2000).

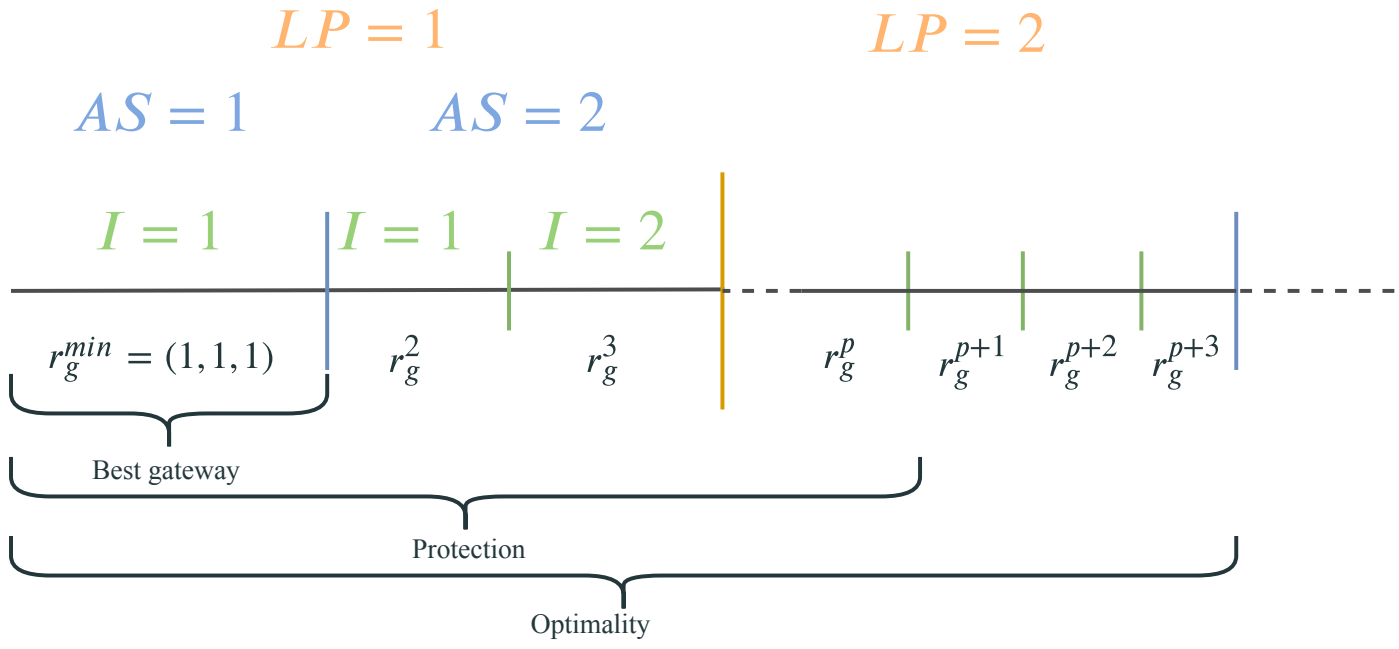

r_d^1
r_d^2
r_d^3
r_d^4
S^1
S^2
S^3
S^4
S^5
Active 1OP sets
1OP sets inactive for t seconds < t^{max}
1OP sets inactive for t seconds > t^{max}


BGP routes towards D
Min local-pref
Min as-path
Min IGP cost
Arbitrary Tie-Break
G(D) =
P2C1 ra[LP] = 1
AS1 ra[AS] = 3
I1 ra[I] = 10
TB1 ra[TB] = 1
... P2C ra[LP] = 1
... AS ra[AS] = 3
... I ra[I] = 10
... TB ra[TB] > 1
P2C, ..., P2C ra[LP] = 1
AS, ..., AS ra[AS] = 3
I, ..., I ra[I] = 100
P2C, ..., P2C ra[LP] = 1
AS, ..., AS ra[AS] = 4
P2P, ..., P2P ra[LP] = 2
C2P, ..., C2P ra[LP] = 3
BGP routes
Relation
AS Path Length
IGP distance
IP TB
Tie-Break out
IGP Cost out
AS Path out
Local Pref out
ra = (LP, AS, I, TB)


α_g
1
g
s
1
1
α_f^1
f
a
α_f^2
1
b
4
2
α_h^1
1
h
α_h^2
3
c


α_g
1
g
s
10
1
f
a
α_f
1
b
4
10
1
h
α_h
3
c


$\mathbb{S}(r_d)$
$\mathbb{S}_o(r_d) = \{r^1, r^2, \ldots, r^k\}$
$\mathbb{S}_g(r_d) = \{r_g^1, r_g^2, \ldots, r_g^k\} \rightarrow_{hash}$
#12K
$\mathbb{S}_w(r_d) = \{r^{k+1}, r^{k+2}, \ldots, r^{k+n}\}$
$\mathbb{S}_g(a_d)$
$\mathcal{G}$


LP = 1
LP = 2
AS = 1
AS = 2
I = 1
I = 1
I = 2
$r_g^{min} = (1,1,1)$
r_g^2
r_g^3
r_g^p
r_g^{p+1}
r_g^{p+2}
r_g^{p+3}
Best gateway
Protection
Optimality