Où un opérateur IP cache-t-il ses détours?

J Del-Fiore, P Mérindol, V Persico, C Pelsser, A Pescape

Où un opérateur IP cache-t-il ses détours?

J. M. Del Fiore[1] et P. Mérindol[1] et V. Persico[2] et C. Pelsser[1] et A. Pescape[2]

[1]*Laboratoire ICube, Université de Strasbourg, France*
[2]*University of Napoli Federico II, Italy*

La quantité de préfixes BGP à manipuler, ainsi que le nombre de routes à sélectionner, commencent à devenir colossaux pour les routeurs aux performances limités. Le nombre de préfixes se rapproche du million, ∼867K préfixes en Mars 2021, avec une augmentation de ≈50K préfixes par an sur les 10 dernières années. Pour pallier à ces problèmes d'extensibilité les systèmes autonomes (AS) peuvent tenter de filtrer certains préfixes, réaliser de l'agrégation ou bien recourir à des routes par défaut. Malgré leurs efficacités, ces astuces peuvent engendrer des détours de commutation (ou *forwarding detours*, FD), i.e., du traffic en transit acheminé via des routes internes à l'AS non optimales. Dans ce travail, nous étudions ce phénomène et proposons une méthode efficace pour détecter et analyser les FDs. En se basant sur un campagne de mesures réalisée depuis 92 moniteurs de l'infrastructure NLNOG RING, nous avons révélé que 25 ASes, sur les 54 suffisamment bien échantillonnés, semblent sujets, au moins partiellement, à de tels détours. En particulier, nous avons observé un motif binaire assez remarquable : pour un couple entrée/sortie d'un AS, soit tout le traffic de transit est détourné, soit aucun préfixe n'y est sujet.

Mots-clefs : Forwarding Detours, Load Balancing, Traffic Engineering, Network management, Scalability

1 Introduction

Over the last 8 years, the full Internet feed has doubled in size, reaching ∼867K prefixes as of Mars 2021 [cid]. The sustained increase in the number of prefixes advertised on the Border Gateway Protocol (BGP) has led Autonomous Systems (ASes) to suffer from scalability issues [512, 768]. Indeed, considering the current trend, maintaining a full Forwarding Information Base (FIB) may be challenging, specially for ASes incapable of upgrading their network devices regularly [*10*].

In this context, networks operators have found alternatives to endure with legacy routers unable to maintain a complete FIB in memory. Among the different options, one consists in making them store partial-FIBs [*09a*, *12*] with default routes redirecting traffic towards more capable routers (e.g. having a full-FIB). While the aforementioned workaround may look effective at first glance, ASes relying on this technique may suffer from *forwarding detours* (FDs), i.e., for some aggregated prefixes, traffic may traverse the network across AS border routers (ASBRs) through sub-optimal paths regarding the cost metric used by the Internal Gateway Protocol (IGP). Generally speaking, and more precisely, FDs occur between two points of an AS when there exists at least one prefix for which traffic flows through a forwarding route not included in the set of paths used for load balancing (LB) or traffic engineering (TE). This prefix is subject to FDs between these two points and we say that the route leading to FD is a detouring one.

In this study we take a close look at this phenomenon. Contrary to hot-potato routing, FDs increase the IGP distance required to traverse an AS, arguably resulting in waste of resource utilization inside the network. As discussed before, FDs may result as a side effect of scalability workarounds. However, misconfigurations or bugs in router's software may also lead to such detours. Prior work has focused on detecting routers relying on backup default routes [*09b*], or identified them as a possible cause of BGP lies [*19*]. However, no study has focused on the impact of such techniques on the forwarding inside ASes. In that sense, to the best of our knowledge, we are the first to tackle the problem of detecting FDs.

2 Problem statement : partial FIBs lead to Forwarding Detours

Fig. 1 illustrates how the use of a partial-FIB router with a default route, here on $ASBR_1$, may result in FDs inside an AS. In this example, with its default route pointing to $ASBR_2$, $ASBR_1$ forwards traffic

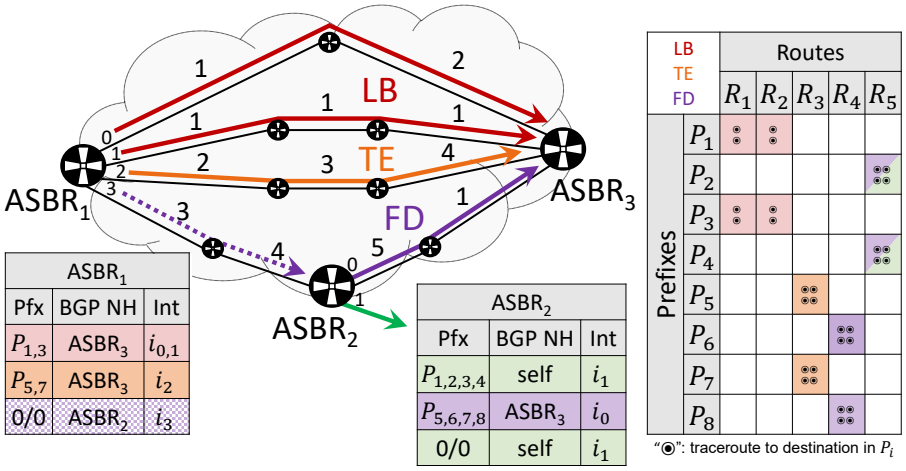

FIGURE 1: From partial-FIB routers to FDs, and the challenge to discriminate them from LB and TE. Path colors highlight routes whereas the numbers on top of the links report IGP weights.

destined to the set of prefixes $\{P_2, P_4, P_6, P_8\}$ towards $ASBR_2$ (dotted violet line). Having a full-FIB, $ASBR_2$ acts as exit point for $\{P_2, P_4\}$, but redirects traffic aiming $\{P_6, P_8\}$ towards $ASBR_3$. While traffic concerning $\{P_2, P_4\}$ flows through the best path connecting $ASBR_1$ and $ASBR_2$, packets targeting $\{P_6, P_8\}$ entering the AS via $ASBR_1$ flow via route R_4 and finish exiting the AS via $ASBR_3$. However, the set of LB and TE routes between $ASBR_1$ and $ASBR_3$ are $\{R_1, R_2\}$ and $\{R_3\}$, that are actually used for prefixes $\{P_1, P_3\}$ and $\{P_5, P_7\}$ respectively. Therefore, the set of prefixes $\{P_6, P_8\}$ is subject to FDs and route R_4 is a detouring route.

Currently, there exists no methodology allowing to determine, leveraging only data-plane information obtained with traceroute, whether FDs occur between any two routers i and e of an AS. Considering $i = ASBR_1$ and $e = ASBR_3$ in Fig. 1, we would like to correctly determine that R_4 is a detouring route, and that the set of prefixes $\{P_6, P_8\}$ is subject to FDs. To carry out this task, the challenge is avoiding to (mis)classify LB and TE paths as FDs. Last but not least, our goal is to design an FD-detector only running `traceroute`, i.e., not requiring privileged access to verify the configuration of the distributed forwarding devices.

3 The first FD-detector or Discriminating FD from LB and TE

In this paper, we propose the first method to detect FDs. Our FD-detector first explores the forwarding pattern inside an AS, identifies the set of LB paths and concludes whether FDs occur or not relying on a simple FD-verdict. The methodology we discuss next is a simplified version of that described in [DFPM+21].

3.1 Forwarding pattern and merging-phase

To detect FDs between two points (i, e) in an AS X, we analyze the **forwarding pattern** collected for (i, e); in other words, which routes in AS X, leading from i to e are revealed when targeting different prefixes with traceroute. When prefixes are subject to destination-based LB, multiple destinations inside each prefix need to be explored. For each of them, different LB routes may be discovered. This is equivalent to seeking for the matrix observed in Fig. 1 where, in practice, the rows concerning prefixes P_2 and P_4 would not be included since for these prefixes, traffic actually traverses $(ASBR_1, ASBR_2)$ and not $(ASBR_1, ASBR_3)$.

Once the forwarding pattern for (i, e) is found, we apply a **merging-phase** : we group prefixes for which same set of routes $\mathcal{R}_i$ are revealed into sets of prefixes $\mathcal{P}_i$. Relying on Fig. 1, then $\mathcal{P}_1 = \{P_1, P_3\}$, $\mathcal{P}_2 = \{P_5, P_7\}$ and $\mathcal{P}_3 = \{P_6, P_8\}$ associated to $\mathcal{R}_1 = \{R_1, R_2\}$, $\mathcal{R}_2 = \{R_3\}$ and $\mathcal{R}_3 = \{R_4\}$ respectively. From a mathematical point of view, this step is thus equivalent to generating a partition $\mathcal{S}$ of the prefix-route space for (i, e), such that $\mathcal{S} = \bigcup_i (\mathcal{P}_i, \mathcal{R}_i)$. In practice, while gathering the forwarding pattern of (i, e), only a sub-set may actually be discovered out of all routes associated to a prefix. Therefore, to generate $\mathcal{S}$, we continuously merge the intersecting sets of routes until no more overlaps exist among them. For each union,

we also merge the sets of prefixes associated to merged routes.

3.2 Collecting the DIR : identifying the LB set of prefixes and routes

Once the partition $\mathcal{S}$ of (i,e) is found, it is actually possible to determine both the set of prefixes subject to LB and their routes. This can be done collecting the **direct internal route (DIR)**, i.e., the route inside X from i to e obtained by running traceroute towards e. The DIR, denoted D, is particularly important since, by definition, it belongs to the LB routes of (i,e). Recalling Fig. 1, the DIR of $ASBR_3$ will either be $D = R_1$ or $D = R_2$, both included in $\mathcal{R}_1$, thus allowing to determine that $(\mathcal{P}_1, \mathcal{R}_1)$ is associated to LB. However, there are cases where the DIR may not be included in any set, e.g. when all prefixes are subject to FDs. When this happens, i.e., $\forall(\mathcal{P},\mathcal{R}) \in \mathcal{S}$, $D \notin \mathcal{R}$, we proceed to add $(D, e/32)$ to $\mathcal{S} : \mathcal{S} = \mathcal{S} \cup (D, e/32)$.

The networking rationale behind such an assumption is that internal prefixes of ASes, such as the internal destination e of AS X, are not subject to FDs. In other words, regarding internal destinations, it is reasonable to assume that all devices are full-FIB routers. Indeed, since the IGP does not suffer from similar scalability issues as BGP does, all internal prefixes are expected to be installed in all routers. Since IGP prefixes constitute the backbone of an AS, removing them from the FIB of any router would represent a minor scalability gain while letting BGP running on top of a flawed IGP network. The DIR is not expected to detour, it should be one of the best IGP paths, which by definition is included in the LB set of routes.

3.3 FD-verdict : focusing on extreme-FDs

Analyzing the partition $\mathcal{S}$ of (i,e), considering $s = |\mathcal{S}|$ the total number of sets, $Q = \sum_{\forall \mathcal{P}_i \in \mathcal{S}} |\mathcal{P}_i|$ the total number of prefixes, and $Q_{LB} = |\mathcal{P}|$ such that $e/32 \in \mathcal{P}$ the number of prefixes subject to LB, we conclude that FDs occur when $Q_{LB}/Q < 1/s$. The reasoning is as follows. In the absence of FDs between (i,e), we expect most prefixes to be associated to LB and only a few to TE, since deploying the latter at a large scale requires a significant control plane overhead. Hence, in these cases, $Q_{LB} \simeq Q$ and thus $1/s$ will usually be largely lower than Q_{LB}/Q (when TE is deployed). On the other hand, when FDs occur, a priori they may affect any number of prefixes. If few prefixes are subject to FDs, then most prefixes are associated to LB. Therefore, in these cases we would likely introduce false negatives. However, in the event of **extreme-FDs**, most prefixes are subject to FDs, leaving a few associated to LB. In these scenarios, we expect to see $Q_{LB} << Q$ as it enables an easy effective threshold.

4 Results : FDs exist and look extreme !

We run measurements from 92 vantage points (VPs) of the NLNOG RING monitoring infrastructure [nln] towards a list of 100K /24 prefixes (one IP address per prefix) on May 26th 2020 and look for FDs for each (i,e) ASBR-couples of an AS (i and e being ASBRs of the same AS). We focus on cases where the limits between ASes are clear and thus the ASBR-couples can be unambiguously identified. We discard ASBR-couples for which the DIR cannot be collected or the number of prefixes in the partition (Q) is less than 100 prefixes. Our FD-detector was able to analyze 3963 ASBR-couples spanning 54 ASes.

We find extreme-FDs in 25 ASes, across 168 ASBR-couples. Fig. 2 shows the breakdown per AS of the 168 ASBR-couples subject to FDs, sorted by increasing relative fraction across ASes. We observe no general trend, indicating that the prevalence of FDs is AS-specific, e.g. depending on both router's hardware and OSes in use. Even though most ASes have few measured couples with FDs, less than 10 in general, the relative values spawn from as low as almost 0% to up to 100%.

4.1 The binary effect of FDs

We are interested in determining the forwarding patterns we found for the 3963 ASBR-couples (i,e) in our dataset. Analyzing the partitions $\mathcal{S}$, we see that for 96% of the couples it holds that $s = 1$, i.e., all prefixes were associated to the DIR ($Q_{LB} = Q$). Exploring further the cases where $s = 1$, we see that in 52% of them the DIR is the unique route used between (i,e), and in the remaining 48% there are other LB routes. For the remaining 4% of ASBR-couples, except for a few exceptions, $s = 2$. When $s = 2$, we would conclude that FDs occur when $Q_{LB} < Q/2$. However, in all cases we actually see that 0% of the prefixes are associated to the DIR ($Q_{LB} = 0$). In other words, we see a remarkable on/off pattern : all measured transit traffic that

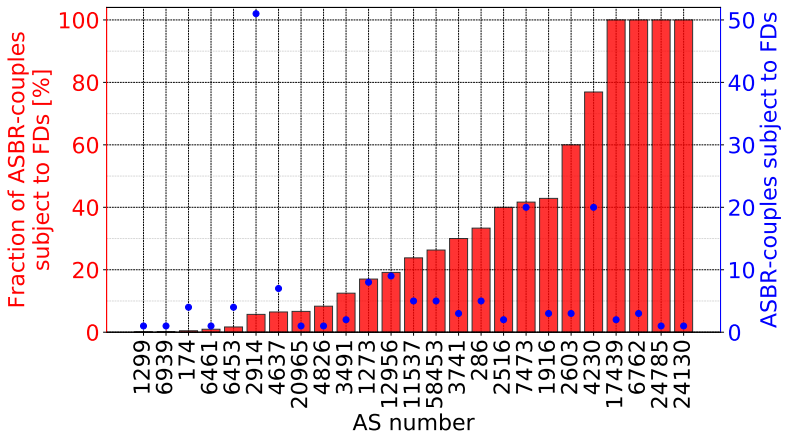

FIGURE 2: Quantification of ASBR-couples subject to FDs per AS.

traverses any ASBR-couple either always detours, or never does. As there are no gray regions, i.e., either $s = 1$ or $s = 2$ with 100% or 0% of the prefixes respectively associated with the DIR, this shows that the threshold $1/s$ described in the FD-verdict has no influence and our results does not seem to include any false positives or negatives.

5 Conclusions and future work

In this paper, we propose a method to detect detours within an AS. More precisely, we show that it is possible to discriminate FDs from LB and TE in cases of multiple prefixes subject to FDs. In a nutshell, we study the forwarding pattern between couple of ASBRs and collect the DIR tracing the exit ASBR. To the best of our knowledge, we are the first to tackle this problem. We build an FD-detector and, using large-scale measurement campaigns, we show that around 50% of the well sampled ASes in our dataset suffer from FDs. In addition, our analysis provides a notable takeaway, since we systematically observe a binary effect : for each affected couple of ASBRs, either all prefixes we measured were subject to FDs, or none were. In future work, we aim to shed light on the detrimental effects that FDs have on routing performance.

Références

[512] What caused today's Internet hiccup. https ://www.bgpmon.net/what-caused-todays-internet-hiccup/.

[768] 768k Day. Will it Happen ? Did it Happen ? https ://labs.ripe.net/Members/emileaben/768k-day-will-it-happen-did-it-happen.

[cid] *CIDR-REPORT* Status summary. `https://www.cidr-report.org`.

[DFPM+21] Julián M. Del Fiore, Valerio Persico, Pascal Mérindol, Cristel Pelsser, and Antonio Pescapè. The Art of Detecting Forwarding Detours. *IEEE Transactions on Network and Service Management*, 2021.

[*09a*] Ballani *et al.* Making Routers Last Longer with ViAggre. In *NSDI*, volume 9, pages 453–466, 2009.

[*09b*] Bush *et al.* Internet Optometry : Assessing the Broken Glasses in Internet Reachability. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09. ACM, 2009.

[*10*] Zhao *et al.* Routing scalability : an operator's view. *IEEE Journal on Selected Areas in communications*, 28(8) :1262–1270, 2010.

[*12*] Karpilovsky *et al.* Practical network-wide compression of IP routing tables. *IEEE Transactions on Network and Service Management*, 9(4) :446–458, 2012.

[*19*] Del Fiore *et al.* Filtering the Noise to Reveal Inter-Domain Lies. In *2019 Network Traffic Measurement and Analysis Conference (TMA)*, pages 17–24, June 2019.

[nln] *NLNOG RING* monitoring infrastructure. `https://ring.nlnog.net`.