The Scope of the IBGP Routing Anomaly Problem

Uli Bornhauser[1], Peter Martini[1], and Martin Horneffer[2]

1 **University of Bonn – Institute of Computer Science 4**
Roemerstr. 164 – D-53117 Bonn – Germany
{ub, martini}@cs.uni-bonn.de

2 **Deutsche Telekom AG – Technical Engineering Center**
Hammer Str. 216 – D-48153 Muenster – Germany
martin.horneffer@telekom.de

Abstract

Correctness problems in the iBGP routing, the de-facto standard to spread global routing information in Autonomous Systems, are a well-known issue. Configurations may route cost-suboptimal, inconsistent, or even behave non-convergent and -deterministic. However, even if a lot of studies have shown many exemplary problematic configurations, the exact scope of the problem is largely unknown: Up to now, it is not clear which problems may appear under which iBGP architectures.

The exact scope of the iBGP correctness problem is of high theoretical and practical interest. Knowledge on the resistance of specific architecture schemes against certain anomaly classes and the reasons may help to improve other iBGP schemes. Knowledge on the specific problems of the different schemes helps to identify the right scheme for an AS and develop workarounds.

1998 ACM Subject Classification Computer-Communication Networks: Network Protocols – Protocol verification, Routing protocols

Keywords and phrases Inter-domain Routing, Border Gateway Protocol, Routing Anomalies

Digital Object Identifier 10.4230/OASIcs.KiVS.2011.2

1 Introduction

The Internet we know today is a system of independent, interconnected Autonomous Systems (ASs). The Border Gateway Protocol (BGP) [12] is the de-facto standard used to exchange global routing information between ASs. The process of spreading global routing information within ASs is usually implemented via internal BGP (iBGP), a particular operational mode of BGP. However, even if iBGP is widely used, it may come along with correctness problems.

1.1 Motivation and Objectives

While classical iBGP is based on a logical full-mesh, expected scalability problems led quickly to the development of alternative information exchange schemes. Two of these schemes, BGP Route Reflection [2] and AS Confederations [14], were finally standardized, implemented, and frequently deployed. Today, it is well known that these schemes are prone to serious correctness problems [6]. Various studies, cf. [6, 9], for example, have shown different classes of conflicts with basic correctness properties; *iBGP anomalies*. However, even if the causes are known [5], the scope of the problem is still unknown: Are iBGP anomalies a direct consequence of using scalability techniques or may similar problems also appear if full-mesh iBGP is used? What causes may appear under which iBGP information exchange schemes? Is – and how is – it possible to address the causes with operational means, i.e. without any protocol modifications? Adequate answers are of high theoretical and practical interest: They help protocol designers to combine the advantages of different schemes and support operators addressing anomalies.

17th GI/ITG Conference on Communication in Distributed Systems (KiVS'11).
Editors: Norbert Luttenberger, Hagen Peters; pp. 2–13
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1.2 Directly Related Work

In 2000, studies of Cisco Systems [6] and McPherson et al. [12] showed that the iBGP routing is prone to consistency problems in general. Two years later, Griffin et al. [9] could show that consistency problems are only the tip of the iceberg: They presented a series of simple and largely realistic configurations revealing further correctness problems. In the same year, Basu et al. [1] proved that at least verifying the consistency of an iBGP configuration is NP-hard. Based on this insight, different concepts which try to alleviate or avoid the observed problems were developed in the following years. However, since none of them can be used unconditionally in production systems, iBGP anomalies are still an issue today. An interesting aspect of all previous studies is that anomalies are usually understood as a side-effect of an information reduction compared to fully meshed iBGP [10, 11]. Some studies even go that far to implicitly assume full-mesh iBGP as inherently anomaly-free [7] – without any proof.

Even if the root causes for iBGP anomalies are well-known [5], the exact scope of the problem was never studied. Today, it is known that scalability techniques facilitates effects inducing anomalies. However, a complete, structured overview of the problem does not exist. Problems in full-mesh iBGP are unknown so far. This is the starting point for our analyses.

1.3 Paper Outline

The rest of the paper is organized as follows. At first, in section 2, we summarize the most important aspects of iBGP relevant in the context of this paper. In section 3, we introduce the problems that may come along with iBGP and sketch their root causes. Knowing the different anomaly classes, in section 4 we study the standardized iBGP schemes in terms of their vulnerability to the anomaly classes. In section 5, we give an outlook on possible workarounds and solutions. Finally, we close with a short conclusion and aspects for future work in section 6.

2 The Border Gateway Protocol

The Border Gateway Protocol is a classical path vector protocol. In this section, we outline the main aspects of BGP that are relevant in the context of the following analyses. Nevertheless, readers of this paper should be familiar with internal BGP, BGP Route Reflection, and AS Confederations. Details on iBGP and its exchange schemes may be found in [2, 12] and [14].

2.1 Basics on the Border Gateway Protocol

Today, BGP is the de-facto standard Exterior Gateway Protocol. *BGP speakers* forward traffic via a unique *best path*. This path is determined by applying the BGP *path selection function* on all known paths. Paths are known due to *local configuration* and *advertisements* of system-*external* and -*internal BGP peers*. BGP peers only advertise their best path and only if this path is not already known. To be able to focus on iBGP issues, we assume that the external and local paths each speaker knows are time-invariant. Traffic forwarding via a path is realized via the *BGP next-hop* it specifies. Under common edge conditions (as assumed in this paper), it specifies the router where the path is externally learned or locally configured. Traffic that reaches this router is correctly forwarded out of the AS. The forwarding to the BGP next-hop is realized on a hop-by-hop basis via the shortest Interior Gateway Protocol (IGP) route.

A *BGP path* is defined by *path attributes*. The Local Preference, AS-path Length, Origin Code, Multi Exit Discriminator (MED), and peer-AS are *global* path attributes. In contrast to *local attributes*, they are equal on every router within an AS. They do not depend on the topological position within the AS. Based on the path attributes, paths can be compared.

2.2 The BGP Path Selection Process

Let π denote the paths known by a BGP speaker v for a certain destination. After removing unresolvable paths from π, v's best path is determined by the following selection process. If π is empty, no best path for the destination exists on router v.

Step 0) Remove all paths from π that are not tied for having the lowest **Local Preference**.
Step 1) If any path is **locally configured** on v, remove all non-local paths from π.
Step a) Remove all paths from π that are not tied for having the shortest **AS-path Length**.
Step b) Remove all paths from π that are not tied for specifying the lowest **Origin Code**.
Step c) Remove all paths from π that do not have the best **MED** for their peer-AS group.
Step d) If paths are **externally learned**, remove all internally learned paths from π.
Step e) Remove all paths from π that are not tied for having the lowest **IGP distance** to the specified BGP next-hop.
Step f) Remove all paths from π which are not tied for specifying the lowest **BGP ID**.
Step g) Choose and return the path from π that specifies the lowest **Peer-address**.

An important aspect of the selection process is the fact that MEDs can only be compared if the paths specify the same peer-AS. The peer-AS specifies the source of a routing information. It can be derived from the AS-path attribute, cf. [12]. As MEDs can only be compared within a peer-AS group, *path rankings* may behave non-transitive: For arbitrary paths p, q, and r, it may hold that $p<q$, $q<r$, and $r<p$, where $<$ is the order defined by the selection process, cf. figure 5.a). Details on the other BGP path attributes may be found in [12].

2.3 Terms and Identifiers

To establish a common language, a few terms and identifiers are introduced at first. A path, i.e. the external or local BGP routing information a router knows, is denoted as *BGP path.* A BGP path is propagated via *signaling paths* according to the applied iBGP scheme through an AS. The tuple of BGP path and BGP next-hop specifies all forwarding information a signaling path contains. It is denoted as *delivery path.* A signaling path resolvable at router v that is propagated according to the applied iBGP scheme is called *permitted* at $v \in V$. The set V covers all routers within the AS. The set of paths permitted at v is written as $\mathcal{P}^v$. The selection process of BGP defined above is formalized by the *selection function* λ. λ is always applied for the local view of a router v, written as λ^v, to a subset of its permitted paths. If λ^v is applied up to and including step e) to all paths $\mathcal{P}^v$ permitted at v, the router's *cost-optimal paths* $\Omega_c^v = \lambda^v(\mathcal{P}^v)$ are calculated. They optimize the forwarding costs. The definition stresses the fact that the BGP ID and Peer-address are only tie-breaker criteria. We label signaling paths usually with the letters p or q. A signaling path p permitted at v is usually identified as p_v. The underlying delivery path is labeled as $\dot{p}_v$, the BGP path as $\bar{p}$.

Once traffic reaches the BGP next-hop, it leaves the system by definition. The forwarding to the BGP next-hop is realized via the AS's IGP. The IGP *forwarding route* from v to the BGP next-hop a delivery path $\dot{p}_v$ specifies is labeled as $f_{v \to \dot{p}_v}$. The route without the BGP next-hop (the last physical hop on the IGP route) is written as $f^1_{v \to \dot{p}_v}$.

In what follows, we study the BGP routing in an *iBGP configuration.* Such a configuration is defined by the routers V, the physical network topology in the AS, the BGP paths available at each $v \in V$, and the AS's iBGP peering structure. The routing decision v has made is defined by the delivery path the selected best path represents. The tuple $s = (\dot{p}_1, \ldots, \dot{p}_n) \mid 1 \leq i \leq n, i \in V$ of delivery paths specifies the *state* a configuration has entered. Over time, the configuration moves from state to state until a *stable* state is *entered*: The delivery paths the routers select do not change any more. The set of all states the configuration may enter is written as S.

3 IBGP Routing Anomalies and their Root Causes

Even if the specific demands on routing protocols and information exchange schemes generally differ, correctness specifies a universally valid requirement. Details on this requirement and the causes for the problems iBGP comes along with are discussed in what follows.

3.1 Correctness Properties and Routing Anomalies

The correctness of a routing protocol is determined by three basic properties: *Expressiveness*, *Global Consistency*, and *Robustness* [5, 7, 8]. A protocol behaves correct, if these properties are fulfilled in arbitrary configurations that match the protocol specification. A definition of the three correctness properties is given in what follows. Details may be found in [5].

3.1.1 Expressiveness

Expressiveness is the first correctness property. It states that the routing implements traffic forwarding according to the “natural forwarding costs”. This means that if a destination is reachable via a policy-conform path, traffic forwarding is implemented via a *cost-optimal path*. The cost-optimal paths are determined by the distance metric the protocol uses. As specified for iBGP, cf. section 2.3, they remain after applying the cost-relevant steps of the protocol’s path selection process to all available paths. Tie-breaker rules are cost-irrelevant. If a router does not comply with expressiveness, it routes *cost-suboptimal*, cf. figure 1.a). Cost-suboptimal routing decisions lead to unnecessary costs and may even foil traffic transition arrangements like – in case of BGP – defined by the MED.

3.1.2 Global Consistency

Global consistency is an essential property for hop-by-hop based routing protocols. It states that the local routing decisions routers make are free of conflicts: If traffic is forwarded by a hop to another to reach a destination via a certain route, the next-hop has to continue the intended route. If a next-hop on an intended route redirects traffic onto another route or to another destination, *inconsistency* is present, cf. figure 1.b). Local policies may be overridden and – in special circumstances [9] – forwarding loops may be induced.

3.1.3 Robustness

Robustness means that the local processes the protocol defines always lead to a stable state. The routing must converge. The state the configuration finally enters must be predictable in advance if the initial state and configuration are known. The routing processes must behave deterministic. If the routing may not converge, it is *vulnerable to divergence*. If the final state is not predictable, the routing is *vulnerable to non-determinism*, cf. figure 1.c). Divergence and non-determinism may cause paket re-orderings, losses, and complicate routing analyses.

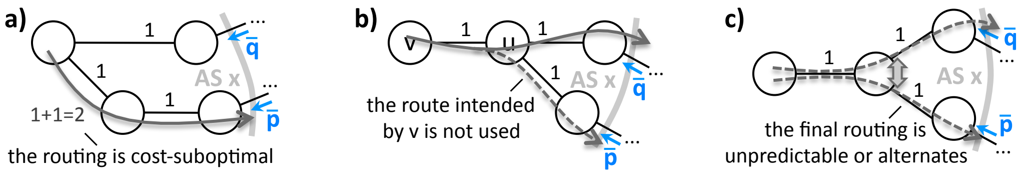

a)
1
1
1
1+1=2
AS x
$\bar{q}$
$\bar{p}$
...
...
the routing is cost-suboptimal
b)
v
u
1
1
1
AS x
$\bar{q}$
$\bar{p}$
...
...
the route intended by v is not used
c)
1
1
1
AS x
$\bar{q}$
$\bar{p}$
...
...
the final routing is unpredictable or alternates

Figure 1 Conceptual illustration of cost-suboptimal routing, inconsistency, and non-determinism.

3.2 IBGP Anomalies and their Root Causes

Using the identifiers defined in section 2.3, anomalies in the iBGP routing can be traced back to six *root causes*, cf. [5]. They are introduced in what follows. Note that in contrast to the generic definitions given in section 3.1, the root causes are specific for the protocol.

3.2.1 Cost-suboptimal Routing

Obviously, it only makes sense to assess a routing decision as cost-suboptimal if the decision is permanently. This is the case, if the configuration has entered a stable state $s \in S$. In this state, every router $v \in V$ selects its best path by applying the selection process to the known paths. Consequently, cost-suboptimal routing is present in state s, if and only if

$$\exists v \in V : \lambda^v(\pi_s^v) \notin \Omega_c^v,$$

where π_s^v denotes the paths known in state s. Due to the design of BGP, this may have two reasons: Obviously, a router cannot select a cost-optimal path as best one if no such path is known. If such paths are known, they may be masked since no transitive ordering relation is defined, cf. figure 5.a). Both root causes can formally be differed as

$$v \in V : (\pi_s^v \cap \Omega_c^v) = \emptyset \quad (1) \qquad \text{and} \qquad v \in V : (\pi_s^v \cap \Omega_c^v) \neq \emptyset \wedge \lambda^v(\pi_s^v) \notin \Omega_c^v \quad (2).$$

3.2.2 Inconsistency

Similar to cost-suboptimal routing, referring to inconsistency only makes sense if a configuration has entered a stable state $s \in S$. By assumption, traffic is correctly delivered once it has reached the BGP next-hop, cf. section 2.1. Inconsistent local views may thus only appear on the IGP route to the BGP next-hop: A router may take away the traffic from the intended route and redirects it to another BGP next-hop. Another delivery path is used. It holds that

$$\exists v \in V, u \in f^1_{v \rightarrow \dot{p}_v} \setminus \{v\} : \dot{p}_v \neq \dot{p}_u,$$

where $p_x = \lambda^x(\pi_s^x)$ denotes the best path of and π_s^x the paths known by router $x \in \{v, u\}$ in state s. If inconsistency is not a side-effect of a cost-suboptimal routing, it is caused by *dissenting local views*. Dissenting views of v and u may have two reasons: Firstly, they may appear due to a simple lack of routing information required for consistent routing decisions. Provided with adequate paths, u and v choose consistent delivery paths. The basic problem is illustrated in a simplified manner in figure 2.a). Secondly, inconsistent views may be a result of *dissenting path rankings*. The BGP paths are not equivalently preferred by v and u. Induced by cost-relevant local path attributes that behave non-isotonicly [13] on a forwarding route, additional routing information is not helpful, cf. figure 2.b). Both root causes can be formalized as

$$\begin{array}{ccc}
\forall p_v, q_v \in \mathcal{P}^v, p_u, q_u \in \mathcal{P}^u \mid u \in f^1_{v \rightarrow \dot{p}_v}: & & \exists p_v, q_v \in \mathcal{P}^v, p_u, q_u \in \mathcal{P}^u \mid u \in f^1_{v \rightarrow \dot{p}_v}: \\
\lambda^v(\{p_v, q_v\}) = p_v \Rightarrow \lambda^u(\{p_u, q_u\}) = p_u \; (3) & \text{and} & \lambda^v(\{p_v, q_v\}) = p_v \not\Rightarrow \lambda^u(\{p_u, q_u\}) = p_u \; (4). \\
\mid \dot{p}_v = \dot{p}_u, \bar{q}_v = \bar{q}_u & & \mid \dot{p}_v = \dot{p}_u, \bar{q}_v = \bar{q}_u
\end{array}$$

3.2.3 Divergence and Non-determinism

In contrast to cost-suboptimal routing and inconsistency, robustness problems do not affect single states. They refer to an unwanted behavior of transitions between states $s_x \in S$. If a configuration is vulnerable to divergence, a cycle may be traversed. Formally, it holds that

$$\exists s_0, s_x, s_y \in S : s_0 \overset{..}{\rightarrow} s_x \overset{..}{\rightarrow} s_y \overset{..}{\rightarrow} s_x \mid s_x \neq s_y,$$

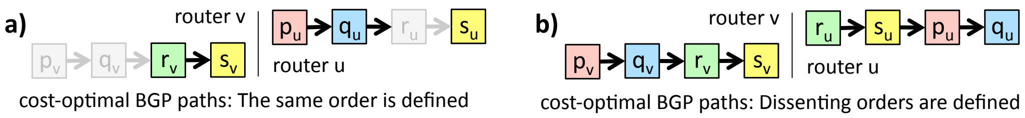

a)
router v
p_v
q_v
r_v
s_v
p_u
q_u
r_u
s_u
router u
cost-optimal BGP paths: The same order is defined
b)
router v
p_v
q_v
r_v
s_v
r_u
s_u
p_u
q_u
router u
cost-optimal BGP paths: Dissenting orders are defined

Figure 2 Inconsistency is either a result of an information lack (b) or dissenting path rankings (b).

where s_0 denotes an initial state of the configuration. If a configuration is vulnerable to non-determinism, state transitions from an initial state to at least two different stable states exist. If s_x and s_y denote stable states, it formally holds that

$$\exists s_0, s_x, s_y \in S : s_x \overset{*}{\leftarrow} s_0 \overset{*}{\rightarrow} s_y \mid s_x \neq s_y.$$

Both, vulnerability to divergence and non-determinism can be traced back to a conceptual defect of the state transition function. It holds that:

An iBGP extension or architecture scheme causes vulnerability to divergence, if and only if its update process definition does not define a partial order on the possible states arbitrary iBGP configuration may enter. [5] (5)

An iBGP extension or architecture scheme causes vulnerability to non-determinism, if and only if the partial order does not define a least element on the states an arbitrary configuration may enter. [5] (6)

The root causes for robustness problems are shown in figure 3.a) and .b). Figure 3.c) illustrates the basic idea behind robust state transitions.

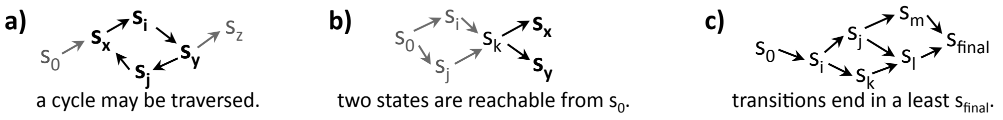

a)
s_0
s_x
s_i
s_j
s_y
s_z
a cycle may be traversed.
b)
s_0
s_i
s_j
s_k
s_x
s_y
two states are reachable from s_0.
c)
s_0
s_i
s_j
s_k
s_m
s_l
s_final
transitions end in a least s_final.

Figure 3 Divergent (a), non-deterministic (b), and robust (partial order, least state) (c) transitions.

4 Studies for Common IBGP Information Exchange Schemes

In literature, correctness problems in the iBGP routing are usually associated with information reduction schemes, cf. section 1.2. Anomalies in fully meshed iBGP configurations were never demonstrated. However, neither the correctness of full-mesh iBGP was ever entirely proven nor systematically studied to which root causes the information reduction schemes are prone to. Evaluating these aspects is tackled in what follows.

4.1 Full-meshed iBGP

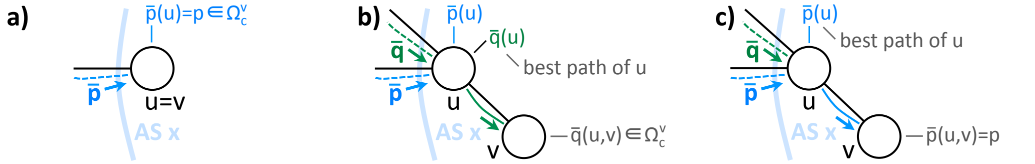
Full-meshed iBGP, the original AS-internal BGP scheme specified in the protocol standard [12], realizes a logical full-mesh for the information exchange. As shown in [3, 5], the state transition process based on such a full-mesh can be mapped to a partial order relation with a least element. This property excludes the root causes (5) and (6) for arbitrary configurations; full-mesh iBGP always behave robust. Whether the remaining root causes can appear is unclear.

4.1.1 Cost-suboptimal Routing

As explained in section 3.2.1, cost-suboptimal routing decisions may have two different causes: Either no cost-optimal path is known (1) or sub-optimal paths mask the cost-optimal ones

(2). In a fully meshed configuration, the first root cause may never appear while the second one may appear. Both properties are proven in what follows.

Firstly, we show that if a configuration is fully meshed, every router knows a cost-optimal path. Let $v \in V$ be a router of a configuration that has entered a stable state. By assumption, it knows none of its cost-optimal paths. Furthermore, let $p \in \Omega_c^v$ be an arbitrary cost-optimal path of v. Let $u \in V$ denote the BGP next-hop where the underlying BGP path $\bar{p}$ is local or external. It knows a signaling path $\bar{p}(u)$ corresponding to the same BGP path $\bar{p}$. Both paths p and $\bar{p}(u)$ specify equal global path attributes. As shown in figure 4, three cases are possible: If $v=u$, cf. figure 4.a), it holds that $p=\bar{p}(u)$. A cost-optimal path is known. If $v\neq u$, u may have chosen path $q \in \mathcal{P}^u \,|\, q \neq \bar{p}(u)$ as its best path, cf. figure 4.b). Compared to $\bar{p}(u)$ and p, this path specifies global attributes of equal preference. Otherwise, as it can be verified easily, u would either not choose q as its best path or p would not be cost-optimal on v. Since $\bar{p}(u)$ is local or external on u, q must be local or external on u, too. If this would not be the case, u would not select q as its best path. Consequently, the signaling path $\bar{q}(u,v)$ permitted and known at v representing $\bar{q}$ specifies u as BGP next-hop. It comes along with the same IGP distance like path p. All in all, $\bar{q}(u,v)$ and p specify cost-relevant path attributes of equal preference. It holds that $\bar{q}(u,v) \in \Omega_c^v$. As this path is known at v, v knows a cost-optimal path. Finally, it may be that $u \neq v$ has selected $\bar{p}(u)$ as its best path. Due to the fact that a full-mesh is applied, this path is expanded to v and thus known. In all three cases, at least one cost-optimal path is known by v. This is a contradiction to the assumptions.



p is external or local on v=u

u provides the cost-optimal path q

u provides the cost-optimal path p

Figure 4 Realizing iBGP via a full-mesh, each router $v \in V$ knows at least one cost-optimal path.

Path maskings, root cause (2), may appear since BGP does not define a transitive ordering relation on arbitrary paths. The basic problem is shown in figure 5.a): Knowing three with respect to the first four attributes equally preferred paths p, q, and r for two peer-AS groups x and y, MEDs are evaluated in step c). For those paths q and r specifying the same peer-AS group y, the MEDs are compared. If different MEDs are specified, the path with the worse MED, here r, is discarded. If path q is now beaten by p in the remaining steps of the selection process, knowledge on q may be essential for a cost-optimal routing decision: If q is unknown, p may lose the decision process against r due to attributes evaluated after step c), the MED. The cost-suboptimal path r can mask path p if q is unknown, even if p is cost-optimal.

As shown in figure 5.b), a situation where a cost-optimal path is masked may also appear if an iBGP configuration is fully meshed. Router u knows all three available exit points. In the decision process, $\bar{q}(u)$ beats $\bar{r}(v,u)$ in step c) (lower MED), and $\bar{p}(u)$ beats $\bar{q}(u)$ in step f) (lower BGP ID). Consequently, v is not aware of $\bar{q}(u,v)$. The known cost-optimal path $\bar{p}(u,v)$ is masked by $\bar{r}(v)$, cf. figure 5.a). For all that is known, this possibility was unknown so far.

4.1.2 Inconsistency

As explained in section 3.2.1, inconsistent routing decisions can have two different root causes: Missing information on relevant paths (3) and dissenting path rankings (4). Using an iBGP full-mesh, a lack of information that leads to inconsistent local views, root cause (3), cannot

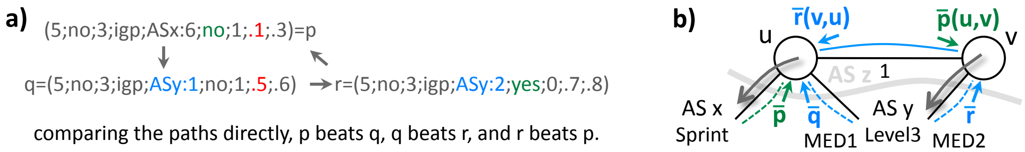

a)
(5;no;3;igp;ASx:6;no;1;.1;.3)=p
q=(5;no;3;igp;ASy:1;no;1;.5;.6)
r=(5;no;3;igp;ASy:2;yes;0;.7;.8)
comparing the paths directly, p beats q, q beats r, and r beats p.
b)
$\bar{r}$(v,u)
$\bar{p}$(u,v)
u
v
AS z
1
AS x
Sprint
$\bar{p}$
$\bar{q}$
MED1
AS y
Level3
$\bar{r}$
MED2

Figure 5 The BGP path selection process does not define a transitive ordering relation (a). This may cause path maskings even in fully meshed iBGP configurations (b).

appear. As depicted in figure 6, the full-mesh structure ensures that v and u learn the same delivery paths from all other routers $w \in V \setminus \{v, u\}$. Since v forwards traffic system-internally, its best path is internally learned. It does not advertise any path. Router u has either chosen an internally learned path as best path and advertises no path or else has chosen a local or external path which is advertised to v. All in all, the information basis v and u have only differ in externally learned and locally configured non-best paths.

By assumption, we know that if we take only the common information basis into account, v and u select a representation of the same BGP path. This is guaranteed by the edge condition defined by equation (3). The routing is consistent. As local paths are preferred in step 1) of the selection process, a router's non-best local path certainly has a lower Local Preference than its best path. Consequently, if such a path is known on v or u, it has no effect on the routers' routing decisions. External paths are preferred in step d). Thus, if a non-best external path is known, the best path specifies better global path attributes: A higher Local Preference (LP), equal LP and shorter AS-path Length (ASpL), equal LP, ASpL, and lower Origin Code (OC), or equal LP, ASpL, OC, and same peer-AS group with lower MED. Again, in all possible cases externally learned paths have no effect on the routing decision. Consequently, even if such paths are known, the same routing decision is made as if only the common information base would be known. Root cause (3) can never occur if full-meshed iBGP is used.

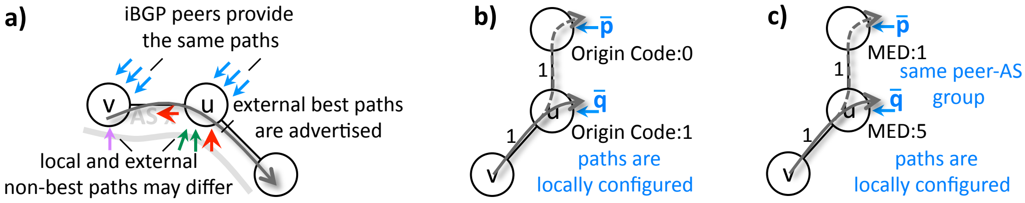


Figure 6 Using a full-mesh, the routing information of v and u differ only by local and external non-best paths (a). Despite of the full-mesh, inconsistency due to root cause (4) may appear (b, c).

Next, we check whether full-mesh iBGP is prone to root cause (4). For that purpose, we analyze the decision process and try to identify paths satisfying the equation. For paths that match the equation, we sketch an exemplary configuration. If equation 4 is satisfied, it holds that $\dot{p}_v = \dot{p} = \dot{p}_u$, $\bar{q}_v = \bar{q} = \bar{q}_u$, and that router v prefers p_v to q_v. Based on these assumptions and the fact that v forwards traffic to a BGP next-hop "behind" u, all possible preferences for the representations permitted at v and u are shown in the RC4-tree, cf. figure 7. Equation (4) is satisfied, if v prefers p_v, while u prefers q_u. Note that $\dot{p}_v = \dot{p} = \dot{p}_u \Rightarrow \bar{p}_v = \bar{p} = \bar{p}_u$.

Local Preference If path p_v is preferred to q_v by v due to a higher Local Preference, path p_u is also preferred over q_u by u. This is the case, since the Local Preference is a global path attribute. It is the same for all representations of a BGP path. If both underlying BGP paths specify the same Local Preference, the subsequent attributes define the path preferences of v and u. Following steps of the decision process are of interest.

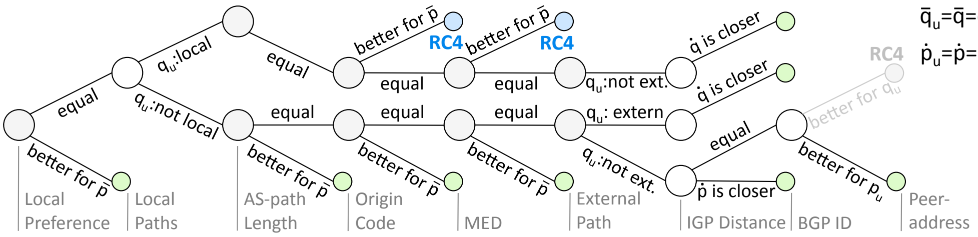


Figure 7 Possible path preferences under the assumptions given by root cause (4). Blue leaves specifying preferences satisfying root cause (4) (RC4), green ones specify a conflict.

Local Paths If step 1) is relevant, neither v nor u prefers any of its paths yet. By assumption, $\bar{p}$ is neither local on v nor on u. Since v does not prefer q_v, $\bar{q}$ is not local on v, too. The cases that $\bar{q}$ is local on u (q_u: local) and that it is not local on u (q_u: not local) remain.
AS-path Length If $\bar{q}$ is local on u, it specifies an AS-path Length of zero. As v prefers path p_v, the AS-path Length of $\bar{p}$ must be zero, too. Following path attributes define the final preferences. This is also the case if $\bar{q}$ is not local on u, while both BGP paths specify the same AS-path Length. If $\bar{p}$ specifies a shorter AS-path, v prefers p_v and u prefers p_u.
Origin Code As p_v is preferred by v, path $\bar{p}$ specifies the same or a better Origin Code than $\bar{q}$. In the former case, independent of whether $\bar{q}$ is local on u or not, v has no preference yet. Following attributes are crucial. In the latter case, v prefers p_u. If $\bar{q}$ is not local on u, u also prefers p_u, equation (4) is not satisfied. If $\bar{q}$ is locally configured on u, u prefers path q_u. This satisfies equation (4). An example that realizes this case is illustrated in figure 6.b).
MED Since p_v is preferred by v, $\bar{p}$ specifies either the same (or an incomparable) or a better MED than $\bar{q}$. In the former case, the following attributes are crucial, independent of whether $\bar{q}$ is local on u or not. In the latter case, p_u is preferred on u if q_u is not local on u. The root cause is not matched. If q_u is local on u, path q_u is already preferred by u. Root cause (4) is satisfied. However, this scenario seems unusual: Since both paths specify an empty AS-path, both paths are locally configured in the AS. For local paths, the MED is usually not specified. A default value is evaluated instead. Nevertheless, a conceivable example is shown figure 6.c).
External Paths According to local paths, no path is externally learned on v and $\bar{q}$ may be external on u. If $\bar{q}$ is local on u, it cannot be external on u, too. If $\bar{q}$ is not local on u, it may be external on u or not. Three cases that may satisfy equation (4) remain, cf. figure 7.
IGP Distance As we assume that IGP distances increase strictly monotonically, the distance from v to u is smaller than the distance to any BGP next-hop behind u. Consequently, if $\bar{q}$ is local or external on u while the preceding attributes did not define a preference on v yet, v prefers path q_v. This is a contradiction to the assumptions. It remains the case that $\bar{q}$ is neither local nor external on u. If the BGP next-hop given by $\dot{p}$ is closer to v than the one given by $\dot{q}$, this also holds for u. Equation (4) may only hold if both IGP distances are equal.
BGP ID In the remaining open case, p_v and q_v as well as p_u and q_u are equally preferred by v and u, respectively. They are internally learned on v and u. Because every router advertises only its best path and $\bar{p} \neq \bar{q}$, both p_v and q_v as well as p_u and q_u specify different BGP IDs. As v prefers p_v, p_v specifies a lower BGP ID than q_v. If path p_u also specifies a lower BGP ID than q_u, consistent rankings are defined. If p_u specifies a worse ID than q_u, root cause (4) is matched. However, in a full-mesh, this case cannot appear in practice: Since p_v and p_u as well as q_v and q_u are advertised by the same router, p_u specifies the same BGP ID like p_u and q_u like q_u. Consequently, p_u has a lower BGP ID than q_u.

All in all, full-meshed iBGP turned out to be prone to expressiveness and global consistency problems. The correctness problems that may appear are summarized in table 1.

Expressiveness Problems		Global Consistency Problems		Robustness Problems	
No known optimal path (1)	Path maskings (2)	Lack of relevant paths (3)	Inconsistent path rankings (4)	Cyclic state transitions (5)	No least state exists (6)
impossible	possible	impossible	possible	impossible	

Table 1 Possible root causes for iBGP anomalies in fully-meshed iBGP configurations.

4.2 BGP Route Reflection AS Confederations

Since Route Reflectors and routers within a member-AS are fully meshed, Route Reflection and AS Confederations are prone to the same problems like full-mesh iBGP. However, as both concepts realize an information reduction compared to a full-mesh, further causes may appear.

Anomalies caused by a lack of information (root cause (1) and (3)) as well as robustness problems are well-known and well-studied phenomenas if Route Reflection or a Confederation is applied. For that reason and due to the space limitation, we confine ourselves to sketch only the conceptual problems for these problem classes.

A cost-suboptimal routing decision due to a lack of cost-optimal paths, root cause (1), is very easy to induce. It appears if the Route Reflectors or the member-AS Border Routers do not reflect a router's cost-optimal paths. This behavior occurs if none of the relevant paths is the best one from the reflectors' points of view. A simple example for such a configuration is shown in figure 8.a). An example for AS Confederations can be designed analogously. Root cause (3), inconsistent routing decisions due to a lack of relevant information, can be induced in a similar way. Inconsistent routing decisions are made, if router u on the IGP forwarding route to the BGP next-hop knows a better path $\bar{q}$ that is not reflected. Again, this is the case if $\dot{q}$ is not the best delivery path from the reflectors' points of view. An example configuration is depicted in figure 8.b). A comparable configuration can be designed for Confederations, too. The fact that information reduction may cause cyclic state transitions and avoid that a least state is defined, root cause (5) and (6), is thoroughly studied. Readers who are interested in the details may be referred to [9], for example.

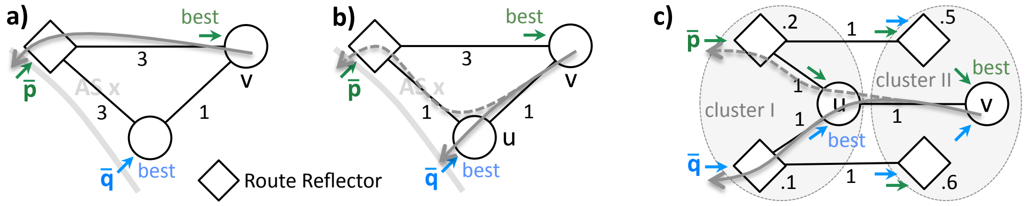


Figure 8 Classical examples for cost-suboptimal routing due to root cause (1) (a), inconsistency due to root cause (3) (b), and dissenting rankings due to the BGP ID (c). Example (b) is taken from [9].

Finally, we take a quick look on inconsistency due to dissenting path rankings, root cause (4), in ASs using Route Reflection and AS Confederations. An important difference between full-mesh iBGP and schemes which reflect paths is that representations of the same delivery path on different routers may specify different BGP IDs. As a result of this property, the third possible path preference that matches root cause (4), cf. the RC4-tree in figure 7, may occur. An example for such a configuration is shown in figure 8.c): Routers v and u both learn two cost-optimal paths $\dot{p}$ and $\dot{q}$ from two different routers. Due to the BGP ID, u prefers $\dot{q}$, while v prefers $\dot{p}$. Note that this case is of high relevance in practice. Network Operators in large ASs usually use several redundant reflectors per cluster to avoid single points of failure. A similar configuration can also be designed for AS Confederations. The root causes for iBGP anomalies that may appear if scalability techniques are used are summarized in table 2.

Expressiveness Problems		**Global Consistency Problems**		**Robustness Problems**	
No known optimal path (1)	Path maskings (2)	Lack of relevant paths (3)	Inconsistent path rankings (4)	Cyclic state transitions (5)	No least state exists (6)
trivial	possible	trivial	possible	possible	

Table 2 Possible root causes for anomalies in ASs using Route Reflection or AS Confederations.

5 Workarounds and Solutions

Understanding the protocol and configuration properties that induce iBGP anomalies, possible countermeasures can be discussed. We differ between *solutions* and *workarounds*: Solutions correct conceptual defects but require protocol changes. Workarounds are quickly deployable patches that avoid unwanted behavior in production systems. Due to the space limitation, we only sketch the basic ideas.

Unknown cost-optimal paths, root cause (1), appear if a router's local view is not matched by any of its BGP peers' best path selection. Information on the cost-optimal paths is not forwarded. To avoid this problem, reflectors could advertise paths matching to the receivers' point of view – independent of their own best path decision. A scheme realizing this concept is the *iBGP Route Server architecture*, cf. [4]. Using common information reduction schemes, problems are avoided if every cluster shares a common local view. To reach this, the distances of the links must ensure that cluster-internal BGP next-hops are always closer than -external ones. However, this workaround is a serious limitation for the IGP routing. Path maskings, root cause (2), are avoided if the selection process defines a transitive ordering relation on arbitrary paths. This can be reached if the MED is not evaluated at all or across all peer-AS, cf. [11]. Alternatively, it can be ensured that every router knows an AS-cost-optimal path for each peer-AS group [4]. However, this is not trivial in arbitrary configurations in general.

Consistency problems due to a lack of relevant routing information, root cause (3), are avoided if routers on a future forwarding route have a comparable information basis for their routing decisions. Improvements can be achieved here if additional cluster-internal BGP sessions are kept. However, this workaround counteracts with the intended information reduction and – which is the greater problem – can be the source of inconsistent routing, cf. [9]. Dissenting rankings, root cause (4), are avoided if all routers on a future forwarding route prefer the same delivery paths. However, to reach this, modifications on the selection process are necessary. Using a full-mesh, operators can achieve consistent rankings by means of a simple workaround, given by two design rules: Firstly, paths for a destination originated within the own AS must specify the same Origin Code. Secondly, MEDs must not be used for local paths. Controlled by the operator, both properties can be realized easily in ASs.

Robustness problems can be addressed in various different ways. Redesigning the protocol and the information exchange scheme, it is possible to ensure that an optimal state always exists and that this state is certainly entered, cf. [4]. Making use of common iBGP schemes, cyclic dependencies between the cost-optimal paths of routers in different clusters can be studiously avoided, cf. [11]. This can be reached by design rules comparable to those specified above. However, also similar problems are induced. Other concepts are based on modified path preferences [15], the exchange of additional information [16], and cycle detection [10].

6 Conclusion and Future Work

In this paper, we systematically analyzed the possibility of anomalies when using the standardized iBGP routing information exchange schemes. The most interesting results were obtained for full-mesh iBGP: While we could prove that full-mesh iBGP is resistant to problems caused by an information lack, we could also reveal that path maskings and inconsistent rankings may appear. This observation clearly falsifies that iBGP anomalies are only a consequence of an information reduction, a position many studies convey. For BGP Route Reflection and AS Confederations, we could give a systematic overview on the for the great part known effects.

Understanding that and why full-mesh iBGP is clearly more resistant to anomalies seems to be an important increase of knowledge. Drafts for new iBGP extensions come along with similar conceptual defects like Route Reflection. Knowledge why certain iBGP schemes are prone or resistant to different root causes could help to avoid that new correctness problems are caused by the final standards. Besides this, the knowledge could also help to improve the existing schemes in the long term. In addition to protocol designers, the results gained here are also of interest for Network Operators. Knowing why certain root causes appear helps to develop and deploy adequate workarounds like network design rules for the existing schemes. Moreover, it becomes easier to evaluate iBGP scheme alternatives and identify the right one for the own AS. Even if we sketched first ideas in section 5, addressing anomalies so that the solution is usable in production systems is an important aspect for future research.

References

1 A. Basu, C. Luke Ong, A. Rasala, F. B. Shepherd, and G. Wilfong. Route Oscillations in I-BGP with Route Reflection. In *SIGCOMM '02*, pages 235–247. ACM Press, 2002.

2 T. Bates, E. Chen, and R. Chandra. BGP Route Reflection - An Alternative to Full Mesh IBGP. April 2006. RFC 4456.

3 U. Bornhauser and P. Martini. A Divergence Analysis in Autonomous Systems using Full-mesh iBGP. In *CNSR 2008*, pages 600–608. IEEE Computer Society, May 2008.

4 U. Bornhauser, P. Martini, and M. Horneffer. An Inherently Anomaly-free iBGP Architecture. In *IEEE LCN 2009*, pages 145–152. IEEE Computer Society, October 2009.

5 U. Bornhauser, P. Martini, and M. Horneffer. Root Causes for iBGP Routing Anomalies. In *IEEE LCN 2010*. IEEE Computer Society, October 2010.

6 Cisco Systems. Field Notice: Endless BGP Convergence Problem in Cisco IOS Software Releases. Technical report, October 2000.

7 N. Feamster and H. Balakrishnan. Correctness Properties for Internet Routing. *Allerton Conference 2005*, September 2005.

8 T. Griffin, A. D. Jaggard, and V. Ramachandran. Design Principles of Policy Languages for Path Vector Protocols. In *SIGCOMM '03*, pages 61–72, 2003.

9 T. Griffin and G. Wilfong. On the Correctness of IBGP Configuration. *SIGCOMM Comput. Commun. Rev.*, 32(4):17–29, August 2002.

10 T. Klockar and L. Carr-Motyckova. Preventing Oscillations in Route Reflector-based I-BGP. In *Proceedings of ICCCN 2004*, pages 53–58, Chicago, IL, October 2004.

11 D. McPherson, V. Gill, D. Walton, and A. Retana. Border Gateway Protocol (BGP) Persistent Route Oscillation Condition. August 2002. RFC 3345.

12 Y. Rekhter, T. Li, and S. Hares. A Border Gateway Protocol 4. January 2006. RFC 4271.

13 J. L. Sobrinho. An Algebraic Theory of Dynamic Network Routing. *IEEE/ACM Trans. Netw.*, 13(5):1160–1173, 2005.

14 P. Traina, D. McPherson, and J. Scudder. Autonomous System Confederations for BGP. August 2007. RFC 5065.

15 M. Vutukuru, P. Valiant, S. Kopparty, and H. Balakrishnan. How to Construct a Correct and Scalable iBGP Configuration. In *IEEE INFOCOM*, Barcelona, Spain, April 2006.

16 D. Walton, A. Retana, E. Chen, and J. Scudder. BGP Persistent Route Oscillation Solutions. May 2010. Internet Draft.