Two Steps Towards Practical Compact Routing

Rolf Winter, *NEC Labs Europe*

Abstract— **The stress on today's inter-domain routing system is constantly and sharply increasing. The mere size of the Internet, operational practices and a number of limitations of the routing protocol itself worsen the scaling behavior of this fundamental service at an alarming rate. This paper introduces 2SIDR, a two step inter-domain routing approach that aims at significantly reducing the state requirements of routers while minimizing the incurred path stretch penalty. This fundamental trade-off has already been analyzed on a graph theoretical basis in the domain of compact routing. In contrast to those theoretical approaches, 2SIDR aims at what we call practical compactness. It deliberately gives up mathematical bounds while only relying on data that is available in practice and adhering to existing business relationships between operators in the Internet.**

Index Terms—**routing protocol, inter-domain routing, compact routing, stretch/state trade-off**

I. Introduction

Inter-domain routing in today's Internet is done using the Border Gateway Protocol (BGP4), which connects an impressive number of networks, each under the control of a distinct administration known as an Autonomous System (AS). Routing itself is based on IP addresses and Internet Service Providers (ISPs) are assigned contiguous IP address blocks (IP prefixes). ISPs can further divide these prefixes in order to assign them to different customers. The provider itself can announce its address space to the global Internet as a single aggregate; this is BGP's main mechanism for achieving scalability today.

However, this approach has a number of important problems as despite aggregation, BGP still suffers from scalability issues regarding not only the amount of routing state but also the amount of updates this state is subject to [1]. Studies have shown that both of these exhibit super-linear growth trends over time, with worst-case predictions exceeding the pace at which critical router hardware components evolve [4]. This is not solely a result of the "natural" growth of the Internet but also of BGP's limitations and operational practice.

The fast growth of routing state comes as a result of several factors. To begin with, IP address aggregation, while simple, fails to support some of today's requirements efficiently, and so operational reality frequently deviates from it. For example, provider-independent prefixes exist that are not part of a prefix hierarchy. These non-aggregated prefixes are advantageous for ASes as they can be ported when changing ISPs, leaving site-internal numbering intact. Indeed, the difficulty associated with the process of re-numbering can be discouraging enough to lead to provider lock-in. Another practice breaking the assumptions about prefix aggregation is multi-homing, whereby prefixes are advertised over parts of the topology where they cannot be aggregated. The increasing need to multi-home stub ASes (ASes that do not provide traffic transit services) is a major cause for routing table inflation [7]. Furthermore, this practice is unlikely to subside, since the Internet has become critical for the operation of many businesses that use multi-homing as a means to achieve reliability.

The routing state scalability problem is further exacerbated by address fragmentation, i.e. different prefixes that originate a single AS might not be aggregatable, and prefix disaggregation, whereby an AS splits an IP prefix to, among other things, force traffic to enter its network from different points. This form of traffic engineering will doubtlessly continue as BGP provides no mechanism to allow for this kind of inbound traffic engineering without resulting in an uncontrollable increase in the global routing state.

This clearly shows IP–prefix aggregation is no longer sufficient to provide good scalability in today's Internet. In addition, on a purely mathematical note, such a hierarchical aggregation approach works best in graphs in which the average distance is increasing noticeably with an increasing number of nodes [5]; the Internet's topology has not proven to exhibit this property [6].

The second problem that BGP suffers from in terms of scalability is an increasing rate of churn. Unfortunately, BGP has poor locality properties, meaning that a large portion of BGP updates are globally visible [1], thus consuming resources such as CPU cycles and memory on routers. Put together, these effects influence the convergence properties of BGP, which, after an incident, might take as much as tens of minutes to converge [8]; proper confinement of routing updates to regions of the network is therefore a strong scaling requirement.

Clearly, BGP is struggling to keep up with the demands of the current Internet. In this paper we present 2SIDR, the Two Step Inter-Domain Routing protocol. 2SIDR achieves a low path stretch, where the stretch is defined as the ratio of the path found by the routing algorithm and the shortest possible path. 2SIDR further limits the propagation of updates throughout the network and achieves scalability without relying on an aggregatable addressing scheme. This paper analyzes 2SIDR's routing approach based on a substantial amount of data [22][14][2] including routing table data from 294 different ASes. This evaluation covers a broad range of aspects including the AS-level stretch, the state requirements, the impact of multi-homing and routing policies.

Manuscript received February 15, 2009. The research results presented herein have received support from Trilogy (http://www.trilogy-project.org), a research project partially funded by the European Community under its Seventh Framework Programme.

II. Two Step Inter-Domain Routing – 2SIDR

BGP is operating at the granularity of IP prefixes and therefore BGP's scaling behavior is dependent on the amount of prefixes being advertised. Clearly, as prefixes can be disaggregated, controlling the overall amount of prefixes is difficult. Only little can be done besides installing filters that discard advertisements above a certain prefix length (today this boundary is typically at /24). There is other routing relevant information available to BGP today such as BGP atoms [16] or AS numbers, but they are either not used at all in the routing process or have a somewhat "subordinate" role to determine the AS path length or to detect loops. This is somewhat unfortunate, as in the current Internet the way networks are interconnected is largely based on commercial relationships [9] between ISPs. An IP prefix in itself does not carry information that would be relevant for such interconnections.

A. Addressing

2SIDR routes based on AS numbers, since at the inter-domain level the AS information and more precisely what the AS represents (i.e. an entity having contractual, organizational and political implications), has much more relevance than a prefix. In 2SIDR, since AS numbers are used to route across AS boundaries, IP prefixes lose their global significance: IP addresses used in one AS are only valid inside that particular AS. As a result, a fully-qualified address for inter-domain routing contains both the local IP address of the destination node's network interface as well as the corresponding AS number.

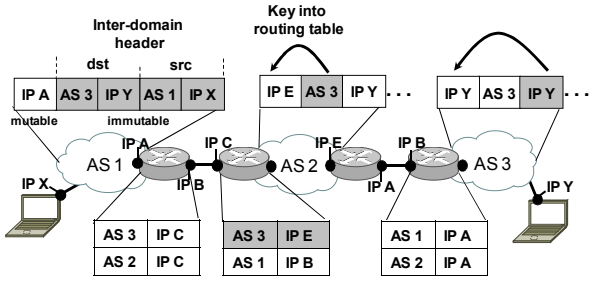

Fig. 1. Addressing scheme

Fig. 1 gives a simple example of addressing under 2SIDR, where host X on the left sends a packet to host Y on the right. The figure shows routing tables at the bottom and simplified packet headers at the top (due to space constraints source fields are only shown once). As shown in the left-most header, the packet is split into an immutable inter-domain part consisting of a destination AS's number and the end-host's IP address, and a mutable IP header used to address routers within an AS.

In the figure, the packet first arrives at interface IP A, whose router performs a look-up based on the destination AS 3 and changes the IP header to IP C. This process repeats itself until eventually the packet arrives at AS 3, where the destination IP address is used to forward the packet locally to host Y. The encapsulation technique used for the outer IP header should be understood as only one possible implementation, since, e.g., MPLS could serve the same purpose. Within this addressing framework, an end-host that wants to establish a connection with another end-host needs to perform a DNS lookup returning a 2SIDR specific resource record. This record contains the end-host's IP address and the respective AS number.

Using routing information at the AS level has a number of technical advantages. Some basic form of aggregation is e.g. implicitly built-in, as an AS naturally aggregates the host and networks it contains. At the AS-level, the problem of provider lock-in through address delegation also vanishes: as the routing information is not aggregatable, there is no reason to force ASes to change their AS number when changing their provider. Another positive property stemming from the non-aggregatable nature of the routing information is failure atomicity [17] which today not always exists. The failure of a prefix today might be "masked" in the global routing system because it has been aggregated.

The downside of switching to AS granularity for inter-domain routing is that inbound traffic engineering as it is done today through prefix disaggregation is not possible any more. On the other hand, disaggregation causes a fair share of the scalability problems observed today. More precisely, every routing approach that relies on traffic engineering being solved using more specific addressing will end up suffering from scalability pressures similar to what we observe on the Internet today. Therefore, a new routing protocol needs to provide a better, more scalable mechanism for traffic engineering.

B. Routing

The general routing approach of 2SIDR was mainly influenced by work on compact routing. The main characteristic of these approaches is to have a low mathematical upper bound on the storage burden on nodes in the network in order to perform routing. At the same time, the path stretch is also bounded by a constant factor. Compact routing has already been investigated for graphs exhibiting the power-law characteristics of the Internet [11][12]. It was shown that certain schemes on Internet-like graphs perform even well below their mathematical bounds [12].

Unfortunately, compact routing schemes usually require full knowledge of the graph, which is not available in the current Internet due to information hiding practices. Besides, from what source should such information come in practice if not from the routing protocol itself? To make matters worse, if the graph changes, compact routing schemes might require substantial topological restructuring to ensure reachability of nodes and to further guarantee the properties mentioned above. These requirements render them impractical as a solution for inter-domain routing.

Despite this, we make use of insights from the general field of compact routing but primarily from [21] and [11] to make an important key finding: when constructing a routing topology on top of a power-law graph such as the Internet's AS-level graph, the construction process should give a central role to the densely-meshed core of the network. As a result, 2SIDR separates inter-domain routing in the sparsely connected edges from routing in the core of the Internet, where the core is formed by a small number of large providers, commonly referred to as tier-1 providers.

1) Step One: Routing in Edge Regions

In order to reduce the amount of routing state, only local information is maintained in edge regions. "Local" here generally refers to information about destination ASes that are reachable over paths that do not include one of the tier-1 provider networks. Keeping only local information not only results in low state requirements, but also shields edge regions from updates that originate somewhere beyond the dense core. Using 2SIDR, an AS primarily maintains routing state for direct neighbors. In purely theoretical compact routing schemes this could already violate strict mathematical state bounds. However, any practical routing protocol that fails to adhere to this requirement might already introduce stretch for routes towards direct neighbors.

In order to reach distant destinations, the tier-1 providers function as *landmark ASes*. ASes in the edge region elect their closest tier-1 provider in terms of AS hops as their associated landmark. In case a destination AS is locally unknown, packets are simply forwarded to the landmark. Because of the structure of the Internet's AS-level graph, a shortest path between distant ASes will, with high probability, go through one of the well-connected tier-1 providers. Further, the average distance between any two ASes in this graph is fairly short, as we will show in section III, and so any tier-1 AS is close to any other point in the AS graph. Even if the chosen landmark is not optimal for a given destination, as the tier-1 providers are all direct neighbors, the path stretch incurred on a long AS path will be small.

Tier-1 ASes advertise themselves as landmarks to direct customer ASes, who, in turn, advertise a chosen landmark towards their customers, and so forth. The Interior Gateway Protocol of an AS needs to make sure that a chosen landmark is propagated to all 2SIDR routers belonging to the same AS. These routers are not to distribute the landmark state beyond the AS boundary, the one exception being other customers; these advertisements follow customer-provider links only in the customer direction.

This simple model makes sure that each customer is aware of at least one landmark; it also adheres to the most fundamental routing policy employed today [23]. In the reverse direction, customer reachability information is propagated towards the chosen landmark, thus ensuring that each AS is reachable through at least one landmark.

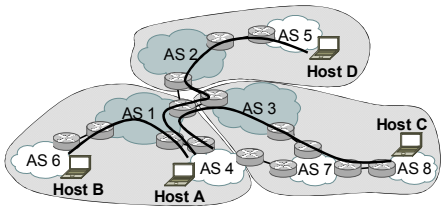

Fig. 2. 2SIDR routing example

To illustrate how 2SIDR routes in edge regions, consider the topology depicted in Fig. 2. ASes 1 to 3 are tier-1 providers, while the ASes in white belong to the edge regions with their associated landmark shown in the same light gray areas. Thus, ASes 4 and 6 in this example share a common landmark, AS 1. Routing between hosts of these two ASes follows a route towards the landmark as they have no explicit routing state available on how to reach each other. As both have chosen AS 1 as their landmark, the packet can be delivered without it having to leave the edge region.

2) *Step Two: Routing Between Landmarks*

In order to ensure global reachability, tier-1 landmark ASes need to maintain state to reach all direct neighbors. That could already be a large fraction of the total number of ASes in the Internet. The highest AS degree observed in our data is over 3000, with over 30,000 ASes in total. These numbers also imply that by just changing the routing granularity from prefixes to AS numbers, the routing table state can be reduced from around 270,000 entries to a mere 30,000 at the landmarks. In the long run, however, this reduction might only be a one-time gain [13], especially since the number of ASes does exhibit a fast growth trend [1] that would require a more scalable algorithm to be applied between landmark ASes. It is unclear whether tier-1 providers would be really willing to only maintain partial routing state. Given this, we introduce and later evaluate two variants of landmark routing, one that maintains full state and another one that distributes the state.

Maintaining full state is the simplest possible strategy that can be applied between landmarks. To establish this state, landmark ASes simply exchange their customer-learned routing state. This would obviously result in routing state in the order of the number of all ASes at each tier-1 AS. This state enables each landmark to forward packets to ASes that are associated with a different tier-1 to that landmark directly, which in turn can deliver the packet down its customer-provider hierarchy.

The other variant of 2SIDR's landmark routing employs routing inspired by distributed hash tables (DHTs), relaxing the full state requirement of landmarks. Customer state is only distributed selectively among the tier-1 ASes based on numerical closeness. In more detail, a landmark has to check whether a customer's AS number is numerically closest to its own AS number among all landmarks (possibly a common hash function is required for an equal distribution of state). If it is not, the customer-learned state is propagated towards the numerically closest landmark exclusively, thus potentially introducing additional stretch. If stretch is incurred, the distributed landmark routing scheme between tier-1 providers only adds one additional AS hop, since tier-1 providers are all neighbors and are therefore in each other's routing tables so that the numerically closest landmark can be reached directly. This distributed variant of 2SIDR's landmark routing is therefore another compromise in the fundamental state/stretch trade-off, reducing the amount of state in exchange for a potential increase in stretch.

As an example of landmark routing, assume host A in Fig. 2 sends a packet towards host D. The packet finds its way to AS 1, AS 4's landmark. Using the full state landmark routing variant of 2SIDR the packet can be forwarded directly to AS 2. On the other hand, in the distributed variant the necessary routing state is not available, since AS 1 is not numerically closest to AS 5 among the landmark ASes. Consequently, AS 1 forwards the packet to AS 3, which is the numerically closest landmark AS to AS 5. Since AS 3 is not responsible for AS 5 it must have received routing information about AS 5's connectivity to AS 2, the target landmark. From AS 2 on, customer-learned state is used to finally deliver the packet.

3) Policy Routing

Today, shortest path routing is only a first order requirement in case there is no applicable policy that pursues a different objective. That is why a large fraction of inter-domain paths are inflated [15][20]. This demonstrates that policy is an important part of today's inter-domain routing, and so 2SIDR needs to be able to depart from technical optimality in order to accommodate policy decisions [18].

One common goal of policy is to optimize routing with respect to the commercial benefit of an ISP. The basic connectivity 2SIDR provides as described so far could be comparatively expensive in monetary terms as routing towards landmarks will most likely go through costly transit links. As a result, 2SIDR needs to provide ways to model today's business relationships.

One such relationship is called peering, which is an agreement [9] between ASes to offer mutual, cost-free transit for traffic between them and possibly towards each other's customers. To achieve this with 2SIDR, peering ASes could simply exchange their customer-learned state. This would increase the routing state at the respective ASes, but it would translate directly into commercial benefit, since the more expensive transit links are not used for the destinations covered under the peering agreement; as a result, this state is very different from much of the state that is kept at BGP routers today. For one, the benefit (reduced cost) is gained where the price is paid (state). In contrast, in the current Internet much of the state in routing tables is a result of disaggregation and other practices whose beneficiaries are not the ones who "pay" for the increase in state, but rather the global Internet does. Another advantage of sharing state with peers is that the churn isolation properties are preserved. As there is no economic incentive to further distribute the routing information learned over peering links [18], the respective churn will not affect ASes that are not covered under the peering agreement.

State through peering is not just "good" in a commercial sense: it also helps to improve 2SIDR's stretch. By extending the knowledge about local destinations through peering agreements, these potentially shorter paths become available. A good example for this can be found in Fig. 2. Packets originating in AS 4 destined for a host in AS 8 would first travel towards AS 4's landmark. As AS 1 is not the landmark for AS 8 it forwards the packet towards AS 3 (in both variants). The packet is then sent to AS 7, AS 8, and finally to the host, resulting in 4 hops in total. The shortest AS path available in the graph is two AS hops long going through AS 7, and so the stretch incurred in this example is 2. In this scenario, a peering agreement between AS 4 and AS 7 would eliminate this path stretch.

Other common policies with a commercial objective, such as preferring customer-learned routes over those learned from peers and providers, can be easily applied to 2SIDR. Customer-learned routes are the basis of 2SIDR and can be tagged with a higher local preference, similar to practices known from BGP [18]. In addition, policy decisions can be applied to how customer state is propagated towards a landmark to, for instance, choose a more economical path over a shorter one. The effect on our model is that the paths towards landmarks might not follow a shortest path anymore and the same path inflation effect as demonstrated in [20] is the result. This is an inherent artifact of policy routing and the economic environment of inter-domain routing, and not a trait of 2SIDR.

III. Experimental Results

In order to evaluate 2SIDR's performance on the current Internet we used two data sets. One consists of AS-level information collected by the Internet Research Lab [2] which combines a large set of sources including BGP routing tables and updates. We applied the heuristics found in [3] to purge outdated information from it. The resulting AS-level graph consists of 30,753 ASes and 106,137 links between them. The graph exhibits the well known scale-free properties regarding the AS degree distribution with an average node degree of around 6.9 and a maximum degree of 3,190. The 7 tier-1 ASes identified by [15] are used as landmark ASes.

Our second set of data consists of an extensive number of routing tables gathered by the RouteViews [14] project and the RIPE NCC's Routing Information Service [22]. Their collection day coincides with that of the previous data set and comprises views from 294 different ASes. We reduced the amount of paths found by extracting BGP atoms [16], since 2SIDR routes based on AS numbers instead of prefixes. Only considering full routing tables, on average over two thirds resulted in unique AS paths; in other words, although there were multiple occurrences of a destination AS in a FIB, the AS paths leading to it were all identical. On average around half of these unique AS paths did not even have multiple occurrences. Overall, the largest fraction (over 80%) of AS paths found in the individual FIBs were redundant. For the remaining destination ASes that were reachable over multiple AS paths, we selected the path with the largest combined IP address coverage. By discarding longer prefixes, we filter out a number of the paths that might exist in today's routing system due to prefix disaggregation, a technique that is not possible under 2SIDR. By utilizing short prefixes for our evaluation, we assume that we preserve policy decisions that are still valid using 2SIDR.

A. Evaluation Methodology

As a first step, we constructed an all-pair shortest path matrix from the AS-level information data set in order to calculate the stretch 2SIDR is causing in terms of AS hops. Based on this matrix, we also derived the topology 2SIDR creates by assuming shortest path routing within edges and between landmarks. To some degree, this assumption is valid as BGP itself is also in principle a shortest path routing protocol. This model is used as it allows us to reconstruct all possible AS pair path combinations.

In section B.4) we reconstruct 2SIDR paths using the routing table data. By using the paths learned from the routing tables we implicitly apply the ASes' policy decisions. Unfortunately, as only limited amount of routing table information is available and more than one observation point is needed to reconstruct 2SIDR paths, relying on a routing table based model alone is limiting compared to the shortest path-based model. However, we use this latter, more accurate model to

confirm the accuracy of shortest path-based model and explain its surprisingly high level of accuracy.

B. *Analysis of 2SIDR*

In order to validate our topology assumptions we first analyzed the underlying AS-level graph. The results confirm the assumed topological characteristics on which 2SIDR hinges, with relatively short distances between any two ASes. On average, ASes are only 3.7 AS hops apart with a maximum distance of 14 AS hops. Additionally, ASes are never far away from a tier-1 network. On average it takes only 1.81 AS hops to reach the closest tier-1 provider and at most 11 AS hops.

1) *Stretch*

For the stretch calculations we created paths between all AS pairs that originate a prefix using 2SDIR and compared each one against the shortest possible path. Using the full state variant of 2SIDR, the stretch amounted to only around 1.17 on average. A little below 50% of all paths were in fact shortest paths, as can be seen in Fig. 3. Close to 98% of the paths had a stretch of equal to or less than 1.5 and the largest stretch observed amounted to 4 with only $1.78*10^{-7}$% of the paths having that stretch factor.

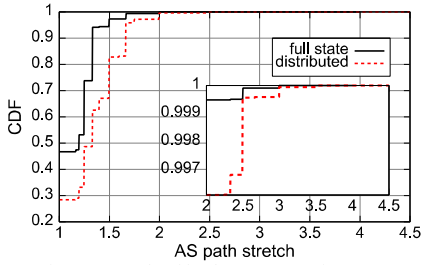

Fig. 3. Stretch of both 2SIDR variants

With the exception of a negligible fraction of paths with stretch 4.5, the distributed approach did not exceed the maximum stretch of 4 found in the full state variant. The reason is simply that only longer paths are actually affected by landmark routing, and the additional hop incurred influences the stretch far less than on shorter paths that do not leave edge regions. However, as can be expected, in this variant a larger fraction of the paths incurred stretch, although 28.5% of the paths still did not incur any stretch at all and over 99.5% of the paths had a stretch equal to or smaller than 2. The average stretch was below 1.32.

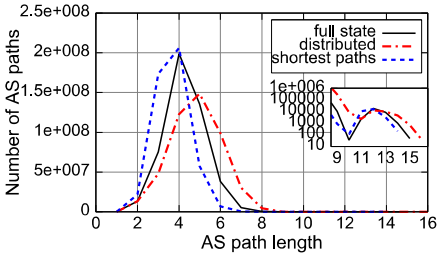

Fig. 4. Path lengths

The stretch statistics alone do not show which paths suffer from stretch the most: if long paths incurred large stretch, it would render 2SIDR impractical. Fig. 4 shows the path length histograms of 2SIDR's two variants together with the histogram of the shortest paths. As can be seen, 2SIDR does not significantly increase the maximum path length, meaning that the stretch is mainly incurred on shorter paths. In more detail, the most notable shift is in paths of length between 3 to 7 AS hops. In the distributed case, the majority of the paths have now 5 instead of 4 AS hops. Both approaches have significantly less paths of length 3, stemming from the fact that the source and destination ASes do not know each other through neighborhood relationships and are also "far enough" apart to potentially have different landmarks.

2) *State and Churn*

In Fig. 5 we show 2SIDR's variant-independent state requirements. The figure shows the customer-learned state (excluding direct neighbors) and the overall state, which includes neighbor state and a default route towards a landmark. The figure further depicts the state requirements for edge and core regions separately.

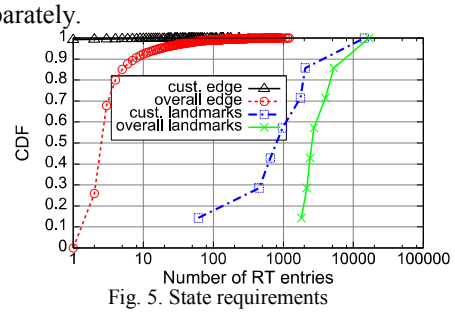

Fig. 5. State requirements

In the edge regions, around 91% of the ASes require at most 8 routing table entries, with a maximum of a little over 1,200 entries. The figure clearly shows that in edge regions, the amount of state is largely a function of the underlying graph structure, where the local connectivity dominantly influences the amount of state. Such state distribution is sensible, as the degree is an indicator of an AS's size [24] and therefore an indicator of an AS's role. Another important result is that distant ASes do not influence the state requirements as it is the fact today, where e.g. disaggregation at the edges impacts upstream and distant routing tables as this state is not propagated beyond the core. This is not only a technical improvement but also corrects an existing economic imbalance to some degree, as this state and the associated churn is resulting in cost inflicted by ASes without direct economic ties.

As can be seen, the customer-learned state at the landmark nodes is very unbalanced, mainly due to the fact that the degree of the seven best connected ASes already varies significantly from a little over 1,300 to over 3,000. Since in our model edge ASes chose the closest tier-1 provider as a landmark, the likelihood that the best connected provider will be chosen is high. On average, 2,900 customer entries are kept at landmarks in addition to the entries for direct neighbors. Finally, the full state variant would add up to n routing table entries at landmarks (n being the number of ASes), while the distributed variant would add in the worst case n/n_l entries, where n_l is the number of landmark ASes.

However, even if full state is maintained at landmarks, the churn rate is much lower than it is today, since 2SIDR isolates landmarks from changes in distant edge regions. In other words, if an AS's local connectivity changes but not its associated landmark, the change does not need to be propagated beyond the local edge. Distant edge regions are generally shielded from such churn as they do not keep *any* state about

distant ASes.

3) Multi-homing

Our data reveals that the majority of the stub ASes today is multi-homed. The most prominent reasons for multi-homing are to increase an AS's reliability and to increase capacity. The latter one requires a traffic engineering (TE) mechanism to influence the way traffic enters and exits a site through the available attachment points.

We evaluate a multi-homing strategy that establishes an additional path towards a second landmark, i.e. one that is different from the already chosen one. This strategy maximizes the resilience of an AS's connectivity, as a failure of the primary landmark path does not cause the disconnection from the larger network. This strategy would also serve to protect against disruptions caused by de-peering events between tier-1 providers such as recently observed between Sprint and Cogent.

With this strategy, most multi-homed stub ASes created state at an additional landmark AS. Only 148 ASes needed to be multi-attached to the same landmark, because they could not reach an additional landmark without traversing their primary landmark.

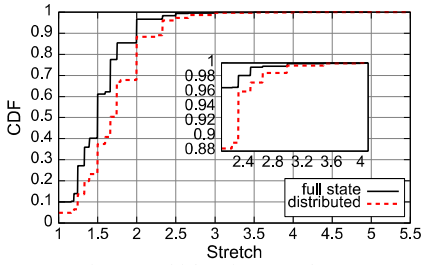

Fig. 6. Multi-homing stretch

Under this scenario, we analyzed the stretch between multi-homed networks and the additional state. For the stretch statistics, we used *only* multi-homing landmark paths and neighbor state for routing and we did not consider the primary landmark paths. As the secondary landmarks are likely further away than the primary landmarks, paths that purely make use of multi-homing state are more likely to incur additional stretch, as shown in Fig. 6. This analysis therefore identifies the absolute worst-case performance.

As we are trying to construct paths optimized with respect to resilience and not length, the stretch between multi-homed ASes using *only* secondary routing state results in only around 10% and 5% of the paths being shortest paths. In both variants, the maximum stretch increases slightly to 5 and 5.5 respectively. But only a handful of paths actually have that maximum stretch. Still, the average stretch for both strategies is not excessive, with slightly above 1.50 for the full state variant and 1.72 for the distributed one. Multi-homing resulted in additional state of around only 1,600 entries at the landmarks on average. In the full state variant, this does not need to translate to additional state at the landmarks of course. Instead of utilizing the state learned from other landmarks, this customer-learned state can be used instead. Even for the distributed case this does not result in all cases in additional state but depends on the numerical closeness to the edge ASes in question.

To evaluate the degree of resilience this multi-homing strategy provides we measured the AS and link overlap of the primary and secondary landmark paths; this overlap is expressed as the fraction of ASes or AS hops in both paths that are equal, excluding the source. The AS overlap between primary and secondary path is 0 for over 94% of all routes and the link overlap is 0 for over 99.5% of all routes. These numbers demonstrate that the multi-homing strategy employed is very successful as an overlap of 0 implies that primary and secondary landmark paths are completely disjoint.

Increasing capacity is a quite different motivation for multi-homing as it requires a mechanism to influence the way traffic flows in and out of the network. While outgoing traffic can be influenced by the AS itself, inbound traffic flow depends on the routing decisions of the ASes on a given path towards a multi-homed AS. Today, ASes at the edges of the Internet can influence these routing decisions at distant ASes by prefix disaggregation. "Abusing" the fact that routers perform longest prefix matching, this practice contributes to the pollution of global routing tables. An AS today has only very limited means to express path preference such as AS path prepending which often results in extremely coarse grained (on/off) control over traffic flows.

The above depicted multi-homing strategy is one way to achieve a certain level of traffic engineering. All traffic that originates within the customer hierarchy of the second landmark will arrive through the upstream provider through which the second announcement was made. In addition, the path selection process needs to be based on more information than just the AS path. This is especially true for AS paths with equal path attributes, where, in the absence of policies, meaningless tie-breaking rules today determine the chosen path. Approaches to achieve this without relying on the dissemination of more specific prefixes have been proposed, even for BGP, such as the inter-AS cost [19] which allows an origin AS and intermediate ASes to express cost, i.e. preference, for a given prefix advertised over different providers. The adjustment of the fine-grained cost metric allows an origin AS to influence the next hop choice of upstream ASes in a way that not all ASes at equal distance will change their next hop choice at the same time (as AS path prepending does). Important to note is that policy is still stronger as policy decisions are placed before the AS-cost evaluation in the decision process. Policy restrictions are unlikely, as the traffic will always be passed on to customers who can be charged and there is no reason to set policy constraints to avoid customer-learned routes. Admittedly, more evaluation is needed to compare such a kind of TE with today's practices.

To evaluate the potential impact of TE, we let every multi-homed AS propagate its reachability information over *every* attached link. We let the state propagate to the closest landmark as seen from the AS that provides the multi-homing. Doing so, the state of 4534 ASes does not end up at an additional landmark. 6599 ASes are known as customers at two landmarks after propagating state over all available links, 734 at three landmarks with the numbers quickly to a maximum of 7 landmarks. The implications of this state are similar to the state discussed in the previous section, i.e. it does not necessarily translate to an overall state increase at landmarks and none of this state will find its way into distant routing tables.

4) Routing Policy

So far, the shortest path matrix of the AS-level Internet topology has served as the basis for our evaluation. But routing

in the Internet does not always follow shortest paths. The reason, as mentioned earlier, is routing policy. When BGP deviates from a simple shortest path, policies have been applied for some reason such as preferring peering links over transit links. We are interested in applying the same policies and observe the effect on 2SIDR.

Therefore, we reconstructed paths as built by the 2SIDR algorithm, but instead of using shortest paths, we use the ones found in the routing tables from RouteViews and RIPE RIS. To achieve this, we needed the tables of both source ASes and of landmark ASes, a limiting factor on the number of paths that we were able to reconstruct. With the given set of routing tables, we were able to reconstruct nearly 3 million paths and evaluated the stretch of 2SIDR's full landmark state scheme. The stretch in the BGP case was calculated by applying the BGP table data to 2SIDR and comparing the resulting paths to the direct path as constructed by BGP, and not the shortest possible path.

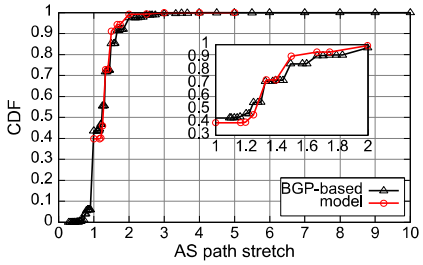

Fig. 7. Model and BGP-based stretch

Fig. 7 shows the stretch statistics of the all paths we were able to reconstruct in comparison to the stretch of the same paths as obtained with our model based on the shortest path matrix. As we are missing many of the routing tables of intermediate ASes, we are missing many of the neighborhood relationships. Consequently, for the comparison against our 2SIDR model, we do not incorporate that state in the routing decisions depicted in Fig. 7. Therefore, the maximum stretch increases significantly for a small fraction of the paths due to the lack of data, not due to the routing algorithm. The overall result in direct comparison is not starkly different confirming a satisfactory accuracy of our previously used model. Interesting to notice is that using the policy as inferred from the routing tables and applying it to 2SIDR, a small fraction of the paths that 2SIDR constructs are in fact shorter than the ones our model finds. The reason is that the policy that has been applied by BGP for this small fraction of ASes has led to paths that did not go through a landmark but to the destination directly. Such routes might become available in 2SIDR once policy is applied as described in section II.B.3). Our model though does not capture such relationships. The consequences of this for our model are positive though as the stretch the model produces is worse than can be expected in the real Internet.

The small number of paths that have a stretch smaller than one at first seem odd. But 2SIDR's routing decision at the source AS is based on the policy applied to the landmark, not to the destination in case it is locally unknown. In certain cases, this can be more efficient in term of AS hops, i.e. the path found this way, by going through the landmark is shorter than the direct path as constructed using BGP resulting in a stretch factor smaller than 1. The policy that has such longer paths as a result is likely to provide economic benefits. Again, by distributing state as described in section II.B.3), these paths can be and will be made available in 2SIDR when being deployed.

IV. Related Work

2SIDR was initially inspired by graph theoretic work carried out by A. Brady and L. Cowen [11] which suggests that routing in the densely connected core of a power-law graph should be different from routing in its edges. Other results described in [21] find as well that selecting landmarks according to their degree for another compact routing schemes results in an optimal state/stretch trade-off. Despite some similarities, 2SIDR differs from common compact routing solutions in a number of aspects. For one, 2SIDR does not make assumptions about the availability of topological knowledge beyond what is known today by BGP routers. In addition, the construction of the routing topology does not aim primarily at optimizing the routing efficiency strictly based on distance, but rather it is guided by the business relationships between and the reachability of ASes in the Internet today.

A large amount of inter-domain routing approaches have been investigated in the recent past that share 2SIDR's goals. Out of those, HLP [10] is the most closely related approach. HLP splits up the routing problem into a link-state part in edge regions and into a path vector part between edges in order to achieve its goals. HLP's route propagation model reduces the amount of information about destinations in distant edge regions, which are only represented as fragmented path vectors. In contrast, 2SIDR still relies on path vectors in edge regions to create strict customer-provider hierarchies. Furthermore, 2SIDR's route propagation model makes explicitly use of the underlying graph which HLP does not mandate this specifically. 2SIDR further does not propagate *any* information about distant ASes into remote edge regions to reduce the amount of state and churn even further. On a general note, at this stage 2SIDR's protocol implementation has deliberately been only vaguely described as we first aimed at studying the properties of the general approach not protocol specifics which are still part of our future work.

Routing is currently also hotly debated in the context of the IRTF's Routing Research Group (RRG). The approaches developed within that forum, such as LISP [26] to name just one, appear to fall into two distinguishable categories termed *separation* and *elimination* as described in [25]. 2SIDR falls into the former category as it separates routing granularity and approaches within a particular AS from routing across domains. 2SIDR differs in many respects from the approaches developed within the RRG. Most prominently 2SIDR is building upon the topological properties of the AS-level graph, something none of the many approaches brought forward by the RRG are considering. Also, AS-based addressing is not particularly interesting to that group as more near term engineering solution are preferred. Nevertheless, there are synergies between the RRG efforts and 2SIDR. As already mentioned we have not evaluated all detailed protocol design options for 2SIDR. For example, instead of changing end hosts

to retrieve 2SIDR-specific resource records from the DNS the ingress/egress routers of the ASes could include/remove 2SIDR shim headers which would require a mapping system. A number of those are being developed within the RRG.

Other efforts try to scale routing within the current framework of routing and addressing. Such solutions include e.g. CRIO [28] and virtual aggregation [27]. CRIO separates topological significance from prefixes by assigning virtual prefixes to routers and building inter-provider tunnels between them. Virtual aggregation even allows to merely "configure" scalability into routers today. While such approaches provide superior migration benefits when compared to 2SIDR, they do not address all scalability issues to the extend 2SIDR does such as the restriction of churn to confined regions of the network. Additionally, abandoning aggregatable addressing allows for improvements at the routing layer that goes beyond scalability. As only one example, it becomes possible to implement security at the routing layer as described in [17]. Generally speaking this is another typical engineering trade-off between ease of migration and potential effectiveness and benefit.

V. Conclusion

Over the past years it has become quite clear that the Internet's inter-domain routing protocol and its addressing scheme have serious shortcomings. In this paper, we presented 2SIDR, a new routing protocol that tackles these issues. Based on insights from the field of compact routing but avoiding its practical limitations, 2SIDR leverages the Internet's topology by using different routing strategies in its densely-meshed core and sparser edge regions.

In order to demonstrate how effective 2SIDR would be on the Internet, we used large sets of data in our analysis. We analyzed two variants of 2SIDR with respect to a number of key characteristics including path stretch and state requirements. Although 2SIDR's performance is as expected lower than its graph theoretical counter parts, we found it to perform very well. 2SIDR was found to have acceptable stretch characteristics and it achieves a significant reduction of routing state. It furthermore shields regions of the network from distant routing events. In addition, we discussed how multi-homing would work under this new scheme and we looked at the effect of policy as observed today.

We are aware that a 2SIDR-like approach will not be easy to deploy in the Internet and we are under no illusion that 2SIDR will be deploy as described here. Rather than an exact blueprint for a BGP replacement, we used 2SIDR as an example to argue on a much broader level. We argued that it seems wrong to have a universal routing scheme that treats all parts of the network the same. Clearly, in the Internet today, ASes have different roles, which is not directly reflected in the routing system. Additionally, the most basic routing that provides global reachability should be based on economic relationships between ASes especially customer-provider relationships. The dissemination of additional routing state needs to be possible but outside this most basic routing approach encouraging an economic balance between the parties that bear the cost and the ones that benefit. We further tried to demonstrate that, at the inter-domain level, aggregatable addressing does not need to be the sole means to achieve scalable routing. What we sincerely hope that 2SIDR achieves is that these aspects will influence future protocol design in fora such as the RRG.

There are still a number of questions that need to be answered in greater detail such as a detailed analysis of traffic engineering and specific protocol design, but in sum, we believe that the fundamental principles implemented in 2SIDR should be part of an effective solution to the problems currently facing the Internet's routing scheme.

Acknowledgment

The author would like to thank Pierre Francoise, Felipe Huici, Iljitsch van Beijnum, Olivier Bonaventure and Thomas Zahn for the review of an earlier version of this paper.

References

[1] G. Houston, BGP Data, http://bgp.potaroo.net/.
[2] Internet Topology Map, http://irl.cs.ucla.edu/topology/.
[3] B. Zhang et al., "Collecting the Internet AS-level Topology," ACM CCR, special issue on Internet Vital Statistics, 2005.
[4] V. Fuller, "Scaling of Internet Routing and Addressing", APRICOT 2007.
[5] L. Kleinrock, F. Kamoun, "Hierarchical routing for large networks: Performance evaluation and optimization", Computer Networks, 1:155–174, 1977.
[6] Q. Chen et al., "The origin of power law in Internet topologies revisited", IEEE INFOCOM 2002.
[7] T. Bu, L. Gao, D. Towsley, "On Characterizing BGP Routing Table Growth", IEEE Global Internet 2002.
[8] C. Labovitz, A. Ahuja, A. Bose, F. Jahanian, "Delayed Internet Routing Convergence", ACM SIGCOMM 2000.
[9] W. B. Norton, "The Art of Peering: The Peering Playbook", white paper.
[10] L. Subramanian et al., "HLP: A Next Generation Inter-domain Routing Protocol", SIGCOMM 2005.
[11] A. Brady, L. Cowen, "Compact Routing on Power Law Graphs with Additive Stretch", ALENEX 2006.
[12] D. Krioukov, K. Fall, X. Yang, "Compact Routing on Internet-Like Graphs", INFOCOM 2004.
[13] D. Krioukov, K. Fall, kc claffy, A. Brady, "On Compact Routing for the Internet", ACM SIGCOMM CCR, 2007.
[14] Route Views Project, http://www.routeviews.org/.
[15] NEC Labs Topology Project, http://topology.neclab.eu/
[16] P. Verkaik et al., "Beyond CIDR Aggregation", CAIDA technical report TR-2004-01.
[17] D. Andersen et al., "Holding the Internet Accountable", HotNets 2007.
[18] M. Caesar, J. Rexford, "BGP routing policies in ISP networks", IEEE Network Magazine 2005.
[19] I. van Beijnum, "A BGP Inter-AS Cost Attribute", I.D. work in progress http://tools.ietf.org/html/draft-van-beijnum-idr-iac.
[20] L. Gao, F. Wang, "The Extent of AS Path Inflation by Routing Policies", IEEE Global Internet Symposium, 2002.
[21] M. Enachescu, M. Wang, A. Goel, "Reducing Maximum Stretch in Compact Routing", IEEE INFOCOM 2008.
[22] Routing Information Service, http://www.ripe.net/ris/.
[23] L. Gao, "On Inferring Automonous System Relationships in the Internet", IEEE Global Internet, November 2000.
[24] H. Tangmunarunkit et al., "Does AS Size Determine Degree in AS Topology?", ACM CCR 2001.
[25] D. Jen et al., "Towards a New Internet Routing Architecture: Arguments for Separating Edges from Transit Core", HotNets 2008.
[26] D. Farinacci et al., "Locator/ID Separation Protocol (LISP)", I.D. work in progress, http://tools.ietf.org/html/draft-farinacci-lisp
[27] H. Ballani et al., "ViAggre: Making Routers Last Longer!", HotNets 2008
[28] X. Zhang et al., "Scaling Global IP Routing with the Core Router-Integrated Overlay," ICNP 2006