Deployment and Scalability of an Inter-Domain Multi-Path Routing Infrastructure

Cyrill Krähenbühl
ETH Zürich

Seyedali Tabaeiaghdaei
ETH Zürich

Christelle Gloor
ETH Zürich

Jonghoon Kwon
ETH Zürich

Adrian Perrig
Anapaya Systems,
ETH Zürich

David Hausheer
OVGU Magdeburg

Dominik Roos
Anapaya Systems

ABSTRACT

Path aware networking (PAN) is a promising approach that enables endpoints to participate in end-to-end path selection. PAN unlocks numerous benefits, such as fast failover after link failures, application-based path selection and optimization, and native inter-domain multi-path. The utility of PAN hinges on the availability of a large number of high-quality path options. In an inter-domain context, two core questions arise. Can we deploy such an architecture *natively* in today's Internet infrastructure without creating an overlay relying on BGP? Can we build a scalable multi-path routing system that provides a large number of high-quality paths?

We first report on the real-world native deployment of the SCION next-generation architecture, providing a usable PAN infrastructure operating in parallel to today's Internet. We then analyze the scalability of the architecture in an Internet-scale topology. Finally, we introduce a new routing approach to further improve scalability.

CCS CONCEPTS

• **Networks** → **Network design principles**; **Control path algorithms**; **Network simulations**; **Routing protocols**.

KEYWORDS

SCION, Next-generation Internet Architecture, BGP, Deployment, Scalability, Control-Plane Algorithm Design, Multi-Path, Inter-Domain Routing, Network Simulation

ACM Reference Format:
Cyrill Krähenbühl, Seyedali Tabaeiaghdaei, Christelle Gloor, Jonghoon Kwon, Adrian Perrig, David Hausheer, and Dominik Roos. 2021. Deployment and Scalability of an Inter-Domain Multi-Path Routing Infrastructure. In *The 17th International Conference on emerging Networking EXperiments and Technologies (CoNEXT '21), December 7–10, 2021, Virtual Event, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3485983.3494862

1 INTRODUCTION

Path aware networking (PAN) is a promising trend in networking [22], where endpoints are given more information and control about the paths their packets traverse. This is unlike the current Internet, where path selection is performed implicitly within the network. PAN enables exciting opportunities to evolve the Internet: multi-path communication can harness inherent Internet path diversity, application-based path selection allows endpoints to influence routing and choose optimal paths, and rapid failover can mask link failures.

Although intra-domain multi-path techniques are already in use, a large body of inter-domain multi-path research [5, 18, 19, 25, 30, 33, 39, 40, 59–62] has not seen large-scale deployment, mainly due to the challenges of deploying inter-domain protocols. Deploying a new inter-domain multi-path architecture requires a sizable financial investment, and the architecture can only gain real-world traction with tangible incentives for early adopters [55]. The PAN Internet architecture SCION is the first inter-domain multi-path architecture in practical use. In this work, we revisit the deployment concept that underlies the SCION production network, along with the incentive scheme it provides to early adopters. In this context, we will answer the following questions:

(1) What are the (early adopters') incentives for using SCION?
(2) How could the deployment of SCION start and grow given the pre-eminent BGP-based infrastructure?
(3) Can SCION scale to the size of the Internet?

To understand why SCION was deployed and how it can be successful in the future, we must answer the first question about the incentives for initial and future SCION adopters. In particular, we discuss both short- and long-term incentives that play a role in the early deployment.

To answer the second question about deployability, we show how SCION has been deployed in production networks in an overlay-free manner, co-existing side-by-side with BGP, without relying on it. Furthermore, as currently eight Internet service providers (ISPs)—with a total market capitalization exceeding $40B—are already offering native SCION connectivity, and real-world traffic is being transported on the network, we document different models that were used for SCION deployment in ISPs and Internet exchange points (IXPs).

With SCION's expansion, its scalability needs to be assessed. This motivated the third question and lead us to analyze the scalability of different SCION components. We find that the path construction / exploration process has the largest impact on scalability. Therefore, we propose a new approach for constructing paths which capitalizes on the extensibility of the SCION control plane. Based on simulations on a realistic large-scale topology, we show that the new approach not only drastically improves the scalability of

CoNEXT '21, December 7–10, 2021, Virtual Event, Germany

ACM ISBN 978-1-4503-9098-9/21/12…$15.00
https://doi.org/10.1145/3485983.3494862

SCION's path construction process, but also finds higher quality paths. The path construction process can be adapted to any given optimality criteria, for instance link disjointness, which more than doubles the resilience of the set of disseminated paths against link failures, compared to the baseline approach.

The main contributions of this work are as follows:

- We report on the real-world deployment of a next-generation Internet architecture in production networks.
- We introduce a new path construction process for SCION, which finds paths that are significantly more resilient to link failures and reduces communication overhead of the beaconing by two orders of magnitude compared to the current path construction process, thus improving scalability.
- We compare it to BGP and BGPsec, and show that it reduces the overhead per constructed path by two and three orders of magnitude, respectively.

Since we use simulations without any personalized measurements such as traffic traces, this work does not raise any ethical issues.

2 BACKGROUND ON SCION

SCION is a next-generation Internet architecture, offering high availability even in the presence of network adversaries. In this Section, we briefly present the key concepts of the SCION architecture, including control- and data-plane features that are relevant to follow the paper. Further details are available in the SCION book [38] or article [8].

2.1 Network Structure and Naming

To maintain the economic principles of today's Internet, SCION reuses the Autonomous Systems (AS) structure, and ensures that network traffic only flows on policy-compliant paths. To achieve scalability and sovereignty, **Isolation Domains (ISD)** are introduced. An ISD groups ASes that agree on a set of trust roots, called the **Trust Root Configuration (TRC)**. An AS can be a member of multiple ISDs. The ISD is governed by a set of **core ASes**, which provide connectivity to other ISDs and manage the trust roots. Typically, the 3–10 largest ISPs of an ISD form the ISD's core. Figure 1 shows a SCION network with 3 ISDs, each containing 2 or 3 core ASes.

Routing is based on the ⟨ISD, AS⟩ tuple, agnostic of local addressing. Existing AS numbers are inherited from the current Internet, but a 48-bit namespace allows for additional SCION AS numbers beyond the 32-bit space in use today. Host addressing extends the network address with a local address, forming the ⟨ISD, AS, local address⟩ 3-tuple. The local address is not used in inter-domain routing or forwarding, does not need to be globally unique, and can thus be an IPv4, IPv6, or MAC address, for example.

2.2 Control Plane

The SCION control plane discovers and distributes AS-level **path segments**. A path segment encodes a network path at the granularity of inter-domain interfaces on either end of an inter-domain link connecting two consecutive ASes on a path. Constructing inter-domain paths at the granularity of inter-AS links increases the number of available paths and enables optimization of paths with regard to different criteria.

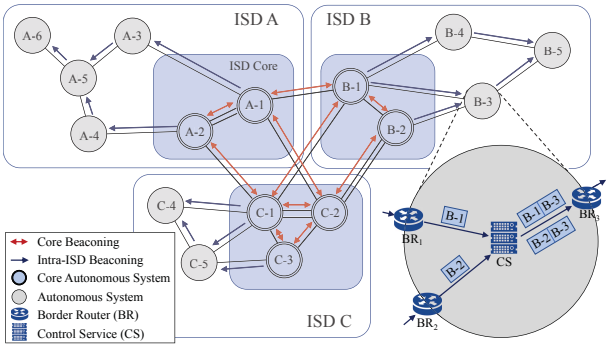


Figure 1: Core and intra-ISD beaconing in a SCION Network.

Each end-to-end path consists of up to three path segments: core-path, up-path, and down-path segments. **Core-path segments** refer to path segments containing only core ASes, an **up-path segment** is a path segment from a customer (leaf AS) to a provider (core AS) inside one ISD, and a **down-path segment** is a path segment from a provider (core AS) to a customer (leaf AS) inside an ISD. For example, if an endpoint in B-3 in Figure 1 connects to an endpoint in A-6, then a possible path consists of: up-path (B-3, B-2), core-path (B-2, B-1, A-1, A-2), and down-path (A-2, A-4, A-5, A-6). To address the suboptimality of hierarchical routing, SCION introduces **peering links** and **shortcuts**. In a shortcut, a path only contains an up-path and a down-path segment, which can cross over at a non-core AS that is common to both paths. Peering links can be added to up- or down-path segments, resulting in an operation similar to today's Internet.

Routing (or path segment construction) is conducted hierarchically on two levels: (1) among all core ASes of all ISDs, which constructs core-path segments, and (2) within each ISD, which constructs up- and down-path segments. The path segment construction process is referred to as **beaconing**, where a **Path-segment Construction Beacon (PCB)** is initiated by core ASes to iteratively construct path segments. Up- and down-path segments are interchangeable, simply by reversing the order of ASes in a segment.

Figure 1 depicts the core and intra-ISD beaconing using the red double-headed arrows and blue arrows, respectively. The beaconing process in each AS is performed by its **beacon server** which is a part of its **Control Service (CS)**, that performs control-plane-related tasks. The beacon server decides which PCBs to propagate on which interfaces based on AS-local policies. Before propagating a PCB, the beacon server appends its AS number and the incoming and outgoing **interface identifiers** of the links connecting to the neighbor ASes. Additionally, each PCB has an expiration timestamp which is specified by the initiator of the PCB, to indicate the validity period of the path. It is important to note, that only control plane messages are processed by the control service, i.e., the data plane scales independently from the control plane.

(1) Core Beaconing. **Core beaconing** is the process of constructing path segments between core ASes. During core beaconing, a core AS either initiates PCBs or propagates PCBs received from neighboring core ASes to all other neighboring core ASes. Since the number of ISDs is expected to be in the hundreds and the number

of core ASes per ISD is typically around 3–10, the total number of entities participating in core beaconing is relatively small.

(2) Intra-ISD Beaconing. **Intra-ISD beaconing** is the second level of the beaconing hierarchy, which creates path segments from core ASes to non-core ASes. For this, core ASes create PCBs and send them to their non-core neighbors (typically customer ASes). Each non-core AS propagates the received PCBs to its respective customers. This procedure continues until the PCB reaches an AS without any customer (leaf AS) and as a result, all ASes receive path segments to reach the core ASes of their ISD. This policy-constrained flooding is highly efficient, as only core ASes initiate PCBs. Non-core ASes can include their peering links in the PCBs, enabling valley-free forwarding if both up- and down-path segments contain the same peering link.

Path Segment Dissemination. A global **path server** infrastructure is used to disseminate path segments. Each AS contains a path server as a part of the control service. The infrastructure bears similarities to DNS, where information is fetched on-demand only. A core AS's path server stores all the intra-ISD path segments that were registered by leaf ASes of its own ISD, and core-path segments to reach other core ASes.

2.3 Data Plane

Name resolution in SCION returns the ⟨ISD, AS, local address⟩ 3-tuple. Core- and down-path segments are fetched based on the ⟨ISD, AS⟩ tuple. Hosts can then combine one of their up-path segments with the received core- and down-path segments. Shortcut paths that avoid a core AS are possible, if the up- and down-path contain the same AS, or if a peering link is available between an AS in the up-path and an AS in the down-path segment. Cryptographic protections ensure authentic path segments and prevent unauthorized path combinations.

The path segments contain compact **hop-fields**, that encode information about which interfaces may be used to enter and leave an AS. The hop-fields are cryptographically protected, preventing path alteration. This so-called **Packet-Carried Forwarding State (PCFS)** replaces signaling to use a path, ensuring that routers do not need any local state on either paths or flows.

3 CASE STUDY: SCION DEPLOYMENT

This Section describes how SCION is deployed in production networks and used for real-world traffic at IXPs, ISPs, and end domains. Before we describe the technical deployment details, we first discuss the stakeholder incentives that led to the first production use of SCION in 2017, and then briefly discuss deployment considerations that are needed to achieve the salient SCION properties.

3.1 Stakeholder Incentives

An important aspect for the deployment of a new Internet architecture are the incentives that drive initial deployment. Similar to the question of "Who bought the first fax machine?", the case for a next-generation architecture is even more challenging, given the plethora of commercial communication offerings.

The initial customer incentive was to test the reliability of SCION and to use a SCION connection to replace a leased line. A leased line—often provisioned via dedicated layer-2 circuit switching or layer-3 MPLS—is a premium connectivity service that provides availability and confidentiality. On the other hand, leased lines often have long lead times (in some cases several months), lack flexibility for short-term changes, and are often expensive to operate. SCION approximates leased line properties, offering geofencing, path transparency, high reliability thanks to fast failover, and flexibility for changes. Furthermore, as SCION adoption grows and converges towards today's pricing, costs will be reduced compared to leased lines in the long term. For instance, to connect N branches with K data centers, which can be implemented using $N \cdot K$ leased lines, $N + K$ SCION connections are required (and for even larger savings if redundancy is needed). Since SCION can reuse the existing IP or MPLS-based network, the additional capital and operational expenditures to run SCION are marginal, requiring only a few standard servers or VMs.

The long-term incentives for using SCION are to achieve higher performance and quality of communication through the use of multi-path and optimized path selection based on application requirements (e.g., latency, bandwidth, jitter, or loss).

Since August 2017, SCION has been in production use by a central bank, with two main goals: test the long-term reliability of SCION, and replace leased lines. Over time, several of their branches have been connected to their data centers over the SCION network. Their positive experiences have fueled adoption by ISPs, as well as by commercial, education, and government entities. Today, eight ISPs offer SCION connections, and several banks and government entities benefit from the BGP-free backbone for production use. This demonstrates that the initial deployment incentives have been sufficient, but additional incentives are needed to further drive deployment to ultimately reach wide-spread native SCION connectivity on endpoints used by applications. Other use cases that will likely benefit from SCION's path awareness and multi-path properties include: industrial control systems ([23], [24]), bulk file transfers, CO2 optimized routing [44], access to cloud environments, communication infrastructure for blockchain systems, and many more.

3.2 Deployment Considerations

An overlay deployment on today's Internet was not desirable as SCION would inherit the vulnerabilities of the weak underlay. Thus, a challenge was to deploy SCION in parallel to existing networks in an economically viable way, while preserving the security properties. In particular, there should not be any dependence on BGP for the SCION network to operate, which we refer to as a "BGP-free" deployment.

Since deploying a completely new network infrastructure is prohibitively expensive, internal AS networks are re-used. However, care needs to be taken that traditional IP traffic cannot be used to crowd out SCION traffic, for instance by causing IP-level congestion.

3.3 ISP Deployment

As Figure 1 depicts, an ISP deploying SCION needs to set up border routers and run instances of the control service. The border router and control service instances are deployed on standard x86 commercial off-the-shelf (COTS) servers, supporting up to 100 Gbps connections, while with P4 hardware it is possible to forward SCION

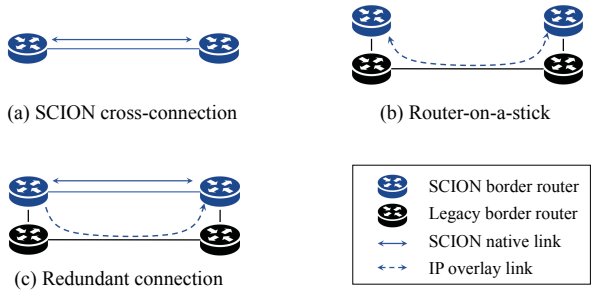


Figure 2: ISP deployment scenarios. The various deployment models support early, intermediate, and full deployment cases.

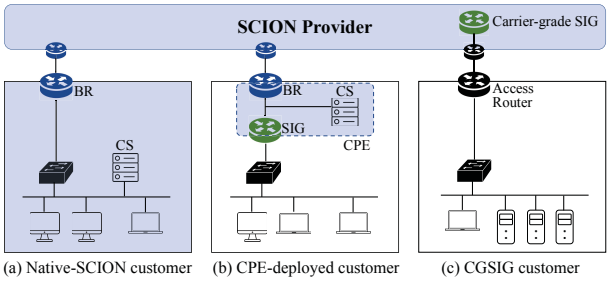


Figure 3: Example deployments for end domains. With the SCION-IP Gateway, end domains enable SCION connections, without requiring any changes to end-hosts or applications.

traffic even at terabit speeds [13, 14]. The internal IP or MPLS-based network can be re-used to enable the SCION infrastructure to communicate within the AS. If dedicated links are not available, queuing disciplines on internal switches can provide separation of IP and SCION traffic.

Customer connections and SCION connectivity between the border routers of neighboring ISPs can be achieved in three different ways. Ideally, SCION-enabled adjacent ISPs would be connected via a native SCION link (Figure 2a). That is, two SCION border routers are directly connected via a layer-2 cross-connection at the same point-of-presence location, achieving connectivity with high reliability, availability, and performance. The native SCION link is unaffected by BGP failures, achieving a "BGP-free" deployment.

To minimize changes to the current infrastructure, ISPs may also reuse existing cross-connections to carry SCION traffic, e.g., in a *Router-on-a-stick* deployment model. As shown in Figure 2b, the SCION border routers can be attached to the existing legacy border routers. A SCION border router encapsulates SCION packets into IP packets and forwards them to a neighboring SCION border router over a short IP connection, which can be "BGP-free" through the setup of host routes. The main advantage of this deployment model is that ISPs can simultaneously use their network infrastructure for the new network architecture. Since the *Router-on-a-stick* model is a short-range direct cross-connection, potential shortcomings of using IP encapsulation, such as non-optimal routing, BGP hijacking, and slow route convergence, are typically not an issue. Given that an adversary could overload the shared link with IP traffic, it is important to define a queuing discipline on the link to ensure that SCION traffic obtains at least a minimum fraction of the link bandwidth to achieve availability properties.

Finally, Figure 2c shows the deployment of a *redundant connection* model, combining the aforementioned two deployment models. The two links can be transparently combined into one logical single link, or exposed as two separate links with different SCION interface numbers, enabling multipath selection for either of the links.

SCION-enabled ISPs should seamlessly communicate with each other even in partial-deployment scenarios; i.e., two SCION-enabled ASes may not be neighbors. Bridging two SCION islands – e.g., by creating an IP tunnel to forward SCION packet through the public Internet – however, introduces BGP vulnerabilities. To this end, Anapaya has established the SCION-transit service [53], a global backbone service for SCION-enabled ISPs. The SCION-transit service provides native SCION connectivity at 100+ data centers located at the largest metropolitan areas across the world. With such distributed points-of-presence, ISPs can readily establish one-hop access to the SCION-transit service, forwarding SCION traffic through the BGP-free network.

3.4 End Domain Deployment

A customer can use SCION in two different ways: (1) native SCION applications, and (2) transparent IP-to-SCION conversion. The benefit of using SCION natively is that the full range of advantages becomes available to applications, at the cost of installing the SCION endpoint stack and making the application SCION-aware. In the short term, approach (2) is preferred, leveraging a *SCION-IP-Gateway* (*SIG*) that encapsulates regular IP packets into SCION packets with a corresponding SIG at the destination that performs the decapsulation.

In the deployments, end domain customers need to decide if their domain becomes a SCION AS (Cases a & b), or if they connect to the provider's AS (Case c).

Case a: Native SCION Customer. So far, all deploying entities elected to become their own AS, as no complex routing (policy) configurations are needed. The required cryptographic certificates are issued by the core ASes, and the AS numbers are re-used from today's AS numbers or, if needed, allocated from the larger 48-bit space of SCION AS numbers.

As shown in Figure 3a, native SCION hosts can send SCION traffic directly to a SCION border router (BR) over the existing AS-internal routing infrastructure. Native SCION hosts are equipped with the SCION stack components, enabling applications to generate SCION packets. The data-plane component (i.e., *SCION dispatcher*) dispatches packets to the corresponding application and performs packet transmission. The control-plane component (i.e., *SCION daemon*) communicates with the AS's *control service* (*CS*) to build end-to-end forwarding paths for applications on their behalf.

Case b: SIG-based deployment. We understand that many customer hosts may initially not be SCION capable. Customers purchasing a SCION-connection from a provider ISP therefore obtain a *customer-premise equipment* (*CPE*) that provides the functionality

of the SIG, BR, and CS. Figure 3b depicts a high-level topology of an end-customer network SCION-enabled with a CPE; the SIG enables legacy hosts to opt into the SCION network.

The SIG is responsible for encapsulating legacy IP packets in SCION packets, to provide interoperability between SCION and legacy networks. When the SIG receives an outgoing packet, it first determines the SCION AS to which the destination IP address belongs. For the mapping between IP address space and ASes, the SIG keeps the *ASMap* table [47]. The SIG then obtains paths to the remote AS from the control service, encapsulates the packet with a SCION header, and routes it via a BR.

Case c: Carrier-grade SIG customer. End domain customers can also be SCION-enabled with a *carrier-grade SIG* (*CGSIG*) as depicted in Figure 3c, requiring no changes to the customer premises; the CGSIG operated by the provider ISP aggregates upstream traffic towards remote ASes and carries out SCION packet routing on behalf of its customers, while legacy hosts residing in the end-domain networks remain SCION-unaware. The CGSIG-driven SCION service is designed to minimize the impact on existing infrastructure and is suitable for small business and home office users.

Internal Routing of SCION Traffic. To transport SCION packets to an egress BR, the customers do not need to change their internal routing infrastructures; the SCION packets are IP-routed by IGP, e.g., OSPF or IS-IS. Given that the AS's internal entities are considered to be trustworthy, the IP overlay for the first-hop routing does not compromise or degrade any properties SCION delivers. To exchange SCION packets with the provider network, the customer-side SCION border routers directly connect to the provider-side border routers using, e.g., fiber cables or layer-2 cross-connections. It is important to note that the current customer connections to the SCION production network are native SCION connections, not shared IP / SCION connections, i.e., while IP is used to route SCION traffic AS internally, those existing SCION customer connections *cannot* forward regular IP packets.

Real Deployment Case. Under the SCI-ED project (2019–2021) [35], SWITCH (Swiss education and research network ISP) provides SCION service to the ETH domain institutions (e.g., ETHZ, EPFL, PSI, and CSCS) and all Swiss universities [52]. Swisscom [51] and Sunrise [50] are the pioneer commercial ISPs that first introduced SCION connectivity as a premium Internet service. Additional ISPs are pilot testing the technology. In the SCI-ED deployment, research institutes were connected via Case a as depicted in Figure 3, while commercial deployments are typically deployed with a setup following Case b. Several end-domain customers, such as banks or government offices, are connected to the provider network through either a layer-2 circuit or an IP link (recall that the first-hop overlay does not hamper the SCION properties), and benefit from the SCION service the ISPs offer. As of late 2021, the customer SCION service is available worldwide, with denser connectivity in Europe and Asia.

3.5 IXP Deployment

Internet Exchange Points (IXP) play an important role in today's Internet, as they let ISPs, content delivery networks (CDNs), and other providers exchange traffic with each other. We envision two

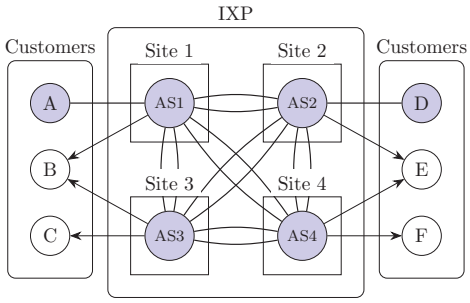


Figure 4: IXP deployment. Shaded circles denote Core ASes.

models describing how the role of IXPs can be reflected in the SCION infrastructure: either as "big switch" or by exposing their internal topology. In the big switch model, IXPs would be considered as a large L2 switch between multiple SCION ASes (i.e., customers of the IXP). The role of the IXP is then to facilitate bilateral (peering) links among those ASes. This role is entirely transparent to the SCION control plane. Today, SwissIX already follows this model by offering a dedicated SCION VLAN on which it prohibits non-SCION traffic. Although SCION does not yet offer native multilateral peering support for the big switch model similar to a BGP route server today, the automatic interconnection of SCION ASes over an IXP can be facilitated with a SCION Peering Coordinator [45].

Figure 4 shows an enhanced model in which the internal topology of an IXP is exposed within the SCION control plane. Here, the IXP operates its own SCION ASes, whereas each AS represents an IXP site and the links between them represent redundant connections between these sites. This enhanced model enables IXP customers to use SCION's multi-path and fast failover capabilities to leverage the IXP's internal links (including backup links) and to select paths depending on the application's needs (e.g., to optimize latency or bandwidth through the IXP's network). This model would entirely replace the IXP's traditional interconnection fabric that is mostly based on Ethernet switching or MPLS today. We believe that IXPs have an incentive to expose their rich internal connectivity as the benefits from SCION's multi-path capabilities would increase their value for customers and provide them with a competitive advantage.

Real Deployment Case. In 2019, the concepts of a Secure Swiss Finance Network (SSFN) based on SCION have been worked out and, throughout 2020 and the first half of 2021, a pilot has been conducted to evaluate the feasibility and effectiveness of the SCION-based SSFN, that culminated in the official announcement of the SSFN in July 2021 [7]. The core network is provided by three independent ISPs (i.e., Sunrise, Swisscom, and SWITCH). Participants connect to the SSFN via one or more of the ISPs and can communicate with every other participant on the network. To this end, the Swiss Internet Exchange (SwissIX) currently interconnects major Swiss ISPs, the Swiss Stock Exchange (SIX), and other enterprises through bi-lateral peering over a dedicated SCION port. SwissIX also provides access to the SCION transit service provided by Anapaya [53]. This deployment currently follows the traditional "big switch" model.

4 SCALABILITY

The SCION production network, currently operated by Anapaya [2], provides native SCION connectivity to customers in collaboration with 8 ISPs, operating side-by-side with existing networks. The current production network explores paths at a low control plane overhead, which is unsurprising given the relatively small size. However, as the network grows and more (core) ASes join, we need to ensure scalability up to an Internet-scale deployment.

In this Section, we first study the scalability aspects of the SCION control plane and show, that topology exploration (beaconing) has the highest overhead among all control plane components: beaconing, path lookup, revocation, and (de-)registration. Then, we show that although the current topology exploration is scalable, it introduces overhead comparable to BGPsec. Fortunately, topology exploration is based on *AS-local* decisions and is designed to be *extensible*. This extensibility allows us to propose a novel approach for path exploration that significantly reduces the overhead, while also improving the quality of the disseminated paths.

4.1 SCION Scalability Analysis

A multi-path routing architecture disseminates multiple paths per destination and thus increases the number of sent control plane messages and the memory requirements for routing tables. In SCION, scalability to an Internet-wide deployment in terms of (I) processing, (II) communication, and (III) state overhead is achieved through the following mechanisms:

(1) Grouping ASes into ISDs (I & II).
(2) Eliminating routing protocol convergence towards stable state, i.e., disseminated paths are stable upon dissemination (I & II).
(3) Routing is performed at the granularity of AS (+ interface[1]) level paths compared to per-prefix routing (I, II, & III).
(4) SCION border routers are simple by design. Packet-Carried Forwarding State (PCFS) removes the need for large inter-domain forwarding tables on routers. Additionally, routers only perform packet forwarding and no control-plane functionalities. This separation allows border routers, which do not require any global state, to focus on efficient packet forwarding (III).
(5) Different control plane strategies on each of the two levels of the routing hierarchy: selective flooding among core ASes (core beaconing) and uni-directional intra-ISD forwarding (intra-ISD beaconing) (I & II).
(6) Parallel to the push-based beaconing, a separate path-server infrastructure operates a pull-based path segment lookup with caching, without the need for global broadcast or dissemination (II).

Several mechanisms, e.g., Mechanisms 1 and 3, have already been proposed by other inter-domain routing protocols [1, 21, 49, 60] and have been shown to improve scalability. We consider each control-plane operation to determine which aspect to focus on for the scalability analysis. Table 1 shows an overview, guiding the discussion.

[1] A path segment in SCION is described by the inter-domain interfaces of the outgoing and incoming border routers of two neighboring ASes (see Section 2.2).

Table 1: Path Management Overhead Comparison

SCION Control Plane Component	Scope: AS	Scope: ISD	Scope: Global	Frequency: Hours	Frequency: Minutes	Frequency: Seconds
Core Beaconing			●		●	
Intra-ISD Beaconing		●			●	
Down-Path Segment Lookup			●			●
Core-Path Segment Lookup		●			●	
Endpoint Path Lookup	●					●
Path (De-)Registration		●		●		
Path Revocation			●	●		

Core Beaconing. In SCION, each ISD is governed by a few "core" ASes, which provide connectivity within and between ISDs (see Section 2.2). In core beaconing, path segments between core ASes are disseminated through *selective flooding*, i.e., an AS selects a subset of received PCBs for each outgoing interface, signs, and forwards them. Core beaconing potentially has the highest impact on scalability among the control plane components. As the number of possible paths in a large densely-interconnected topology can be extensive, disseminating all received PCBs would introduce an overwhelming amount of communication and computation overhead. However, in a topology containing n core ASes, propagating at most a constant threshold k PCBs per origin AS in each beaconing interval results in at most kn PCBs being sent on each interface, an overhead linear in the number of core ASes. The core beaconing is thus scalable even to large topologies with thousands of ASes. Since the current production deployment and the SCIONLab testbed [28] are too small to infer scalability, we use simulations on large topologies to study the communication overhead of the core beaconing. Section 5.2 shows that the overhead of the baseline core beaconing algorithm is in the same order as BGPsec. With the improvements we propose in Section 4.2, we reduce this overhead by more than two orders of magnitude, resulting in a one order of magnitude *lower* overhead than BGP even for finding 60 paths between any pair of ASes.

Intra-ISD Beaconing. Intra-ISD beaconing is initiated by the core ASes, which disseminate PCBs uni-directionally to the leaf ASes (see Mechanism 5). PCBs are forwarded along provider-customer links, limiting the number of PCBs and leading to an overhead linear in the number of interfaces. Since the number of PCBs received by non-core ASes in an ISD only depends on the topology of that ISD, regardless of the size and topology of the entire network, intra-ISD beaconing is scalable to any network structured like today's Internet (see Mechanism 1). In particular, the overhead of intra-ISD beaconing is two orders of magnitude lower than BGP, as we show in Section 5.2.

Down-Path Segment Lookup. Down-path segment lookup consists of path servers fetching down-path segments from other ISDs, to enable construction of end-to-end data plane paths. Fetching path segments is a unicast operation to the origin AS's path server and is amortized by the large amount of data-plane traffic. To further reduce overhead, path servers and endpoints cache path segments to serve subsequent requests for a given origin AS, which is effective in SCION due to the long lifetime of a path (on the order of several hours). Additionally, due to the Zipf distribution of Internet traffic's

destinations [34], scalability is further improved by caching path segments for popular origin ASes, such as CDN providers.

Core-Path Segment Lookup. A Core-path segment lookup consists of path servers (in non-core ASes) fetching path segments between core ASes from a core AS within their ISD. In addition to the previously mentioned points for down-path segment lookups, a core-path segment lookup requires only intra-ISD communication. The communication overhead thus scales with the size of the core network and the size of the ISD (and not the size of the global network).

Endpoint Path Lookup. An endpoint path lookup consists of endpoints fetching path segments from their local path server, which is an intra-AS operation and thus not influenced by the size of the network.

Path Registration and De-registration. Path (de-)registration is typically performed every tens of minutes and consists of sending around 10 KBytes of data to the core path server in the same ISD. Similar to a core-path segment lookup, communication only occurs between ASes within the ISD.

Path Revocations. Path revocations triggered by failing links have two reactions depending on where the failure occurred. The AS in which the failing link is located revokes the affected path segments at the core path server, which is an intra-ISD operation. Endpoints and border routers that use a path containing a failed link are informed of the link failure through SCION Control Message Protocol (SCMP) messages sent by the border router observing the failed link. The SCMP messages typically produce much less traffic than the data plane traffic prior to the link failure, and hosts switch to a different path as soon as the SCMP message is received.

4.2 Improving Beaconing: Quality-Aware Path Construction

In a multi-path routing architecture, exploring all path combinations is not only unnecessary, but also results in overwhelming communication overhead as the number of paths can be extensive in densely-interconnected topologies. Therefore, a viable path dissemination algorithm must optimize paths to each destination for specific optimality criteria, while keeping the cost in terms of control plane communication overhead as small as possible. Given the relatively small size of the initial SCION production network and SCIONLab testbed, a simple baseline path construction algorithm is used, which optimizes paths for the same metric as BGP, which is (AS) path length. However, there are two shortcomings:

- An AS cannot optimize the disseminated paths for any optimality criteria other than AS-path length, since only the P shortest paths are disseminated at each interval. However, compared to BGP, it is still an improvement as it provides *multiple* shortest paths.
- The algorithm sends a set of paths irrespective of previously sent paths. Paths that are sent too frequently on an egress interface cause unnecessary redundancy and waste bandwidth.

With the increasing size of the SCION production network, a more sophisticated path exploration algorithm is needed to improve scalability and enhance path quality.

Therefore, we introduce a new algorithm, called *Path-Diversity-Based Path Construction Algorithm*, which takes the diversity of paths into account to construct path segments. This algorithm optimizes for link-disjointness of constructed paths by only using AS and interface identifiers already available in PCBs. Henceforth, when mentioning links and link-disjointness, we consider *inter-domain* links between two interfaces of neighboring ASes. Link-disjointness is considered to be an essential property for multi-path network architectures as it increases the reliability of communication [20]. In SCION, an endpoint directly benefits from having a diverse set of paths, since after detecting a link failure (e.g., via an SCMP message), it can immediately switch to an alternative path not containing the failed link. The path-diversity-based path construction algorithm is a distributed greedy algorithm maximizing the disjointness of paths, while reducing the overhead by inhibiting redundant path retransmissions. The rationale of the algorithm is to prefer PCBs with few overlapping links, PCBs containing new links, and PCBs with a long remaining lifetime. This is enabled by keeping track of recently sent PCBs on each egress interface. We choose link instead of AS disjointness as a metric for diversity, since AS failures are unlikely events.

Similar to the baseline path construction algorithm, the path-diversity-based path construction algorithm is triggered periodically by the beacon server to select and disseminate PCBs. The algorithm iteratively tries all combinations of received PCBs and egress interfaces, and selects the k-highest-score combinations (k being a constant threshold) per *[origin AS, neighbor AS]* pair if their scores are above a certain threshold. The score is based on a PCB's age, lifetime, and link disjointness with regard to previously disseminated PCBs for the same *[origin AS, neighbor AS]* pair. Appendix A contains a more detailed description of the algorithm.

Link Diversity Score Calculation. To perform the *link diversity score* calculations, the algorithm stores a *Link History Table* per *[origin AS, neighbor AS]* pair. Each table is a one-to-one map from *link_id*s to their associated *counter*s where the *link_id* is an identifier for a link between two ASes, and the *counter* counts the number of times the link is part of a valid path from the origin AS to the neighbor AS. When a PCB initiated at an origin AS is disseminated to a neighbor on an outgoing link, the associated counters are incremented for every link on its path, as well as the one associated with the outgoing link in the *Link History Table* of that *[origin AS, neighbor AS]* pair. If a link has not been visited on any previously-disseminated PCB's path from the origin AS to the neighbor AS, a new entry for that link is created. With the help of this *Link History Table*, the algorithm calculates the *link diversity score* of a path from an origin AS to a neighbor AS, by finding the geometric mean of the counters of all links on the path. This mean is scaled to the interval $[0, 1]$ by dividing it by the maximum acceptable geometric mean. The geometric mean of link counters shows the degree of jointness of this path with other paths, since the counter of each link is equal to the number of paths having that link in common with the current path. However, the algorithm does not calculate the *link diversity score* of a path if its PCB has previously been sent

and is still valid. Instead, it reuses the *link diversity score* of the path at the time its PCB was sent. To that end, the algorithm stores the *link diversity score* as well as the age and the lifetime of every PCB it disseminates to each egress interface in the *Sent PCBs List* associated with that egress interface. If a path is sent again, its corresponding timers in *Sent PCBs List* get updated.

Final score calculation based on PCB age and lifetime. In SCION, core ASes initiate PCBs periodically. Each PCB has an initiation timestamp and an expiration timestamp. A PCB is valid only between these two timestamps. Therefore, every AS desires to have paths to any destination that are valid at any given time and will not expire in the near future. But, receiving a newer instance of a PCB with the same path as its previous instance is a waste of bandwidth. Therefore, we need a scoring function that minimizes the number of repetitive PCBs that are sent over each link and assures that all ASes always have valid high-quality paths to any origin AS. Therefore, the final score of a PCB is calculated as a function of its path's *link diversity score* and the age and the lifetime of its current instance and its previously-sent instance as shown in Equation (1).

If a PCB has not been sent before, the exponent is proportional to the ratio of the PCB's age to its lifetime as shown in Equation (2). If a PCB has been sent before, the exponent is proportional to a power of the ratio of the remaining lifetime of the previously-sent PCB to the remaining lifetime of the current PCB as shown in Equation (3). The different functions are due to the following three objectives which cannot be satisfied with a single function or two functions of the same form.

- **Preserve connectivity** by prioritizing previously-sent PCBs over not-previously-sent PCBs from the same origin AS, when the previously-sent PCB instance is about to expire.
- **Discover new paths** by prioritizing not-previously-sent PCBs over previously-sent PCBs from the same origin AS whenever the previously-sent PCB's expiration time is far away.
- **Save bandwidth** by not sending recently-sent PCBs by lowering their score.

The parameters α, β, γ, and the score threshold are selected such that the above three objectives are achieved and depend on the topology and the lifetime of a PCB. For a given topology, we find suitable parameters by first performing a grid search with exponentially spaced values to narrow down the set of parameters followed by a grid search with linearly spaced values to find a set of well-performing parameters.

$$\text{score} = \begin{cases} (\text{diversity score})^{g} & \text{if previously sent} \\ (\text{diversity score})^{f} & \text{otherwise} \end{cases} \tag{1}$$

f and g are calculated using Equations (2) and (3) respectively:

$$f = \alpha \frac{\text{PCB's age}}{\text{PCB's lifetime}} \tag{2}$$

$$g = \left(\beta \frac{\text{sent PCB's remaining lifetime}}{\text{current PCB's remaining lifetime}}\right)^{\gamma} \tag{3}$$

We evaluate the performance of the path-diversity-based path construction algorithm with respect to how close to optimal it performs regarding its failure resilience and how much overhead it incurs in Section 5.

Optimizing for other Criteria. With additional information transmitted through PCBs or other channels, the path construction can optimize paths for multiple optimality criteria. To optimize for latency for example, the currently disseminated information, i.e., interface numbers and traversed ASes, is insufficient. If additional information, such as border router locations or latency measurements were made available, then path construction could optimize for low latency paths.

Disseminating additional information is a non-trivial task. The measured and possibly aggregated metrics, must be distributed efficiently with low overhead, sensitive information must be filtered by the origin, and the veracity of the information should be verifiable. Moreover, for dynamic metrics (e.g., latency), which depend on the network's state, the traffic volume endpoints send impacts the experienced metric for future packets. In these cases, optimization might lead to unexpected and even counter-intuitive performance changes. The complete design and analysis of additional optimization metrics is therefore left for future work.

5 EVALUATION

In this Section, we evaluate the communication overhead and the quality of paths constructed by the path-diversity-based path construction algorithm in comparison to the baseline path construction algorithm for both core and intra-ISD beaconing. Furthermore, we compare to BGP as the most widespread routing protocol in the Internet. We additionally compare to BGPsec, a secure inter-domain routing protocol based on BGP.

Note that, since there is no convergence phase in SCION, we cannot compare to BGP's convergence time. SCION path-segments are stable as soon as they are disseminated.

5.1 Simulation Setup

Although SCION is deployed in the real world, the current production networks are not yet large enough for Internet-scale inferences. Therefore, we have developed a scalable SCION control plane simulator using the ns-3 network simulation framework [36, 54]. We simulate the core and intra-ISD beaconing on large-scale topologies derived from the CAIDA AS relationship with geolocation data set [17] (*AS-rel-geo*), which contains the relationships between 12000 ASes as well as their interconnection locations. This dataset allows us to infer the relationships and number of links between neighboring ASes, giving us a realistic view of the Internet's core AS-level topology. We use the results collected from simulating beaconing on this topology to approximate an Internet-scale deployment of SCION.

In all SCION experiments, we simulate six hours of beaconing with a beaconing interval of ten minutes and a PCB lifetime of six hours. The *PCB dissemination limit*, which is the maximum number of PCBs per origin AS to disseminate in a beaconing interval, is set to 5 for all experiments. For the baseline path construction algorithm, the limit is applied to each interface and for the path-diversity-based path construction algorithm, the limit is applied to each neighbor AS. The *PCB storage limit*, which is the maximum

number of PCBs per origin AS to store at each beacon server, varies in different experiments.

Core Beaconing. Since establishing an ISD and operating a core AS requires substantial effort, we expect that a global SCION deployment will have a few hundred ISDs with typically fewer than 10 core ASes per ISD. In our core beaconing simulation, we assume 200 ISDs with 10 core ASes each, resulting in 2000 core ASes. We use the subset of the 2000 highest-degree ASes from the topology of 12000 ASes in the CAIDA AS-rel-geo topology, by incrementally pruning the 10000 lowest-degree ASes. We simulate SCION core beaconing using both the baseline path construction algorithm and the path-diversity-based path construction algorithm.

Intra-ISD Beaconing. We simulate intra-ISD beaconing on a large ISD topology to evaluate its scalability in an extreme case. To construct such an ISD, we first select its core ASes by picking the 11 highest-rank American ASes (by customer cone size) from the CAIDA AS Rank [15] data set. Then, we add their direct or indirect customers to the ISD by iterating down the Internet hierarchy starting with the core ASes. The result is a large ISD with 11 core ASes and 7017 non-core ASes, which is one of the largest ISDs we can construct using the CAIDA AS-rel-geo topology. For the intra-ISD beaconing simulation, we only employ the baseline path construction algorithm. The path-diversity-based path construction algorithm, which has lower overhead, would scale even better.

Since SCION ISDs provide routing isolation, the intra-ISD beaconing process of each ISD is independent from other ISDs and from core beaconing, rendering simulations of multiple, connected ISDs superfluous. In a global-scale deployment of SCION, the intra-ISD beaconing process is expected to run on similar-sized topologies.

BGPsec. Since there is no worldwide deployment of BGPsec, we simulate BGPsec on the entire CAIDA AS-rel-geo topology using the SimBGP simulator [46]. We organize the border routers of each AS in a star topology. Therefore, each border router has two interfaces, one connected to the border router of the neighboring AS and the other to the internal BGPsec speaker of its own AS. In our configuration, each BGPsec speaker has a Minimum Route Advertisement Interval (MRAI) timer of 15 seconds and a processing delay of 5 ms for each incoming update message. Within an AS, only the internal BGPsec speaker has LOC_RIB, and border routers just forward traffic between the interfaces.

5.2 Control Plane Overhead

To obtain the ground truth for BGP, we leverage data from the RouteViews [37] update messages dataset collected by RouteViews2 collector in May 2020. This collector peered with 42 monitors of which we consider the 26 that are included in the CAIDA AS-rel-geo topology. To compare the control-plane signaling cost of BGP, BGPsec, and SCION, we need to compare the received control-plane traffic in the same ASes and during the same time period. We compare our six-hour simulations to the BGP measurements collected over a month, by leveraging the periodicity of announcements and multiplying the traffic by the number of periods in a month. Note that we estimate the control plane communication overhead based on real-world measurements for BGP. However, we simulate SCION using a smaller core network of 2000 core ASes. This comparison

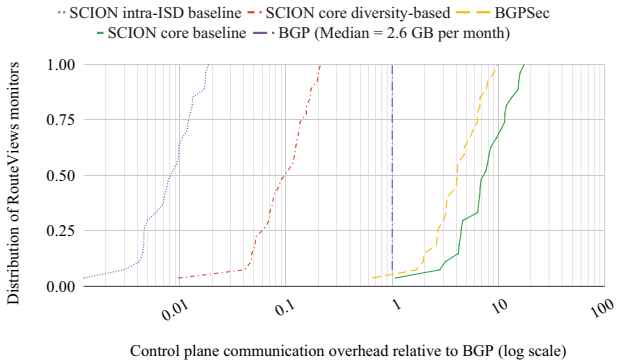


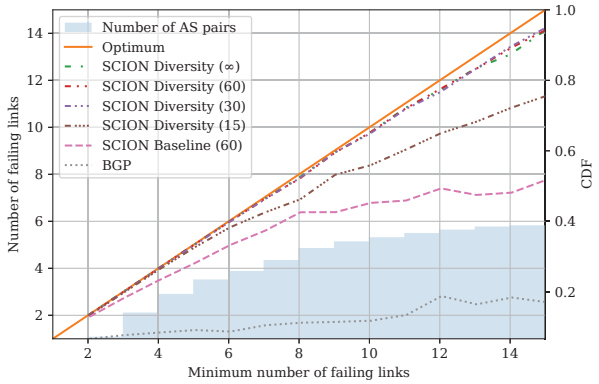
Figure 5: Distribution of control plane overhead of RouteViews monitors for BGPsec, SCION (baseline and diversity-based) core beaconing, and SCION baseline intra-ISD beaconing relative to BGP during one month.

is fair to BGP and BGPsec due to SCION's hierarchical structure, however, it is pessimistic for SCION as the number of core ASes per ISD is expected to be below 10, resulting in a smaller core network than in our experiments.

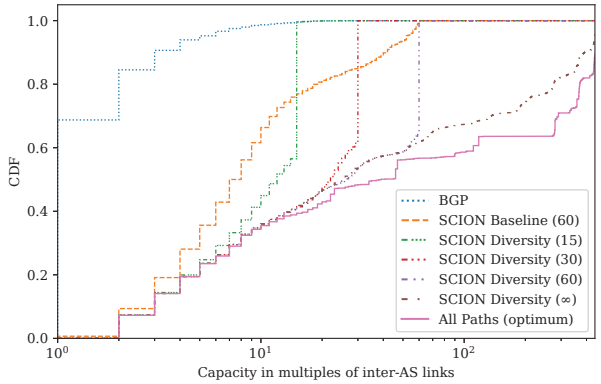
BGP. We measure communication overhead using the RouteViews dataset. We calculate the size of update messages based on the individual field sizes defined in RFC 4271 [41].

SCION. To measure the amount of traffic used for core and intra-ISD beaconing in SCION, we observe the amount of PCB traffic sent on each inter-domain interface.

BGPsec. We measure BGPsec's communication overhead using SimBGP. We first monitor update messages received by the central BGPsec speaker of each AS. Then, we derive BGPsec's overhead per destination prefix per monitor based on the BGPsec update message specifications [29]. Since every AS in the simulation only announces one prefix, we multiply the overhead for each destination prefix by the number of prefixes its AS announces, to arrive at the overhead per origin AS. We use the RouteViews dataset to find the number of prefixes each AS announces [37]. Since the CAIDA AS-rel-geo topology contains only 12000 ASes, the calculated overhead is not comparable with BGP's overhead observed in the real world. Therefore, we extrapolate the overhead resulting from simulations on this topology to the entire Internet topology inferred from CAIDA AS relationships data set (*AS-rel*) [16], which contains the majority of ASes and their relationships but *not* their border routers' geolocations. We assume that for a prefix in AS A outside the AS-rel-geo topology, a router receives the same number of update messages as for a prefix in A's lowest-tier provider *within* the AS-rel-geo topology. Additionally, we assume that the routes originated from A are longer than the routes originated from its lowest-tier provider by their hop difference to their nearest Tier-1 provider. With these assumptions, we derive the overhead for prefixes in ASes outside the CAIDA AS-rel-geo topology. Assuming a re-beaconing period of one day [48], the resulting overhead is multiplied by 30 to find the monthly BGPsec overhead.



(a) Minimum number of failing links disconnecting two ASes.



(b) Maximum capacity in terms of multiples of link capacities.

Figure 6: Path quality of SCION path selection algorithms and BGP. The PCB storage limit is indicated in the parentheses.

Results. Figure 5 shows the overhead, relative to BGP, of BGPsec, the path-diversity-based path construction algorithm for core beaconing, and the baseline path construction algorithm for both core beaconing and intra-ISD beaconing during one month with PCB storage and dissemination limits of 60 and 5, respectively. In this experiment, we assume the use of ECDSA384 signatures in both SCION and BGPsec. The overhead of BGPsec is one order of magnitude higher than BGP due to larger update messages and lack of aggregation in BGPsec. The overhead of core beaconing (baseline) is slightly higher than BGPsec. The advantage of the path-diversity-based path construction algorithm is evident as the overhead of core beaconing (diversity-based) is one order of magnitude lower than BGP, which indicates that core beaconing scales to global deployment. Compared to the currently used baseline path construction algorithm, the path-diversity-based path construction algorithm reduces the control plane overhead by more than two orders of magnitude. Finally, as expected, the overhead of SCION's intra-ISD beaconing is very low, i.e., two orders of magnitude lower than BGP, since PCBs are only sent uni-directionally.

5.3 Path Quality

We evaluate the quality of paths provided by SCION based on the criteria of inter-AS link failure resilience and available inter-AS link bandwidth, assuming no intra-AS link failures and bandwidth limitations. Failure resilience is defined as the minimum number of links whose failures disconnect two ASes. We compare to BGP and to the optimally achievable path quality. We can thus evaluate how close to optimal the path dissemination is, and showcase the increased resilience of the path-diversity-based path construction algorithm. We consider the best possible case for BGP, by choosing the best path present in RouteViews and assuming full BGP multi-path support between every AS pair for bandwidth aggregation and fast failover.

Link Failure Resilience. Figure 6a shows the link failure resilience. We evaluate the performance with different PCB storage limits. Although BGP has limited link failure resilience due to the use of BGP multi-path and the frequency of parallel links in the core topology, it is outperformed by the baseline path construction algorithm. For at most 15 failing links, which covers almost 40 % of the cases, the baseline path construction algorithm on average more than doubles the link failure resilience compared to BGP.

Maximum Bandwidth. We measure the total available capacity between each AS pair in terms of how many inter-AS links can be saturated, assuming that all inter-AS links have uniform capacity (since we cannot precisely infer inter-domain link capacity from our dataset). It is important to note, that the objective function of the path-diversity-based path construction algorithm is to maximize the number of links which can fail before connectivity is lost, which is *equivalent* to maximizing the number of parallel links on which traffic can be sent without experiencing congestion. Figure 6b confirms that the maximum bandwidth of BGP using multi-path is the lowest and that the path-diversity-based path construction algorithm outperforms the baseline path construction algorithm due to prioritizing PCBs with new links. Furthermore, we can see that the capacity of the path-diversity-based path construction algorithm is close to the optimal capacity until the PCB storage limit is almost reached. In particular, the path-diversity-based path construction algorithm achieves 99%, 97%, 95%, and 82% of the optimal capacity for the PCB storage limits (i.e., 15, 30, 60) and unlimited storage, respectively. This shows that the algorithm effectively finds paths with a diverse set of links and performs close to optimal for small PCB sizes.

5.4 Measurements on SCIONLab

We evaluate path quality and overhead of the SCIONLab research testbed [28] control plane to cross-validate the simulation results. We analyze the disseminated paths, and the number and average size of PCBs sent at each core AS. The evaluation shows that the baseline path construction algorithm provides link failure resilience in over 90% of the cases and provides optimal link failure resilience in over 30% of the cases. We simulate the SCIONLab topology with different PCB storage limits and show that increasing the PCB storage limit over 15 provides negligible benefits in terms of resilience. The beaconing overhead in SCIONLab is less than

4 KB/s per interface for almost 80% of all core interfaces, which is negligible compared to the capacity of a typical inter-domain link. However, it will be beneficial to apply the path-diversity-based path construction algorithm for core beaconing as the SCIONLab network expands. Appendix B provides a more detailed evaluation of the SCIONLab testbed.

6 RELATED WORK

In this Section, we review work in the areas of deployment, design of new Internet architectures, and multi-path routing.

6.1 Deployment

VINI [9] is a virtual network infrastructure that enables researchers to test new networking protocols in realistic but controlled settings. GENI [32] is a distributed virtual laboratory for the development, deployment, and validation of transformative, at-scale concepts in network science, services, and security, deployed at around 50 US sites. FABRIC [6] is a programmable networking infrastructure facilitating experiments with novel network designs and applications. VINI, GENI, and FABRIC are testbeds that allow the evaluation of Internet architectures on a large-scale network, but have so far not been used to build the SCION production network. The vast effort on deployment concepts and test beds for next-generation routing infrastructures demonstrates the challenge of deploying a novel Internet architecture. The design choices of SCION have made it possible to overcome this challenge, without creating an overlay on today's Internet.

Trotsky [31] proposes a backward-compatible architectural framework to deploy new Internet architectures. The Internet is described as a collection of layered overlays, with the only exception of intra- and inter-domain communication. There is a single inter-domain communication protocol, the *narrow waist* of the Internet, hindering innovation at this layer so far. Trotsky describes how to design new inter-domain protocols, by constructing the inter-domain layer as an overlay on the intra-domain communication.

The abstractions used in Trotsky and SCION are similar, since both propose novel inter-AS control planes, while treating intra-AS connectivity as logical pipes. The main goals of Trotsky, incremental deployment and extensibility, are reflected in SCION's deployment model and flexible path dissemination approach, enabling different path selection algorithms per AS. Despite the conceptual similarities, the goals of these efforts are quite different, and SCION could benefit from additional deployment in Trotsky's infrastructure.

6.2 New Internet Architecture Proposals

Resilient Overlay Network (RON) [3], is a proposal to improve both the resilience and the performance (with respect to application requirements) of the Internet. RON is an application-level (BGP-)overlay network composed of RON nodes, that monitor the performance of different paths and quickly reroute traffic in case of a link outage that would otherwise require re-convergence of BGP. However, due to the fact that RON is an overlay network, it cannot achieve the same guarantees as a natively deployed architecture.

In Plutarch [12], the global network is divided into contexts, and interstitial functions are used to communicate between them. This allows interconnecting heterogeneous networks, instead of imposing the same L3 protocols everywhere, which might not always be feasible (e.g., for sensors) while the interstitial functions translate between the different contexts, still providing global interconnectivity. The authors argue that a model with explicit interaction at context boundaries is more accurate and extensible. SCION follows the same approach by clearly separating communication between ISDs (core-path segments) and within ISDs (down- and up-path segments). SCION also distinguishes between inter-AS communication (through AS + interface level granularity path segments) and intra-AS communication (independent intra-AS namespaces).

HLP [49] is a proposed next-generation routing architecture which uses a hybrid approach between link-state and path-vector algorithms in order to improve the scalability, convergence and isolation properties compared to BGP. SCION follows their approach of hierarchically partitioning the network and opting for AS-based routing instead of prefix-based routing. HLP combines a link-state algorithm inside with a path-vector algorithm between hierarchies and obscures AS-path information present in BGP through a generic cost metric to reduce overhead. In SCION, the beaconing mechanism constituting the control plane allows for stateless routers, and the beaconing scales fundamentally better than BGP as shown in this work. Even though a simplified version of the HLP protocol can be implemented in BGP-speaking routers by changing the current operational practise of BGP, allowing for incremental deployment, not all BGP policies are supported by HLP. The consequences of such a mixed deployment remain unclear.

XIA [1, 21], which paved the way for the initial work on SCION, unifies different networking paradigms, such as content- and service-centric networking, using a generic principal-centric networking approach. XIA intends to be extensible and evolvable to support new types of communication and facilitate deployability. This is achieved by enabling applications or protocols to start using new principal types before the network develops inherent support for them. Instead, network entities unfamiliar with the new principal use a fallback mechanism, and still provide global connectivity.

The Framework for Internet Innovation (FII) [26] proposes a clean-slate redesign of the Internet to remove deployment barriers for innovations. It defines three primitives: an inter-domain routing architecture, a network API, and an interface for hosts to protect themselves against DoS attacks. The latter requires a trusted third party, with the ability to *shut up* a host which is attacking another host or network entity. While this party does not need to be globally valid, and there can be multiple parties, any misbehaving trusted third party would have a severe impact on the system. The authors propose to use pathlet routing as the inter-domain routing architecture.

The NEBULA project [4] shared a similar vision to SCION in terms of using diverse paths to meet the high reliability and privacy requirements. However, their approach differs from SCION, as it depends on ultra-reliable core routers interconnecting data centers to support cloud-based applications. These new router components complicate incremental deployment, and limit applicability to the cloud context.

ChoiceNet [43, 58] introduces an economy plane that allows the establishment of dynamic business relationships to create a competitive marketplace for innovative solutions. The authors briefly

touch on the scalability question of ChoiceNet, stating that its scalability would depend on the spectrum of choices provided, how the choices were made, the frequency with which choices change, and the threat model to be protected against. However, it is unclear if the system could scale to a network of the size of the Internet. ChoiceNet requires the network to generate alternatives for users to choose from and provides monetary rewards for alternatives that address the user's needs. SCION could provide an instantiation in the form of paths with different properties presented to users.

Route Bazaar [11] introduces a contractual system based on a public ledger, where ASes and customers agree on QoS-aware routes in the form of BGP-overlay pathlets. Route Bazaar is orthogonal to SCION, as it proposes a new contractual system, which could be used to offer SCION paths (instead of pathlets) to customers, circumventing the disadvantages of overlay connections.

6.3 Multi-path Routing

Although numerous research teams worked on multi-path routing, only few approaches were developed beyond a proof of concept, and even fewer were deployed. We compare the most relevant approaches to SCION and highlight the differences. We refer to Singh et al. [57] for an in-depth survey of approaches that were not further developed or deployed.

BGP-based Approaches. BGP Add-Path is a deployed extension to BGP [42], which allows announcing additional paths for a certain prefix without implicitly revoking the existing path. The two main drawbacks of BGP Add-Path are increased border router memory requirements for storing additional paths, and lack of path control for endpoints. Other proposals using BGP include BGP-XM [10], which explores existing redundant routing paths provided by BGP, and Path Splicing [33] and STAMP [30], which provide multiple paths by running k ($k = 2$ for STAMP and configurable for Path Splicing) parallel BGP sessions to explore multiple routing paths. The main drawback of these approaches is the overhead of running multiple BGP sessions, which typically requires network operators to purchase additional hardware. A large number of BGP-based inter-domain multi-path routing approaches have similar limitations: DIMR [62], AMIR [39], YAMR [18], BGP-XM [10], MIFO [63], D-BGP and B-BGP [56], and R-BGP [27].

Source-Based Routing. Source-based routing protocols allow a sender to (partially) control the packets' forwarding paths. BANANAS [25] encodes partial paths as PathIDs, and a packet specifying a PathID is sent along the specified path. BANANAS supports incremental deployment, by enriching the link-state tables of upgraded routers with the knowledge of which other routers are multi-path-capable. This way, the set of available paths can be computed locally by enriched routers. In order to avoid the situation where a path computed by one router does not exist in another, they employ a distributed path validation algorithm. The additional information in upgraded routers increases the size of routing tables, since each PathID forwarding rule adds an additional entry, and the validation algorithm increases computational complexity. Platypus [40] enhances source routing with per-packet capabilities to enable policy compliance among operators. However, path exploration and path selection are not defined, and thus it is difficult to reason about its scalability. Wide-Area Relay Addressing Protocol (WRAP) [5] is based on loose source routing, i.e., WRAP packets specify a list of IP addresses of AS edge routers that packets should traverse. Since each AS edge router maintains at least two AS paths to each other AS, WRAP must maintain multiple routing paths per destination prefix, hampering scalability. New Inter-Domain Routing Architecture (NIRA) [60] constructs end-to-end paths from up- and down-segments connected at a core network similar to SCION, but only supports a single core network and no isolation domains which are essential for the scalability of SCION. Routing Deflection [61] allows endpoints to deflect their traffic at certain BGP routers to choose different paths. While this approach can be incrementally deployed with minimal changes to BGP, it only provides coarse-grained path control. Multipath Interdomain Routing (MIRO) [59], is a mix between source-based and tunneling-based routing. ASes can negotiate the advertisement of alternative paths pairwise, for the purpose of avoiding a specific AS (e.g., for security reasons). This keeps the increased state small and MIRO could in principle be incrementally deployed, as long as the most densely-interconnected ASes adopt it first. Pathlet Routing [19], allows (partial) paths (Pathlets) to be constructed from a set of routers. These Pathlets are then disseminated in a similar way as BGP disseminates routes to prefixes today. Incremental deployment is hindered by routers needing to understand the pathlet vocabulary. Additionally, policies in pathlet routing are no longer destination based, making it non-interoperable with BGP policies.

7 CONCLUSION

SCION provides rich inter-domain multi-path, a core component for a path-aware Internet which benefits from the Internet's extensive path diversity. By improving the scalability and utility of SCION's path exploration, we further enhance SCION's viability. Since the first adopters desired strong security properties, designing SCION as an overlay network was not an option for the SCION production network. The insights gained while deploying a BGP-free infrastructure will hopefully prove beneficial for other overlay-free technology deployments. As SCION's production network grows, we anticipate that the multi-path routing system will offer a rich variety of path choices, which will in turn enable opportunities for application-based path optimizations.

8 ACKNOWLEDGMENTS

We gratefully acknowledge support from ETH Zürich, the Zürich Information Security and Privacy Center (ZISC), and the European Union's Horizon 2020 research and innovation programme under grant agreements No 825310 and 825322. This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korea government (MSIT) under grant agreement No 2019-0-01697 (Development of Automated Vulnerability Discovery Technologies for Blockchain Platform Security).

REFERENCES

[1] Ashok Anand, Fahad Dogar, Dongsu Han, Boyan Li, Hyeontaek Lim, Michel Machado, Wenfei Wu, Aditya Akella, David G. Andersen, John W. Byers, Srinivasan Seshan, and Peter Steenkiste. 2011. XIA: An Architecture for an Evolvable and Trustworthy Internet. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks* (Cambridge, Massachusetts) *(HotNets '11)*. Association for Computing Machinery, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/2070562.2070564

[2] Anapaya. 2021. Anapaya - Next-Generation Internet. https://www.anapaya.net

[3] David Andersen, Hari Balakrishnan, Frans Kaashoek, and Robert Morris. 2001. Resilient Overlay Networks. In *Proceedings of the ACM Symposium on Operating Systems Principles* (Banff, Alberta, Canada) *(SOSP '01)*. Association for Computing Machinery, New York, NY, USA, 131–145. https://doi.org/10.1145/502034.502048

[4] Tom Anderson, Ken Birman, Robert Broberg, Matthew Caesar, Douglas Comer, Chase Cotton, Michael J. Freedman, Andreas Haeberlen, Zachary G. Ives, Arvind Krishnamurthy, William Lehr, Boon Thau Loo, David Mazières, Antonio Nicolosi, Jonathan M. Smith, Ion Stoica, Robbert van Renesse, Michael Walfish, Hakim Weatherspoon, and Christopher S. Yoo. 2013. The NEBULA Future Internet Architecture. In *The Future Internet*, Alex Galis and Anastasius Gavras (Eds.). Springer Berlin Heidelberg, 16–26. https://doi.org/10.1007/978-3-642-38082-2_2

[5] Katerina Argyraki and David R. Cheriton. 2004. Loose Source Routing as a Mechanism for Traffic Policies. In *Proceedings of the ACM SIGCOMM Workshop on Future Directions in Network Architecture* (Portland, Oregon, USA) *(FDNA '04)*. Association for Computing Machinery, New York, NY, USA, 57–64. https://doi.org/10.1145/1016707.1016718

[6] Ilya Baldin, Anita Nikolich, James Griffioen, Indermohan Inder S Monga, Kuang-Ching Wang, Tom Lehman, and Paul Ruth. 2019. FABRIC: A National-Scale Programmable Experimental Network Infrastructure. *IEEE Internet Computing* 23, 6 (2019), 38–47. https://doi.org/10.1109/MIC.2019.2958545

[7] Swiss National Bank. 2021. SNB and SIX launch the communication network Secure Swiss Finance Network. https://perma.cc/PU5L-ALPM

[8] David Barrera, Laurent Chuat, Adrian Perrig, Raphael M. Reischuk, and Pawel Szalachowski. 2017. The SCION Internet Architecture. *Commun. ACM* 60, 6 (June 2017), 56–65. https://doi.org/10.1145/3085591

[9] Andy Bavier, Nick Feamster, Mark Huang, Larry Peterson, and Jennifer Rexford. 2006. In VINI Veritas: Realistic and Controlled Network Experimentation. *ACM SIGCOMM Computer Communication Review* 36, 4 (Aug. 2006), 3–14. https://doi.org/10.1145/1151659.1159916

[10] Jose M. Camacho, Alberto García-Martínez, Marcelo Bagnulo, and Francisco Valera. 2013. BGP-XM: BGP eXtended Multipath for transit Autonomous Systems. *Computer Networks* 57, 4 (2013), 954 – 975. https://doi.org/10.1016/j.comnet.2012.11.011

[11] Ignacio Castro, Aurojit Panda, Barath Raghavan, Scott Shenker, and Sergey Gorinsky. 2015. Route Bazaar: Automatic Interdomain Contract Negotiation. In *Proceedings of the USENIX Conference on Hot Topics in Operating Systems* (Switzerland) *(HOTOS'15)*. USENIX Association, USA, 9.

[12] Jon Crowcroft, Steven Hand, Richard Mortier, Timothy Roscoe, and Andrew Warfield. 2003. Plutarch: An Argument for Network Pluralism. *ACM SIGCOMM Computer Communication Review* 33, 4 (Aug. 2003), 258–266. https://doi.org/10.1145/972426.944763

[13] Joeri de Ruiter and Caspar Schutijser. 2021. Future internet at terabit speeds: SCION in P4. https://perma.cc/SJ3G-YDQ2

[14] Joeri de Ruiter and Caspar Schutijser. 2021. Next-generation internet at terabit speed: SCION in P4. In *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies* (Germany) *(CoNEXT '21)*. Association for Computing Machinery, New York, NY, USA, 7 pages. https://doi.org/10.1145/3485983.3494839

[15] Center for Applied Internet Data Analysis. 2021. *CAIDA AS-Rank*. https://asrank.caida.org/

[16] Center for Applied Internet Data Analysis. 2021. *CAIDA AS-relationships*. https://www.caida.org/data/as-relationships/

[17] Center for Applied Internet Data Analysis. 2021. *CAIDA Geolocation Data*. https://www.caida.org/data/as-relationships-geo/

[18] Igor Ganichev, Bin Dai, P. Brighten Godfrey, and Scott Shenker. 2010. YAMR: Yet Another Multipath Routing Protocol. *ACM SIGCOMM Computer Communication Review* 40, 5 (Oct. 2010), 13–19. https://doi.org/10.1145/1880153.1880156

[19] P. Brighten Godfrey, Igor Ganichev, Scott Shenker, and Ion Stoica. 2009. Pathlet Routing. In *Proceedings of the ACM SIGCOMM Conference on Data Communication* (Barcelona, Spain) *(SIGCOMM '09)*. Association for Computing Machinery, New York, NY, USA, 111–122. https://doi.org/10.1145/1592568.1592583

[20] Nikola Gvozdiev, Stefano Vissicchio, Brad Karp, and Mark Handley. 2018. On Low-Latency-Capable Topologies, and Their Impact on the Design of Intra-Domain Routing. In *Proceedings of the ACM SIGCOMM Conference on Data Communication* (Budapest, Hungary) *(SIGCOMM '18)*. Association for Computing Machinery, New York, NY, USA, 88–102. https://doi.org/10.1145/3230543.3230575

[21] Dongsu Han, Ashok Anand, Fahad Dogar, Boyan Li, Hyeontaek Lim, Michel Machado, Arvind Mukundan, Wenfei Wu, Aditya Akella, David G. Andersen, John W. Byers, Srinivasan Seshan, and Peter Steenkiste. 2012. XIA: Efficient Support for Evolvable Internetworking. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation* (San Jose, CA) *(NSDI'12)*. USENIX Association, USA, 23. https://dl.acm.org/doi/10.5555/2228298.2228330

[22] IRTF. 2021. Path Aware Networking Research Group (PANRG). https://datatracker.ietf.org/rg/panrg/about/

[23] Tony John, Piet De Vaere, Caspar Schutijser, Adrian Perrig, and David Hausheer. 2021. Linc: Low-Cost Inter-Domain Connectivity for Industrial Systems. In *Proceedings of the ACM SIGCOMM Poster and Demo Sessions (SIGCOMM '21)*. Association for Computing Machinery, New York, NY, USA, 68–70. https://doi.org/10.1145/3472716.3472850

[24] Tony John and David Hausheer. 2021. S3MP: A SCION based Secure Smart Metering Platform. In *Proceedings of the 17th IFIP/IEEE International Symposium on Integrated Network Management* (Bordeaux, France) *(IM '21)*. IEEE, 944–949. https://ieeexplore.ieee.org/document/9463922

[25] H. Tahilramani Kaur, S. Kalyanaraman, A. Weiss, S. Kanwar, and A. Gandhi. 2003. BANANAS: An Evolutionary Framework for Explicit and Multipath Routing in the Internet. In *Proceedings of the ACM SIGCOMM Workshop on Future Directions in Network Architecture* (Karlsruhe, Germany) *(FDNA '03)*. Association for Computing Machinery, New York, NY, USA, 277–288. https://doi.org/10.1145/944759.944766

[26] Teemu Koponen, Scott Shenker, Hari Balakrishnan, Nick Feamster, Igor Ganichev, Ali Ghodsi, P. Brighten Godfrey, Nick McKeown, Guru Parulkar, Barath Raghavan, Jennifer Rexford, Somaya Arianfar, and Dmitriy Kuptsov. 2011. Architecting for Innovation. *ACM SIGCOMM Computer Communication Review* 41, 3 (July 2011), 24–36. https://doi.org/10.1145/2002250.2002256

[27] Nate Kushman, Srikanth Kandula, Dina Katabi, and Bruce M. Maggs. 2007. R-BGP: Staying Connected In a Connected World. In *Proceedings of the 4th USENIX Conference on Networked Systems Design and Implementation* (Cambridge, MA) *(NSDI '07)*. USENIX Association, USA, 25. https://dl.acm.org/doi/10.5555/1973430.1973455

[28] Jonghoon Kwon, Juan A. García-Pardo, Markus Legner, François Wirz, Matthias Frei, David Hausheer, and Adrian Perrig. 2020. SCIONLab: A Next-Generation Internet Testbed. In *Proceedings of the 28th IEEE International Conference on Network Protocols (ICNP '20)*. https://doi.org/10.1109/ICNP49622.2020.9259355

[29] M. Lepinski and K. Sriram. 2017. *BGPsec Protocol Specification*. RFC 8205. IETF. https://tools.ietf.org/rfc/rfc8205.txt

[30] Yong Liao, Lixin Gao, Roch Guerin, and Zhi-Li Zhang. 2008. Reliable Interdomain Routing through Multiple Complementary Routing Processes. In *Proceedings of the 4th International Conference on emerging Networking EXperiments and Technologies* (Madrid, Spain) *(CoNEXT '08)*. Association for Computing Machinery, New York, NY, USA, Article 68, 6 pages. https://doi.org/10.1145/1544012.1544080

[31] James McCauley, Yotam Harchol, Aurojit Panda, Barath Raghavan, and Scott Shenker. 2019. Enabling a Permanent Revolution in Internet Architecture. In *Proceedings of the ACM SIGCOMM Conference on Data Communication* (Beijing, China) *(SIGCOMM '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3341302.3342075

[32] Rick McGeer, Mark Berman, Chip Elliott, and Robert Ricci (Eds.). 2016. *The GENI Book*. Springer International Publishing. https://doi.org/10.1007/978-3-319-33769-2

[33] Murtaza Motiwala, Megan Elmore, Nick Feamster, and Santosh Vempala. 2008. Path Splicing. In *Proceedings of the ACM SIGCOMM Conference on Data Communication* (Seattle, WA, USA) *(SIGCOMM '08)*. Association for Computing Machinery, New York, NY, USA, 27–38. https://doi.org/10.1145/1402958.1402963

[34] Johannes Naab, Patrick Sattler, Jonas Jelten, Oliver Gasser, and Georg Carle. 2019. Prefix Top Lists: Gaining Insights with Prefixes from Domain-Based Top Lists on DNS Deployment. In *Proceedings of the Internet Measurement Conference* (Amsterdam, Netherlands) *(IMC '19)*. Association for Computing Machinery, New York, NY, USA, 351–357. https://doi.org/10.1145/3355369.3355598

[35] ETH Zurich Network Security Group. 2021. SCI-ED Project. https://scied.scion-architecture.net/

[36] nsnam. 2021. ns-3. https://www.nsnam.org/

[37] University of Oregon. 2021. University of Oregon Route Views Archive Project. http://routeviews.org/

[38] Adrian Perrig, Pawel Szalachowski, Raphael M. Reischuk, and Laurent Chuat. 2017. *SCION: A Secure Internet Architecture*. Springer International Publishing. https://doi.org/10.1007/978-3-319-67080-5

[39] Donghong Qin, Jiahai Yang, Zhuolin Liu, Jessie Wang, Bin Zhang, and Wei Zhang. 2012. AMIR: Another Multipath Interdomain Routing. In *Proceedings of the IEEE 26th International Conference on Advanced Information Networking and Applications (AINA '12)*. 581–588. https://doi.org/10.1109/AINA.2012.83

[40] Barath Raghavan and Alex C. Snoeren. 2004. A System for Authenticated Policy-Compliant Routing. *ACM SIGCOMM Computer Communication Review* 34, 4 (Aug. 2004), 167–178. https://doi.org/10.1145/1030194.1015487

[41] Y. Rekhter, T. Li, and S. Hares. 2006. *A Border Gateway Protocol 4 (BGP-4)*. RFC 4271. IETF. https://tools.ietf.org/rfc/rfc4271.txt

[42] Alvaro Retana. 2015. *Advertisement of Multiple Paths in BGP: Implementation Report*. Internet-Draft draft-ietf-idr-add-paths-implementation-00.

IETF Secretariat. http://www.ietf.org/internet-drafts/draft-ietf-idr-add-paths-implementation-00.txt
[43] George N Rouskas, Ilia Baldine, Ken Calvert, Rudra Dutta, Jim Griffioen, Anna Nagurney, and Tilman Wolf. 2013. ChoiceNet: Network innovation through choice. In *Proceedings of the 17th International Conference on Optical Networking Design and Modeling (ONDM '13)*. 1–6. https://ieeexplore.ieee.org/document/6524925
[44] Simon Scherrer, Markus Legner, Tobias Schmidt, and Adrian Perrig. 2021. Footprints on the path: how routing data could reduce the internet's carbon toll. https://perma.cc/PRW7-675A
[45] Lars-Christian Schulz and David Hausheer. 2021. Towards SCION-enabled IXPs: The SCION Peering Coordinator. In *Proceedings of the Conference on Networked Systems (NetSys '21)*. https://doi.org/10.14279/tuj.eceasst.80.1159
[46] João Luís Sobrinho, Franck Le, and Laurent Vanbever. 2014. SimBGP. https://github.com/network-aggregation/dragon_simulator
[47] Supraja Sridhara, François Wirz, Joeri de Ruiter, Caspar Schutijser, Markus Legner, and Adrian Perrig. 2021. Global Distributed Secure Mapping of Network Addresses. In *Proceedings of the ACM SIGCOMM Workshop on Technologies, Applications, and Uses of a Responsible Internet (TAURIN '21)*.
[48] K. Sriram. 2018. *BGPsec Design Choices and Summary of Supporting Discussions*. RFC 8374. IETF. https://tools.ietf.org/rfc/rfc8374.txt
[49] Lakshminarayanan Subramanian, Matthew Caesar, Cheng Tien Ee, Mark Handley, Morley Mao, Scott Shenker, and Ion Stoica. 2005. HLP: A next Generation Inter-Domain Routing Protocol. *ACM SIGCOMM Computer Communication Review* 35, 4 (Aug. 2005), 13–24. https://doi.org/10.1145/1090191.1080095
[50] Sunrise. 2021. The Innovative solution for a secure Internet. https://www.sunrise.ch/en/business/products-and-solutions/connectivity/scion.html
[51] Swisscom. 2021. The secure, high-speed Internet under your control. https://www.swisscom.ch/scion
[52] SWITCH. 2021. A quantum leap in cybersecurity. https://www.switch.ch/stories/a-quantum-leap-in-cyber-security/
[53] Anapaya Systems. 2021. Anapaya CONNECT: The SCION-transit service. https://www.anapaya.net/anapaya-connect-for-service-providers
[54] Seyedali Tabaeiaghdaei and Christelle Gloor. 2021. Beaconing Simulator.
[55] D. Thaler. 2017. *Planning for Protocol Adoption and Subsequent Transitions*. RFC 8170. IETF. https://tools.ietf.org/rfc/rfc8170.txt
[56] Feng Wang and Lixin Gao. 2009. Path Diversity Aware Interdomain Routing. In *Proceedings of the 28th IEEE Conference on Computer Communications (INFOCOM '09)*. 307 – 315. https://doi.org/10.1109/INFCOM.2009.5061934
[57] Robert Wójcik, Jerzy Domundefinedał, Zbigniew Duliński, Grzegorz Rzym, Andrzej Kamisiński, Piotr Gawłowicz, Piotr Jurkiewicz, Jacek Rzundefinedsa, Rafał Stankiewicz, and Krzysztof Wajda. 2016. A Survey on Methods to Provide Interdomain Multipath Transmissions. *Computer Networks* 108, C (Oct. 2016), 233–259. https://doi.org/10.1016/j.comnet.2016.08.028
[58] Tilman Wolf, James Griffioen, Kenneth L. Calvert, Rudra Dutta, George N. Rouskas, Ilya Baldin, and Anna Nagurney. 2014. ChoiceNet: Toward an Economy Plane for the Internet. *ACM SIGCOMM Computer Communication Review* 44, 3 (July 2014), 58–65. https://doi.org/10.1145/2656877.2656886
[59] Wen Xu and Jennifer Rexford. 2006. MIRO: Multi-Path Interdomain Routing. *ACM SIGCOMM Computer Communication Review* 36, 4 (Aug. 2006), 171–182. https://doi.org/10.1145/1151659.1159934
[60] Xiaowei Yang, David Clark, and Arthur W. Berger. 2007. NIRA: A New Inter-Domain Routing Architecture. *IEEE/ACM Trans. Netw.* 15, 4 (Aug. 2007), 775–788. https://doi.org/10.1109/TNET.2007.893888
[61] Xiaowei Yang and David Wetherall. 2006. Source Selectable Path Diversity via Routing Deflections. *ACM SIGCOMM Computer Communication Review* 36, 4 (Aug. 2006), 159–170. https://doi.org/10.1145/1151659.1159933
[62] Xia Yin, Dan Wu, Zhiliang Wang, Xingang Shi, and Jianping Wu. 2015. DIMR. *Computer Networks* 91, C (Nov. 2015), 356–375. https://doi.org/10.1016/j.comnet.2015.08.028
[63] Ming Zhu, Dan Li, Ying Liu, Dan Pei, KK Ramakrishnan, Lili Liu, and Jianping Wu. 2015. MIFO: Multi-path Interdomain Forwarding. In *Proceedings of the 44th International Conference on Parallel Processing (ICPP '15)*. 180–189. https://doi.org/10.1109/ICPP.2015.27

A PATH-DIVERSITY-BASED PATH CONSTRUCTION ALGORITHM PSEUDO CODE

Algorithm 1 shows the pseudo code of the path-diversity-based path construction algorithm mentioned in Section 4.2. The path-diversity-based path construction algorithm is triggered periodically by the beacon server to select paths and disseminate them only if the path quality is above a certain threshold. The algorithm selects the best paths iteratively by selecting at most one path from each origin AS to each neighbor AS. To find the best path, the algorithm first creates paths by appending all the egress interfaces connecting to the neighbor AS to every path from the origin AS stored in its own beacon database. Then, it calculates the scores of all these new paths based on their age, lifetime, and link disjointness with regard to previously disseminated paths for the same neighbor and origin AS pair. At the end of each iteration, the best possible path is selected, provided that its score is above the threshold. Then, it considers the selected path as a sent path and updates the algorithm's data structures. The algorithm iterates until either the number of selected paths meets a maximum threshold, or the best path's score in the last iteration is less than the score threshold.

Algorithm 1: Path-diversity-based path selection

```
Result: Paths from origin o to neighbor n
selected_paths ← [];
sent_PCBs_cnt ← 0;
while sent_PCBs_cnt < PCB_dissemination_limit do
    max_score ← 0;
    max_score_path ← null;
    for p ∈ received paths with origin o do
        for iface ∈ interfaces to neighbor n do
            p_new ← [p, iface];
            diversity_score ← calculate_diversity_score(p_new, o,
              n);
            if p_new ∈ Sent_PCBs_List[iface] then
                score ← diversity_score^g;
            else
                score ← diversity_score^f;
            if score > score_threshold and
            score > max_score then
                max_score ← score;
                max_score_path ← p_new;
    if max_score_path == null then
        break;
    else
        selected_paths.append(max_score_path) ;
        sent_PCBs[egress_iface].append(max_score_path);
        for link ∈ max_score_path do
            link_history_table[o][n][link] ←
              link_history_table[o][n][link] + 1
        sent_PCBs_cnt ← sent_PCBs_cnt + 1
```

B DETAILED SCIONLAB TESTBED EVALUATION

We evaluate the SCIONLab testbed by fetching a snapshot of all the paths stored in the path server for all 21 core ASes and analyze their failure resilience and maximum capacity and compare to the optimal failure resilience and capacity. Figures 7 and 8 show, that the baseline path construction algorithm of SCION provides multi-path opportunities and increased resilience to link failures. As expected, the behavior of SCION Baseline with a PCB storage limit of 5 closely resembles the data gathered from SCIONLab, since the baseline path construction algorithm is modeled after the current path selection algorithm. In general, there is limited benefit for

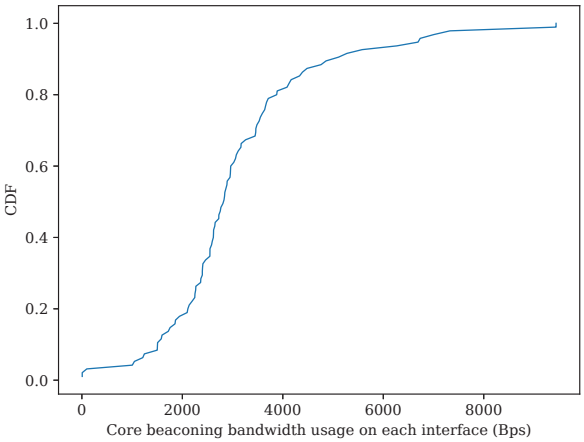


Figure 9: Overhead of core beaconing per interface in SCIONLab.

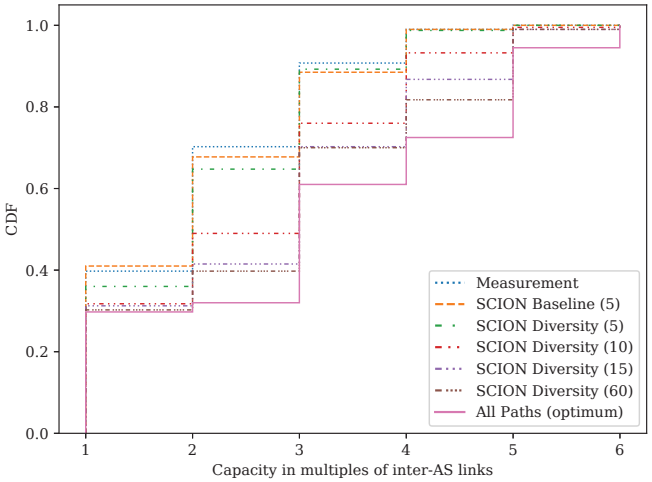


Figure 8: Maximum capacity in terms of multiples of link capacities.

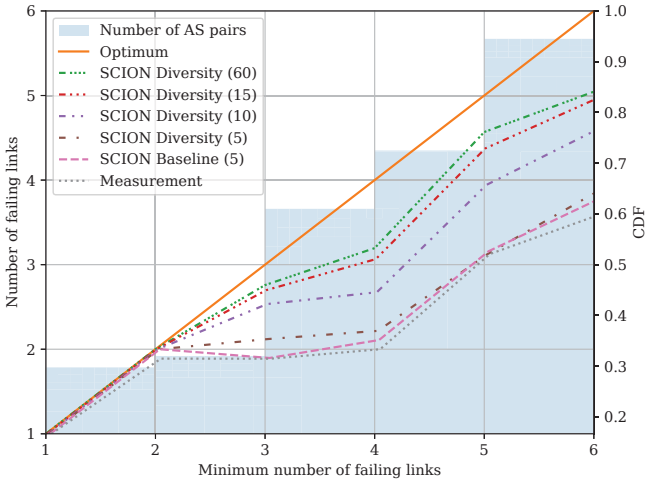


Figure 7: Minimum number of failing links disconnecting two ASes.

the path-diversity-based path construction algorithm in SCIONLab, since for our non-densely-interconnected core ASes, where on average, a core AS has 2 neighbors, choosing the shortest paths often yields paths without overlapping links. The path-diversity-based path construction algorithm with a PCB storage limit of 5, 10, 15, and 60 achieves a better link failure resilience than the current SCIONLab path selection algorithm in 17%, 42%, 52%, and 55% of cases, respectively. This indicates that for the current SCIONLab topology, increasing the PCB storage limit over 15 provides negligible benefits. The overhead of beaconing in SCIONLab is shown in Figure 9 and we can observe that for the majority of interfaces, it is below 4 KB/s.