Pathlet Routing

P. Brighten Godfrey[†], Igor Ganichev[‡], Scott Shenker[‡§], and Ion Stoica[‡*]
[†]University of Illinois at Urbana-Champaign [‡]UC Berkeley [§]ICSI
pbg@illinois.edu, {igor,shenker,istoica}@cs.berkeley.edu

ABSTRACT

We present a new routing protocol, pathlet routing, in which networks advertise fragments of paths, called pathlets, that sources concatenate into end-to-end source routes. Intuitively, the pathlet is a highly flexible building block, capturing policy constraints as well as enabling an exponentially large number of path choices. In particular, we show that pathlet routing can emulate the policies of BGP, source routing, and several recent multipath proposals.

This flexibility lets us address two major challenges for Internet routing: scalability and source-controlled routing. When a router's routing policy has only "local" constraints, it can be represented using a small number of pathlets, leading to very small forwarding tables and many choices of routes for senders. Crucially, pathlet routing does not impose a global requirement on what style of policy is used, but rather allows multiple styles to coexist. The protocol thus supports complex routing policies while enabling and incentivizing the adoption of policies that yield small forwarding plane state and a high degree of path choice.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Packet-switching networks; C.2.2 [**Network Protocols**]: Routing Protocols; C.2.6 [**Internetworking**]: Routers

General Terms

Design, Experimentation, Performance, Reliability

1. INTRODUCTION

Challenges for interdomain routing. Interdomain routing faces several fundamental challenges. One is *scalability*: routers running the Internet's interdomain routing protocol, Border Gateway Protocol (BGP) [25], require state that scales linearly in the number of IP prefixes advertised in the Internet. This is particularly a concern in the data plane where the router stores the routing table, or forwarding information base (FIB). Because it has to operate at high speeds and often uses SRAM rather than commodity DRAM, FIB memory is arguably more constrained and expensive than other resources in a router [22]. Moreover, the number of IP prefixes is increasing at an increasing rate [15], leading to the need for expensive hardware and upgrades. The Internet Architecture Board Workshop on Routing and Addressing recently identified FIB growth as one of the key concerns for future scalability of the routing system [22].

A second challenge for interdomain routing is to provide *multipath routing*, in which a packet's source (an end host or edge router) selects its path from among multiple options. For network users, multipath routing is a solution to two important deficiencies of BGP: poor reliability [1, 14, 17] and suboptimal path quality, in terms of metrics such as latency, throughput, or loss rate [1, 27]. Sources can observe end-to-end failures and path quality and their effect on the particular application in use. If multiple paths are exposed, the end-hosts could react to these observations by switching paths much more quickly and in a more informed way than BGP's control plane, which takes minutes or tens of minutes to converge [19, 21]. For network providers, multipath routing represents a new service that can be sold. In fact, route control products exist today which dynamically select paths based on availability, performance, and cost for multi-homed edge networks [3]; exposing more flexibility in route selection could improve their effectiveness. Greater choice in routes may bring other benefits as well, such as enabling competition and encouraging "tussles" between different parties to be resolved within the protocol [6].

But providing multiple paths while respecting network owners' policies is nontrivial. BGP provides no multipath service; it selects a single path for each destination, which it installs in its FIB and advertises to its neighbors. Several multipath routing protocols have been proposed, but these have tradeoffs such as not supporting all of BGP's routing policies [32, 30], exposing only a limited number of additional paths [28], making it difficult to know which paths will be used [31, 23], or increase the size of the FIB [28, 31, 23], which would exacerbate the scalability challenge.

Our contributions. This paper addresses the challenges of scalability and multipath routing with a novel protocol called *pathlet routing.* In pathlet routing, each network advertises *pathlets*—fragments of paths represented as sequences of virtual nodes (vnodes) along which the network

*The first and fourth authors were supported in part by a Cisco Collaborative Research Initiative grant.

SIGCOMM'09, August 17–21, 2009, Barcelona, Spain.

Platypus [24] is similar to loose source routing except each waypoint can be used only by authorized sources to reach either any destination, or a specified IP prefix. Pathlet routing supports a different set of policies and enforces these using the presence or absence of forwarding tables, rather than cryptography.

R-BGP [18] adds a small number of backup paths that ensure continuous connectivity under a single failure, with relatively minimal changes to BGP. However, it somewhat increases forwarding plane state and is not a full multipath solution. For example, sources could not use alternate paths to improve path quality.

LISP [9] reduces forwarding state and provides multiple paths while remaining compatible with today's Internet. Although it can limit expansion of forwarding table size, LISP's forwarding tables would still scale with the size of the non-stub Internet, as opposed to scaling with the number of neighbors as in our LT policies.

MPLS [26] has tunnels and labels similar to our pathlets and FIDs. It also shares the high level design of having the source or ingress router map an IP address to a sequence of labels forming a source route. However the common use of these mechanisms is substantially different from pathlet routing: tunnels are not typically concatenated into new, longer tunnels, or inductively built by adding one hop at a time. To the best of our knowledge MPLS has not been adapted to an interdomain policy-aware routing.

Metarouting [13], like pathlet routing, generalizes routing protocols. It would be interesting to explore whether pathlet routing can be represented in the language of [13].

8. CONCLUSION

Pathlet routing offers a novel routing architecture. Through its building blocks of vnodes and pathlets, it supports complex BGP-style policies while enabling and incentivizing the adoption of policies that yield small forwarding plane state and a high degree of path choice. We next briefly discuss some limitations and future directions.

We suspect it is possible to optimize our path vector-based pathlet dissemination algorithm. The techniques of [16] may be very easy to apply in our setting to reduce control plane memory use from $O(\delta\ell)$ to $O(\ell)$ per pathlet, where δ is the number of neighbors and ℓ is the mean path length. Routers could also pick dissemination paths based on heuristics to predict stability, which for common failure patterns can significantly reduce the number of update messages [12]. The more radical approach of [30] could also be used to dramatically reduce state in Internet-like environments.

Traffic engineering is an important aspect of routing that we have not evaluated. One common technique—advertising different IP to different neighbors to control inbound traffic—is straightforward to do in our LT policies. But source-controlled routing would dramatically change the nature of traffic engineering, potentially making it more difficult for ISPs (since they have less control) and potentially making it easier (since sources can dynamically balance load).

Acknowledgements

We thank the authors of [7] for supplying the Internet-like topologies.

9. REFERENCES

[1] D. G. Andersen, H. Balakrishnan, M. F. Kaashoek, and R. Morris. Resilient overlay networks. In *Proc. 18th ACM SOSP*, October 2001.

[2] Routing table report. http://thyme.apnic.net/ap-data/2009/01/05/0400/mail-global.

[3] Avaya. Converged network analyzer. http://www.avaya.com/master-usa/en-us/resource/assets/whitepapers/ef-lb2687.pdf.

[4] B. Awerbuch, D. Holmer, H. Rubens, and R. Kleinberg. Provably competitive adaptive routing. In *INFOCOM*, 2005.

[5] CAIDA AS ranking. http://as-rank.caida.org/.

[6] D. Clark, J. Wroclawski, K. Sollins, and R. Braden. Tussle in cyberspace: defining tomorrow's Internet. In *SIGCOMM*, 2002.

[7] X. Dimitropoulos, D. Krioukov, A. Vahdat, and G. Riley. Graph annotations in modeling complex network topologies. *ACM Transactions on Modeling and Computer Simulation (to appear)*, 2009.

[8] J. P. (ed.). DARPA internet program protocol specification. In *RFC791*, September 1981.

[9] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis. Locator/ID separation protocol (LISP). In *Internet-Draft*, March 2009.

[10] B. Ford and J. Iyengar. Breaking up the transport logjam. In *HOTNETS*, 2008.

[11] L. Gao and J. Rexford. Stable Internet routing without global coordination. *IEEE/ACM Transactions on Networking*, 9(6):681–692, December 2001.

[12] P. B. Godfrey, M. Caesar, I. Haken, S. Shenker, and I. Stoica. Stable Internet route selection. In *NANOG 40*, June 2007.

[13] T. Griffin and J. Sobrinho. Metarouting. In *ACM SIGCOMM*, 2005.

[14] K. P. Gummadi, H. V. Madhyastha, S. D. Gribble, H. M. Levy, and D. Wetherall. Improving the reliability of internet paths with one-hop source routing. In *Proc. OSDI*, 2004.

[15] G. Huston. BGP routing table analysis reports, 2009. `http://bgp.potaroo.net/`.

[16] E. Karpilovsky and J. Rexford. Using forgetful routing to control BGP table size. In *CoNEXT*, 2006.

[17] N. Kushman, S. Kandula, and D. Katabi. Can you hear me now?! it must be BGP. In *Computer Communication Review*, 2007.

[18] N. Kushman, S. Kandula, D. Katabi, and B. Maggs. R-BGP: Staying connected in a connected world. In *NSDI*, 2007.

[19] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet routing convergence. In *ACM SIGCOMM*, 2000.

[20] K. Lakshminarayanan, I. Stoica, S. Shenker, and J. Rexford. Routing as a service. Technical Report UCB/EECS-2006-19, UC Berkeley, February 2006.

[21] Z. M. Mao, R. Bush, T. Griffin, and M. Roughan. BGP beacons. In *IMC*, 2003.

[22] D. Meyer, L. Zhang, and K. Fall. Report from the iab workshop on routing and addressing. In *RFC2439*, September 2007.

[23] M. Motiwala, M. Elmore, N. Feamster, and S. Vempala. Path splicing. In *ACM SIGCOMM*, 2008.

[24] B. Raghavan and A. C. Snoeren. A system for authenticated policy-compliant routing. In *ACM SIGCOMM*, 2004.

[25] Y. Rekhter, T. Li, and S. Hares. A border gateway protocol 4 (BGP-4). In *RFC4271*, January 2006.

[26] E. Rosen, A. Viswanathan, and R. Callon. Multiprotocol label switching architecture. In *RFC3031*, January 2001.

[27] S. Savage, T. Anderson, A. Aggarwal, D. Becker, N. Cardwell, A. Collins, E. Hoffman, J. Snell, A. Vahdat, G. Voelker, and J. Zahorjan. Detour: Informed Internet routing and transport. In *IEEE Micro*, January 1999.

[28] W. Xu and J. Rexford. MIRO: Multi-path Interdomain ROuting. In *SIGCOMM*, 2006.

[29] X. Yang. NIRA: a new Internet routing architecture. Technical Report Ph.D. Thesis, MIT-LCS-TR-967, Massachusetts Institute of Technology, September 2004.

[30] X. Yang, D. Clark, and A. Berger. NIRA: a new inter-domain routing architecture. *IEEE/ACM Transactions on Networking*, 15(4):775–788, 2007.

[31] X. Yang and D. Wetherall. Source selectable path diversity via routing deflections. In *ACM SIGCOMM*, 2006.

[32] D. Zhu, M. Gritter, and D. Cheriton. Feedback based routing. *Computer Communication Review (CCR)*, 33(1):71–76, 2003.