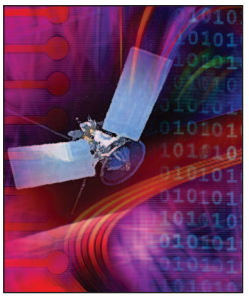
Editor: Chris Metz • chmetz@cisco.com



The Transition to IPv6, Part II

The Softwire Mesh Framework Solution

Yong Cui, Jianping Wu, Xing Li, and Mingwei Xu • *Tsinghua University*
Chris Metz • *Cisco Systems*

Part I of this series described a prototype solution that provides dynamic IPv4 routing and forwarding across the IPv6-based China Education and Research Network (CERNET2). This work spawned an effort in the IETF to develop a generalized method for routing and tunneling different address families across uniform IPv4 or IPv6 backbone networks. Inspired by the CERNET2 effort, the IETF Softwires working group has introduced a framework for a solution that offers a generalized, network-based capability for routing and tunneling multiple address families across native IPv4 or IPv6 backbone networks.

With the transition under way from IPv4 to IPv6, some countries (including China) are establishing large-scale native IPv6 backbone networks. Among the significant challenges in such efforts is the need to support large numbers of IPv4-based Internet applications and services across these native IPv6 backbones. CERNET2 researchers developed the Border Gateway Protocol (BGP)-based 4over6 implementation, described in Part I of this series, to address this requirement.[1,2]

At the same time, other ISP and enterprise backbone providers are being asked to support IPv6 routing and forwarding across their IPv4-based backbone networks. They can employ one of the many existing IPv6-over-IPv4 transition tunneling schemes that have been defined (and in some cases implemented) through the years. Yet, some of these schemes are limited in their functional scope (working only in LAN environments, for example), others fail to scale because they require extensive manual configuration, and still others require special IPv6 addressing schemes to work effectively. What the industry really needs to support the transition to IPv6 is a generalized, network-based client IP(*i*)-over-backbone IP(*j*) solution (in which *i* and *j* denote different IP address families [AFs]).

The softwire mesh framework is an extension of the China Education and Research Network (CERNET2) 4over6 solution. By employing IP tunnels or Mulitprotocol Label-Switching (MPLS) tunnels, called softwires, it can enable connectivity between islands of IPv6, IPv4, or dual-stack networks across single IPv4 or IPv6 backbone networks. This solution can reuse existing multi-AF routing mechanisms such as BGP as well as existing IP (and label) tunnel encapsulation schemes where appropriate. The intent is to encourage multiple, interoperable vendor implementations in the hope that operators will find it easier and more attractive to support the transition to IPv6.

Softwire Mesh Solution

Following the requirements set forth in the "Softwire Problem Statement" Internet draft,[3] a generalized, network-based client AF(*i*)-over-backbone AF(*j*) routing and forwarding solution needs to support the following functions[4]:

- The AF(*j*) backbone network forwards packets with headers or labels based on AF(*j*).
- Local provider edge (PE) routers discover sets of AF(*j*) tunnel-encapsulation parameters and tunnel endpoints located on remote PE routers.
- A set of PE routers dynamically establishes a

 Published by the IEEE Computer Society 1089-7801/06/$20.00 © 2006 IEEE

mesh of inter-PE tunnels with tunnel headers based on AF(*j*). Ingress PE routers direct client AF(*i*) packets into the appropriate tunnels according to destination client AF(*i*) prefixes and next hops reachable through the other end of the tunnel.
- Local PE routers store client AF(*i*) prefixes, AF(*i*) next hops, and tunnel-identifier/next-hop addresses and distribute them in a scalable fashion to interested remote PE routers. (The tunnel identifier/next-hop addresses bind the advertised client AF(*i*) prefix/next hops with established inter-PE tunnels leading to that prefix and terminating in the tunnel next-hop address.)
- Ingress PE routers encapsulate client AF(*i*) packets in backbone AF(*j*)-based tunnel headers (IP or labels) and forward them across the backbone AF(*j*) network.

These functions must operate with an interchangeable mix of different AF and tunnel-encapsulation types. For example, client AF(*i*) prefixes could be native IPv4, native IPv6, virtual private network (VPN) IPv4, or VPN IPv6. The backbone AF(*j*) could be native IPv6 or native IPv4. Moreover, the tunnel-encapsulation types could be IP-IP,[5] Generic Routing Encapsulation (GRE),[6] or Layer-2 Tunneling Protocol, version 3 (L2TPv3),[7] and other alternatives are certainly possible. MPLS tunnels could possibly even progress client AF(*i*) packets across the AF(*j*) backbone network, consistent with MPLS VPN solutions already deployed in many networks today.

Softwire Mesh Architecture

When describing the softwire mesh framework, we refer to PE routers as AF border routers (AFBRs). These dual-stack AF(*i*, *j*) routers, positioned at the edge of the transit core, peer with one or more customer edge (CE) routers located inside the AF access island to exchange AF access-island reachability information. AFBR nodes also peer with each other directly or via BGP route reflectors to exchange softwire configuration information, perform softwire signaling, and advertise routing information for AF access islands that can be reached through softwires.

The single AF(*j*) transit core is an IPv4 or IPv6 backbone network surrounded by a periphery of AFBRs. It provides inter-access-island connectivity across a mesh of softwires (hence the term *softwire mesh*). Single AF(*j*) access islands (same AF as the core) can communicate across the transit core using softwires or normal default routing functions, depending on the operator's wishes and the system's routing configuration.

Whether single AF(*i*) or dual-stack AF(*i*, *j*), access islands rely on the transit core for connectivity to remote access-island networks of the same AF. Routers inside an access island will run a routing protocol, and a subset of access island CE routers will peer with upstream AFBRs to exchange client AF(*i*) or AF(*i*, *j*) reachability information.

Softwire tunnel configuration information, which we refer to as softwire encapsulation sets (SW-encap sets), comprises the one or more tunnel-encapsulation types and parameters supported on a given AFBR. Softwire signaling involves the local definition of SW-encap sets on the AFBRs as well as the dynamic establishment of softwires in which peering AFBRs exchange their configured SW-encap sets plus their own IP addresses. Once the sets are in place, each AFBR has sufficient information to encapsulate and then forward packets to prefixes that are reachable via a softwire through any other AFBR in the mesh.

Using BGP to Set Up Tunnels and Advertise Prefixes

Multiprotocol-BGP (MP-BGP) is an ideal choice for softwire signaling.[8] First, it supports the one-to-many signaling paradigm required by the egress AFBR to communicate softwire information to multiple ingress AFBRs. Second, BGP need only operate between softwire-capable AFBR nodes, given that these are the only devices that maintain softwire tunneling state. This saves the routers inside the transit core from having to process softwire-specific messages. Finally, BGP is extensible and so can easily carry softwire information between AFBRs.

An Internet draft coauthored by Cisco engineers has defined a new subsequence address family identifier (SAFI), called the tunnel SAFI,[9] as a method for using BGP to communicate tunnel-specific information among BGP-speaking routers, including AFBRs. The tunnel SAFI comprises the following elements:

- A new SAFI value (equal to 64) contained in the BGP `MP_REACH_NLRI` attribute indicates that the network layer reachability information field pertains to an IPv4 (AFI=1) or IPv6 (AFI=2) tunnel. (We refer to the `MP_REACH_NLRI` attribute message with a SAFI value equal to 64 simply as the *tunnel SAFI*.)

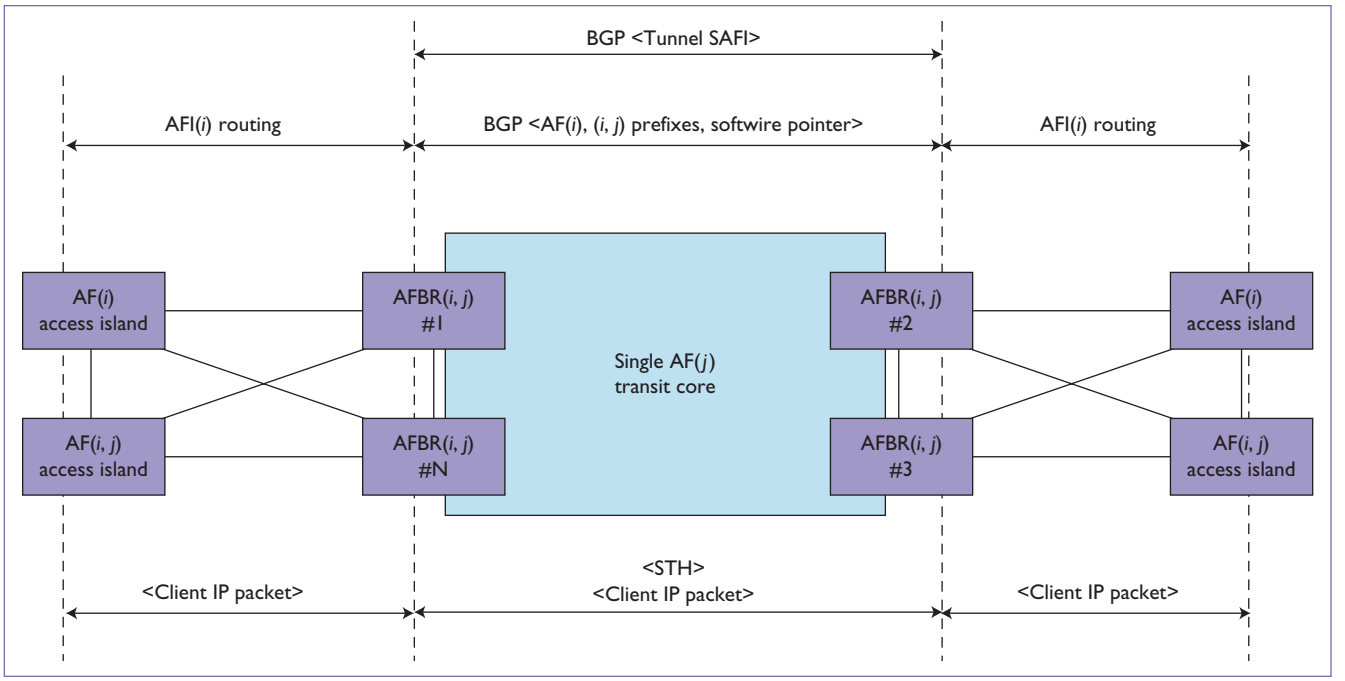


Figure 1. Border Gateway Protocol (BGP) tunnel SAFI (subsequence address family identifier) message exchanges between dual-stack address family border routers (AFBRs) that automatically establish the mesh of inter-AFBR softwire tunnels. Egress AFBR nodes will advertise BGP updates with pointers to the softwire tunnels to use to reach the AF access-island networks included in the update messages.

- The tunnel SAFI's NLRI is encoded with a tunnel identifier and the tunnel endpoint's IP address (the egress AFBR's address, in this case). To associate reachability with an advertised prefix through a given softwire, subsequent BGP prefix advertisements will include pointers to index the identifiers and IP addresses that in turn point to the given softwire to use.

The tunnel SAFI might also be associated with one or more additional attributes, including payload AFI and SAFI and softwire-encapsulation parameters (L2TPv3 header parameters, for example). Including the payload information tells the AFBR up front what types of IP packets it needs to process upon their exit from the softwire. The egress AFBR employs MP-BGP to distribute the tunnel SAFI and associated attributes (from the encapsulation parameters) to signal softwire setups to interested AFBR nodes.

Once the softwire mesh is in place, any ingress AFBR can forward packets over a softwire to any egress AFBR. In essence, the egress AFBR nodes use normal MP-BGP routing mechanics to advertise client AF access-island reachability to the set of interested ingress AFBRs. But they also include "pointers" to existing softwire tunnels, basically informing the ingress AFBR which softwire to use to reach the egress AFBR's prefix. The IETF Softwires working group is still working to define the specific "pointer" mechanism. Figure 1 illustrates the components of the softwire mesh framework architecture and BGP flows for softwire signaling and prefix advertisements.

Existing IP Tunnel Encapsulations

In forwarding packets from an AF access island onto a softwire originating in an AFBR, the ingress AFBR takes the following steps:

- look up AF access-island IP destination addresses in the respective AF access-island routing and forwarding table;
- encapsulate the IP packet in the appropriate softwire transport header (STH); and
- transmit the softwire-encapsulated packets across the single AF transit core based on the STH.

When packets arrive, the egress AFBR removes the STH, performs a lookup of the original IP packet in the corresponding AF access-island routing and forwarding table, and transmits the native AF access-island IP packet toward the respective downstream CE router. The softwire mesh frame-

Table 1. Comparison of IP(x)-over-IP(y) transition solutions.

Solution	Scenario	Advantage	Disadvantage
IPv6 configured tunnels[10]	Manual inter-router configured IPv6-over-IPv4 tunnels	Stable and commonly deployed	Configuration burden
Auto 6to4 tunnel[11]	Offers IPv6 site connectivity to the IPv6 Internet across IPv4 networks	Simple configuration	Security risks because relay routers must accept packets from anywhere on the Internet
Intra-site Automatic Tunnel Addressing Protocol (ISATAP)[12]	Applied to IPv6 hosts and routers for connectivity over IPv4 networks	Simple IPv6 transition technique in small enterprise or campus networks	Suboptimal routing and no IPv6 multicast
China Education and Research Network (CERNET) 4over6[1]	Dynamic IPv4-over-IPv6 tunneling between provider edge (PE) routers using the Border Gateway Protocol (BGP)	Simple dynamic tunnel setup using BGP	Limited to IPv4 (or IPv6)-over-IPv6 tunnels (IP-in-IP or Generic Routing Encapsulation)
Multiprotocol Label Switching (MPLS) VPNv6[13]	Dynamic IPv6 VPN connectivity over MPLS backbones	Coexistence of VPNv4 and VPNv6 across MPLS backbones	No IPv4-over-IPv6 support; access island prefixes must be stored in virtual private network routing tables on PE routers
Softwire mesh framework[4]	Applied to PE routers with IPv6 or IPv4 backbones	Automatic tunneling, good scalability to support any AFI/SAFI over IPv4 or IPv6	Some details still under development in IETF

work is designed to accommodate any form of IP tunnel encapsulation, including IP-in-IP, GRE, L2TPv3, and MPLS encapsulation.

Comparison to Other Solutions

Table 1 compares various solutions that let providers tunnel IPv6 packets across IPv4 backbones; the main limitation evident among these approaches is that they fail to perform the converse. The softwire mesh framework is a network-based solution that exploits the scalability of inter-AF communications by using existing IP tunnel-encapsulation methods and extending the BGP protocol to enable providers to pass packets of either IP AF across a backbone network of the other AF.

The initial revision of the softwire mesh framework draft was presented at the Montreal IETF meeting in July 2006 and adopted as an official working group document. Efforts are under way in the Softwires working group to settle some details, including how to associate IP access-island reachability with softwires and how to advertise IPv4 reachability with IPv6 tunnel endpoints. We expect these issues to be resolved over the next couple of meetings, at which point the document will become a formal RFC standard and operators will be able to test or deploy the techniques to improve IPv6 transition.

In the next installment in this series, we'll describe the details of the softwire mesh solution as they could be applied in several different provider backbone networks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (no. 60403035, 90604024) and the National Major Basic Research Program of China (no. 2003CB314801).

References

1. J. Wu, Y. Cui, and X. Li, "4over6 Transit Using Encapsulation and BGP-MP Extension," IETF Internet draft, Feb. 2006; work in progress.
2. J. Wu et al., "The Transition to IPv6, Part I: 4over6 for the China Education and Research Network," *IEEE Internet Computing*, May 2005, pp. 54–59.
3. S. Dawkins, "Softwire Problem Statement," IETF Internet draft, Dec. 2005; work in progress.
4. J. Wu et al., "A Framework for Softwire Mesh Signaling, Routing and Encapsulation across IPv4 and IPv6 Backbone Networks," IETF Internet draft, June 2006; work in progress.
5. A. Conta and S. Deering, *Generic Packet Tunneling in IPv6 Specification*, IETF RFC 2473, Dec. 1998; www.ietf.org/rfc/rfc2473.txt.
6. D. Farinacci et al., *Generic Routing Encapsulation* (GRE), IETF RFC 2784, Mar. 2000; www.ietf.org/rfc/rfc2784.txt.
7. J. Lau, M. Townsley, and I. Goyret, *Layer-Two Tunneling Protocol – Version 3* (L2TPv3), IETF RFC 3931, Mar. 2005; www.ietf.org/rfc/rfc3931.txt.
8. T. Bates et al., *Multiprotocol Extensions for BGP-4*, IETF RFC 2858, June 2000; www.ietf.org/rfc/rfc2858.txt.
9. G. Nalawade et al., "BGP Tunnel SAFI," IETF Internet draft, June 2006; work in progress.

10. E. Nordmark and R.E. Gilligan, *Basic Transition Mechanisms for IPv6 Hosts and Routers*, IETF RFC 4213, Oct. 2005; www.ietf.org/rfc/rfc4213.txt.
11. B. Carpenter and K. Moore, *Connection of IPv6 Domains via IPv4 Clouds*, IETF RFC 3056, Feb. 2001; www.ietf.org/rfc/rfc3056.txt.
12. F. Templin et al., "Intra-Site Automatic Tunnel Addressing Protocol (ISATAP)," IETF Internet draft, May 2004; work in progress.
13. J. De Clercq et al., "BGP-MPLS VPN Extension for IPv6 VPN," IETF Internet draft, Feb. 2005; work in progress.

Yong Cui is an assistant professor in the computer science department at Tsinghua University, Beijing. His current research interests include Internet architectures, IPv6 transition, and quality of service. Cui has a PhD in computer science from Tsinghua University. Contact him at cy@csnet1.cs.tsinghua.edu.cn.

Jianping Wu is a full professor in the computer science department at Tsinghua University, Beijing, and the director of the China Education and Research Network. His current research interests include computer network architectures, next-generation Internet, and formal methods. Wu has a PhD in computer science from Tsinghua University. Contact him at jianping@cernet.edu.cn.

Xing Li is a full professor in the electronic engineering department at Tsinghua University, Beijing, and a vice director of the China Education and Research Network. His current focus is on Internet architectures, IP multicast, and routing architectures. Li has a PhD in electrical engineering from Drexel University. Contact him at xing@cernet.edu.cn.

Mingwei Xu is a full professor in the computer science department at Tsinghua University, Beijing. His current research interests include Internet architectures, IPv6 transition, multicast, and mobility. Xu has a PhD in computer science from Tsinghua University. Contact him at xmw@csnet1.cs.tsinghua.edu.cn.

Chris Metz is a technical leader in the Routing Technology Group for Cisco Systems, based in San Jose, California. His current areas of interest include Internet architectures and services, IP/MPLS, transport-layer protocols, and layer-2/layer-3 virtual private networks. Contact him at chmetz@cisco.com.