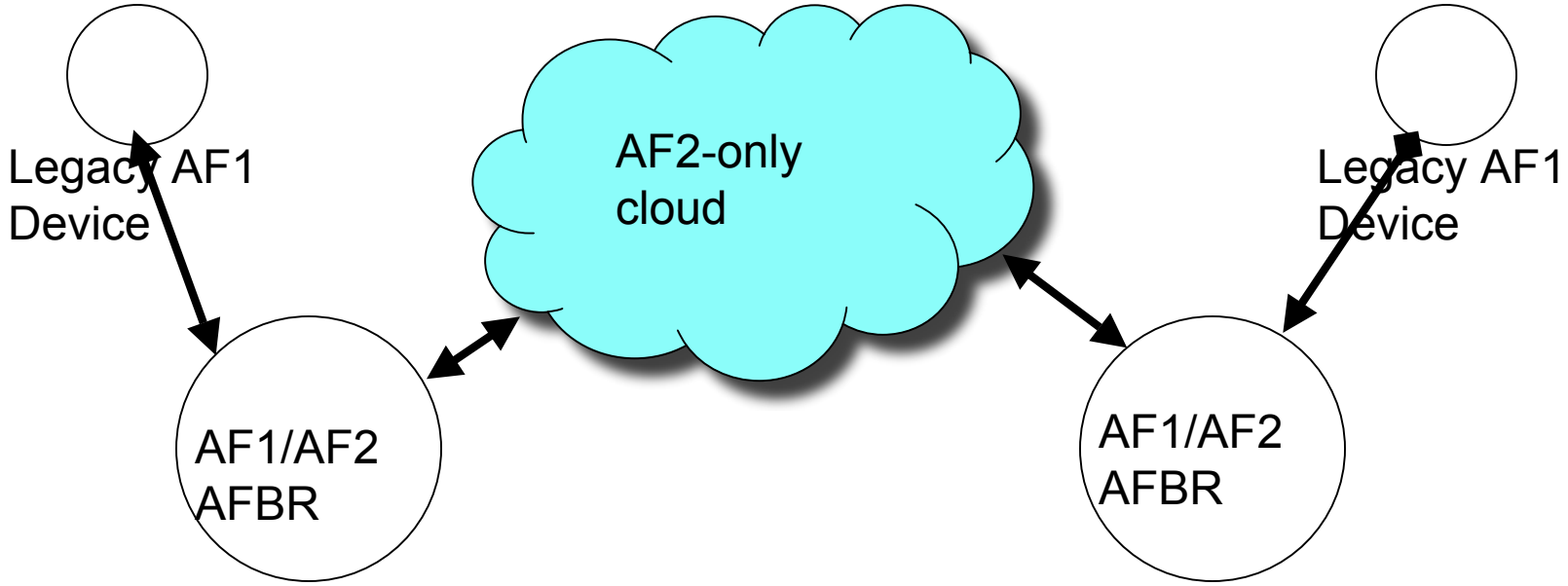
Mesh Update
Softwires Interim Meeting
Barcelona
Sept 2006

jgs@cisco.com
erosen@cisco.com
Presented by David Ward

Softwires “Mesh” Scenario



- Problem: To carry the AF1 traffic over the AF2 network, obeying certain constraints on solution

Solution Constraints

- No changes to legacy AF1 equipment
- Only AFBRs are dual-AF
- No AF1 routing protocols within core
- No native AF1 data packets within core
- Legacy doesn’t know that core does not support AF1 natively: no explicit mesh of legacy-legacy adjacencies
- If the solution involves tunneling, no one tunneling technology can be required

Design Center

- We talk in general of AF1 over AF2 core, but design center is:
  - AF2 either IPv4 or IPv6, AF1 is the other
- N.B.: If AF1 is a “VPN address family”, this becomes the already-solved L3VPN problem, which is out of scope for this WG.

Solution Components

- Routing:
  - Ingress AFBR must know egress AFBR for AF1 prefix
  - Core routers do not have AF1 routing
  - So AFBRs use BGP to distribute AF1 routes among themselves
  - Well understood model
- Data Plane
  - BGP provides AFBR next hop for each AF1 prefix
  - Since core does not support AF1 packets:
    - AFBRs must tunnel packets to each other,
    - Using tunneling technology that works in AF2 core

Routing: AF1 Prefixes with AF2 Next Hops

- AFBR “perspective” on Next Hops for AF1 prefixes:
  - next hop across the core is another AFBR
  - since the AFBRs can only communicate with each other via AF2, the next hop for an AF1 prefix is an AF2 address
- BGP Terminology:
  - the address or address prefix whose route is being distributed is known as *NLRI*.
  - each route associates a *next hop (NH)* with an NLRI.
  - our model allows AF1 NLRI to have AF2 NH
- Several precedents for BGP NLRI and NH in different AF
  - non-IP NLRI with IPv4 and/or IPv6 NH
  - VPN-IP NLRI with IPv4 and/or IPv6 NH
- Issue: precedents use different methods of NH encoding

NH Encoding Issue

- Existing precedents never converged on single way to determine the AF of the NH, given that it is a different AF than the NLRI
  - The text says that "the network layer protocol associated with the network address of the next hop is identified by a combination of <AFI, SAFI>"
  - Original intention: a particular AFI/SAFI could be defined to mean, e.g., that the NLRI is VPN-IPv4 and the NH is IPv4.
  - It does seem to rule out the use of the length field to determine the NH address family, but if we want to amend it to formally allow use of the length field; I don't think anyone will oppose
- No one proposes to invalidate the installed base, but in retrospect the NRLI-NH relationship was incorrectly specified, and the need to have IPv6 NHs for NLRI that is shorter than an IPv6 address forces the issue.

NH Encoding Issue.2

- Some techniques make NH look like it is in AF of NLRI, even if it isn't:
  - IPv6 prefix with IPv4 NH, NH address can be coded as v6. (Doesn't work in reverse.)
  - VPN-IP prefix with IPv4 NH, NH address is coded as if it were VPN-IP prefix, with special bytes set to zero.
  - This technique generally regarded as confusing and silly.
- Other existing models use the length of the NH to distinguish IPv4 from IPv6.
- Other possibility: use TLV encoding to specify AF of NH.

Encoding Alternatives

- Length-based encoding:
  - doesn’t add new syntax to BGP
  - does add new semantics
  - old syntax plus new semantics = not backwards compatible
  - E.g. you can't just start sending IPv6 NHs for IPv4 NLRI (distinguishing the NH type by the length field) and expect everyone to be able to process the messages correctly.

- TLV-based encoding:
  - new syntax and semantics (so also not backwards compatible)
  - better for extensibility and future-safety
  - new syntax does create some protocol issues, e.g., what if new and old are both present

- Both schemes require BGP capability:
  - All AFBRs must be able to understand new NH address semantics (deployment restriction)

Encoding Alternatives.2

- New SAF - why exp/informational?
  - SAF is used for something else and causes unintended consequences
  - OPS: use of multiple SAFIS is visible at mgmt level (config, troubleshooting) in a way other techniques are not.
    - Consider the route for prefix P between point A and point B. Suppose that along that path are some v6 NHs and some v4 NHs. This means that update which advertises the route to P will sometimes have one SAFI, sometimes another. So updates with one SAFI affect updates with the other.
  - FWDING: Routers with different with different NH will not be comparable in BGP but, will be in the forwarding table
  - MS-BGP: different BGP sessions for different AFI/SAFIs, must recognize that certain AFI/SAFI pairs should never be separated from others

Encoding Alternatives.3

- New SAF - con’t
  - SAFI Alloc: new SAFI for its real purpose, you actually have to allocate two.
    - E.g. SAFI for "multicast” which is somewhat misnamed; it is used to pass unicast routes,where the routes are only to be used as RPF routes for multicast. These are deliberately made non-comparable with ordinary unicast routes, so that the multicast topology can be made non-congruent to the unicast topology.
      - you'd need to get a second SAFI for that also, and we'd have the same issues about what's supposed to be comparable to what and what the relationship between the two SAFIs is really supposed to be.
  - Future: next hop in a VRF instead of the global table, we'll need VPN-IPv4 and VPN-IPv6 next hops, that will be two more SAFIs for each AFI.

Selecting an Encoding

- Encoding discussions can get very heated (long history in IDR):
  - everyone has an opinion
  - it doesn't matter much which choice is made
  - it's hard to prove that one way is always best
  - this combination of factors leads to a lot of noise
- But there really is no fundamental issue dividing proponents of the two alternatives
- Should be settled fairly readily in IDR WG after the standard rituals.

Data Plane: Encapsulation and Policy

- AF1 data packets must be tunneled through AF2 core from one AFBR to another
- How do we choose an appropriate tunneling technology?
- Typical case probably very simple:
  - administration selects tunneling technology to be used through its core
  - administration only deploys AFBRs that provide that technology
- Policy is configured at AFBR which does encapsulation (i.e., tunnel head end)

Conditional Policies

- Policy *cannot* be automatically deduced from information about capabilities of head end and tail end:
  - E.g., what if they both support MPLS, but core doesn't?
  - Policy must be configured at head end
- Policies *can* be made conditional on information gathered in real time about tail end or even about individual prefix

Example Conditional Policy

- Example:
  - use GRE to talk to “type X” AFBRs,
  - but use L2TP to talk to “type Y” AFBRs
- Conditional policy pre-configured
- Information about “type” of AFBR gathered in real time
- BGP can be used by an AFBR to distribute the information that it is, e.g., type Y.

BGP-Based Distribution of Information about an AFBR

- Like BGP-based auto-discovery used in L2/L3VPN
- Useful to have special BGP update to carry arbitrary information about originator of update: *Information-SAFI*
  - Carries factual info about AFBR
  - Info is opaque to BGP
  - Info often representable as arbitrarily assigned (ext.) community
  - Info and route distribution can be constrained independently
  - Factual info from tail end used as input to conditional policy configured at head end
- Very general but simple mechanism, allows administrators full control of policy

More on Policy

- Head end and tail end do *not* negotiate policy, or even suggest policies to each other
- Policies depending on facts about individual prefixes could be configured:
  - possibly based on attributes of prefixes
  - no need for tail end to dynamically assign prefixes to tunnels
- Polices with respect to QoS, TE, Security, could also be configured at head end

Tunnel Setup and Signaling

- Some tunneling technologies don't need anything but the tail address:
  - GRE (without optional key)
  - IP-in-IP
  - LDP-based MPLS
- Others have native signaling which can be used:
  - RSVP-TE
- But …

BGP for Tunnel Setup

- For some tunneling technologies:
  - tail needs to pass info for head to place in the encapsulation header,
  - but native signaling either doesn't exist or isn't right for this application
- Examples:
  - GRE if optional key is to be used
  - L2TP

BGP for L2TPv3 Tunnel Setup

- Tail end must pass session id and cookie value to head ends
  - for given tail end, all head ends can use same session id and cookie
  - best done via p2mp signaling
    - but L2TPv3 native signaling is p2p
    - makes sense to pass this info via BGP.
  - good use for the information Update.

What's Not Needed

- Prioritized lists of encapsulation types
  - policies configured at head end
  - no negotiation of policy between head and tail
  - so prioritized list from tail doesn't make much sense
- Alt: Simplification over some previous proposals,without loss of generality

Payload Type Identification

- Some tunnel technologies have native means of identifying payload type
- Others require demultiplexor value to be distributed by BGP with other encaps info
- Details to be worked out for each encaps type in new alt

Security?

- Scope is Internet traffic, not VPN traffic
  - If confidentiality or integrity is required inside the tunnel, it's also required outside the tunnel, so no new confidentiality requirement
- Spoofed encapsulation header is possible
  - but without the tunnel, a spoofed payload packet would be possible, so no new authentication requirement

Security??

- Security folks didn't take "no" for an answer:
  - Allow tunnel-mode IPsec as one of the IP-based tunneling technologies
  - Given large number of tunnels, requires automated key distribution (IKE)
- Need for security probably not prefix-dependent or dependent on tunnel endpoints
- In progress

Multicast

- Next time!

Next Steps

- Review status in San Diego softwires meeting
- Show alt encoding of info safi
  - Agree on functionality, contents and encoding
- Present issues/framework in IDR WG
  - NH seems first order of business
- Post San Diego
  - Finish Multicast and Security portions of framework
  - Present info safi in IDR
- Question: Once framework is done do we want to move all work to IDR?