Shim6: Network Operator Concerns

Jason Schiller
Senior Internet Network Engineer
IP Core Infrastructure Engineering
UUNET / MCI

Not Currently Supporting IPv6?

Many parties are going forward with IPv6

- Japan has a large scale adoption
- US government and government contractors are actively pursuing IPv6
- Some application providers are running IPv6 to develop applications
- Some customers are experimenting to gain comfort with IPv6 and migration
- Some customers are concerned about future migration to IPv6 and want a provider with IPv6 experience

Not Currently Supporting IPv6?

- IETF is currently addressing IPv6 problems and protocols to solve them
- RIRs are currently setting IPv6 policy

You will be running IPv6 eventually

If you don’t get involved now, you may have a solution which is non-useful

It is easier to start with good protocols and policies, rather than changing them later

Background: IPv6 Address Size

- IPv4 has 2^32 IP addresses (4,294,967,296)
- IPv4 largest unicast Internet routable block /24 (16,777,184)
- Concerns about address exhaustion in some countries
- Use of Network Address Translation (NAT) to reduce consumption

- IPv6 has 2^128 IP addresses
- 64 bits reserved for host, 64 bits reserved for network
- IPv6 Unicast routable space 2000::/3 (2,305,843,009,213,693,952 /64s)
- 137,439,215,616 times more IPv6 /64s than IPv4 /24s

Background: IPv6 Impact

Extra routing state:

- Consumes routing memory (RIB)
- Consumes forwarding memory (FIB)
- Affects forwarding rate
  (FIB lookup as a function of memory speed and size)
- Affects convergence
  (SPF, RIB rewrite, RIB to FIB population)

Background:
IPv6 Routing Table Size Predictions

- Predictions about IPv6 routing table size vary greatly
- Predictions about time to wide spread adoption vary greatly
  - –US Federal Government mandates 2008 IPv6 capable
  - – Speed of Japanese IPv6 adoption
  - –Predictions of exhaustion of IPv4 space in the US (2016 – 2009)

Background: Current IPv4 Route Classification

- Three basic types of IPv4 routes
  - Aggregates
  - De-aggregates from growth and assignment of a non-contiguous block
  - De-aggregates to perform traffic engineering
- Tony Bates CIDR report shows:

DatePrefixes	Prefixes	CIDR Agg
23-09-05	166,976	112,062

- Can assume that 55K intentional de-aggregates

Background: Current IPv4/IPv6 Routing Table Size

- Assume that tomorrow everyone does dual stack
- Current IPv4 Internet routing table 166K routes
- Current tier 1 ISP internal routes 50K-100K routes
- 20-30K IPv6 Aggregates (1/active AS or 1/assigned AS)
- 55K intentional IPv6 de-aggregates for traffic engineering
- Internal IPv6 de-aggregates (1/static customer, 3/multihomed customer)
- Tier 1 ISPs require IP forwarding in hardware (6Mpps)
- Easily exceed the current FIB limitations of 300K-350K prefixes

IPv6 Route Table Explosion Solutions

- Throw hardware at the problem
  - Commit to scaling router memory size and speed to support very large RIB and FIB sizes
  - Commit to faster processors for SPF of larger tables
  - Optimize FIB storage and SPF processes
  - Hope hardware / software solution is available at least 5 years before wide spread adoption
  - Use 5 years to depreciate and replace current hardware with new hardware capable of holding larger routing information

IPv6 Route Table Explosion Solutions

- Minimize de-aggregation. Only allow service providers to announce a single aggregate.
  - –Solve multi-homing, host mobility, and provider independent space without de-aggregating
- The tradeoff is the information lost from the routing state will result in sub-optimal routing, or will need to be replace by some forwarding plane measurement.

Shim6 Solution For IPv6 Multi-homing

Goal: allow for IPv6 multi-homing without de-aggregation

- Provide multiple IP addresses to multi-homed end host
- Separate “Locator” from “Upper Layer ID”
- “Locator” – behind what network(s) on the Internet the host resides. Also IP address on end host interface.
- “Upper Layer ID” – what upper layer session (TCP/UDP) terminates on
- Allows IP source and / or destination (locators) to change without impacting upper layer communication
- Depend on source host to choose locator and this path to load
- Multi-homed destination has no ability to traffic engineer links as currently used

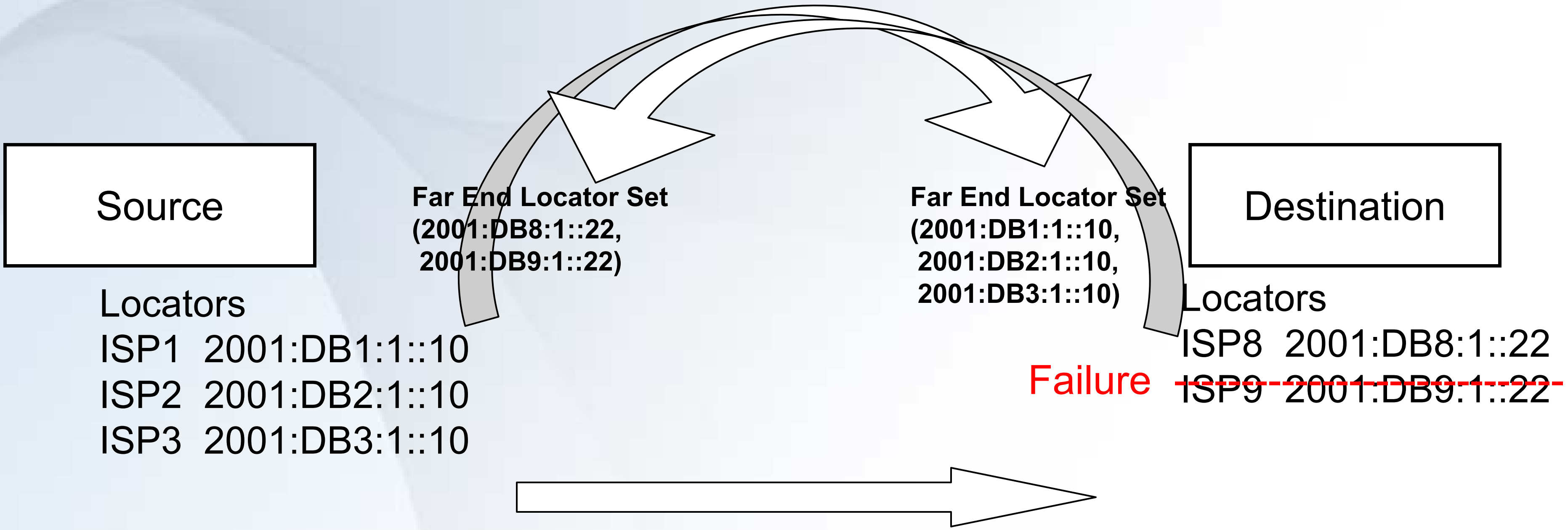
L3 Shim Approach

- Current solution documented in draft-ietf-multi6-l3shim-00.txt
- Multi-homed destinations and sources require one unique IP address (Locator) for each upstream ISP
- Multi-homed destinations and sources also require an Upper Layer ID (ULID)
- A shim containing the Upper Layer ID is inserted between the transport layer and the IP layer

L3 Shim Function

- AAAA DNS query provides a (possibly incomplete) set of Locator addresses
- Source chooses a Locator address to establish conversation on
  - Source is required to choose a different Locator ID if the first attempt is not successful
  - Once upper layer communication begins, end hosts can signal shim6 capabilities and pass a complete set of Locators.
  - Current source and destination Locators are used as source and destination Upper Layer IDs in the shim6
  - In the event of an outage source or destination can detect path failure in the forwarding plane, and change source or destination Locator to any in the Locator set.
  - Existing sessions continue uninterrupted using the unchanged Upper Layer ID

L3 Shim Function



Source

Locators
ISP1 2001:DB1:1::10
ISP2 2001:DB2:1::10
ISP3 2001:DB3:1::10

Far End Locator Set
(2001:DB8:1::22,
2001:DB9:1::22)

Far End Locator Set
(2001:DB1:1::10,
2001:DB2:1::10,
2001:DB3:1::10)

Destination

Locators
ISP8 2001:DB8:1::22
Failure ~~ISP9 2001:DB9:1::22~~

IP Header		L3 Shim			
IP Source	**IP Destination**	**Source ULID**	**Destination ULID**		
2001:DB1:1::10	~~**2001:DB9:1::22**~~ **2001:DB8:1::22**	**2001:DB1:1::10**	**2001:DB9:1::22**	**TCP/UDP/ICMP Header**	**Payload**
Source Locator	**Destination Locator**				

Inter-AS Traffic Engineering as a Requirement?

- Inter-AS TE is not currently a requirement for a shim6 solution
- Thought the problem was a lack of understanding of the basic inter-AS traffic engineering requirements
  - Previous attempts left out certain approaches (primary/backup, shortest path)
  - Documented specific cases instead of basic concepts
  - I attempted to document the basic approaches as building blocks
- Attempts to make Inter-AS traffic engineering a requirement failed
  - RFC3582 Site multi-homing requirements down graded to “goals”

Current IETF Focus

- Network operators where all of their consumer customers connect only to them are unconcerned with inter-AS traffic engineering and believe that simple fail-over solves the problem for 90% of the Internet.
- Focus is currently on forwarding path failure detection.
- Currently only concerned with designing the protocol. Believe that inter-AS traffic engineering is just a function of how a source orders the locator address set, and has no impact on the protocol.

Site Multi-homing vs. Host Multi-homing

- Current IPv4 inter-AS traffic engineering is accomplished at the network level
- Shim6 is done at the host level
  - –Shim6 is host multi-homing not site multi-homing
- How do you add network wide traffic engineering preferences to the host to host shim6 solution?

Site Multi-homing vs. Host Multi-homing

- Host multi-homing may be useful for consumer customers
  - Number of hosts at location is small
  - End user own host and network configuration
  - Routing equipment may have limited capabilities or is owned by service provider
- Site mutli-homing is more useful for large commercial customers
  - Large number of hosts
  - Complex routed network
  - End users do not own network or traffic engineering preferences

Inter-AS Traffic Engineering Impact on Shim6 Protocol

- In IPv4 multi-homing a destination can influence routing to determine what traffic will load inbound links local to the destination.
- In IPv4 multi-homing, a source can constrain routing and/or alter IGP metrics to influence how traffic will load outbound links local to the source.

- In IPv6 multi-homing the source has no BGP routing information, and has no IGP routing information
- In IPv6 multi-homing the source will need to be sent “clues” by the destination on how to order the destination locator address set, if the destination wants the ability to control what traffic will load inbound links local to the destination.
- In IPv6 multi-homed sources will not leverage BGP routing or IGP distance in choosing which source locator address to choose.
  - Upstream ISPs may filter on source IP address and only allow packets with a source address as part of the aggregate /32

IPv4 TE Inbound to Destination

In IPv4 multi-homing a destination can influence routing to determine what traffic will load inbound links local to the destination.

- Can use approximate shortest path based on AS path length
- Can designate links as primary, secondary, tertiary using community triggered local-pref to different upstream ASes, or MED to a single AS
- Can attempt to share traffic across links by advertising multiple more specifics
- Can dial traffic by shuffling multiple more specifics
- Can bias some traffic away from best path by AS pre-pending

IPv6 TE Inbound to Destination

- In IPv6 multi-homing a destination has one IP address for every upstream ISP
- Source makes initial session with one address at random, begins traffic exchange, and signals shim6 capabilities
- Can insert layer 3 shim at any time and migrate upper layer session to use Upper Layer ID
  - Current approach is to wait for an outage, but could be immediate
- Traffic inbound to the destination is determined by source
  - Prior to shim inbound traffic based on DNS choice
  - Post shim inbound traffic based on locator set order

IPv6 TE Inbound to Destination

Post layer 3 shim insertion, inbound traffic to destination determined by source's sorting of locator set

- Source lacks BGP routing information
- Source lacks any destination metric or preference information
- Source will require to be sent "clues" about destination's inbound preferences

IPv4 TE Outbound From Source

If source network receives full BGP routes then source network will:

- Honor destination preferences (more specific prefixes, local-pref, AS pre-pending)
- If destination sets no preferences, the source will use shortest AS path
- If the source hears multiple paths through the same AS it will choose lowest MED
- All things equal source will choose shortest exit
  - –Can modify out bound link traffic by adjusting IGP

IPv4 TE Outbound From Source

- Source network can filter inbound routes from some links to reduce outbound traffic
- Source network can set MED inbound to make links primary, secondary, tertiary… for one or more prefixes
- Source network can learn only default routes in order to load all links outbound
- Source network can modify IGP metrics with each above approaches to bias traffic away from outbound links

IPv6 TE Outbound From Source

- Source host must choose which locator to use as source IP address
- ISPs may ingress filter on source address
- If the host chooses the locator of one ISP but the network routes the traffic to another ISP the traffic may get dropped by the ISP
- Source may need to be sent “clues” from the local network to choose the correct source IP address
- Or Internet border routers will need to re-write IP source address on egress if the wrong locator is used

IPv4 Outbound / Inbound Conflicts

- If destination and source both connect to the same upstream ISP, then outbound source policy over rides destination
- If destination and source connect to different upstream ISPs, then outbound source policy will be used, and transit AS will honor destination inbound policy
- Transit ASes can traffic engineer traffic towards upstream transit providers to more equally load different links
- Transit ASes can distance one or more prefixes towards a specific AS

Four Approaches to Providing Destination Based Traffic Engineering

1. Destination host sends inbound preferences (TLVs) to source host along with locators

2. Allow routers to insert inbound preferences (TLVs) to host based locator exchange

3. Move shim function to routers

4. 8+8 and shim6 solution where routers re-write IP source or destination

In Bound preferences Exchanged By TLV

- A multi-homed source or source router can send “clues” about how to order the locator set to the destination
- Can attach a Type Length Value (TLV) to a locator or set of locators

  Can encode things like:

  – Link ordering information
  – Recommendation for source to choose “best” path based on some forwarding plane measurement such as round trip time
  – Weight for a locator or set of locators with regard to a particular “best” path forwarding plane measurement
  – Weight for a locator or set of locators with regard to a simple round-robin

Shim6 Locator ID and TLV Exchange Example

Destination is multi-homed to four upstream ISPs with one link each

- Destination wants links to ISP1 and ISP2 to be used as primary links
- Destination wants source to use best path inbound between ISP1 and ISP2 as measured by round trip time
- Destination wants to bias a small amount of inbound traffic away from the overloaded ISP1 link by dereferencing round trip measurements by 3ms
- Destination want links to ISP3 and ISP4 to be used as secondary links in a simple round-robin

Locator set	TLV information encoded
(ISP1, ISP2)	link order 1 – best measured by round trip
(ISP3, ISP4)	link order 2 – best chosen by round-robin
(ISP1)	round trip biased + 3ms

TLV “best” Path Problem

Determination of shortest path (similar to shortest AS) cannot be based on routing information

- End hosts lack routes
- Routing table lacks more specific prefixes

Shortest path must be based on forwarding plane measurement

- Similar problem space to path failure detection
- Measuring multiple paths may create a significant amount of traffic
- N^2 problem. Source and destination with 4 links each have 16 paths to test.

Management on End Hosts

Configuring inbound policy on every host may be problematic

- In IPv4, inter-AS traffic engineering policy is managed on the Internet facing routers and on each end host in IPv6.
- The Internet facing routers and end hosts may not be managed by the same group of operators
- If the Internet facing routers and the end hosts are managed by the same group of operators, those operators may want to manage the inter-AS traffic engineering policy in a few places (Internet facing routers) as opposed to may locations (every host)
- Can manage outbound preferences on Internet routers or TE server
  - Will require some protocol to push out TE preferences to all end hosts
  - May create additional security problems
  - Is added complexity justified?
  - What if end host has a locally configured TE preference?

Routers Sending TLV information

Routers can intercept locator set exchange and insert TLVs containing inbound traffic engineer preferences on behalf of the end host

- Easily lends itself to network wide inbound TE policy
- Can leverage information about routing outages
- Will require routers to re-write shim6 locator exchanges to add TLVs (at least one per session)
- Adds complexity to routers which may be difficult to support for consumer customers

Move Shim Function to Routers

- Allow routers to insert shim on behalf of end hosts
  - May create additional security / authentication problems
- Allow routers to insert a additional “network” shim
  - May create additional security / authentication problems
  - Routers will need to recognize which packets should be shimmed
  - Not all hosts will be multi-homed (embedded devices with small IP stacks)
  - Will require routers to insert shim into all transit packets from multi-homed host at line rate
  - Will need to support line rate
    - Current routers support 600Mpps
    - Largest measured link 6Mpps throughput
- Which routers to do the shimming? Ingress? Egress?

8+8 Solution: Rewriting IP Source & Destination

- Allow routers to re-write IP source and / or destination
- Router will need to able to map multiple networks to a given destination.
  - –Router can replace the network portion of the IP destination of one network for another
    - All hosts are required to be multi-homed
    - All multi-homed hosts must have the same “host” address on all networks or the router must maintain lots of address mapping state
  - –Router will need to keep locator set and do a NAT function
    - May be a considerable amount of state
- Rewriting source may solve ISP filtering issues

8+8 Problems

- Router will need to re-write packets at line rate
  - Current routers support 600Mpps
  - Largest measured link 6Mpps throughput
- Which router to do re-writing? Ingress? Egress? Transit?
  - Will all routers know about all source networks?
  - Transit routers may need to look past MPLS labels
- May break HBA/CGA
- Rewriting network address
  - Will break privacy addressing
  - Will break non-multi-homed hosts
- IPv6 NAT type solution
  - Requires all re-writing routers to keep source locator set state

Non-useful Transit AS Traffic Engineering

- Inter-As traffic engineering in IPv4 is accomplished by sending more specific routes to the Internet, and allowing these more specific routes to be reachable across all connected ASes.
  - This allows each transit AS to make its own decision about what is the "best" path to take. Each transit AS can manipulate which is the best path by manipulating route announcements heard from its peers.
- In IPv6 transit ASes can only manipulate routing for an ISP aggregate affecting all customers using the ISP aggregate as the routing table lacks more specifics

IPV6 Transit AS Traffic Engineering

Useful IPv6 transit AS traffic engineering

- If transit ASes are aware of the fact that a destination may be reachable through alternate locators and can forward to alternates if they are better
- If transit ASes can reach into the locator set exchanges and further poison TLV metrics then locator ordering by the source can be influenced

Packet Filtering and Firewall Issues

- IP source and destination address will not change on non-shimmed packets
- IP source and destination may change on packets with a layer 3 shim
- Packet filters may need to match on IP source, IP destination, layer 3 source ULID, Layer 3 destination ULID, protocol, and port numbers
- Packet filters may require additional logic to map TCP established sessions when ULID is inserted and IP source or IP destination changes
- Stateful firewalls may need to match on IP source, IP destination, layer 3 source ULID, Layer 3 destination ULID, protocol, and port numbers
- Stateful firewalls will require additional logic to map sessions established with non-shimmed packets that migrate to shimmed packets with possible changing IP source and IP destination addresses.

Consider Approaches, Operational Requirements, and Trade-offs

1. Destination host sends extra information to the source host choosing the locators

   1A. This could be full routing information

   1B. This could be only the needed parts of the routing table

   1C. This could be clues sent by the destination host in TLV like link preference metrics

2. An 8+8 type solution, where the end hosts choose a locator, and allow the transit routers to record locator set information (additional state). Allow routers to recognize a destination address as one out of the locator set, and replace it with a different destination address if there is a different locator which is better.

3. Let the routers reach into the locator set exchange and add additional information or modify the locator set exchange in some way.

4. Move the shim to be a router function

   4A. Router inserts the shim on behalf of the end hosts

   4B. Router inserts a network level shim in additional to a host level shim

Get Involved!

- Join the shim6 working group email list shim6-request@psg.com
- Read the shim6 working group email list archive

  http://psg.com/lists/shim6/
- Come to the IAB's IPv6 multi-homing BOF

  2:00PM Sierra B C D
- Provide feed back to the IPv6 WG at IETF
- Weigh in on RIR policies that support or deny de-aggregation of IPv6 addresses
- Do some research on how big the IPv6 routing table will get if we de-aggregate and when wide spread adoption will occur
- Work with your vendors to determine the cost and feasibility for them to provide you will have enough memory and processor capabilities to allow for de-aggregation

Questions

?