Inter-AS Traffic Engineering Case Studies as Requirements for IPv6 Multihoming Solutions

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

Goals

- Describe the various cases for BGP inter-AS traffic engineering
- Provide simple cases of common operational practices to IPv6 multihoming developers
- Clearly define operational requirements for multihoming solutions

IPv6 Multihoming Conflict

- Current IPv4 multihoming depends on carrying customer's BGP announcement in the global routing table
- Permitting deaggregation for IPv6
  - IPv6 with 128 bit address has huge potential to increase the global route table
  - Can be solved by adding large amounts of memory + CPU
- Current IETF recommendation and RIR policy require the announcement of a single aggregate.
- Current IETF drafts in the multi6 WG do not depend on deaggregation
  - Provide multiple IP addresses to multihomed end host
  - Depend on source host to choose which link to load traffic on for multihomed destinations.
  - Multihomed destination has no ability to traffic engineer links as currently used.

IETF Multihoming Requirements

There are various IETF RFCs and drafts on IPv6 multihoming requirements

- RFC3582

  Section 3.1.1 requires failover

  Sections 3.1.2 and 3.1.3 address the ability to dial traffic around

- draft-ietf-multi6-v4-multihoming-03.txt

  Section 4.1 addresses failover

  Section 4.2 addresses shifting traffic across shared links

  Section 4.3 requires the ability to shift traffic around performance degradation

  Section 4.4 addresses the ability to policy route

No clear IPv6 multihoming solution that supports the current BGP inter-AS traffic engineering functions

Operational Multihoming Cases
(Proposed IPv6 Multihoming Requirements)

There are three basic flavors of BGP multihoming

- Primary / backup
- Load sharing across all links
- Best path

Additional requirement to shift traffic between links with any of the three options, such as pushing traffic away from over utilized links

Complex combinations of the three cases

Case 1: Primary / Backup

Requires the ability to designate a link or set of links as the primary link to use for all traffic to one or more destination prefixes. Primary link should carry all traffic for the designated prefixes. The backup link should only carry traffic if the primary goes down.

Common reason for primary / backup configuration is one link is more expensive than the other

- Cost
- Latency
- Performance
- Bandwidth

Case 1: Primary / Backup – Implementation

If all links are to the same AS, inbound traffic manipulated by setting higher MED on backup link(s)

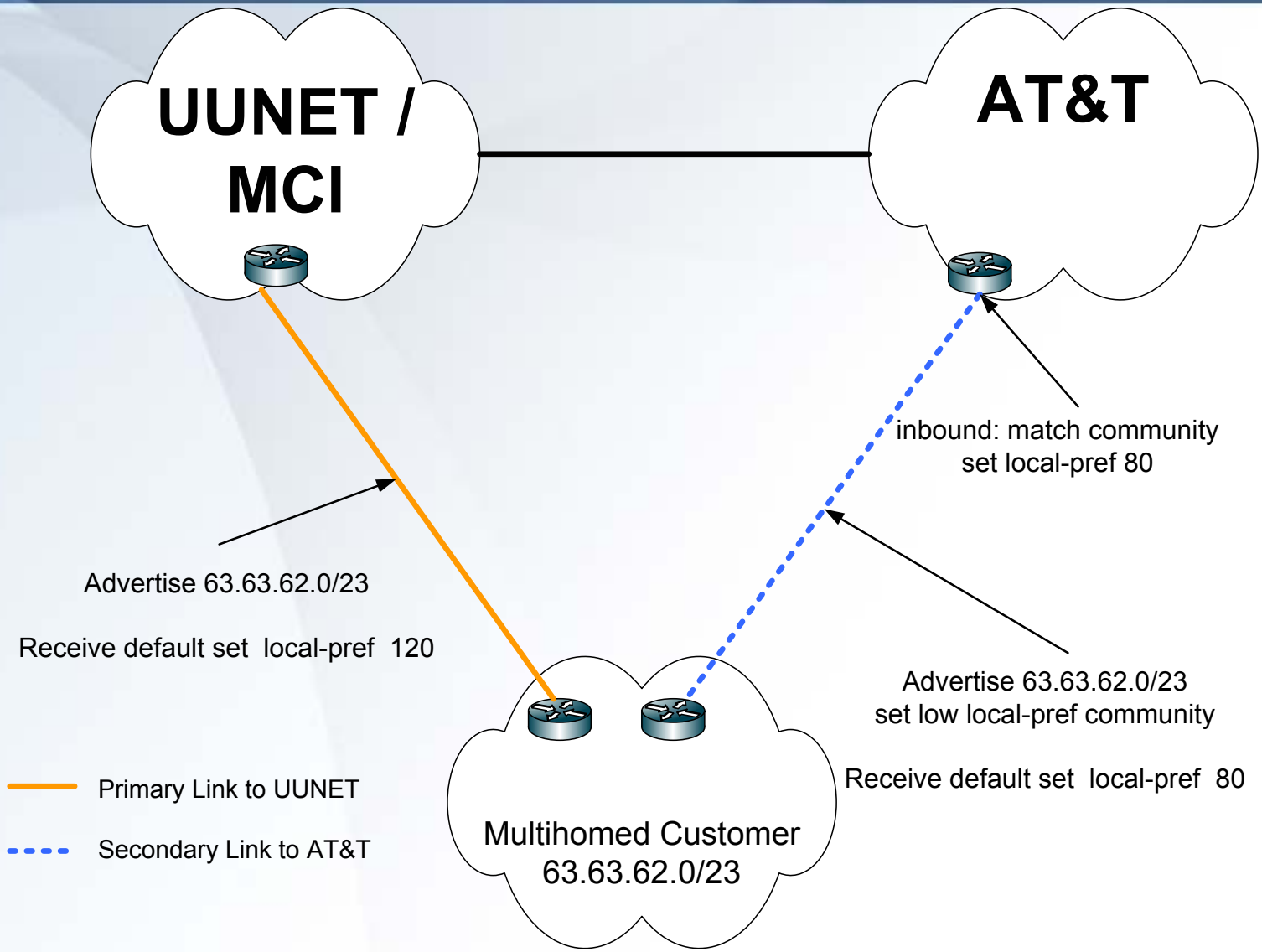
If links are to different ASes, inbound traffic manipulated by having the provider set low local-preference on backup link(s)

Outbound traffic manipulated by learning a default route and setting higher MED or lower local-preference on the backup link. (can set lower MED or higher local-preference on primary link)

Outbound traffic can be manipulated by weighted static default routes.

Can be configured to have multiple level of backup links (secondary, tertiary, etc..)

Case 1: Primary / Backup



Case 2: Load Sharing

Requires loading traffic on all links. The goal is to load the links as evenly as possible without negative impact on traffic flows.

Common reason for load sharing is to squeeze as much bandwidth out of multiple links as possible. This is often the case where larger links are cost prohibitive such as for small companies or locations where circuit cost is high.

Case 2: Load Sharing – Implementation

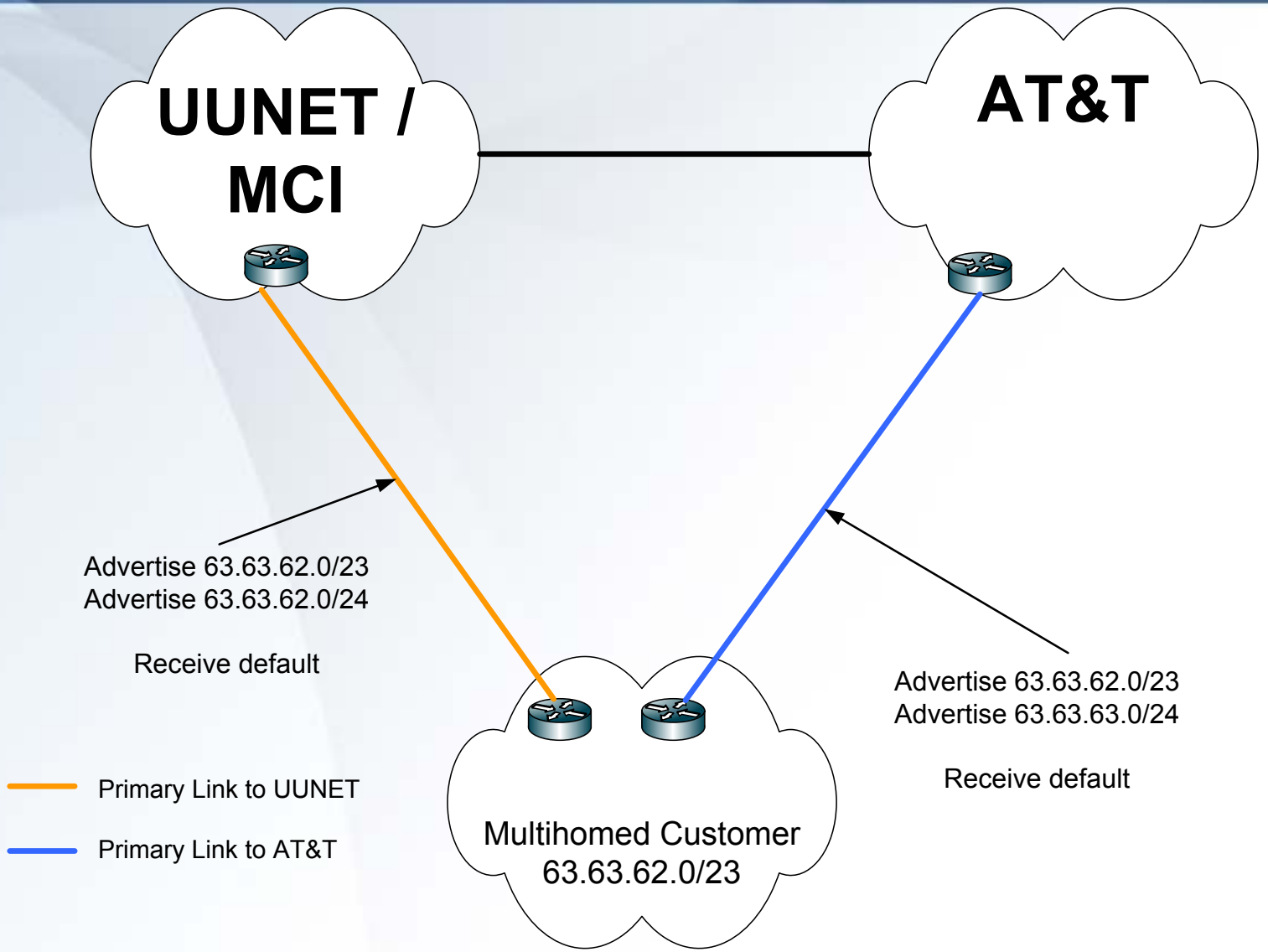
Inbound traffic manipulated by dividing IP space and making particular more specific route announcements across different links in addition to the aggregate.

Outbound traffic manipulated by depreferencing certain inbound route announcements by setting a MED value or local-pref inbound.

Outbound traffic manipulated by adjusting IGP metrics to make certain hosts closer to certain exit points.

Outbound traffic can be manipulated by equal cost static default routes.

Case 2: Load Sharing



Case 3: Best Path

Requires the ability to use a non-random “best” path (For some definition of best).

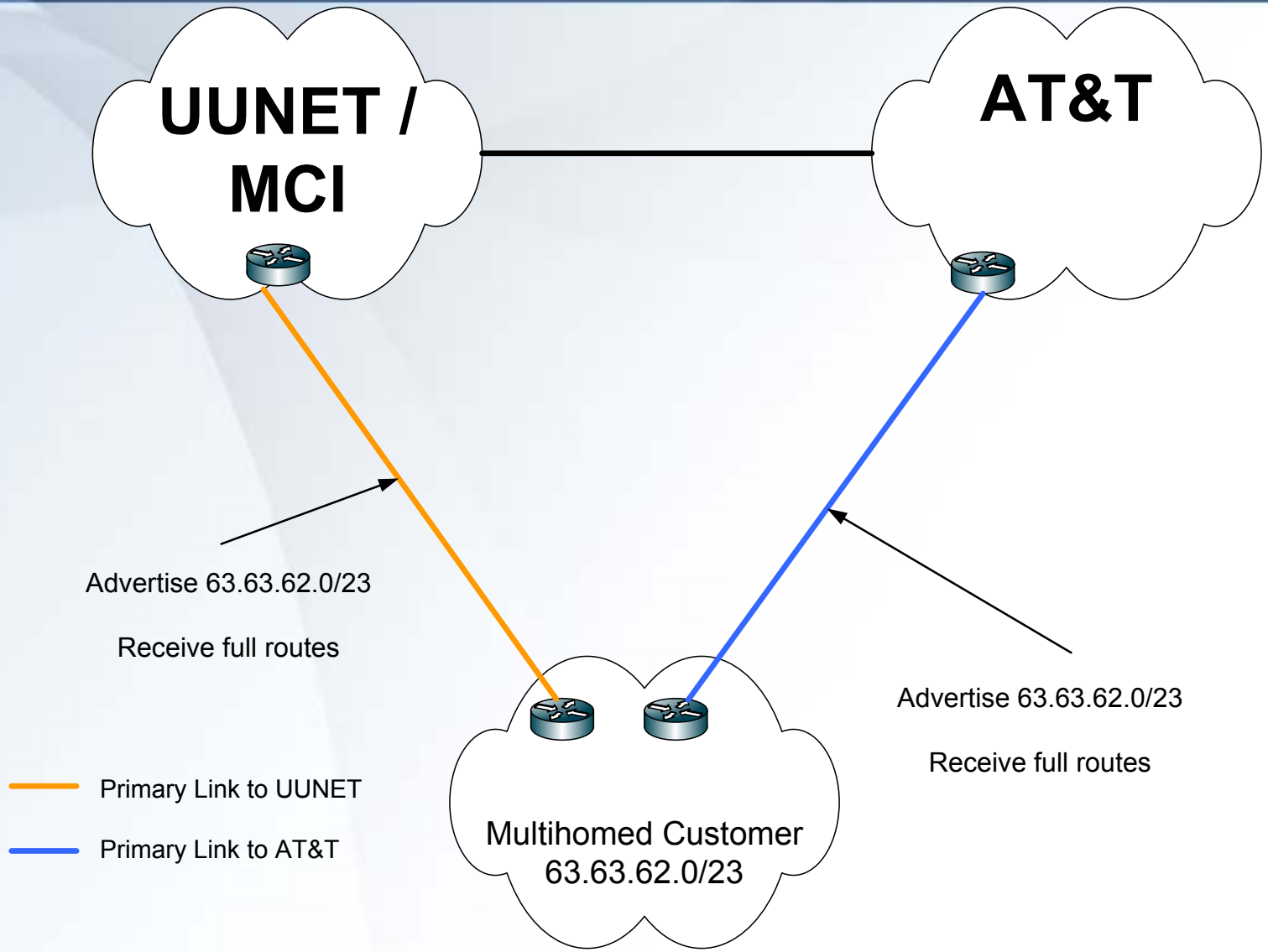
Current best path is based on routing information based on BGP path selection algorithm (LocalPref, AS-path, origin code, MED, eBGP over iBGP, IGP distance, RR cluster length, RID, lowest neighbor)

Best path approximates “shortest” path to host

Inbound traffic manipulated by source BGP table best path selection

Outbound traffic manipulated by learning full BGP routes from all upstream ISPs

Case 3: Best Path



Dialing Traffic

BGP lacks a congestion control mechanism (I'm not suggesting it be added).

- To avoid congestion operators will shift traffic away from over-utilized outbound links, and attempt to draw traffic to under-utilized inbound links. This is a manual process to avoid congestion.

Case 1 & 2: Drawing traffic to underutilized / backup link – traffic dialing

In case 1, the primary link is becoming over-utilized. The goal is to migrate some, but not all, traffic to the backup link. The end result is the primary link being nearly full and the backup link carrying the spillover.

In case 2, one of multiple links is over utilized. The goal is to migrate small amounts of traffic to the under-utilized links. The end result is that no link is over utilized and all of the links are roughly equally full.

Inbound traffic is drawn to the backup link or underutilized link(s) by advertising some, but not all, destinations as more preferred across the links to be loaded. These destinations will be more preferred due to being more specific, having a higher local-preference, shorter AS-path, or lower MED value, etc.

Out bound traffic is pushed away from the over utilized link(s) by increasing the IGP distance to the over utilized link(s) for some sources.

Outbound traffic is pushed away from the over utilized link(s) by depreferencing some inbound BGP paths associated with the over utilized link(s) .

Case 3: Best path – traffic dialing

In this case the goal is to use the best path, but to bias some small amount of traffic away from the best path to the second best path. The traffic which migrates should be the traffic which is closest to the second best path.

Inbound traffic is drawn to the second best path by slightly distancing the best path. This distancing is achieved by AS prepending.

Additional inbound traffic may be fine tuned by advertising some more specifics across the under utilized links.

Out bound traffic is pushed away from the over utilized link(s) by increasing the IGP distance to the over-utilized link for some sources.

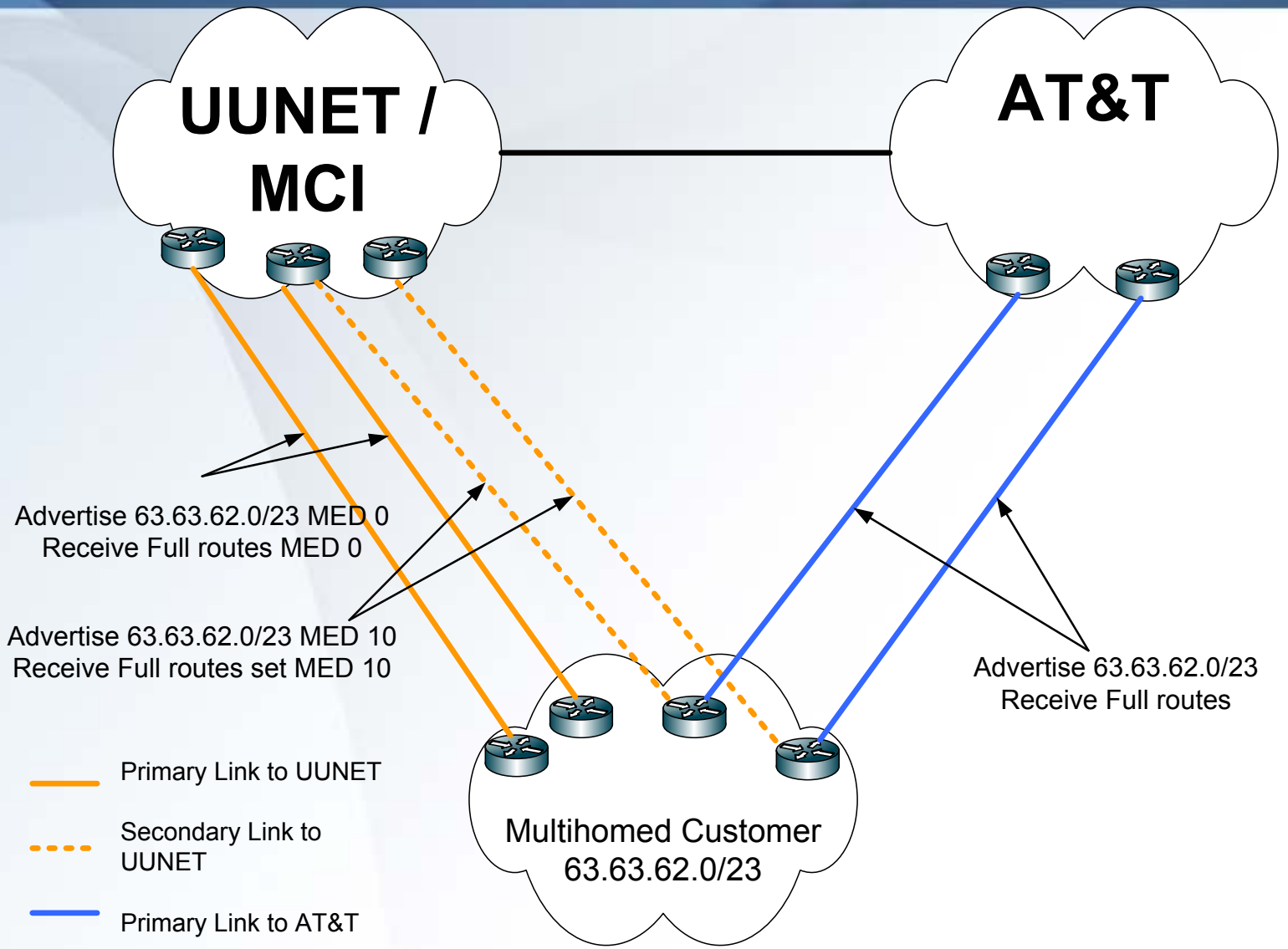
Outbound traffic is pushed away from the over utilized link(s) by depreferencing some inbound BGP paths associated with the over-utilized link.

Complex Combination of Cases

Inter-AS BGP traffic engineering can be a combination of the 3 cases and further refined by dialing the traffic up or down.

You could imagine a customer with links to two ISPs say UUNET / MCI and AT&T, where best path is used inbound and outbound between the customer and both upstream ISPs. Also imagine the connection to UUNET / MCI consists of a pair of primary links and a pair of backup links while the links to AT&T consist of a pair of primary links.

Complex Combination



IETF IPv6 Multihoming Solutions

- Many solutions based on a similar approach
  - LIN6 -- Location Independent Network Architecture
  - HIP – Host Identity Protocol
  - MAST – Multiple Address Service For Transport (MAST)
  - Shim6 – Multihoming L3 Shim Approach

- All IPv6 multihoming that do not utilize deaggregation focus on the separation of Locator ID and Upper Layer IDs
- **Locator ID** – the IP address associated with the host interface and is what routers use to route traffic. In IPv6 a unique Locator ID is required for each upstream ISP
- **Upper Layer ID** – the address used by upper layer protocols (e.g. TCP) to communicate
- In IPv4 the host IP address is both the Locator ID and the Upper Layer ID

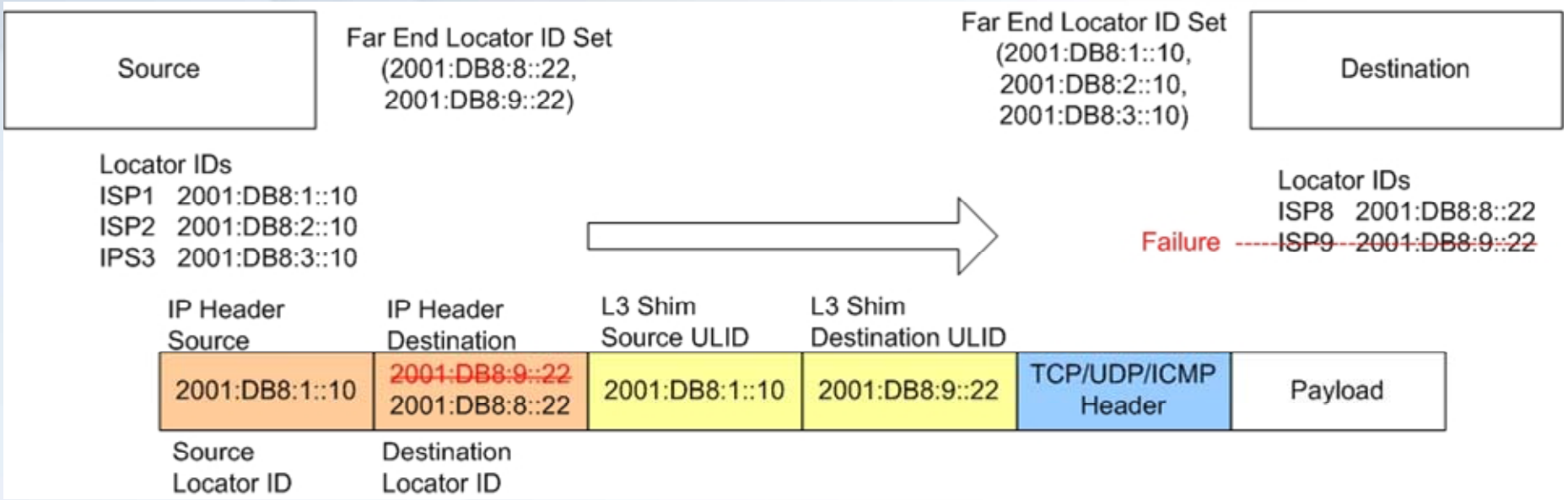
- The goal is for applications to communicate with each other via the Upper Layer ID, while the Locator ID may change to route traffic without impacting the session

L3 Shim Approach

- Current solution documented in draft-ietf-multi6-l3shim-00.txt
- Multhomed destinations and sources require one unique IP address (Locator ID) for each upstream ISP
- Multihomed 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

- AAAA DNS query provides a (possibly incomplete) set of Locator IDs
- Source chooses a Locator ID 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 multi6 capabilities and pass a complete set of Locator IDs. Current source and destination Locator IDs are used as source and destination Upper Layer IDs.
  - In the event of an outage source or destination can change source or destination Locator ID to any in the Locator ID set. (existing sessions continue to use unchanged Upper Layer ID)

L3 Shim Approach



Current IETF Solution

- All traffic engineering decisions are made by the source
- Sources and destinations have provisions for failover
- Traffic should load all links and failover only onto working links
- Failure detection through forwarding plane mechanisms

- Without more specific routes, all destinations of particular ISP can be distanced
- Without more specific routes, all destinations of a particular ISP are reachable or unreachable
- Destinations or intermediate routers have no ability to traffic engineer
  - No provision for destination to designate a source should use “best” path which approximates shortest path between source and destination. (need not be based on routing information)
  - No provision for destination to designate secondary paths

Conclusions

Current IETF solutions provide support for the following operational IPv4 BGP cases:

- Blind load sharing across all links
- Failover

The following are not currently supported

- Primary / Backup
- “Best” Path
- Dialing traffic (is currently a requirement, but is not supported)

Additional proposed IPv6 multihoming requirements

- Ability to designate paths as primary, secondary, tertiary, …
- Ability to instruct source to use “best” path
  - Best need not be based on routing information or BGP path selection algorithm
  - Best needs to approximate shortest path between source and destination

Conclusions

- Need to provide IETF operational requirements
- Need to carefully consider inter-AS traffic engineering with regard to your IPv6 deployment

Questions

?