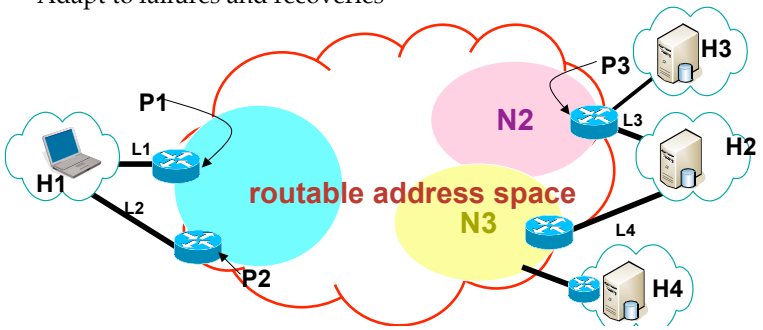
APT: A Practical Transit Mapping Service

Dan Jen, Michael Meisel,
Dan Massey, Lan Wang, Beichuan Zhang, Lixia Zhang

Routing Research Group
IETF69

What APT does

- Assumption
  - PI (or equivalent) prefixes of edge sites are not routed globally
  - Packets are tunneled from ITRs to ETRs
- APT
  - Provide PI prefixes to ETRs mapping
  - Adapt to failures and recoveries



Recall the questions

Q1: How to get mapping info

- Q1.1 How to inject the mapping info into the system
- Q1.2 Where to distribute, who holds the mapping info
- Q1.3 Where/who makes selection decision from multiple (Pi → Hi)

Q2: How to detect failure

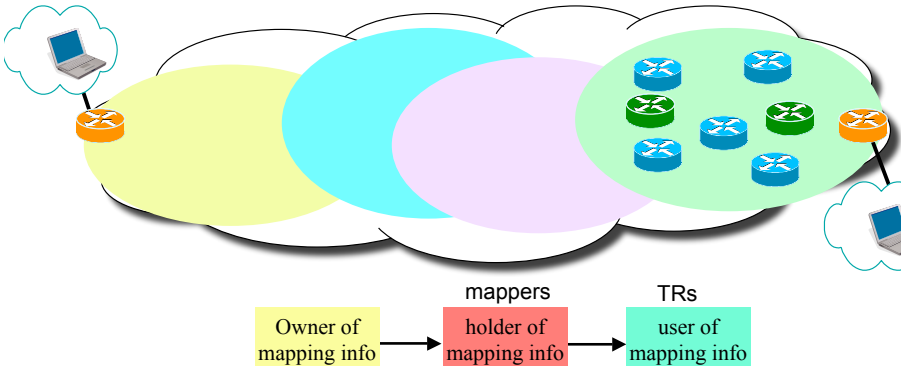
Q3: How to handle failure

- Q3.1: Which nodes to inform
- Q3.2: How to handle in-flight packets
- Q3.3: which party holds the temporary failure info, and how t promptly remove it when failure recovered?

Three Types of Nodes in Transit Space

(no change to edges!)

- Standard routers (routers, blue)
- Tunnel routers (TRs, oragen)
- Default mappers (mappers, green)

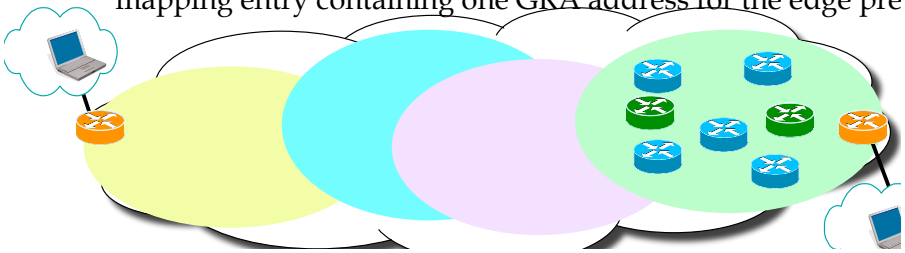


Default Mappers

- These are a new device
- Store *all* edge prefix to transit-space (GRA) addres mappings
- Each edge prefix maps to a non-empty set of GRA addresses
  - Each GRA address has a priority
  - Same priority? Use the shortest path
- At least one per AS
  - Use multiple for robustness, load sharing, shorter data path
  - Use anycase to reach nearest mapper
- Mappers tell ITRs which mapping entries to use

Tunnel Routers (TRs)

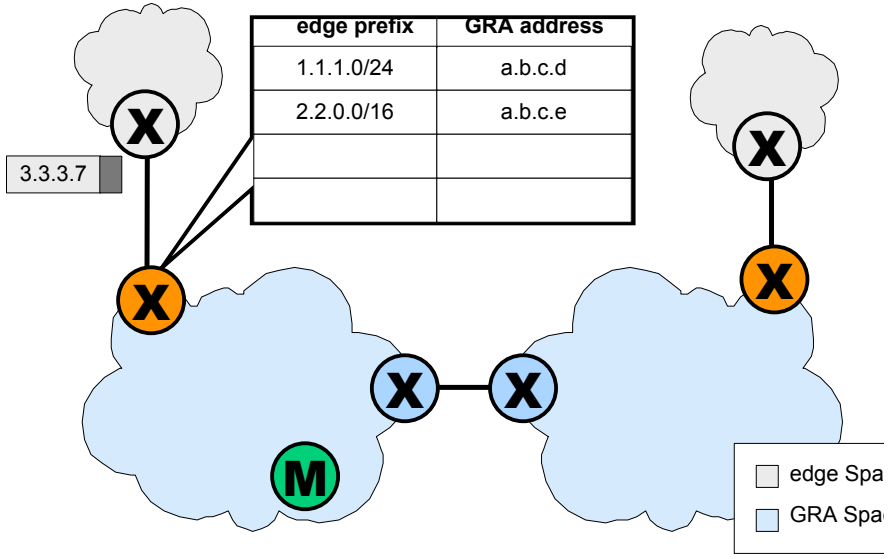
- Design goals for TRs: *minimal changes, stay simple*
  - Encapsulate outgoing packets (ITR mode)
  - Decapsulate incoming packets (ETR mode)
- Cache only mapping entries that are currently in us
  - No mapping entry? Tunnel packet to mapper's anycast addr
  - Mapper (1) forwards the packet, and (2)responds with a mapping entry containing one GRA address for the edge pre



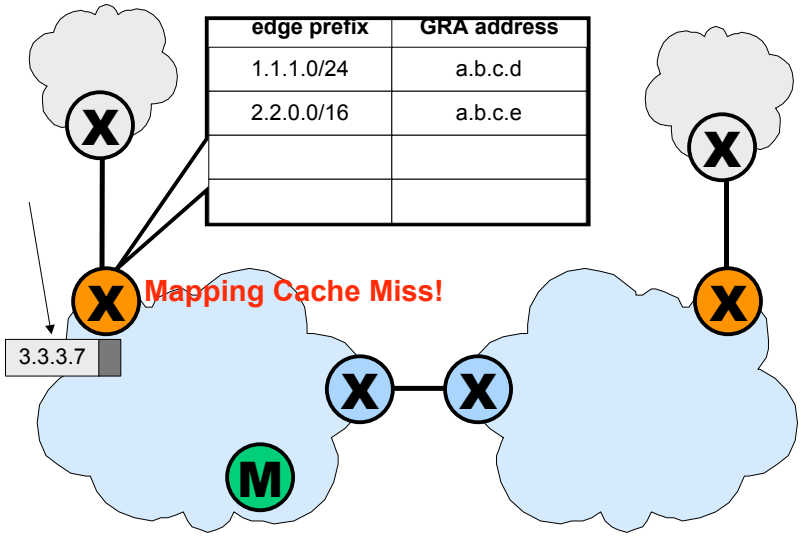
Standard Routers ("Routers")

- These are the rest of the existing routers
- (roughtly) no changes required to support APT

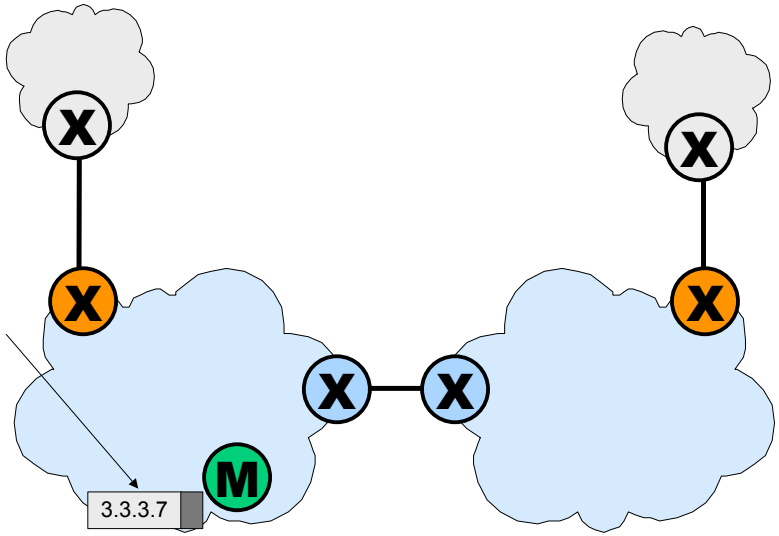
Default Mapping Example



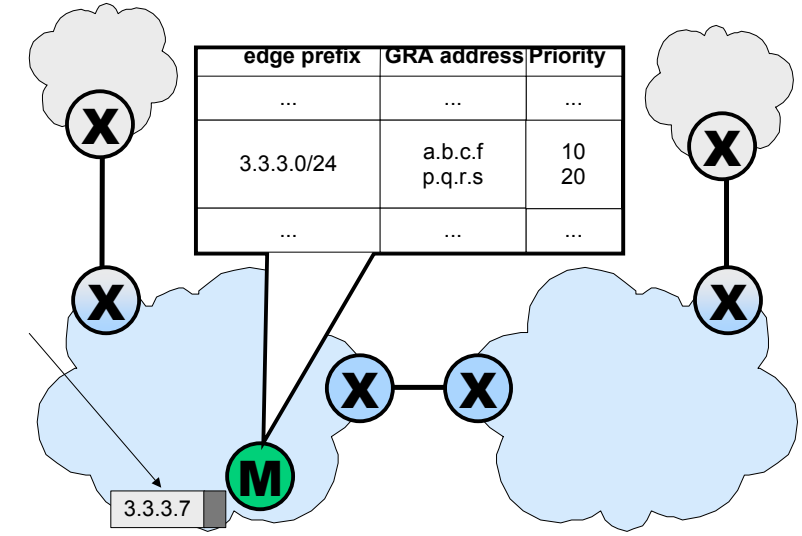
Mapping Not in Cache



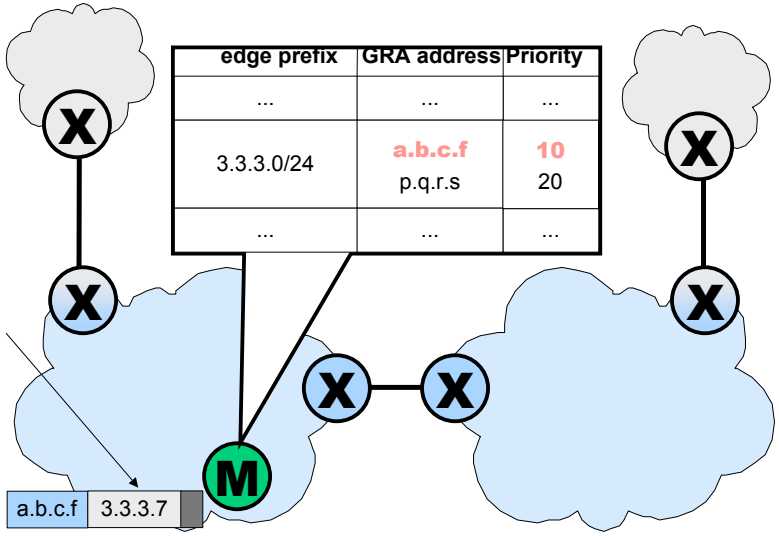
Use the Default Mapper



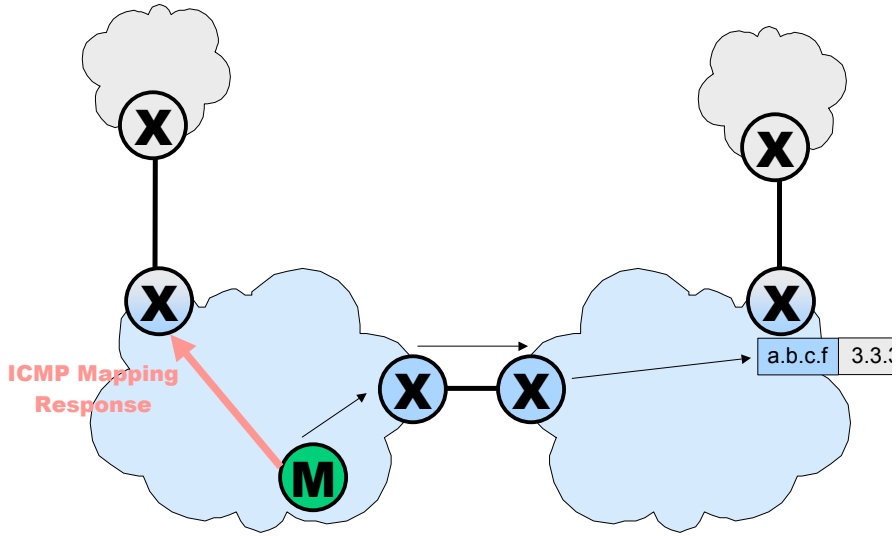
edge prefix is Multihomed



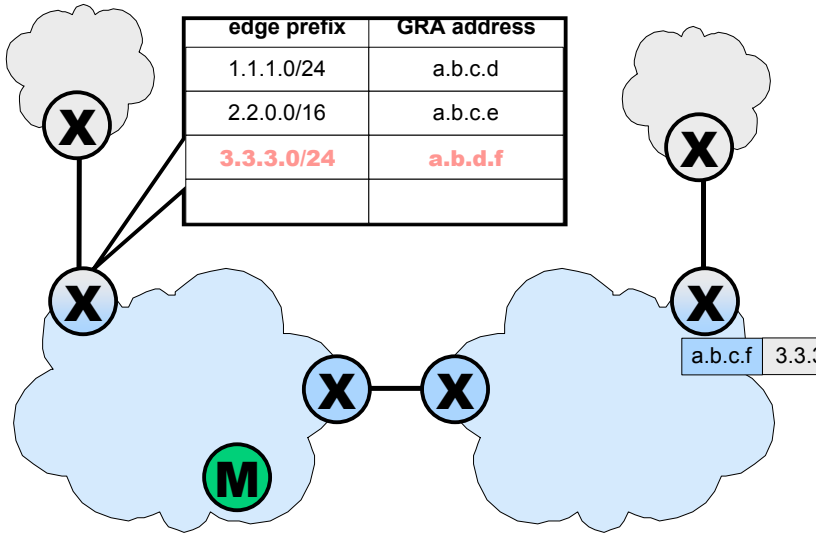
Default Mapper Selects a Mapping



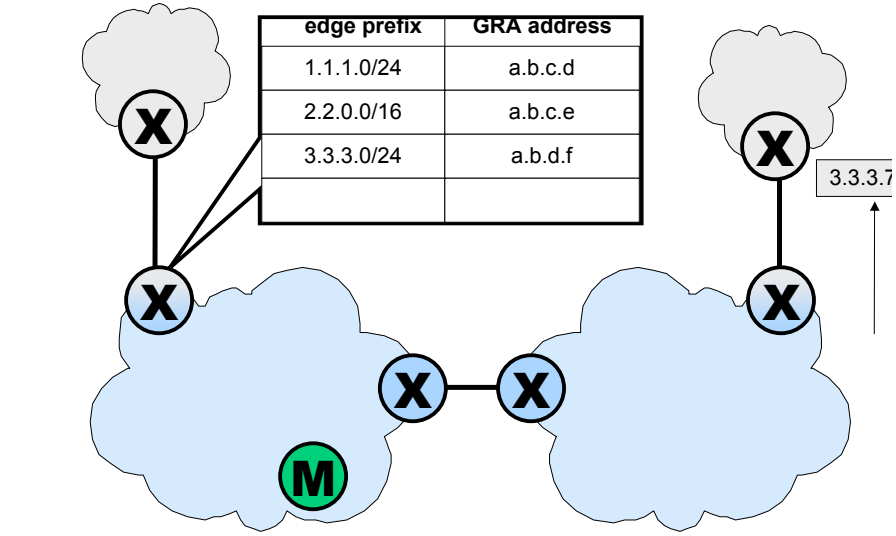
Default Mapper Responds with Mapping and Delivers Packet



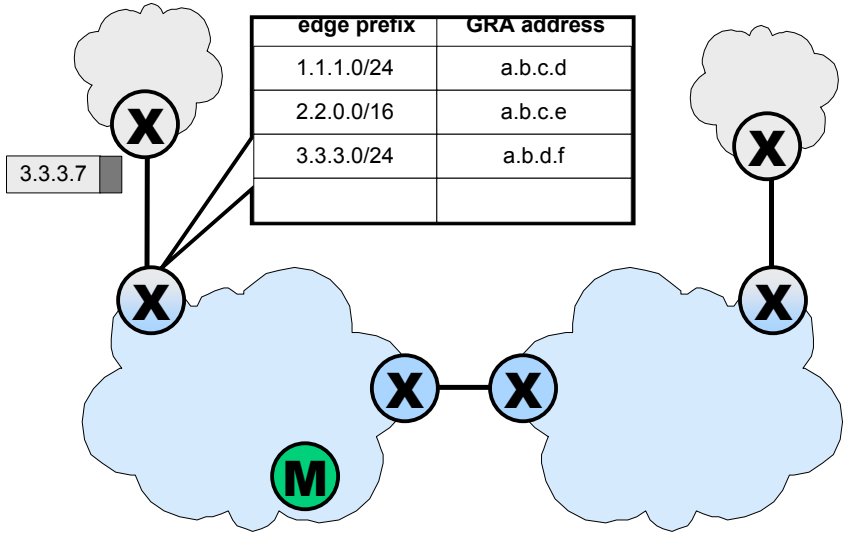
Mapping Added to Cache



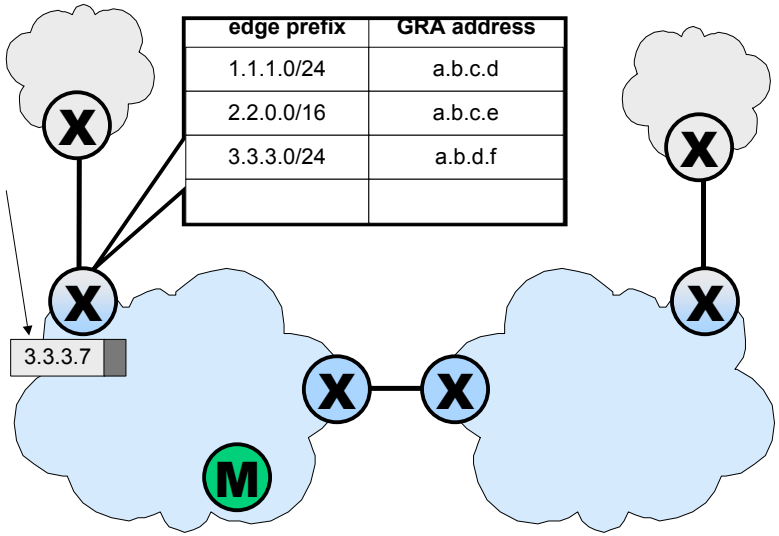
Packet Decapsulated and Delivered



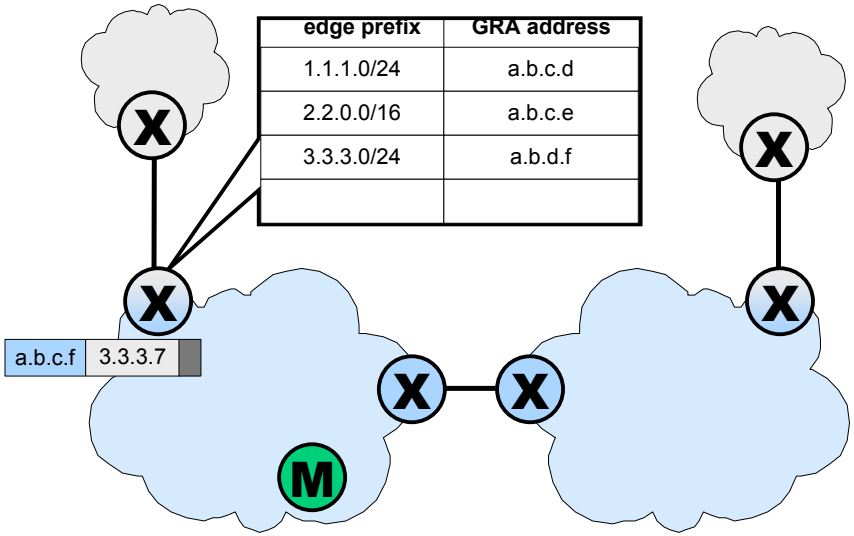
Next Packet



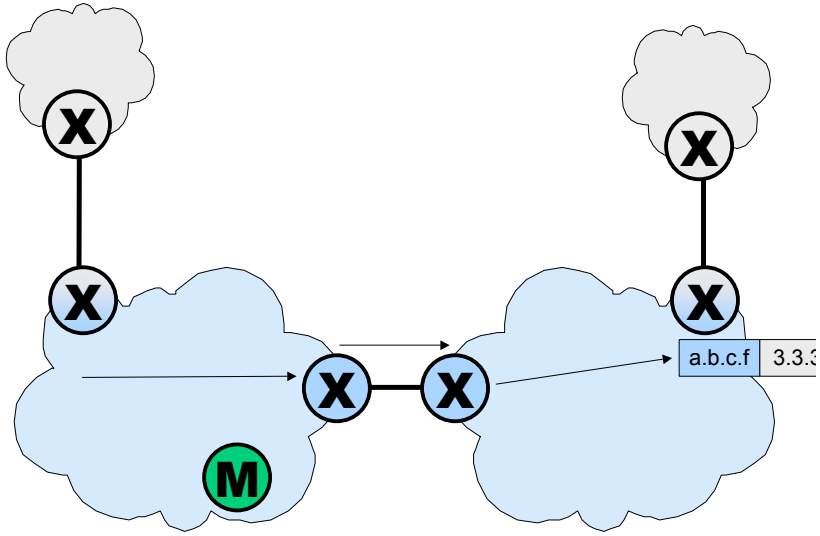
Mapping Already in Cache



Packet Encapsulated



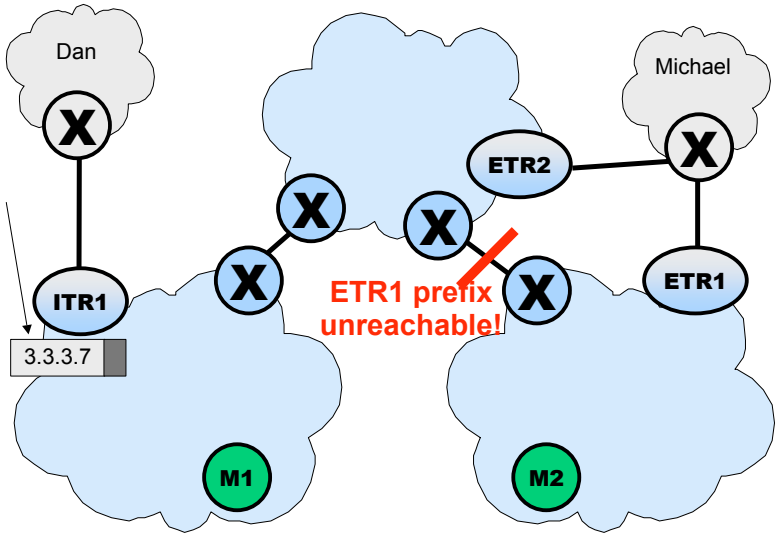
Packet Delivered



Handling Temporary Failures

- Three situations require failover to alternate ETR addresses
  1. A transit space prefix is unroutable via BGP
  2. A single transit space address becomes unreachable
  3. A link between an ETR and user space fails
- Basic approach:
  - Temporarily invadidate the corresponding mapping entrie
  - Do not change the mapping table
- Additional info at default mappers
  - Reverse mapping table: ETR to all PI-prefixes reachable th
  - Time Till Retry (TTR) for each mapping entry

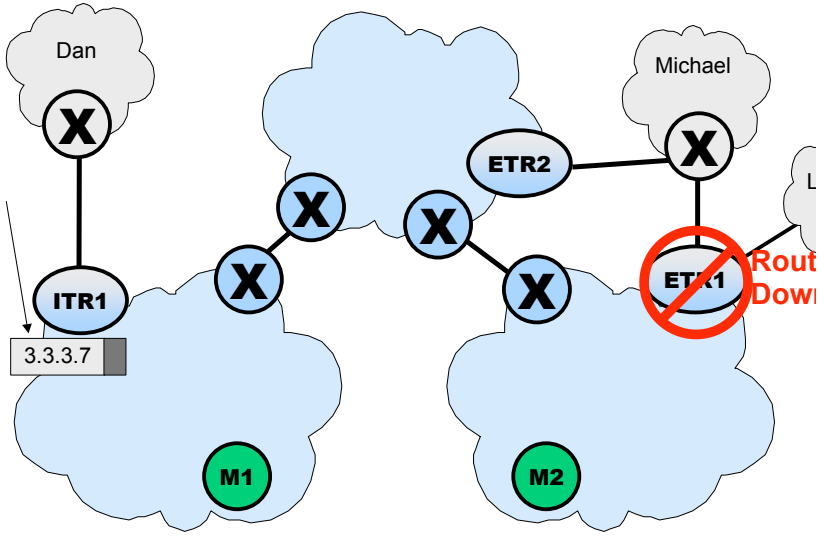
Situation 1: GRA Prefix Unroutable



Situation 1: GRA Prefix Unroutable

- ITRs forward packets with unroutable destination to their default mapper
- Default mappers use mapping priorities to pick a routable GRA destination address
  - And reply to ITR with a new mapping entry of a short TTL

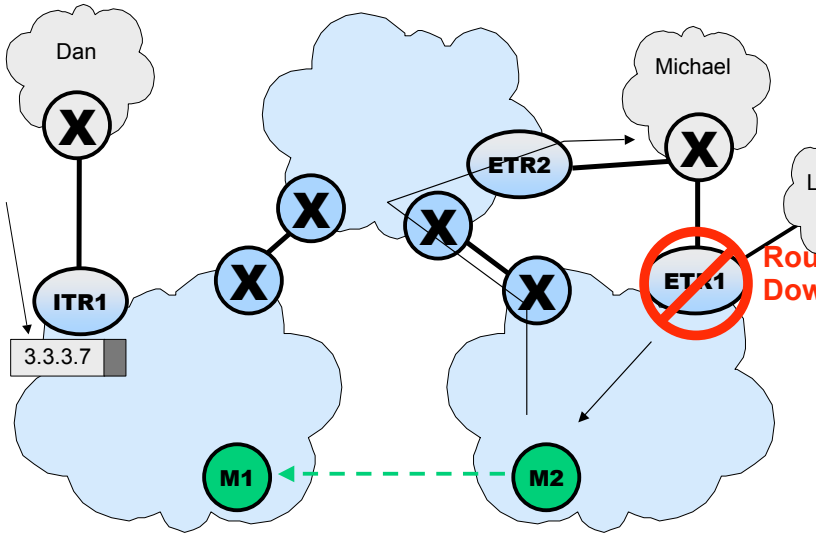
Situation 2 Example



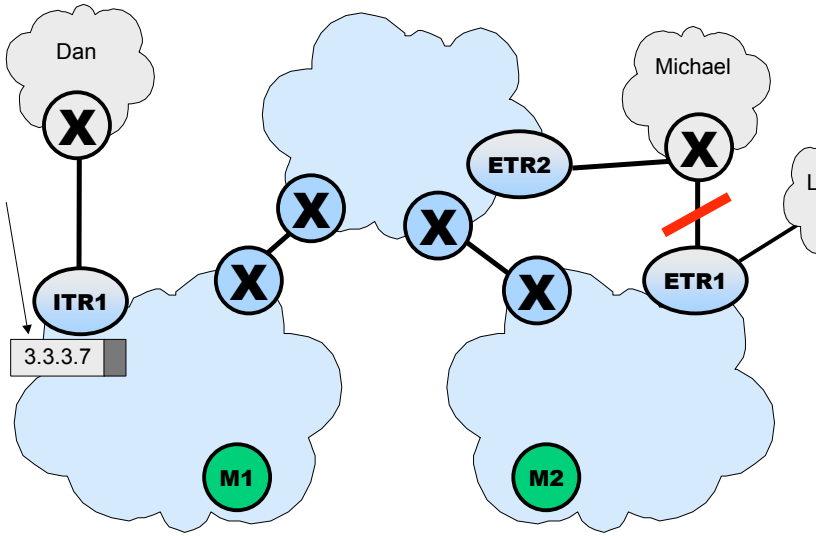
Situation 2: Single GRA address Failure

- Handling packets in-the-fly: minimizing losses
  - In the ETR domain: Forwards packets destined to ETR to its default mapper
  - At the ETR's mapper: Tries to find an alternate GRA destination address to tunnel packet to
- Informing the sender: 2 options
  1. The involved router sends an ICMP destination-unreachabl msg to sending ITR, which in turn forwards to its mapper
  2. (with a wellknown mapper address definition) ETR domair mapper sends the ICMP msg to ITR's mapper; the ITR map informs the ITR
- In either case: ITR's mapper temporarily avoids corresponding mapping entries
  - Set the TTR in the reverse mapping table

Situation 2 Example



Situation 3 Example



Situation 3: Border Link Failure

- Handling packets in-the-fly: minimize losses
  - At the ETR: Forwards the data packet to its default mappe
  - At the ETR's default mapper: Tries to find an alternate GR destination address to tunnel packet to
- Informing the sending AS: 2 options
  1. ETR sends an ICMP Border Link Failure msg to ITR
  2. ETR's mapper sends the ICMP msg to ITR's mapper; the mapper informs ITR
- In either case: ITR's mapper invalidates mapping entry by setting its TTR for the particular edge prefix mapping entry

Distributing Mappings Between ASes

- APT has two distinct parts
  - Data forwarding
  - Mapping info distribution to mappers
- The latter can take any new distribution protocol once we have one
  - e.g. NERD, or CONS
- The current option: APT floods mapping info by piggybacking on BGP announcements

Distributing Mappings Between ASes

- Define a new BGP transitive attribute
  - mapping entry: edge prefix to GRA address mapping
- An edge network sends signed mapping to all its provider
- A provider network floods their customers' mappings to other provider networks via BGP
  - this GRA address may not have any relation with the prefix being announced
- All APT nodes (ITRs and mappers) listen
  - Default mappers store all incoming mappings
  - ITRs just invalidate cache entries that match incoming mappings

In Defense of piggybacking on BGP

- Mapping updates far less problematic than BGP routing updates
  - It only matters where mapping messages go, not what path they take
  - Only require processing at APT nodes
  - No path exploration for mapping messages
- Eases incremental deployment

Security and Robustness

- Wins
  - Transit space is not directly addressable from user space
  - Mapping announcements are only accepted from configure BGP peers
- Issues
  - ICMP packets are unreliable and can be spoofed
  - Mappings can be misconfigured

Security and Robustness for ICMP Packets

- Mapping messages
  - Only used within an AS,
  - drop them at AS boundaries if any trying to cross borders
- Border Link Failure messages
  - Can only be sent by GRA routers
  - Signature field allows easy addition of cryptographic securi

Incremental Deployment

- The user address space will not be affected
- Some edge prefixes will simply not have mapping
  - Packets destined for unmapped addresses are sent via the current infrastructure
  - TRs keep negative cache entries

Regular Mapping Refresh

- Newly added default mappers will need to get the full mapping table
- Allows stale mappings to expire
- Each provider re-announces its customers' mappings on a regular basis
  - Daily? Weekly?
- New default mappers boostrapping from other mappers

(near) Future Work

- Finish an incremental deployment design
  - Borrow ideas from other work (e.g. IvIP)
- Understanding TR cache size using real-world dat
  - Help us get real data !!!
- Reliable key distribution/discovery
  - Edge network keys
  - Provider keys
- Securing ICMP msgs

Questions?

bgpng@cs.ucla.edu