Evolving towards Routing Scalability

Combining APT, Virtual Aggregation, and ideas from Paul Francis, Robert Raszuk, and the APT team

presented by Dan Jen

Our Mission: Control Routing Scalability

- FIB size
- RIB size
- Update rate
- Update cost

RRG proposals save us!

- Assuming universal host changes...
- ... and/or renumbering...
- ... and/or multiparty agreements...
- ... and/or a new 3rd party infrastructure...
- ... run by someone we trust...
- ... and before this all happens...
- ... may not get scalability benefits.

- “How can all this happen?” we ask.

No Revolution, Only Evolution

- Someone proposes to change Internet into S, a scalable architecture.
- Fine, but there must be a realistic story to **evolve** from today's Internet into S.
- What does this mean?

No Revolution, Only Evolution

- No mass coordination among parties to deploy
  - Each should deploy in his own time **for his own selfish scalability gains.**
- No charitable 3rd party infrastructure requirement
- No Renumbering
  - RRG voted on this
- **Scalability should come incrementally with incremental deployment of new game.**

APT designed for Evolvability

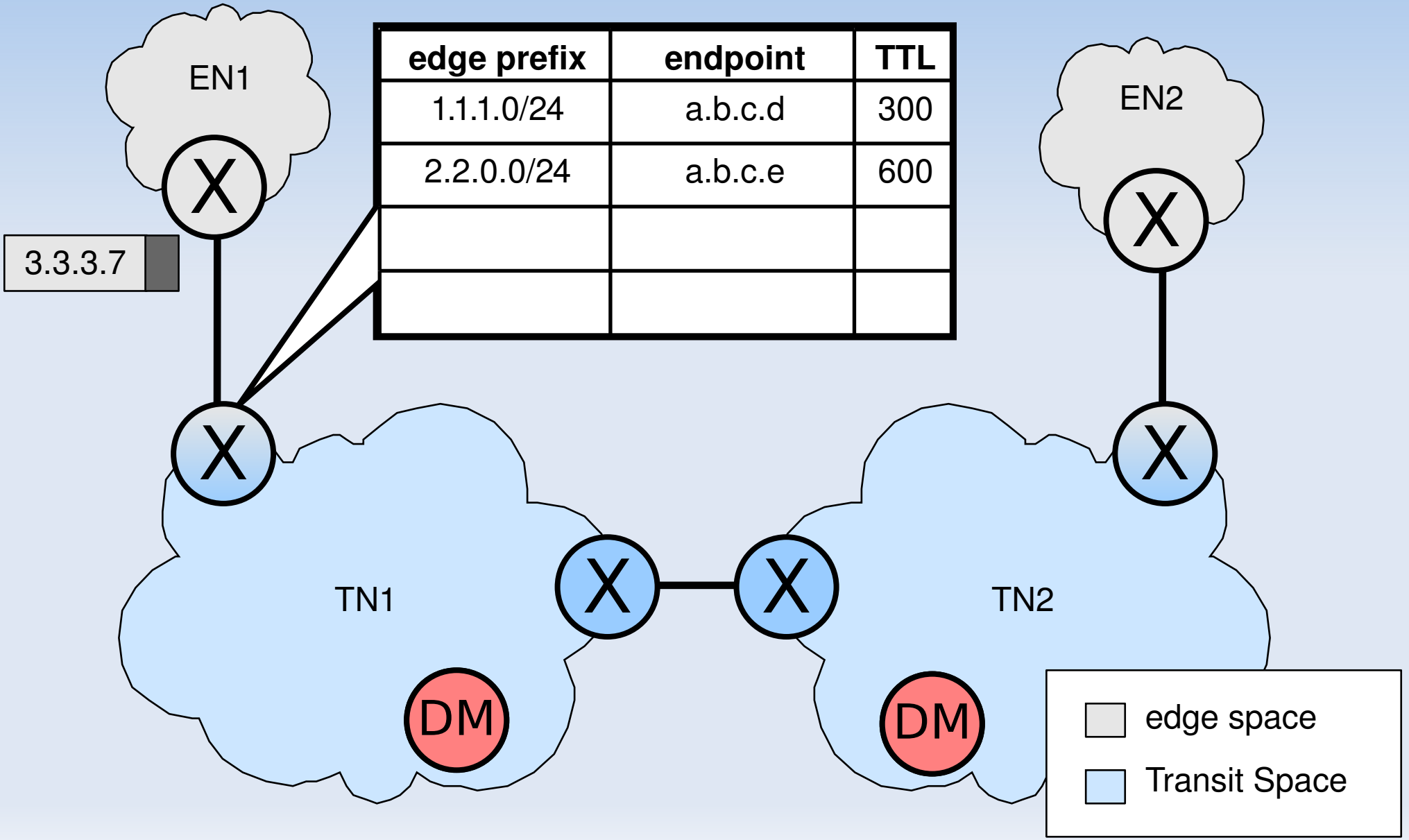
- ISPs can deploy unilaterally for their own scalability gains
- No 3rd party infrastructure required
- No Renumbering
  - Edge Sites just change mapping info to new providers

www.cs.ucla.edu/~meisel/apt-tech.pdf

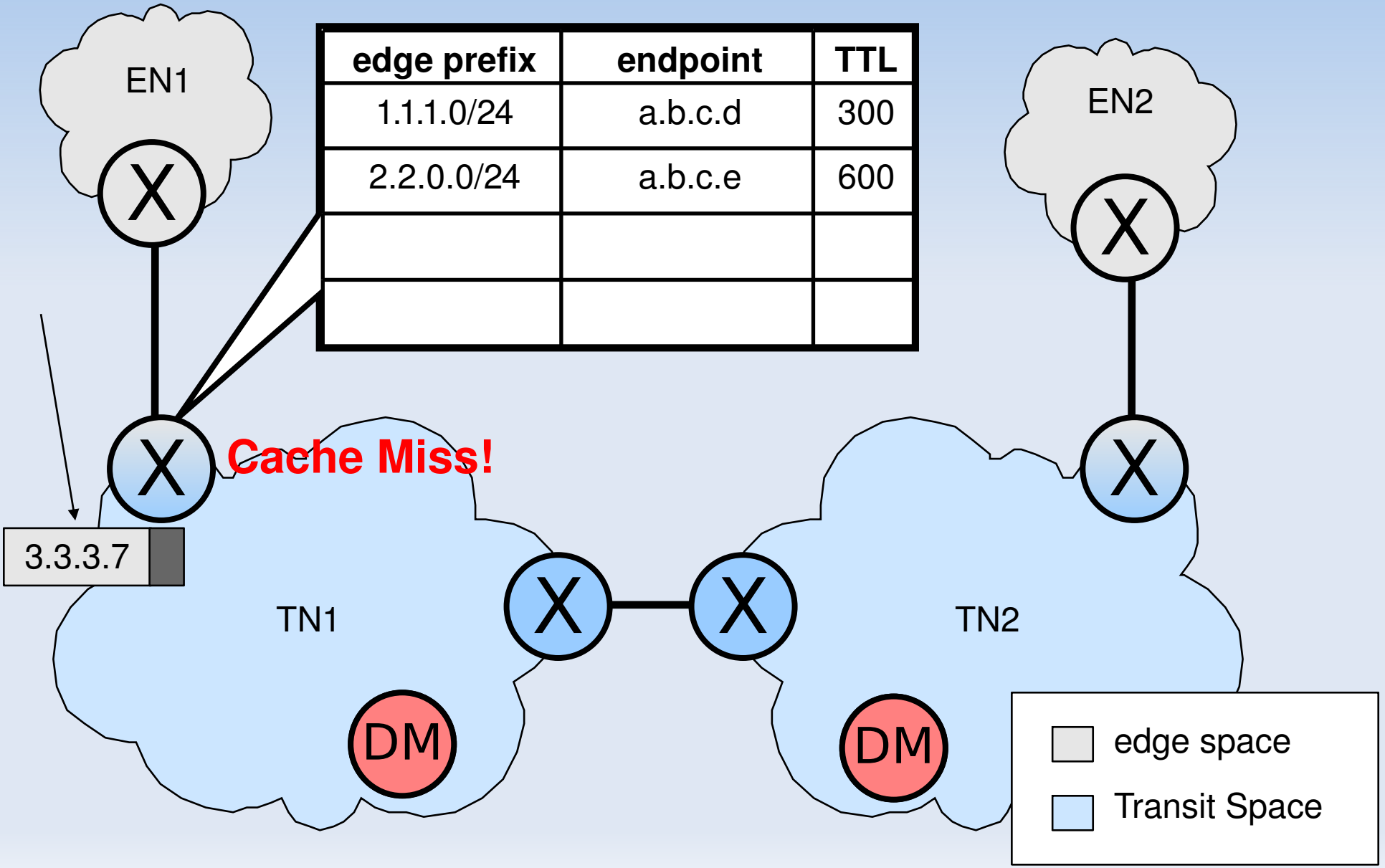
APT Features

- Default mappers in each ISP
  - Carries the full mapping/routing table
- Simple Encap/Decap border routers
  - Only carries mapping cache and ISP routes
  - Does not carry routes to edge prefixes
- Map & Encap to deliver to edge prefixes

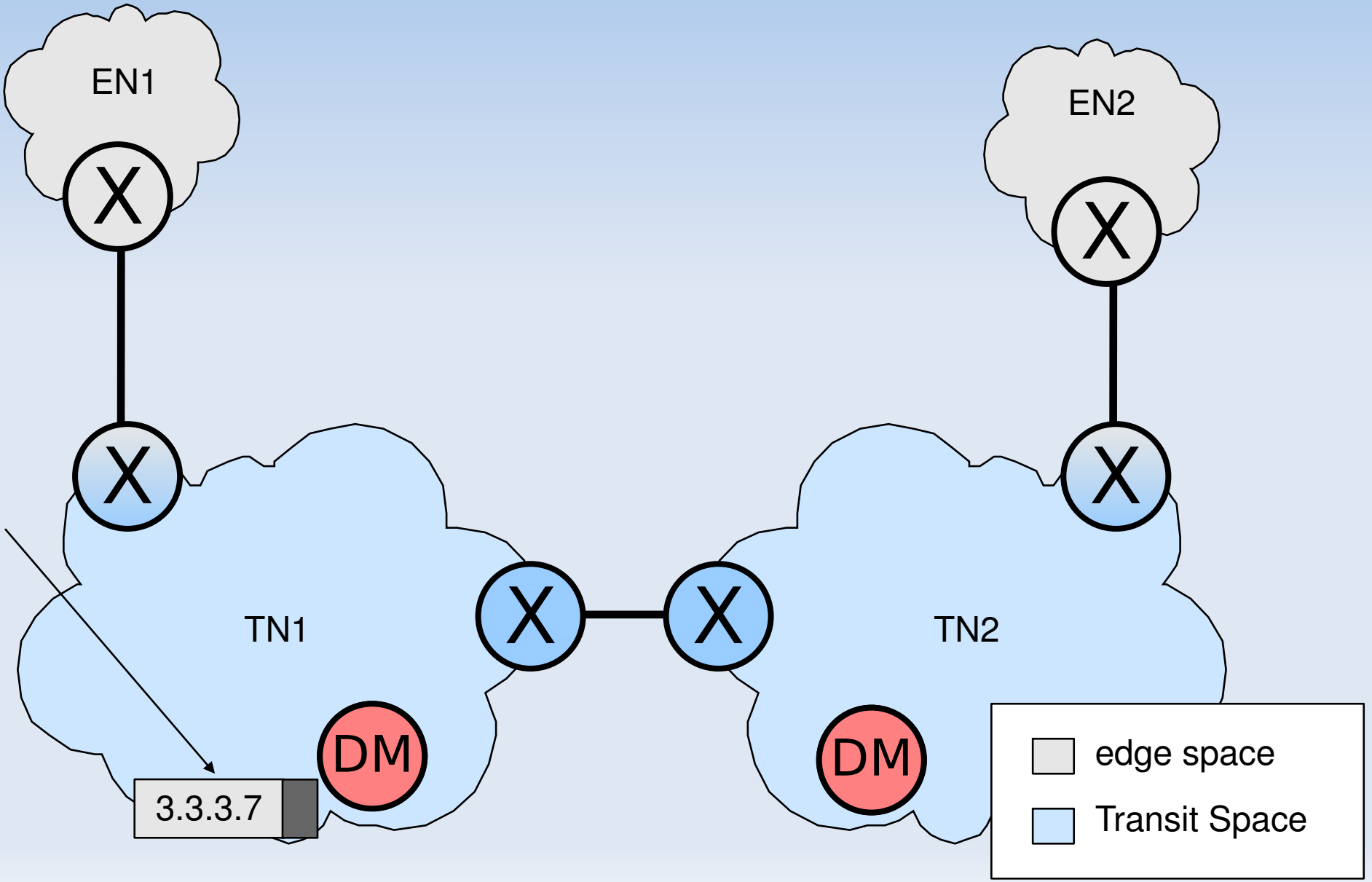
APT Review


EN1
3.3.3.7
edge prefix
endpoint
TTL
1.1.1.0/24
a.b.c.d
300
2.2.0.0/24
a.b.c.e
600
EN2
TN1
TN2
DM
DM
edge space
Transit Space

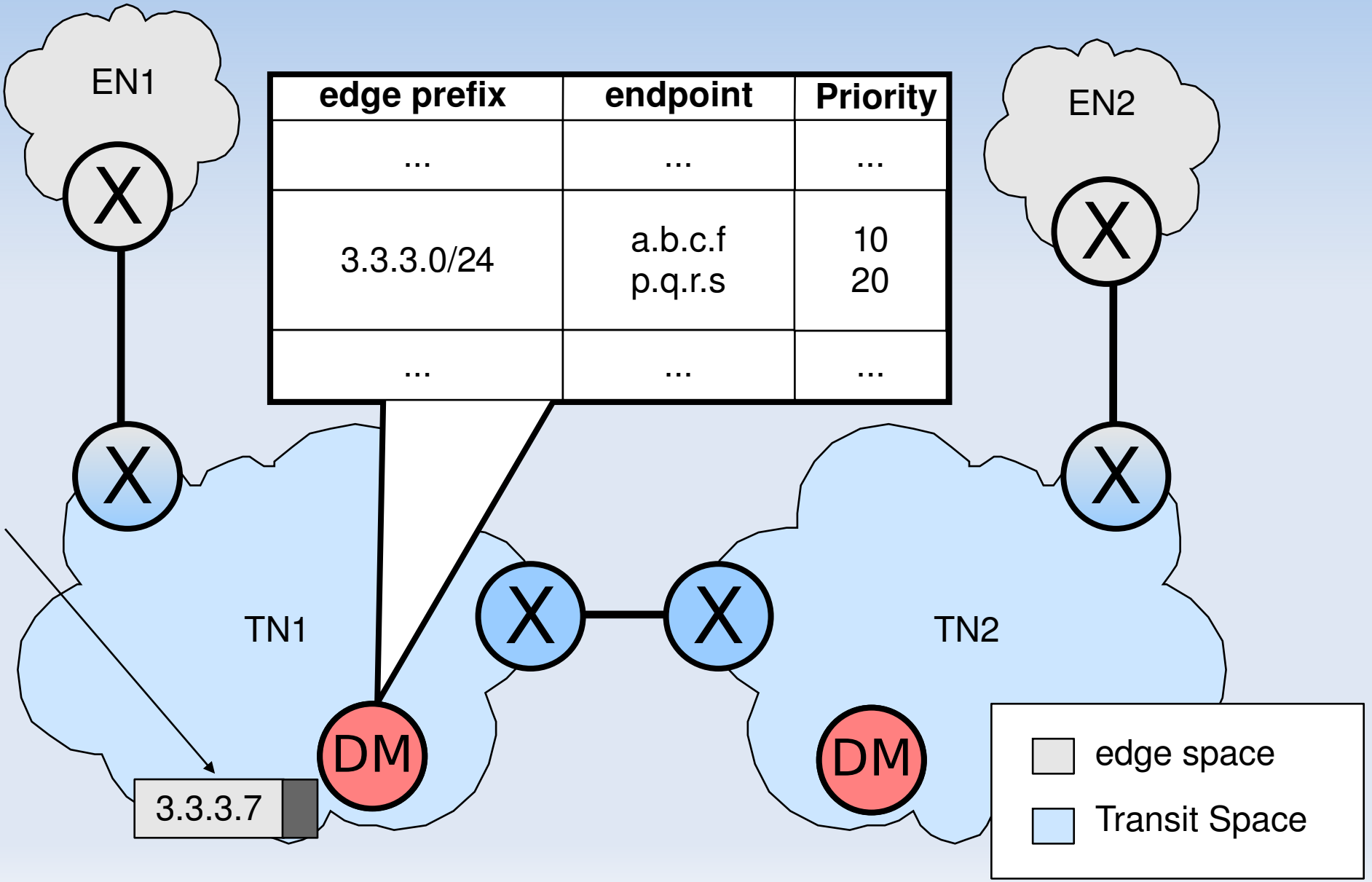
APT Review


EN1
edge prefix
endpoint
TTL
1.1.1.0/24
a.b.c.d
300
2.2.0.0/24
a.b.c.e
600
EN2
Cache Miss!
3.3.3.7
TN1
TN2
DM
DM
edge space
Transit Space

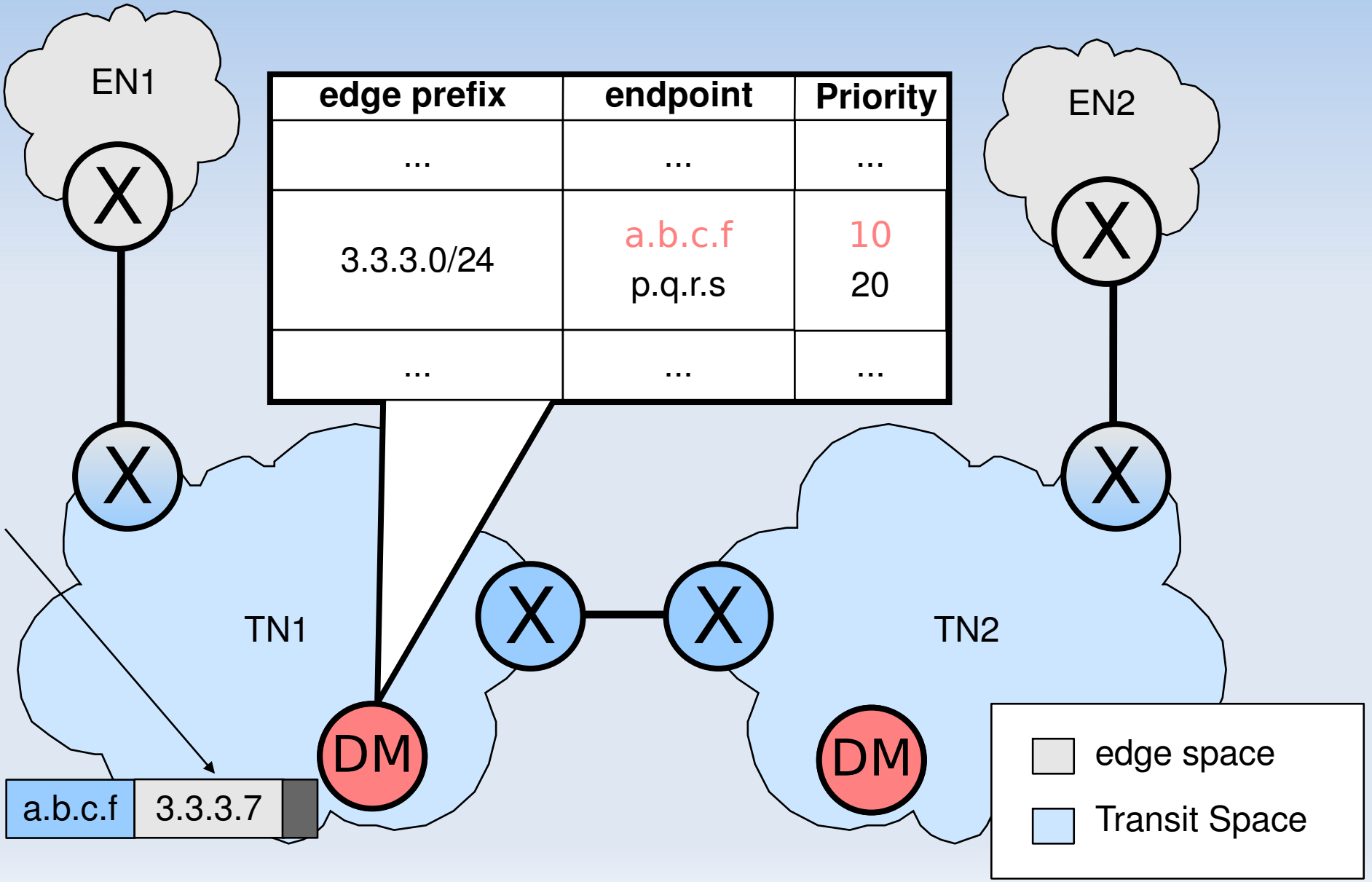
APT Review



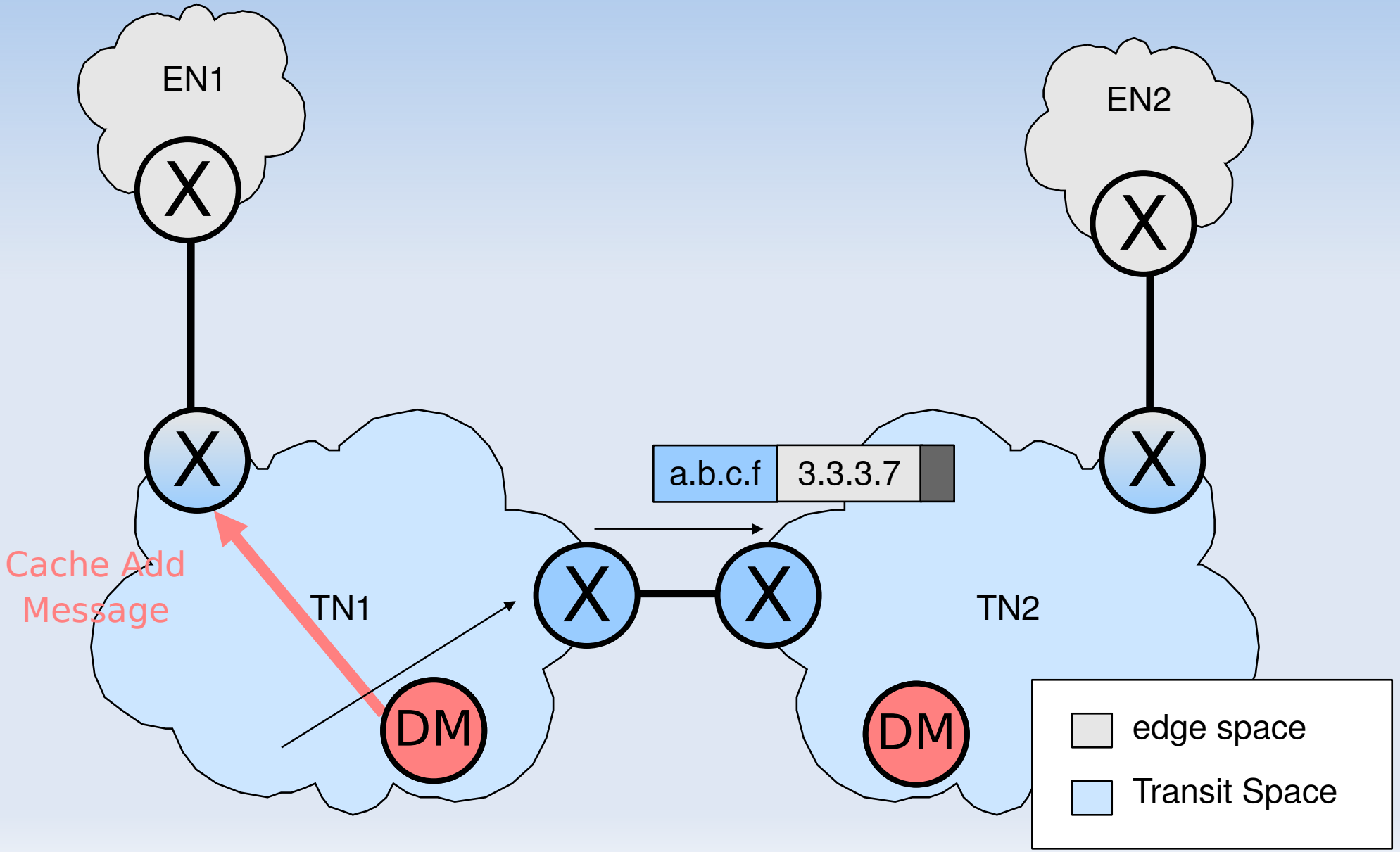
APT Review


EN1
EN2
edge prefix
endpoint
Priority
...
...
...
3.3.3.0/24
a.b.c.f
p.q.r.s
10
20
...
...
...
X
X
X
X
X
X
TN1
TN2
DM
DM
3.3.3.7
edge space
Transit Space

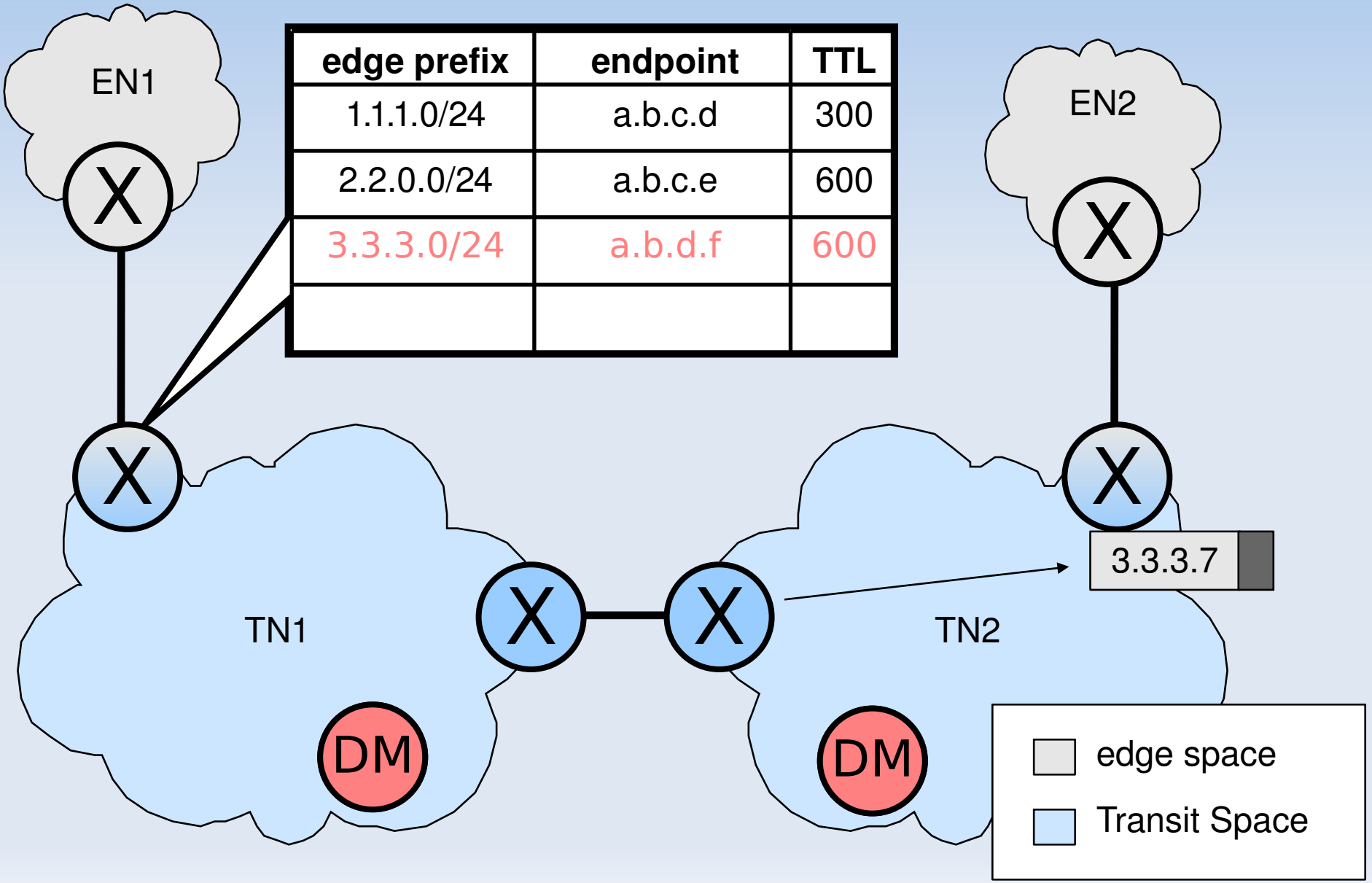
APT Review


EN1
EN2
edge prefix
endpoint
Priority
...
...
...
3.3.3.0/24
a.b.c.f
p.q.r.s
10
20
...
...
...
X
X
X
X
X
X
TN1
TN2
DM
DM
a.b.c.f
3.3.3.7
edge space
Transit Space

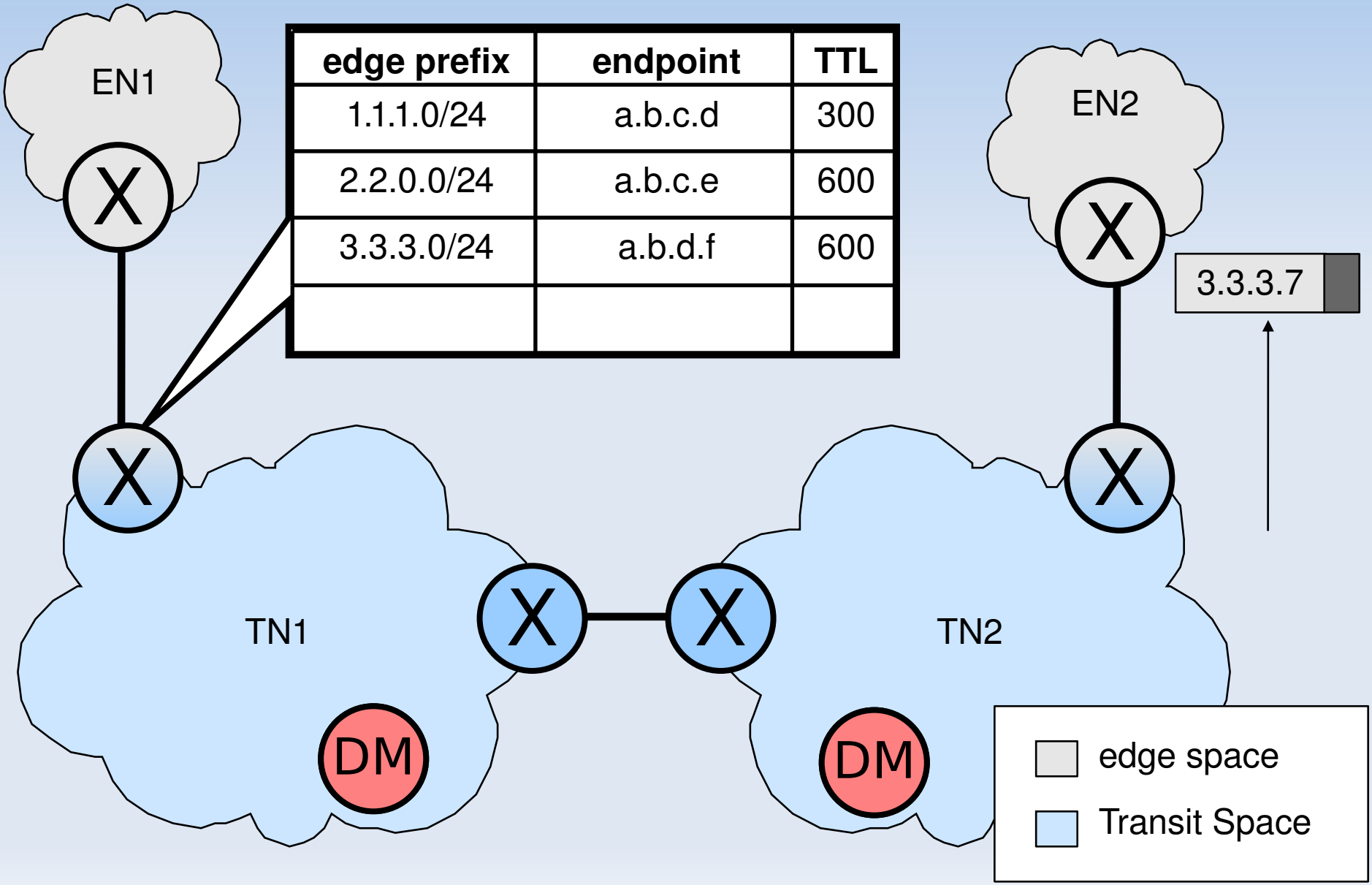
APT Review


EN1
EN2
X
X
X
X
X
X
a.b.c.f
3.3.3.7
Cache Add
Message
TN1
TN2
DM
DM
edge space
Transit Space

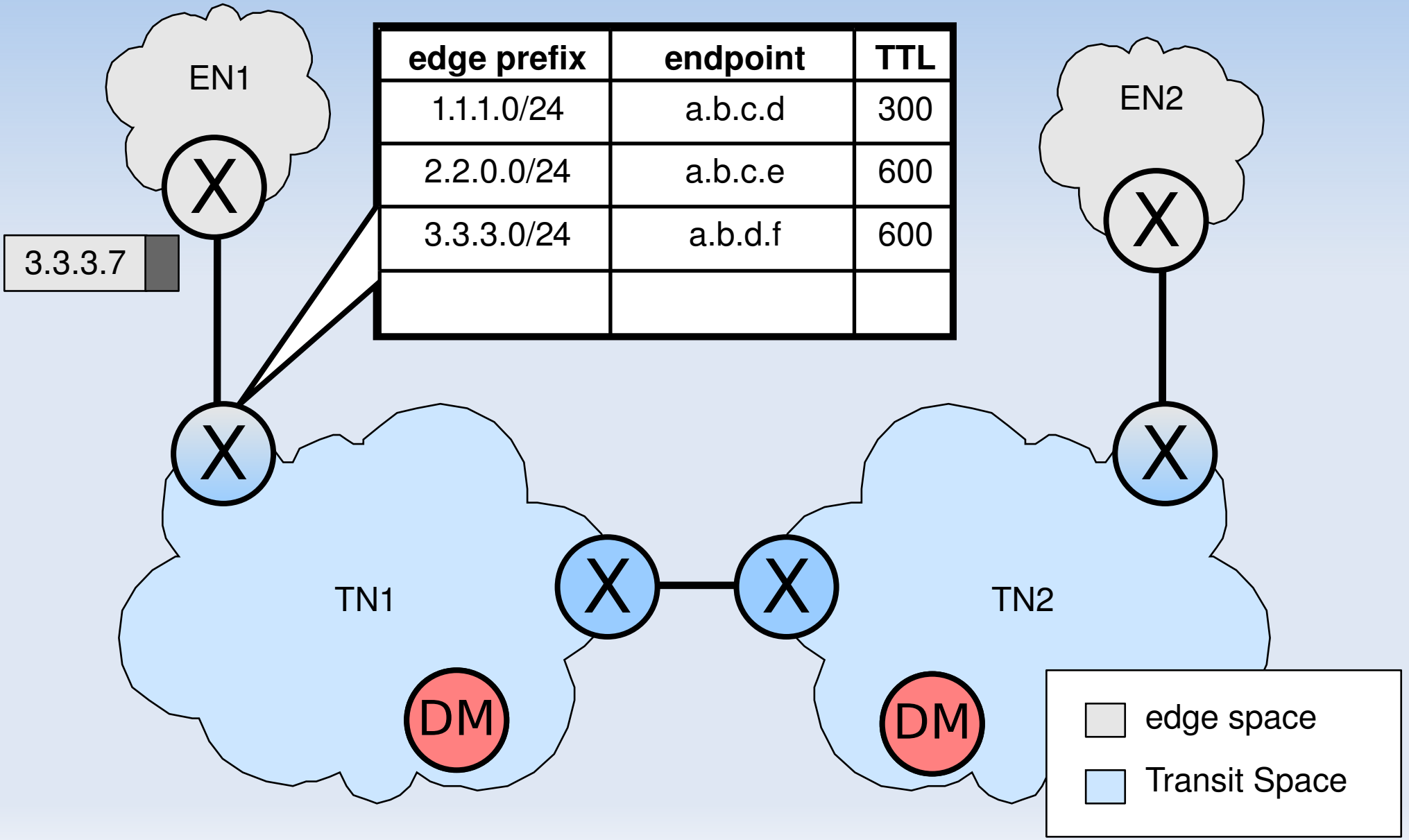
APT Review


EN1
EN2
edge prefix
endpoint
TTL
1.1.1.0/24
a.b.c.d
300
2.2.0.0/24
a.b.c.e
600
3.3.3.0/24
a.b.d.f
600
3.3.3.7
TN1
TN2
DM
DM
edge space
Transit Space

APT Review

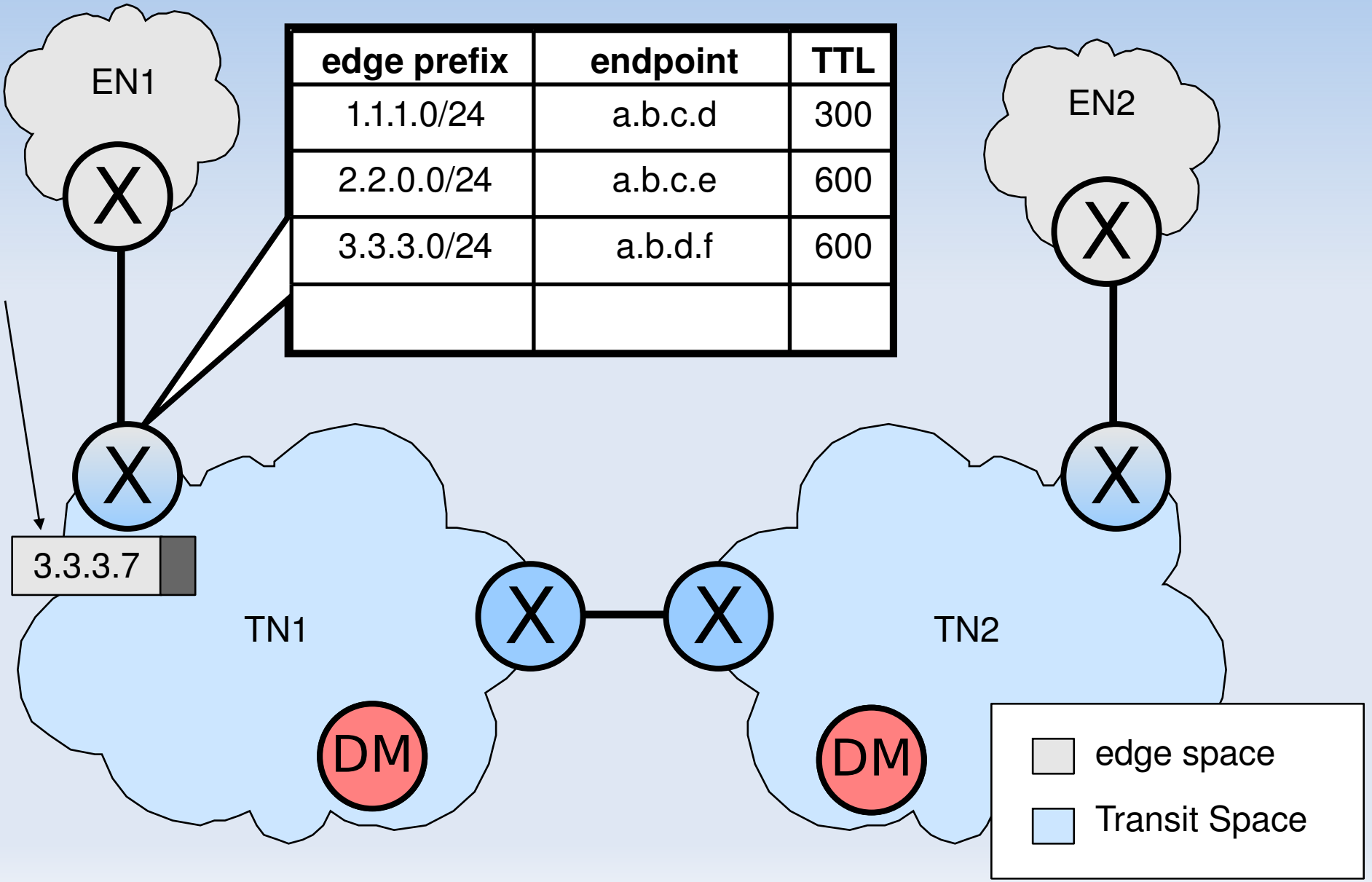

EN1
EN2
edge prefix
endpoint
TTL
1.1.1.0/24
a.b.c.d
300
2.2.0.0/24
a.b.c.e
600
3.3.3.0/24
a.b.d.f
600
3.3.3.7
TN1
TN2
DM
DM
edge space
Transit Space

APT Review



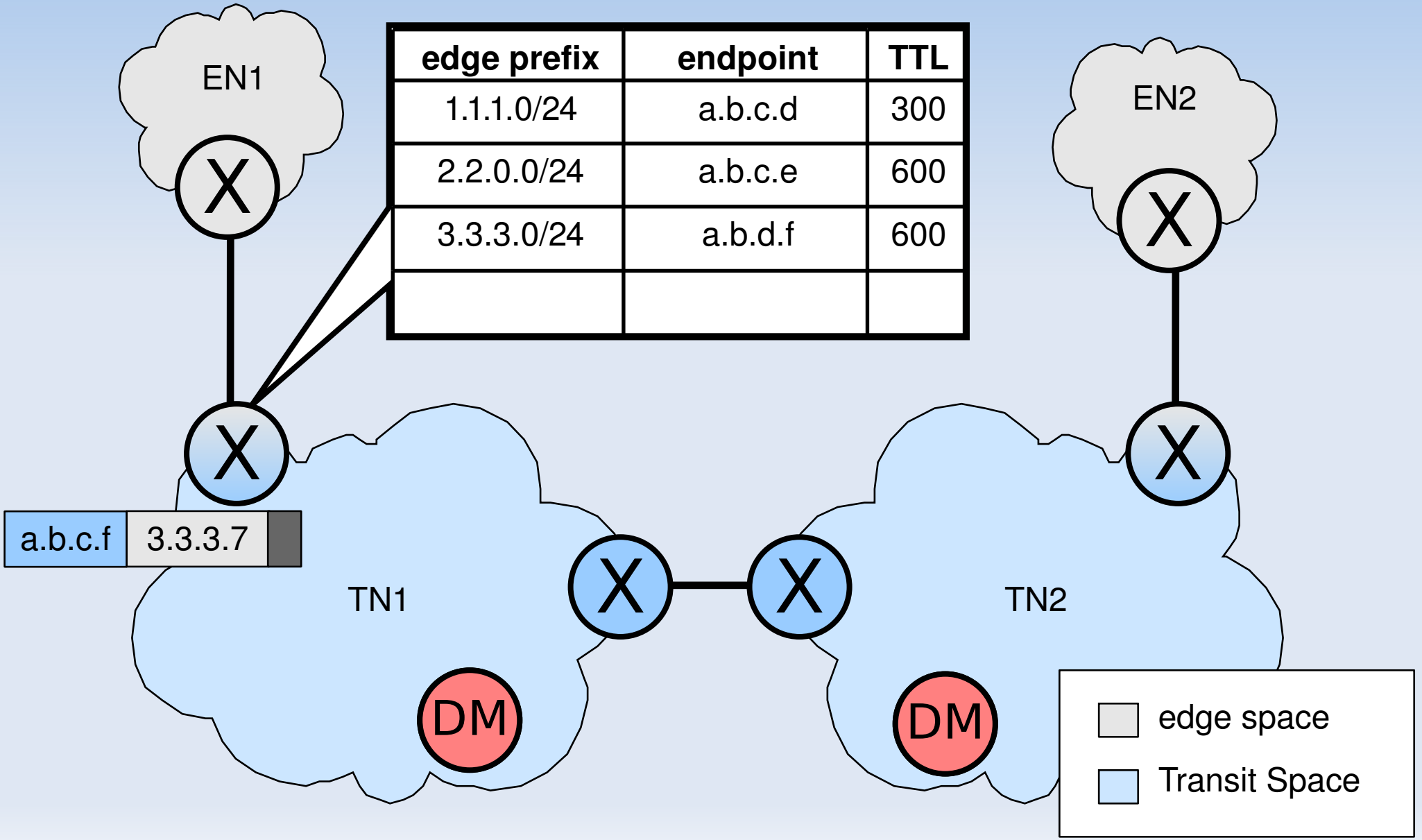
edge prefix	endpoint	TTL
1.1.1.0/24	a.b.c.d	300
2.2.0.0/24	a.b.c.e	600
3.3.3.0/24	a.b.d.f	600

APT Review

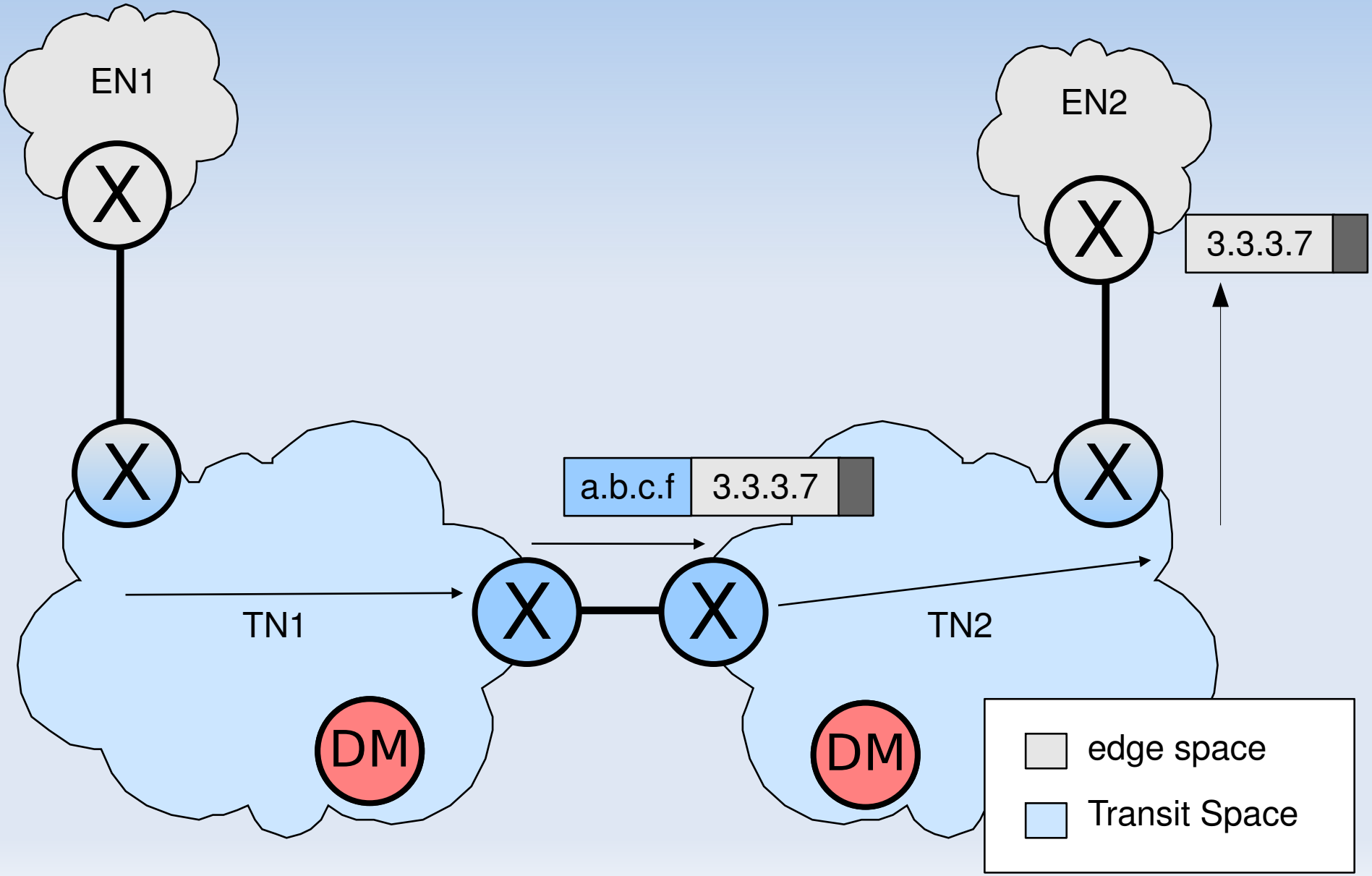


edge prefix	endpoint	TTL
1.1.1.0/24	a.b.c.d	300
2.2.0.0/24	a.b.c.e	600
3.3.3.0/24	a.b.d.f	600

APT Review


EN1
EN2
edge prefix
endpoint
TTL
1.1.1.0/24
a.b.c.d
300
2.2.0.0/24
a.b.c.e
600
3.3.3.0/24
a.b.d.f
600
a.b.c.f
3.3.3.7
TN1
TN2
DM
DM
edge space
Transit Space

APT Review


EN1
EN2
X
X
X
X
X
X
3.3.3.7
a.b.c.f
3.3.3.7
TN1
TN2
DM
DM
edge space
Transit Space

How to Evolve to APT?

- How does scalability come incrementally with the incremental deployment of APT?
- What do 1st movers get?
- What about 2nd movers?

1st Mover Incentive

- Virtual Aggregation
  - Reduce FIB table size

1st Mover Incentive: Virtual Aggregation

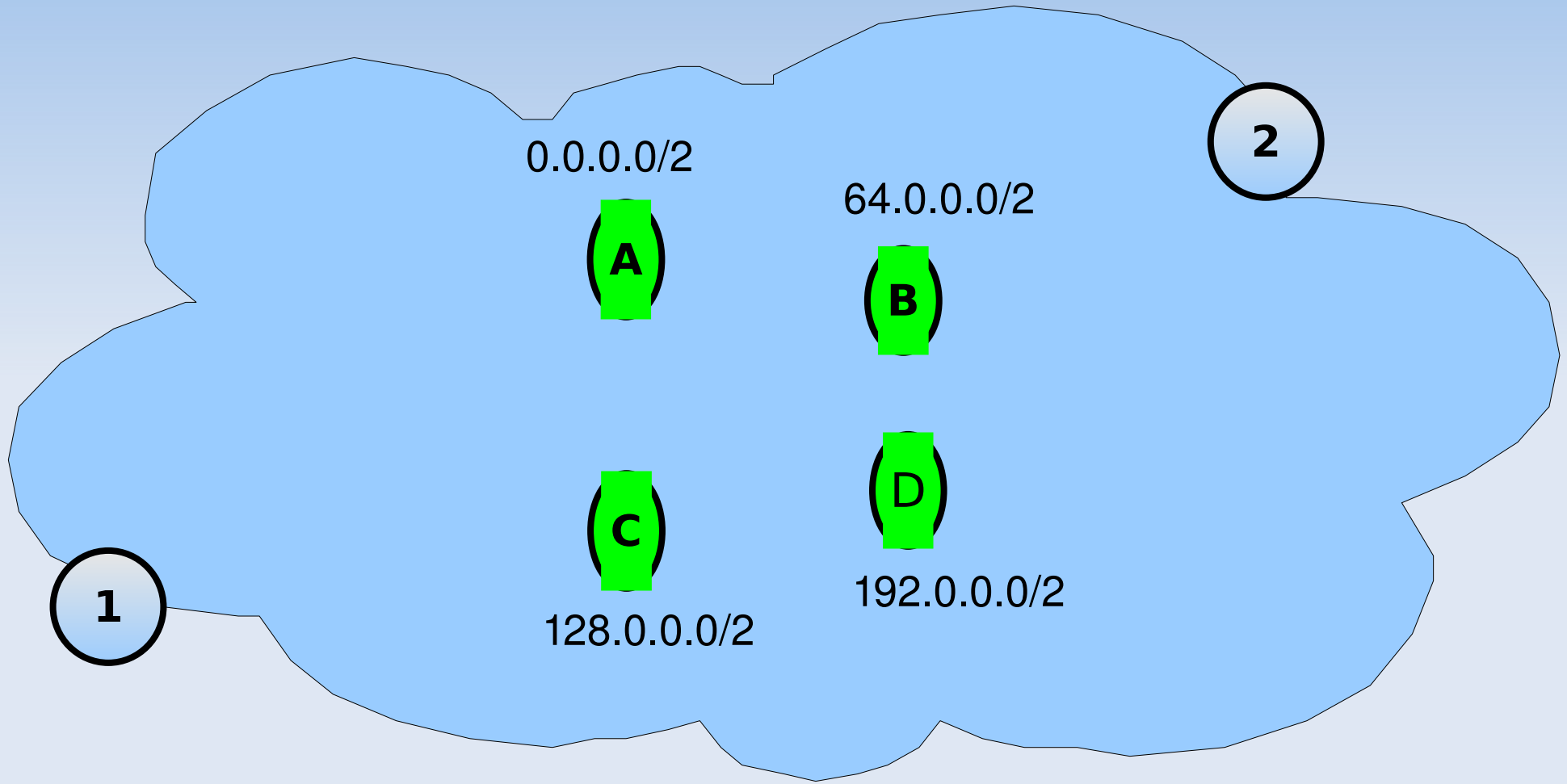
http://tools.ietf.org/html/draft-francis-idr-intra-va-01

- Allows ISPs to tune the FIB size in their routers
  - Scalability benefit
- Divide IP address space into N parts, which we call Virtual Prefixes (VPs).
  - ISPs can divide the address space any way they see fit.
  - We are currently doing measurements on which parts of the address space contain the most global prefixes.

What is Virtual Aggregation?

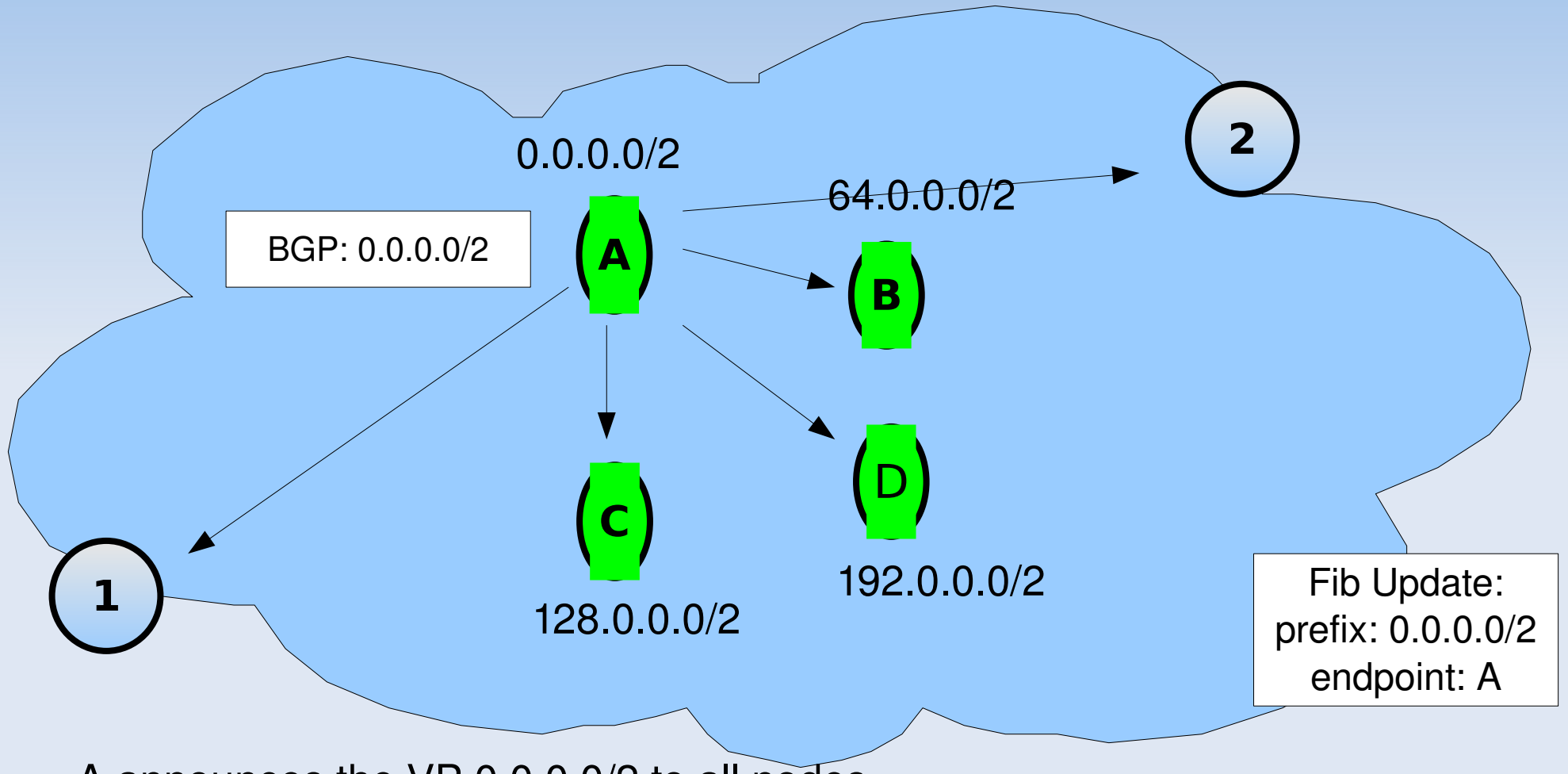
- Per Virtual Prefix, assign 1 router to be the Aggregation Point Router(APR) for the VP.
  - Must store in FIB all prefixes that fall into the range of the VP.
  - Announces the VP to other nodes.
- Edge routers FIB-install only routes to these APRs, and a default route to a core
  - See Paul Francis' IDR talk on VA and tunnel endpoints

VA Example Architecture



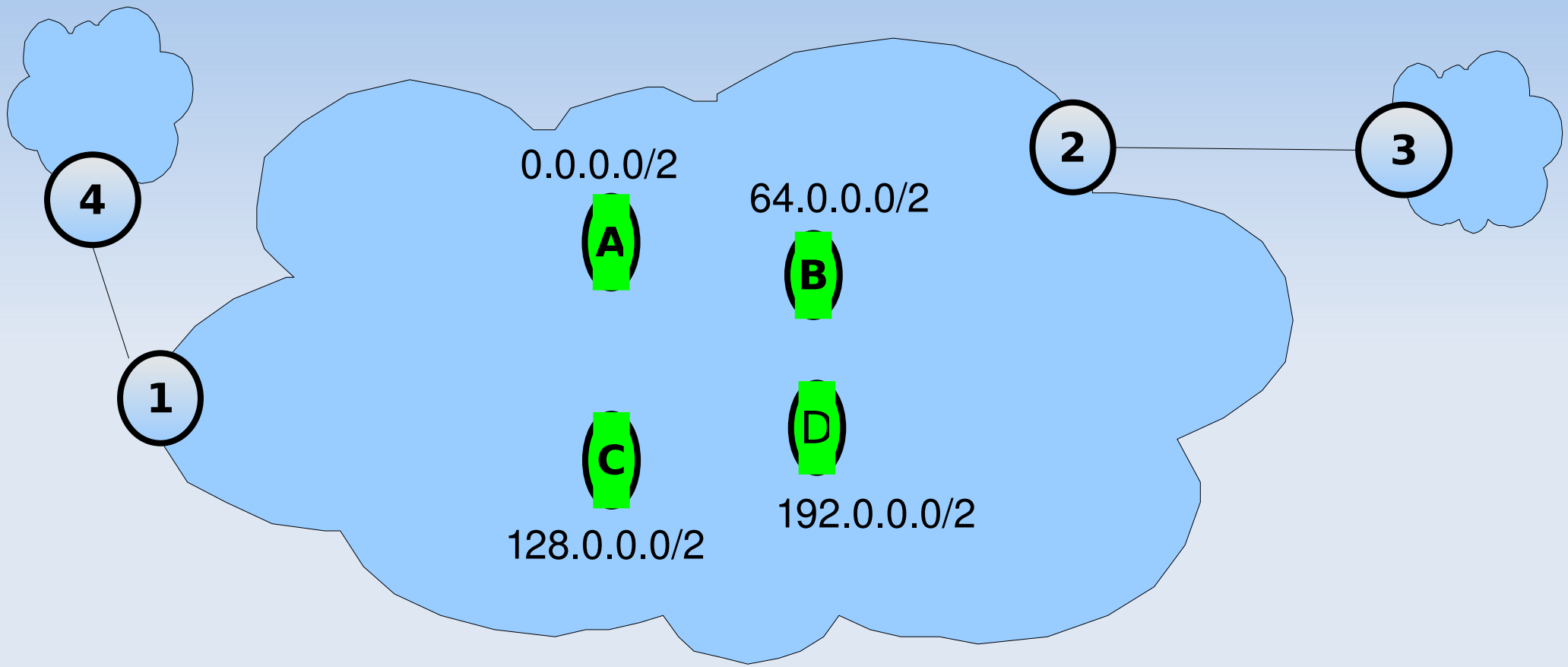
- Above is a single AS running VA
- numbered nodes are EDRs
- lettered nodes are APRs and announce VPs

APRs announce VAs



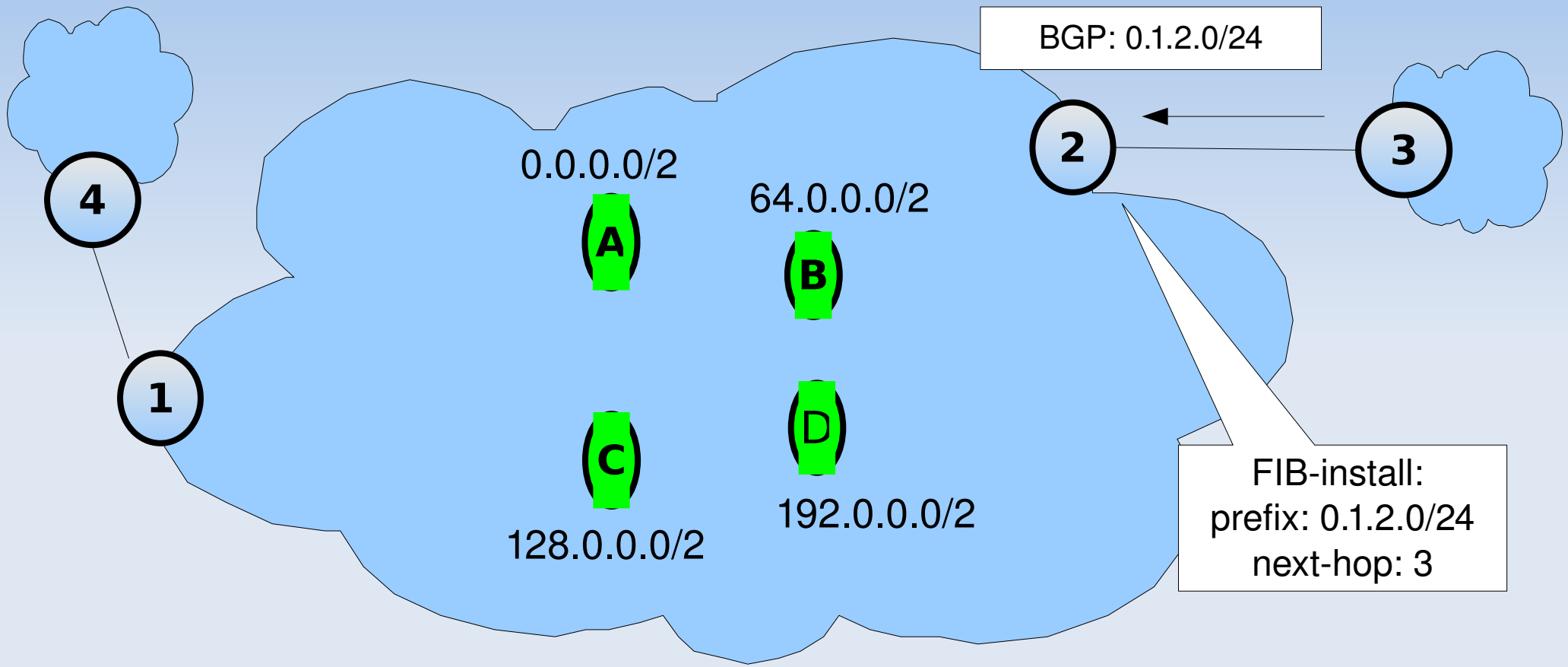
- A announces the VP 0.0.0.0/2 to all nodes
- All routers store VP in both RIB and FIB (FIB entry shown on right)
- B, C, D would also announce their VPs, and other routers would store the update in both RIB and FIB

FIB supression example



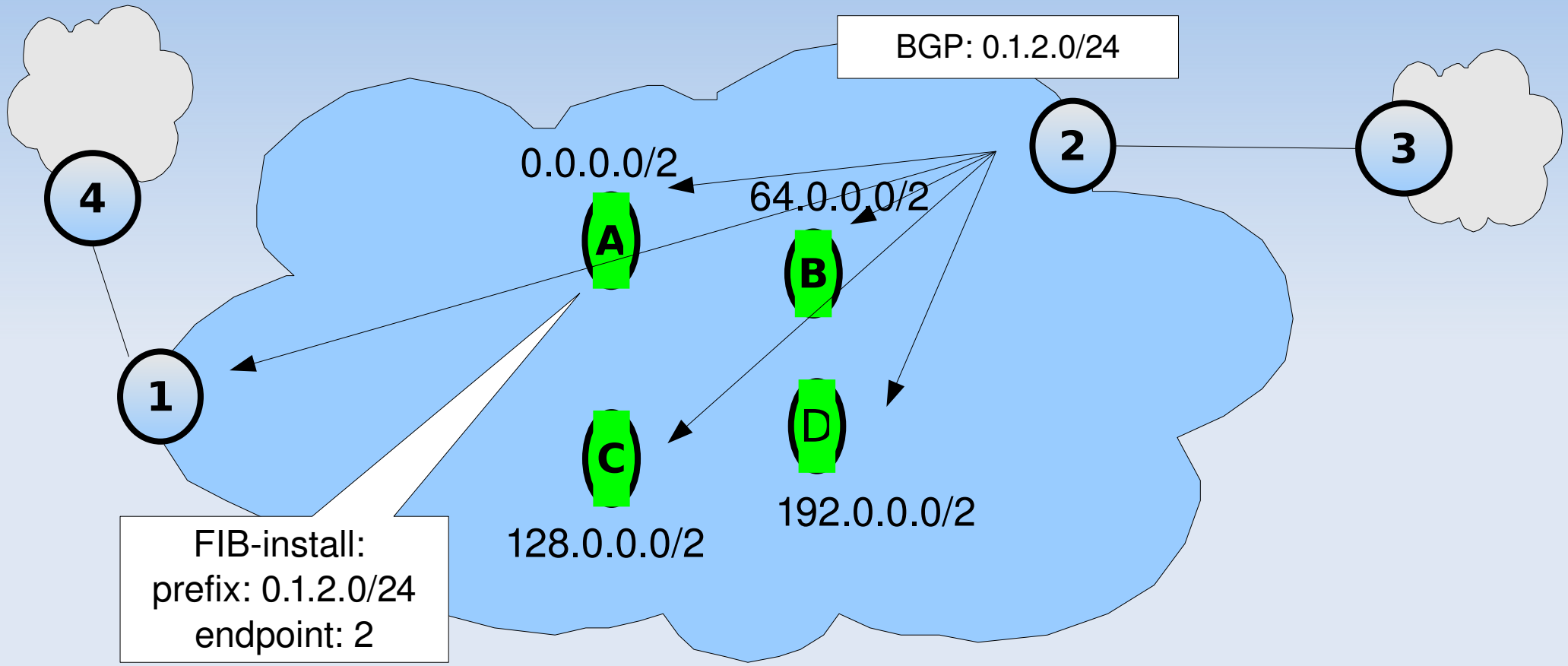
- 3, 4 are routers in peer ASes
- A,B,C,D make up DM system
- 1,2 are EDRs

FIB supression example



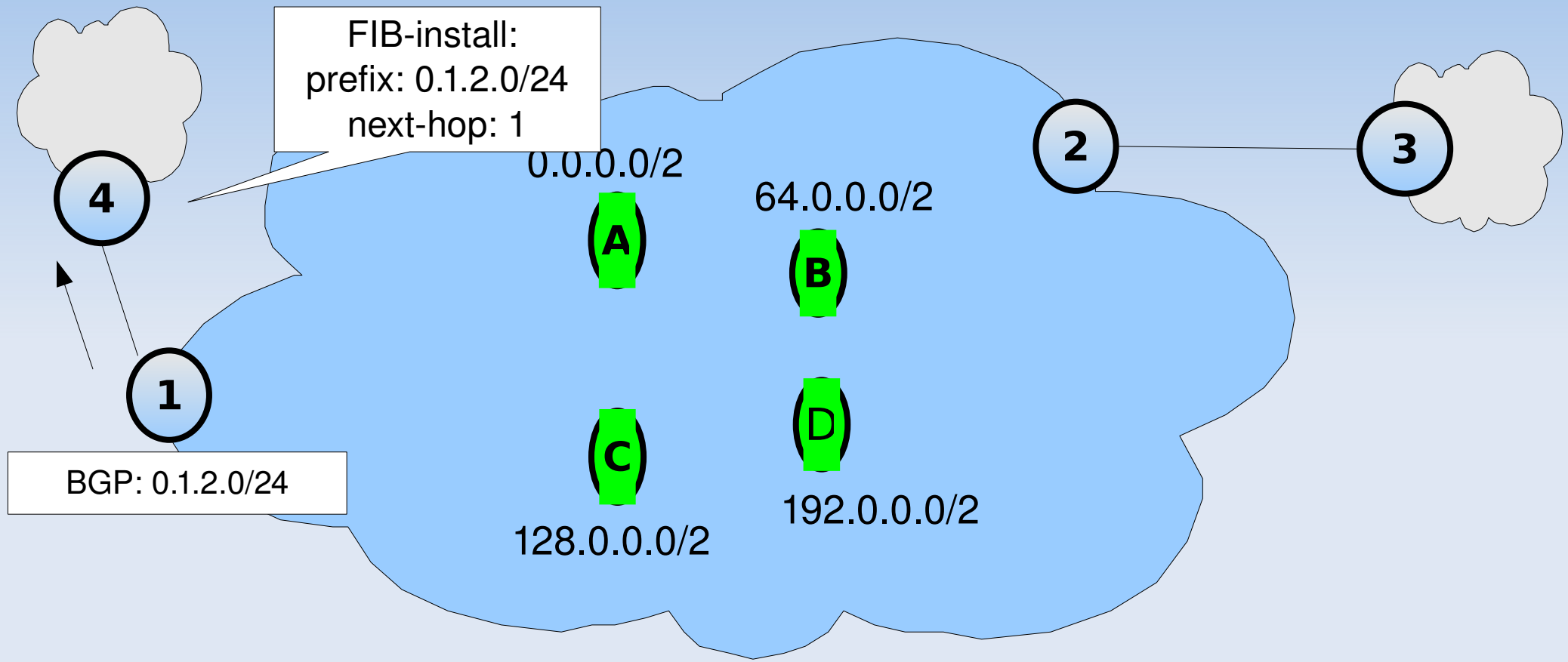
- 2 receives a bgp update from external peer 3
- 2 stores update in its FIB and RIB
- next hop in 2's FIB is the external peer 3

FIB supression example



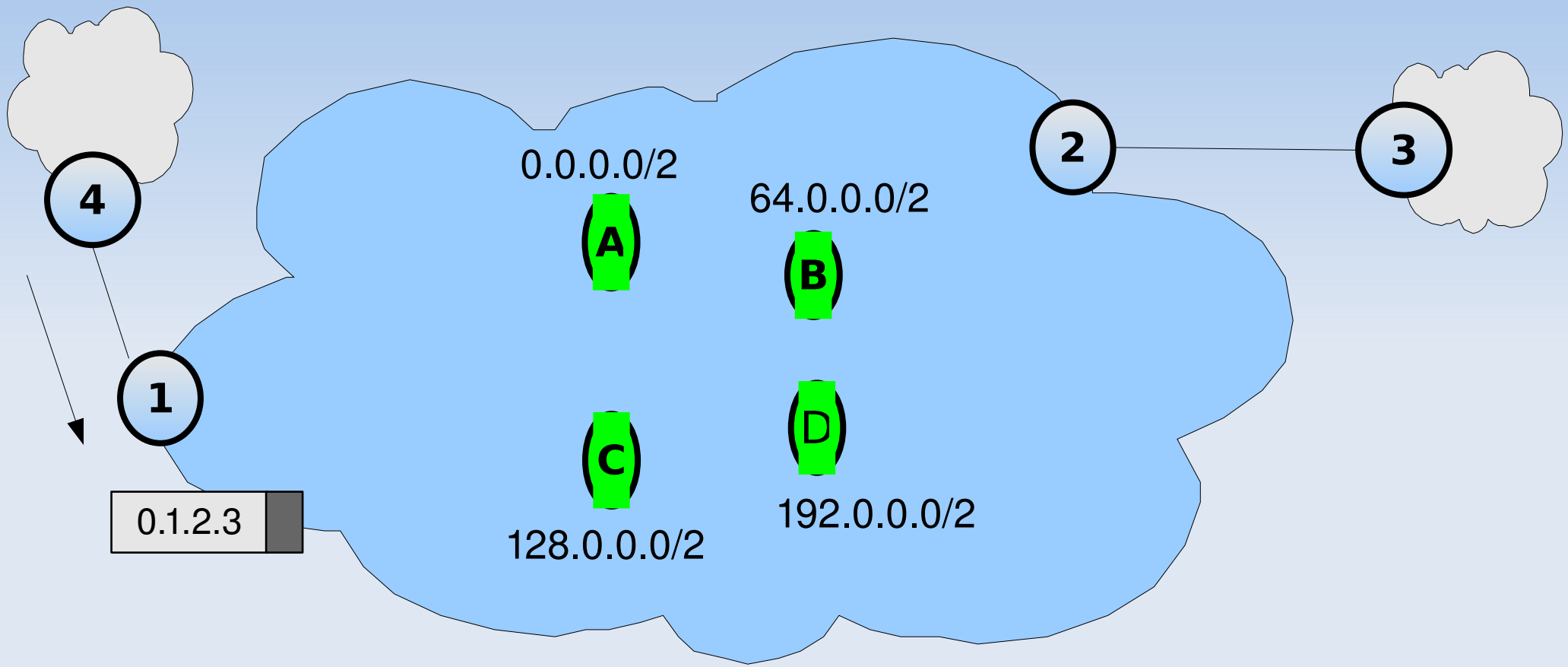
- 2 forwards updates to internal peers
- All routers store update in RIB but only A stores update in its FIB
- Other routers supress this FIB entry, saving memory

FIB supression example



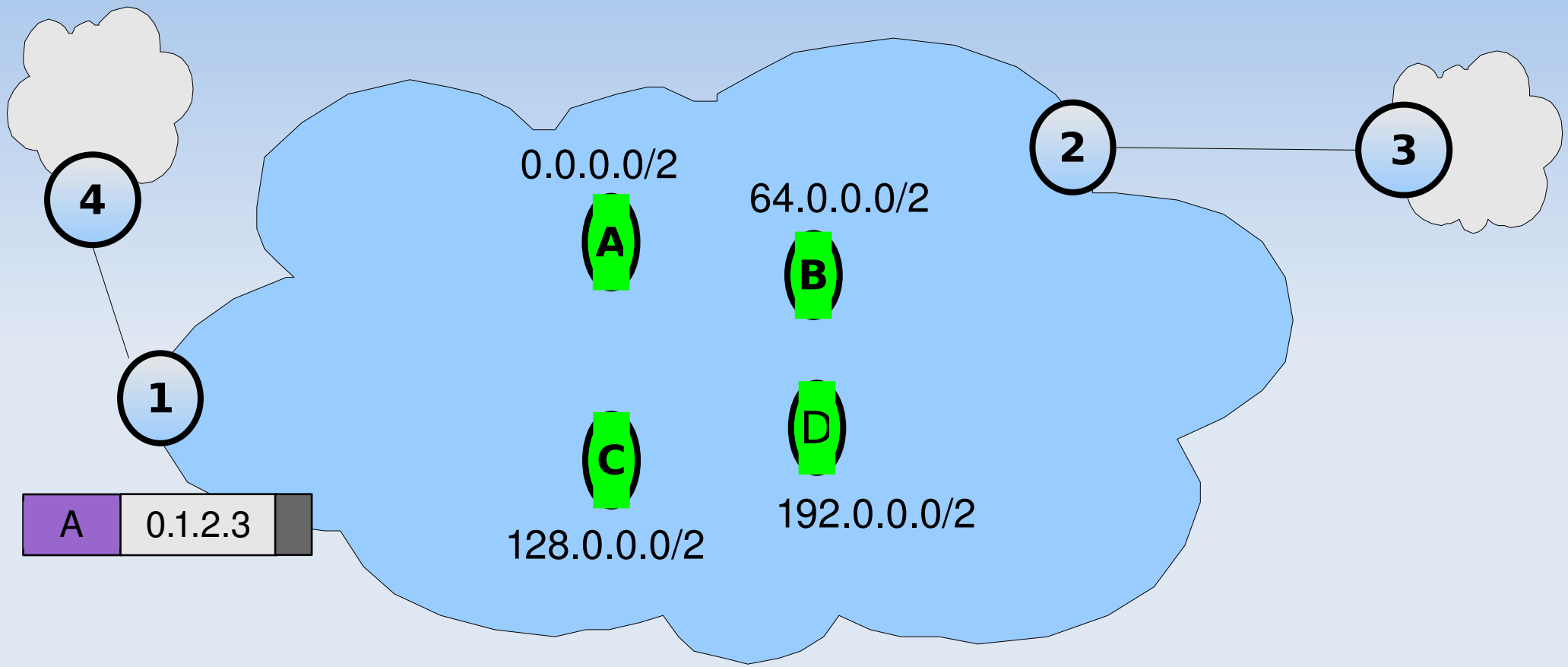
- 1 still forwards update to 4

Packet Delivery under VA



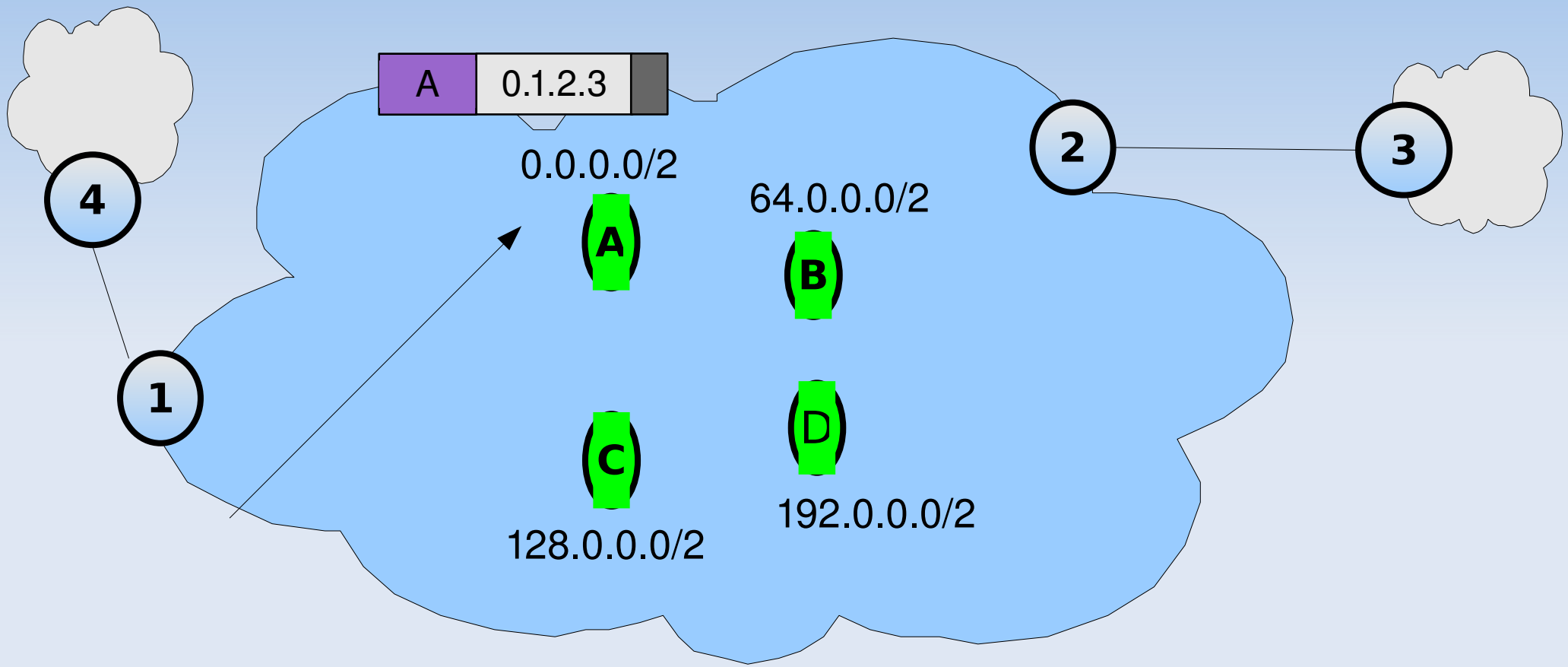
- 4 sends packet destined to 3's network

Packet Delivery under VA

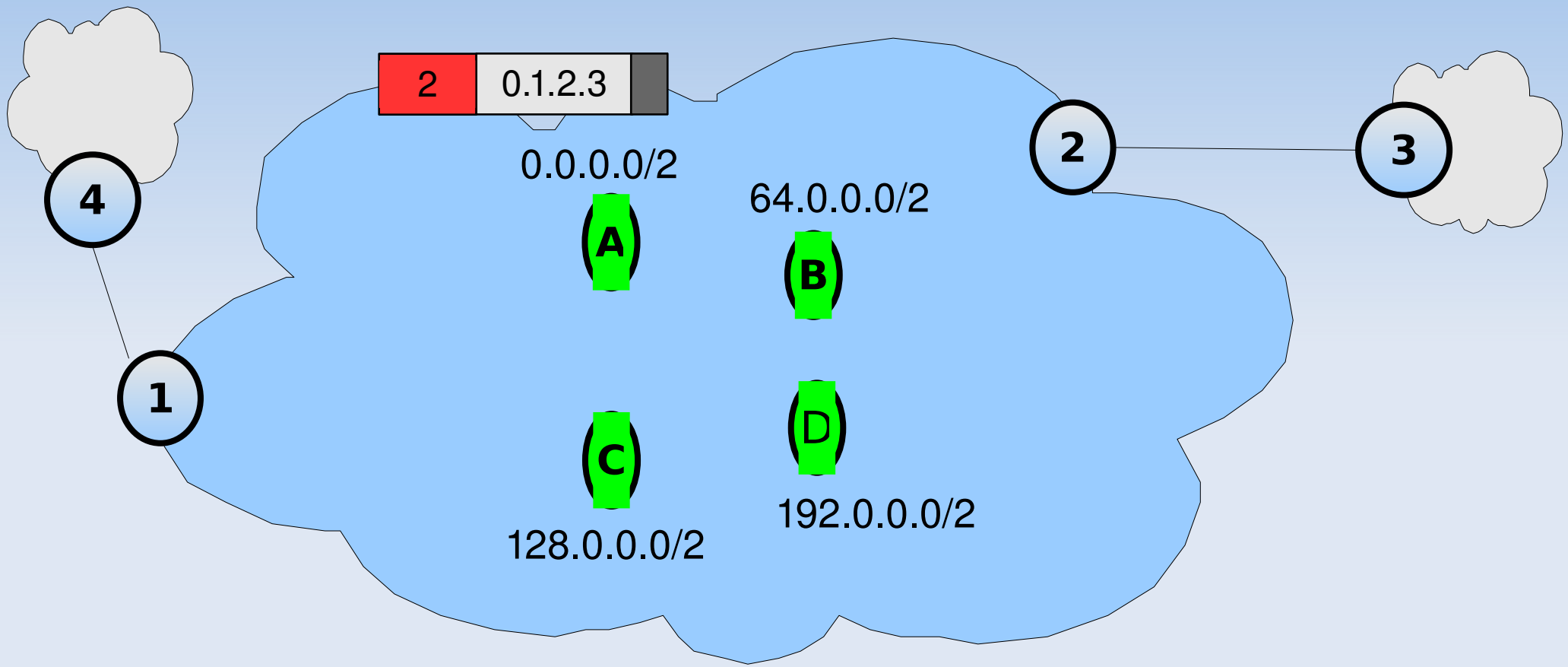


- 1 encaps to APR of the 0.0.0.0/2 VP (node A)

Packet Delivery under VA

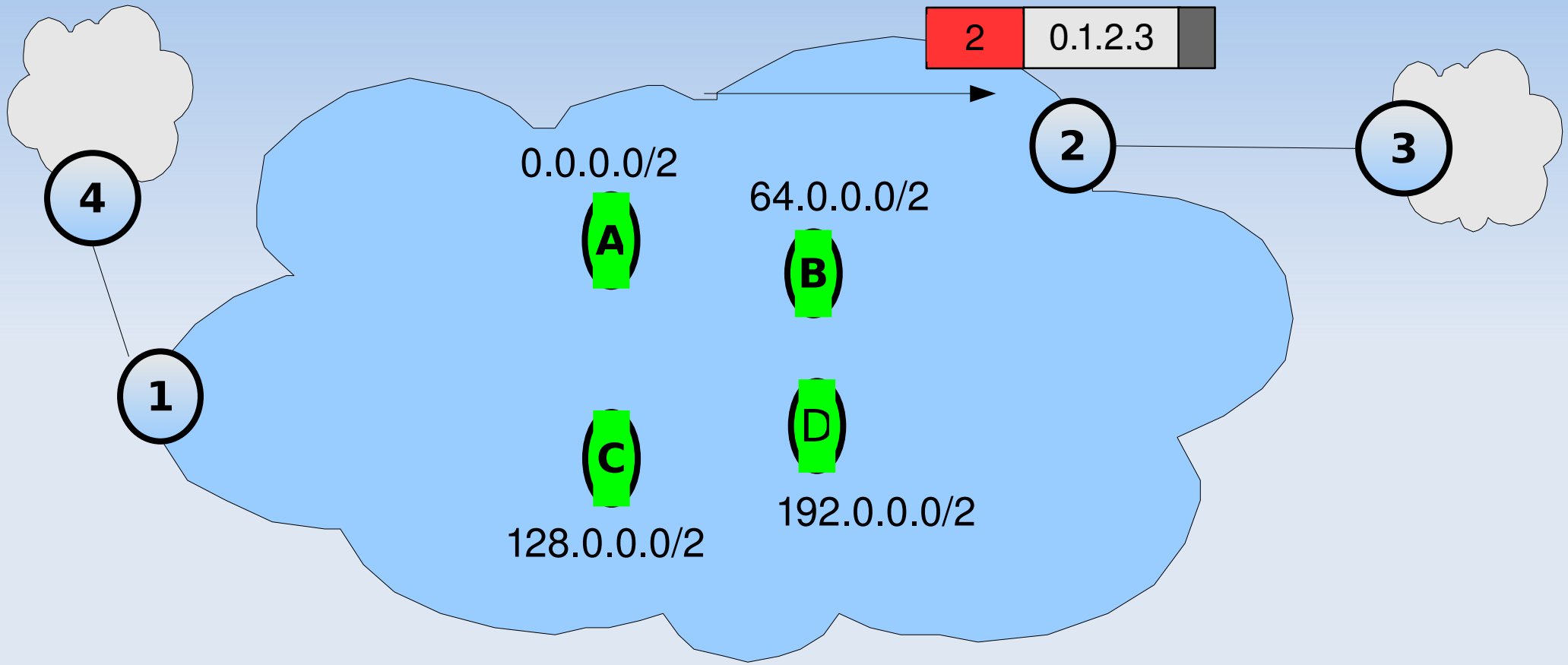


Packet Delivery under VA



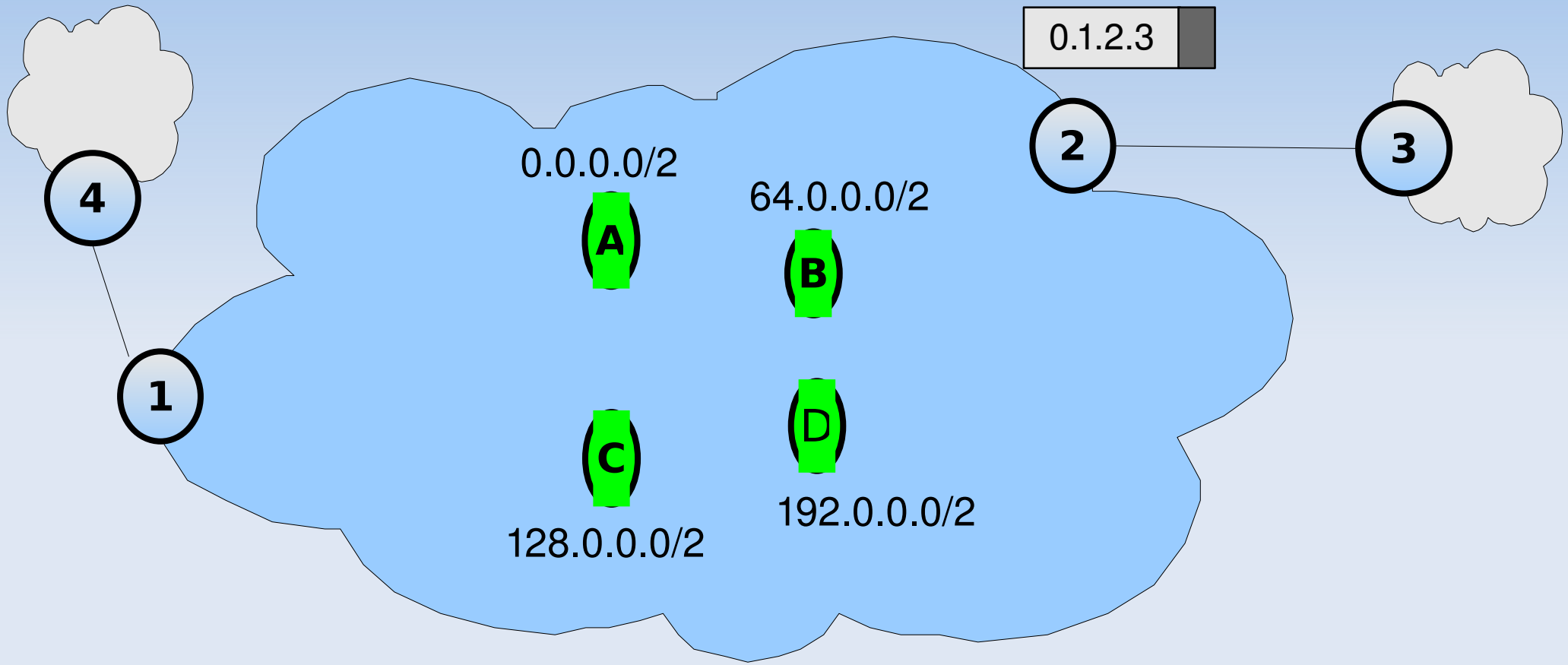
- A knows that the endpoint for the packet is 2
- A re-encaps to 2

Packet Delivery under VA



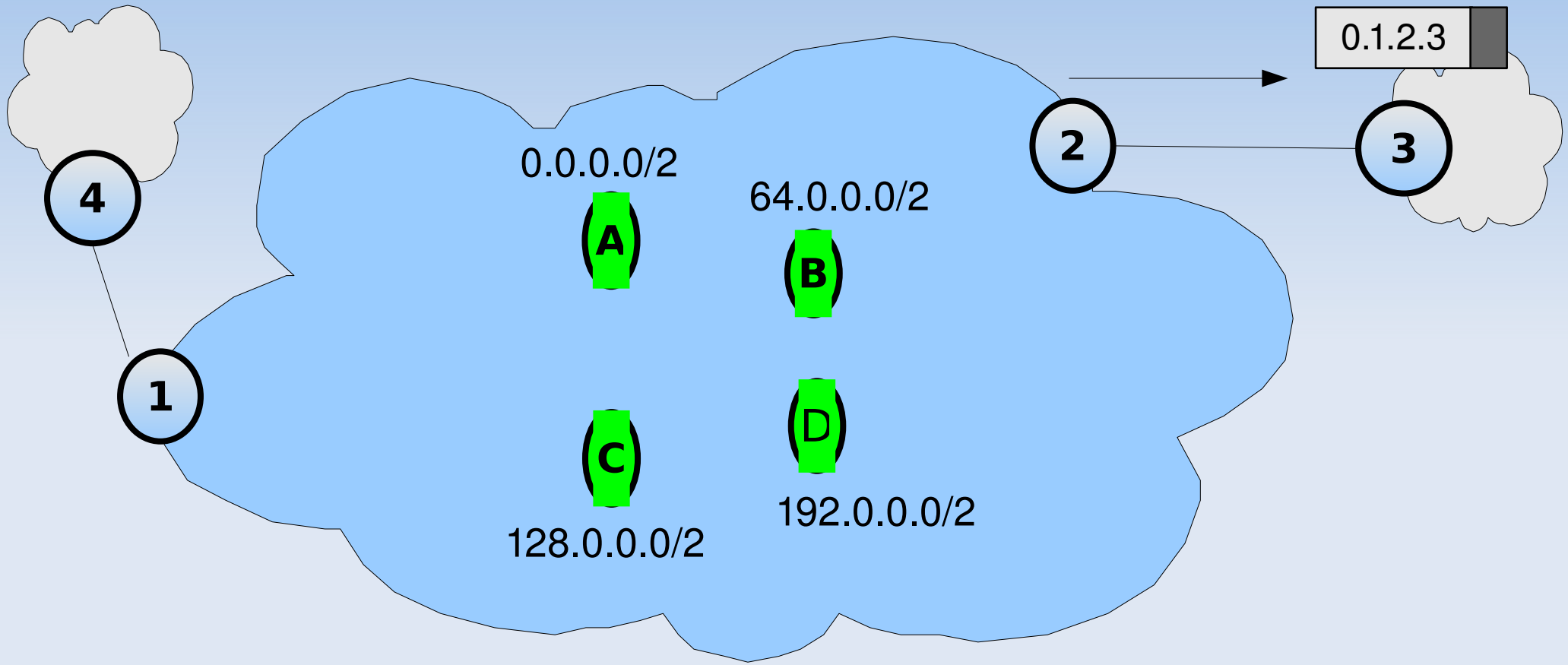
- Packet goes to 2

Packet Delivery under VA



- 2 decaps packet

Packet Delivery under VA



- 2 knows that next hop to packet is 3
- 2 delivers packet

How are VA ISPs considered APT 1st movers?

- VA architecture consists of routers with 'mappings' from a smaller routing table to a larger one with specific information.
  - It's intra-ISP map & encap
- APT Default Mapper = the set of APRs.
- ISPs doing VA get FIB scalability, which is the 1st scalability benefit in the APT evolution.

Great, but what about 2nd movers?

- Some ISPs now have default mappers and have controlled FIB size with intra-ISP map & encap.
  - But not all ISPs.
- What are our 2nd mover incentives?
- And where does inter-ISP map & encap come in?

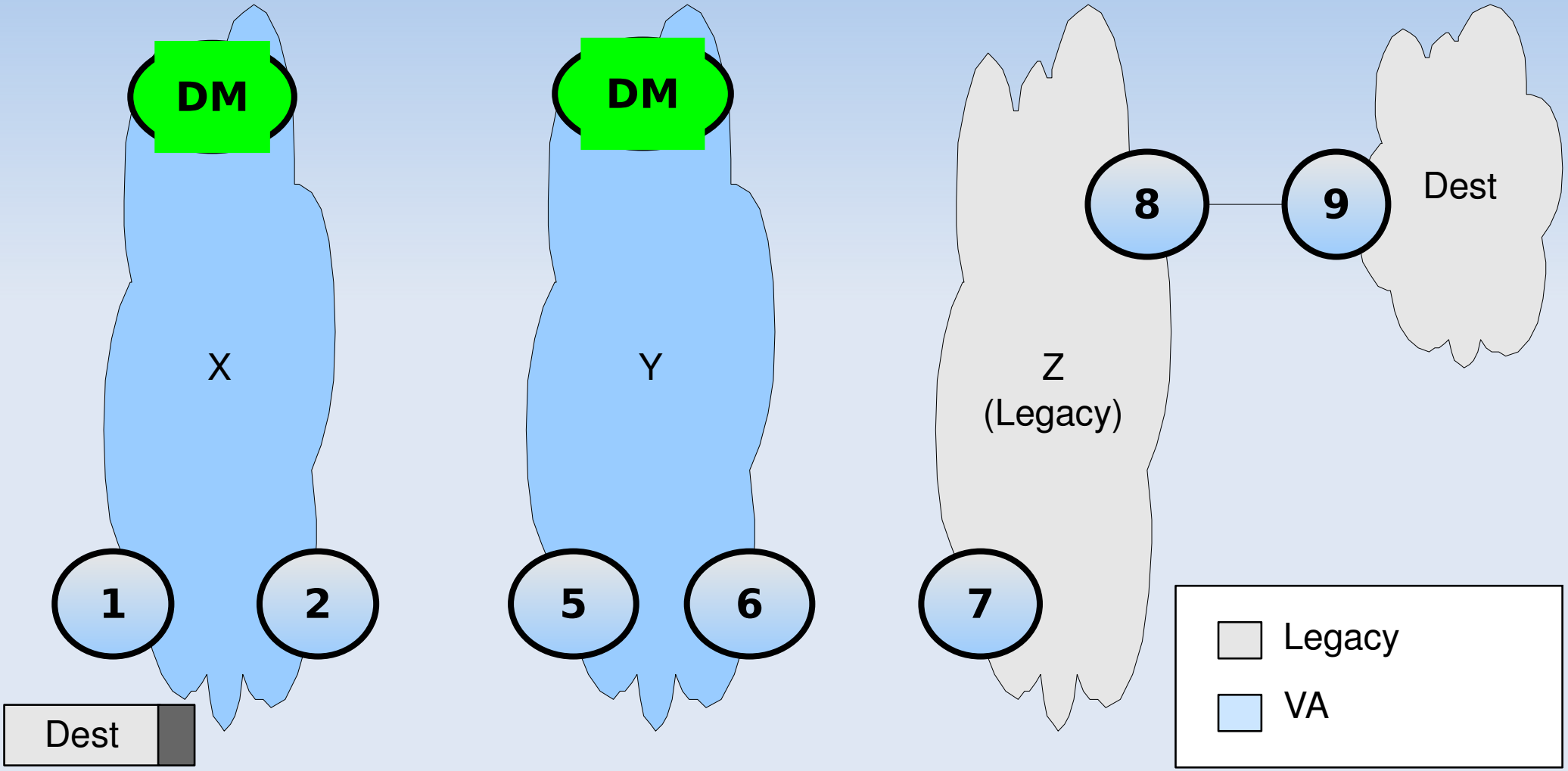
2nd Mover Incentives

- Performance benefits for their customers using map & encap
  - 2nd Movers can avoid the 'stretch issue'
  - Also get TE benefits.
- How?
  - Inter-ISP map & encap

The Stretch Issue

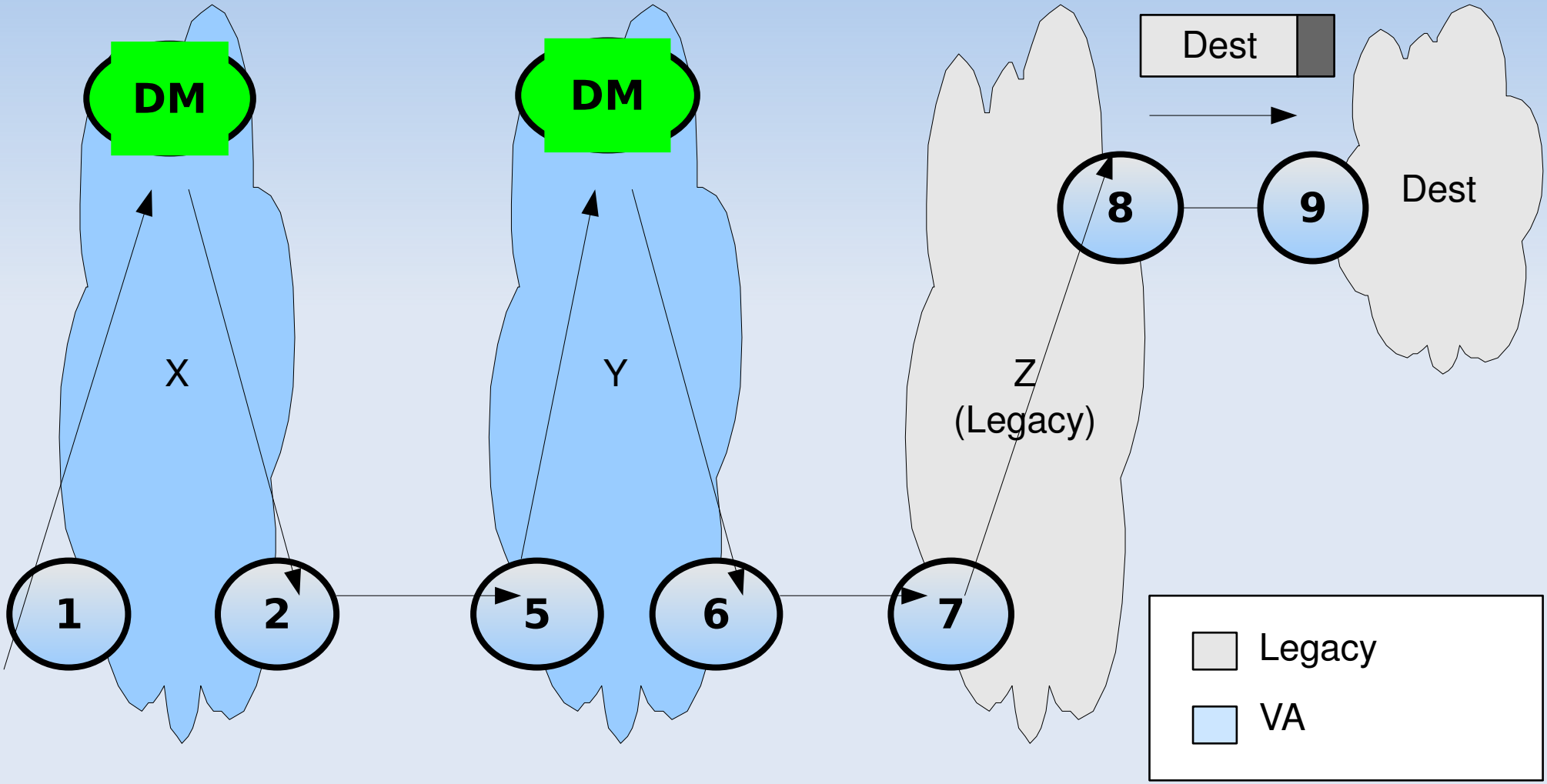
- Packets traveling through an ISPs running VA will travel extra hops before exiting the AS
- Not an issue if only a handful of ISPs run VA.
- But with a significant fraction of 1st movers, stretch can add up.

The Stretch Issue



- Dest is customer of legacy ISP Z
- DM = Default mapper

The Stretch Issue



Avoiding Stretch Issue

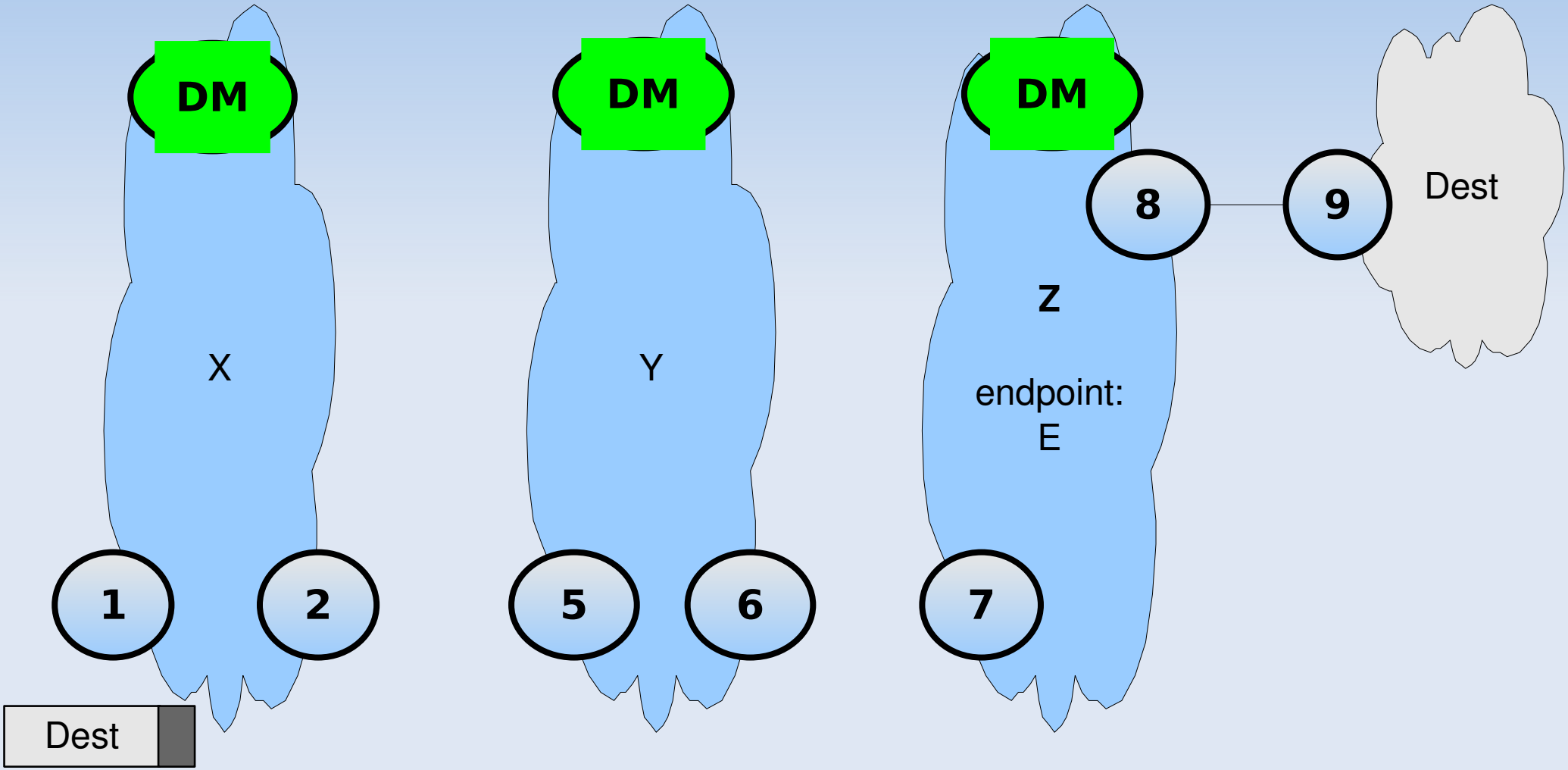
- ISPs doing VA each announce a globally routable tunnel-endpoint prefix and map their customer prefixes to this endpoint
- Mappings are exchanged between these ISPs.
  - Perhaps using Mapped-BGP
  - Or “tunnel endpoints in BGP”

    http://www.ietf.org/internet-drafts/draft-xu-idr-tunnel-00.txt

Avoiding Stretch Issue

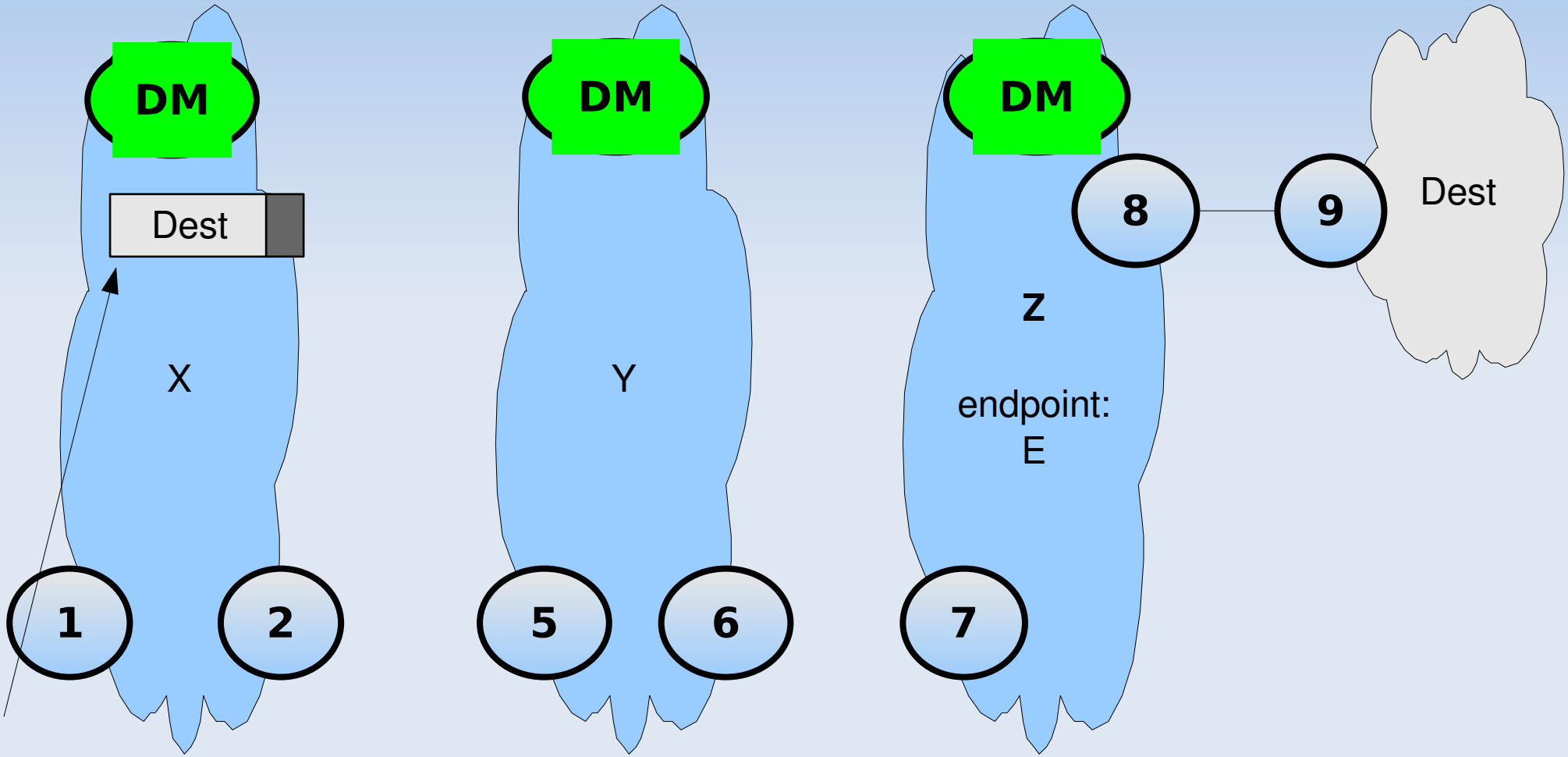
- Mappings ensure that packets stretch <= once
- Z doesn't want his customer packets stretched, so Z has incentive to participate in inter-ISP map & encap.

No Stretch Issues for Upgraded ISPs



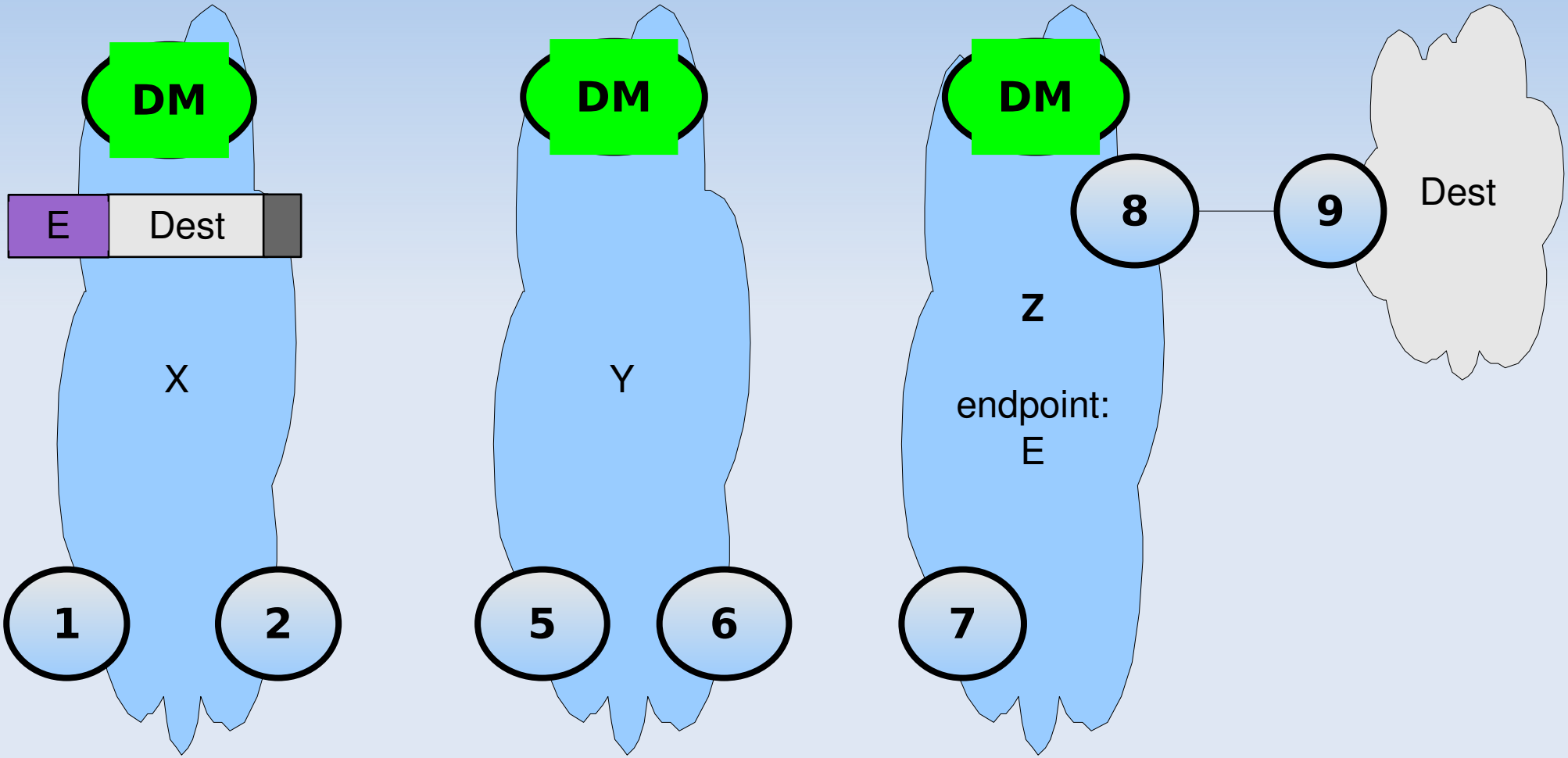
- Now Z exchanges mappings with X and Y
- Provider endpoint prefix is 'E'

No Stretch Issues for Upgraded ISPs



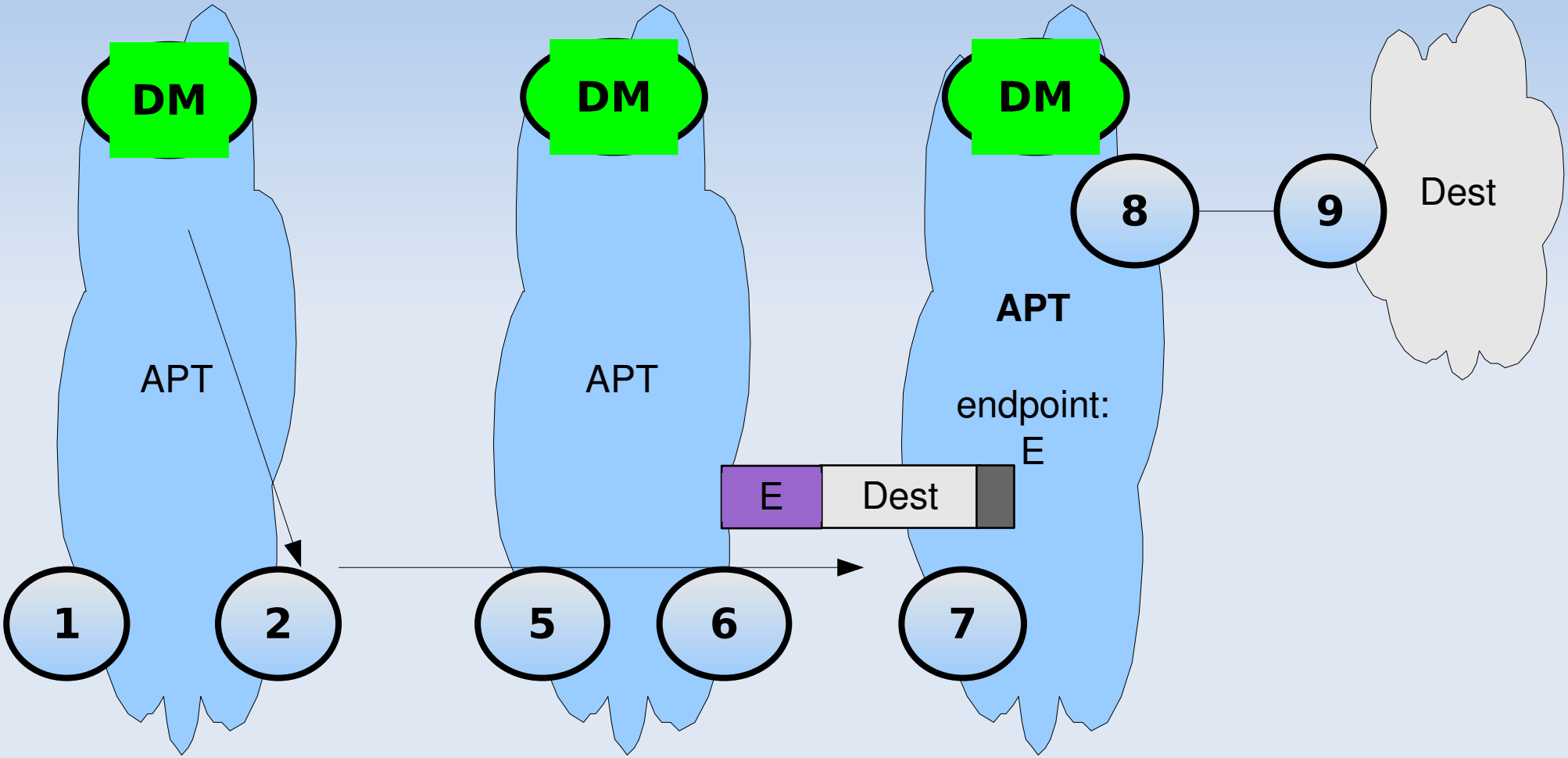
- 1 Sends to DM

No Stretch Issues for Upgraded ISPs



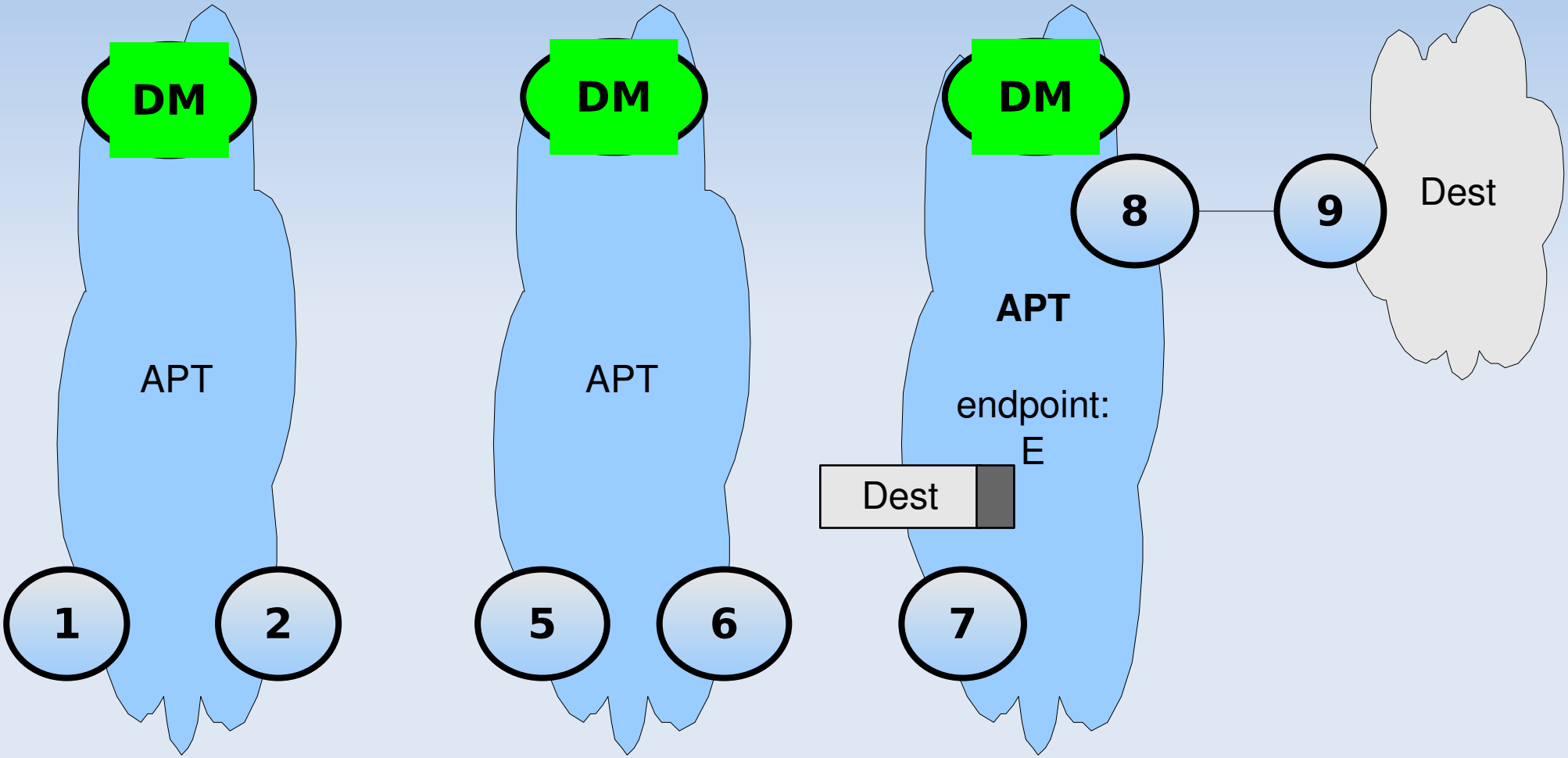
- DM encaps packet with globally routable endpoint

No Stretch Issues for Upgraded ISPs



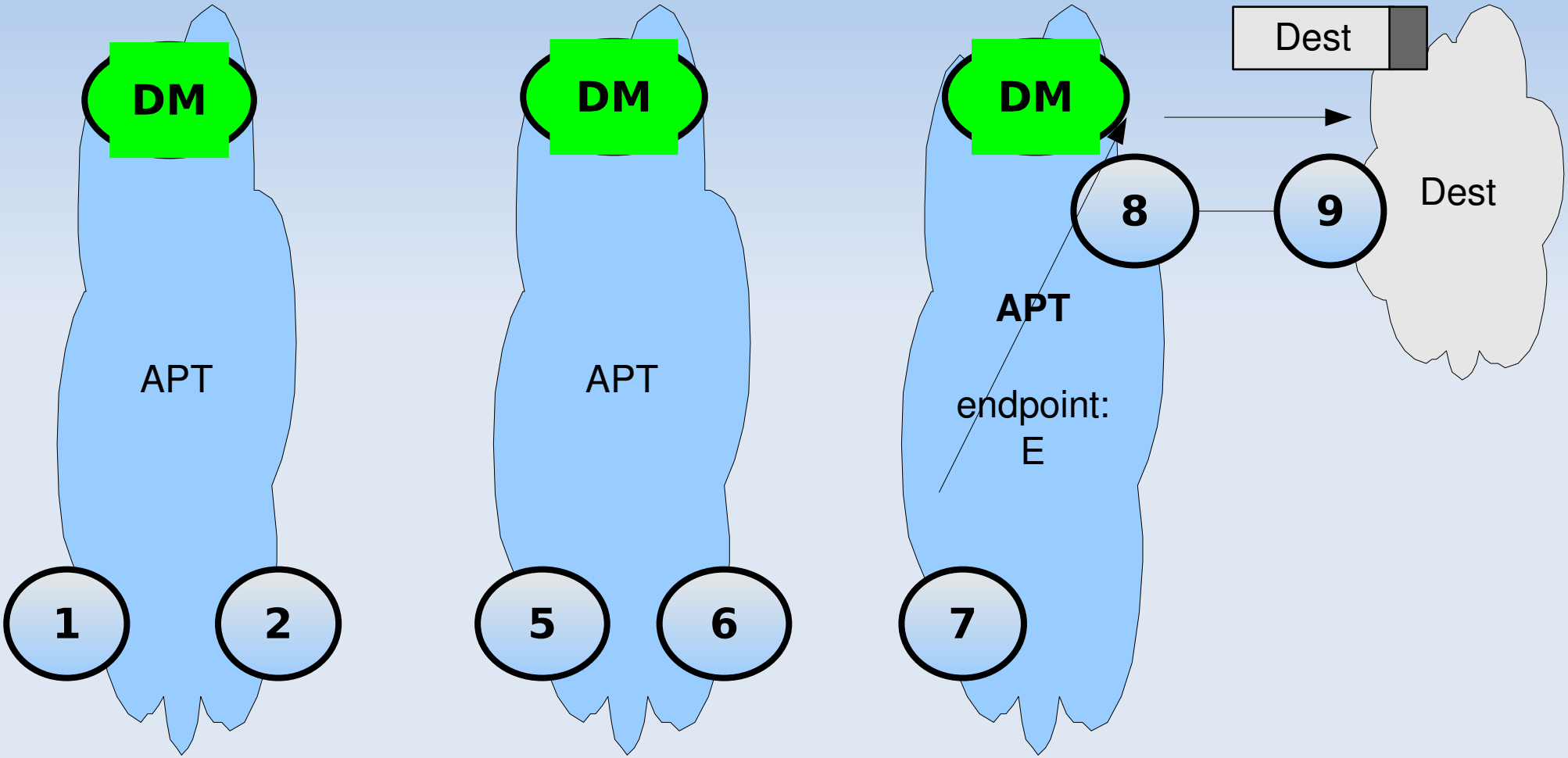
- Once packet is encapped, it is routed to 7 w/o stretch

No Stretch Issues for Upgraded ISPs



- 7 decaps packet, has dest entry in FIB
  - Customer prefixes should be in provider's FIBs for performance reasons

No Stretch Issues for Upgraded ISPs



- provider delivers packet to customer
  - Customer prefixes should be in provider's FIBs for performance reasons

TE Incentives for 2nd Mover

- Mappings allow for better traffic engineering options.
  - Explicit Ingress TE and path selection to tunnel endpoints
- Details can be found in Paul's draft:

http://www.ietf.org/internet-drafts/draft-xu-idr-tunnel-00.txt

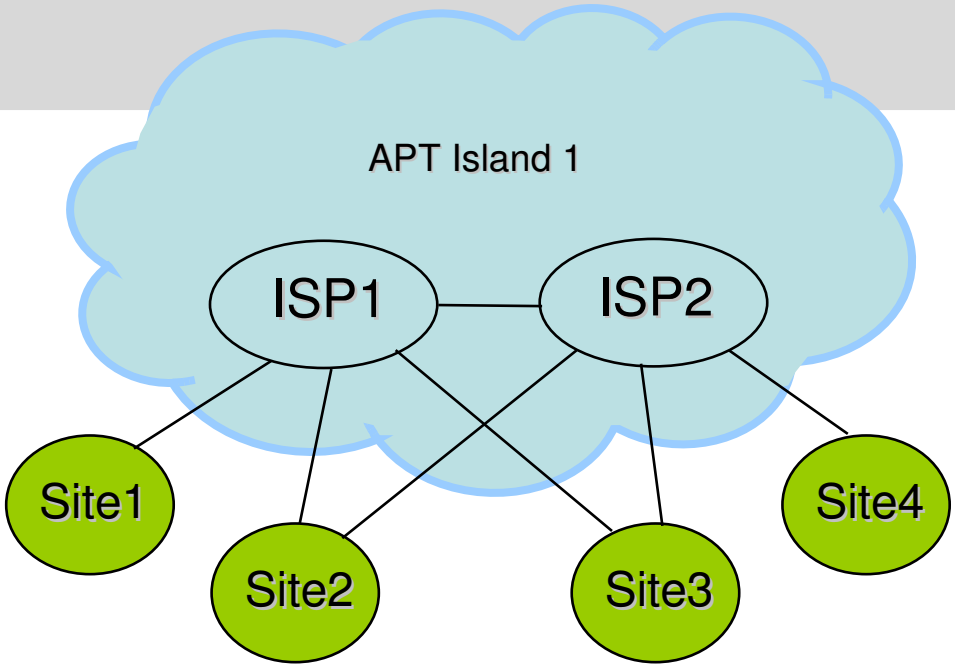
Also...

- Border routers only need to FIB-install routes to tunnel endpoints, and a default route to a default mapper.
- Border routers (usually older machines at ISPs) get the most relief.

What next?

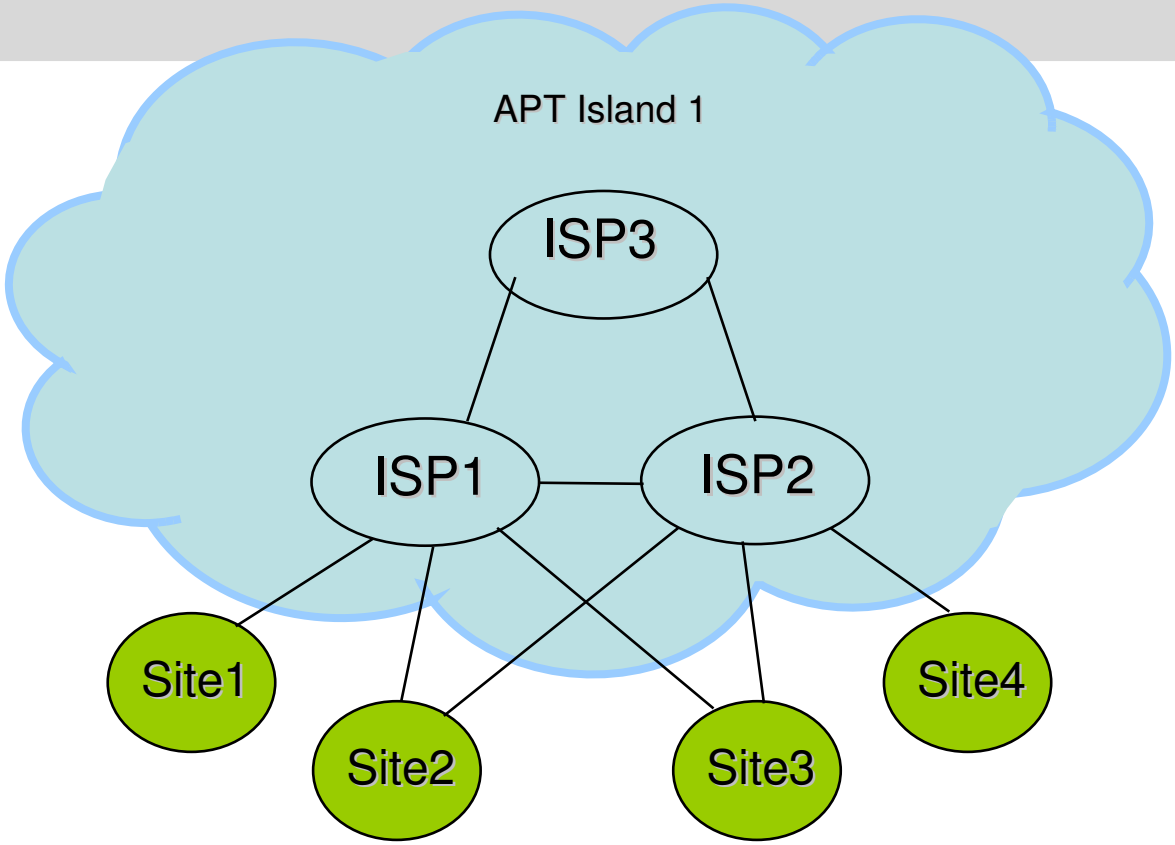
- So now we have the full APT architecture deployed in some ISPs via incremental steps.
  - Default mappers
  - Simple border encap/decap routers
  - Map & encap between APT ISPs
- FIB table is under control
  - Via virtual aggregation
- **RIB and update dynamics need to scale!**

APT Failure-Handling Feature



- Providers of multihomed edge sites can encap to the other provider.
- If the ISP1-site2 link goes down, ISP1 can deliver packets going to Site2 by encapping to ISP2

APT Island RIB Reduction



- Prefixes for Sites 1, 2, 3, and 4 removed from the ISP3's RIB table
  - Larger and larger reduction in RIB table size as deployment grows.

Dynamics scale with RIB here

- As incremental deployment leads to further decrease of the RIB, dynamics will reduce as well.
  - Less edge site reachability to monitor
  - Less edge prefix flappings to worry about

RIB reduction benefits create snowball effect

- With more ISPs adopting APT, we get more scaling benefits.
- With more scaling benefits, we get more ISPs adopting APT.

In closing...

- There is an evolutionary path towards routing scalability
  - Via ideas from APT, Virtual Aggregation, Mapped-BGP
  - Selfish scalability carrots can get people to make the changes towards scalability.
- Scalability can come incrementally via Map & Encap.

The_end

- Thanks!
- Questions? Comments?