Designing A New Routing and Addressing Architecture for Future Internet

Lixia Zhang

UCLA

Asia Future Internet Summer School

August 2008

Disclaimer First

- Wearing no hat
- Personal opinion
- The eFIT/APT design mentioned later is joint effort with colleagues and students: Dan Massey, Lan Wang, Beichuan Zhang, Dan Jen, Michael Meisel

The real challenge:

- The marching order is out: designing a Future Internet
- The question: How?
- The real challenge:

Where do we start to look for answers ???

- Don't be afraid of sticking your neck out to answer this question

Don’t be afraid of sticking your neck out !

- “The only utility of science is to go on and to try to make guesses. So what we always do is to stick our necks out.
- “Of course this means that science is uncertain; the moment that you make a proposition about a region of experience that you have not directly seen then you must be uncertain.
- But we always must make statements about the regions that we have not seen, or the whole business is no use.”

—from "Character of Physical Laws" by R. Feynman

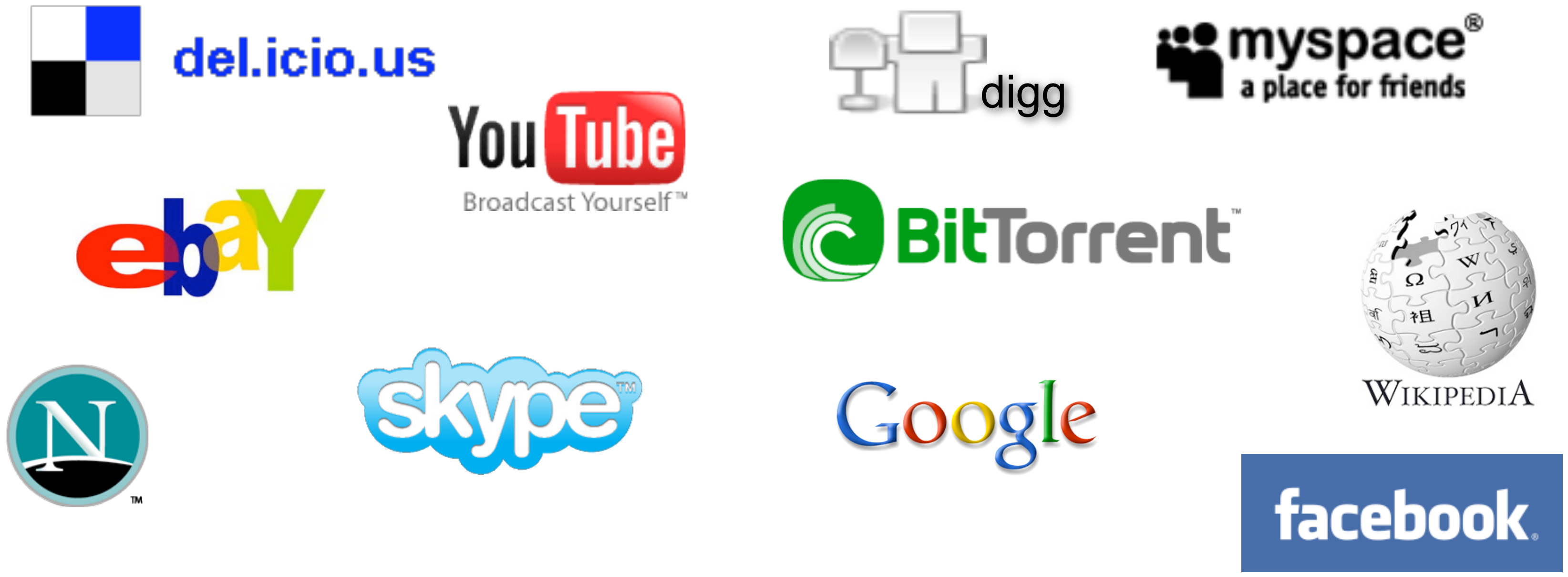
Setting the Stage

As network researchers, how much does our job differ from physicists?

- For them: Someone built the physical world, all is left is to understand it
  - No dispute over whether the world was designed right or wrong
- For us: No one built Internet for us
  - There is even no bable on how to do it
- We create this artifact ourselves through (informed) trials and errors

The right direction to look

- To find out how ***the Future Internet*** should look like is to look back
  - and look around (to other scientific endeavors)
- The success of Internet = the success of Internet applications



What to learn from past success

- Internet's success driven by innovations from user community
- A fundamental enabler: end-to-end reachability
  - Network engineering over the years: bigger networks, delivering bits faster, cheaper, more reliable
  - Together with Moore's Law which puts computers (of all forms) into everyone's hand
- Predicting next killer app? No glorious record

⇒The Future Internet must remain as an enabler to continuous innovations

Today's Hurdles in Routing & Addressing

- Running out of IPv4 address
- Pervasive NAT deployment
  - Block new applications
- Global routing table size growing at super-linear speed
  - Together with high update churns
  - One of the concerns with IPv6 deployment
- Ever increasing security threats
  - DDoS
  - Redirection attacks utilizing spoofed source addresses
  - Route hijacking
  - Attacks directly aimed at routing infrastructure

How did these problems creep in?

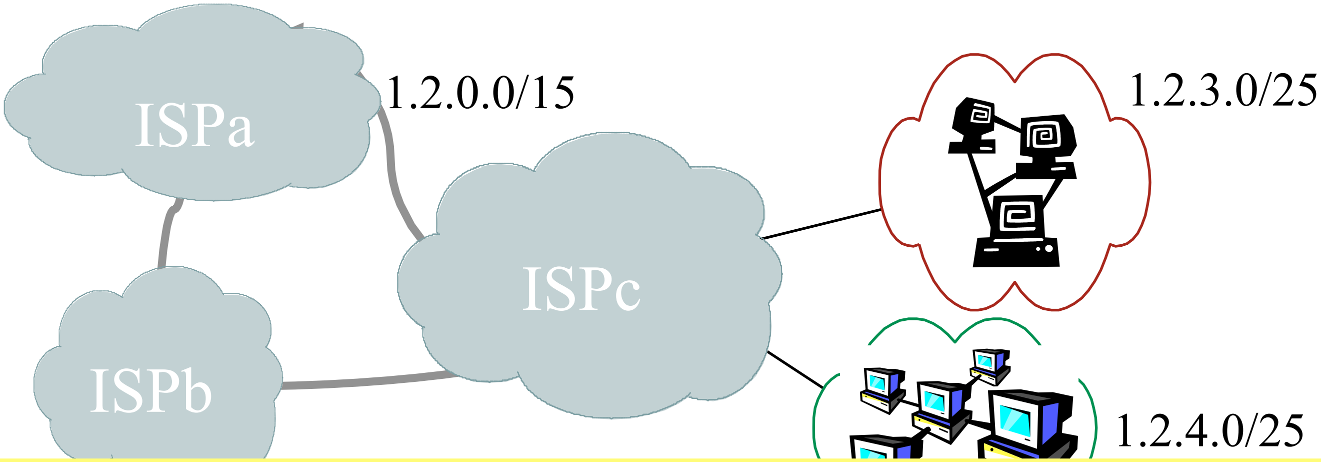
- Running out of IPv4 address
- Pervasive NAT deployment
  - Block new applications
- Global routing table size growing at super-linear speed
  - Together with high update churns
  - One of the concerns with IPv6 deployment
- Ever increasing security threats
  - DDoS
  - Redirection attacks using spoofed source addresses
  - Route hijacking
  - Attacks directly aimed at routing infrastructure

1.Success disaster

2.Signal the need for further architecture changes

There have been changes in the past

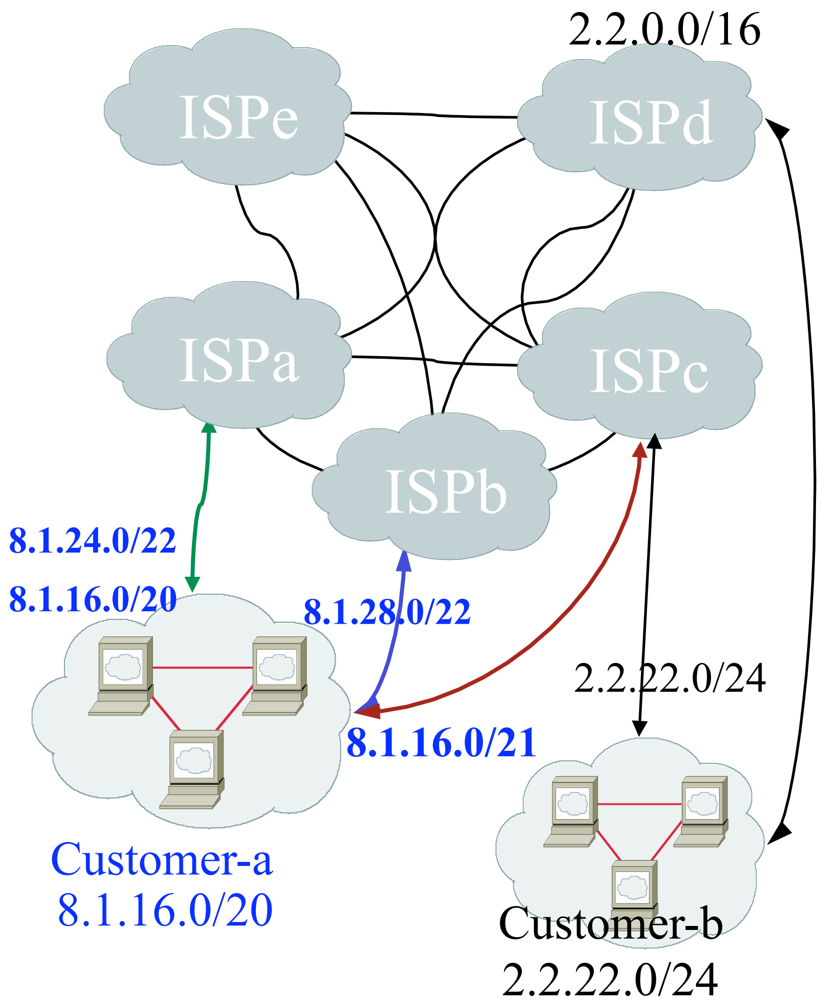
- Autonomous Systems (80's): to match network management and control with administrative boundaries
- CIDR (Classless InterDomain Routing, early 90's): to enable flexible address allocation size and topological aggregation



Today: pervasive multihoming and TE deployment

Driving factors behind the routing growth

- Multihoming
- Traffic engineering (TE)



Routing Table

……

8.1.16.0/20

2.2.0.0/16

…… .

Routing Table

……

8.1.16.0/20

8.1.24.0/22

8.1.28.0/22

8.1.16.0/21

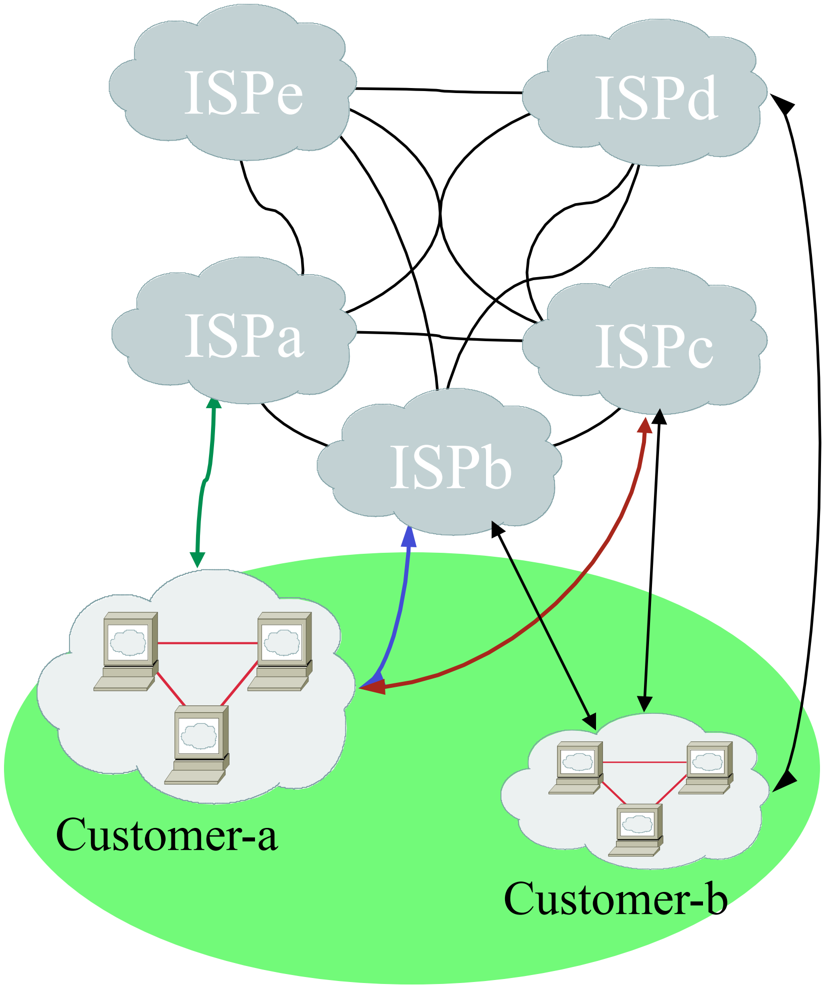
2.2.0.0/16

2.2.8.0/22

Who benefits and who pays

- Whoever doing multihoming & traffic engineering benefit
- The routing system as a whole bears the cost
- Lack of alignment of cost and benefit, leading to ever increasing prefix de-aggregation
  - The expectation: situation is likely getting worse as we approach IPv4 address exhaustion

Customers needs



- Enough IP addresses
- Changing providers without renumbering

→NAT offers these advantage

- Freedom for TE
- Real IP addresses!

IPv6 solves the first & last problems, but not the middle two.

RIR policies?

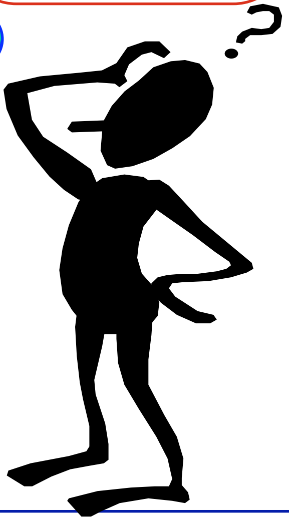
Service providers

Need topologically aggregatable address allocations to scale the routint infrastructure

RIRs
ARIN
RIPE NCC
APNIC
LACNIN
AfriNIC

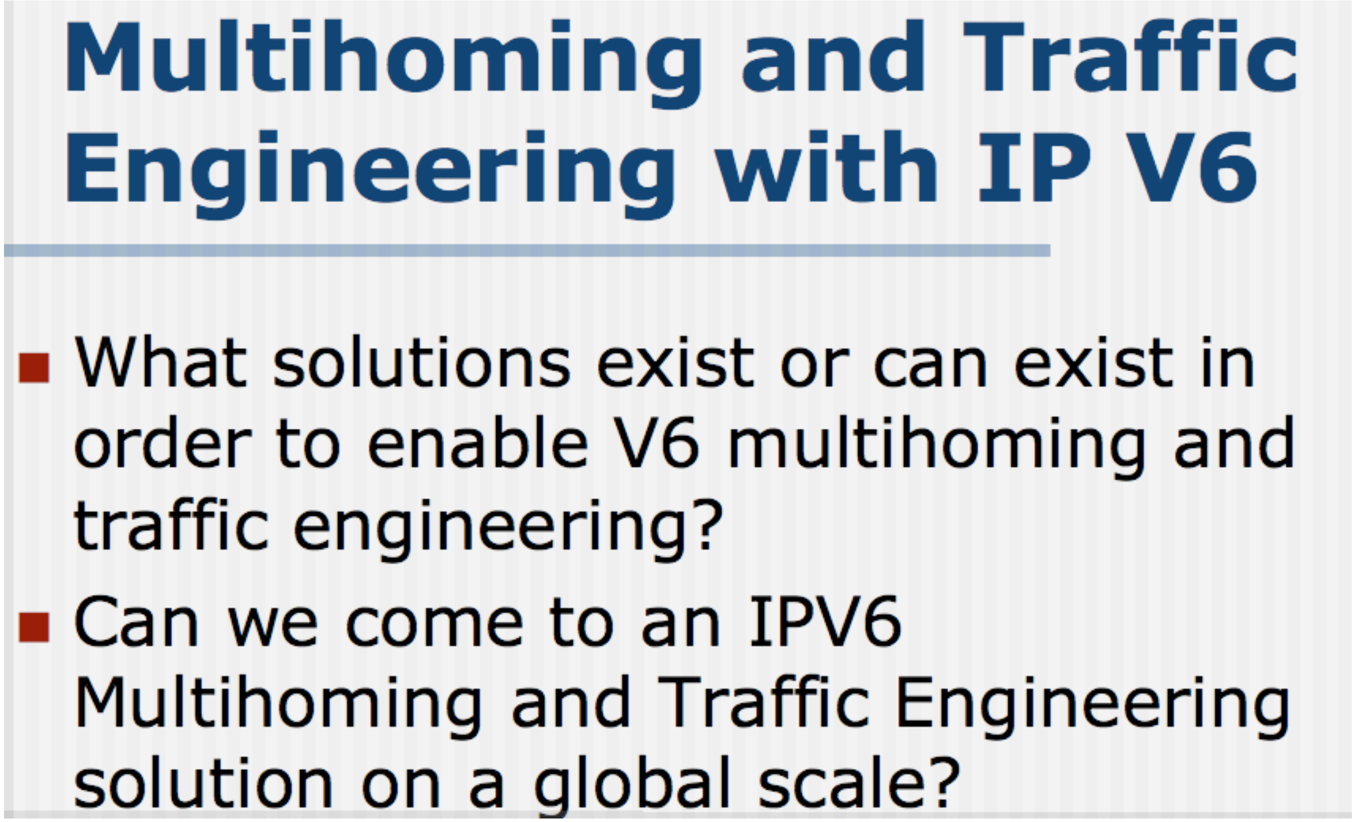
Internet customers

Given me provider-independent (PI) address blocks!



At IETF67, November 2006

Marla Azinger, ARIN Advisory Council:

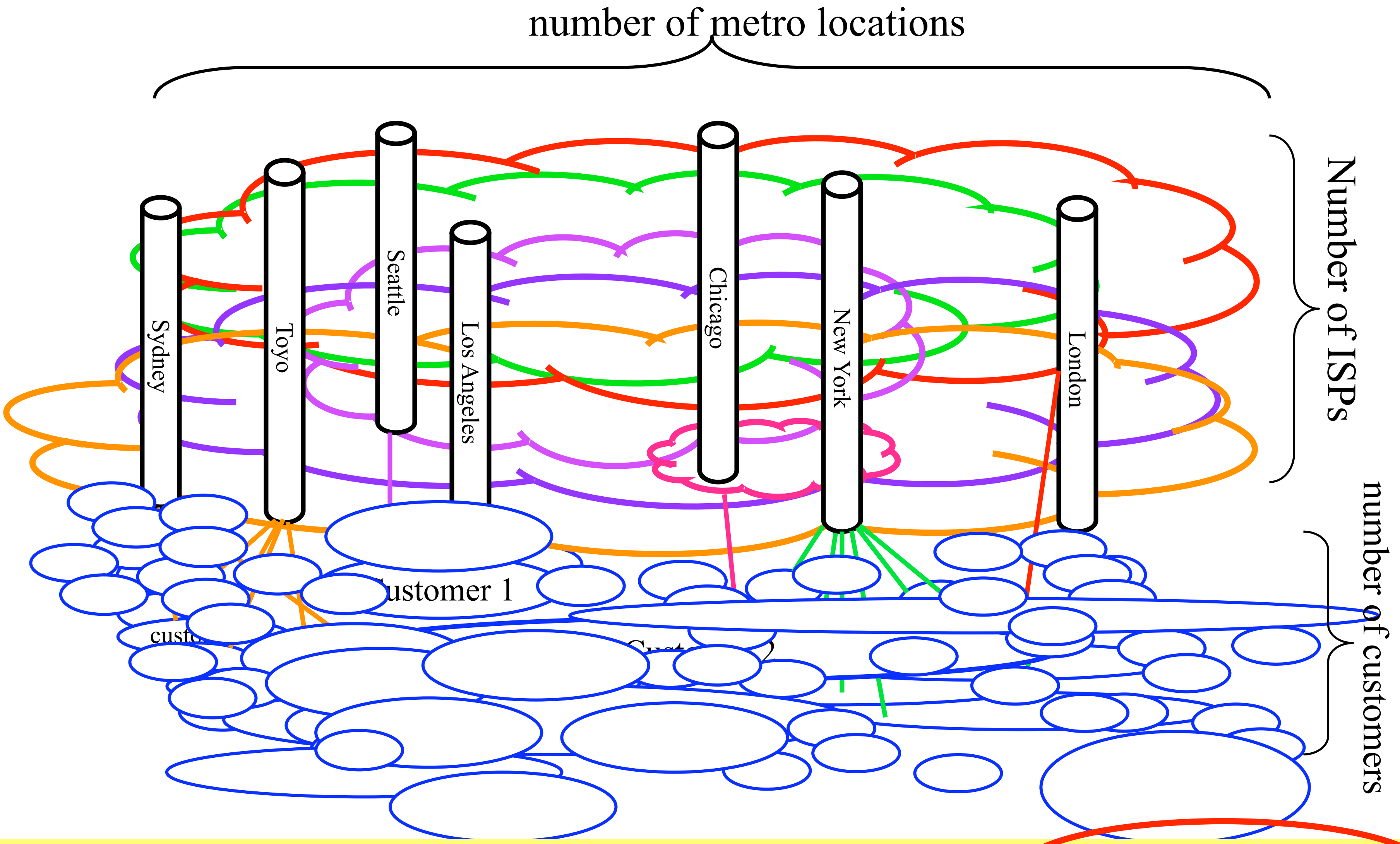


Multihoming and Traffic Engineering with IP V6

- What solutions exist or can exist in order to enable V6 multihoming and traffic engineering?
- Can we come to an IPV6 Multihoming and Traffic Engineering solution on a global scale?

RIRs have been handing out lots PI prefixes lately

How real topology looks like today



DFZ Routing table size = Function(# of ISPs X # of PoPs X # of user sites X TE)

Do We Have A Problem to Solve?

- “Now the vendors are producing or planning to produce greater capability and higher performance routers. Is it possible to produce routers with higher performance to meet the rapid routing table growth?”

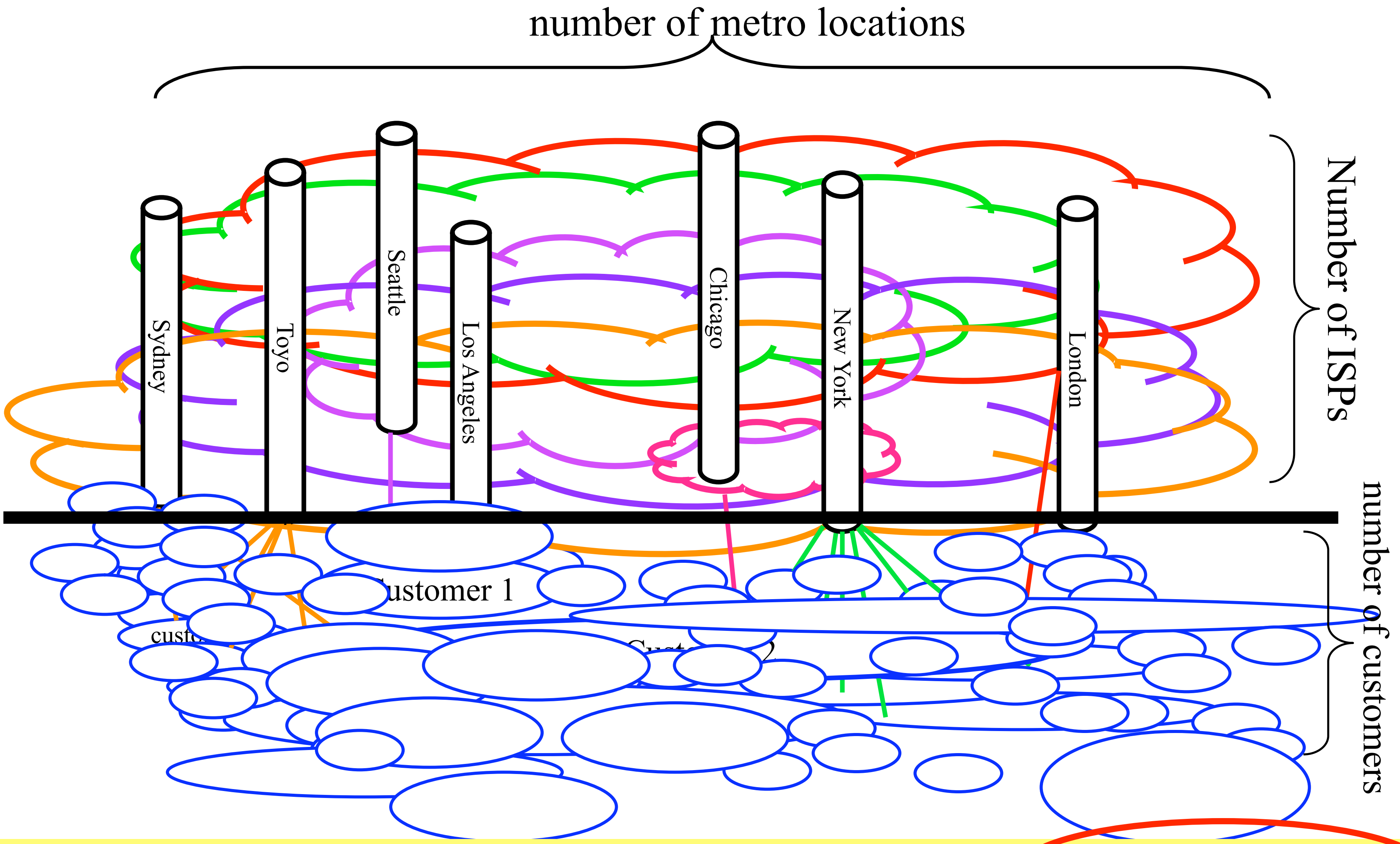
Do We Have A Problem to Solve?

- No: current router technology can support a routing table size at least 10 times bigger
  - And with Moore's Law, this number will continue to go up over time
- Yes: the global routing table size facing *uncontrolled* growth at super-linear speed
- Facts:
  - Current routing table size does not reflect demand
  - No glorous record in prediting the future
  - It is not just availability, it is affordability
    - Network economics
    - Power consumption: the ultimate limitation

Defining Scalable Routing

- Being able to control the scale of the routing system
  - The ability to control, rather than any specific numbers
- Allowing the global transit core to route on aggregatable prefixes only

Proposed solution:
Removing PI prefixes from global routing system



DFZ Routing table size = Function(# of ISPs X # of PoPs X ~~# of user sites X TE)~~

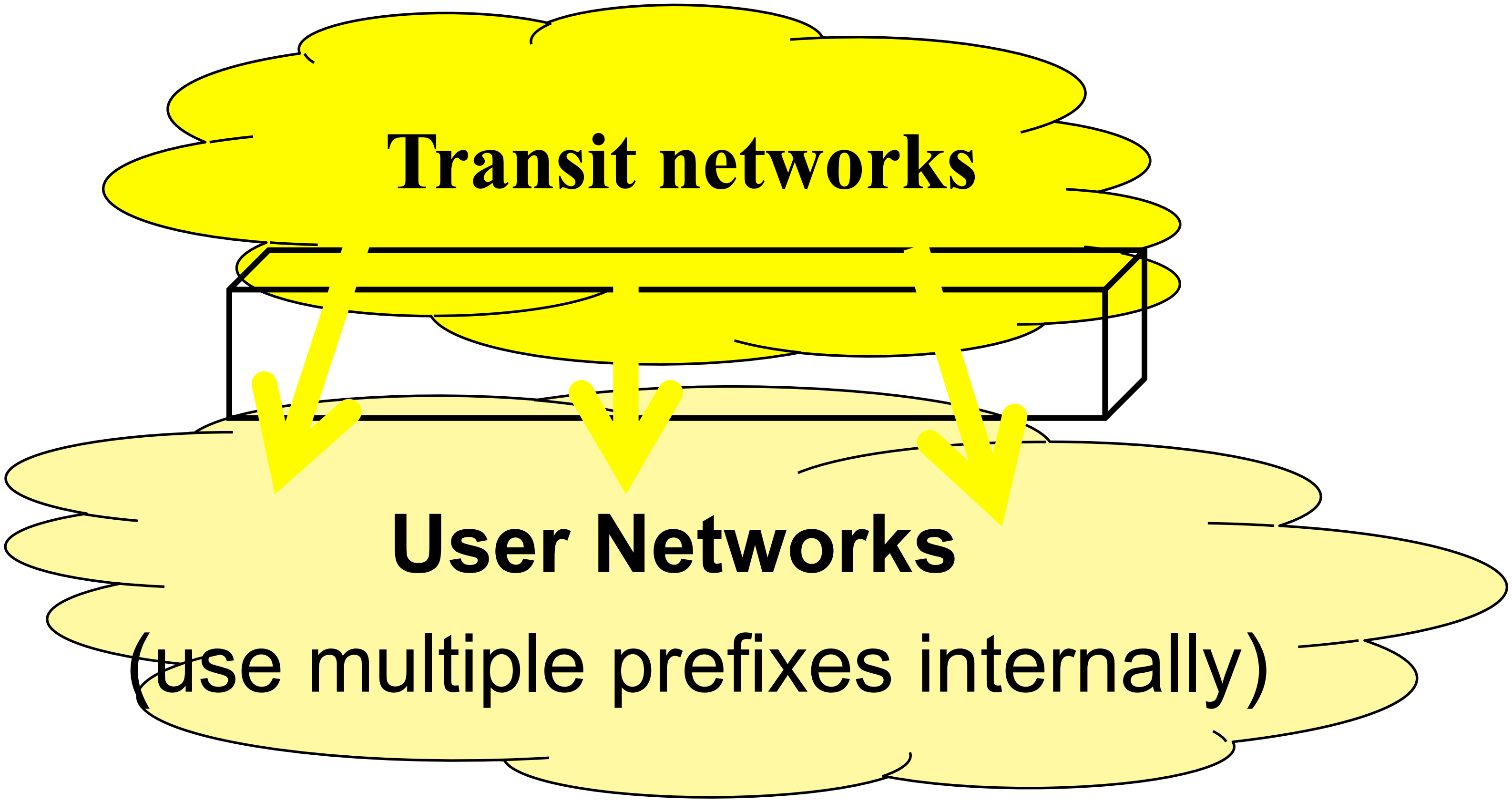
Scalable Routing: Solution Space

Two ways to get there

- **Elimination**: eliminating non-aggregatable PI prefixes from the entire Internet
- **Separation**: separating (removing) non-aggregatable PI prefixes from the global routing system

Eliminating Provider-Independent Address

- All user sites take PA addresses
  - multihomed sites take multiple PA addresses

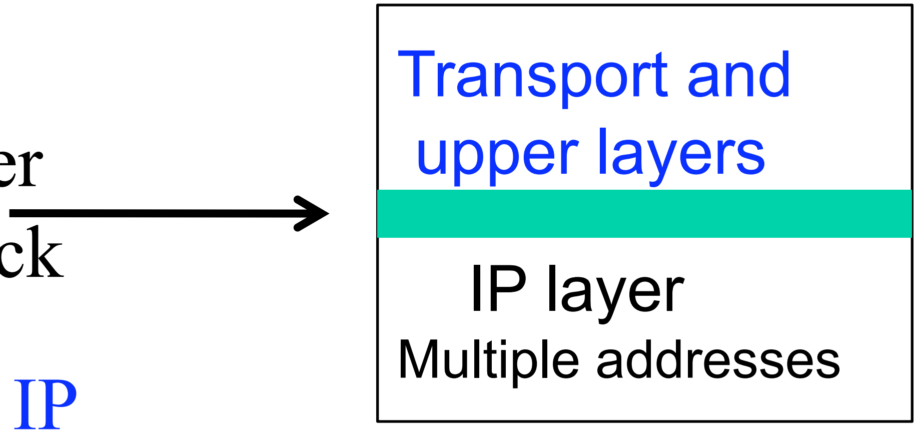


2 proposed solutions on the table

Elimination Solution 1

- **SHIM6:** Multiple PA addresses stop at shim layer in a host
  - Lots of hard work has been done here

adding a shim layer
to the protocol stack

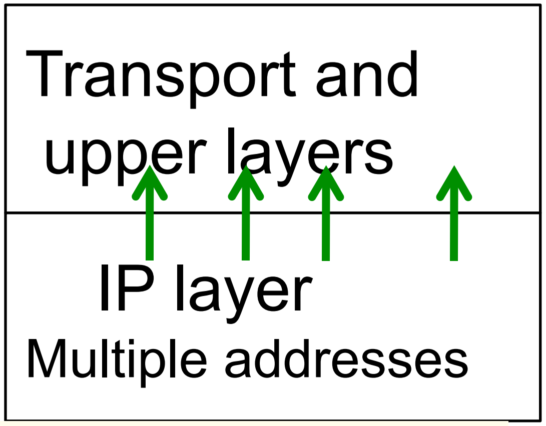


Choose one of the IP addresses to be used by upper layer

Elimination Solution 2

- **Multipath transport**: Push multiple PA addresses all the way up to transport layer
  - See Mark Handley's talk at last RRG meeting: "Multipath Transport, Resource Pooling, and implications for Routing"

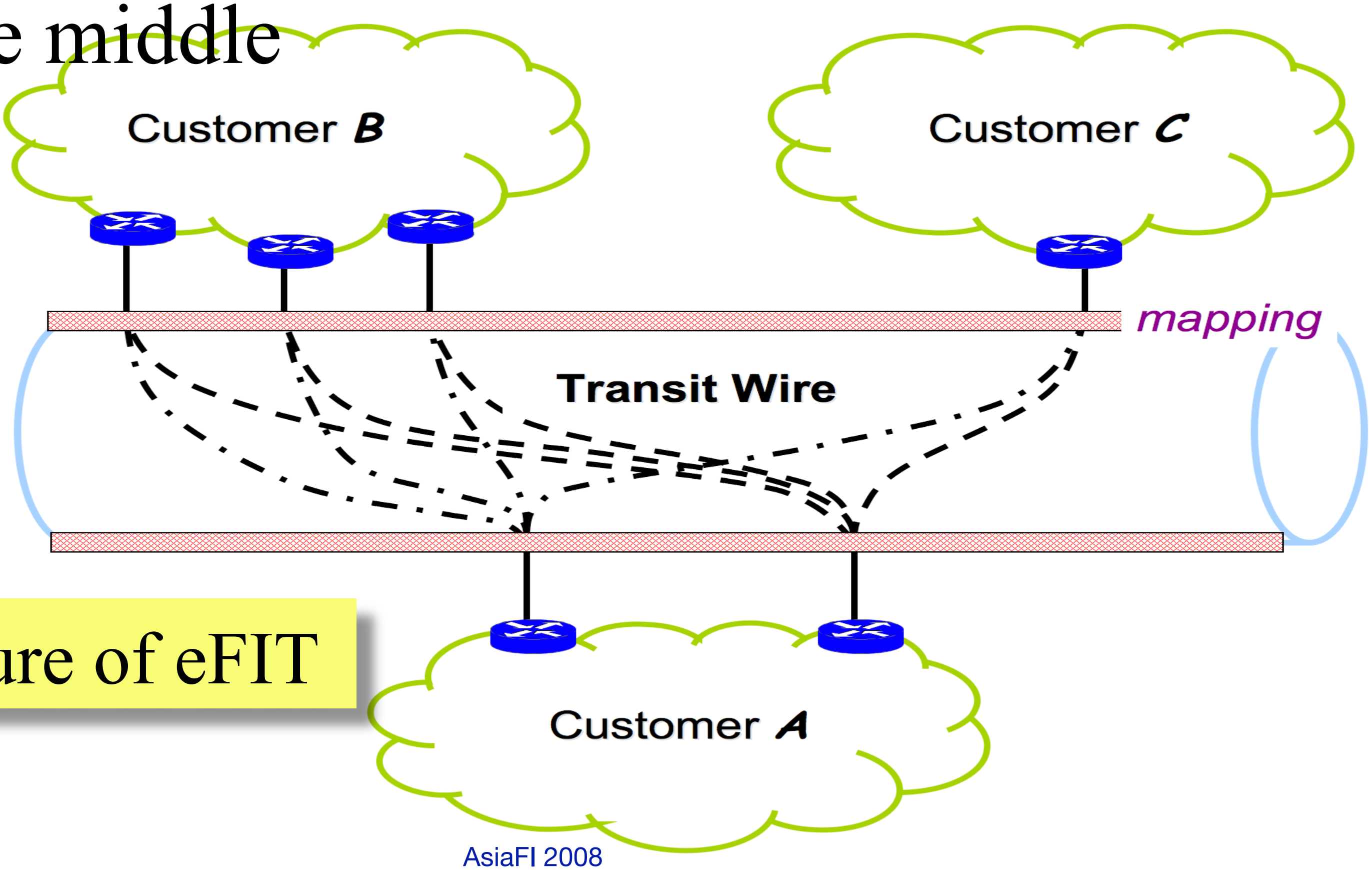
http://www3.tools.ietf.org/group/irtf/trac/wiki/RRGagendaDublin



TCP may use multiple parallel paths simultaneously to improve throughput, delay, and robustness

Separating edge prefixes from transit core

- Map & Encap: A number of proposed solutions
  - APT, IVIP, LISP, TRRP
- Requires a mapping system to glue the edges through the middle



The basic picture of eFIT

First question: Elimination, or Separation?

If elimination:

- No new work need to be done at network layer
- however there is a conservation of hard work
  - New designs to effective utilize multiple parallel paths via multiple addresses by host/transport
  - Changes to all hosts
  - Site renumbering whenever changing providers
- Need to be effective in controlling routing table growth
  - Incentives for majority of user sites to deploy the new design within some finite time period

Elimination, or Separation?

If separation:

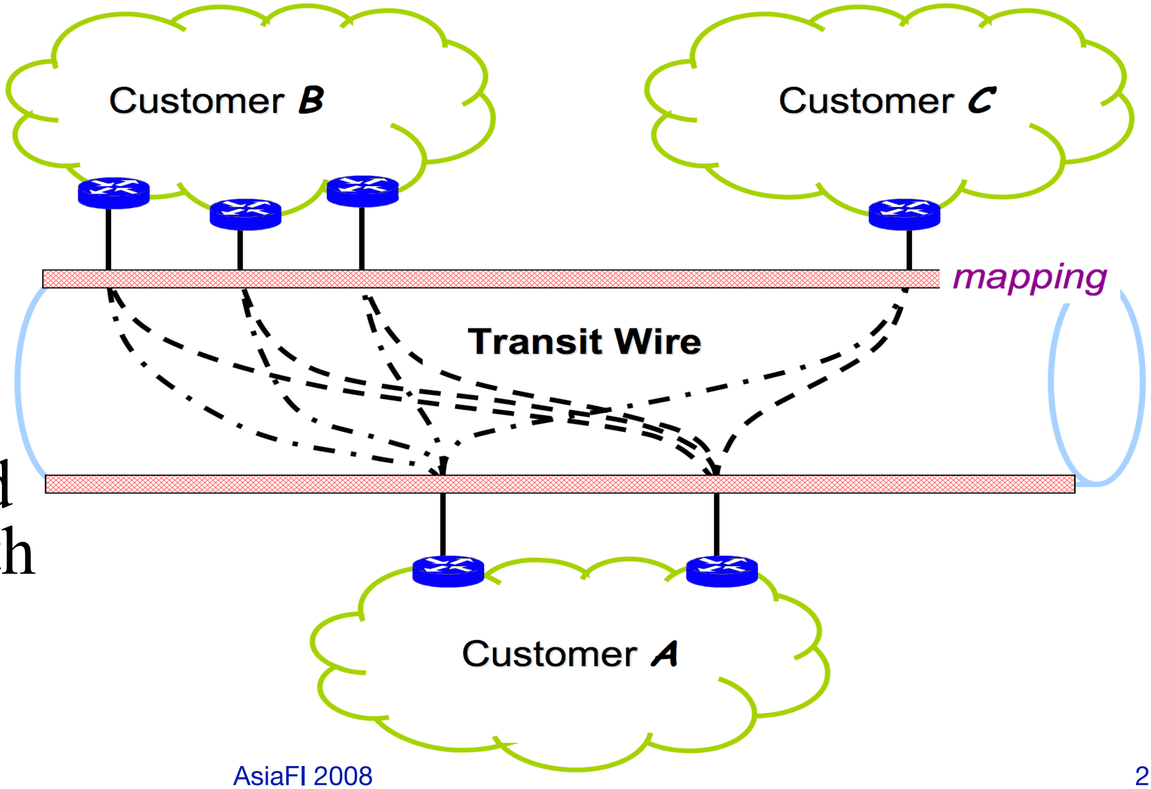
- Need to work out a mapping system design
  - Map an edge destination address to the edge network’s attachment point to the transit core
  - Mapping info must be distributed to all entry points to the core
- Need to design effective detection and recovery mechanism for failures occuring between the core and edge networks
- Need an effective incemental deployment strategy
  - The benefits to first movers should offset the cost

Aside: Why new issue in failure recovery?

- Today: B injects its prefix into global routing system, it can be reached as long as at least one of its 3 attachment points works

With Map&Encap:

- B’s prefix no longer in global routing, but in the mapping table
- Propose not to update mapping system by transient failure
- Require solutions that can detect failures and switch to alternate path promptly, if any is available



Elimination vs Separation: Which way to go?

- Some people believe all hosts can be changes within reasonable time frame
  - Assuming the multipath transport solution getting developed quickly
- Some people believe renumbering is a nonstarter
- The real answer: The future is uncertain

If we choose elimination

- And indeed all edge networks will take in PA addresses in next 5 years
- We would reach the goal of scalable routing without working hard!
  - Of course transport people will work hard to roll out multipath transport, and
  - Sites will have to adopt multiple-addressing and renumbering
- But what if we guessed wrong?

  In next 5 years
  - IPv4 routing table will continue to grow
  - IPv6 deployment would progressing
  - We could be facing real serious routing scalability crisis...

If we choose separation

- We will have to work really hard to solve the three major challenges
- If we choose wrong: all the hard work would be wasted!
  - But we don’t do any worse than that
- If we choose right: the hard work will be worthwhile
  - Resolving a decades long problem

See ftp://ftp.ietf.org/ietf-online-proceedings/95jul/presentations/allocation/pre.allocation.txt

IETF33 Plenary on IP Address Allocation
(July 1995)

- up to now, the IP address has served as an invariant, unique identification for the end host. TCP design makes use of this assumption, so do many other protocols and applications.
- As a result, nobody today has a complete list of all the possible places in the protocol architecture that have the IP address hard wired or embedded in it.
- Therefore, contradicting Peter(Ford)'s assumption that most customers do not care about permanent IP addresses, dynamically changing addresses, as required by provider-based assignment, changes the architecture we used to know and causes serious problems at the user ends.

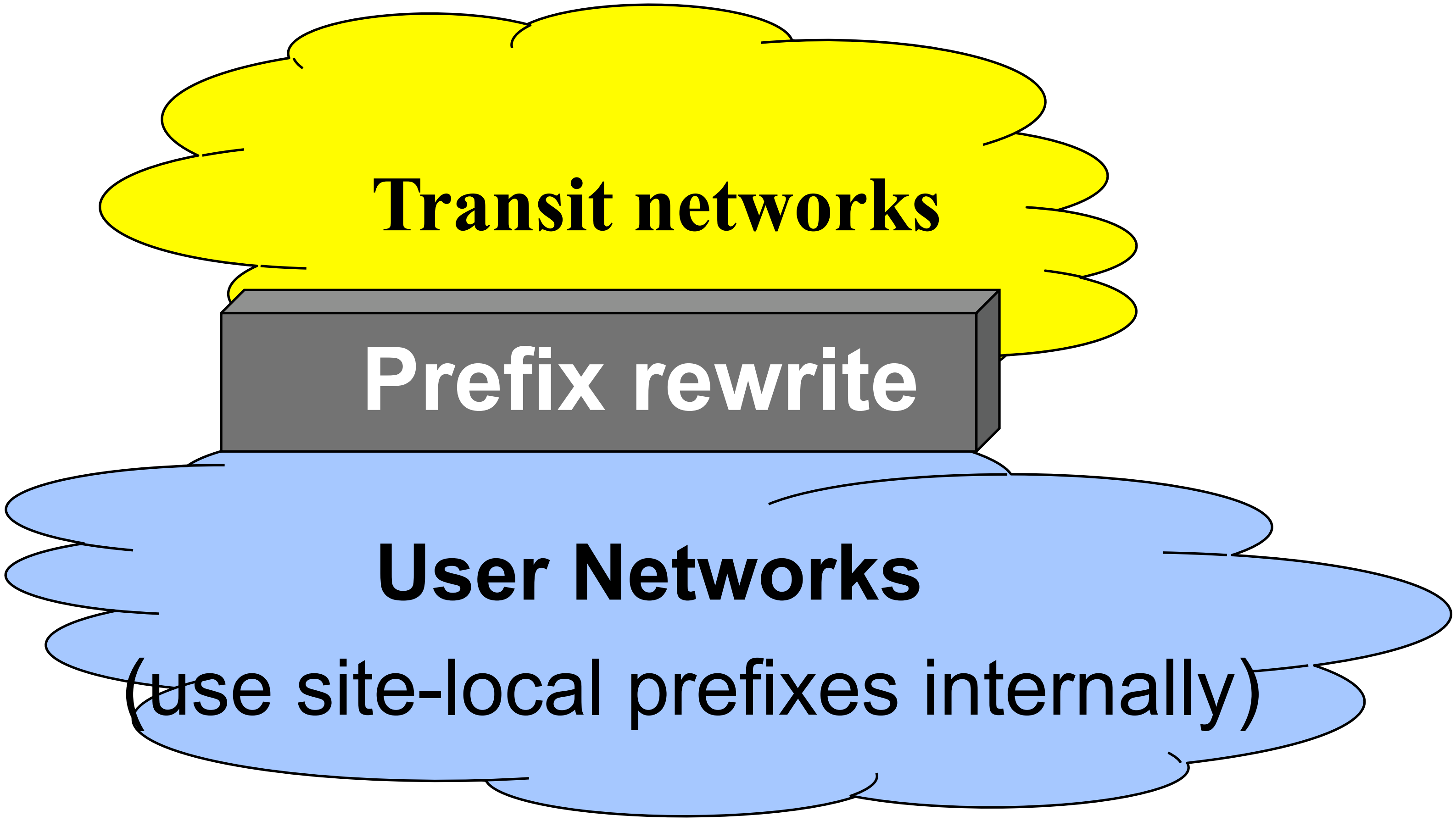
The conclusion?

Although the jury is still out (regarding whether multipath transport + site renumbering will give us effective control over global routing system growth within next few years)

The action to take <u>now</u> is clear: developing an effective and efficient separation solution to global routing scalability problem

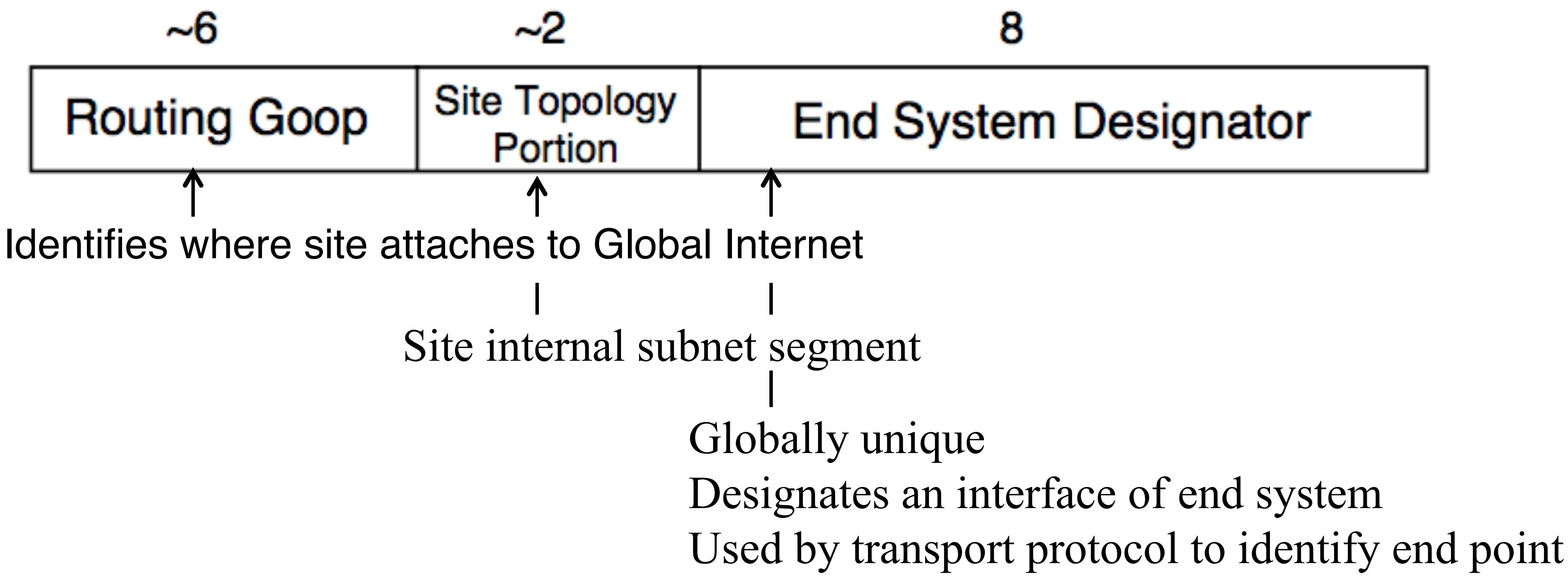
An early proposal for separation: GSE

GSE: Global, Site, and End-system address elements (also called 8+8)



1. Address Prefix Rewriting: GSE

Proposed IPv6 address structure:



~6 ~2 8

Routing Goop | Site Topology Portion | End System Designator

Identifies where site attaches to Global Internet

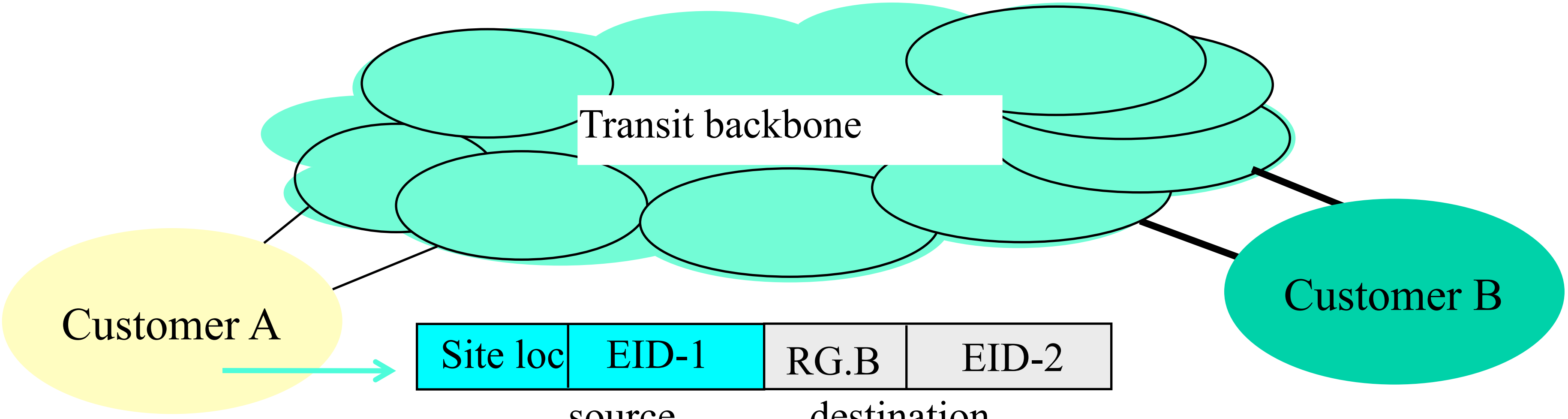
Site internal subnet segment

Globally unique
Designates an interface of end system
Used by transport protocol to identify end point

- Multihomed sites get multiple RGs
- Internal packets: use site-local RG
- External packets: defer/hide external RGs

How GSE Works

- For outbound traffic:
  - Get destination address and RG from DNS lookup
  - Put on source RG when packets exiting local site

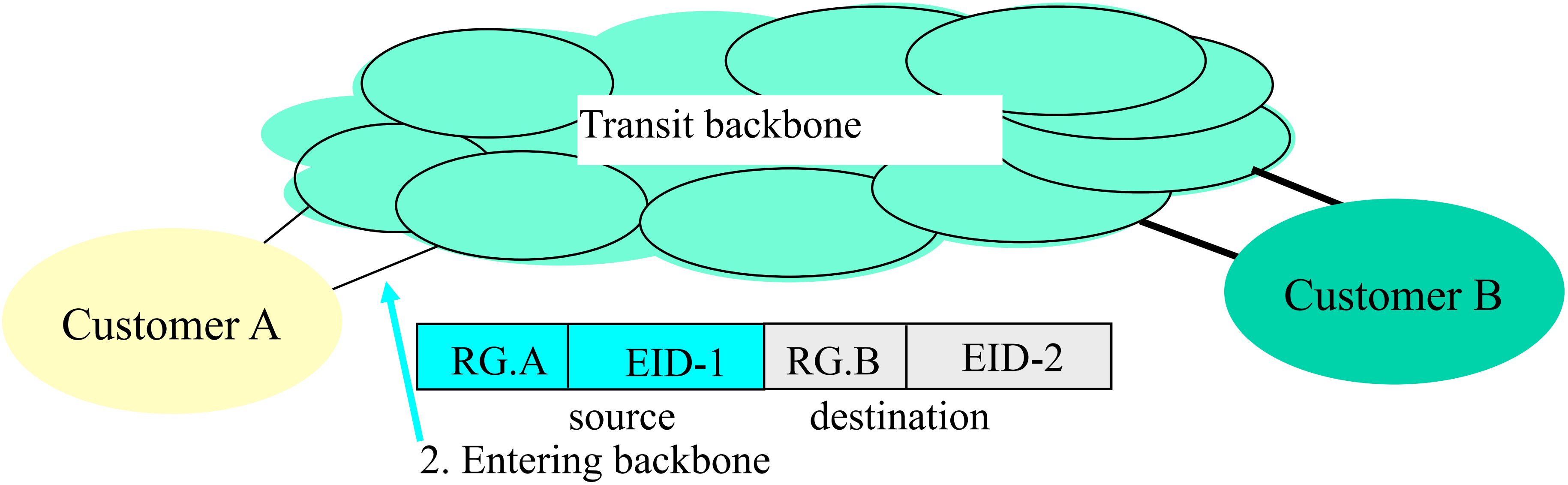


source		destination	
Site loc	EID-1	RG.B	EID-2

1. Inside A

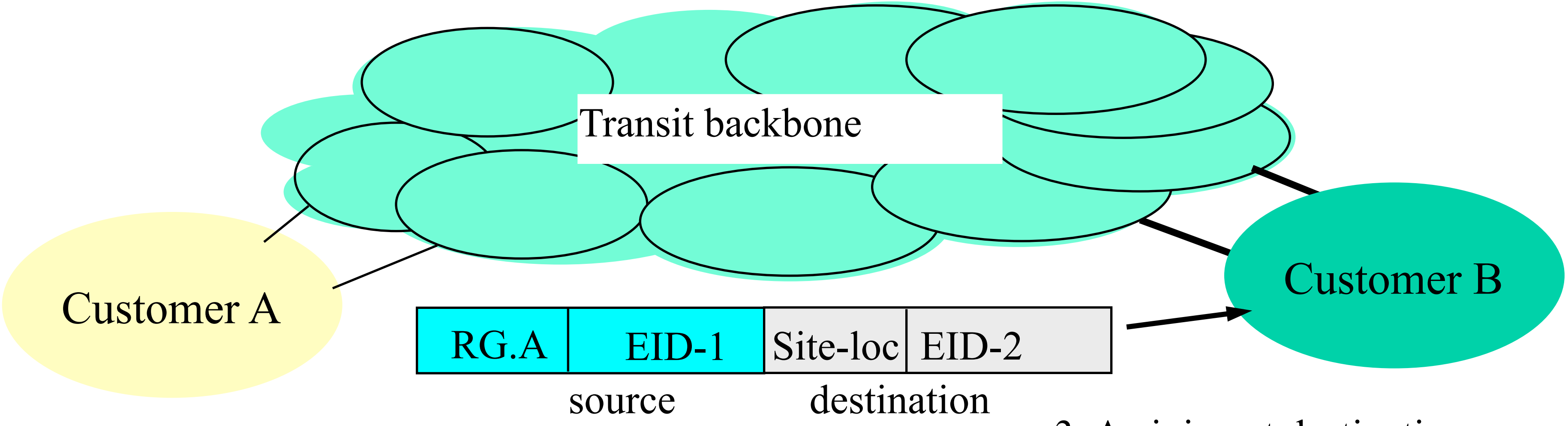
How GSE Works

- For outbound traffic:
  - Get destination address and RG from DNS lookup
  - Put on source RG when packets exiting local site



How GSE Works

- For outbound traffic:
  - Get destination address and RG from DNS lookup
  - Put on source RG when packets exiting local site
- For inbound traffic
  - Take off destination RG at entrance to destination site
  - Keep source RG for returning traffic



Transit backbone

Customer A

Customer B

RG.A	EID-1	Site-loc	EID-2

source

destination

3. Arriving at destination

What Problems GSE Solved

- Making customer sites unaware of the transit backbone or provider change
    - Eliminate renumbering caused by change of providers
- Providing ISPs freedom for performing aggregation as needed in the provider space

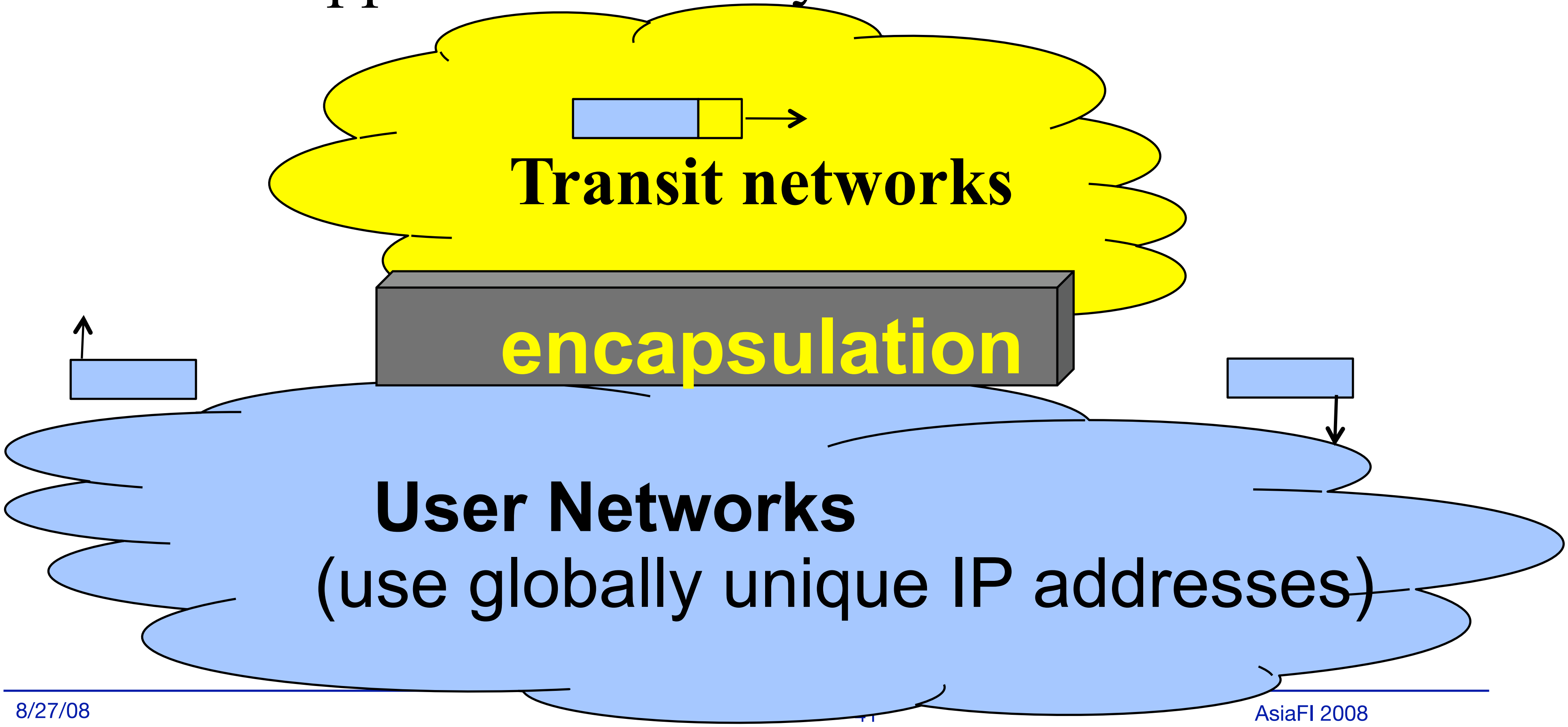
But major Issues Left Open

- Inbound traffic engineering: which destination RG to put on each packet?
- Is 64-bit long enough for globally unique ID space?
    - How is that space managed, and more importantly, enforced?
- Issues from using site-local prefixes
- **Failure recovery?**
- **Incremental deployment?**

See my GSE review article:
http://www.isoc.org/tools/blogs/ietfjournal/?p=98#more-98

A more promising solution: Map-n-Encap

- Keep the user packets intact
- No changes to user sites
- Can be applied recursively if needed



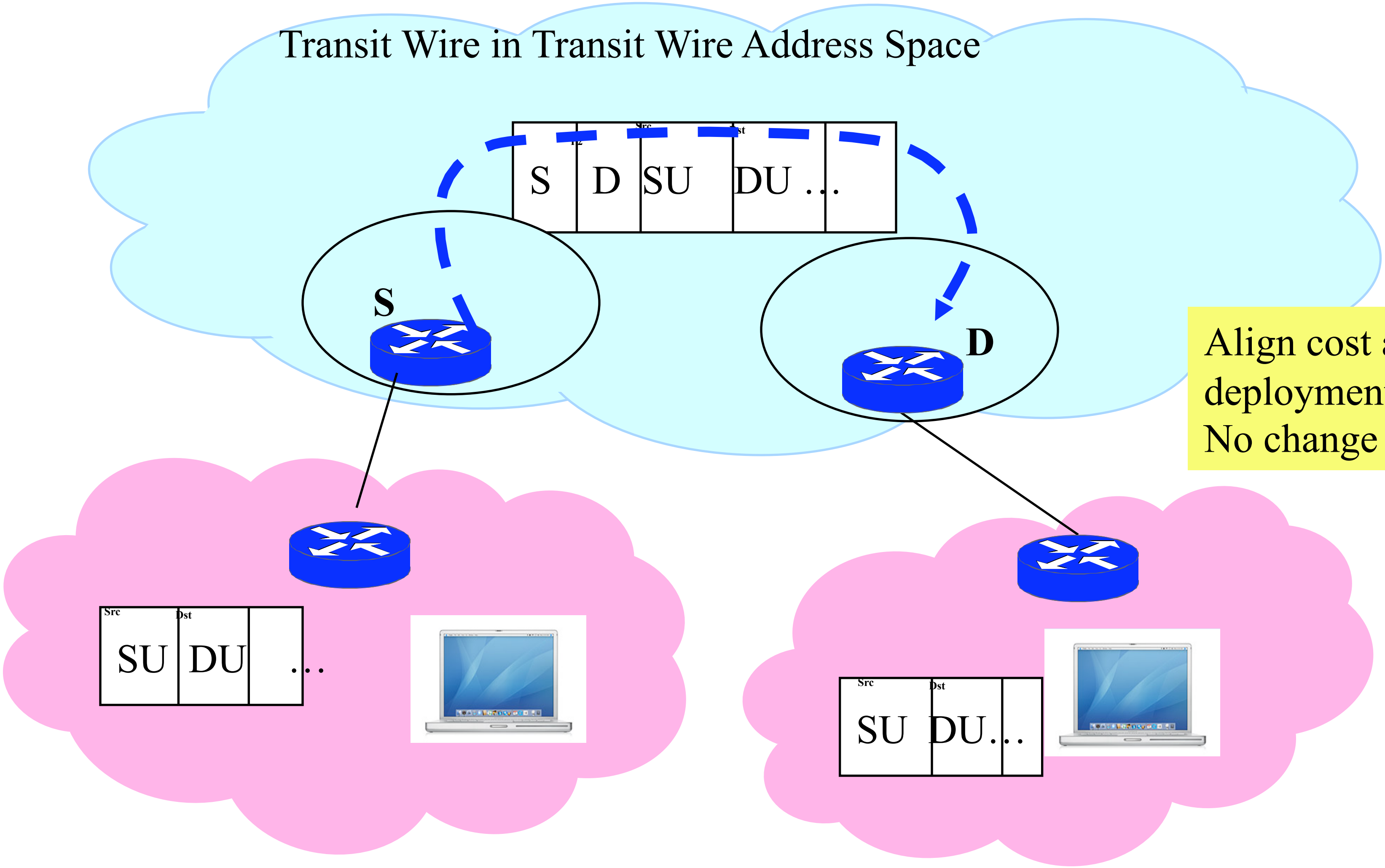
Transit networks

encapsulation

User Networks
(use globally unique IP addresses)

APT: A Practical Transit-Mapping Service

Transit Wire in Transit Wire Address Space



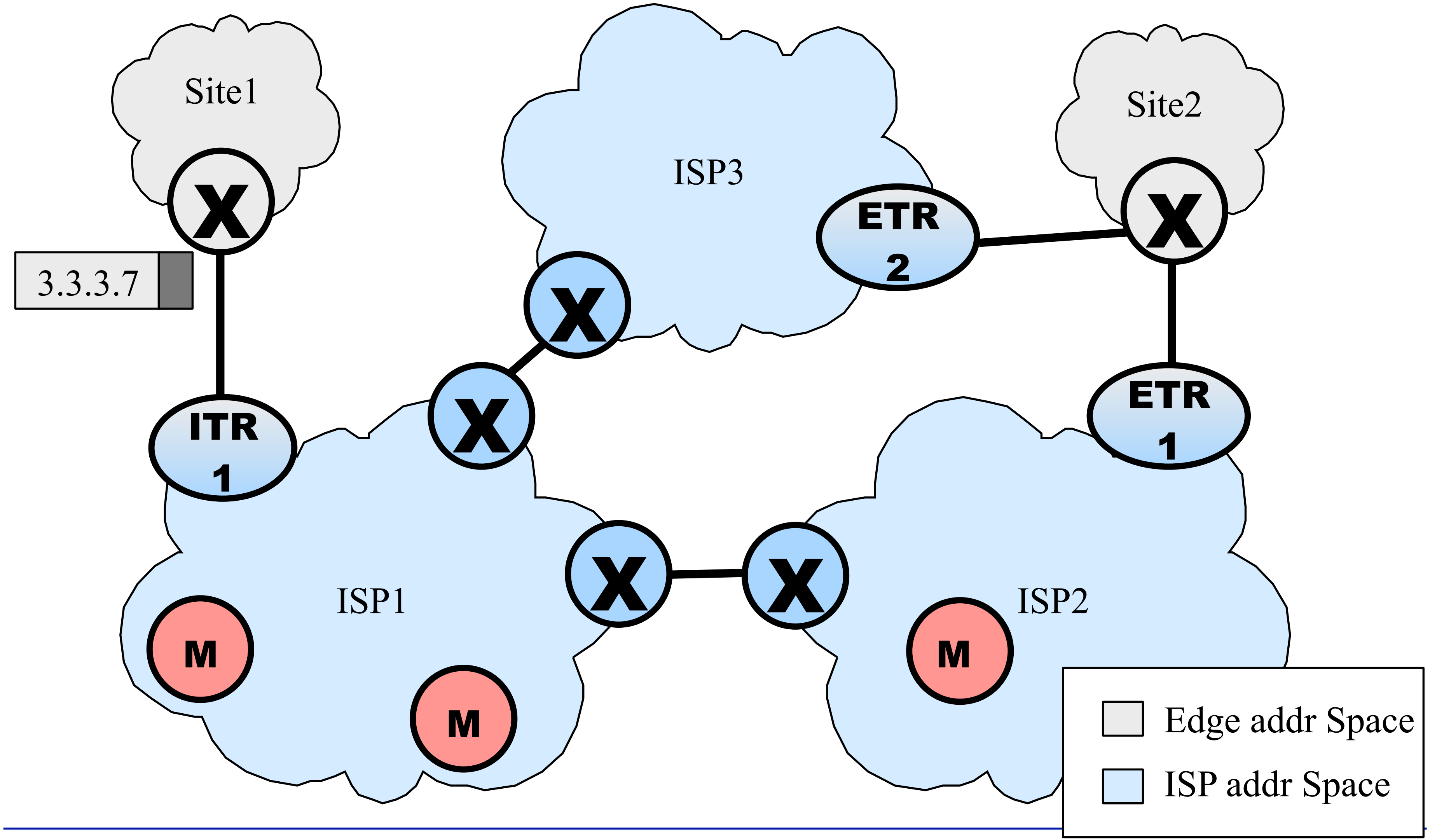
Align cost and benefit:
deployment inside ISP
No change to edge site

Source and Destination Networks in User Address Space

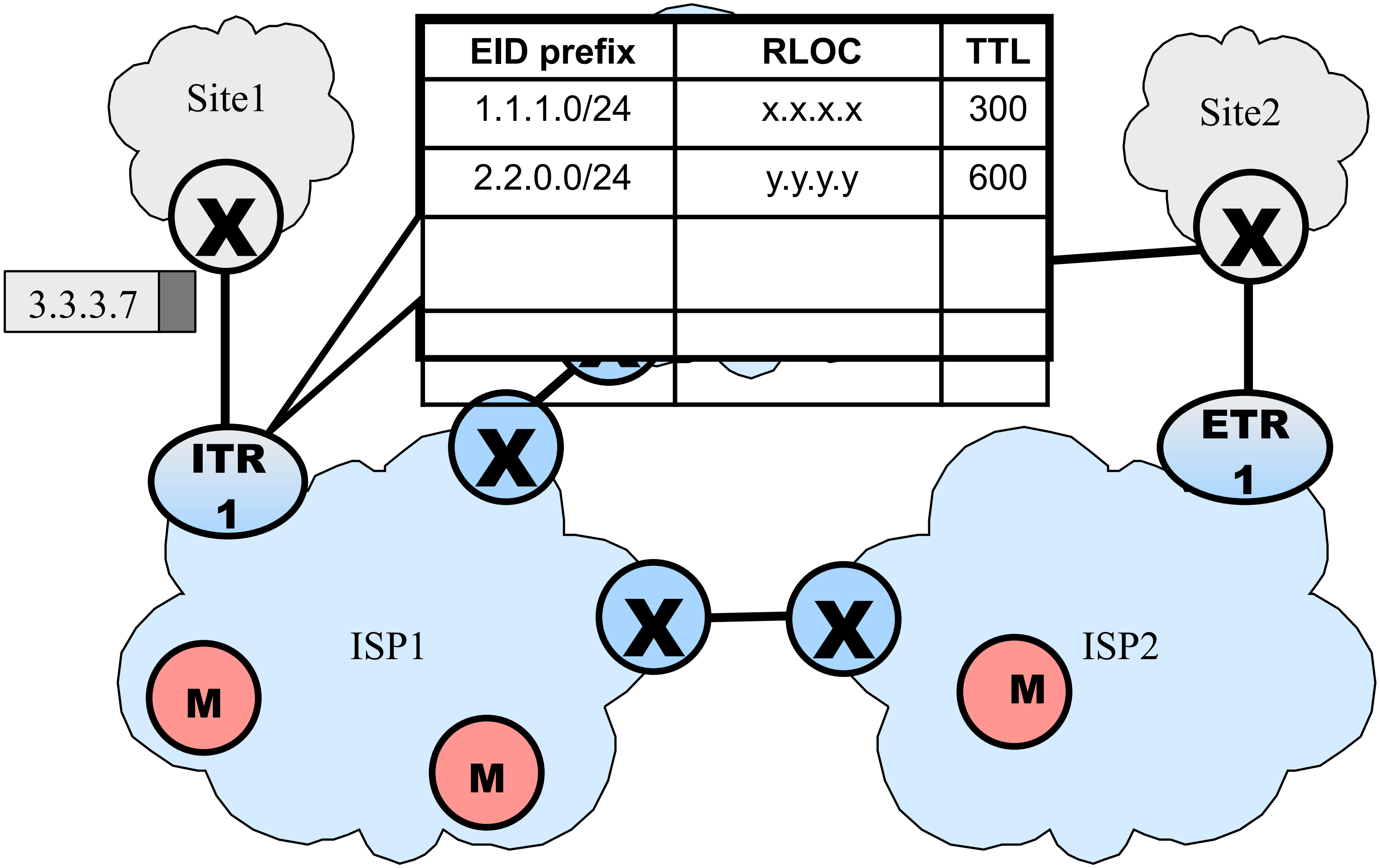
Challenge 1: Mapping information distribution

- Principle: no dependency between ASes regarding mapping information availability
- Basic idea:
  - Each user site provides its ISPs with mapping info
  - ISPs use flooding to distribute mapping info globally
    - Also looking into alternative starting point
- Individual ASes decide how to make mapping info available to all its edge routers
- basic design:
  - each AS runs a few default mappers with full mapping table
  - edge routers retrieve the info as needed

APT Example

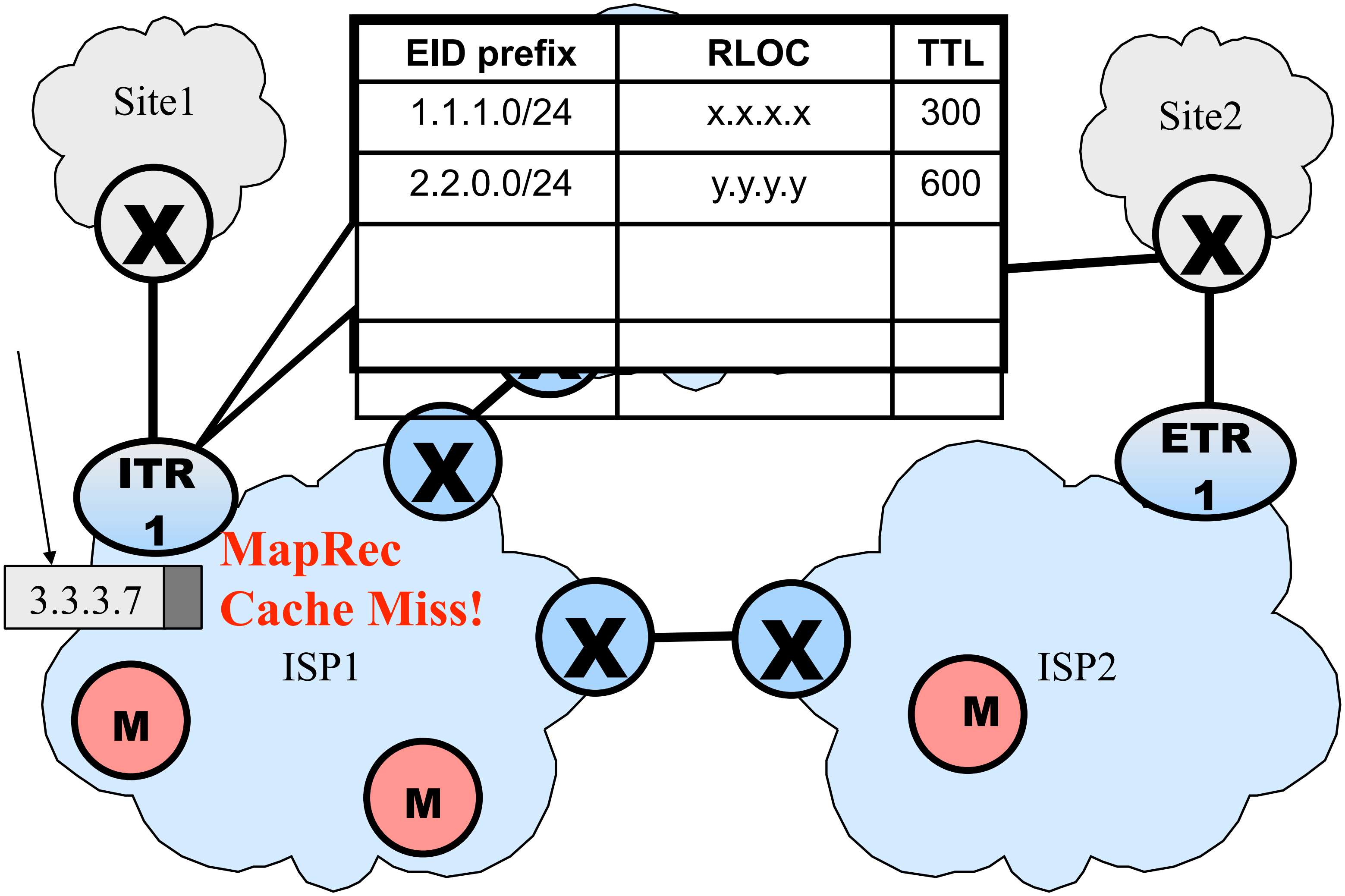

Site1
X
3.3.3.7
ITR 1
ISP3
X
X
ETR 2
Site2
X
ETR 1
ISP1
X
X
ISP2
M
M
M
Edge addr Space
ISP addr Space

APT Example



EID prefix	RLOC	TTL
1.1.1.0/24	x.x.x.x	300
2.2.0.0/24	y.y.y.y	600

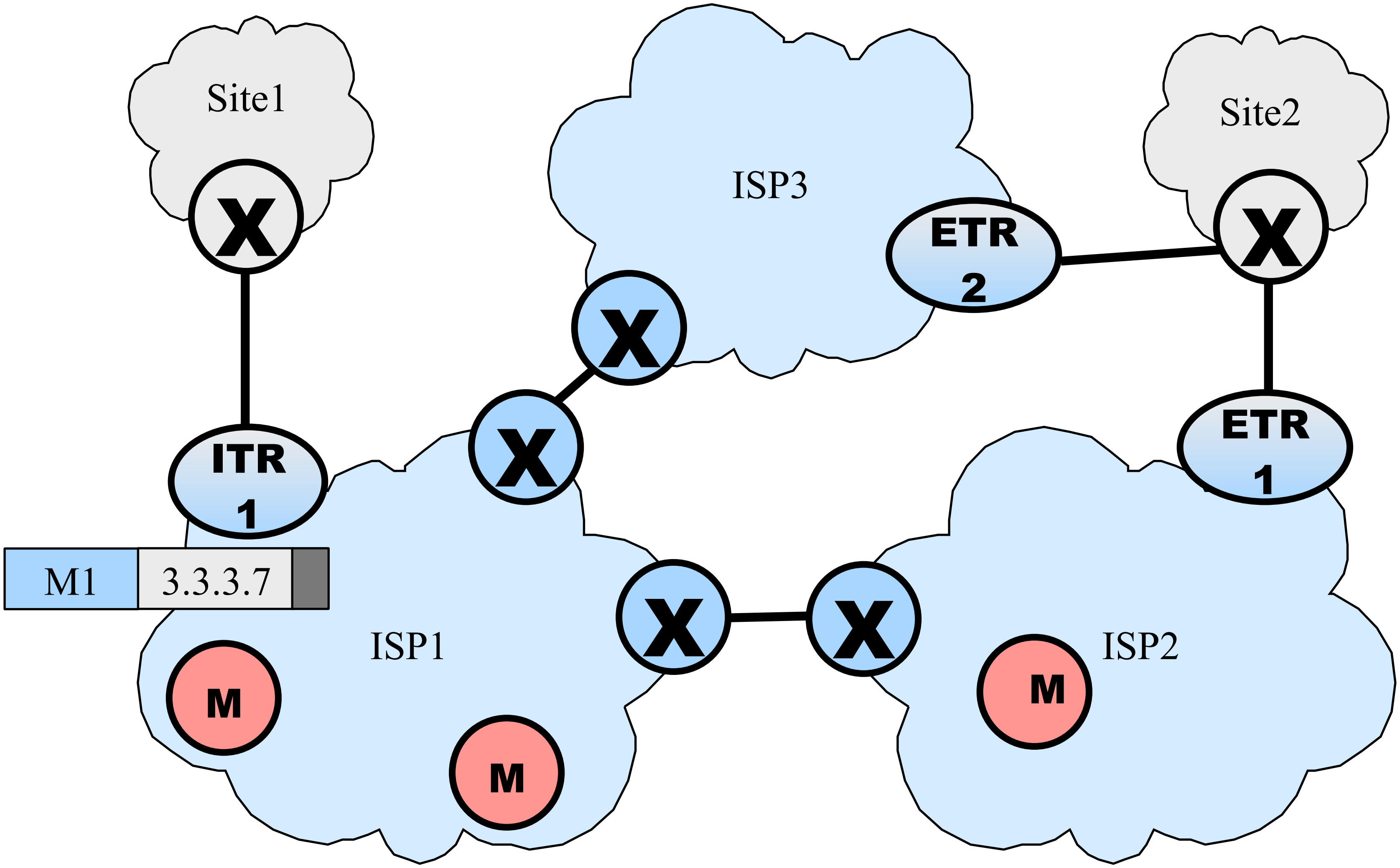
MapRec Not in Cache



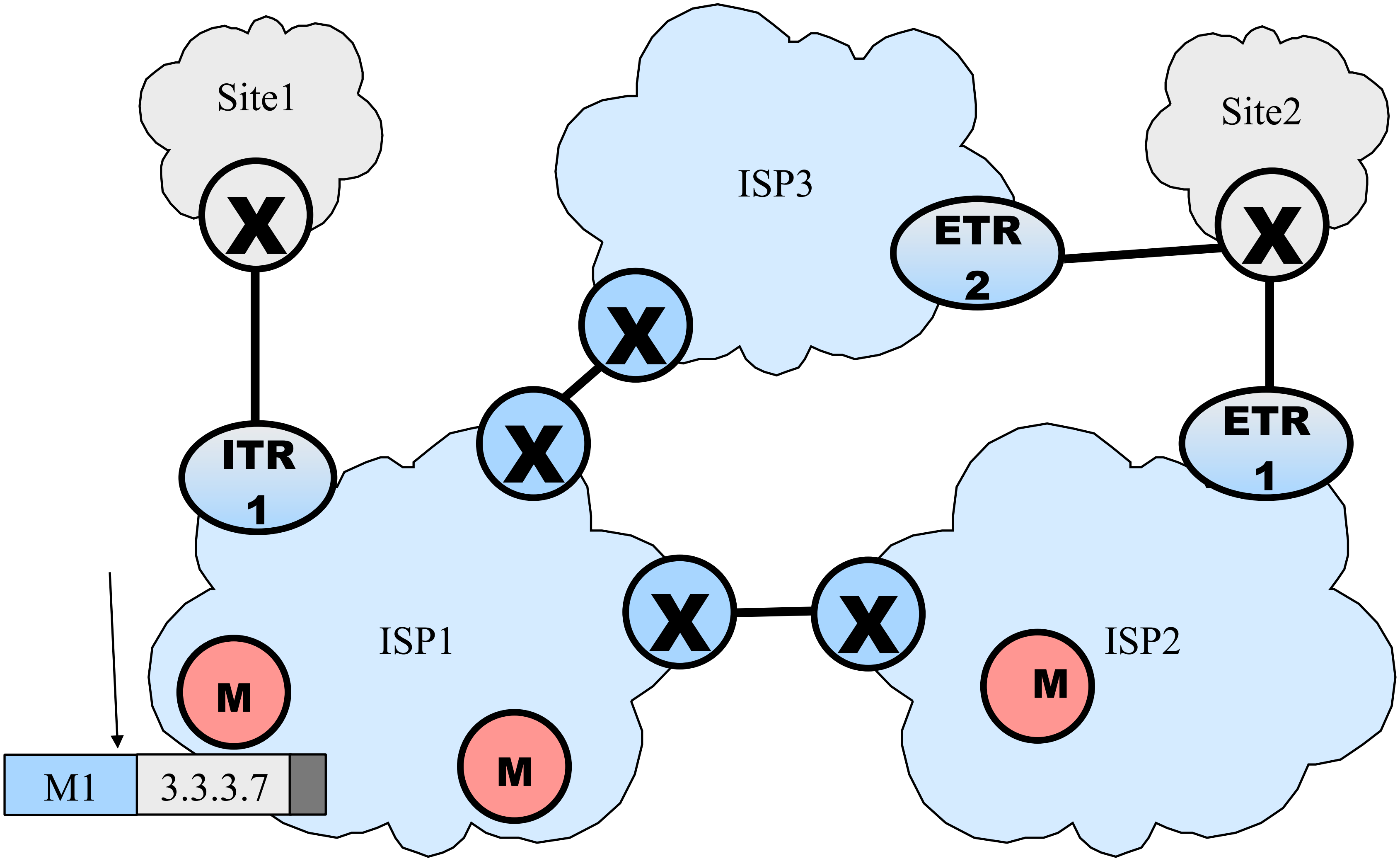
EID prefix	RLOC	TTL
1.1.1.0/24	x.x.x.x	300
2.2.0.0/24	y.y.y.y	600

MapRec
Cache Miss!

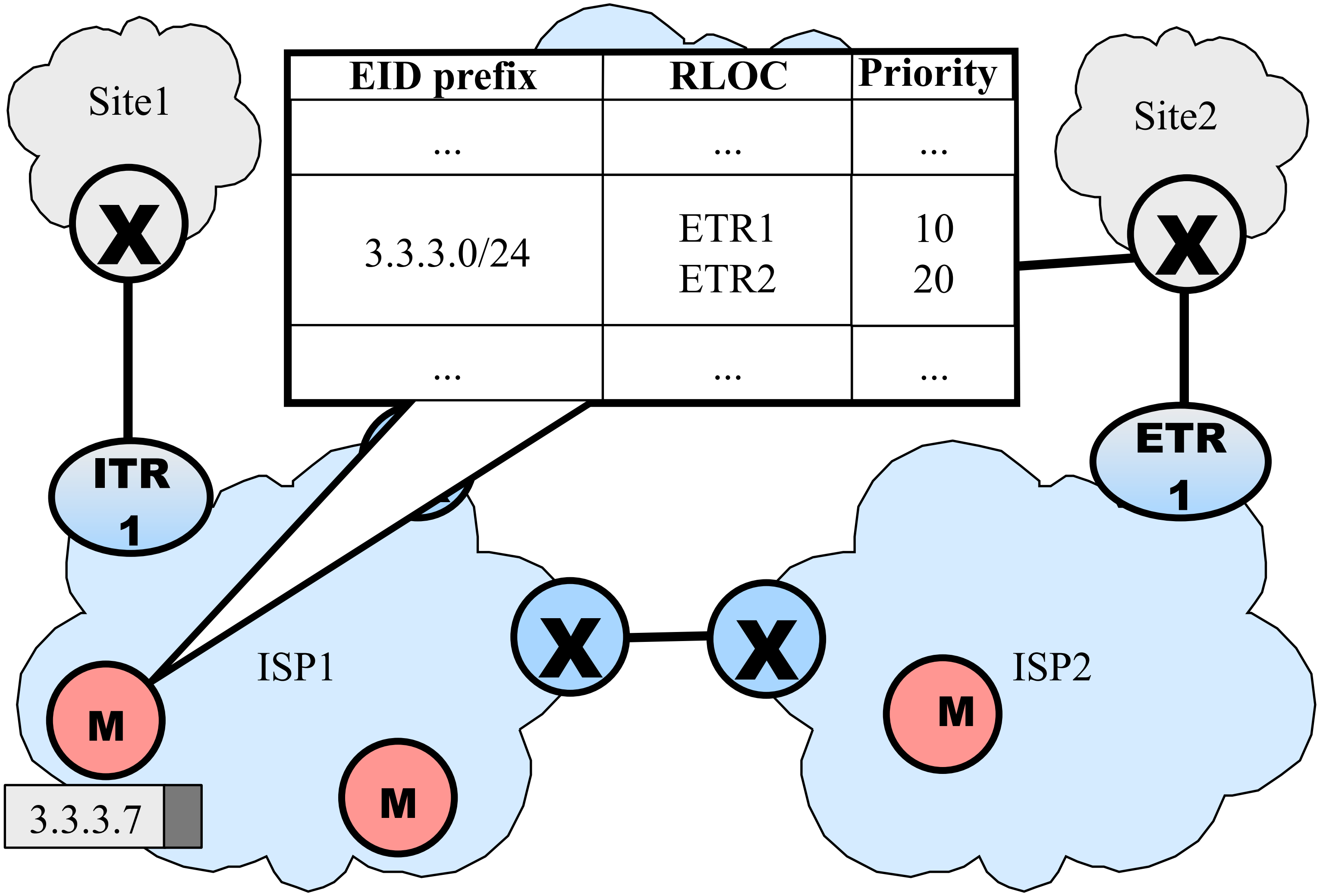
Encap with the Default Mapper Anycast Address



Default Mapper Decaps the Packet

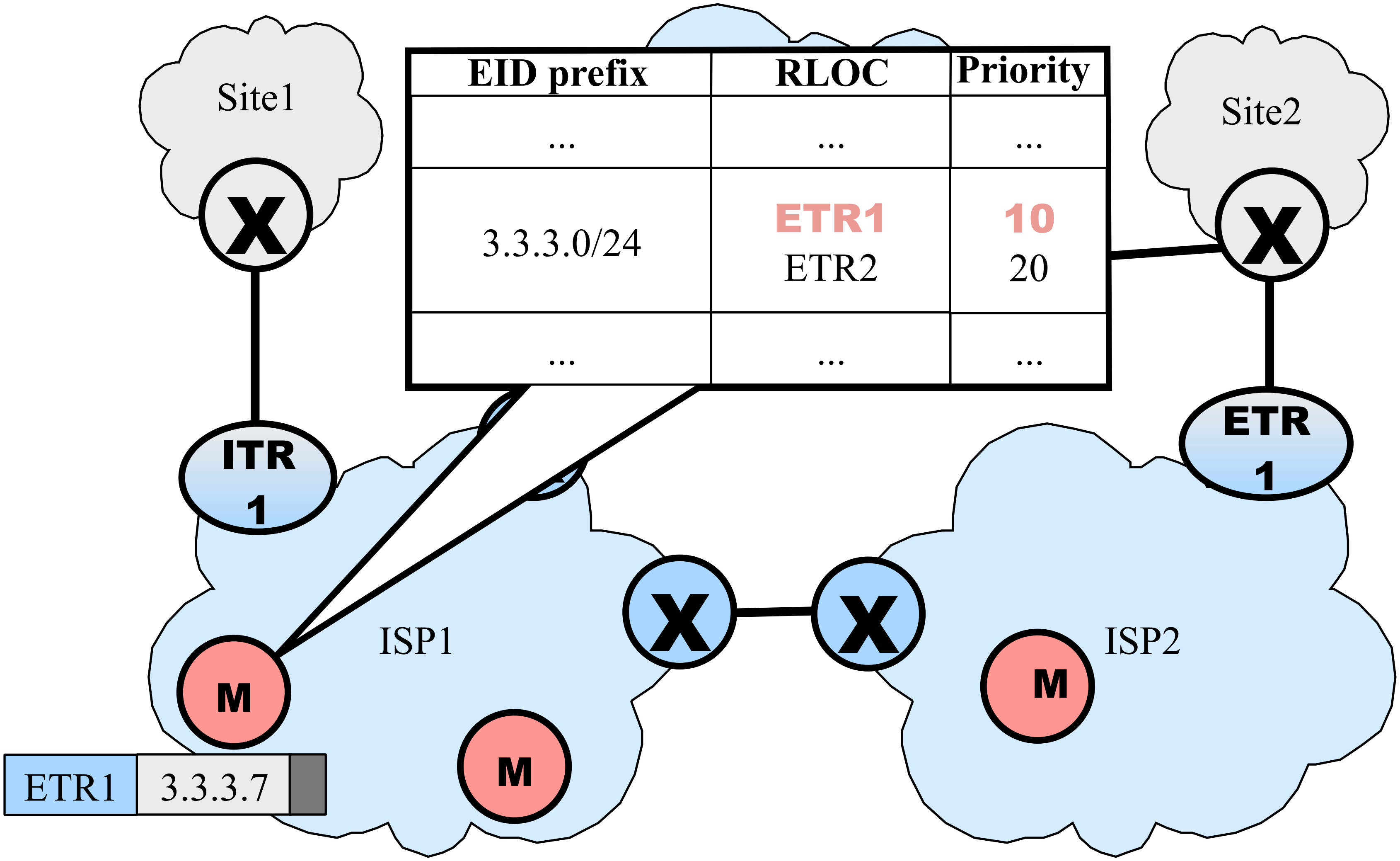


EID Prefix is Multihomed



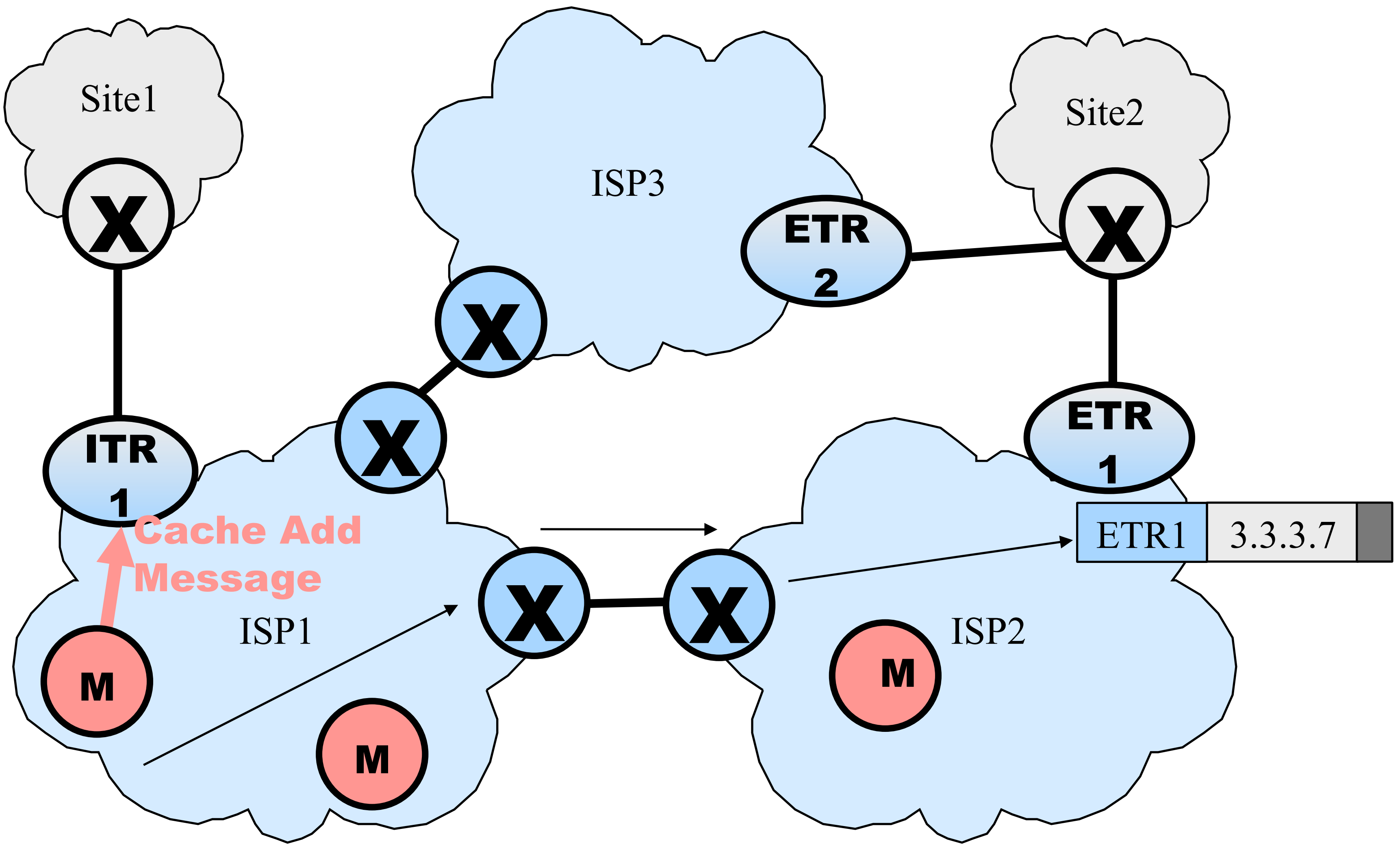
EID prefix	RLOC	Priority
…	…	…
3.3.3.0/24	ETR1 ETR2	10 20
…	…	…

Default Mapper Selects a MapRec

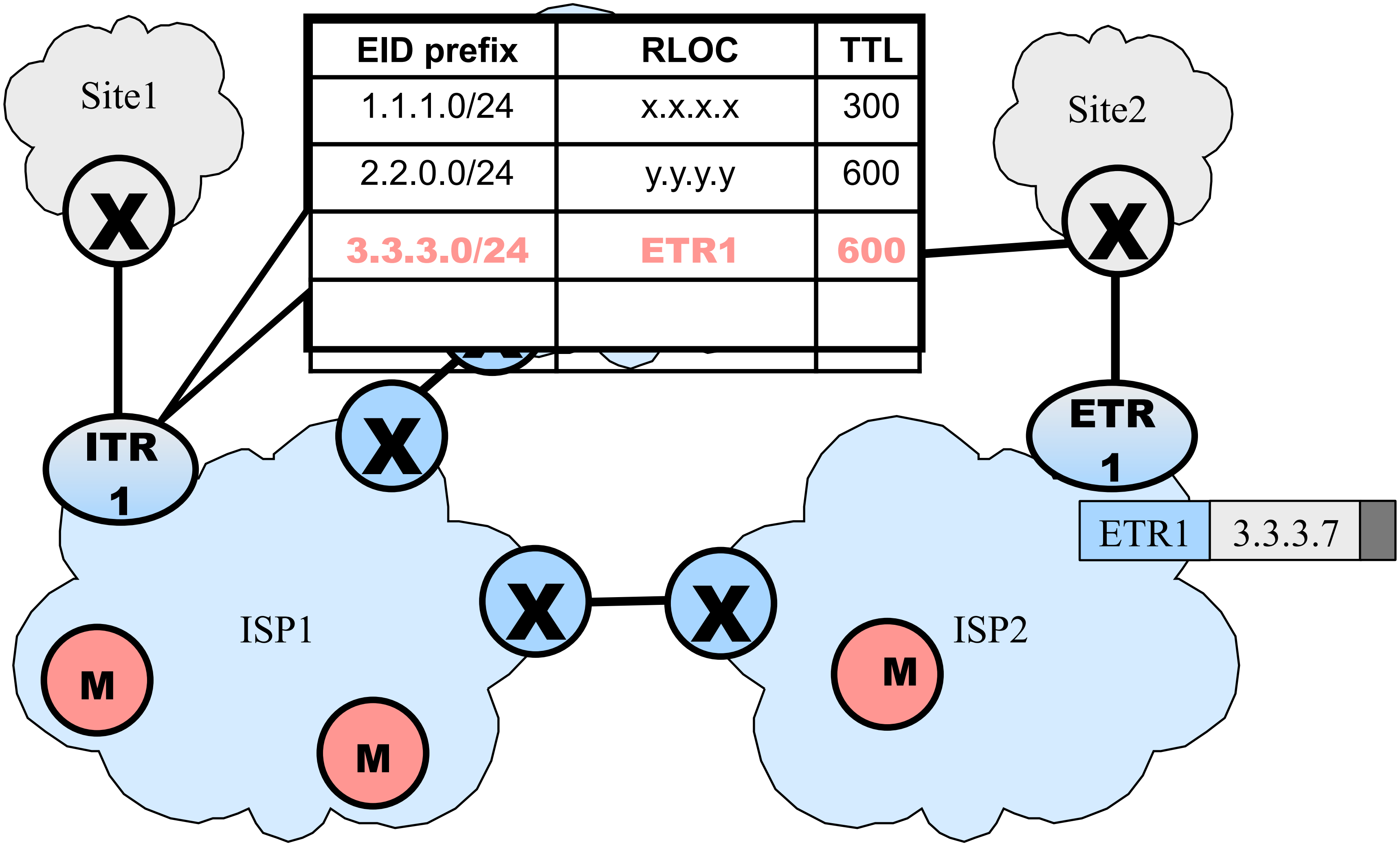


EID prefix	RLOC	Priority
…	…	…
3.3.3.0/24	ETR1 ETR2	10 20
…	…	…

Default Mapper Responds with MapRec and Delivers Packet

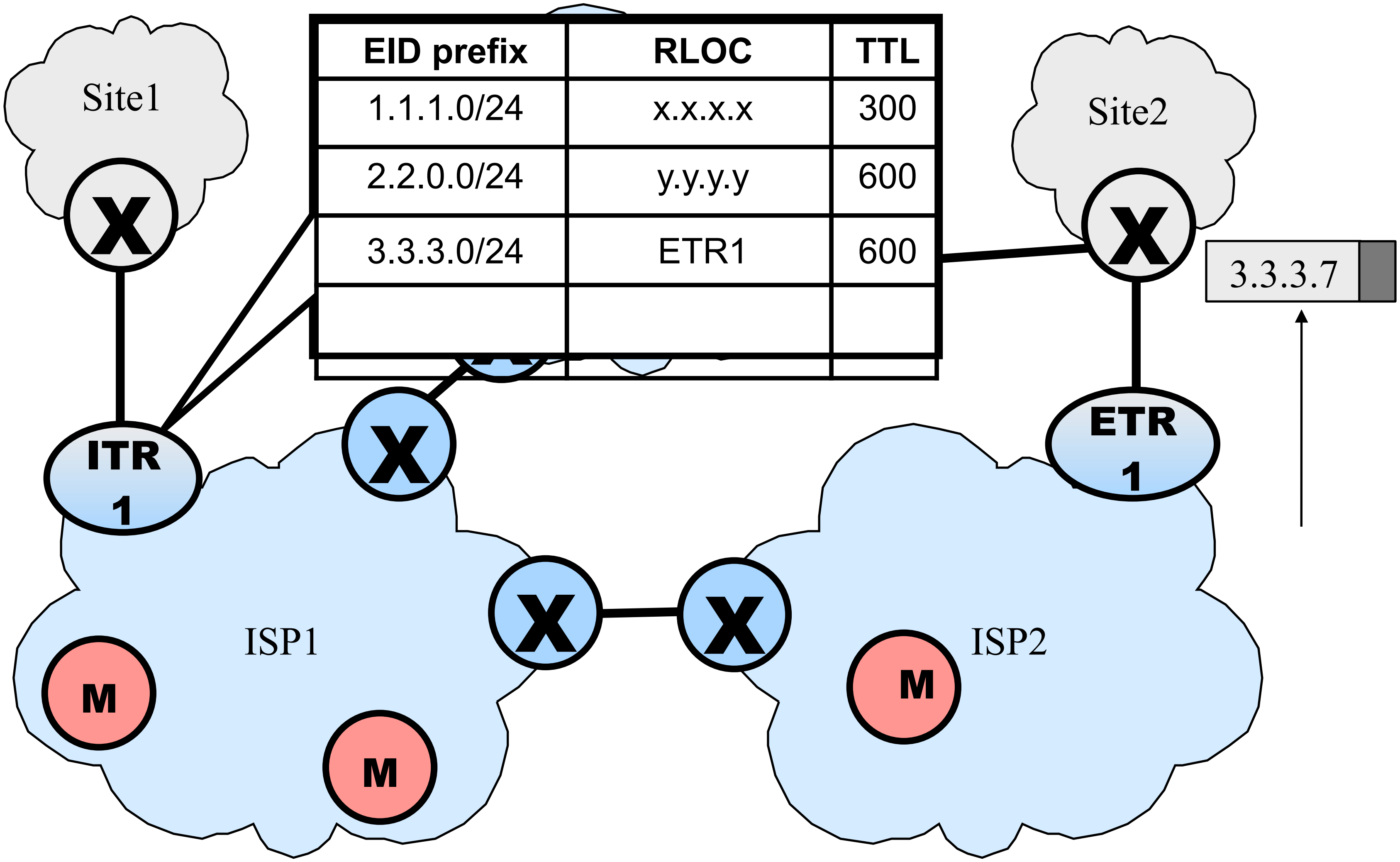


MapRec Added to Cache



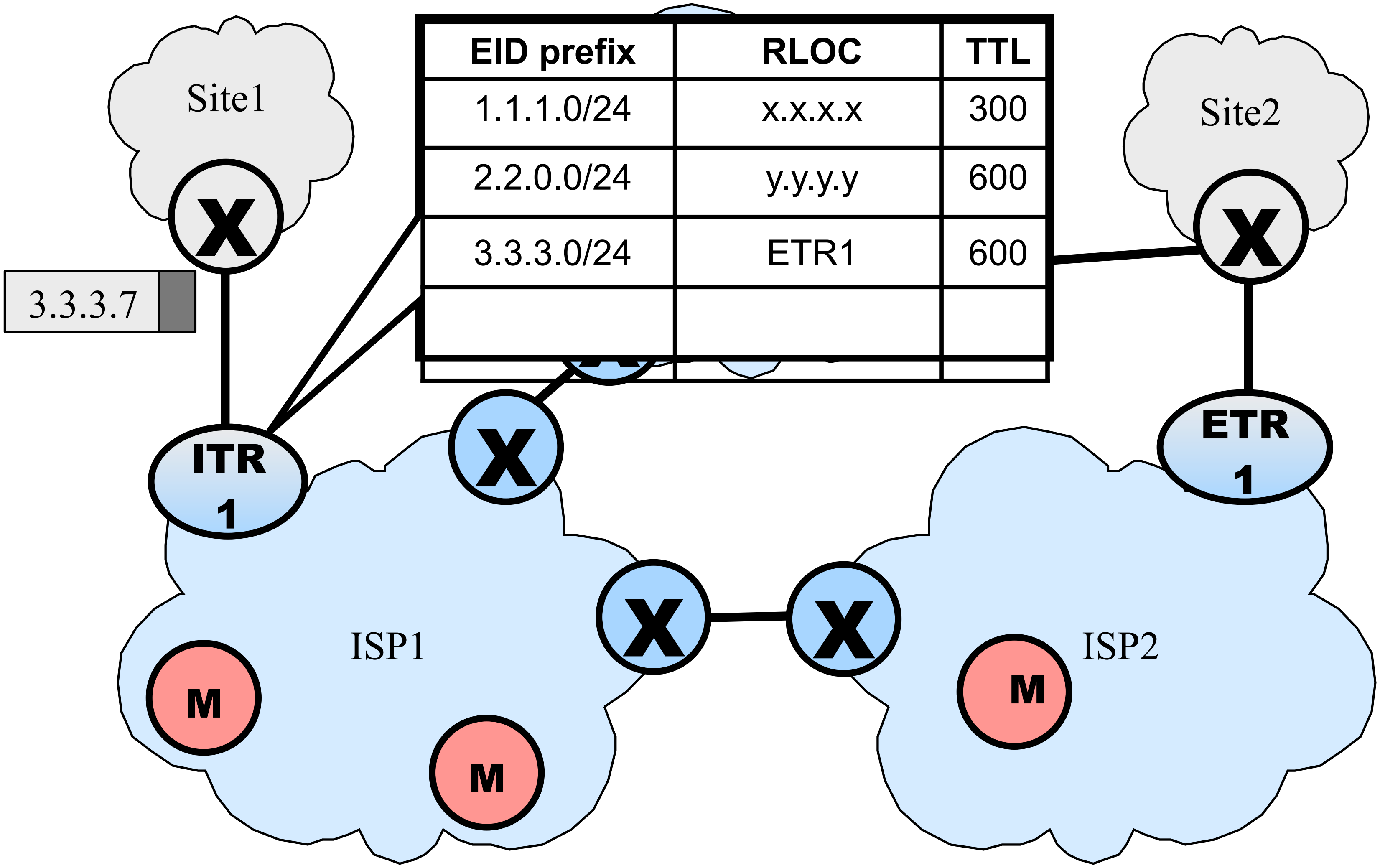
EID prefix	RLOC	TTL
1.1.1.0/24	x.x.x.x	300
2.2.0.0/24	y.y.y.y	600
3.3.3.0/24	ETR1	600

Packet Decapsulated and Delivered



EID prefix	RLOC	TTL
1.1.1.0/24	x.x.x.x	300
2.2.0.0/24	y.y.y.y	600
3.3.3.0/24	ETR1	600

Next Packet

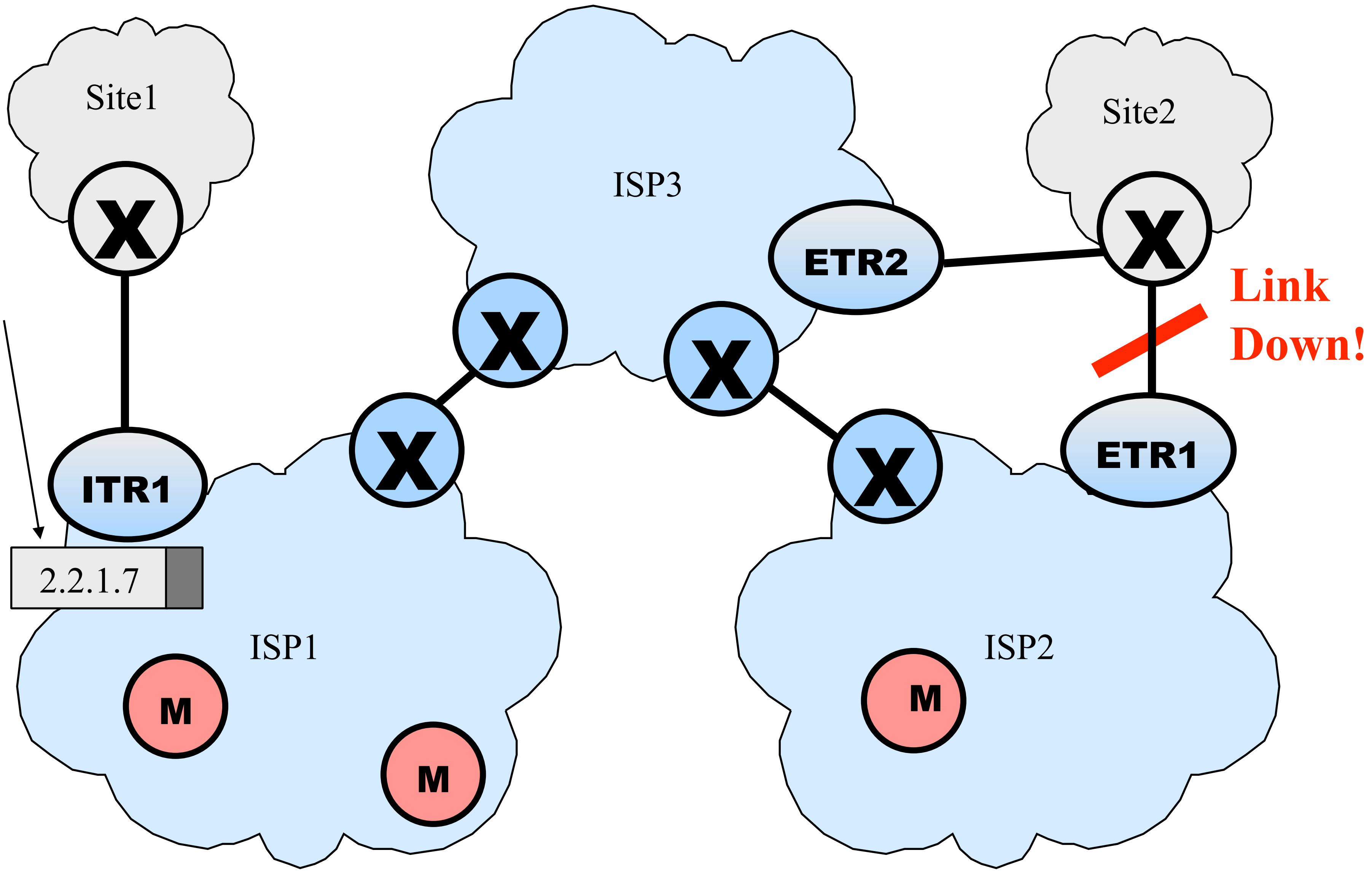


EID prefix	RLOC	TTL
1.1.1.0/24	x.x.x.x	300
2.2.0.0/24	y.y.y.y	600
3.3.3.0/24	ETR1	600

Challenge 2: failure detection and recovery

- Goal: minimize packet losses
- Approaches:
  - Reachability change is learned via data-triggered control messages
  - Reachability state is stored in default mappers
  - Reroute packets that are in transit to unreachable destinations

A Failure Recovery Example



Challenge 3: Incremental Deployment

Most important factors that will determine whether a new design will get rolled out

- The new design must offer enough incentive to the first mover
- One party must be able to deploy the solution unilaterally (and benefit)
- Do not count on majority to move during any reasonable time period

APT Incremental Deployment: basic ideas

- Day-0: Must be a unilateral decision by a single party to turn on APT
  - Map-n-encap: need both tunnel points under one party's control
- To provide incentives for the first mover: being able to reduce its own BGP table size
  - remove internal customers' prefixes from routing to mapping
  - Apply virtual aggregation approach (by Paul Francis) to reduce external prefixes
- Sketched out a smooth transition stratege
  - Snowball rollout as more ISPs deploy APT
  - no disturbance to the rest of the global routing system

Separation + Multipath Transport

- Potential benefit from multipath transport solution
  - End hosts can use multiple paths simultaneously, or choose their favorite path(s)
  - End hosts see the end-to-end picture in load balancing
  - End-to-end resilience against individual path failures
- Separation works well with Multipath transport
  - Edge multihomed site can split its prefix into multiple subprefixes, each subprefix corresponds to one of the site's providers
- Separation for global routing scalability, without dependency on the assumption that all/majority sites would adopt PA addresses any time soon (or ever)
  - multipath transport for end host benefits

Additional Benefits from Separation

Mapping layer as Insulation

- Growth of Internet user population → growth of the mapping table
  - Without affecting the core routing system scalability
- Allow edge protocol innovations
- Allow the core to evolve independently from edge
  - E.g. optical path switching

Mapping layer as a Layer of Protection

- disallow end hosts from communicating directly with routers in the core
  - Most attacks come from end user machines
  - Raising the barrier against attackers targeting at routing infrastructure.
- help trace back offending traffic to the source (the ingress tunnel router)
- *Does not protect against bad parties inside the core*
  - However significantly reduces the scope of the problem

Mapping layer for new functions

- An ideal place to implement DDoS mitigation
  - Encaptulation exit points can monitor and control traffic volume from tunnel entry points
- Other functions to explore

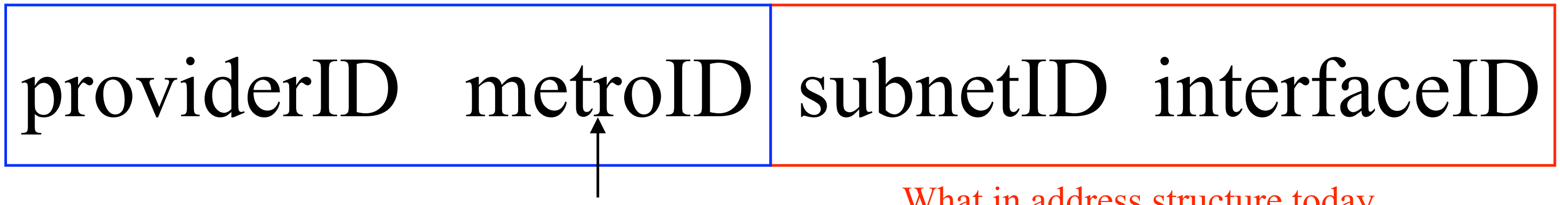
We are excited about the great potential APT may offer

We are seeking collaborations in its further development and implementation

What about routing inside the core

Or routing in general?

- Plan: combining the separation idea with encoding geo-location into IP address

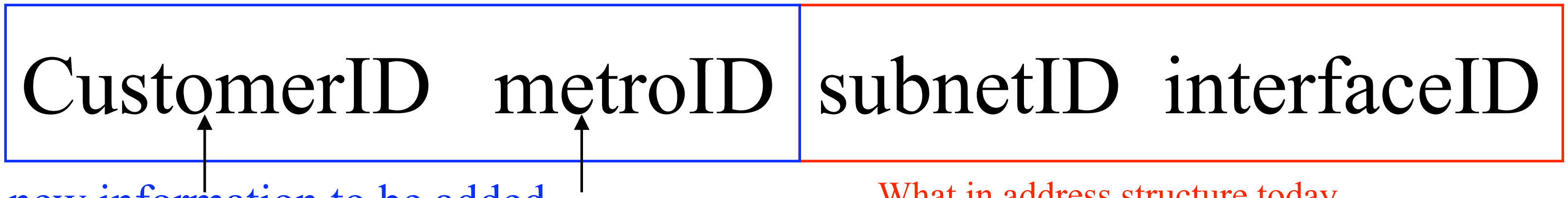


providerID metroID subnetID interfaceID

new information to be added

What in address structure today

Similarly:



CustomerID metroID subnetID interfaceID

new information to be added

What in address structure today

This topic will necessarily make another talk on its own!

Is it worth taking all this trouble

- Given that the existing system seems to be working just fine?
- Why is it *necessary* to plan changes to the existing routing architecture?

"Being the Right Size" by J. B. S. Haldane, 1928

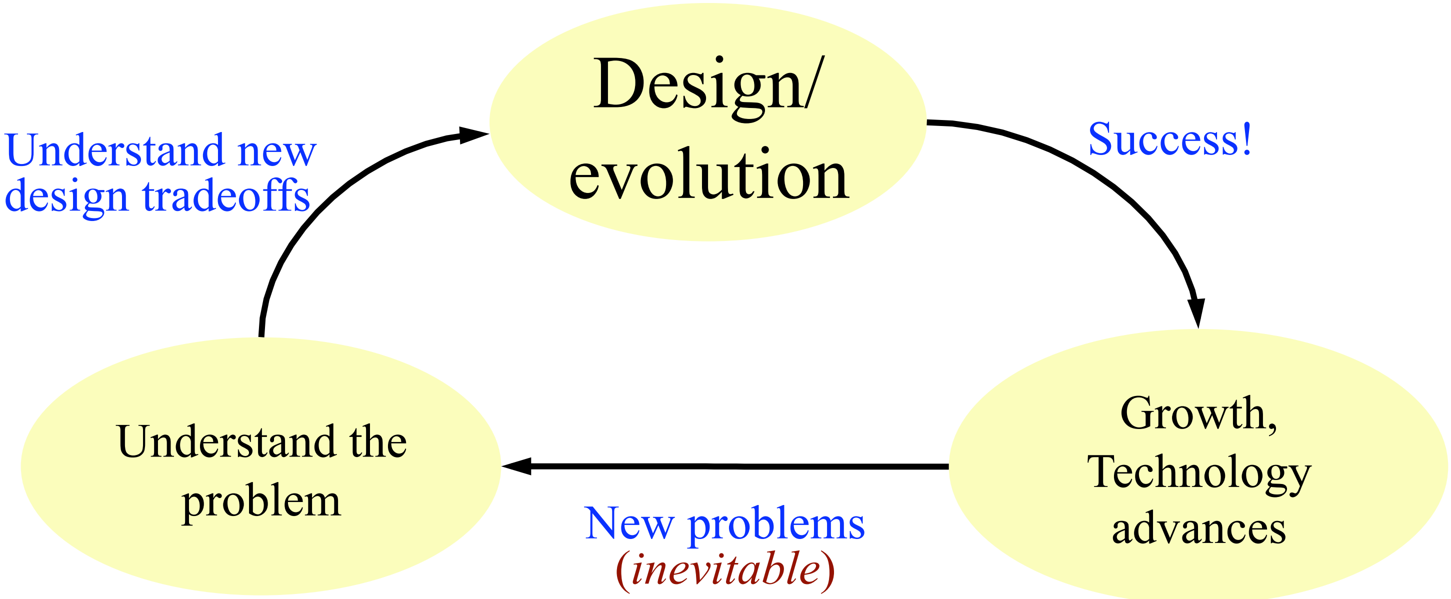
- "A typical small animal, say a microscopic worm or rotifer, has a smooth skin through which all the oxygen it requires can soak in
- "Increase its dimensions tenfold in every direction, and its weight is increased a thousand times, so that if it is to use its muscles as efficiently as its miniature counterpart, it will need a thousand times as much food and oxygen per day
- "Now if its shape is unaltered its surface will be increased only a hundredfold, and ten times as much oxygen must enter per minute through each square millimeter of skin..."

Change in size ⇒ change in form

“For every type of animal there is a most convenient size, and a large change in size inevitably carries with it a change of form.”

Look Back and Look Foreward

- All new systems start small
- Success ⇒ growing large ⇒ change in requirements
- To continue the success ⇒ go through evolution cycle
- Challenge: design with the expectation of future evolutions



Questions?

lixia@cs.ucla.edu