UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

# Internet Scalability

Albert Cabellos
(acabello@ac.upc.edu)

# Some representative numbers

| Forwarding time (TCP ACK at 10Gbps) | 1 cycle (at 1GHz) | Memory Latency (DDR3-2000MHz, 1 Word) |
|---|---|---|
| 35ns | 1ns | 32ns |

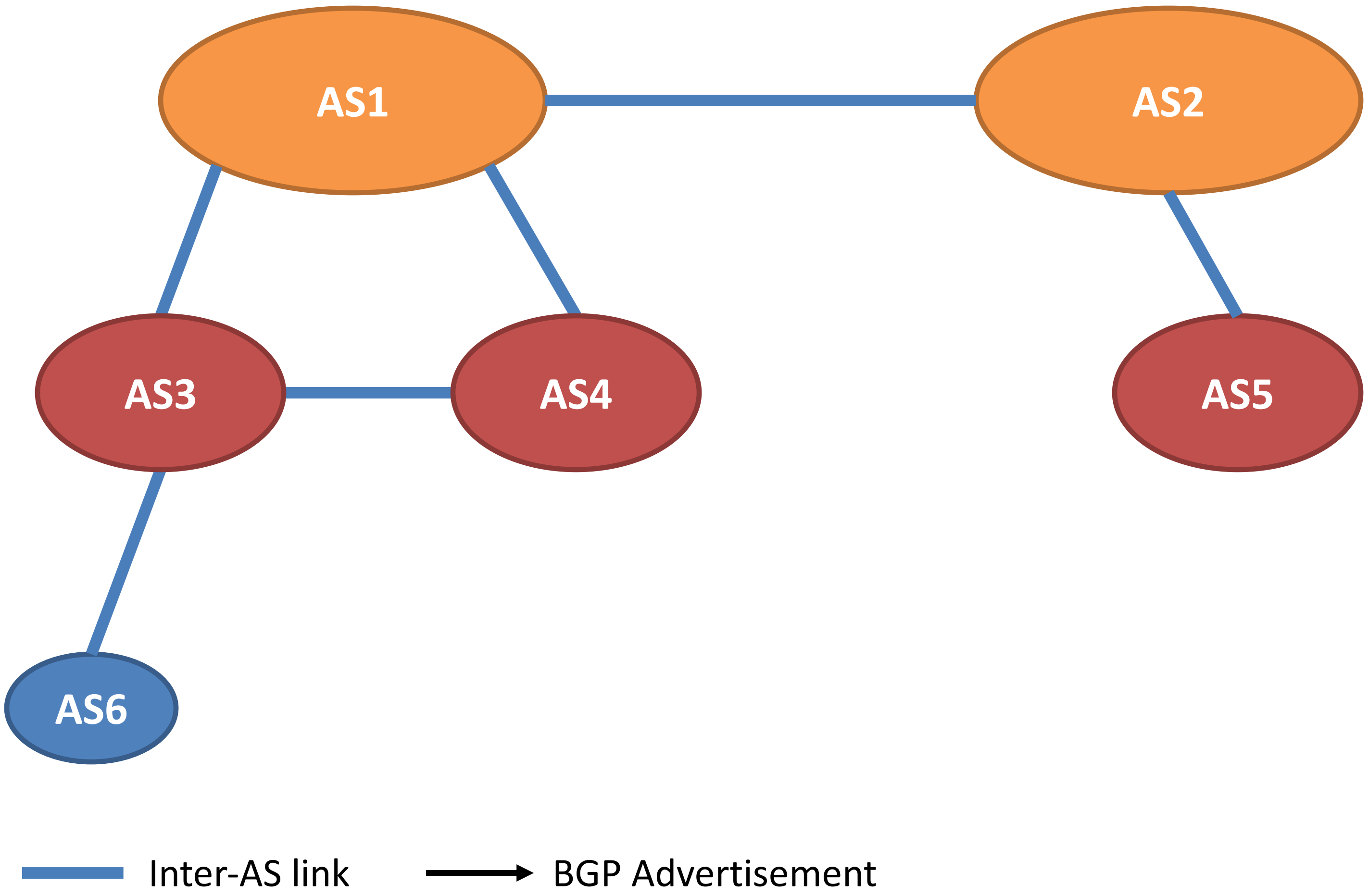
# Agenda

- **The Inter-Domain Routing Problem**
  - Review of Inter-Domain Routing
  - Limitations of Current Internet (IP) Architecture
- **Representative Solutions**
  - Clean-Slate: NNC
  - Evolutionary: LISP

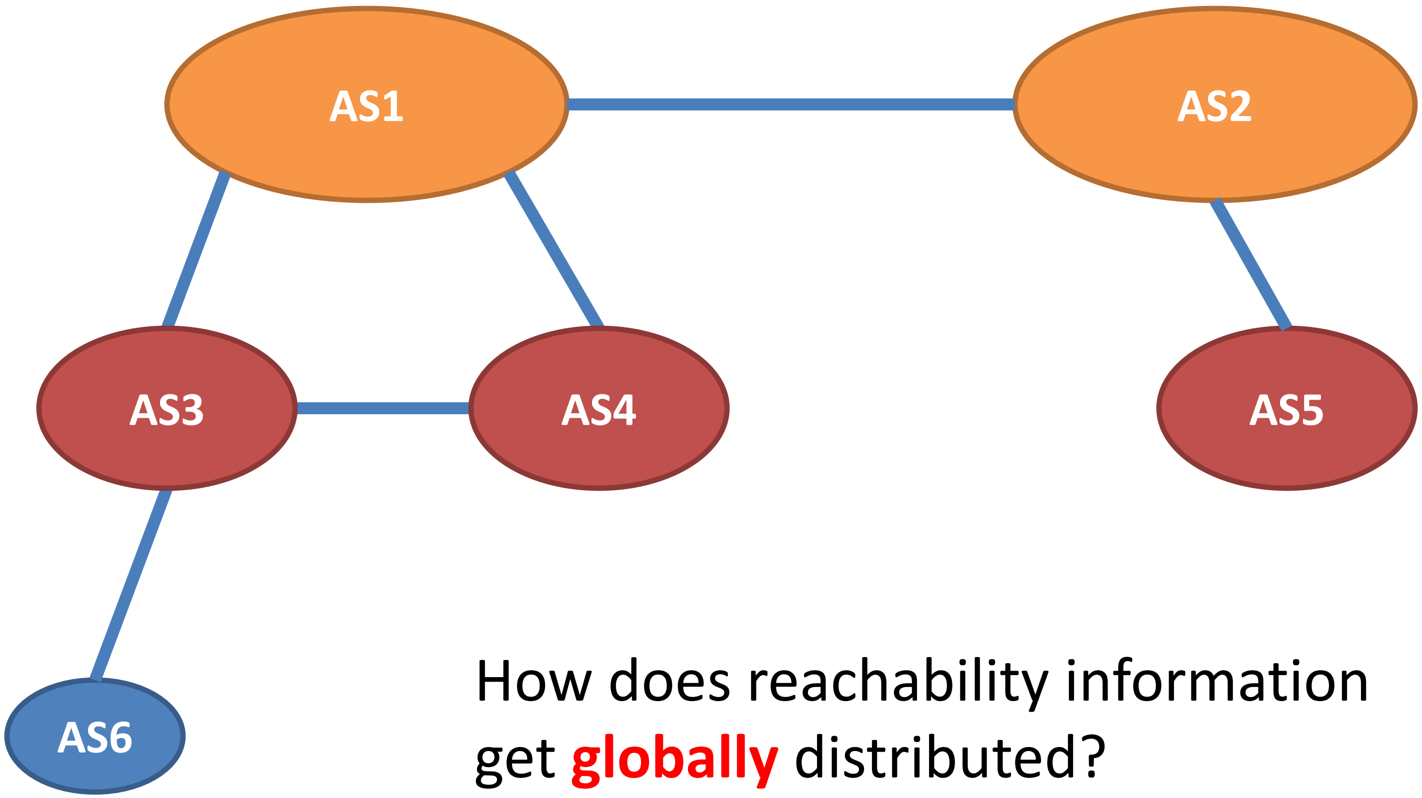
# Inter-Domain Routing

- Allows the exchange of data between peers along the best path that possibly crosses several transit provider domains and fulfils the **routing policies** of each domain independent of its network topology.
- Each peer is an Autonomous System (**AS**)
  - A connected group of one or more IP prefixes run by one or more network operators which has a **single** and **clearly defined** routing policy [RFC1930].
  - Has a globally unique number associated to it (AS number)
- Border Gateway Protocol (**BGP**)
  - Common Inter-AS routing protocol
  - The primary function of a BGP speaking system is to exchange **network reachability** information with other BGP systems [RFC4271]

# Reachability Advertisements



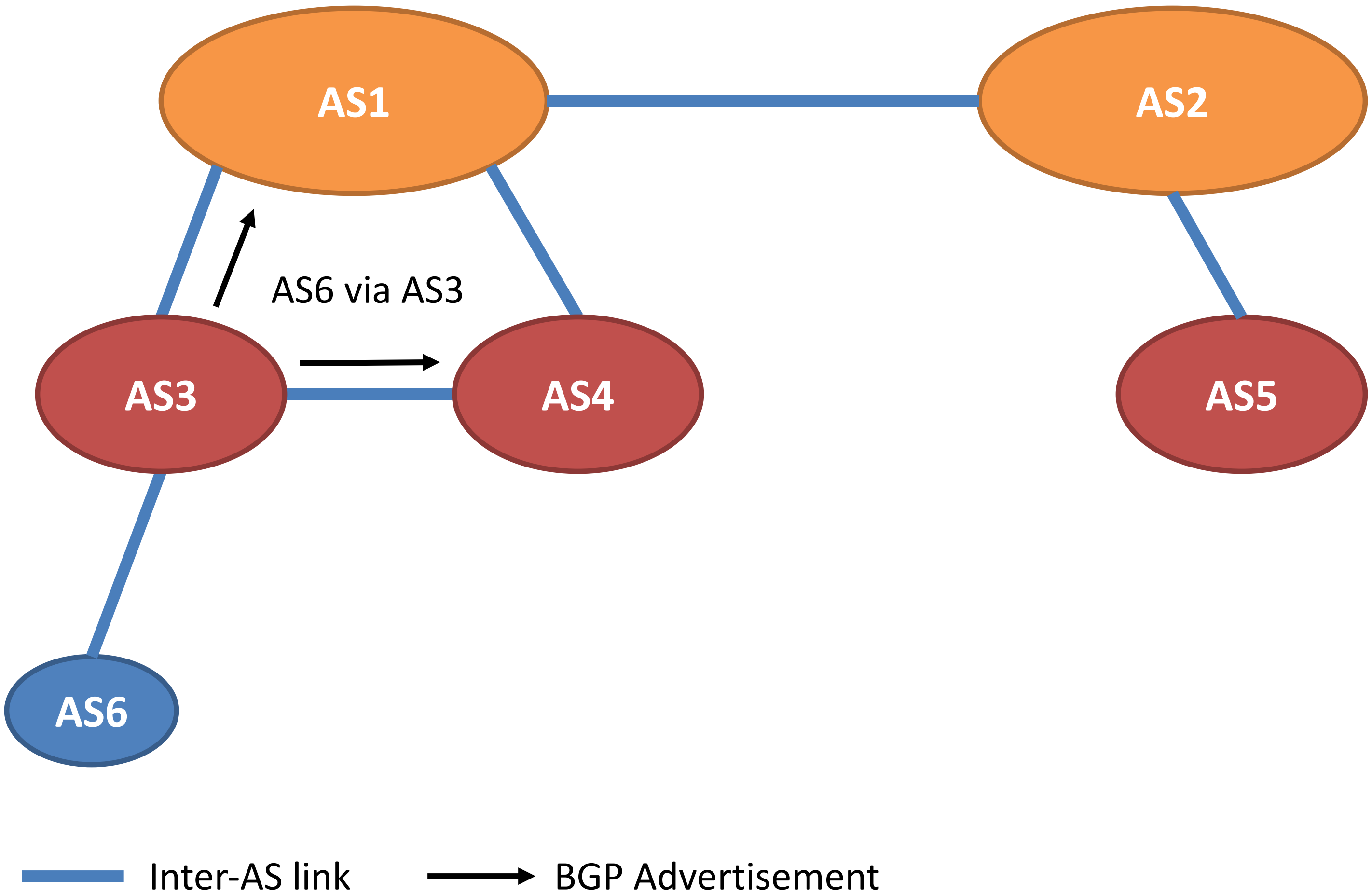
# Reachability Advertisements




How does reachability information get **globally** distributed?

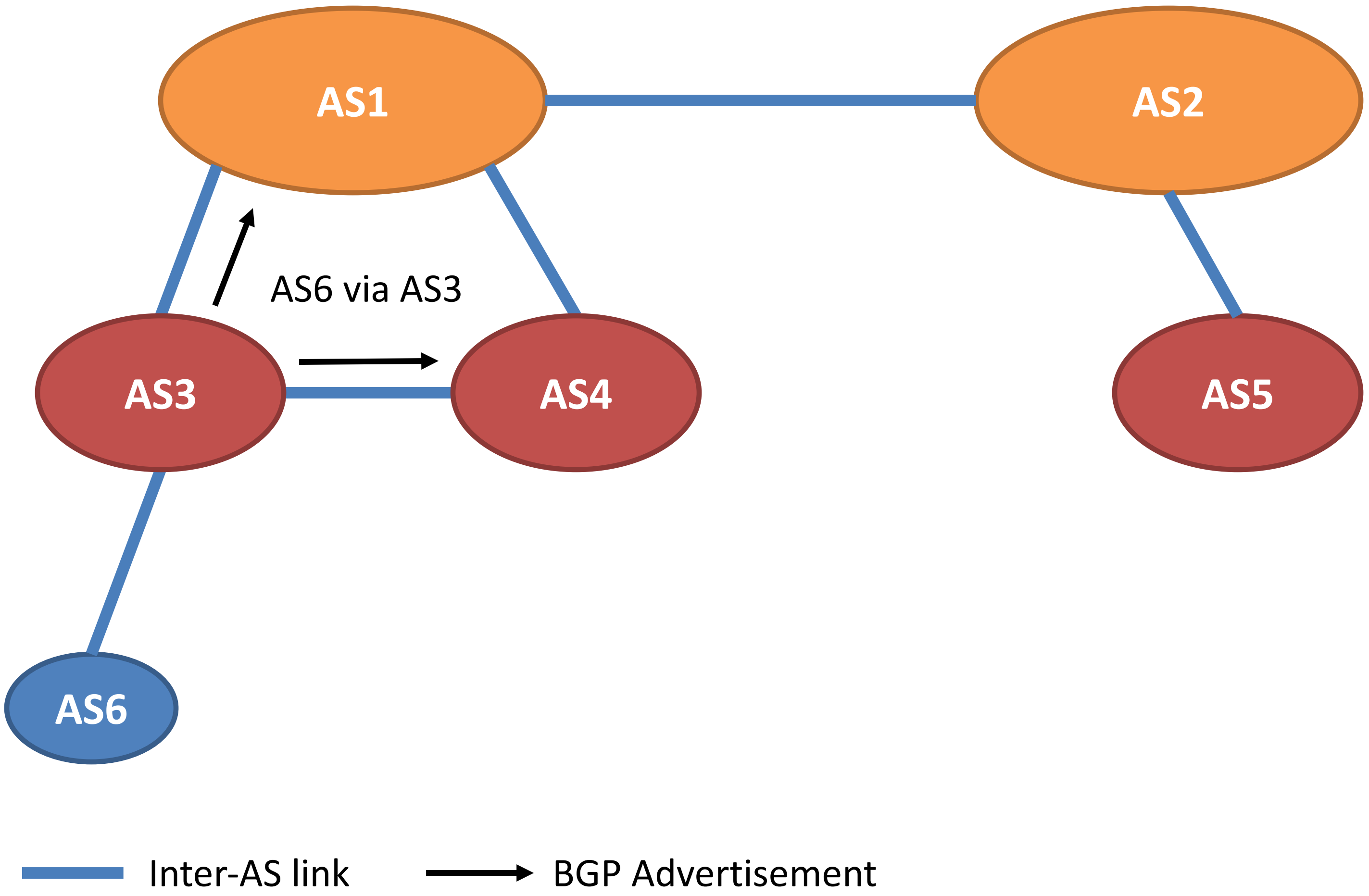
Inter-AS link BGP Advertisement

# Reachability Advertisements



# Reachability Advertisements

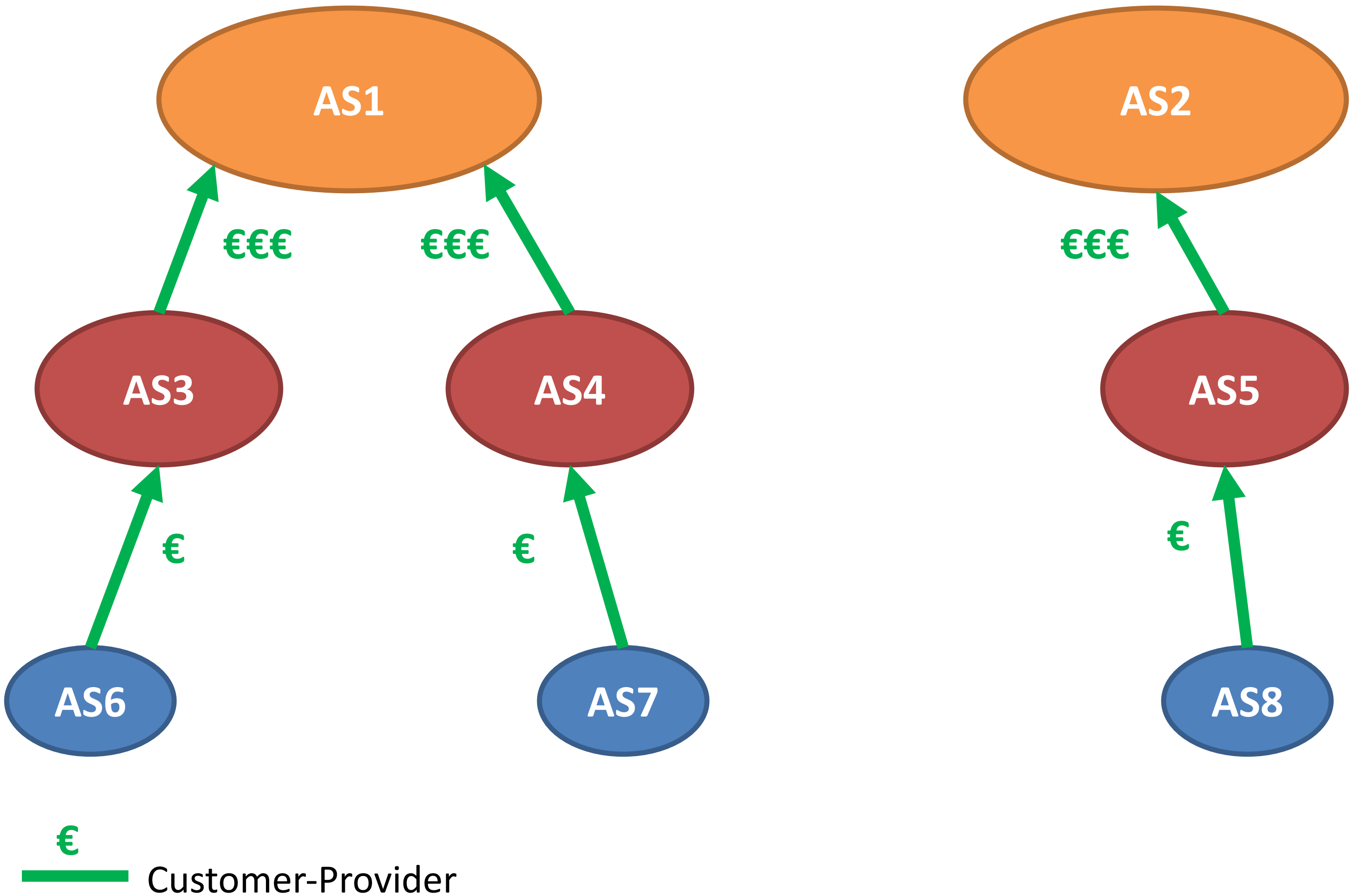
**How to further propagate the routes?**



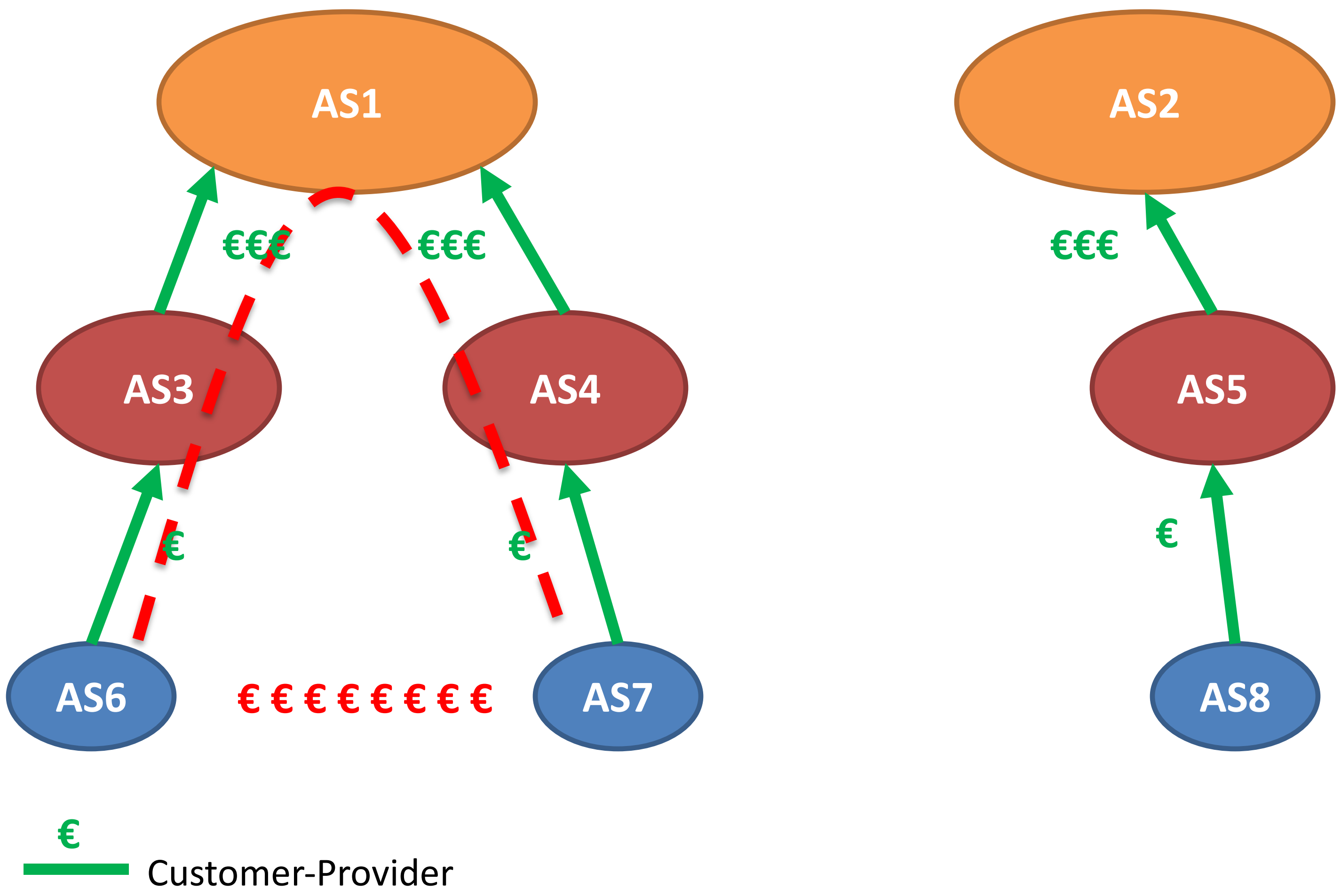
# Routing Policies

- Routing information exchanged via BGP supports only the **destination-based** forwarding paradigm
  - Each AS announces only what it considers its best path
- BGP allows each domain to define its own routing policy
- In practice there are two common policies
  - **Customer-provider peering**: customer c buys Internet connectivity from provider p. We say that p learns c’s customer routes and p provides paid transit for c.
  - **Shared-cost peering**: AS x and y agree cu exchange reachability information (customer routes) by using a direct shared link through an inter-connection point.

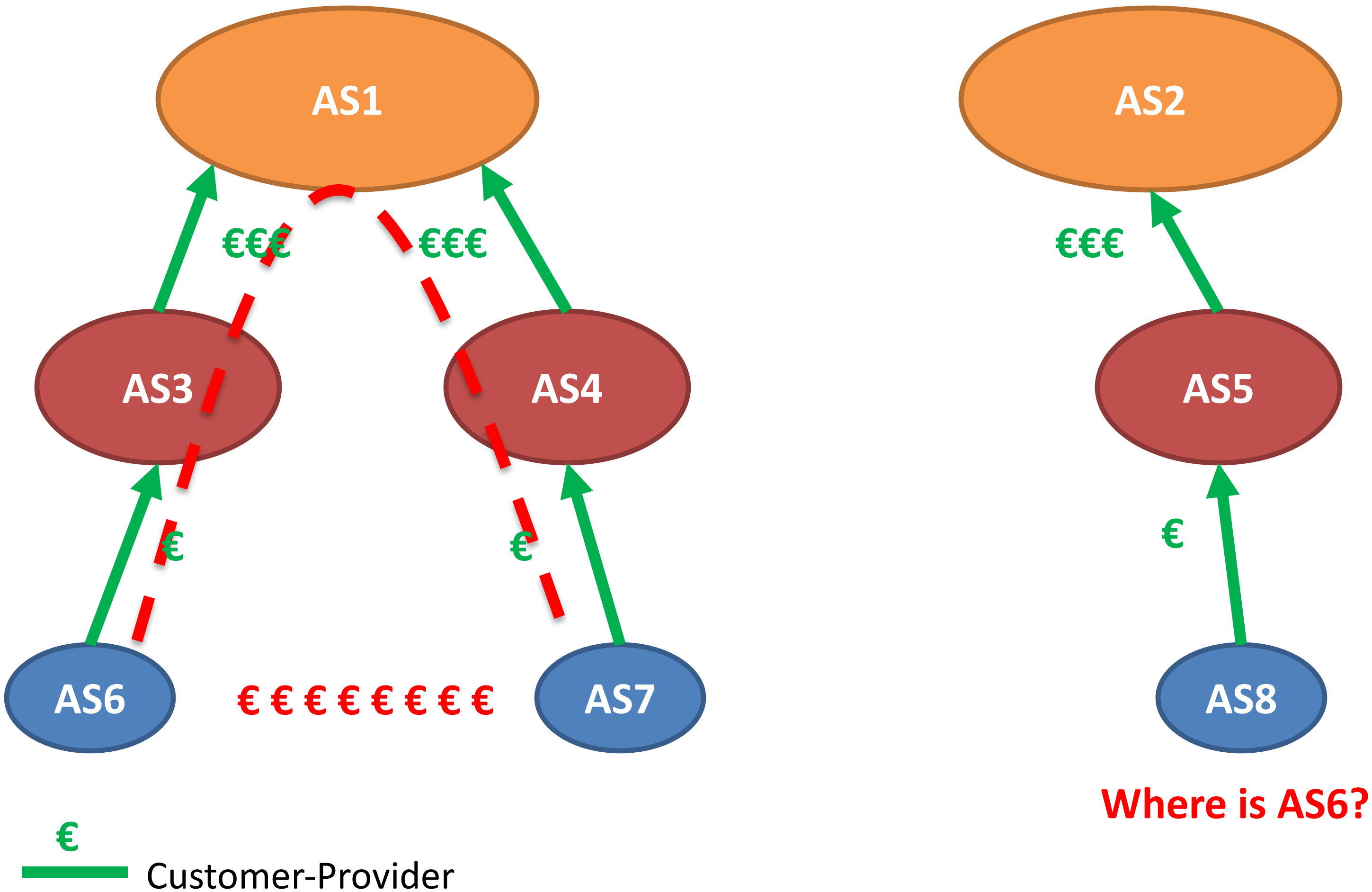
# Customer-Provider Peering



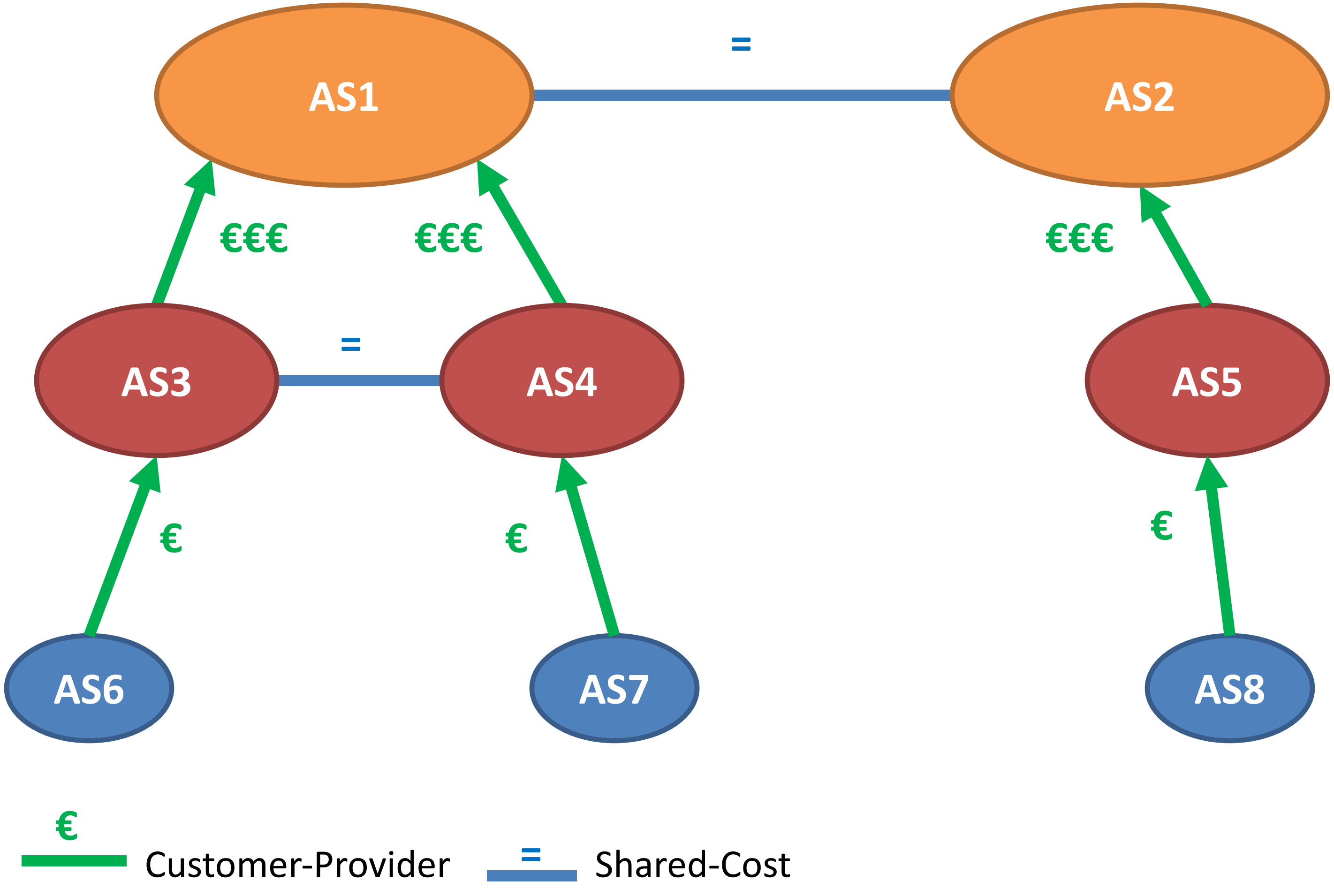
# Customer-Provider Peering



# Customer-Provider Peering

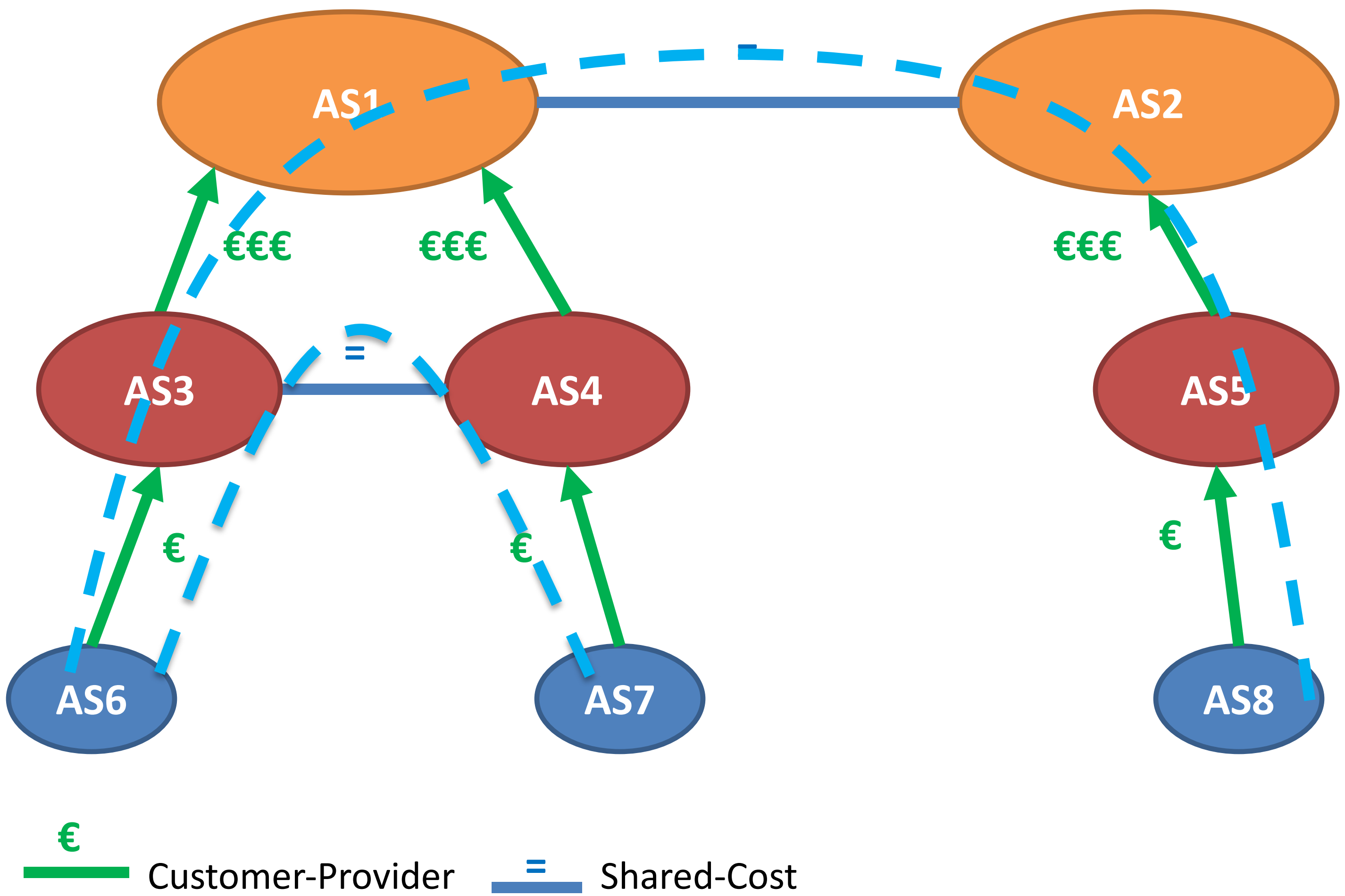


# Shared-Cost Peering

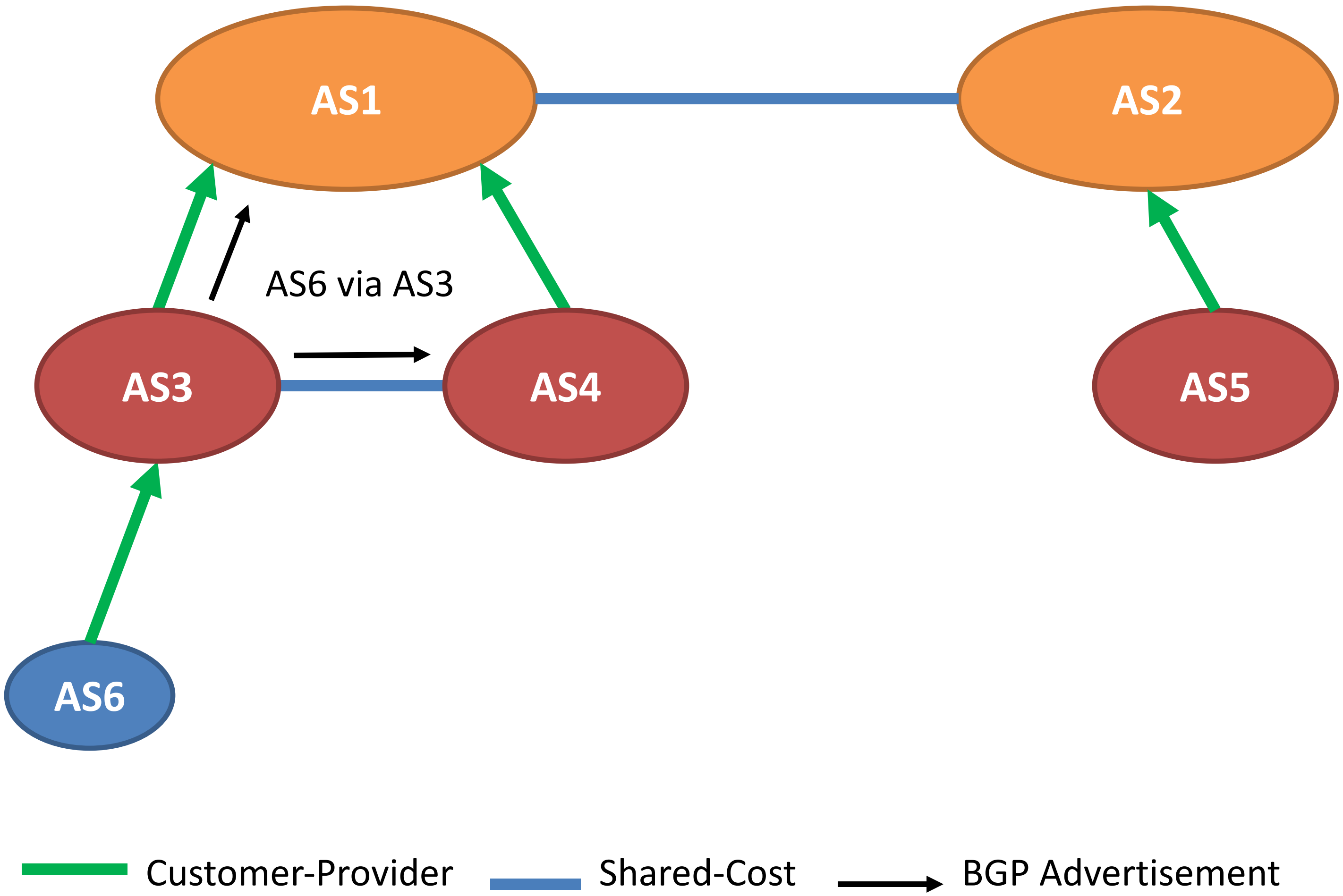


# Shared-Cost Peering

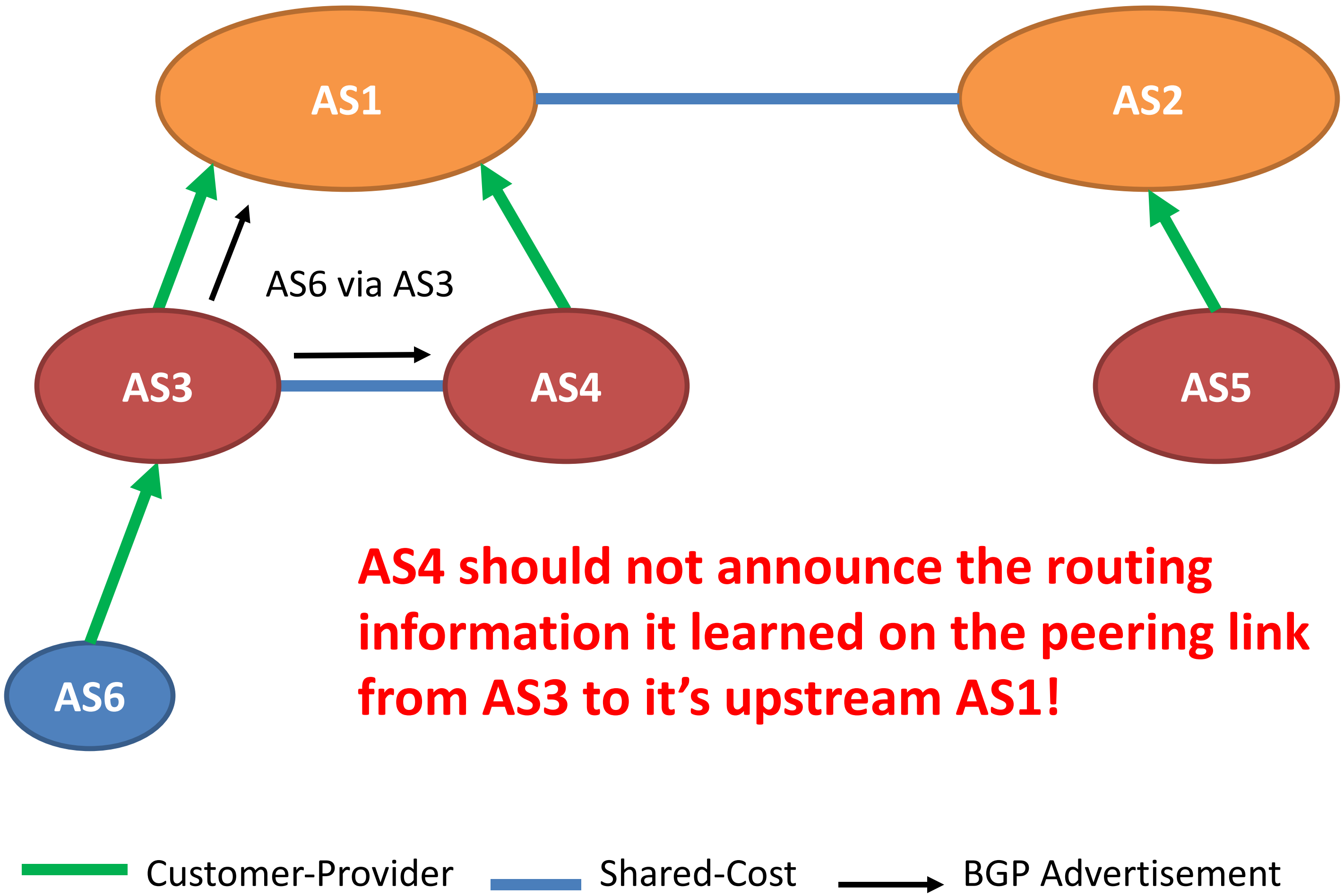
**Note: Valley-free routes!**



# Reachability Advertisements (cntd)

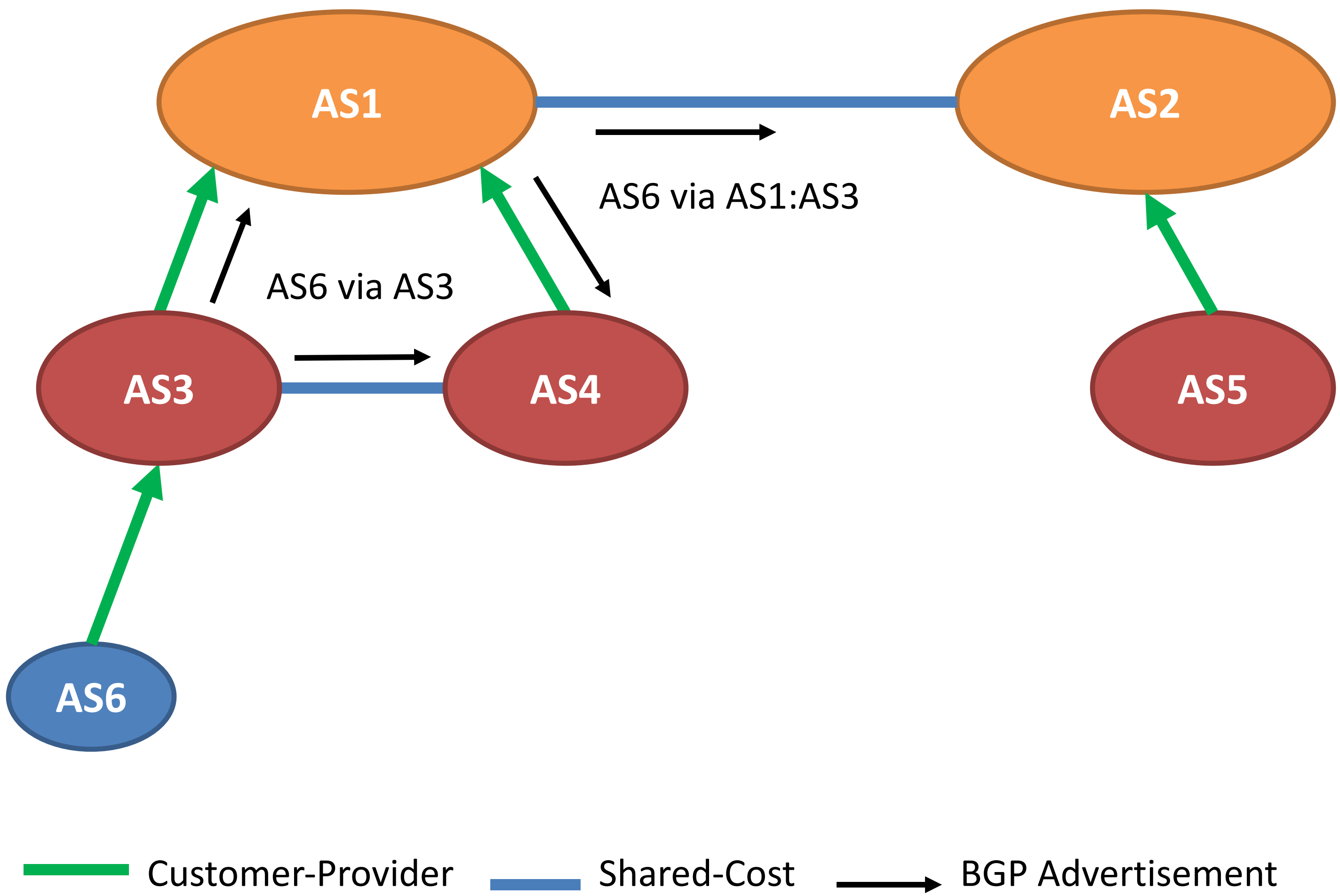


# Reachability Advertisements (cntd)

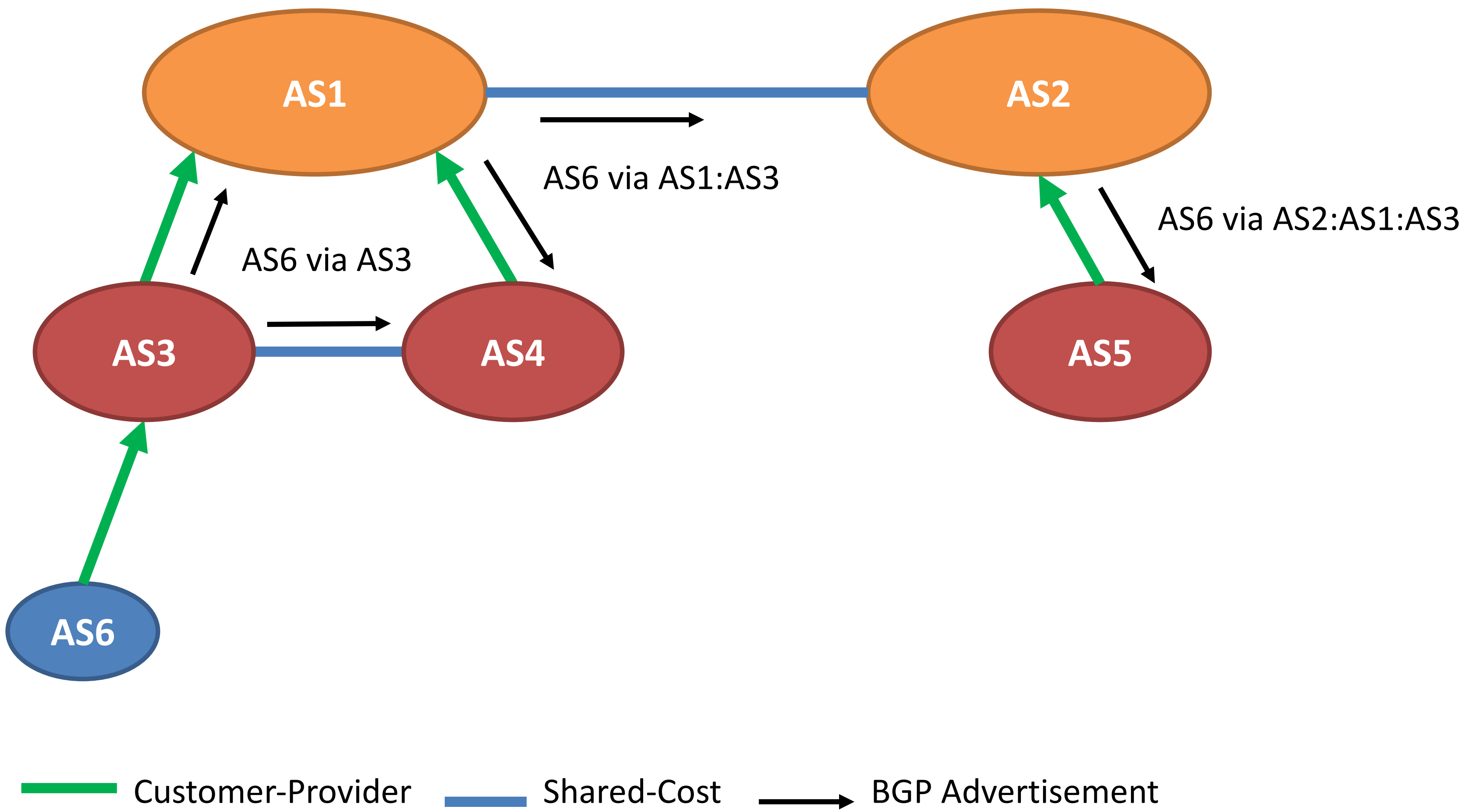



**AS4 should not announce the routing information it learned on the peering link from AS3 to it's upstream AS1!**

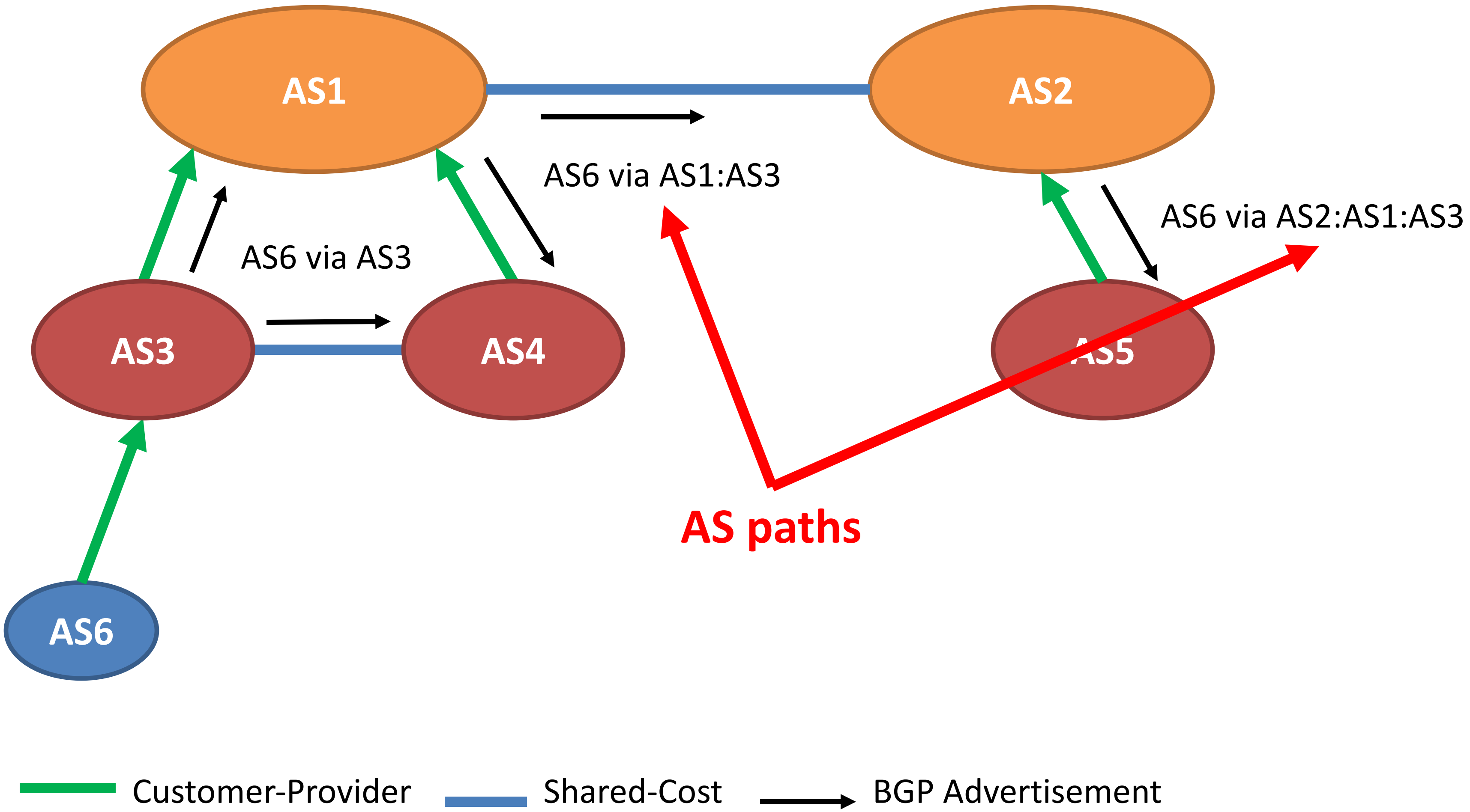
# Reachability Advertisements (cntd)



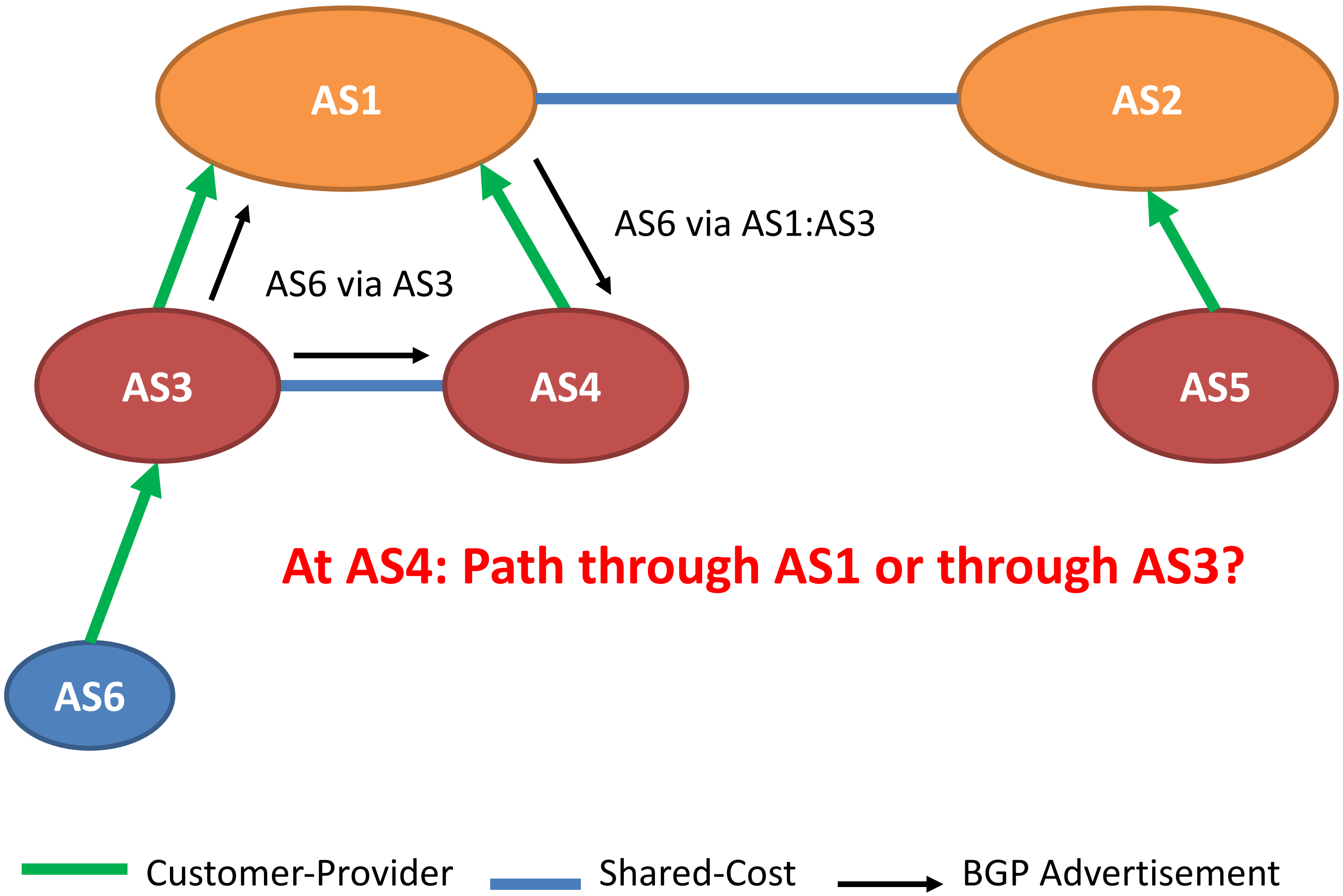
# Reachability Advertisements (cntd)



# Reachability Advertisements (cntd)



# Decision Process (Simplified)




**At AS4: Path through AS1 or through AS3?**

# Decision Process (Simplified)

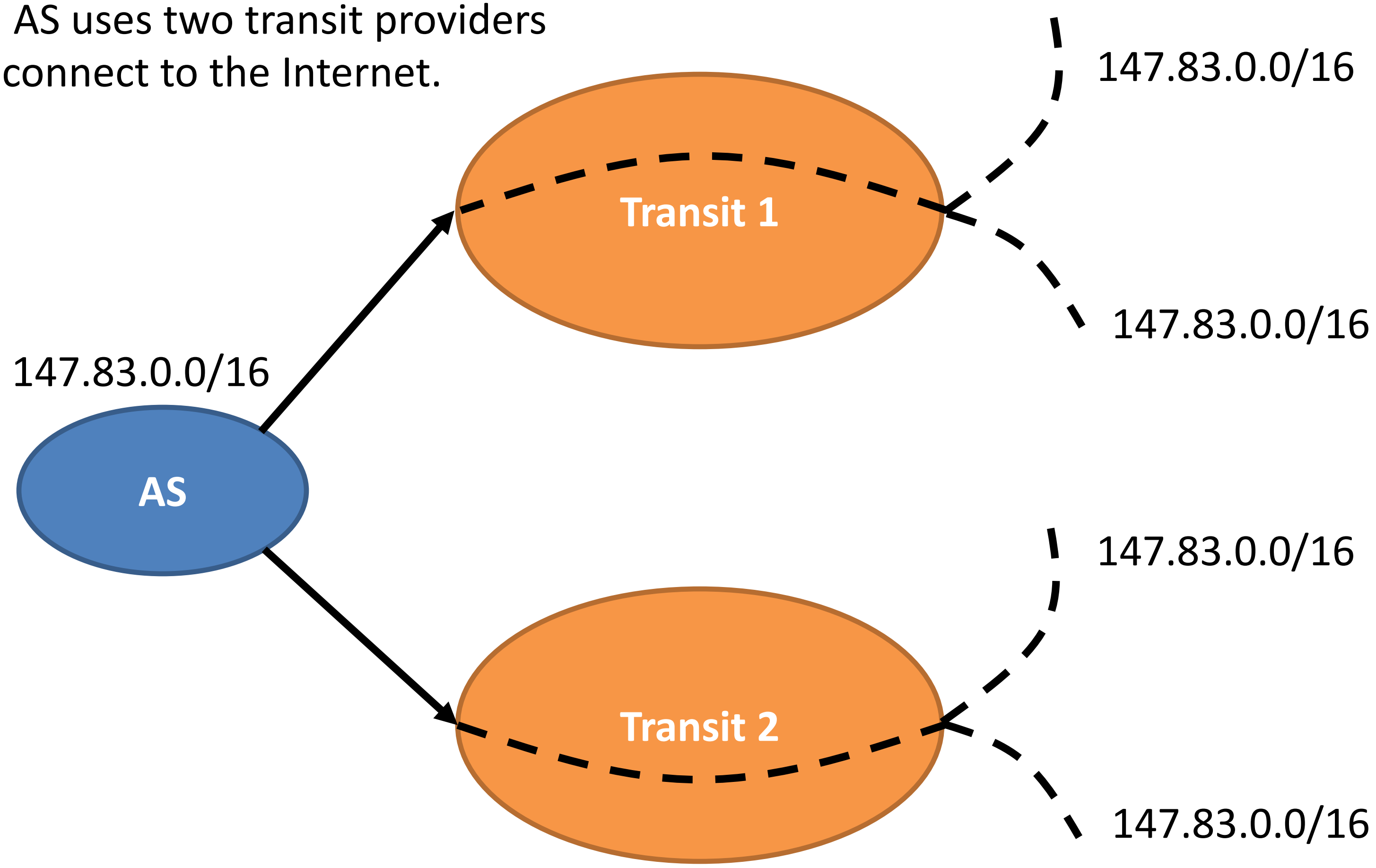
1. Select preferred routes (**local-pref**)
   – Manually configured
2. Select **shortest AS path** route
   – Topology dependent
3. In case of ties use **tie-breaking** rules

# BGP Routing Tables

- BGP Routing Information Base (**RIB**)
  - Aggregates all BGP reachability announcements
  - Saved in control plane memory
- BGP Forwarding Information Base (**FIB**)
  - The output of the BGP decision process ran on the RIB. It contains one route per destination prefix
  - Used when forwarding packets
  - Saved in data plane memory (**fast memory**)

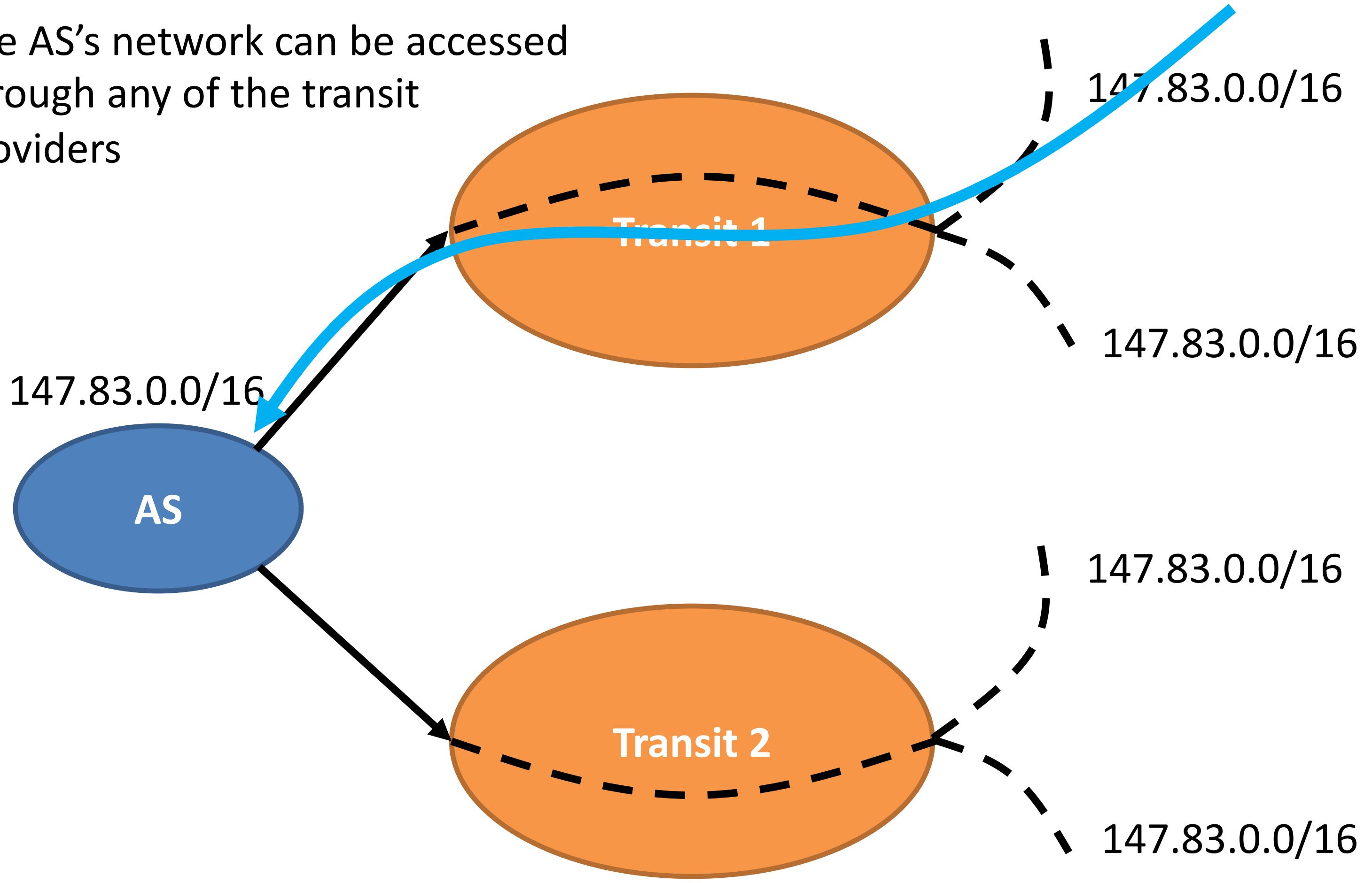
# Multihoming

An AS uses two transit providers
to connect to the Internet.



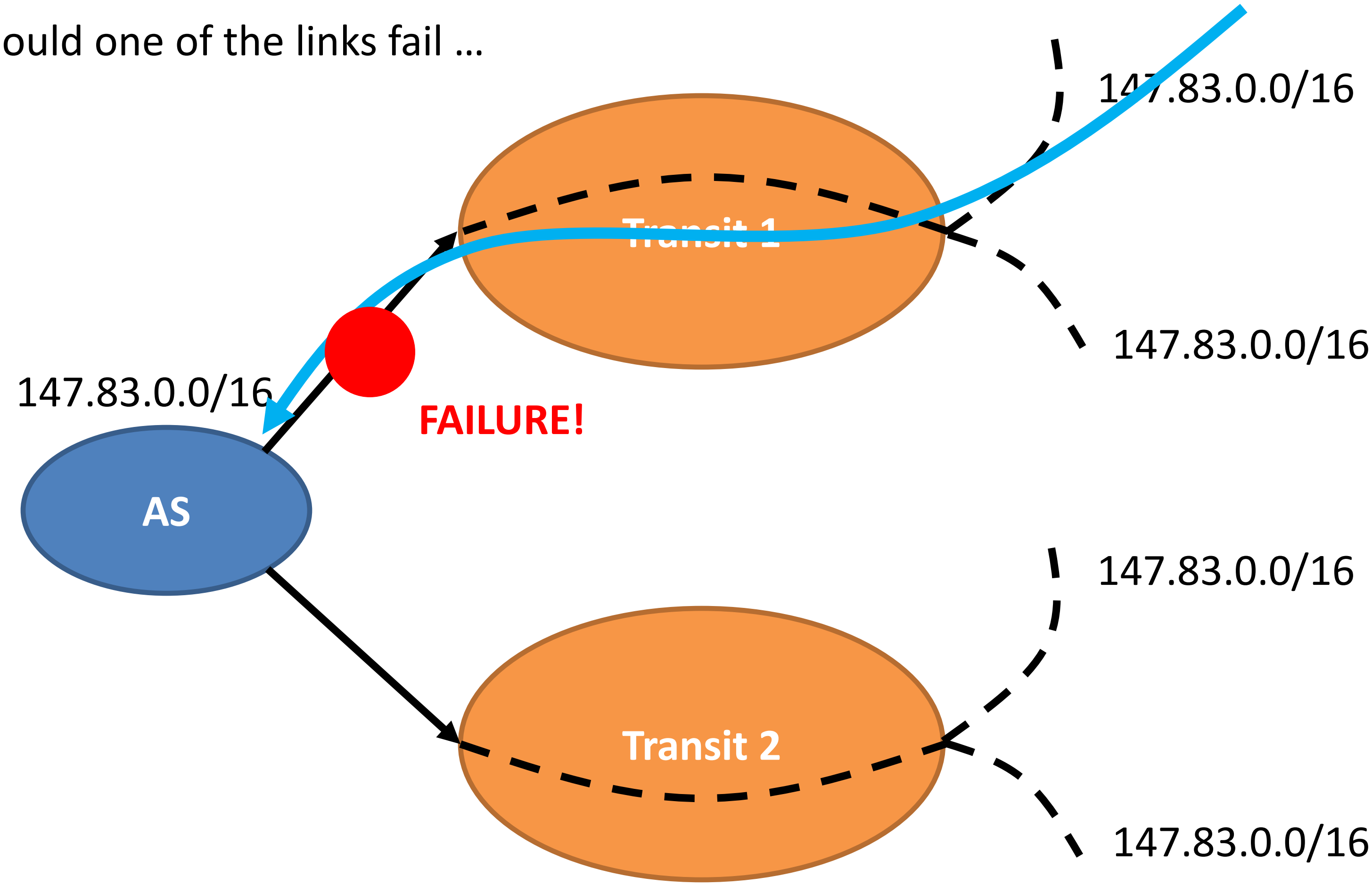
# Multihoming

The AS's network can be accessed through any of the transit providers



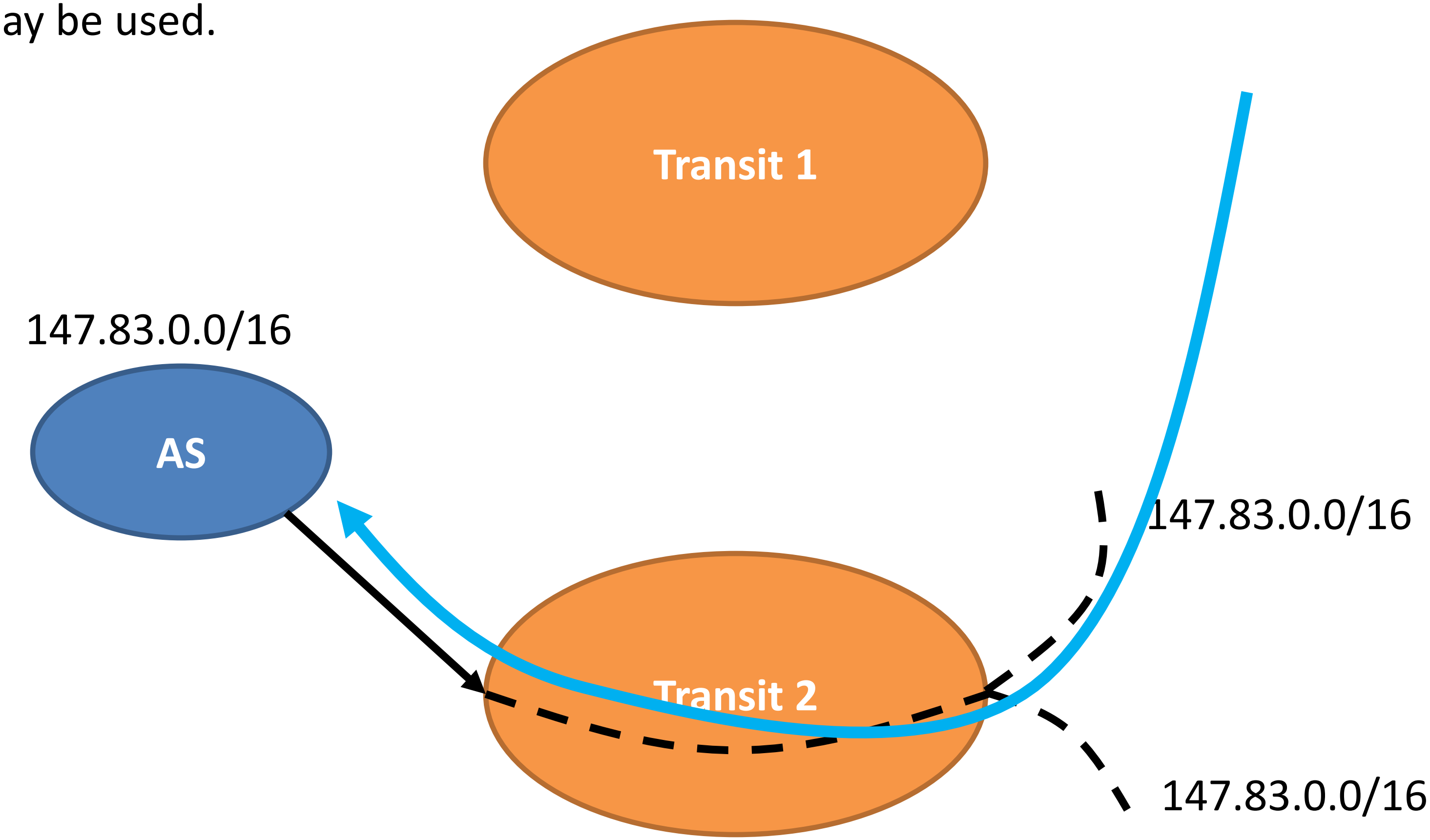
# Multihoming

Should one of the links fail …



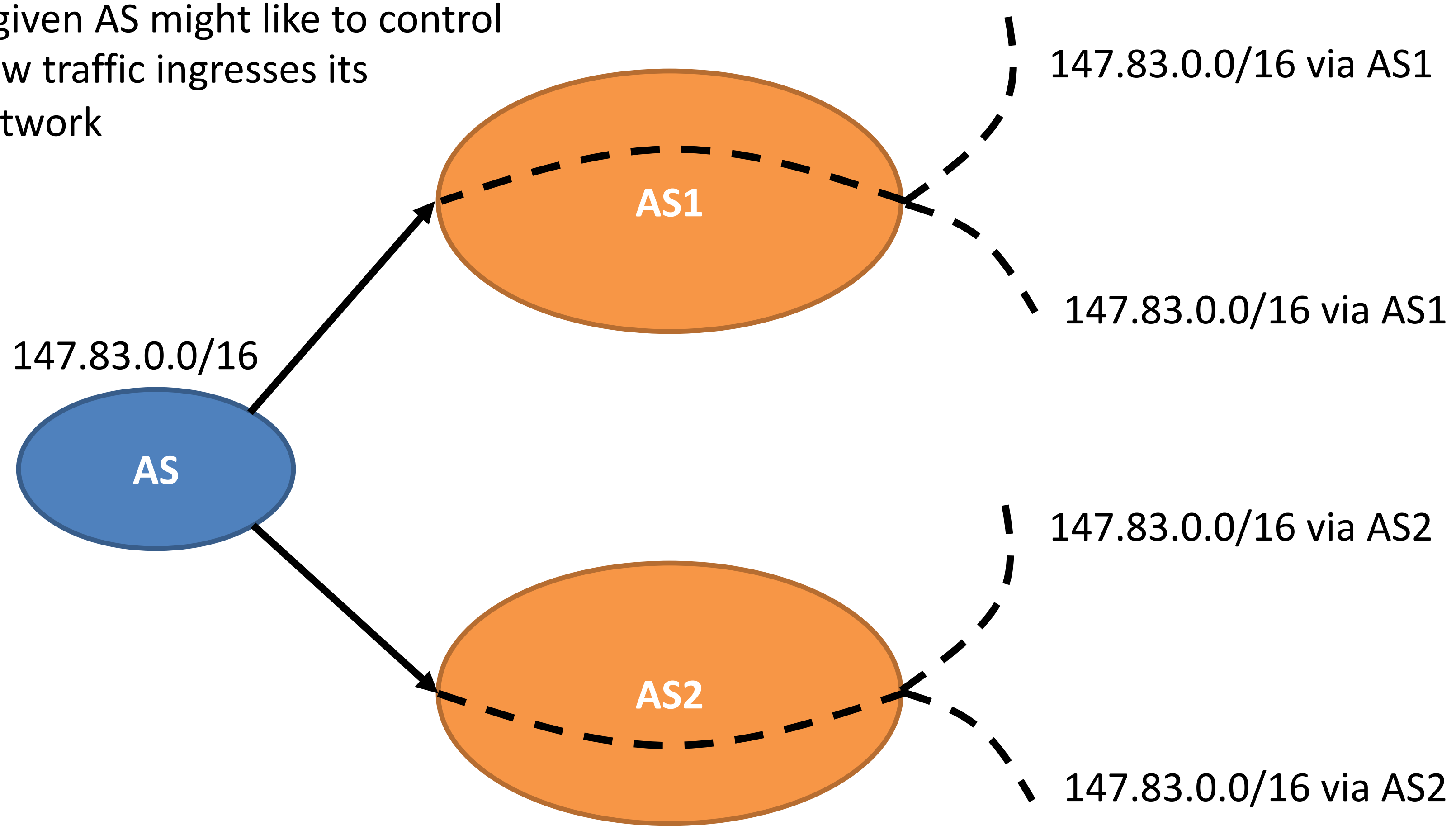
# Multihoming

... the one still available
may be used.



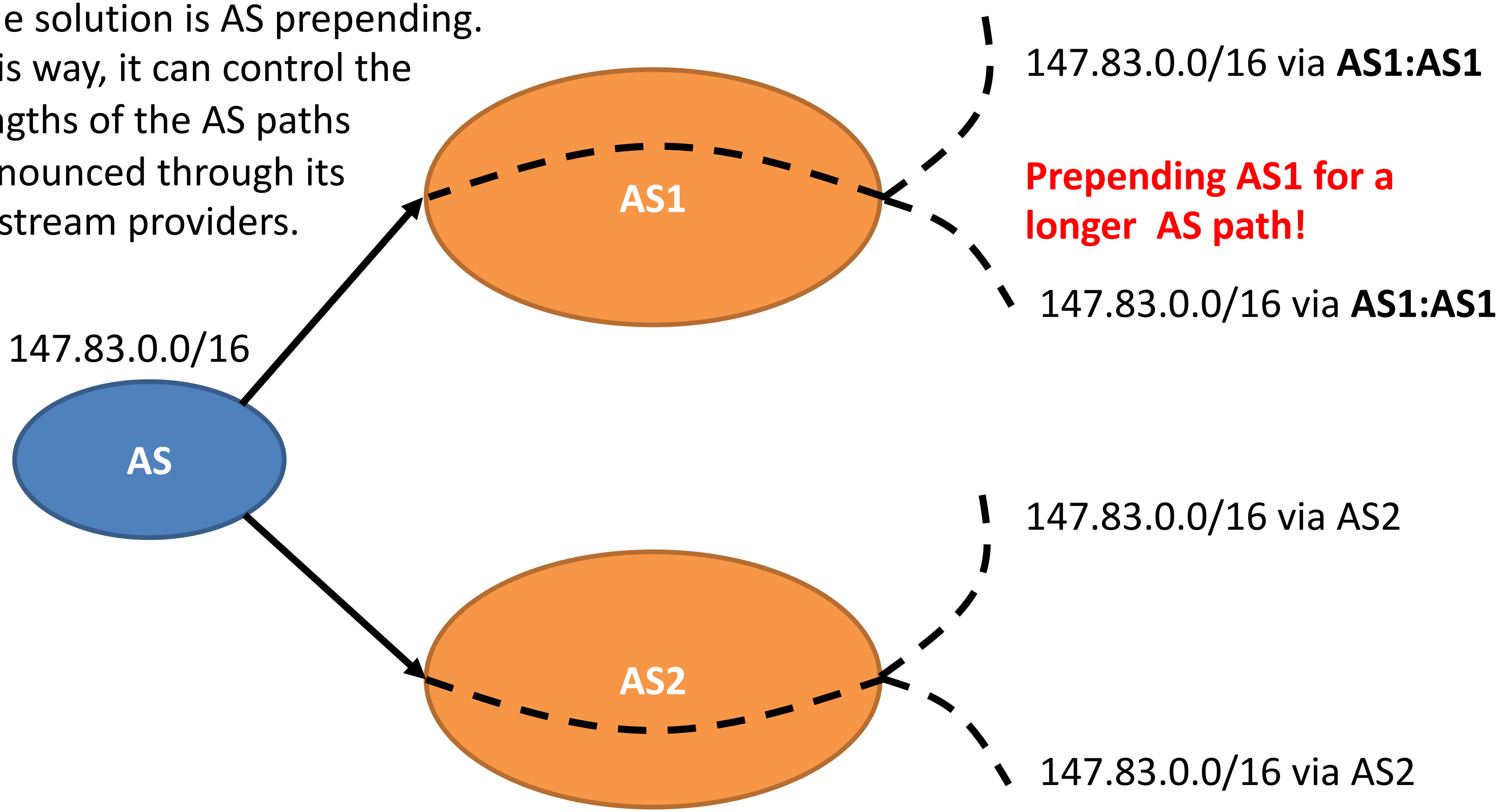
# Ingress Traffic Engineering

A given AS might like to control how traffic ingresses its network



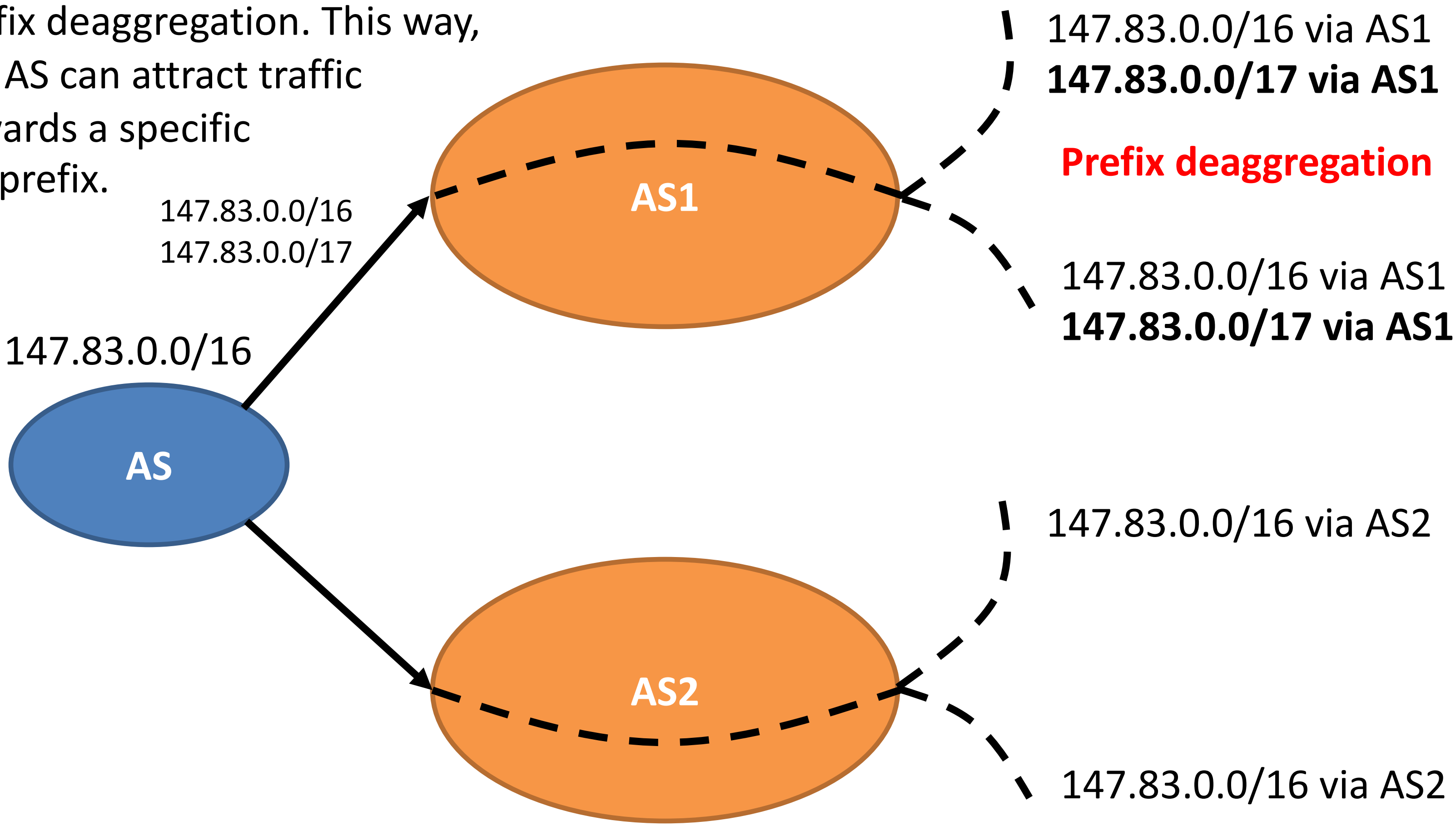
# Ingress Traffic Engineering (1)

One solution is AS prepending. This way, it can control the lengths of the AS paths announced through its upstream providers.



# Ingress Traffic Engineering (2)

The second solution is to perform prefix deaggregation. This way, the AS can attract traffic towards a specific subprefix.



# Internet Scalability

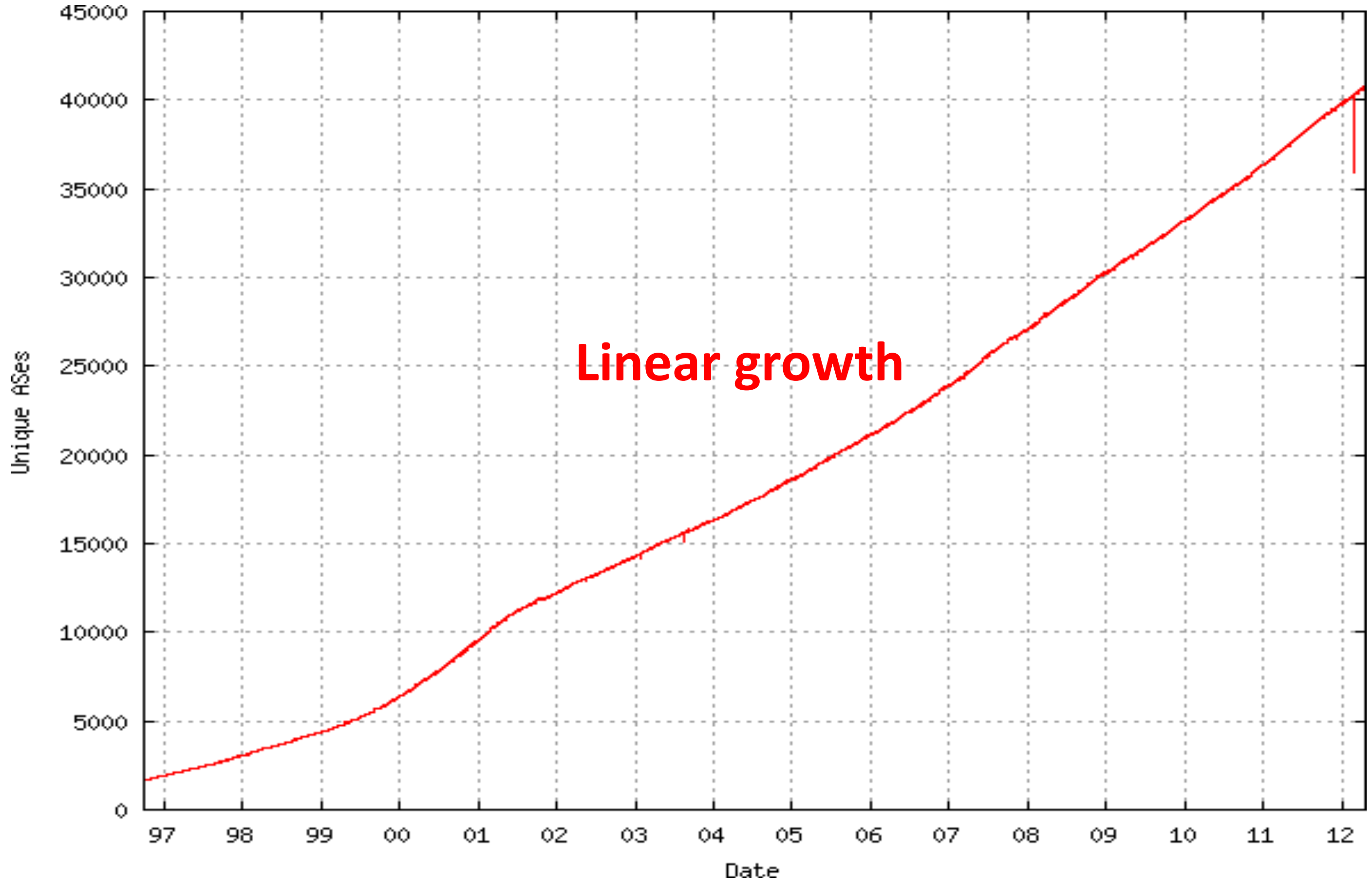
## The Inter-Domain Routing Problem

**Limitations of Current Internet (IP) Architecture**

# How is the Internet Doing?

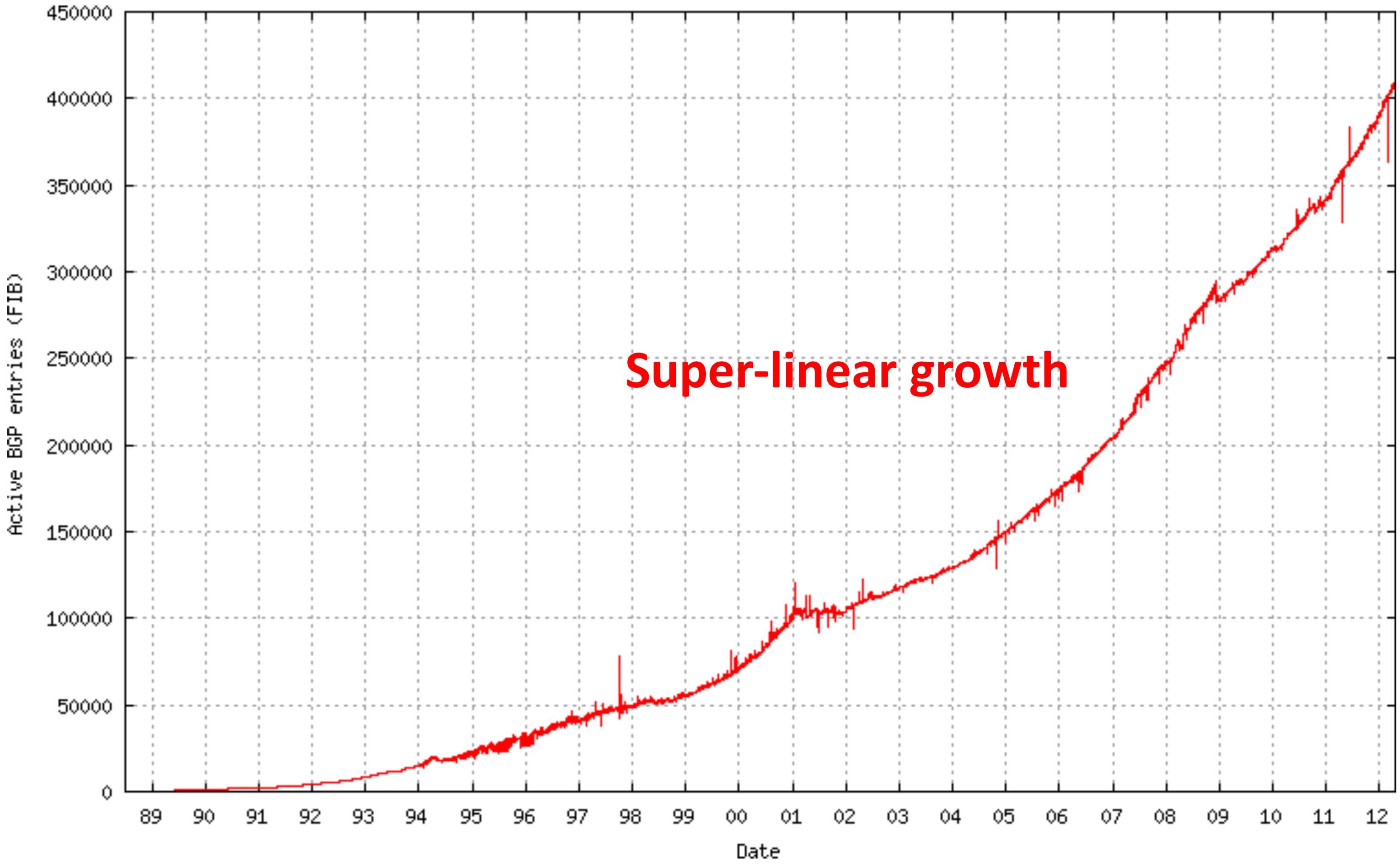
- Routers need to store information about all the destinations in the Internet
  - How does the number of prefixes evolve with time?
  - How is the number of prefixes influenced by the number of ASes?
  - How does the number of AS evolve with time?
- Are the current routing mechanisms enough?
- What’s the effect of all these on the routing tables?

# More Autonomous Systems




[src: www.potaroo.net]

# More Prefixes




[src: www.potaroo.net]

# Why So Many Prefixes? (1)

- More than half of the prefixes are /24's
  - The smallest possible – small companies
- Multihoming
  - More than half of the stubs are multihomed
  - Prefixes can't be aggregated
- Traffic Engineering
  - Many ASes perform prefix de-aggregation

# Why So Many Prefixes? (2)

- IANA allocation policy
  - Initially classful (/8, /16, /24) allocation and few constraints
    - Space quickly exhausted
  - Classless Inter-Domain Routing (CIDR) allocation
    - Provider Aggregatable (PA) addresses to providers
      - Owned by provider and **leased** to clients
      - Aggregatable
      - Renumbering PA is hard
    - Provider Independent (PI) addresses to stubs
      - Owned by stub and announced by providers
      - Not aggregatable

# Existing Routing Mechanisms

- Ingress TE is problematic
  - Deaggregation is limited to /24s and prone to aggregation
  - Prepending and deaggregation are tweaks

# Existing Routing Mechanisms

- Ingress TE is problematic
  - Deaggregation is limited to /24s and prone to aggregation
  - Prepending and deaggregation are tweaks

**Might not work and it affects the Internet's scalability!**

# Routing Table

- Reachability is announced to all Internet
  - The size of the FIB may become a problem
- Updates potentially propagate to everybody
  - Churn (fast sequence of updates) is a problem

# Routing Table

- Reachability is announced to all Internet
  - The size of the FIB may become a problem
    **Requires fast (expensive) memory**
- Updates potentially propagate to everybody
  - Churn (fast sequence of updates) is a problem
    **CPUs required to process updates**

# Summary of problems

- Many prefixes and their count is growing too fast
  - Memory limitations
  - CPU limitations
- Hard to perform ingress traffic engineering
- IPv4 address exhaustion
  - Simultaneous use of both IPv4 and IPv6 (dual stack) worsens the prefix count growth (need to store two routing tables)

# Summary of problems

- Many prefixes and their count is growing too fast
  - Memory limitations
  - CPU limitations
- Hard to perform ingress traffic engineering
- IPv4 address exhaustion
  - Simultaneous use of both IPv4 and IPv6 (dual stack) worsens the prefix count growth (need to store two routing tables)
- **And these are just the problems affecting the scalability of the inter-domain routing!**

# Solutions

- **Disruptive (clean-slate) solutions**
  – Rethink the paradigms
- **Evolutionary solutions**
  – Enhance current architecture
  – Provide inter-working mechanisms

# Internet Scalability

## Representative Solutions

## Clean-Slate Architectures: NNC

# Networking Named Content

- Motivation/History
- Content Centric Networking
- Transport
- Routing
- Evaluation

# History

- Internet engineering principles: ‘60s-’70s
- Problem to be solved: resource sharing
- Communication model: point-to-point conversation
- Central abstraction host identifier

# Issues (1)

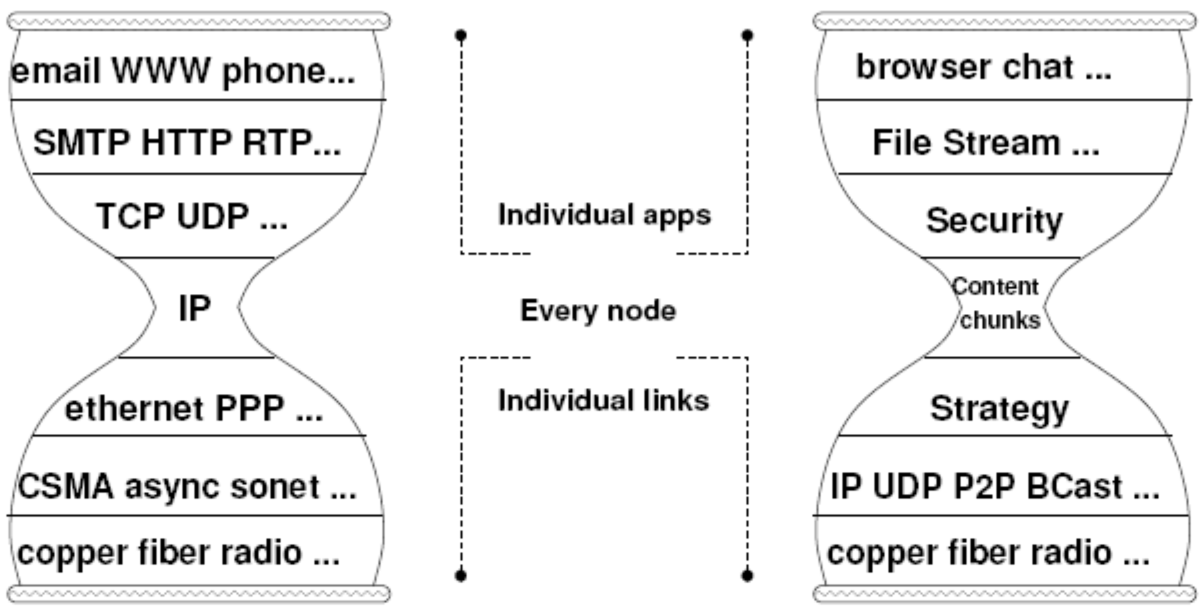
- Content Availability
  - CDNs, P2P networks
  - Excessive bandwidth costs
- Security
  - Trust is easily misplaced
  - Connection information can be false
- Content location dependence
  - Content mapped to host locations

# Issues (2)

- where vs. what
  - Wrong abstraction for obtaining content
- named hosts vs. named data
  - Proposed change for the communication abstraction
- host-to-host vs. many-to-many
  - Evolve from the old ‘server-client’ model

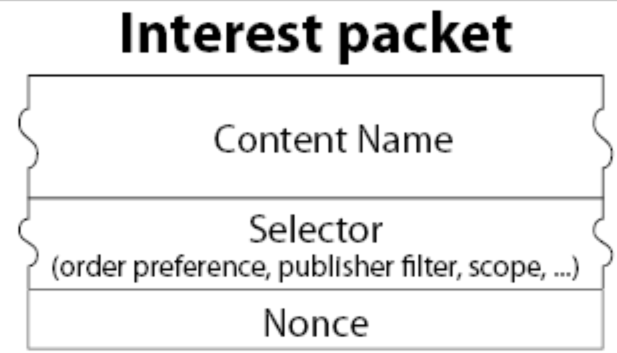
# Content Centric Networking

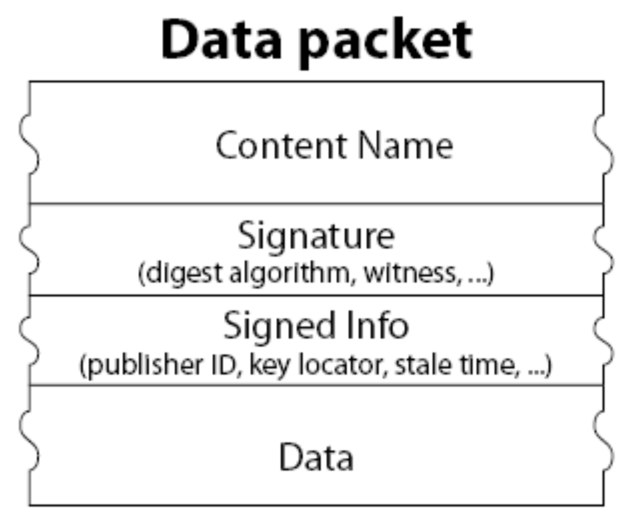
- Communication built on named data
- No notion of hosts



# Content Node Model (1)

- Two packet types
  - Consumers broadcast **Interests** over all interfaces
  - Nodes that receive interests and possess data to satisfy interests reply with **Data Packets**
  - Data satisfies interests if the interest's ContentName is a prefix for the ContentName in the data packet
  - Content that satisfies an interest might be generated on-the-fly (active content)
- Content (Data) Name
  - Follows the hierarchical model used with IP
  - Explicit structure
  - Context dependent
    - Global: /upc.edu/fcoras
    - Local: /thisroom/projector



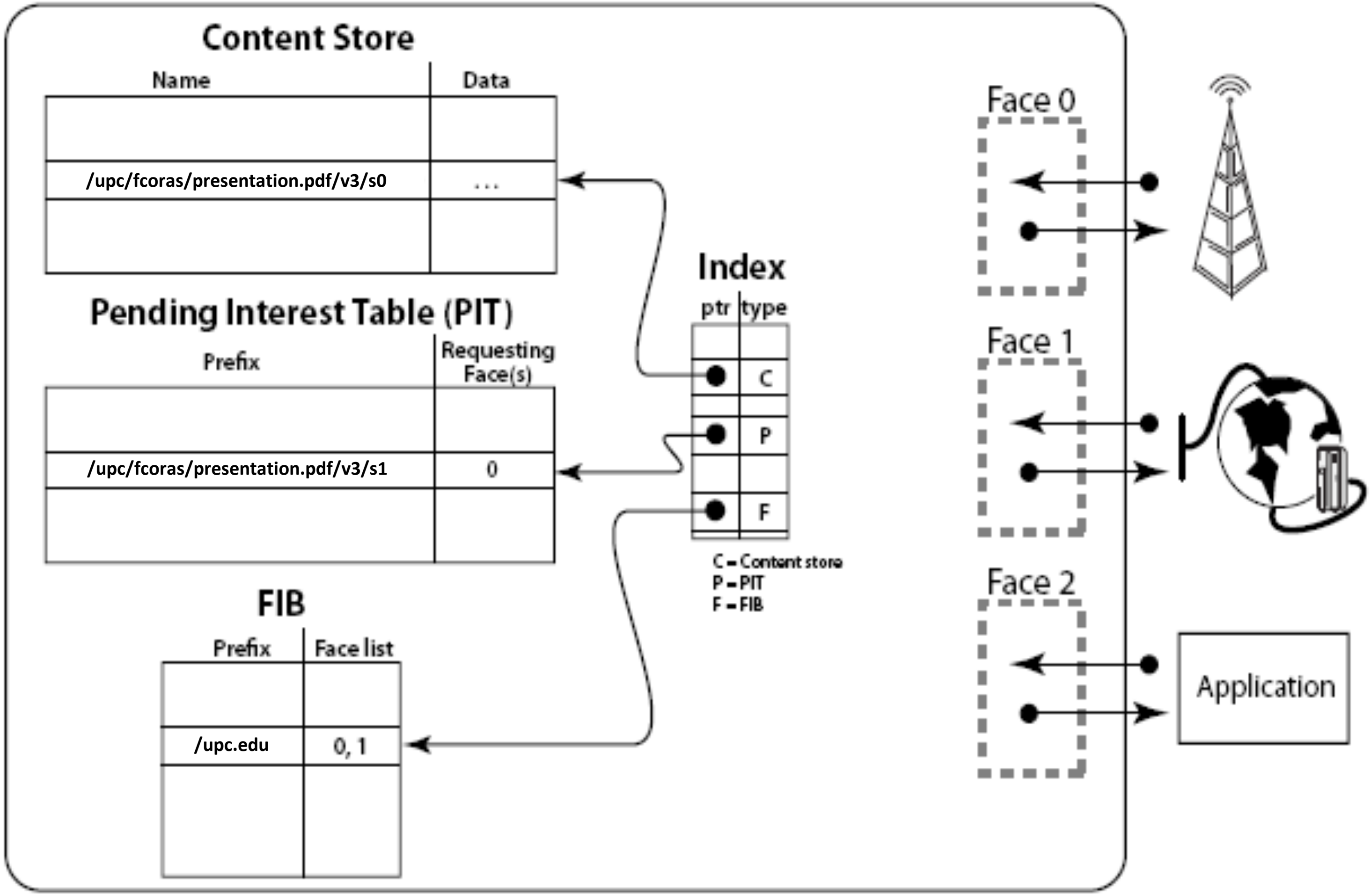
# Content Node Model (2)

- Interest and Data packets are one-for-one
  - Strict flow balance similar to the one in TCP
- Longest match lookups for each packet
  - Fast lookups due to the explicit structure of data names
- Data packets carry security information
  - Self-certifying names
- interface vs. face

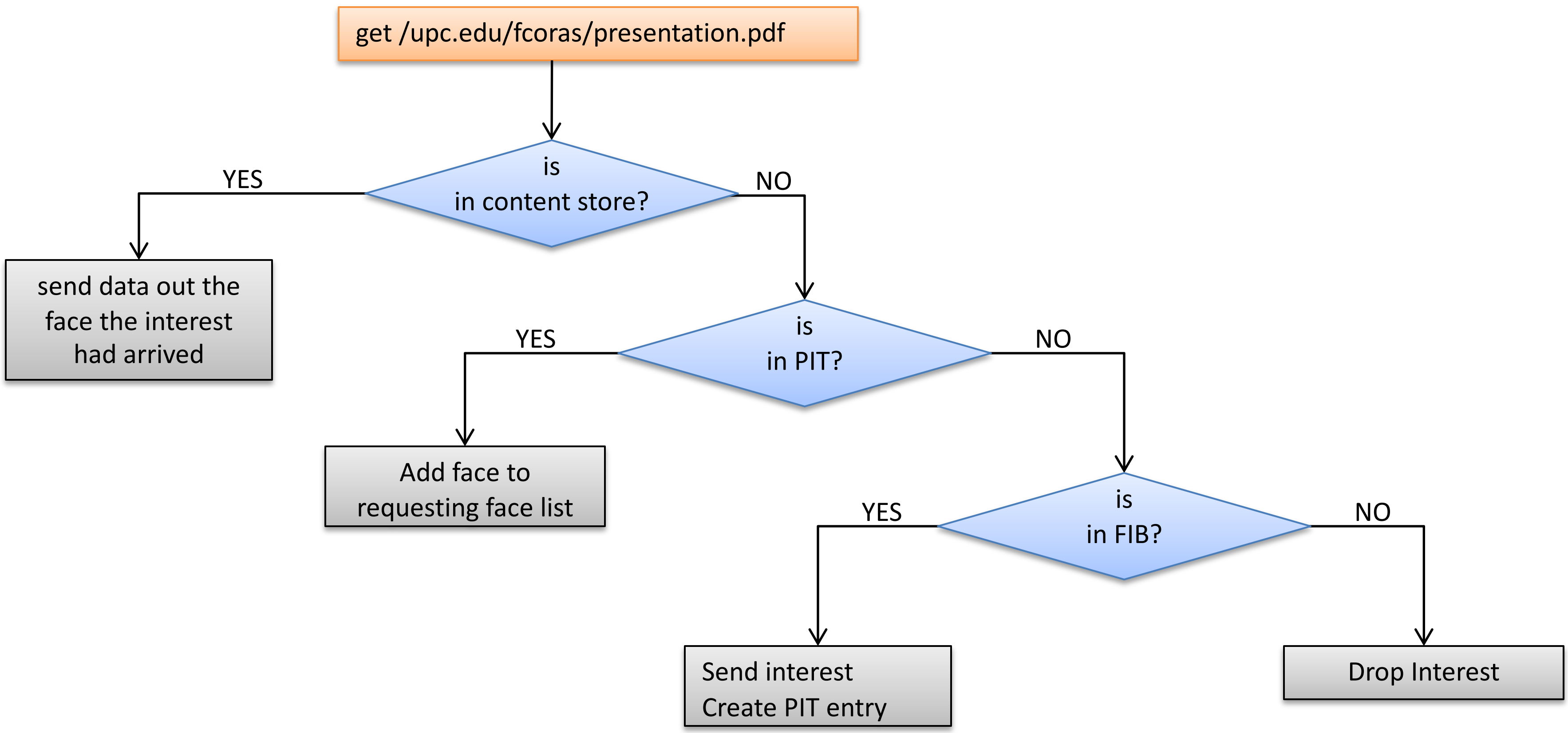
# Overview: Forwarding Engine (1)

- Forwarding Information Base (FIB)
  - Forwards interests towards potential sources of matching Data
  - Similar to an IP router FIB
  - Not constrained to just one (inter)face
  - Query multiple sources of data in parallel
- Content Store
  - Similar to the buffer memory of an IP router
  - Replacement policies may be LRU/LFU in contrast to MRU used for IP
- Pending Interest Table(PIT)
  - Keeps track of interests sent upstream (form a ‘trail’)
  - The entries are erased as soon as Data packets that satisfy them arrive

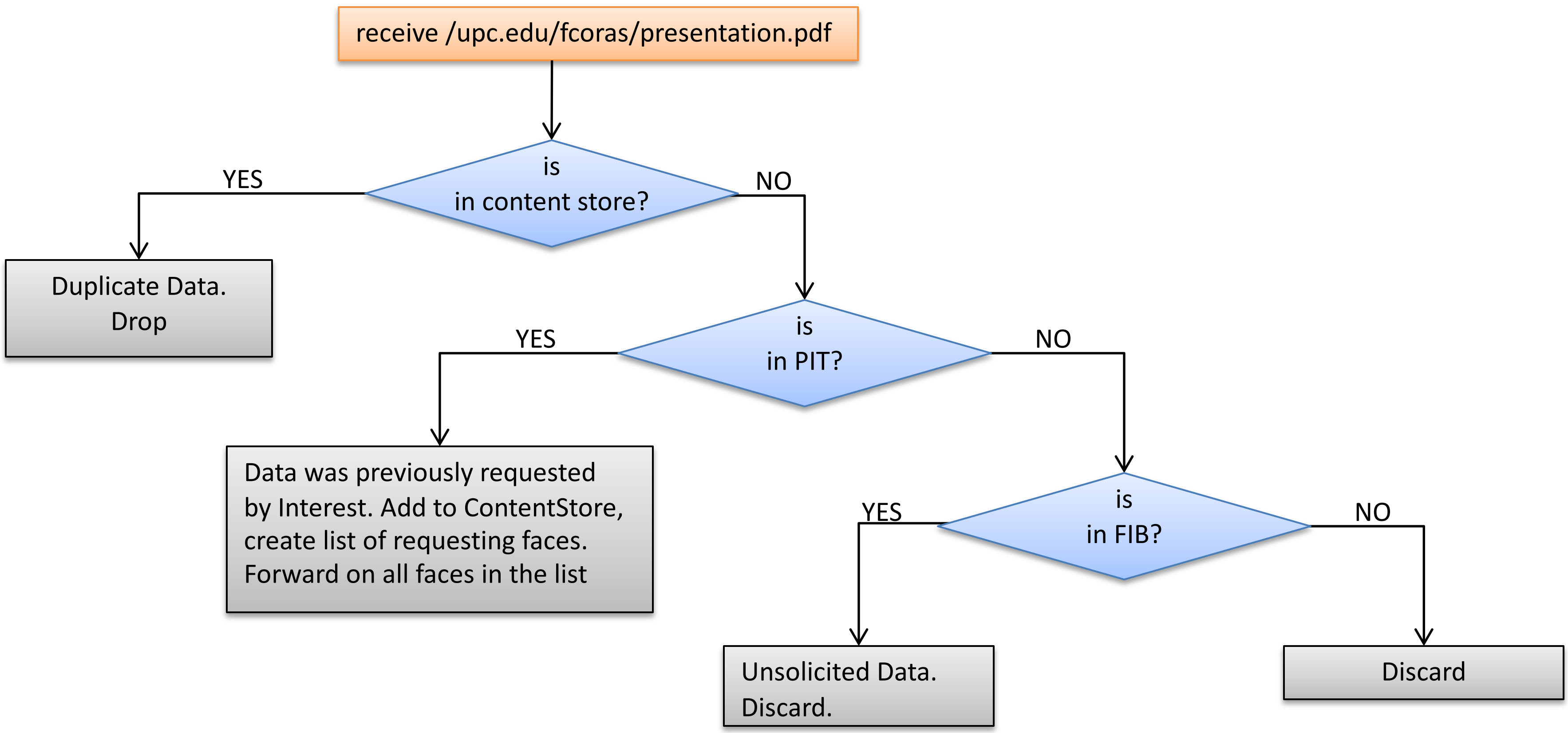
# Overview: Forwarding Engine(2)



# Overview: Interest Forwarding



# Overview: Data Forwarding

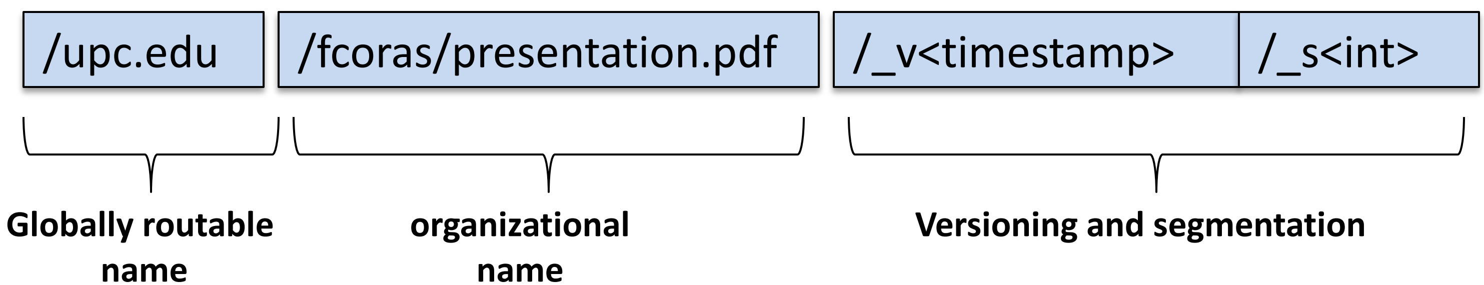


# Transport (1)

- Operates on top of unreliable packet delivery services
- Flow Control
  - CCN senders are stateless -> consumer must require retransmission (controlled by strategy layer)
  - Interests perform same flow control and sequencing as TCP Acks
    - Interests are like window advertisments in TCP
    - Flow control through Interests pipelining
  - CCN operates in “local” flow balance. TCP in end-to-end flow balance
- CCN can take advantage of multiple interfaces
  - Strategy layer controls how Interests are forwarded

# Transport: Sequencing (2)

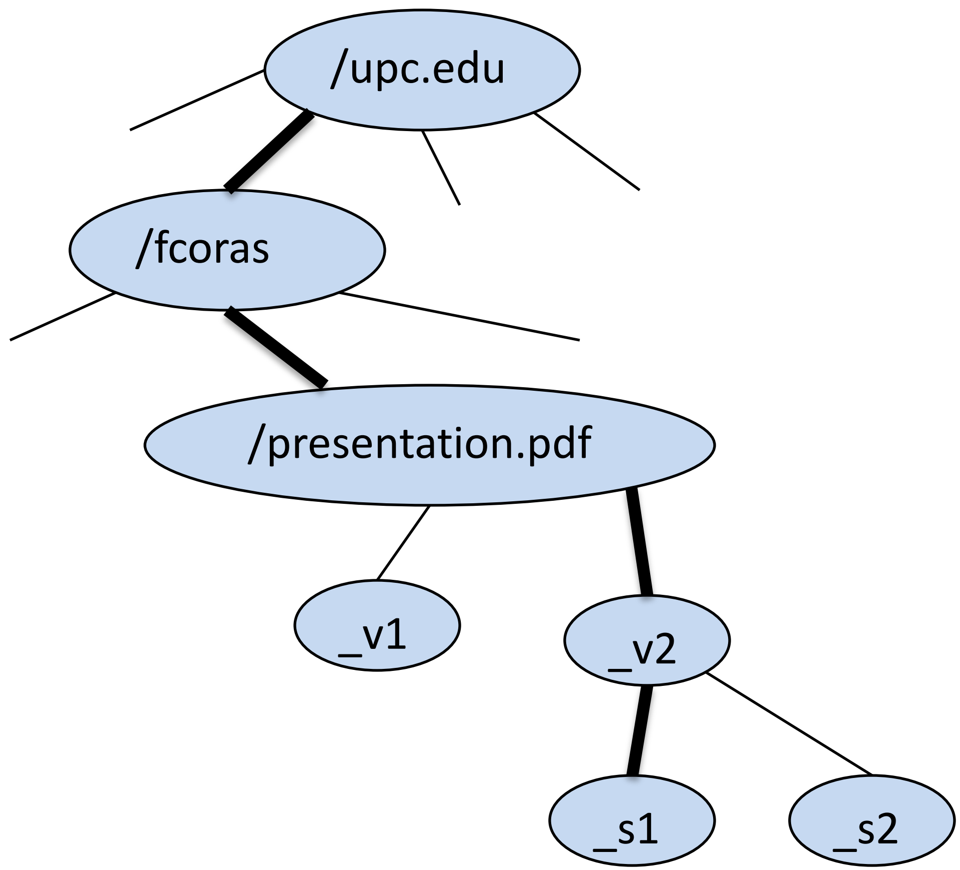
- CCN names are composed of *components*




- "relative" access to data in totally ordered tree
  - previous, next, RightmostChild…

# Transport (3)

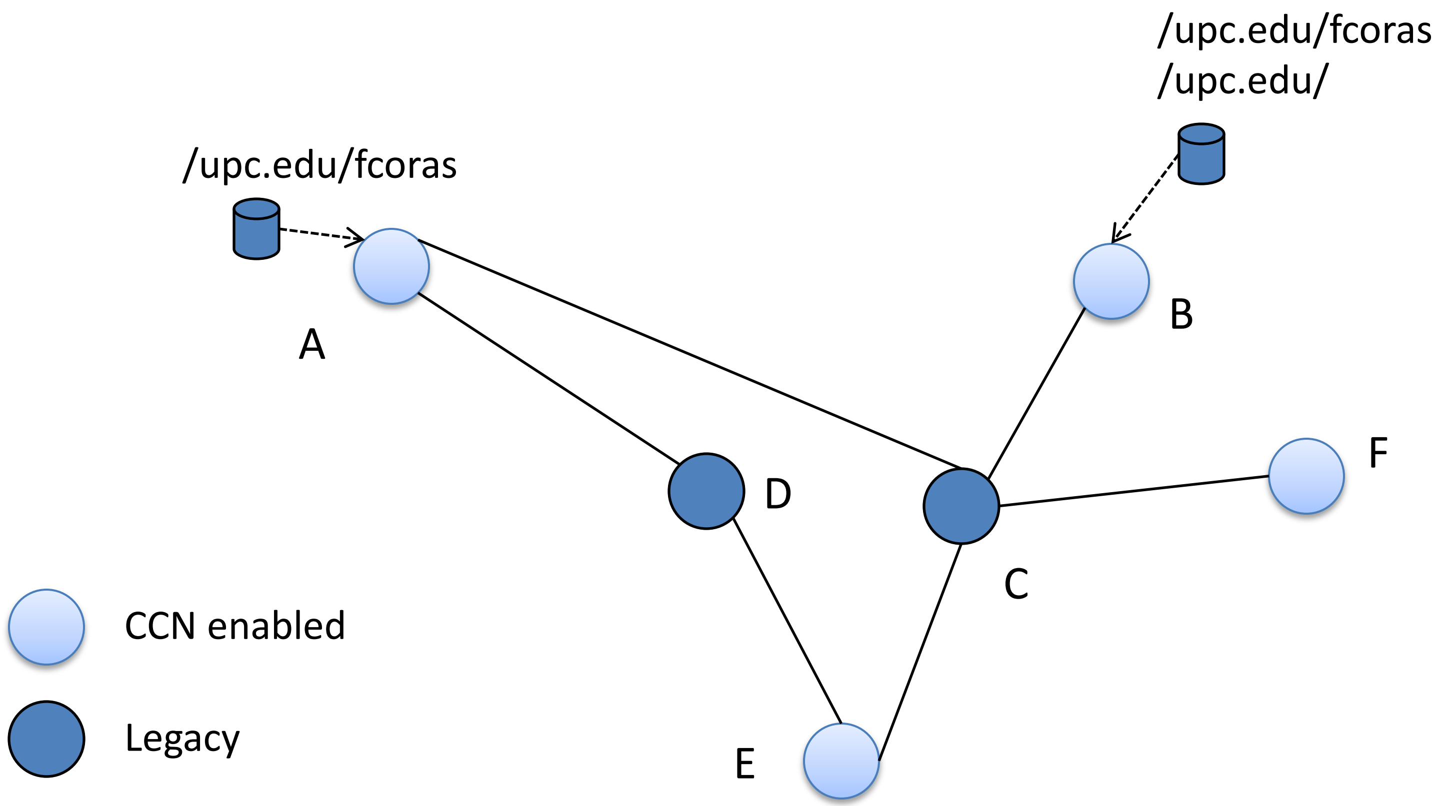
**Name tree traversal for: /upc.edu/fcoras/presentation.pdf/_v2**



# Routing: Intra-Domain (1)

- Works with link state IGP: OSPF, IS-IS
- Customize {Type, Label, Value} link-state announcements to advertise ContentName Prefixes
  - Unknown LSAs are dropped by default so incremental deployment is possible
- Behavioral difference from IP
  - In IP, a prefix advertiser can reach all hosts in the prefix
  - CCN does not need shortest path tree
  - In CCN a content advertiser can reach **some** of the content matching a prefix

# Routing (2)




| /upc.edu/ | B |
| --- | --- |
| /upc.edu/fcoras | A,B |

# Routing: Inter-Domain (3)

- BGP has the equivalent of IGP TLV
  - Use domain-level content prefixes
- Topology at the AS rather than network prefix level could be built

# Open Issues

- Caching
  - How much memory is enough?
  - What is the best replacement policy?
  - How are chunk sizes computed?
- Economical model
- Inter-domain routing scalability

# Bibliography


- [1] Y. Rekhter, T. Li, S. Hares, “A Border Gateway Protocol 4 (BGP-4)”, RFC 4271
- [2] J. Hawkinson, T. Bates, “Guidelines for creation, selection, and registration of an Autonomous System (AS)”, RFC 1930
- [3] D. Meyer, L. Zhang, K. Fall, “Report from the IAB Workshop on Routing and Addressing”, RFC 4984
- [4] Rexford, J. and Dovrolis, C. “Future internet architecture: clean-slate versus evolutionary research”, Communications of the ACM, vol. 59, pp. 36-40, 2010
- [5] Jacobson, V.; Smetters, D. K.; Thornton, J. D.; Plass, M. F.; Briggs, N.; Braynard, R. Networking named content. Communications of the ACM. 2012 January; 55 (1): 117-124.
- [6] http://www.parc.com/work/focus-area/content-centric-networking/