# 5617, Spring 2019
# computer networking and communication

anduo wang, Temple University

TTLMAN 401 A, R 17:30-20:00

# to do

## homework 5

- due April 18
- submit in class

# routing in the wild

# Interdomain Traffic Engineering with BGP

https://inl.info.ucl.ac.be/system/files/commag-may2003.pdf

# why traffic engineering (TE)

## the (research) Internet

- designed with best effort service in mind
  - connectivity was the most important issue
- but, the best effort service was used for mission critical applications with stringent service level agreement (SLA)

## to meet SLAs

# why traffic engineering (TE)

## the (research) Internet

- designed with best effort service in mind
  - connectivity was the most important issue
- but, the best effort service was used for mission critical applications with stringent service level agreement (SLA)

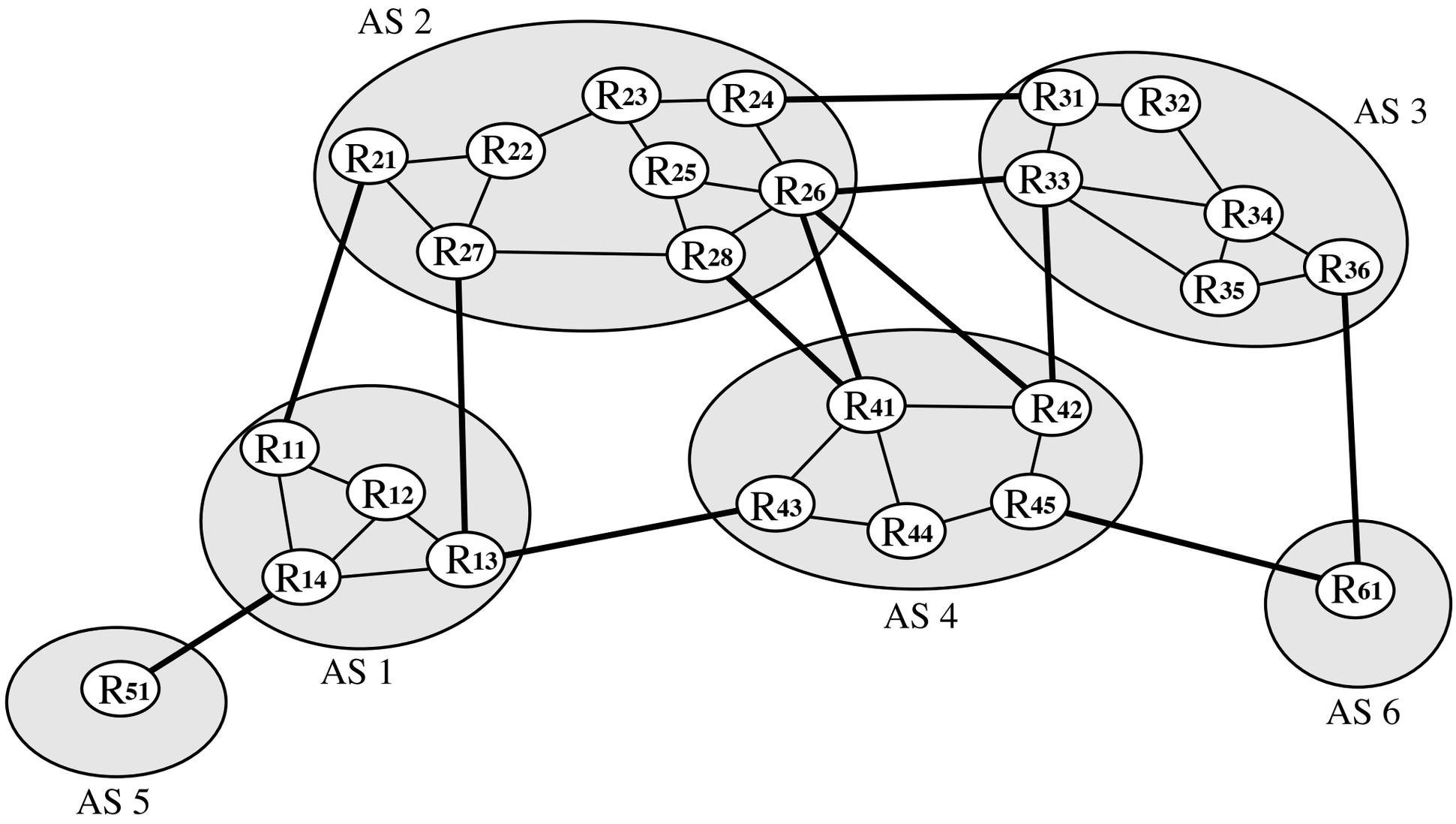
## to meet SLAs, ISPs rely on TE to better control inter domain traffic

- by tuning the configuration of BGP

# interdomain TE with BGP

what — better control the flow of interdomain packets inside an IP network

- control the flow of incoming and outgoing traffic



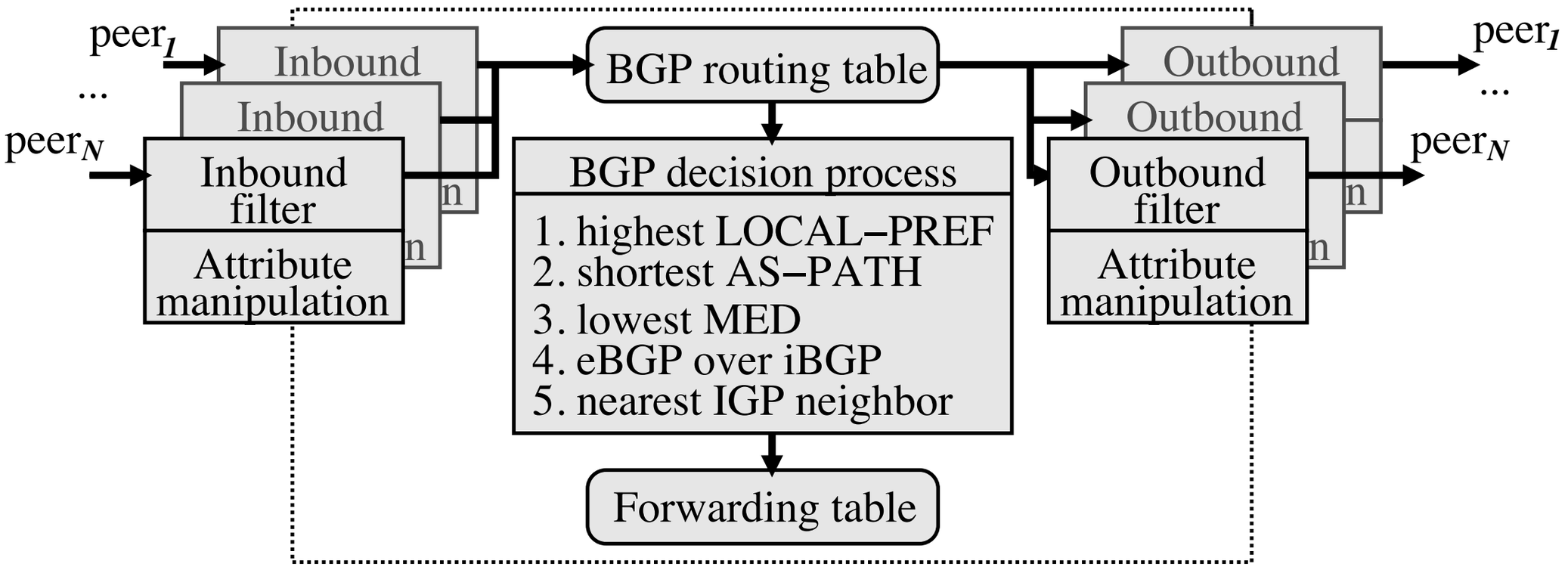
# interdomain TE with BGP

what — better control the flow of interdomain packets inside an IP network

- control the flow of incoming and outgoing traffic
- careful running of the route advertisements sent via BGP

# BGP route

route = (prefix, next_hop, AS_path, optional attributes)



# characteristics of interdomain traffic

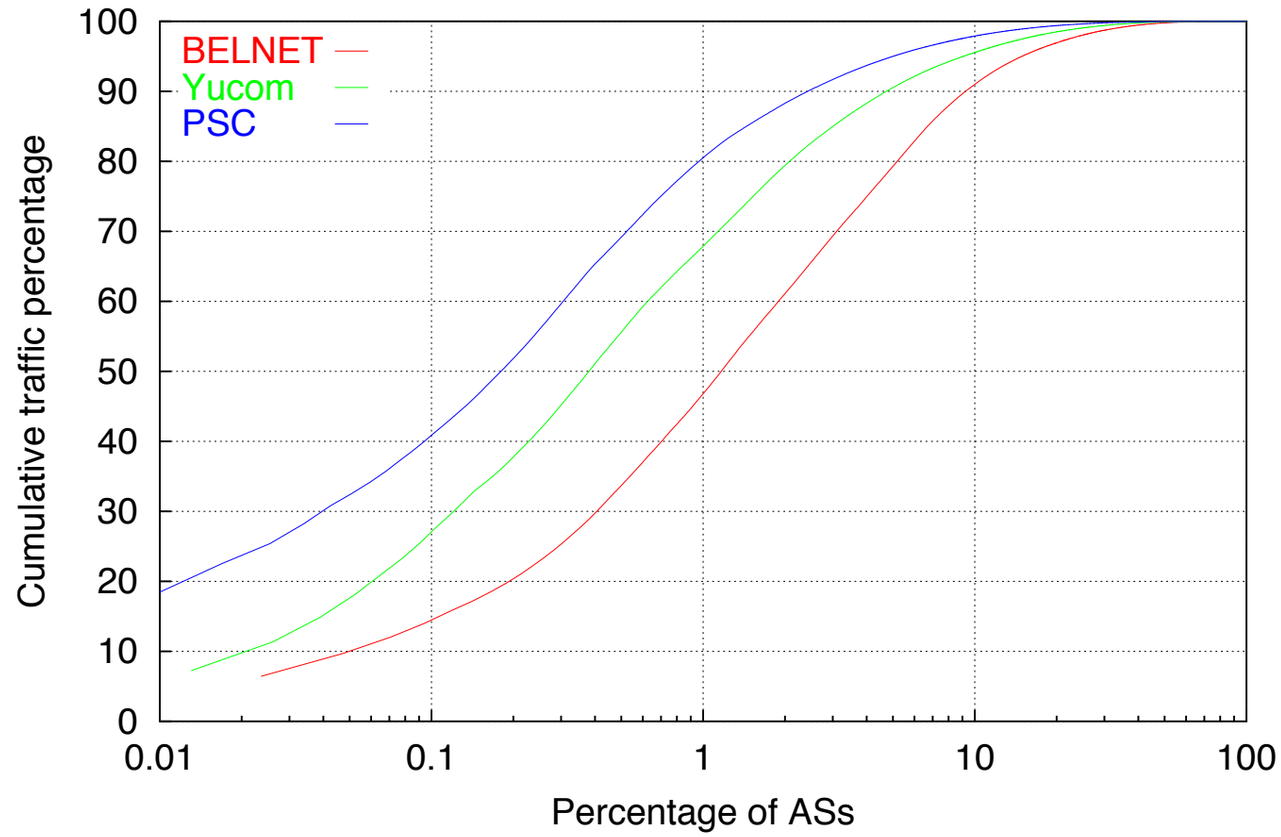
each ISP exchanges IP packets with a large fraction of the Internet

interdomain TE would appear difficult? (since an AS needs to influence most of the Internet to control its traffic?

# characteristics of interdomain traffic

cumulative distribution of the traffic by an ISP

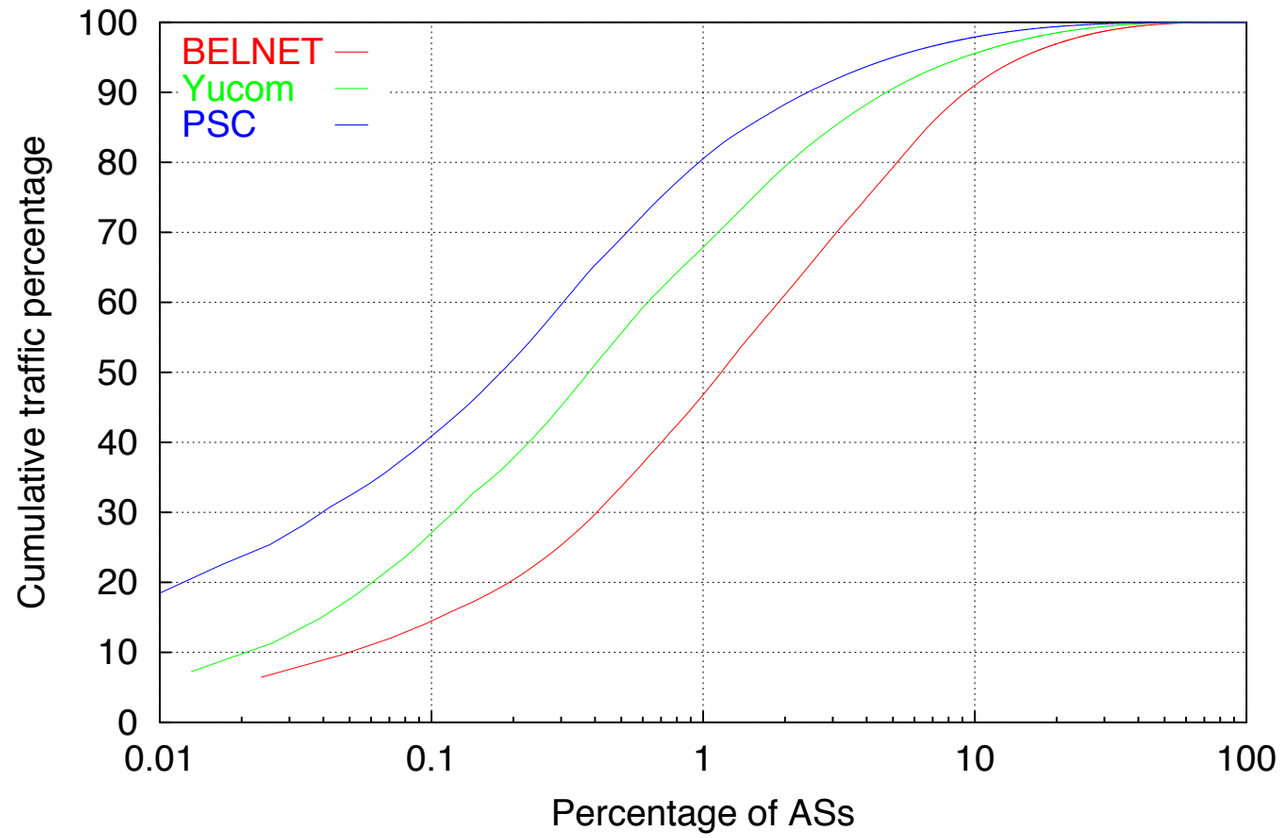
- does not exchange the same amount of traffic with each remote AS



# characteristics of interdomain traffic

## cumulative distribution of the traffic by an ISP

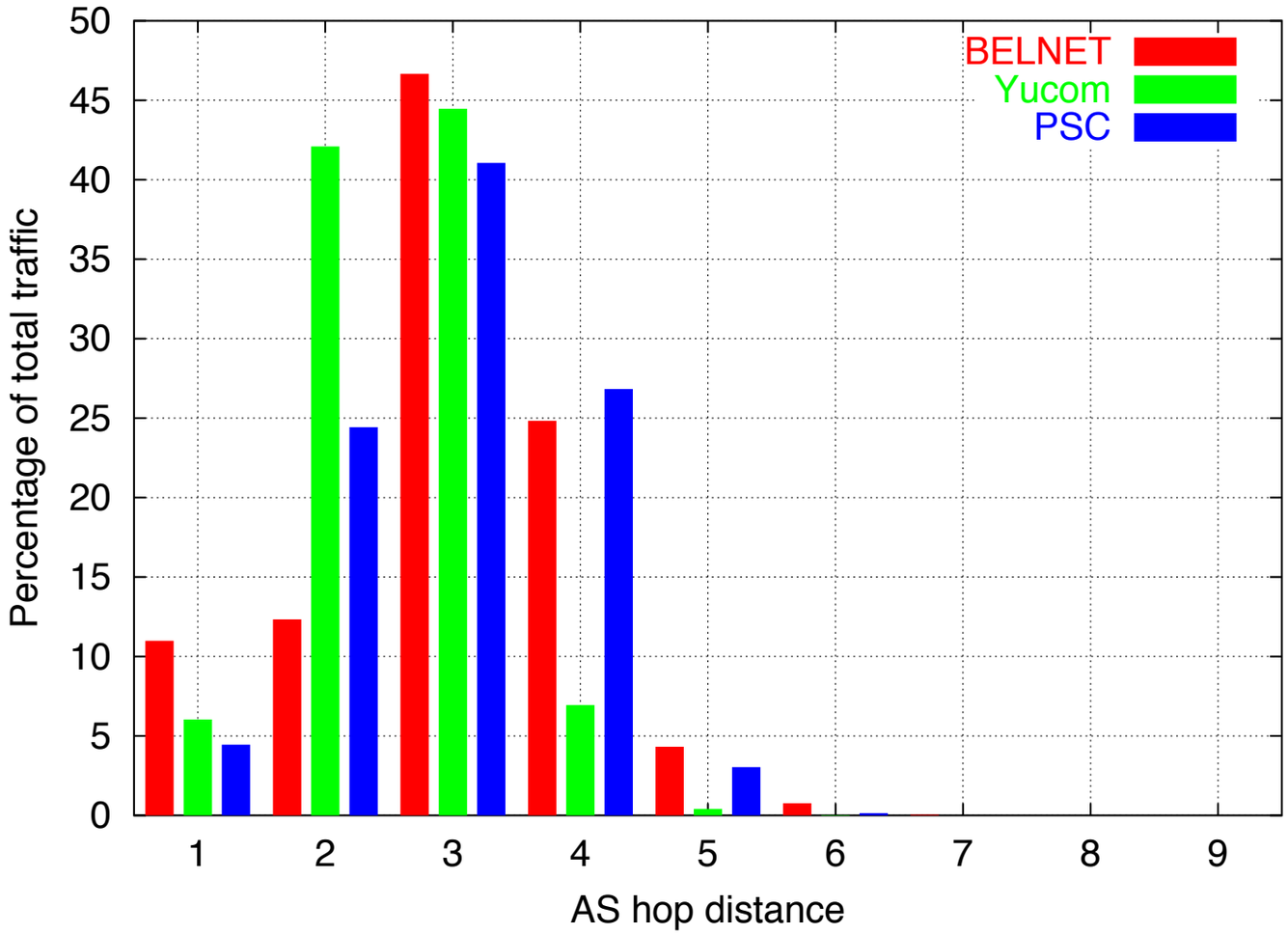
- does not exchange the same amount of traffic with each remote AS
- Yucom — a Belgium ISP
  - top (resp, 100) largest sources contributes for >30% (resp, 72%) traffic



# characteristics of interdomain traffic

## AS path length

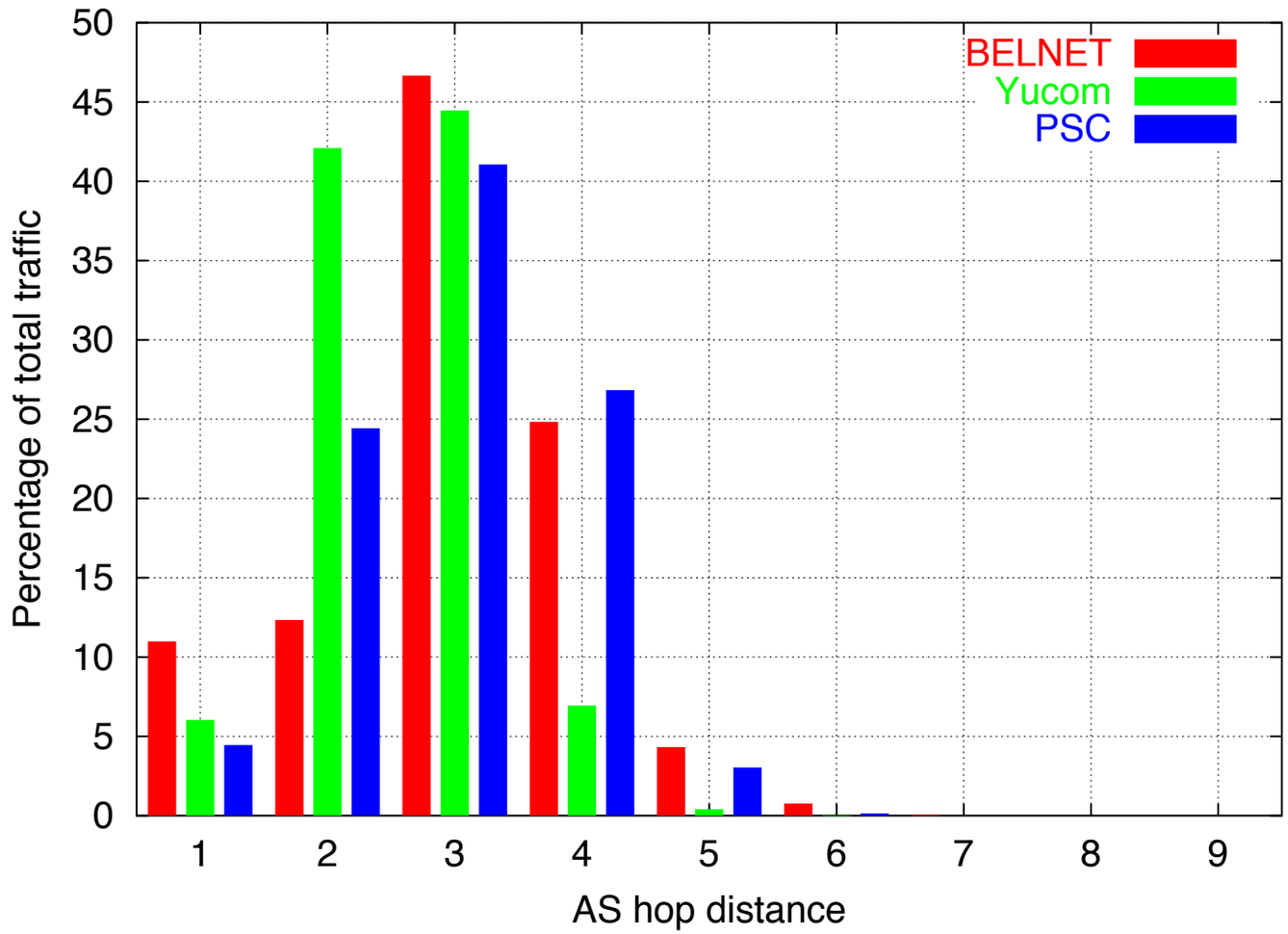
- most of the packets are exchanged with ASes that are only a few hops away



# characteristics of interdomain traffic

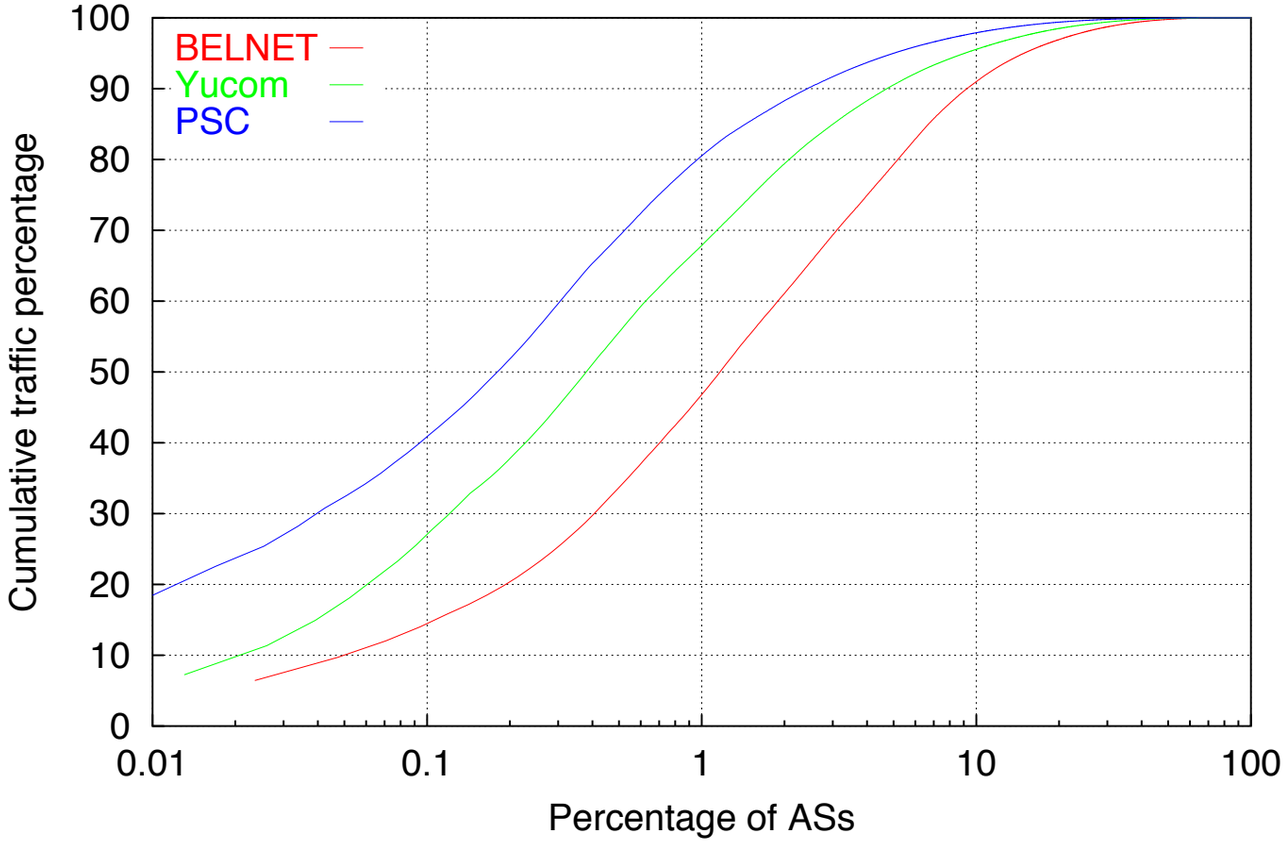
## AS path length

- most of the packets are exchanged with ASes that are only a few hops away
- Yucom
- receives traffic from sources two or three hops away



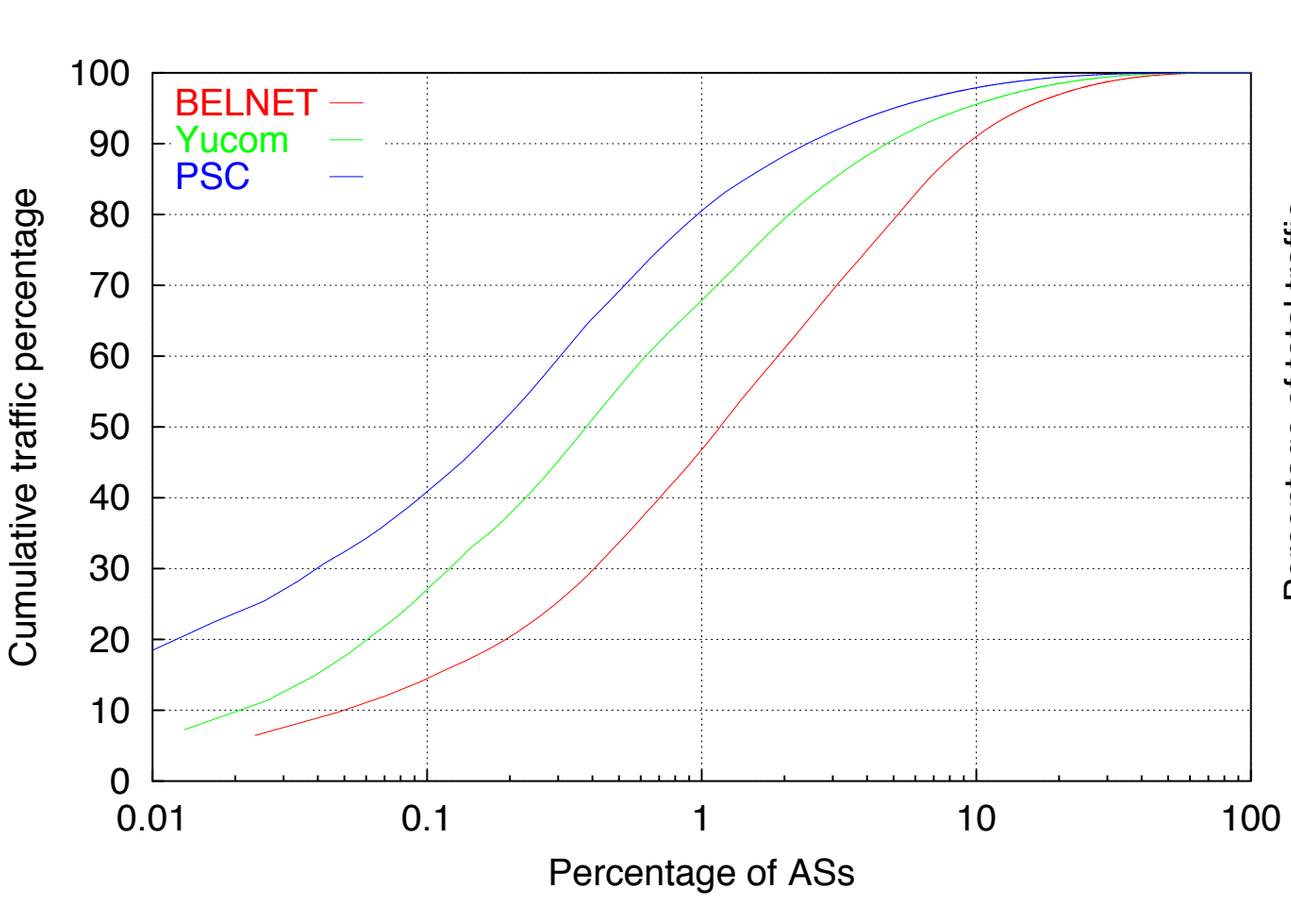
# impact ons TE

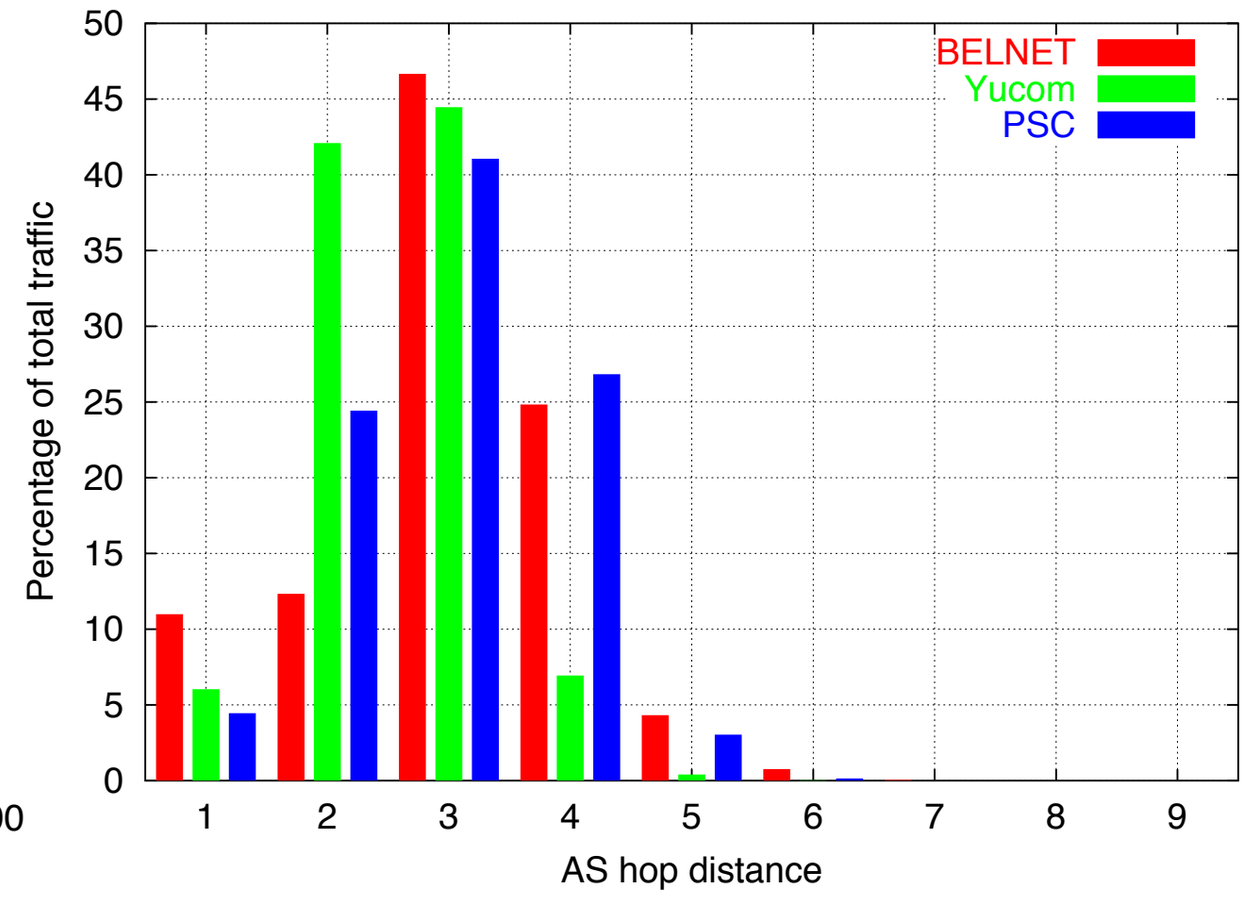
- an AS can move a large amount of traffic by influencing a small number of distant ASes



# impact ons TE

- an AS can move a large amount of traffic by influencing a small number of distant ASes
- an AS needs to influence distinct ASes a few hops away beyond their upstream providers and direct peers



# interdomain TE requirements

diverse, but (often) motivated by the need to

- balance traffic on links with other ASes
- reduce the cost of carrying traffic on those links

depend on

- connectivity of an AS with others
- type of business handled by the local AS

# interdomain TE requirements

| | | optimizes |
|---|---|---|
| content provider | have several customer -provider relationships with transit ASes | how traffic leaves |
| access provider | | how traffic enters |
| transit AS | carry traffic on behalf of others | balance traffic on multiple links it has with its peers |

# interdomain TE requirements

| | | optimizes |
|---|---|---|
| content provider | have several customer -provider relationships with transit ASes | how traffic leaves |
| access provider | | how traffic enters |
| transit AS | carry traffic on behalf of others | balance traffic on multiple links it has with its peers |

performed by tweaking BGP routes of the AS

optimize the way traffic enters or leaves

# interdomain TE requirements

| | | optimizes |
|---|---|---|
| content provider | have several customer -provider relationships with transit ASes | how traffic leaves |
| access provider | | how traffic enters |
| transit AS | carry traffic on behalf of others | balance traffic on multiple links it has |

favor one link over another to a given destination or receive traffic from a given source

performed by tweaking BGP routes of the AS

optimize the way traffic enters or leaves

# control outgoing traffic

control how traffic leaves



local AS chooses which route to use through its peers

# control outgoing traffic

control how traffic leaves

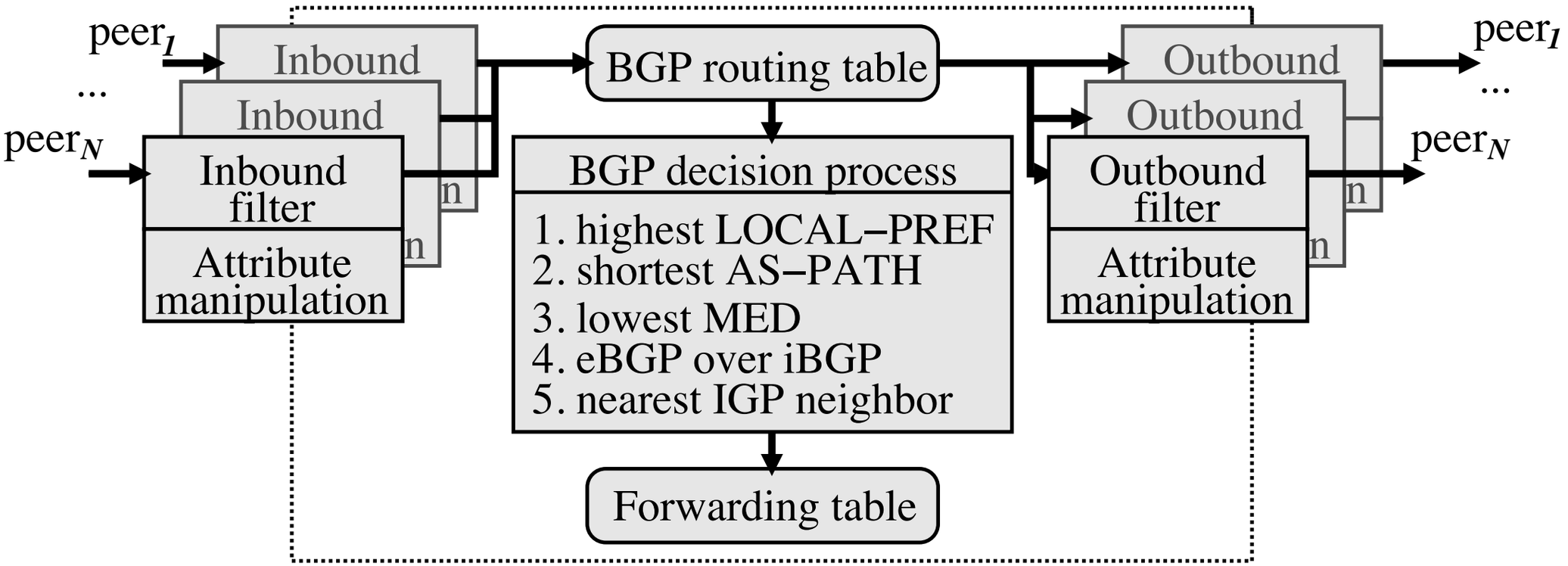
local AS chooses which route to use through its peers

easy: an AS only needs to control the local BGP decision process via local preference

# control outgoing traffic

control how traffic leaves → local AS chooses which route to use through its peers

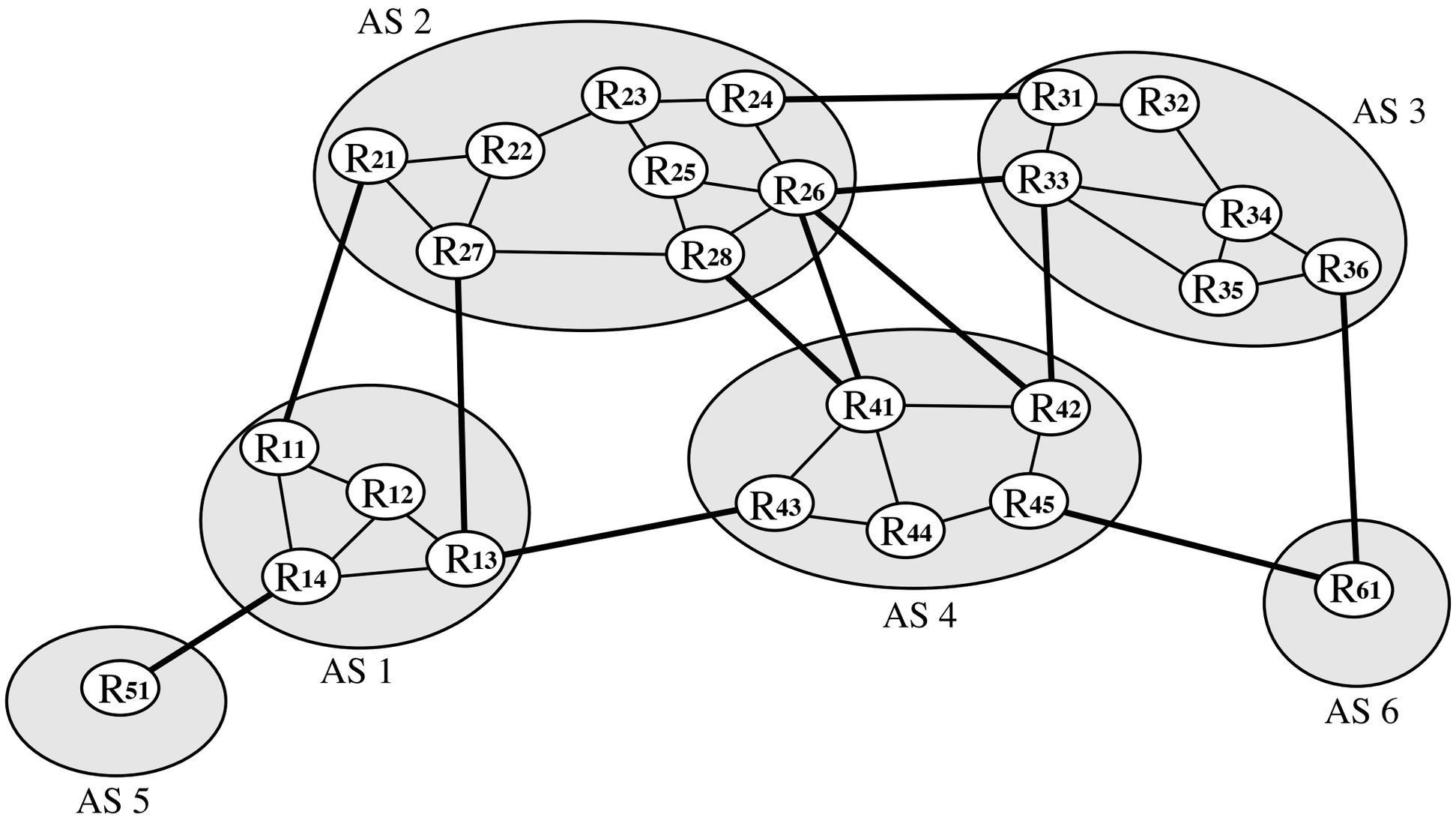
easy: an AS only needs to control the local BGP decision process via local preference



# control outgoing traffic

control how traffic leaves

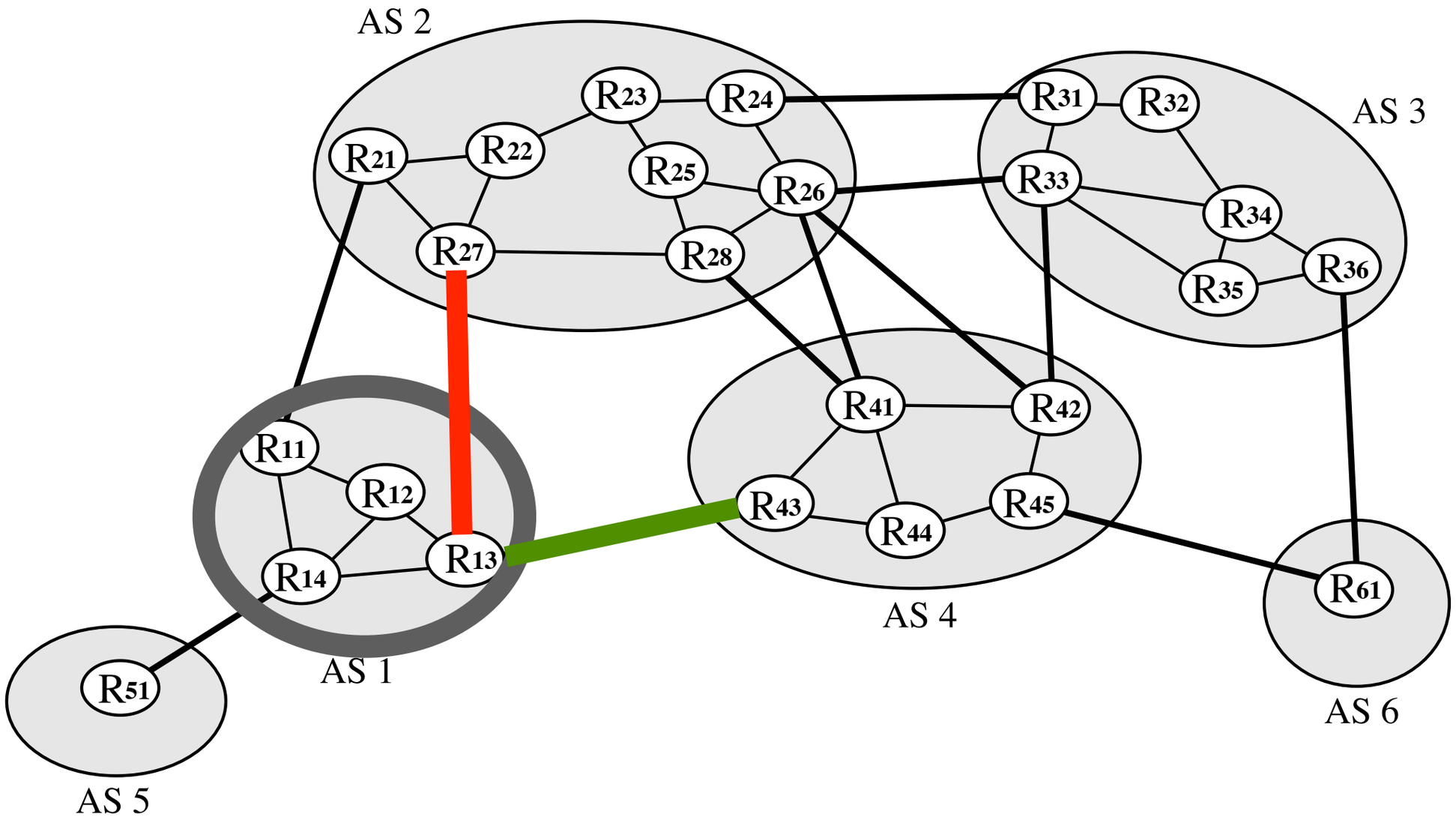
→ local AS chooses which route to use through its peers



# control outgoing traffic

control how traffic leaves

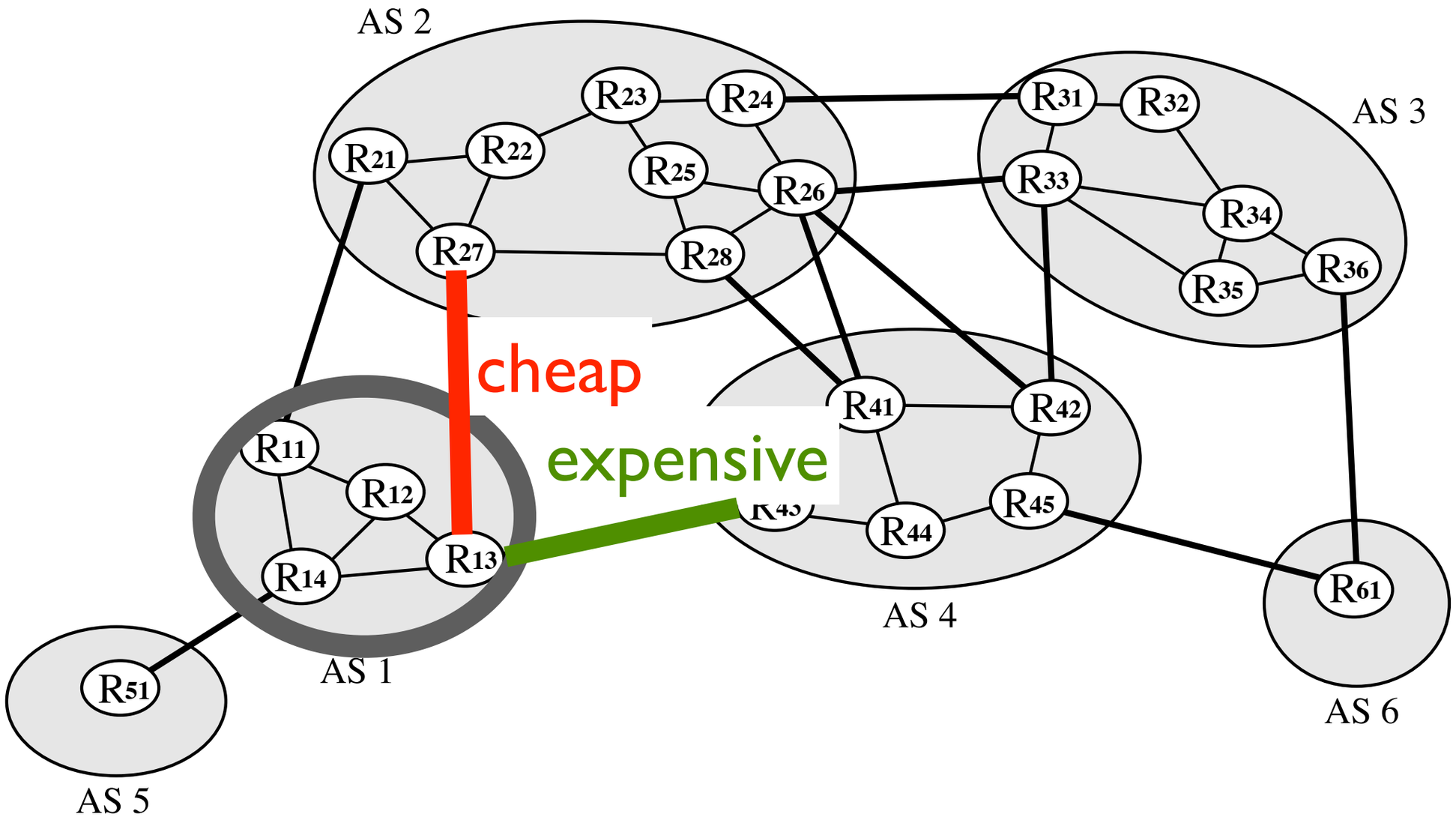
→ local AS chooses which route to use through its peers



# control outgoing traffic

control how traffic leaves

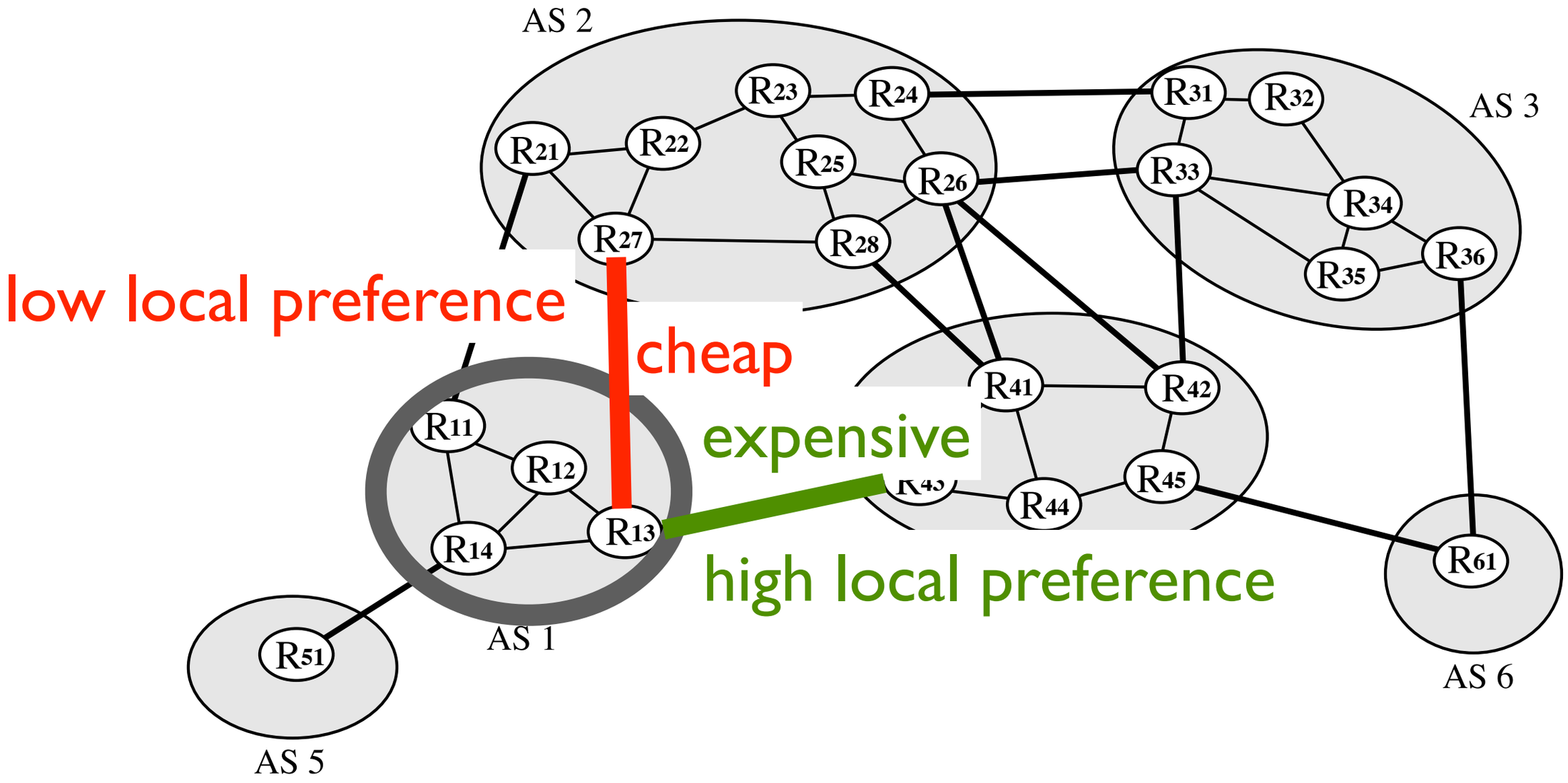
→ local AS chooses which route to use through its peers



# control outgoing traffic

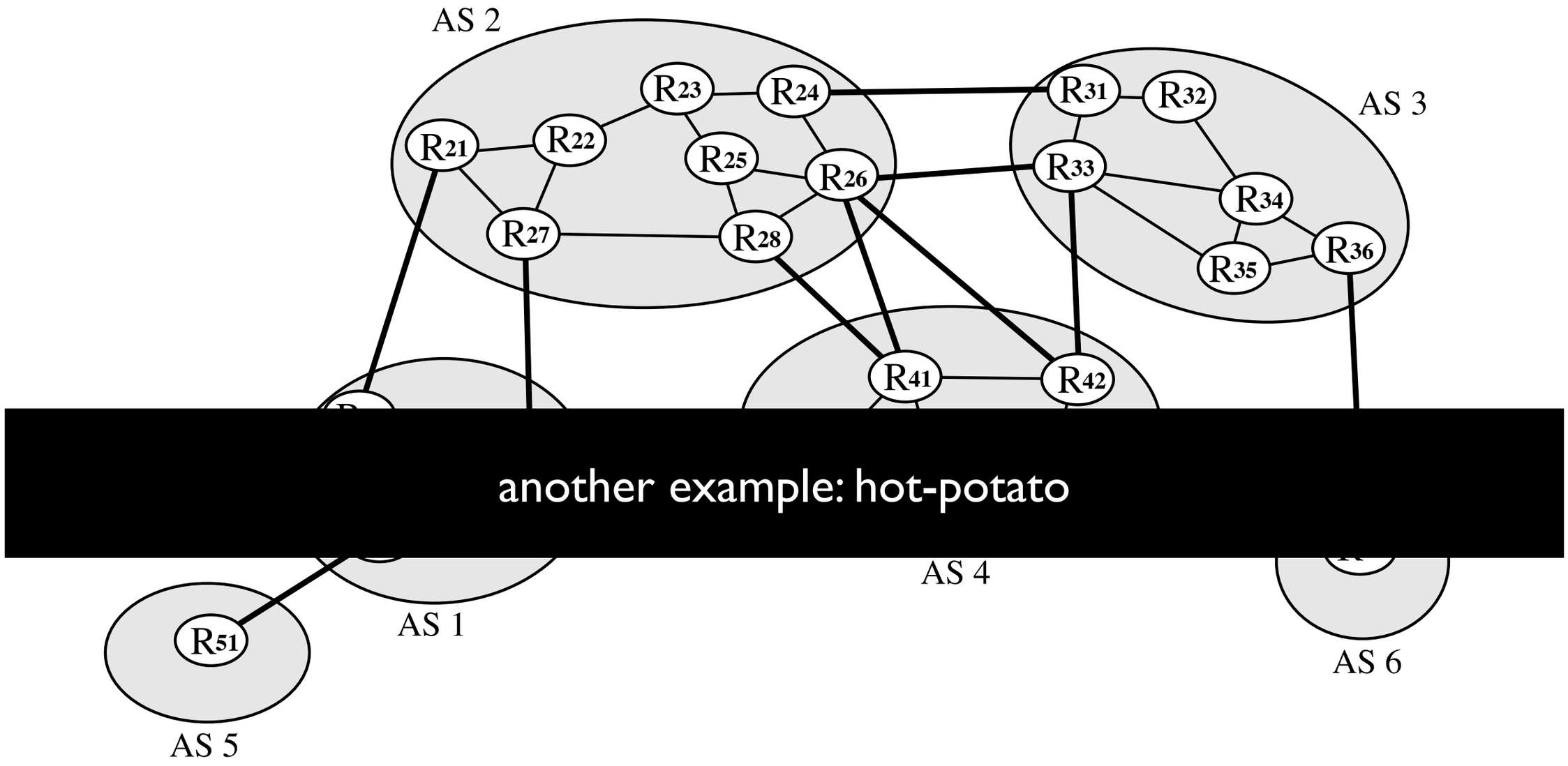
control how traffic leaves

→ local AS chooses which route to use through its peers



# control outgoing traffic

control how traffic leaves → local AS chooses which route to use through its peers




another example: hot-potato

# control incoming traffic

control how traffic enters

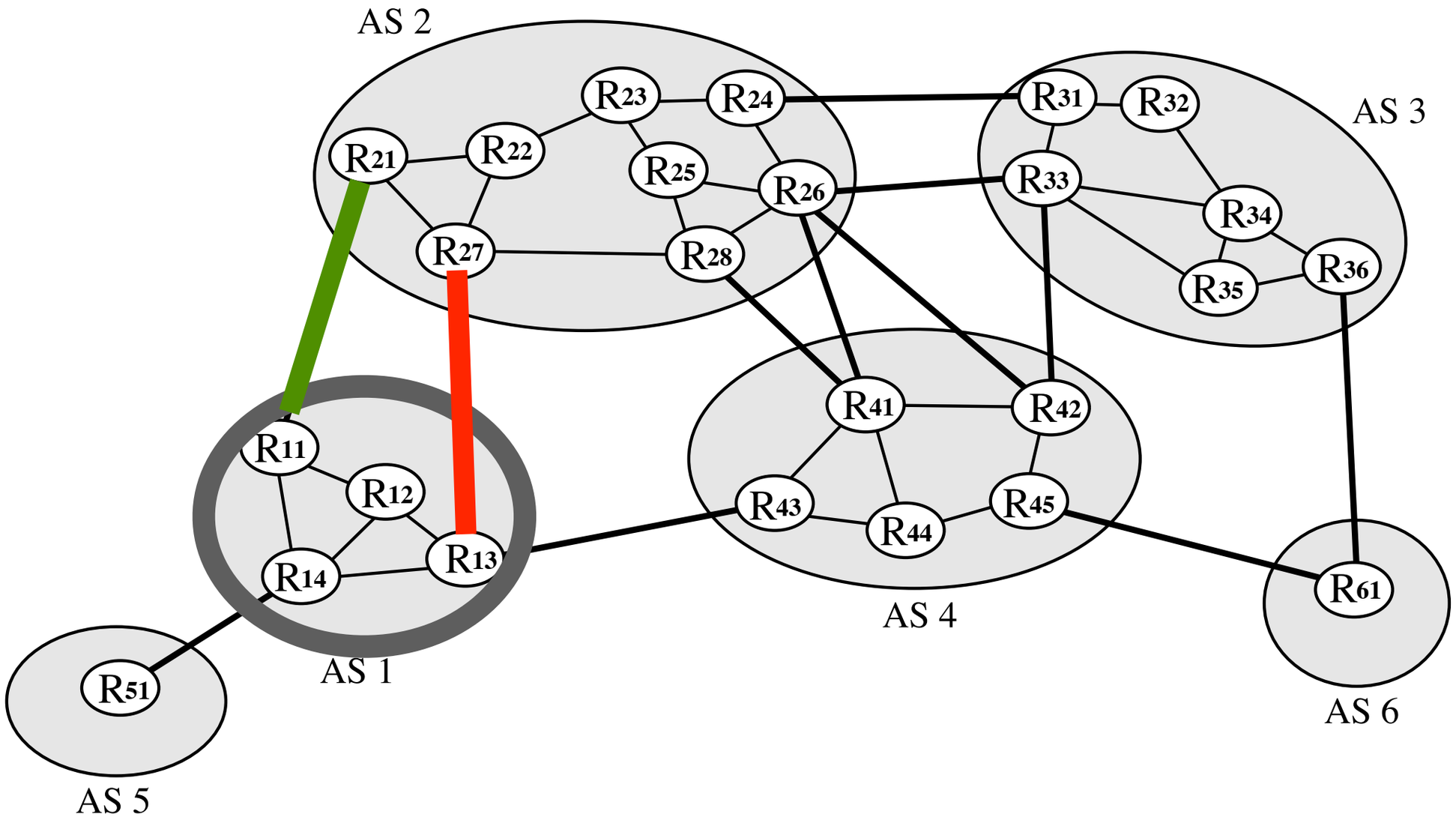


selective advertisements & announce different routes on different links

# control incoming traffic

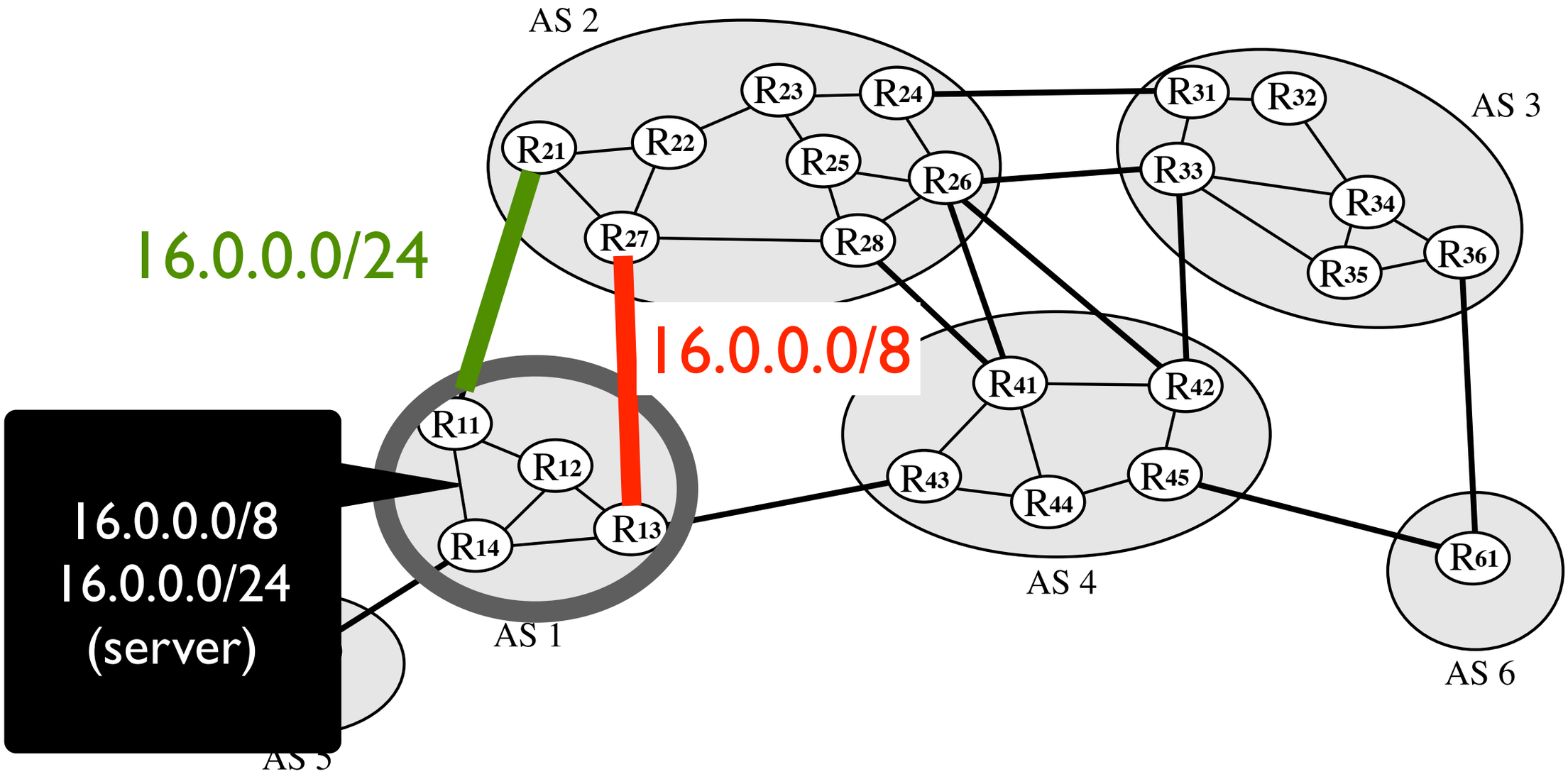
control how traffic enters

→ selective advertisements & announce different routes on different links



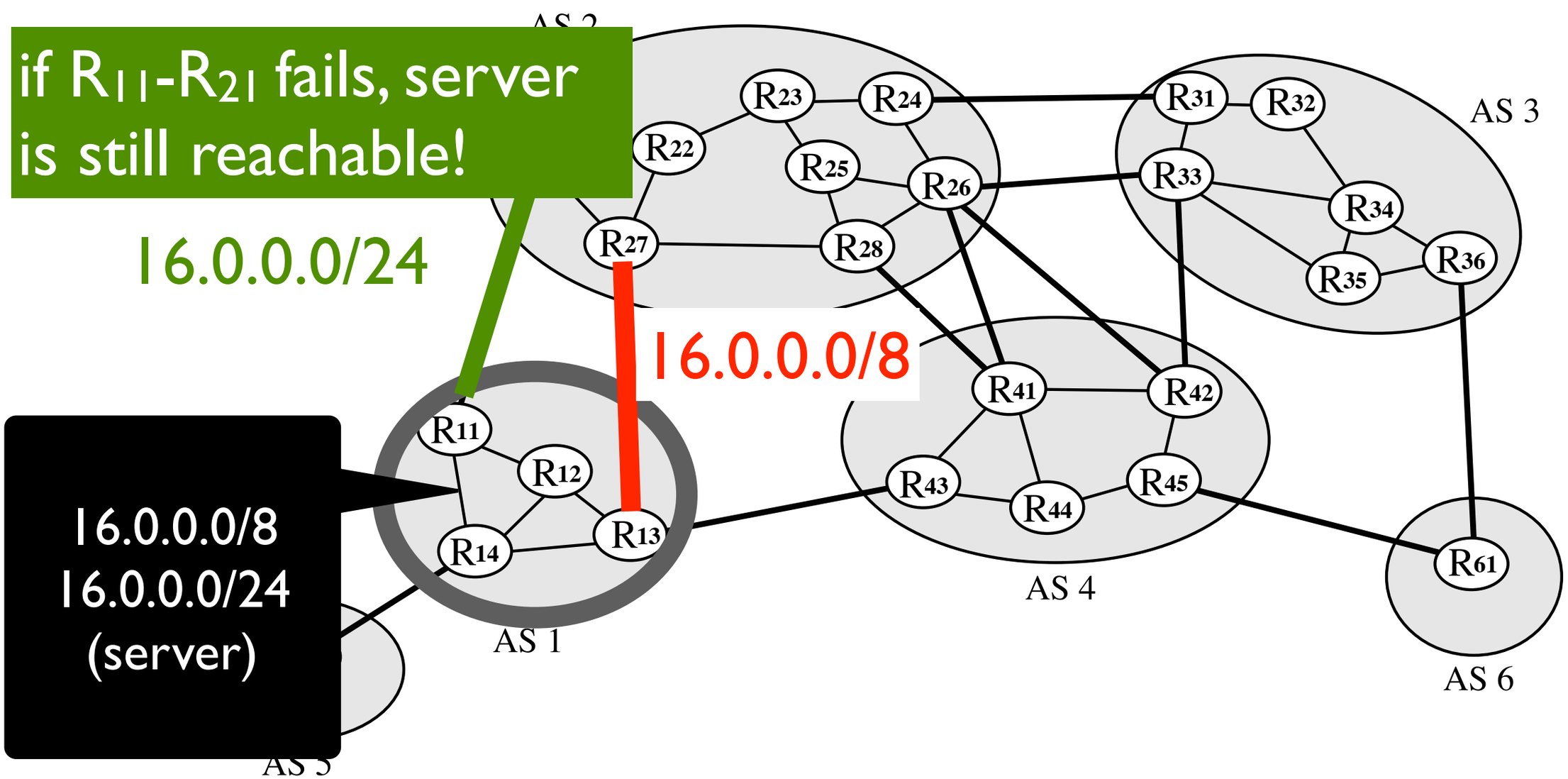
# control incoming traffic

control how traffic enters → (variant) advertise more specific prefixes



# control incoming traffic

control how traffic enters → (variant) advertise more specific prefixes



# limitation

control outgoing traffic based on the selection of best route

- limited by (availability) diversity of routes received from upstream providers

# limitation

control outgoing traffic based on the selection of best route

- limited by (availability) diversity of routes received from upstream providers

control incoming traffic with more specific prefixes

- all prefixes propagated throughout the Internet
  - inflating BGP table
  - instability

# Commentary on Inter-domain Routing

https://tools.ietf.org/html/rfc3221

# goal

the longer term trends of the BGP table

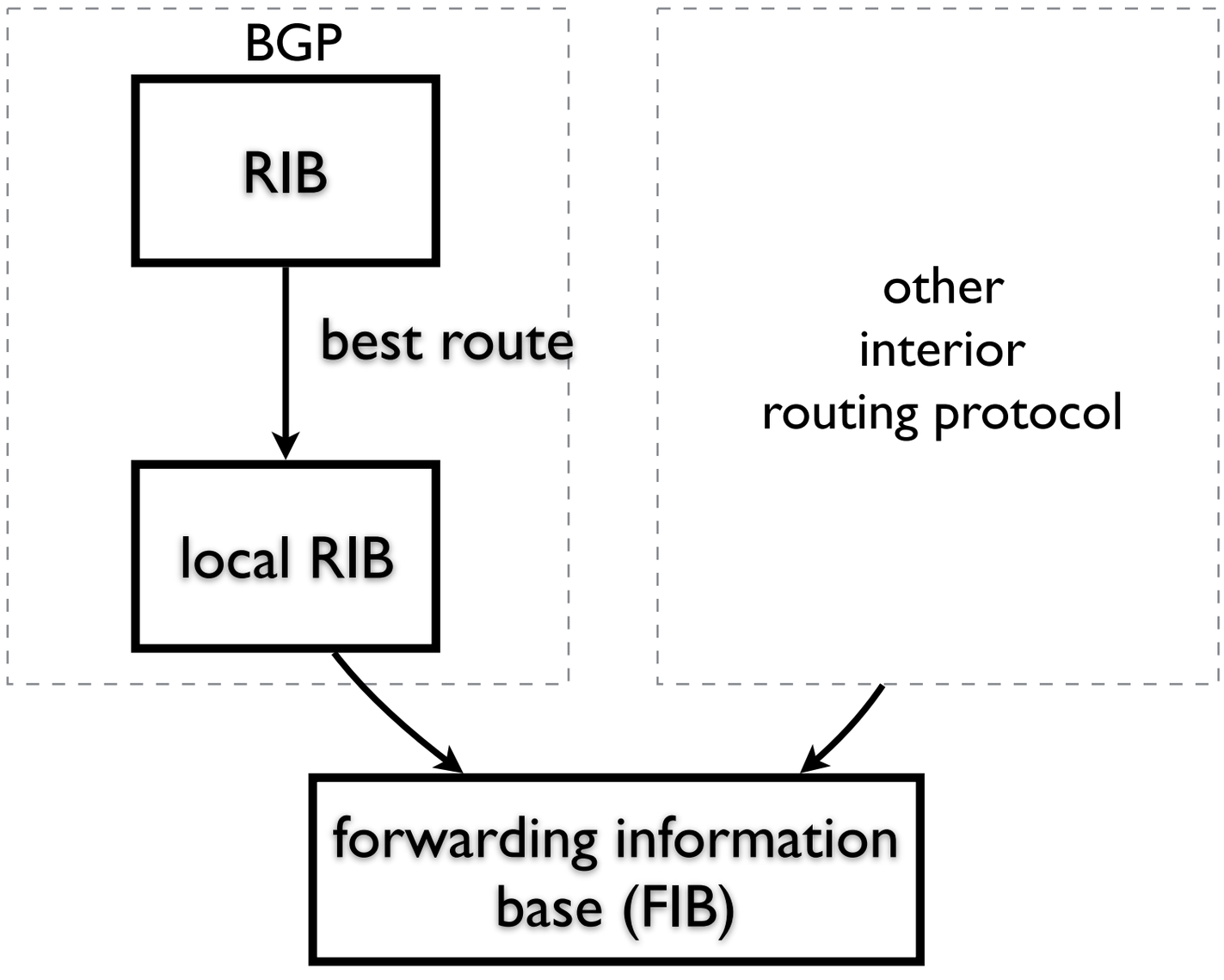
- understand the visible characteristics
- identify the contributing sources

impacts on the ability of the Internet to scale

# BGP routing table

## routing inforation base (RIB)

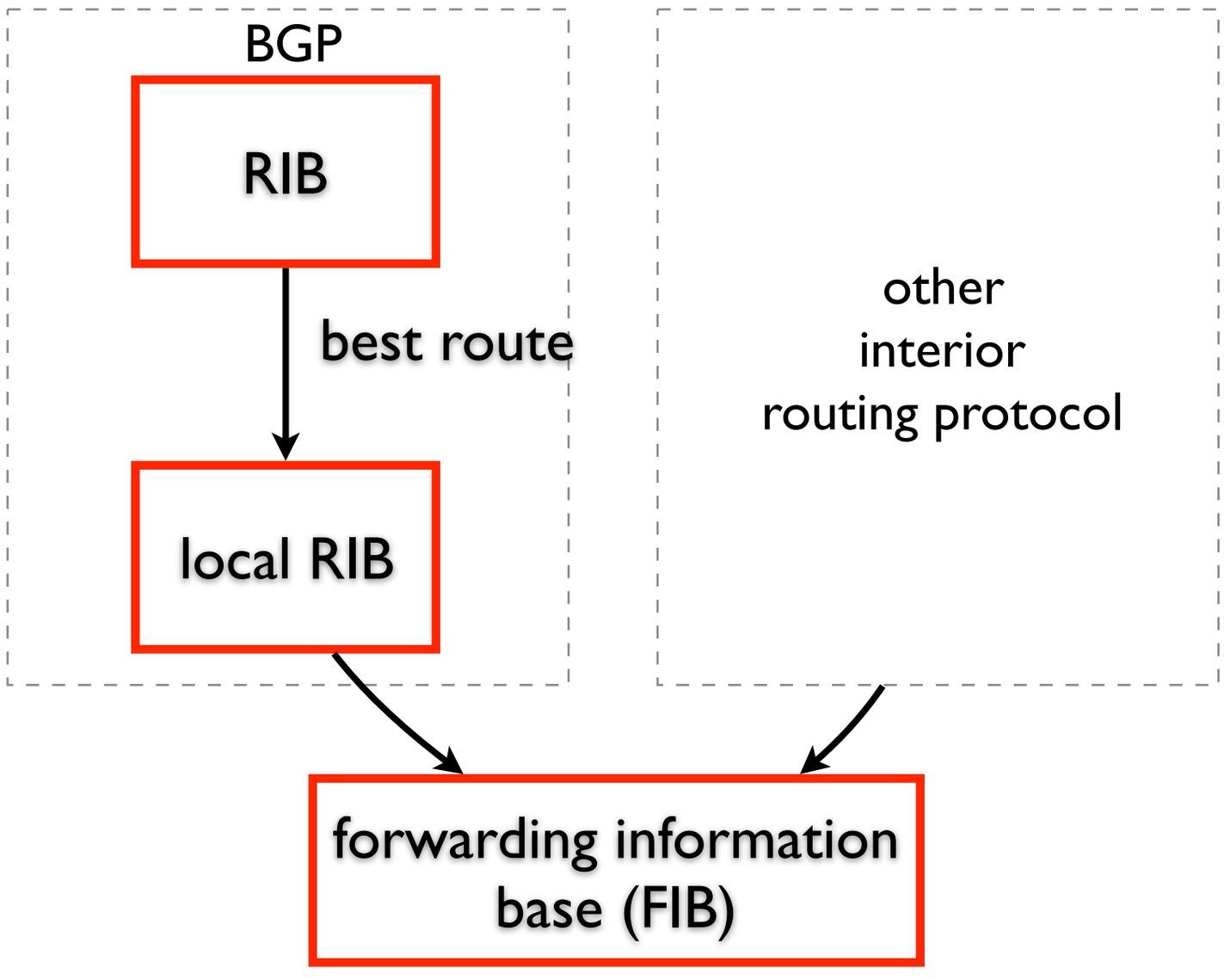
- describe the network-wide connectivity



# BGP routing table

## routing inforation base (RIB)

- describe the network-wide connectivity



# policy routing

an AS advertises a route to a neighboring AS

- the local AS offers to accept traffic from the neighbor
  - the local AS originates the route, or
  - the local AS willing to undertake the role of transit provider

# policy routing

## an AS advertises a route to a neighboring AS

- the local AS offers to accept traffic from the neighbor
  - the local AS originates the route, or
  - the local AS willing to undertake the role of transit provider

## an AS accepts a route

- the local AS will use the neighboring AS to reach addresses spanned by the route

# BGP table — connectivity of the Internet

BGP table maintains a coherent view of the connectivity of the inter-AS domain

- connectivity expressed as a preference for "shortest paths" to reach any destination address
- modulated by AS (connectivity) policies

# BGP table — connectivity of the Internet

BGP table maintains a coherent view of the connectivity of the inter-AS domain

- connectivity expressed as a preference for "shortest paths" to reach any destination address
- modulated by AS (connectivity) policies

coherence

- none of the paths — collection of BGP entries — contains loops or dead ends

# BGP table size

routing entries are the elements of the BGP routing domain

- each entry for a span of addresses
- each shares a common origin AS + policy

total size

- number of distinct routes within the Internet

each BGP route describes a contagious set of addresses that share a common origin AS and a common policy

# routing space

cross product of

- complexity of the inter-AS topology
- number of distinct AS policies
- degree of fragmentation of the address space

# BGP table and CIDR

## classless inter-domain routing (CIDR)

- introduces hierarchy into inter domain
- allows a provider to merge the routing entries for its customers

# BGP table and CIDR

## classless inter-domain routing (CIDR)

- introduces hierarchy into inter domain
- allows a provider to merge the routing entries for its customers

} hierarchical provider address aggregation

# BGP table and CIDR

## classless inter-domain routing (CIDR)

- introduces hierarchy into inter domain
- allows a provider to merge the routing entries for its customers

} hierarchical provider address aggregation

the provider announces its entire block (spanning its entire customer base) into the BGP table as a single entry with a single policy

# BGP table and CIDR

*until the start of 1999,* CIDR proved effective in damping unconstrained growth of the BGP table

- a greater level of stability
  - instability at the edge not immediately propagated into the routing core
  - instability at the last hop, absorbed by an aggregate route

## 1998 - ? towards a compound growth model

- 42% growth of the BGP table per year

# BGP table and CIDR

*until the start of 1999,* CIDR proved effective in damping unconstrained growth of the BGP table

- a greater level of stability
  - instability at the edge not immediately propagated into the routing core
  - instability at the last hop, absorbed by an aggregate route

1998 - ? towards a compound growth model

- 42% growth of the BGP table per year

causes?

- weakening of the hierarchical model in the Internet

# BGP table and CIDR

*until the start of 1999,* CIDR proved effective in damping unconstrained growth of the BGP table

- a greater level of stability
  - instability at the edge not immediately propagated into the routing core
  - instability at the last hop, absorbed by an aggregate route

1998 - ? towards a compound growth model

- 42% growth of the BGP table per year

causes?

- hierarchical addresses allocation and CIDR unable to keep pace with the levels of growth of the Internet

# the compound growth of BGP table

## contributing factors

- number of ASes
- number of distinct AS paths
- range of addresses spanned by the table
- average span of each routing entry

# the compound growth of BGP table

## who needs an AS number?

# the compound growth of BGP table

## who needs an AS number?

each network multi-homed and expresses a distinct policy needs a unique AS number to associate its advertised address with such policy

# the compound growth of BGP table

## who needs an AS number?

each network multi-homed and expresses a distinct policy needs a unique AS number to associate its advertised address with such policy

## number of ASes in the routing table

- tracks the number of entries that have unique policies

# the compound growth of BGP table

## who needs an AS number?

each network multi-homed and expresses a distinct policy needs a unique AS number to associate its advertised address with such policy

## number of ASes in the routing table

- tracks the number of entries that have unique policies

## trend

- deployment of AS number (16-bit to 32-bit) grows exponentially

# the compound growth of BGP table

## address space within the BGP table

- 2001, around 25% of the total IPv4 — 25% of the usable unicast public address

## trend

- the growth in the amount of addresses advertised is far lower, compared to the growth in the number of routing advertisements

# the compound growth of BGP table

## address space within the BGP table

- 2001, around 25% of the total IPv4 — 25% of the usable unicast public address

## trend

- the growth in the amount of addresses advertised is far lower, compared to the growth in the number of routing advertisements

## causes

- NAT: smaller address fragment supporting distinct policies, encompassing large networks located behind NATs
- discrete policies applied to finer addresses blocks

# the compound growth of BGP table

## granularity of table entries

per routing entry

- prefix length
- average span of individual addresses

# the compound growth of BGP table

## granularity of table entries

| | | | |
|---|---|---|---|
| per routing entry | prefix length | /18.3 | |
| | average span of individual addresses | 16,000 | Nov, 1999 |

# the compound growth of BGP table

## granularity of table entries

| | | | |
|---|---|---|---|
| per routing entry | prefix length | /18.3 | /18.44 |
| | average span of individual addresses | 16,000 | 12,000 |
| | | Nov, 1999 | Dec, 2000 |

# the compound growth of BGP table

## granularity of table entries

| | | | | |
|---|---|---|---|---|
| per routing entry | prefix length | /18.3 | /18.44 | /18.6 |
| | average span of individual addresses | 16,000 | 12,000 | 10,700 |
| | | Nov, 1999 | Dec, 2000 | Sept, 2001 |

# the compound growth of BGP table

## granularity of table entries

| | | | | |
|---|---|---|---|---|
| per routing entry | prefix length | /18.3 | /18.44 | /18.6 |
| | average span of individual addresses | 16,000 | 12,000 | 10,700 |
| | | Nov, 1999 | Dec, 2000 | Sept, 2001 |

## trend

- towards finer grained entries

# the compound growth of BGP table

## granularity of table entries

| | | | | |
|---|---|---|---|---|
| per routing entry | prefix length | /18.3 | /18.44 | /18.6 |
| | average span of individual addresses | 16,000 | 12,000 | 10,700 |
| | | Nov, 1999 | Dec, 2000 | Sept, 2001 |

## trend

- towards finer grained entries

ca

smaller networks multi-homed without hierarchical structure

- increasingly dense interconnectivity
- networks with a single-homed connection and hierarchical routing → multi-homed without hierarchical structure

# aggregation and holes with CIDR

advertise a more specific prefix of an existing aggregate

- "punch" a hole in the policy of the larger aggregate announcement
  - creating a different policy for the specifically referenced address prefix

# scalable inter-domain routing

compound growth trend with the BGP table

+

finer granularity of routing entries

# scalable inter-domain routing

compound growth trend with the BGP table

+

finer granularity of routing entries

can the BGP system scale adequately to continue to undertake the role of the inter-domain routing system

# scalable BGP? — CIDR

## trend

- denser interconnectivity mesh, but CIDR deployment assumes a single-homed network with a strict hierarchy of supply providers

## casualty

- CIDR-induced dampened growth of the BGP table

# scalable BGP? — CIDR

trend

- smaller networks, advertised as a /24 prefix entry, multi-homed with a number of peers and upstream provider
- accepted as a substitute for upstream service resiliency

# scalable BGP? — resiliency

trend

- smaller networks, advertised as a /24 prefix entry, multi-homed with a number of peers and upstream provider
- accepted as a substitute for upstream service resiliency

# (problem with) service resiliency

responsibility of the provider → responsibility of the customer

# scalable BGP? — resiliency

trend

- smaller networks, advertised as a /24 prefix entry, multi-homed with a number of peers and upstream provider
- accepted as a substitute for upstream service resiliency

# (problem with) service resiliency

responsibility of the provider → responsibility of the customer

function of the bearer (switching subsystem) → function of the BGP routing system

# scalable BGP? — TE

## trend

- traffic engineering (TE) via selective advertisements of smaller prefixes along different paths within a multi-homed environment

## problem

- additional fine-grained prefixes into the routing table

# scalable BGP? — lack of uniformity

## trend

- lack of common practice among the advertisers and recipients

## problem

- advertisement appear to be propagated well beyond their intended domain of applicability
- withdraw/advertisement not adequately damped close to the origin of the route flap

# scalable BGP?

- denser interconnectivity mesh
- multi-homing with smaller addresses
- traffic engineering
- lack of common practices

# scalable BGP?

- denser interconnectivity mesh
- multi-homing with smaller addresses
- traffic engineering
- lack of common practices

*compound* (rather than liner) *growth* in the total size of BGP table

# scalable BGP?

- denser interconnectivity mesh
- multi-homing with smaller addresses
- traffic engineering
- lack of common practices

*compound* (rather than liner) *growth* in the total size of BGP table → millions of small entries (rather than a hierarchical routing space of 10 of thousands larger addresses

# some requirements

## a scalable inter-domain routing system

- reachability entries
- policy entries
- dynamic change
- time to converge

# some requirements — stability

routing change propagated only as far as necessary to reach a new stable state

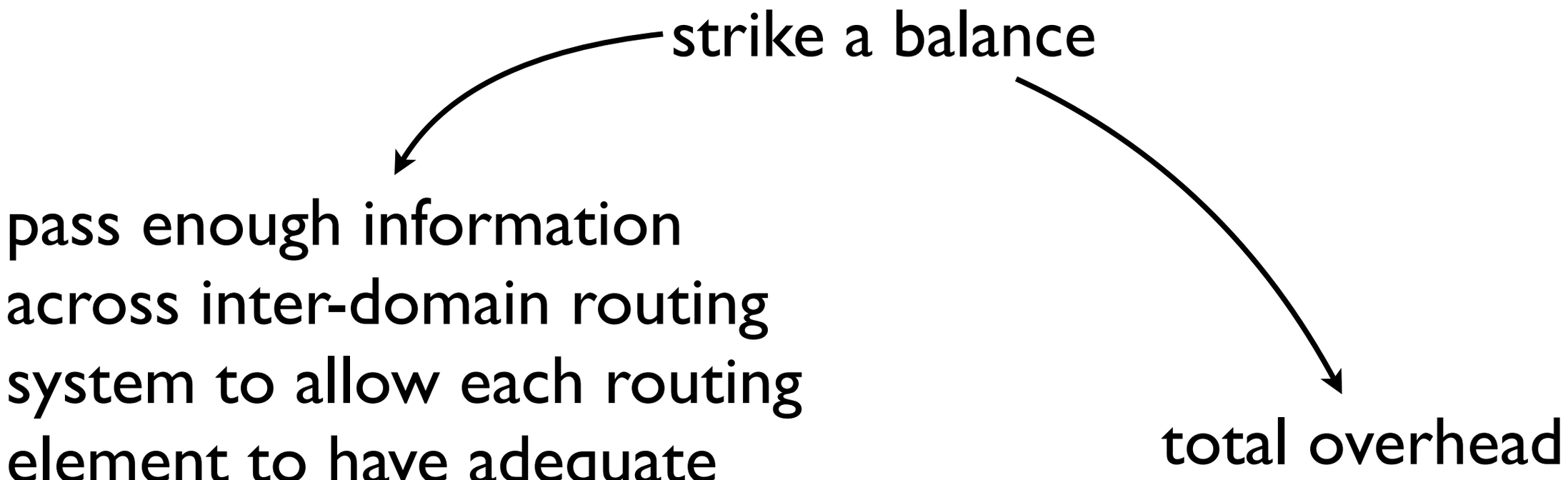
locality

# some requirements — convergence

upper limit reflects the requirement of the routing system

- to support a broad range of application classes, must be of the order of seconds

# some requirements — overhead



# recap

the longer term trends of the BGP table

- understand the visible characteristics
- identify the contributing sources

impacts on the ability of the Internet to scale