Customized BGP Route Selection

Laurent Vanbever, *Cristel Pelsser*
UCLouvain, *Internet Initiative Japan*

laurent.vanbever@uclouvain.be, *cristel@iij.ad.jp*

Pierre François (UCLouvain, BE), Olivier Bonaventure (UCLouvain, BE)
and Jennifer Rexford (Princeton, USA)

WIDE Camp
Tuesday, March 9 2010

Customized BGP Route Selection

Introduction and motivation

Implementing CRS

Practical considerations and solutions

Conclusion

Customized BGP Route Selection

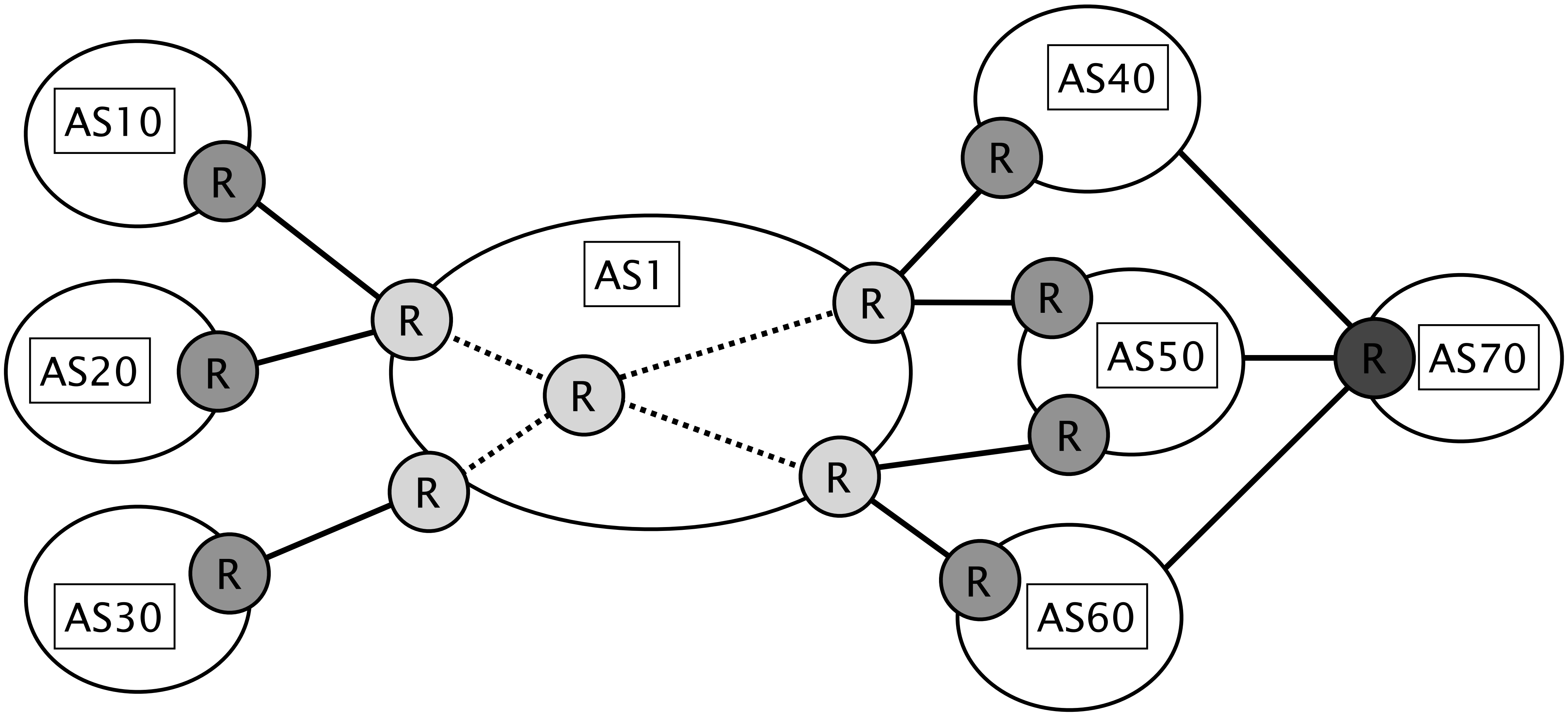
Introduction and motivation

Implementing CRS

Practical considerations and solutions

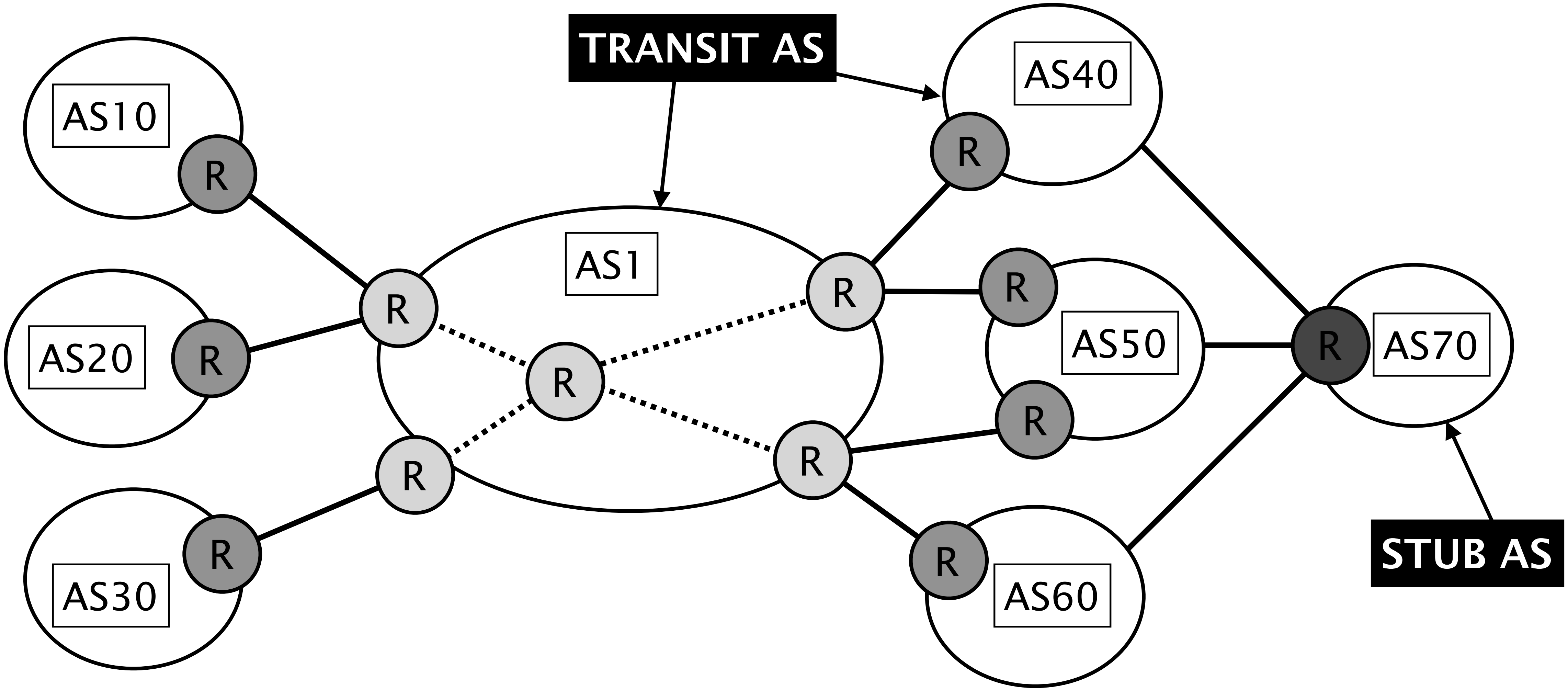
Conclusion

The Internet is a collection of Autonomous Systems (AS)



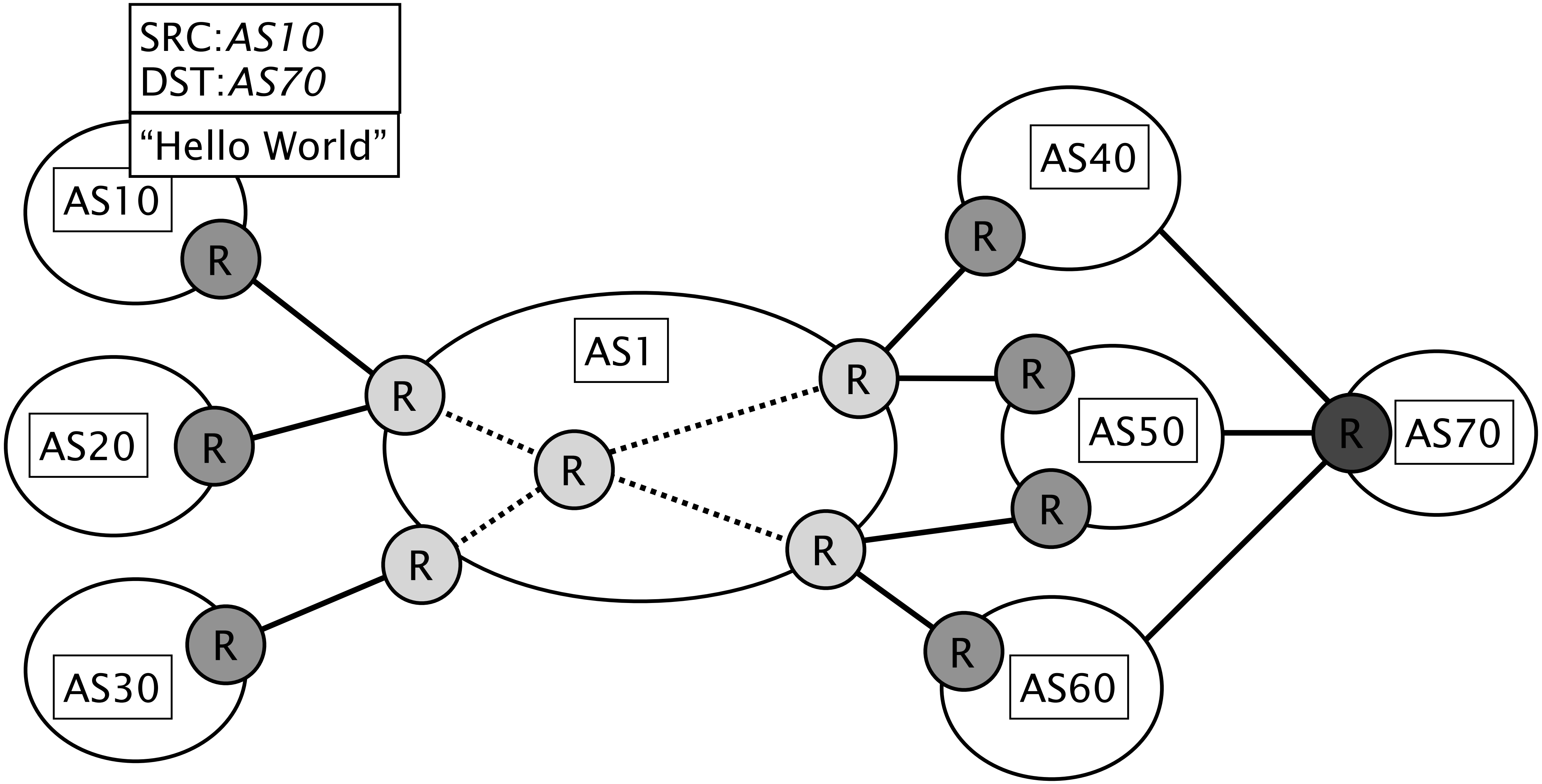
- An AS is a set of routers managed by a single administrative entity
  - Today, there are approximately 30.000 ASes

The Internet is a collection of Autonomous Systems (AS)



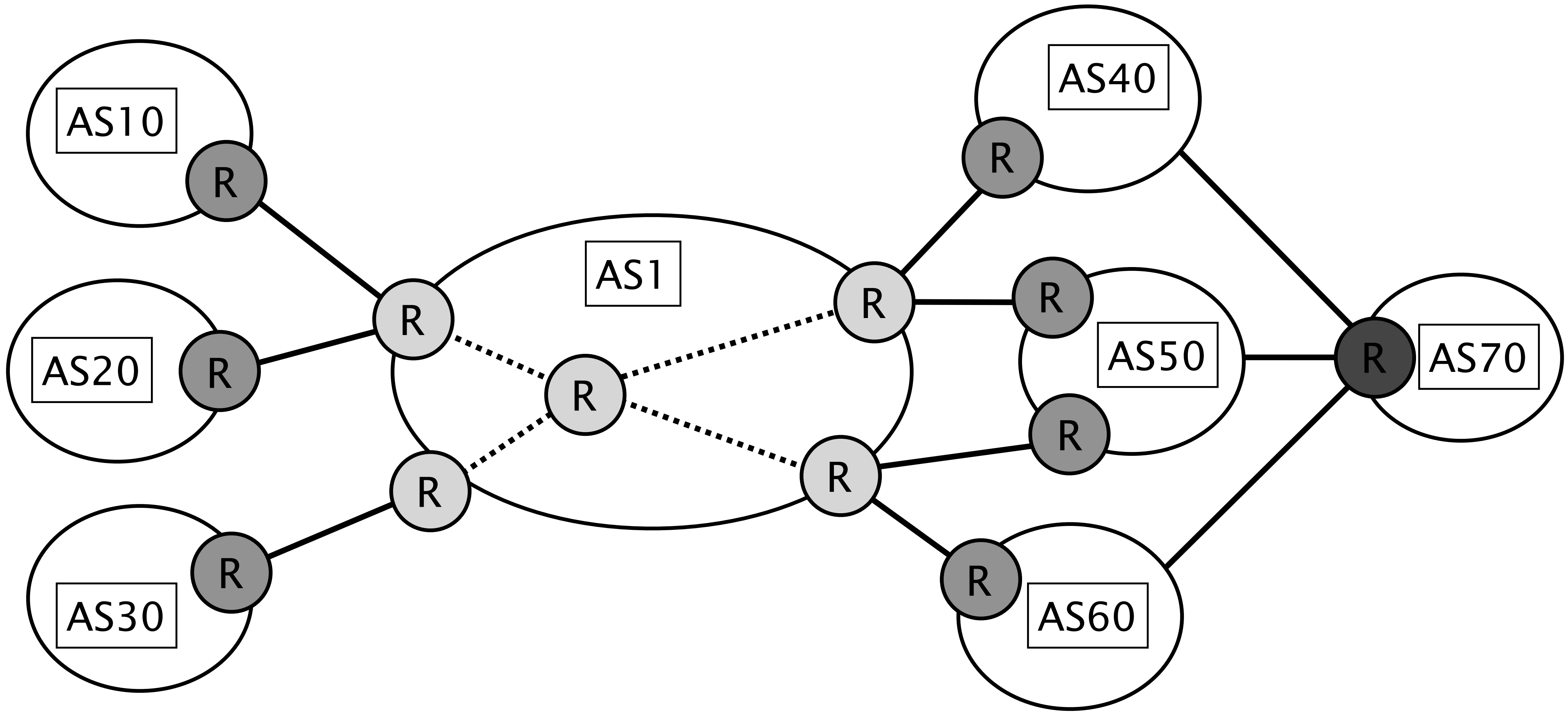
- An AS is a set of routers managed by a single administrative entity
  - Today, there are approximately 30.000 ASes

The Internet is a collection of Autonomous Systems (AS)



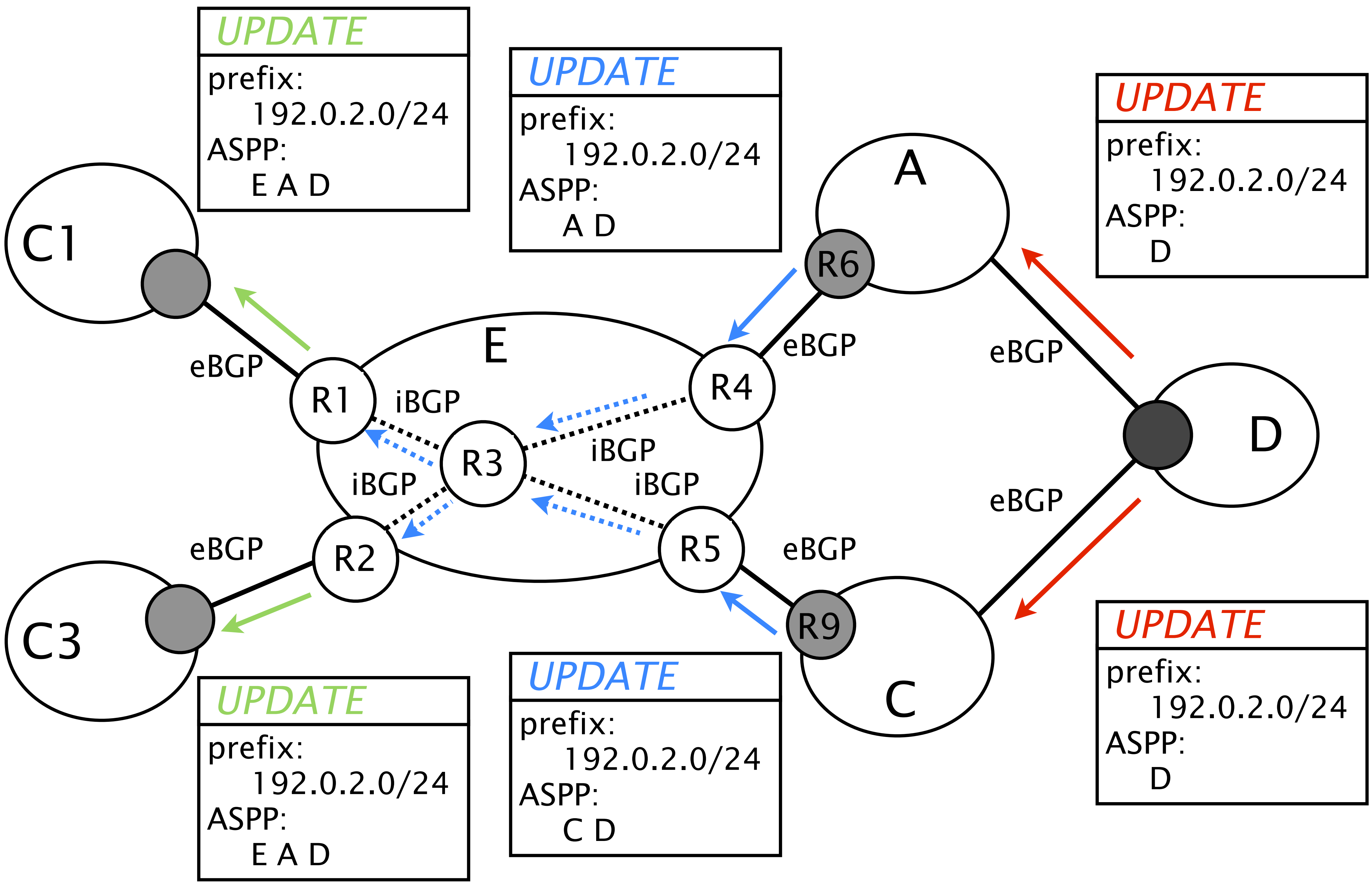
- An AS is a set of routers managed by a single administrative entity
  - Today, there are approximately 30.000 ASes

The Internet is a collection of Autonomous Systems (AS)

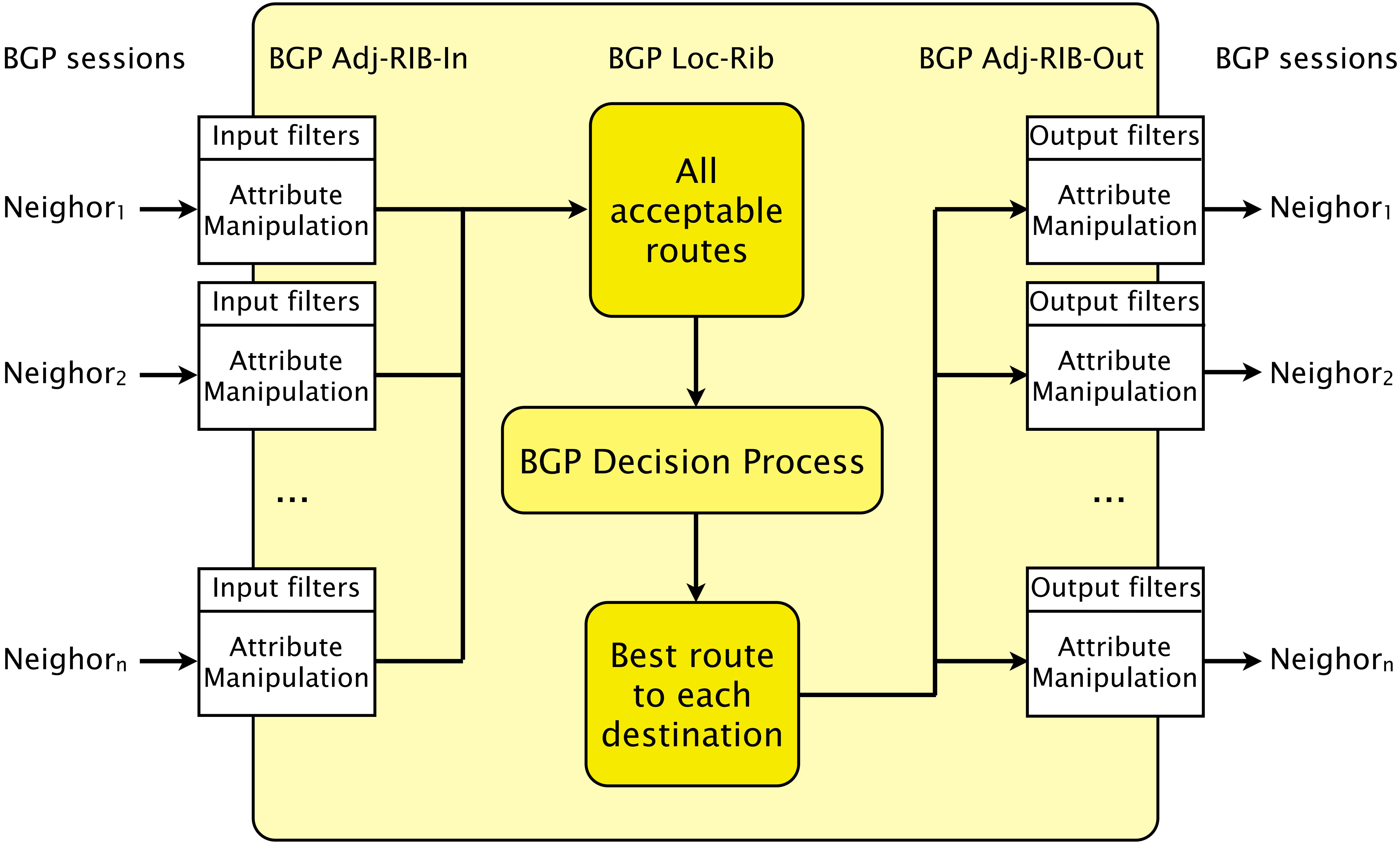


- An AS is a set of routers managed by a single administrative entity
  - Today, there are approximately 30.000 ASes

BGP is the *path-vector*, *policy-based* interdomain routing protocol

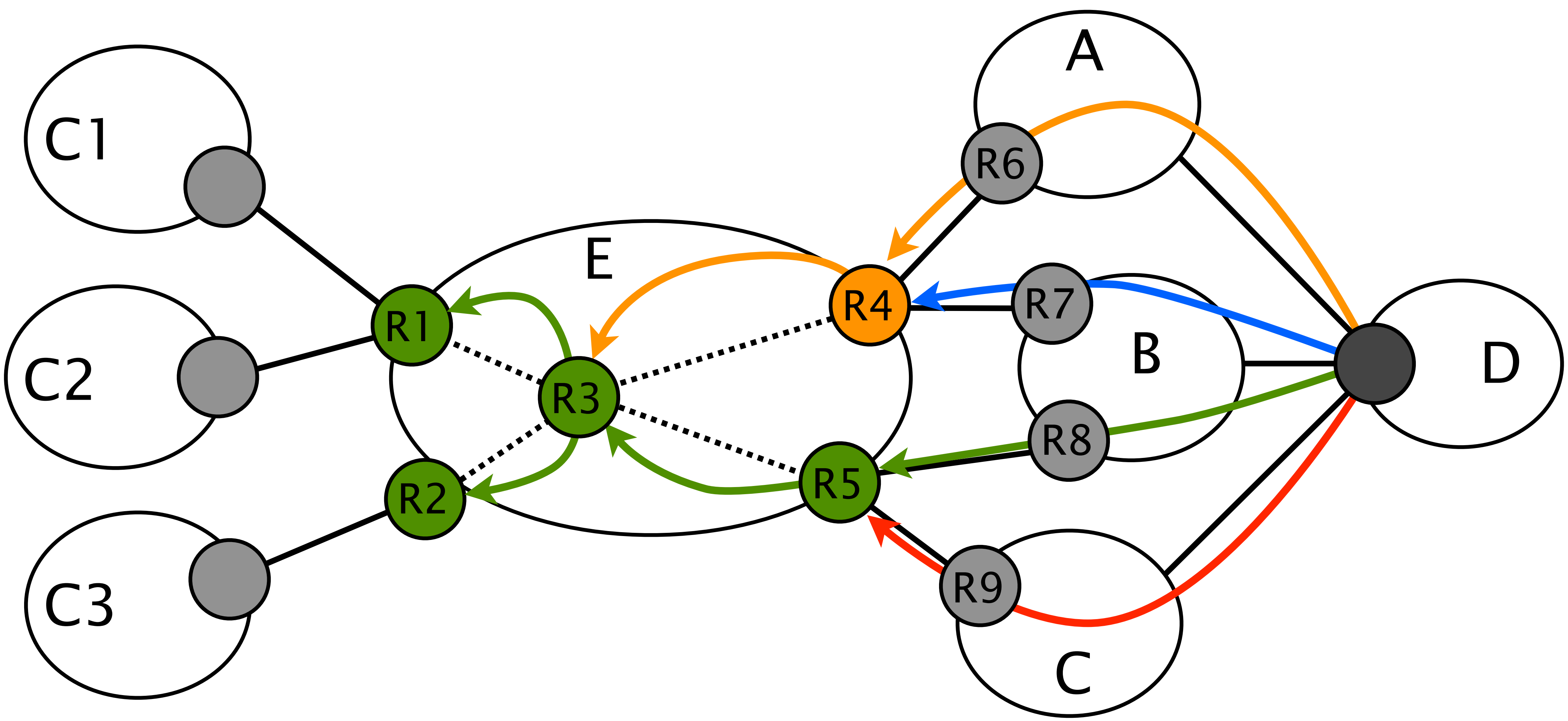


BGP is based on *sessions*, *policies* and a *decision process*



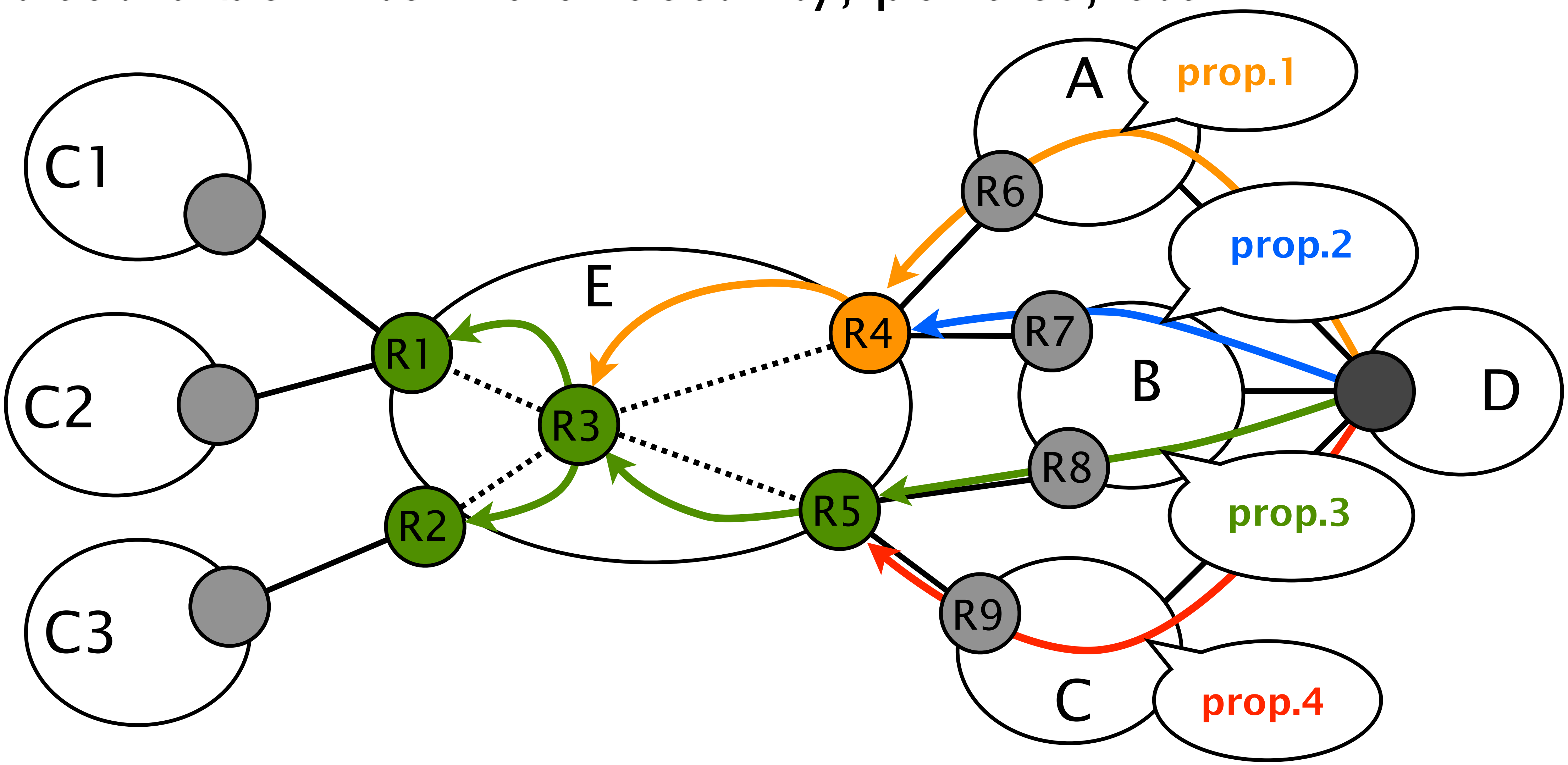
BGP Route Selection: *One-route-fits-all* model

- A BGP router selects **one** best route for each destination
- Globally, AS E knows 4 paths towards D
  - Locally, some routers only know one path (C1...C3, R1, R2)



BGP Route Selection: *One-route-fits-all* model

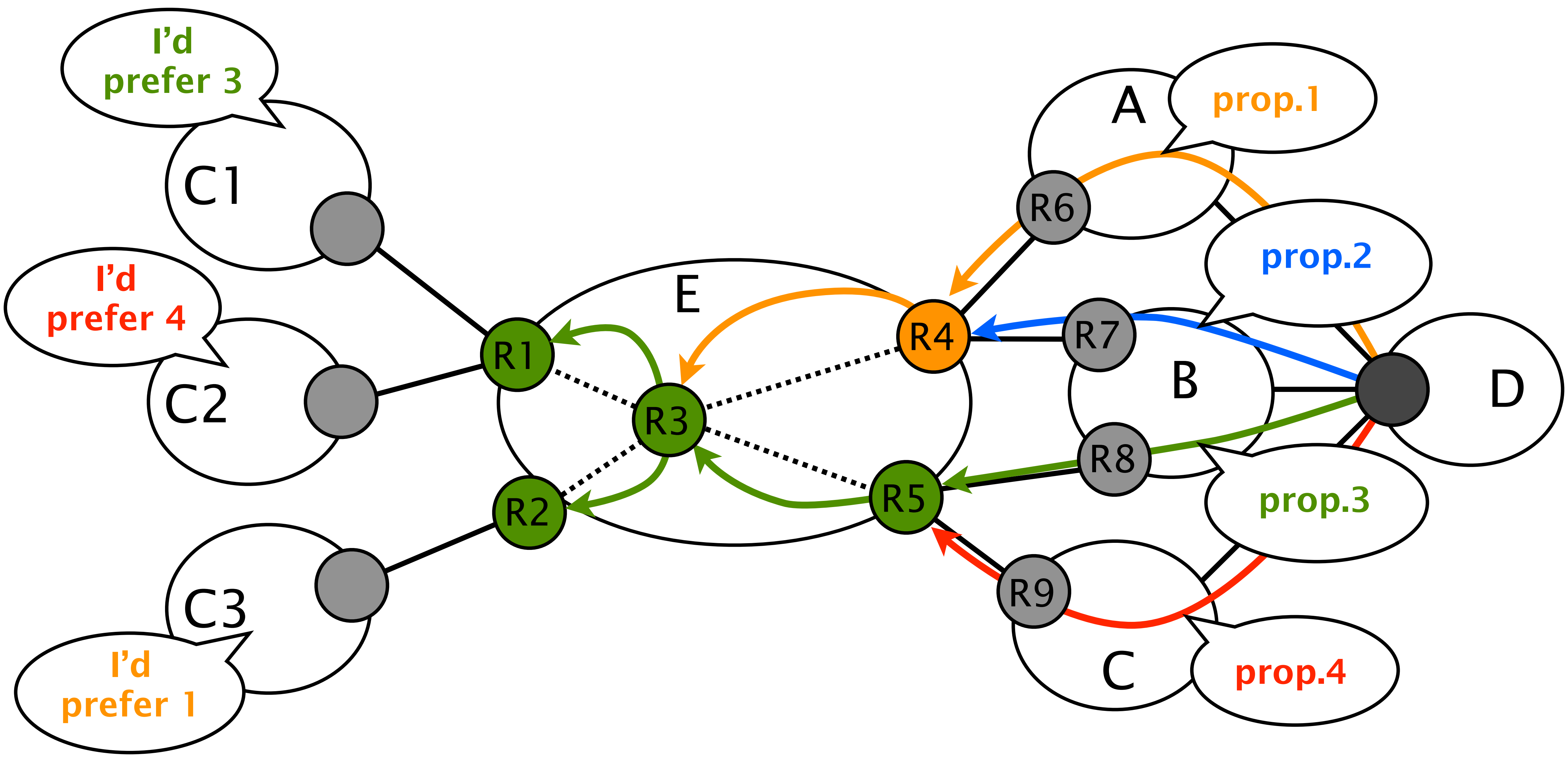
- Many ISPs have a rich path diversity
  - It is common to have 5-10 paths *per prefix*[1]
- Different paths have different properties
  - It could be in terms of security, policies, etc.



[1] W. Muhlbauer, A. Feldmann, O. Maennel, M. Roughan, and S. Uhlig. Building an AS-topology model that captures route diversity. *In Proc. ACM SIGCOMM*, 2006.

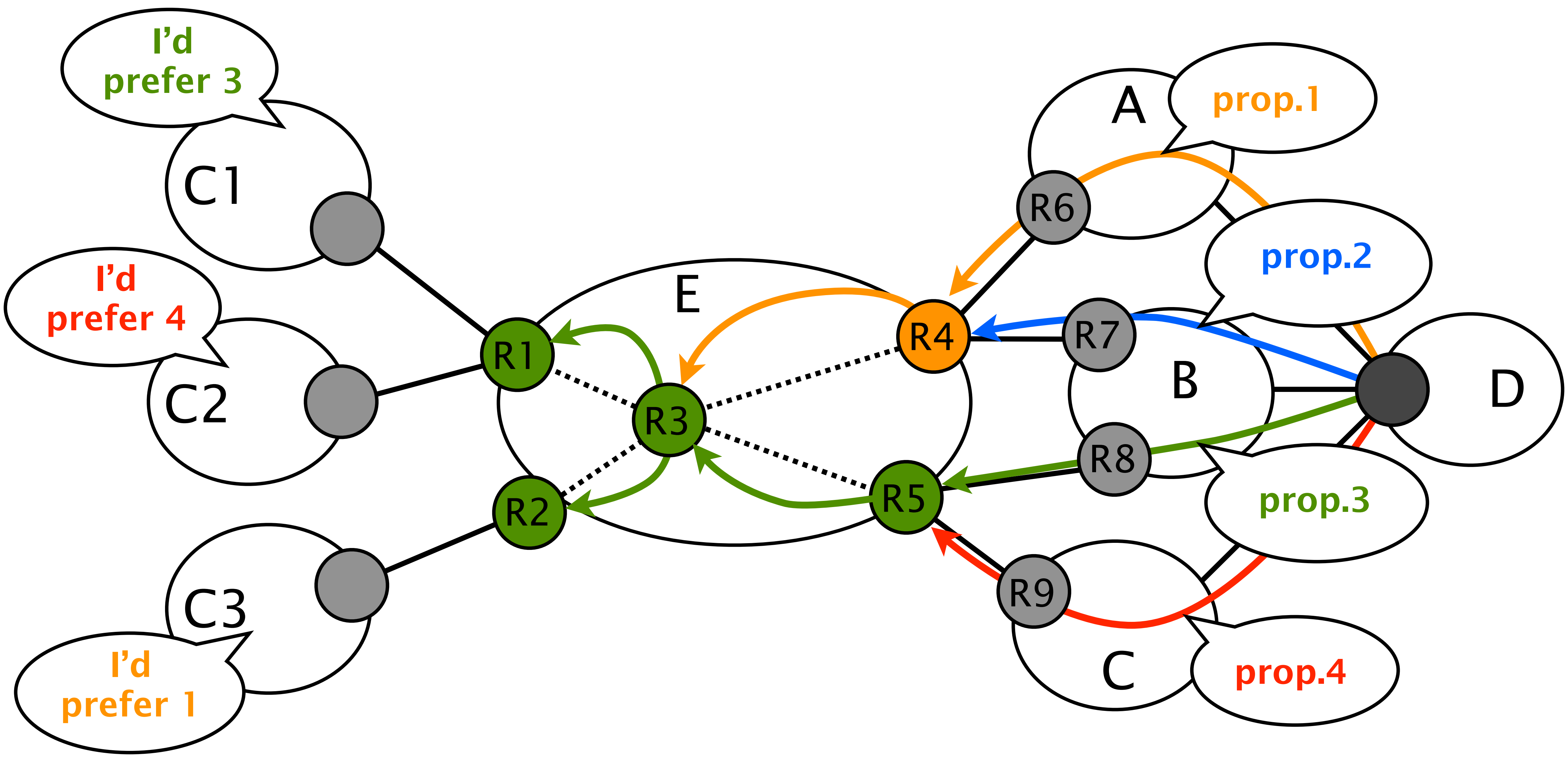
BGP Route Selection: *One-route-fits-all* model

- Clients may want different paths to the same prefix
  - If C1 is a competitor of C, he'd prefer to reach D via A or B
  - C1 may even want to pay an extra fee for that



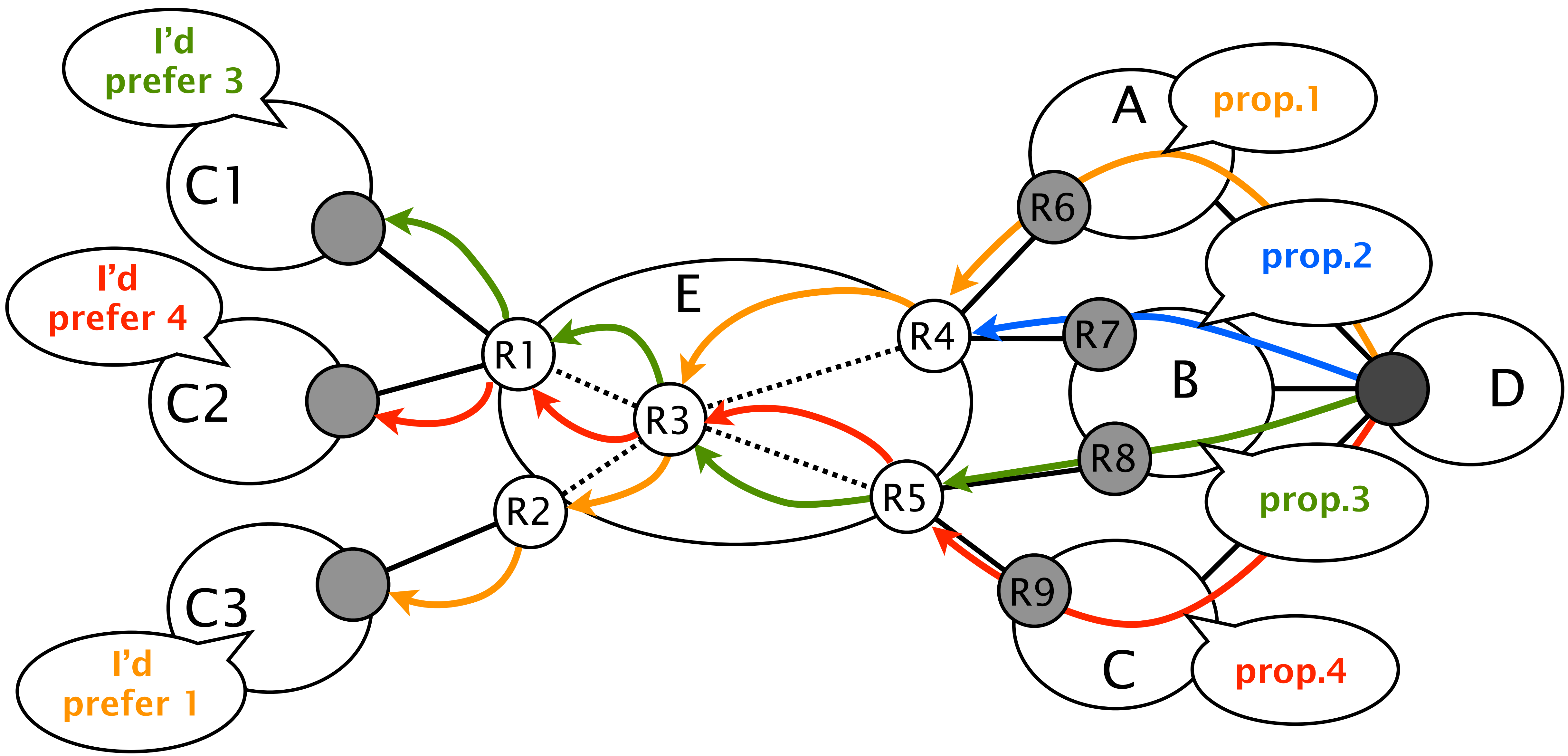
BGP Route Selection: *One-route-fits-all* model

- With vanilla BGP, you *can't* match customers' preferences to available paths
  - Customers of a given PE receive the same path



CRS: *Customized Route Selection*

- Under CRS, one router can offer *different* interdomain routes to *different* neighbors
  - C1 reaches D via B, C2 reaches D via C



Customized BGP Route Selection

Introduction and motivation

Implementing CRS

Potential issues and solutions

Conclusion

Under CRS, routes are *colorized* based on their properties

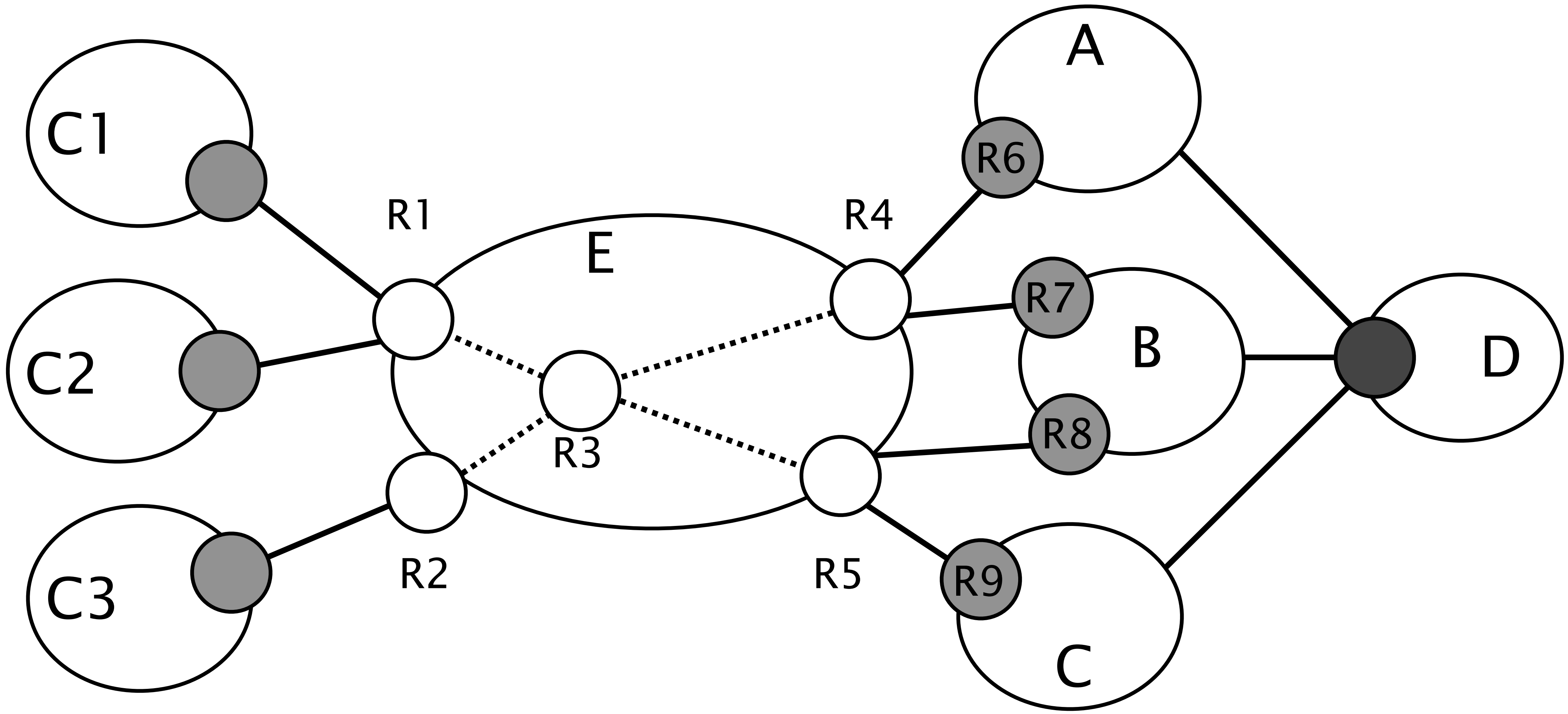
- A *color* denotes a set of routes sharing a property
  - *e.g.*, color *red* is associated to all *high-bandwidth* routes learned on *national* peerings
  - one route can have multiple colors

- Colors are “tags” associated to routes
  - we use the well-known BGP community field

What do we need to implement CRS with BGP MPLS VPNs ?

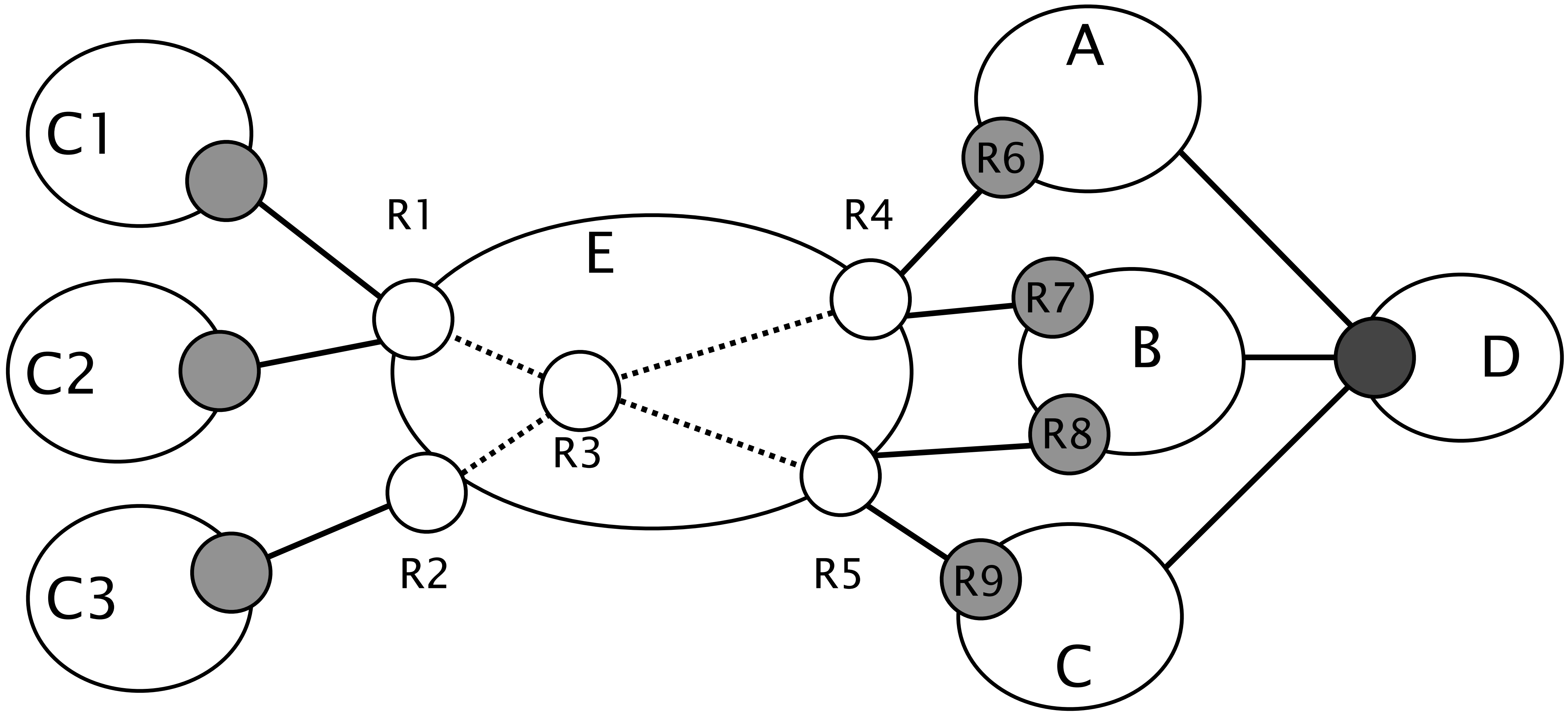
- Mechanisms to *disseminate* and *differentiate* paths
  - Multiprotocol BGP is used as dissemination protocol
  - Route Targets (RT) are used to identify colors
  - Route Distinguishers (RD) are used to ensure diversity
- *Customized* route selection mechanisms at ASBR
  - Use of Virtual Routing and Forwarding (VRF) instances
- Traffic forwarding on the chosen paths
  - MPLS tunneling

How do we implement CRS with BGP MPLS VPNs ?

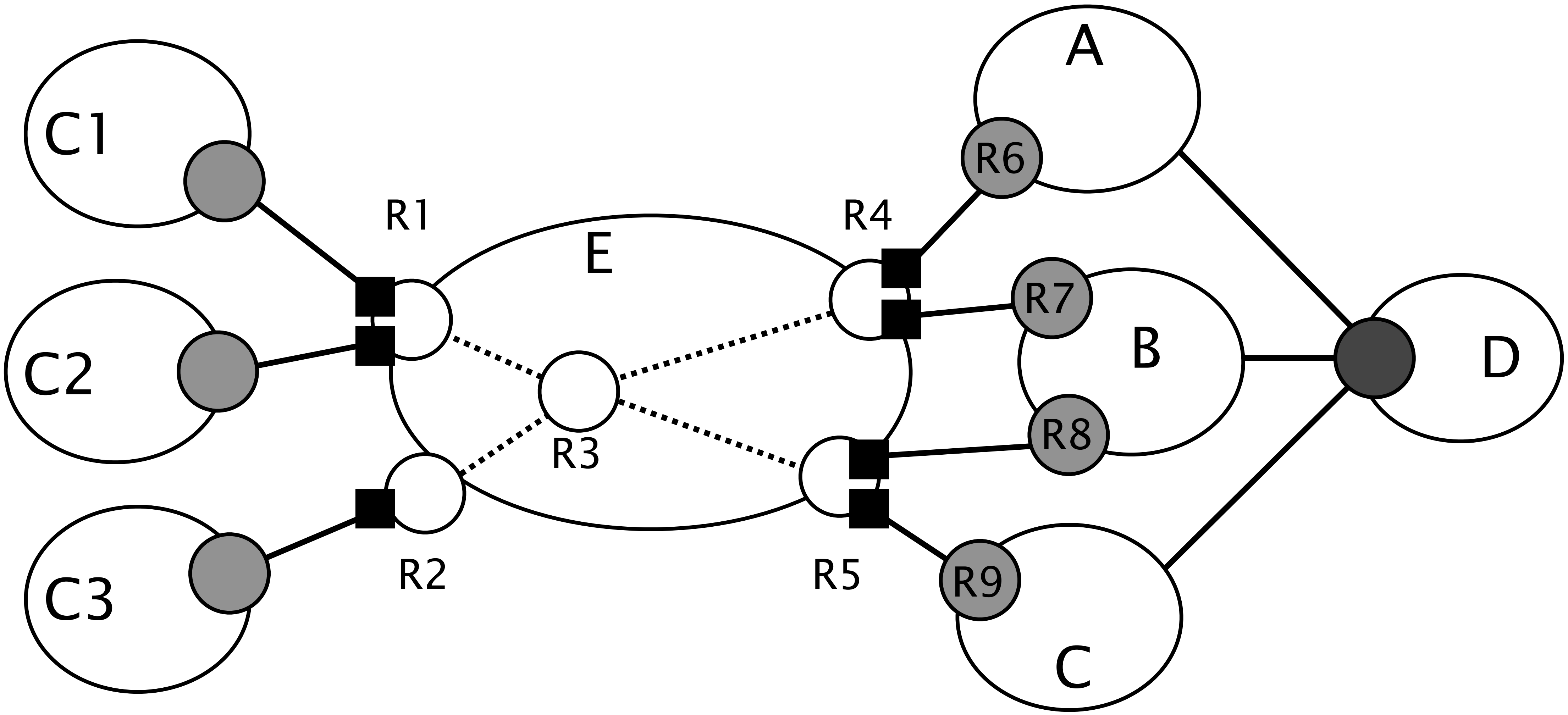


- C1 wants to reach D via B, C2 via C
- Define 3 colors: routes learned via *A (green)*, *B (red)* and *C (blue)*
- Announce *red* routes to C1, *blue* routes to C2

How do we implement CRS with BGP MPLS VPNs ?



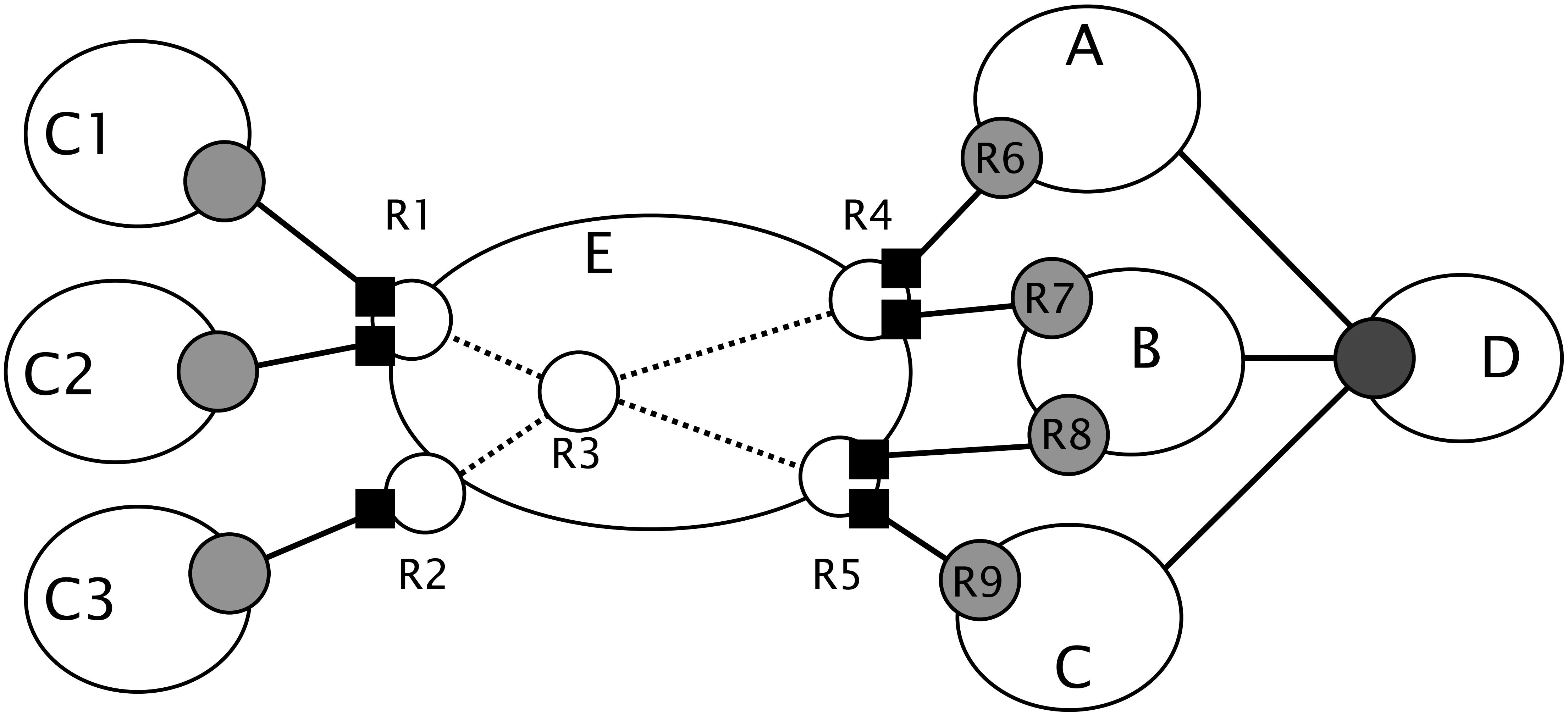
How do we implement CRS with BGP MPLS VPNs ?



- Consider peers as VPNs and put them in VRFs

How do we implement CRS with BGP MPLS VPNs ?

Route Targets	
green	*learned via A*
red	*learned via B*
blue	*learned via C*

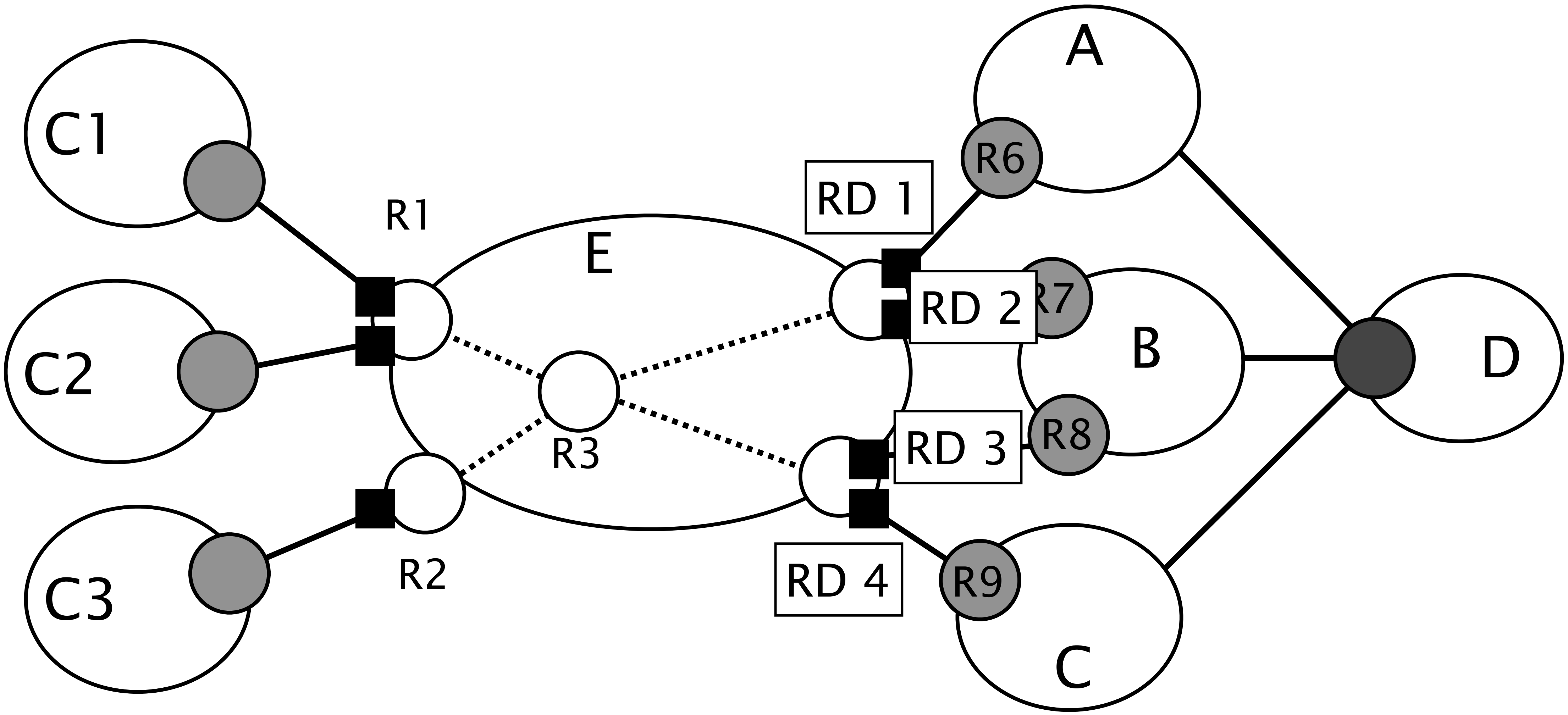


- Consider peers as VPNs and put them in VRFs
- Use RT to identify *colors*

How do we implement CRS with BGP MPLS VPNs ?

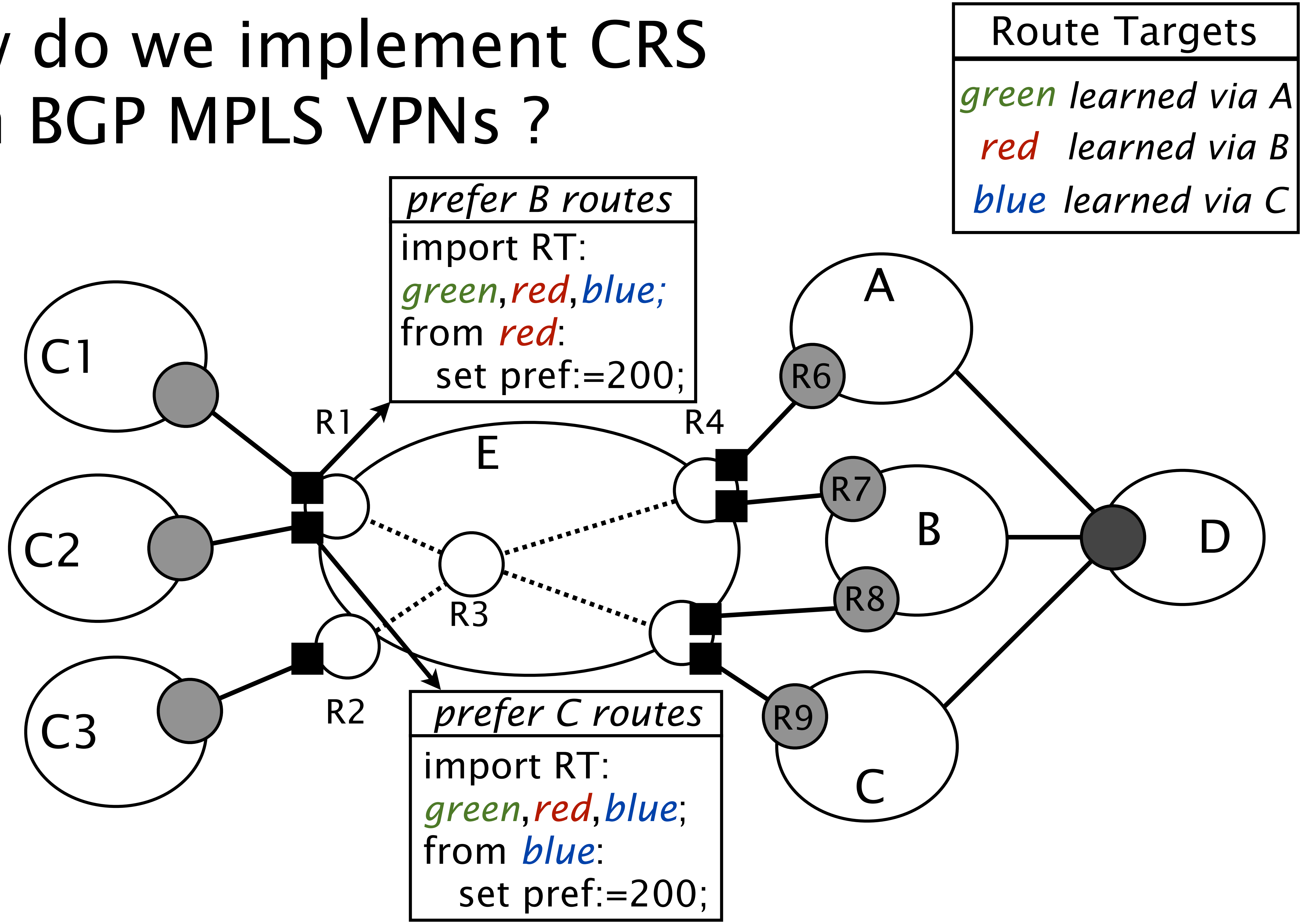
Route Targets

green *learned via A*
red *learned via B*
blue *learned via C*



- Consider peers as VPNs and put them in VRFs
- Use RT to identify *colors*
- Use different RD to differentiate routes

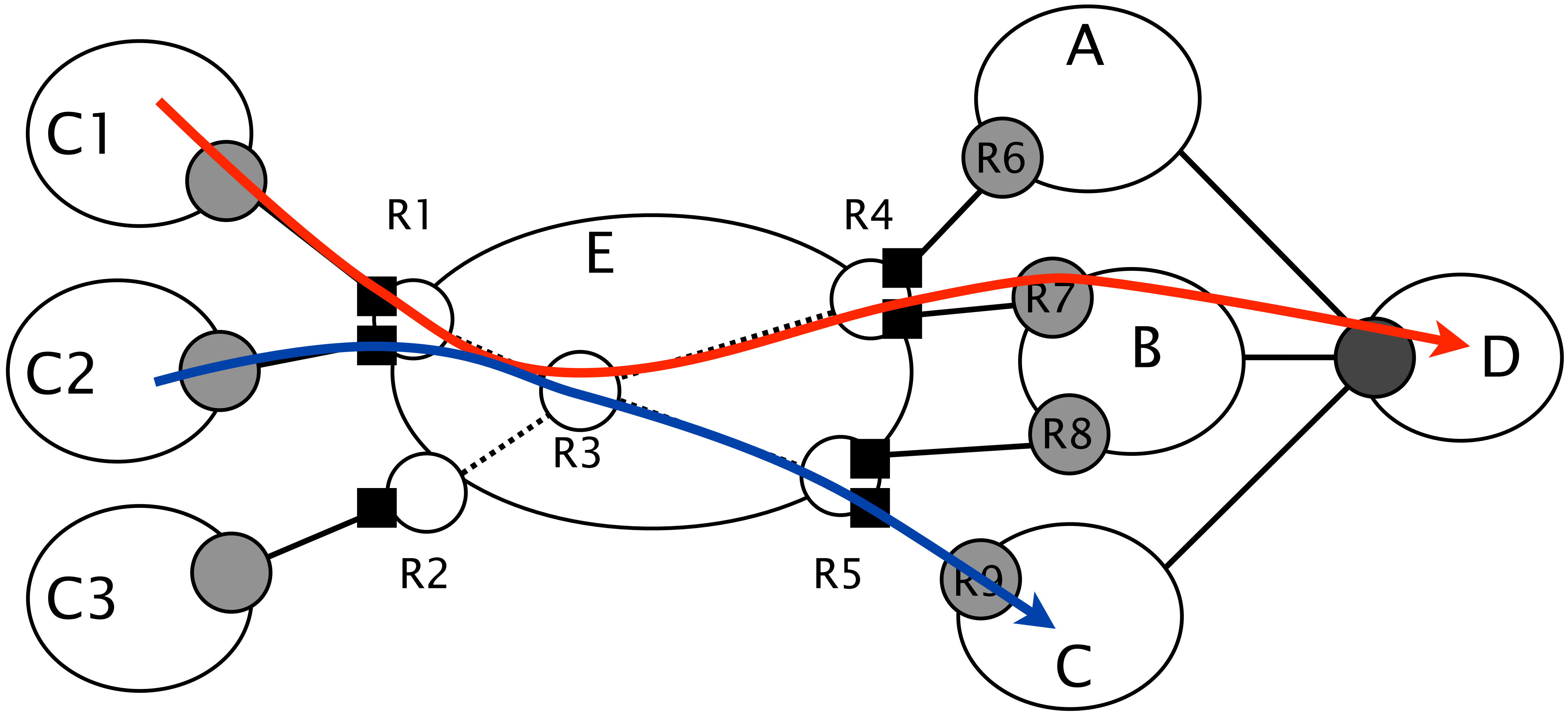
How do we implement CRS with BGP MPLS VPNs ?



- In each VRF, prefer certain routes via import filters

How do we implement CRS with BGP MPLS VPNs ?

Route Targets
green learned via A
red learned via B
blue learned via C



- MPLS is used for forwarding
  - Two levels label stack
  - R3 only knows label to reach the PEs

Customized BGP Route Selection Using BGP/MPLS VPNs

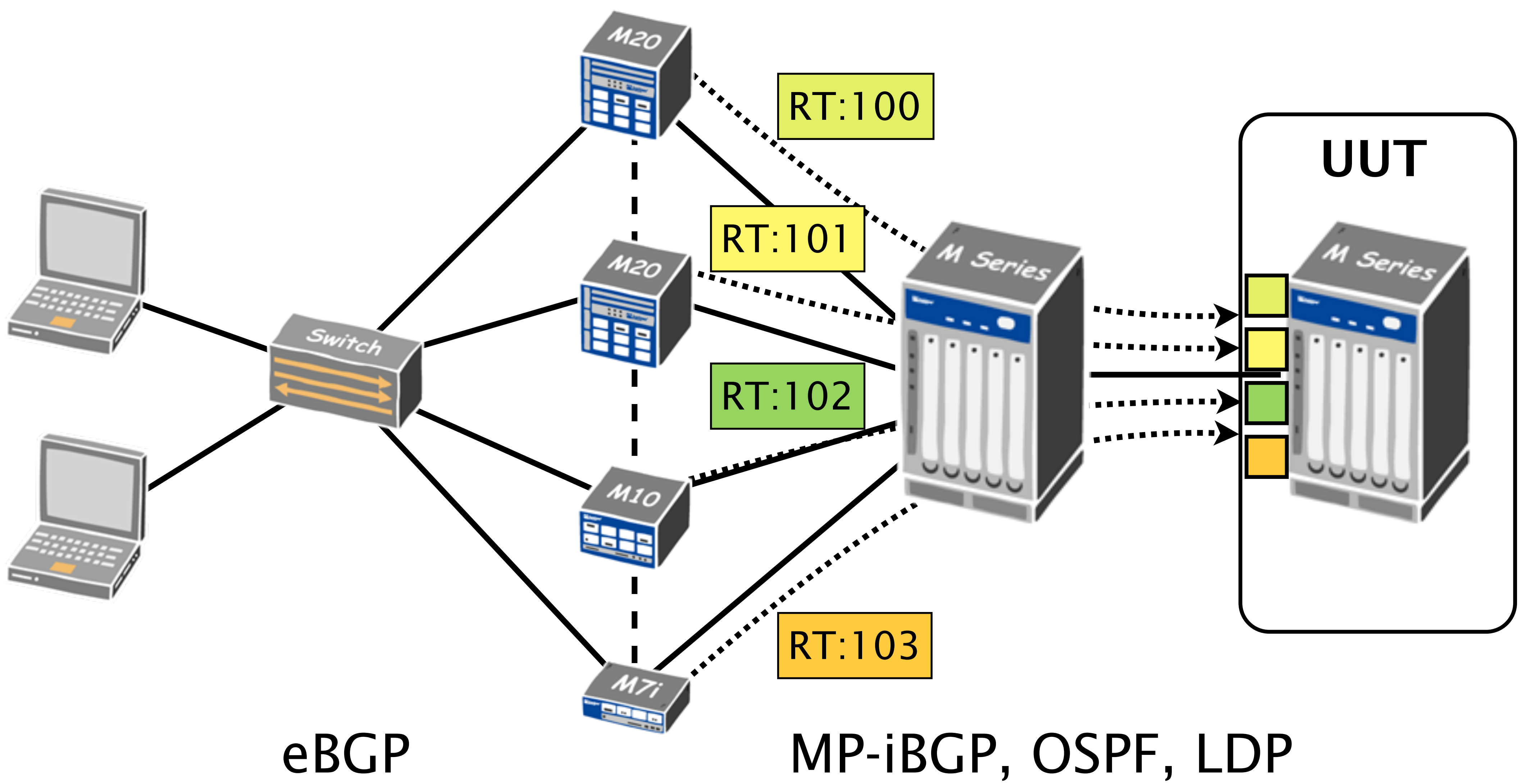
Introduction and motivation

Implementing CRS

Practical considerations and solutions

Conclusions

Is CRS pushing a M120 to the limit ?



Four tables are defined on the Unit Under Test (UUT)

- Each table is fed with one color (one RT)
- In each color, ~300k routes (1 path per route)
- In the end, 1.200.000 routes in **RIB** & **FIB**

Is CRS pushing a M120 to the limit ?

- UUT was a Juniper M120 [JunOS 9.3R2.8]
  - Routing Engine (RE) has 4 GB DRAM
  - Forwarding Engine Boards (FEB) have 512 MB DRAM

	RE	FEB
empty	*17%*	*9%*
fully-loaded (1.200.000 routes)	*38%*	*39%*

- FIB could handle more than 2.000.000 routes
  - Enough to support a few services *without* modifications

More services ?
scalability and...*scalability*

- Routes *dissemination* overhead
  - **All** PEs receive **all** VPN routes

- Routes *storage* overhead
  - RIB
    - Modest performance demand
    - Add more DRAM to support CRS ?
  - **FIB**
    - CRS's biggest challenge
    - Sharing between the VRFs in the FIB ?

How could we improve CRS FIB's scaling: *Selective VRF Download*

- By default, *all* VRFs are installed on *all* line cards

Slot	State	Temp (C)	CPU Utilization (%) Total	CPU Utilization (%) Interrupt	Memory DRAM (MB)	Utilization (%) Heap	Utilization (%) Buffer
2	Online	24	1	0	512	**39**	59
3	Online	28	1	0	512	**39**	59

- Customers ask for the same colors ?
  - Connect them on the same line card
  - Download VRFs only to line cards that need them
- It could be a management nightmare...

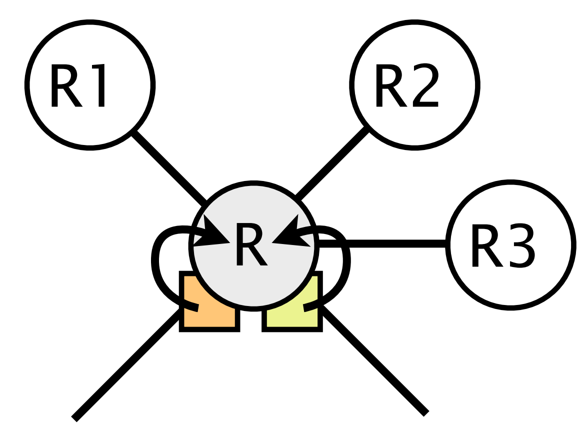
How could we improve CRS FIB's scaling: *Cross-VRF Lookup*

- Specific routing for a small set of prefixes ?
  - Create one small VRF *per color*
  - Add default entry towards a default VRF
- The price to pay is 2 IP lookups

VRF1
*>10/8 via *R1*
0/0 via *default*

VRF2
*>10/8 via *R2*
0/0 via *default*

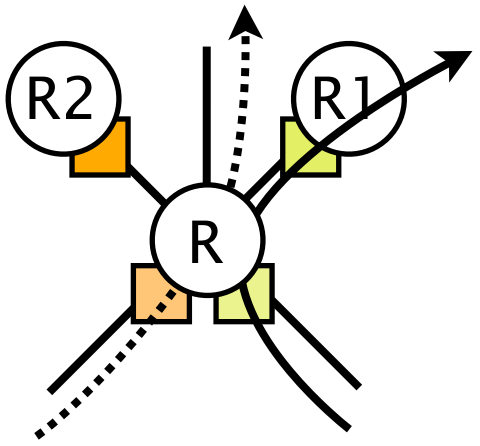
Default
…
*>10/8 via *R3*
…



How could we improve CRS FIB's scaling: *Distributed VRF*

- Distribute VRFs among routers which can afford extra load
  - PEs do not maintain complete VRFs anymore
  - PEs default route traffic towards these routers
- Increase in latency and load
- Distributed version of *Cross-VRF Lookup*

R maintain small VRFs
and default rest to R1 or R2
→ detour path
⋯→ direct path



Customized BGP Route Selection

Introduction and motivation

Implementing CRS

Practical considerations and solutions

Conclusion

CRS is feasible

- *Implementable*
  - It can be realized on today's routers
  - It uses well known BGP MPLS/VPNs techniques
- *Scalable (for a few services*)
  - "Modest" message and storage overhead
  - Lab experiments tend to confirm that
  - Full BGP tables are needed to complete our evaluation
- *Guaranteed interdomain convergence*
  - Extra flexibility does not compromise global routing stability[1]

[1] Proof in SIGMETRICS'09 paper by Y. Wang, M. Schapira, and J. Rexford

Customized BGP Route Selection

Questions ?

Please, come and see our poster !