lecture 05: centralized control —opportunities and challenges

5590: software defined networking

anduo wang, Temple University

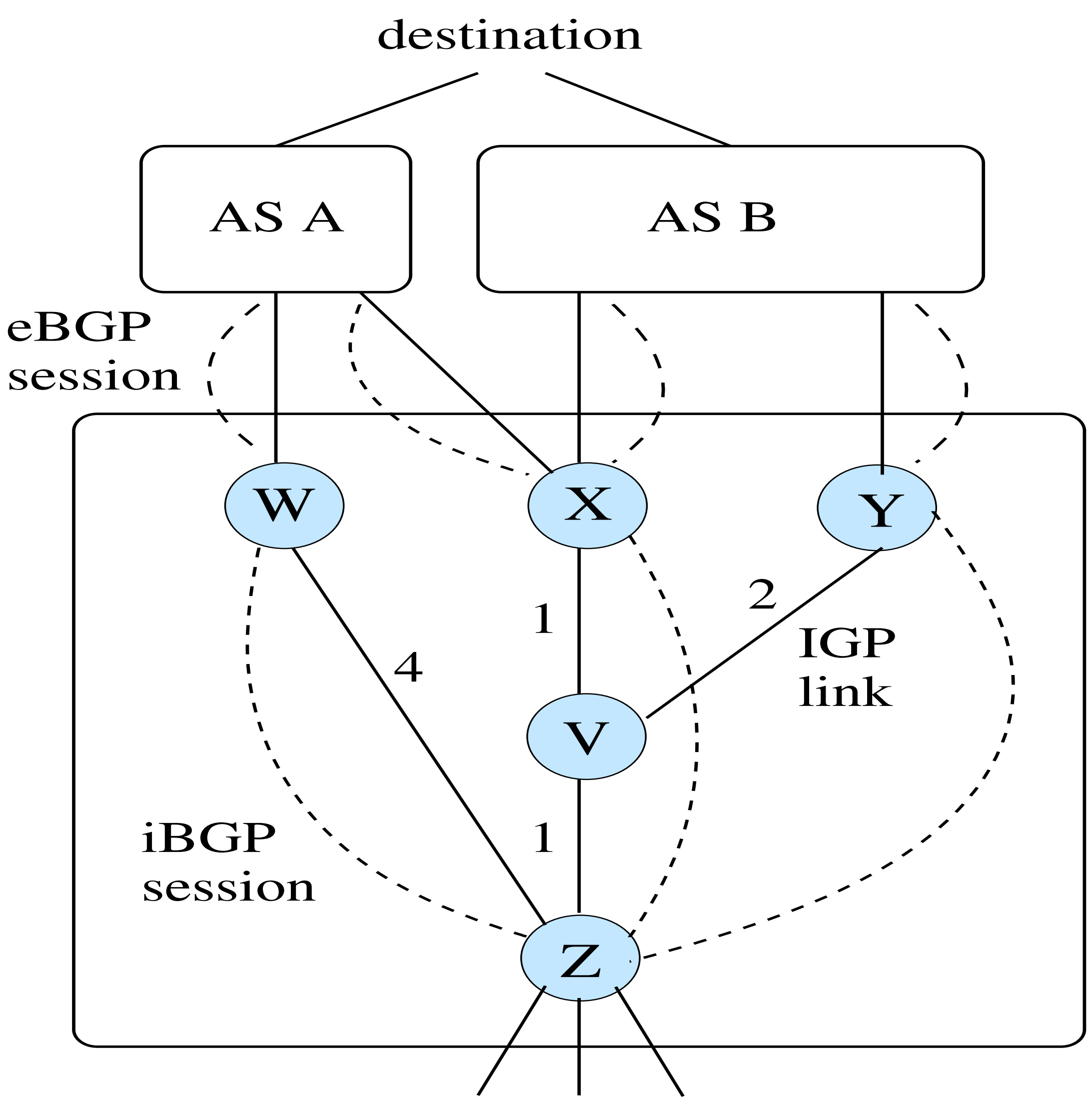
TTLMAN 402, R 17:30-20:00

some materials in this slide are based on lectures by

Jennifer Rexford https://www.cs.princeton.edu/courses/archive/fall13/cos597E/

RCP

BGP background



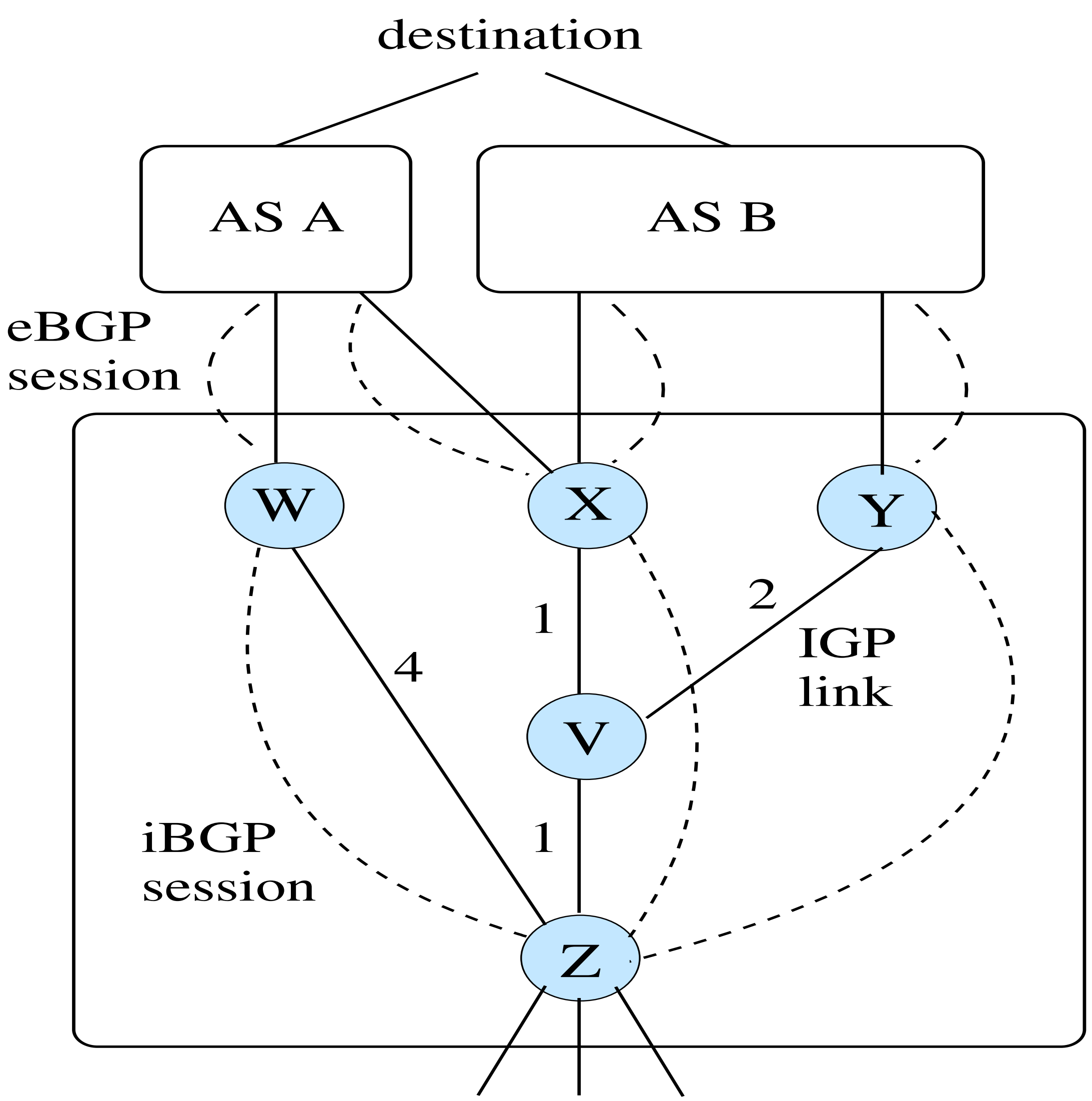
BGP

- de-facto inter-domain (inter-AS) routing protocol

functionality partitioned across routing protocols

- eBGP
- iBGP
- IGP

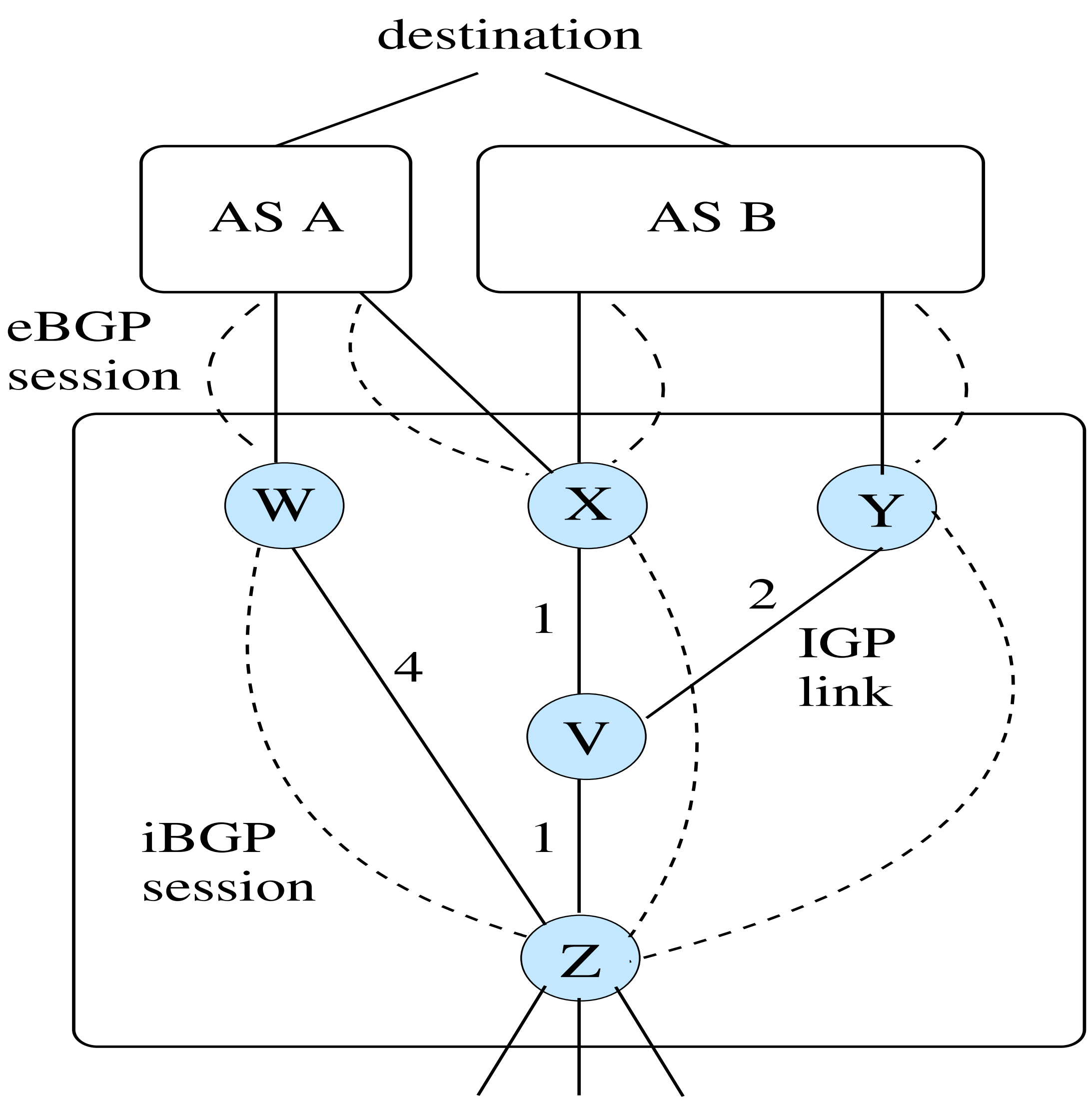
BGP background



BGP route-selection

1. highest local preference
2. lowest AS path length
3. lowest origin type
4. lowest MED (with next hop)
5. eBGP-learned over iBGP-learned
6. lowest path cost to egress
7. lower router ID

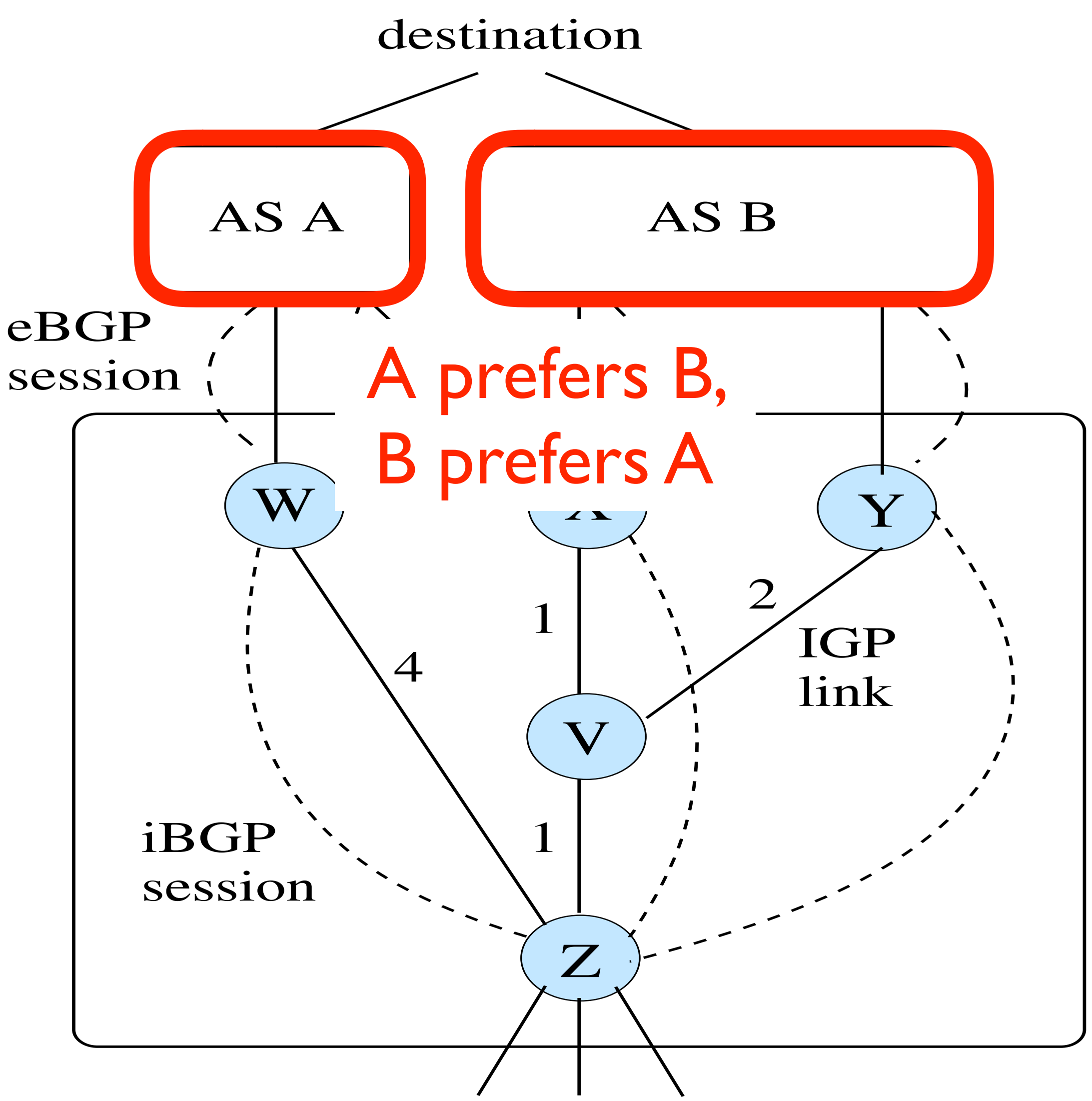
BGP: shortest path routing



BGP route-selection

1. highest local preference
2. lowest AS path length
3. lowest origin type
4. lowest MED (with next hop)
5. eBGP-learned over iBGP-learned
6. lowest path cost to egress
7. lower router ID

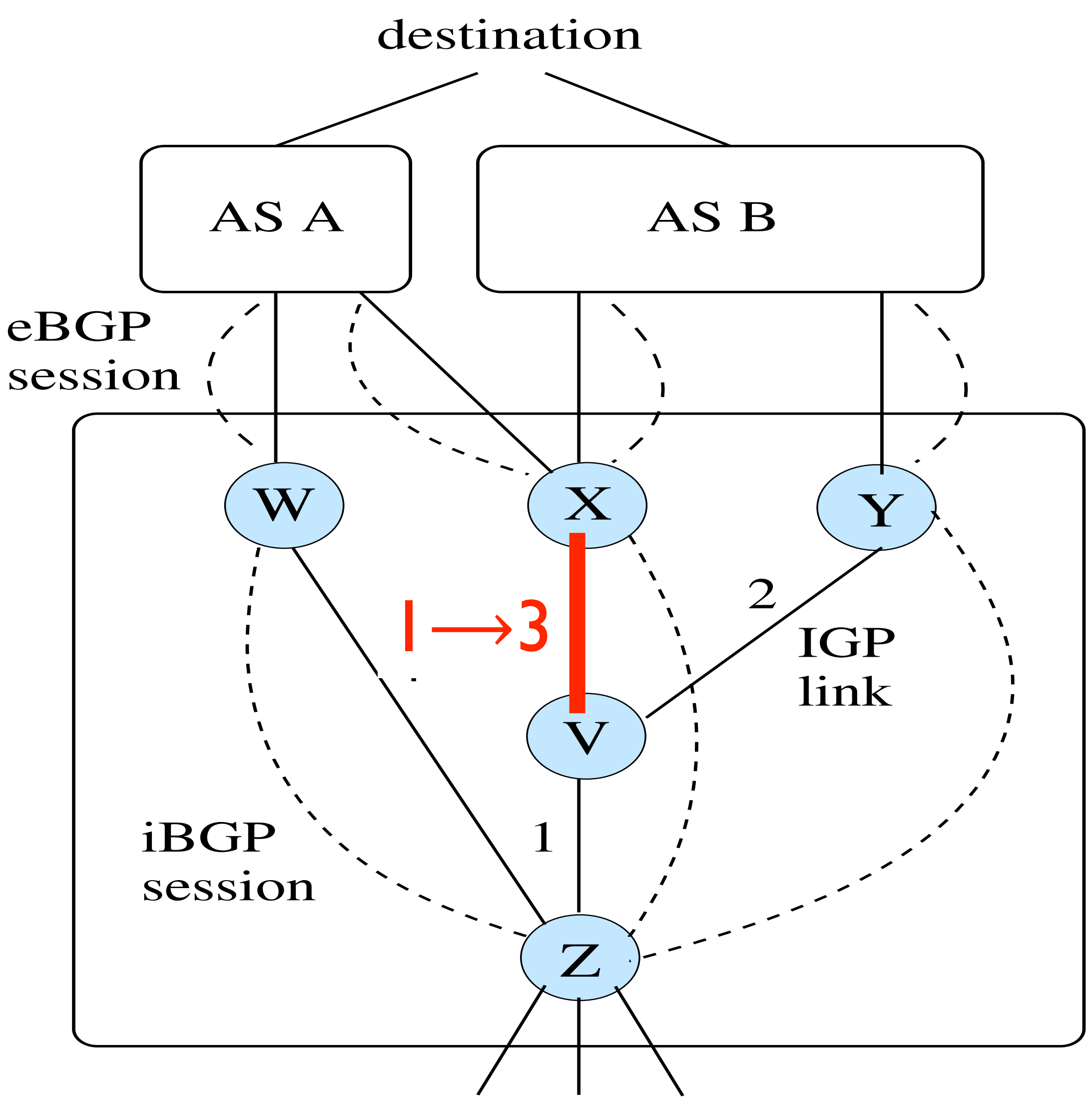
BGP problem: oscillation



BGP route-selection

1. highest local preference
2. lowest AS path length
3. lowest origin type
4. lowest MED (with next hop)
5. eBGP-learned over iBGP-learned
6. lowest path cost to egress
7. lower router ID

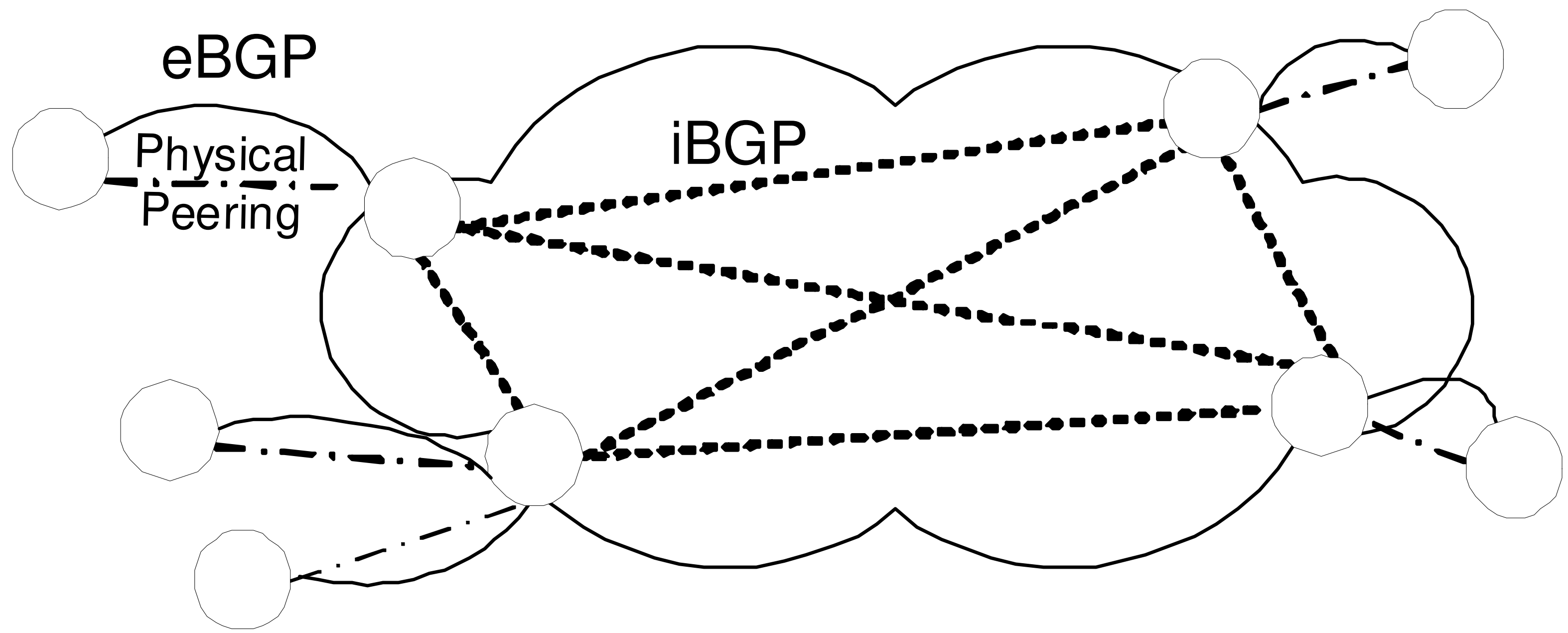
BGP problem: hot-potato



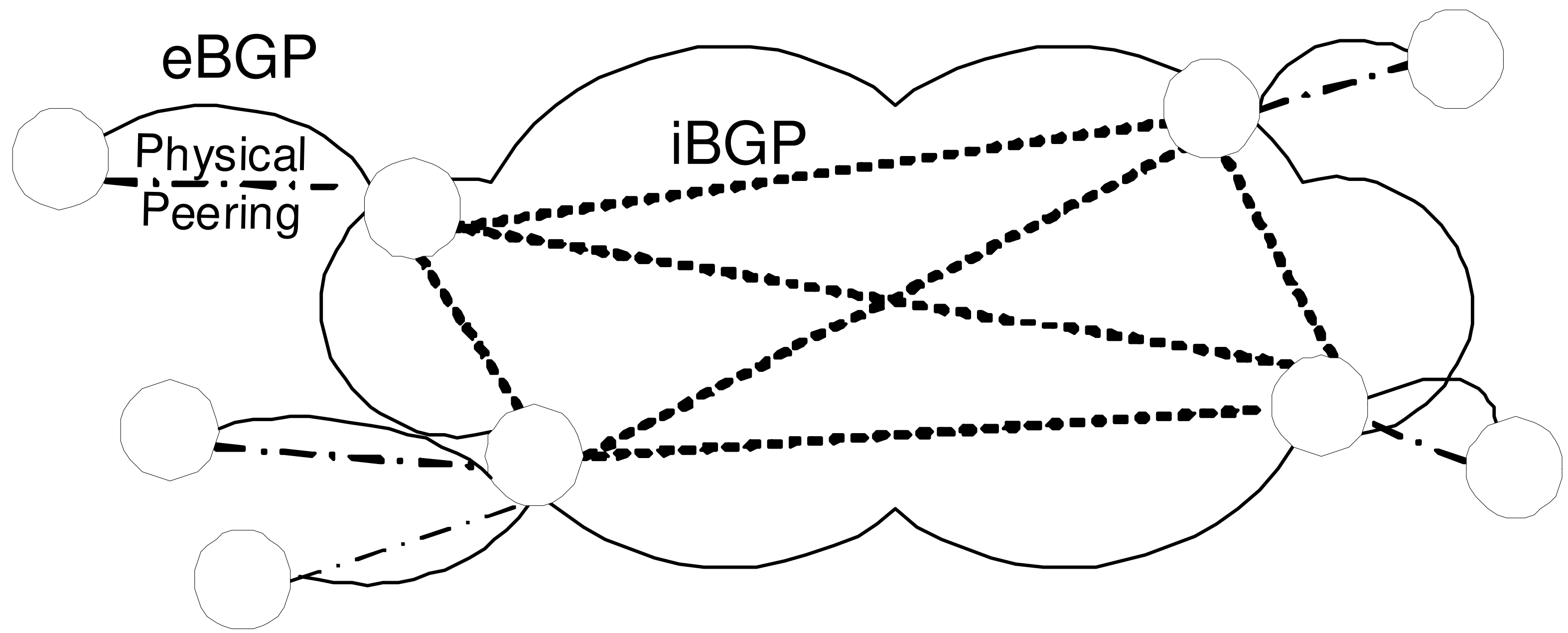
BGP route-selection

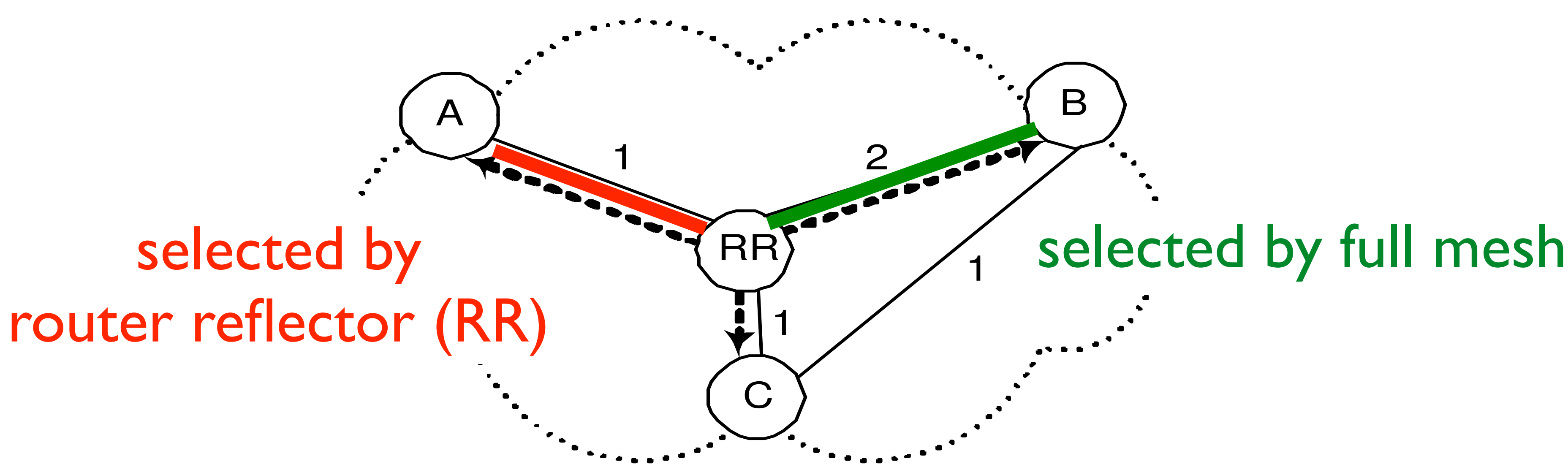
1. highest local preference
2. lowest AS path length
3. lowest origin type
4. lowest MED (with next hop)
5. eBGP-learned over iBGP-learned
6. lowest path cost to egress (hot-potato, early-exit)
7. lower router ID

BGP problem: RR ≠ full-mesh



BGP problem: RR ≠ full-mesh





BGP problems

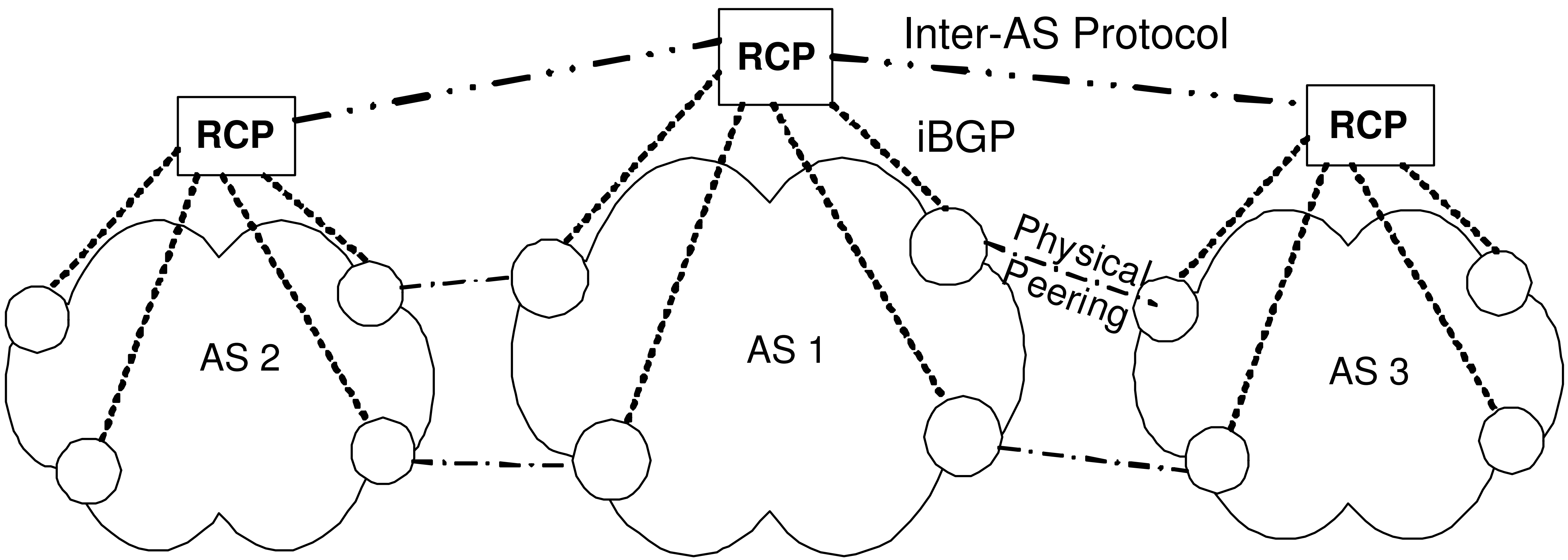
BGP is broken

- converge slowly, sometimes not at all
- routing loops
- misconfigured frequently
- traffic engineering is hard

fixing BGP is hard

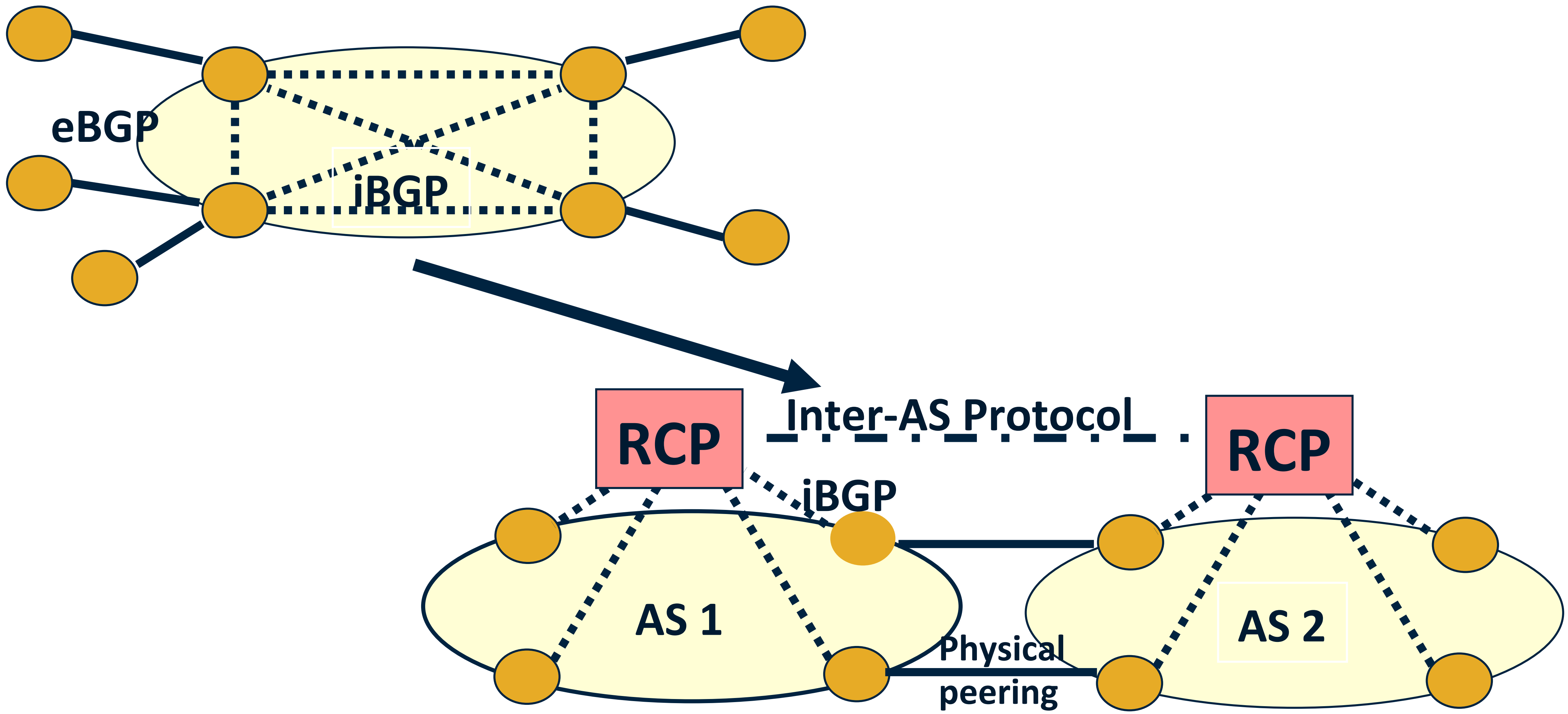
- incremental fixes: even more complex
- deployment of new inter-domain protocol almost impossible

solution: RCP



use centralized controller to customize control

- controller computes routes on behalf of routers
- uses existing routing protocol for control traffic

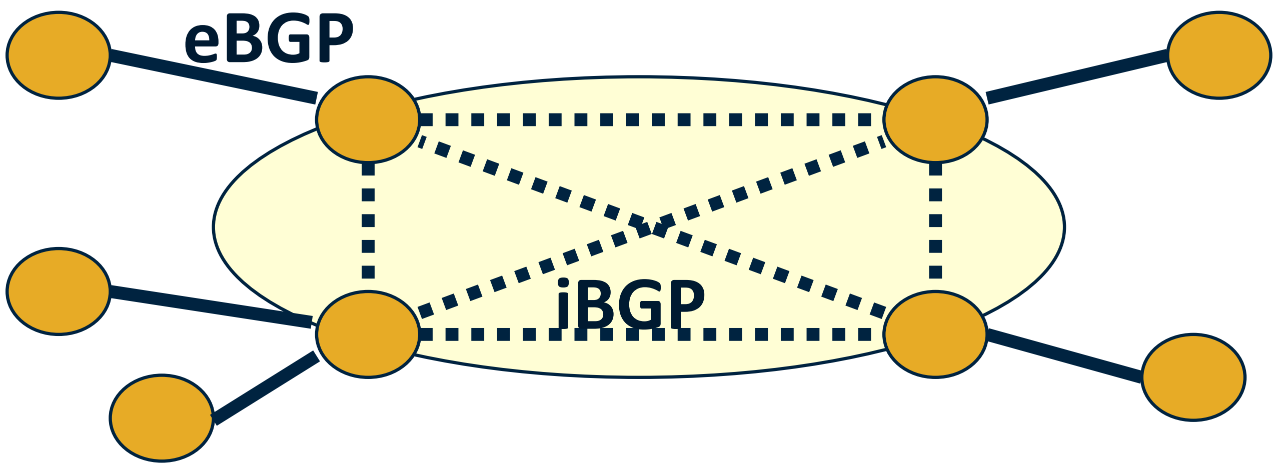


3 phases to achieve

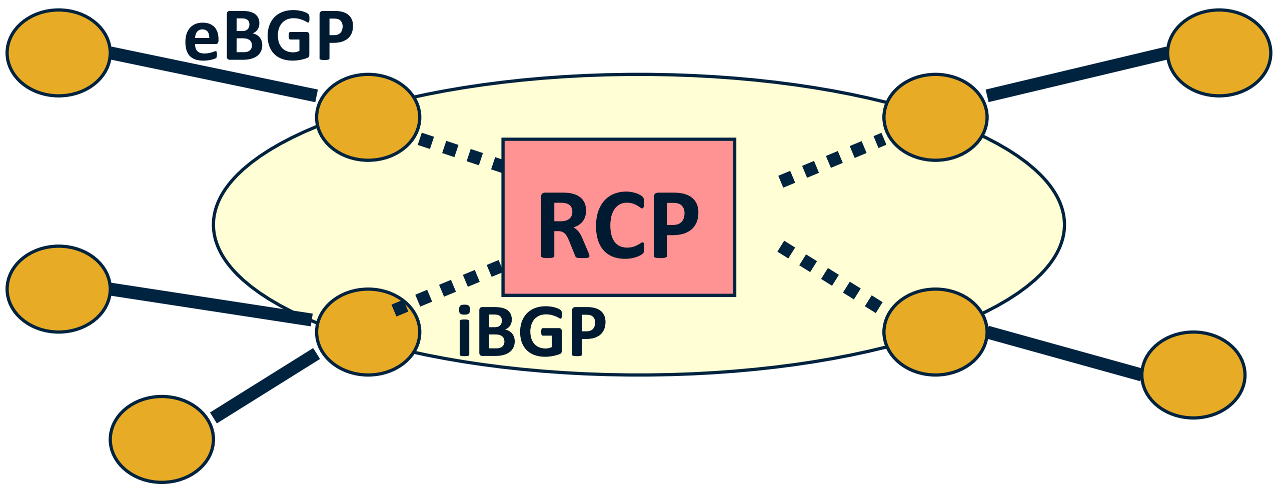
- backward compatibility, deployment incentives

phase 1: control protocol interactions

Before: conventional iBGP



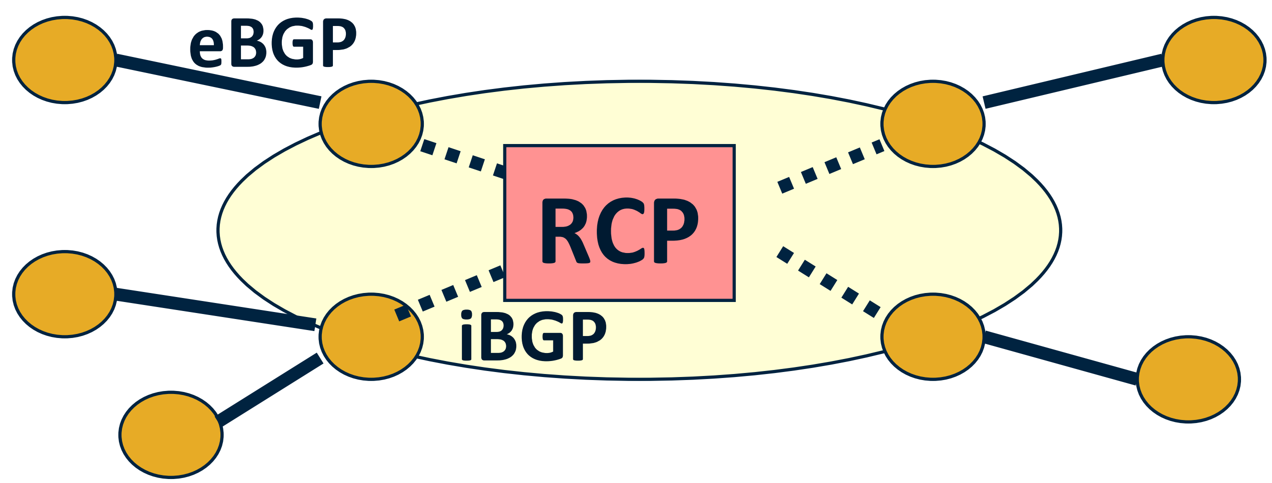
After: RCP gets "best" iBGP routes (and IGP topology)



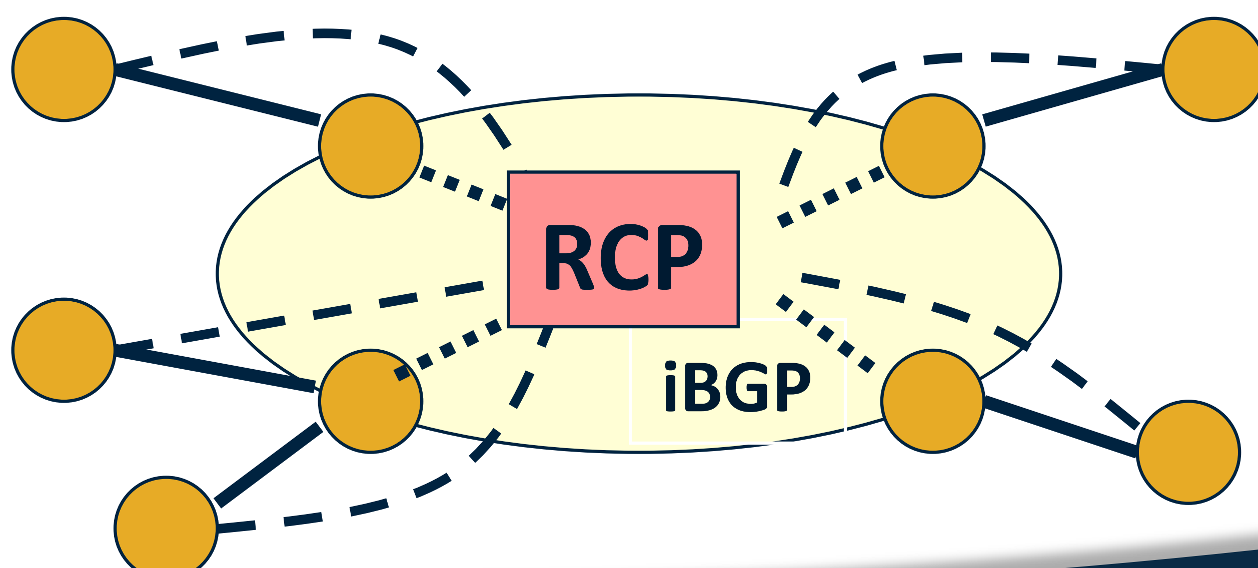
only one AS has to change

phase 2: AS-wide policy

Before: RCP gets "best" iBGP routes (and IGP topology)

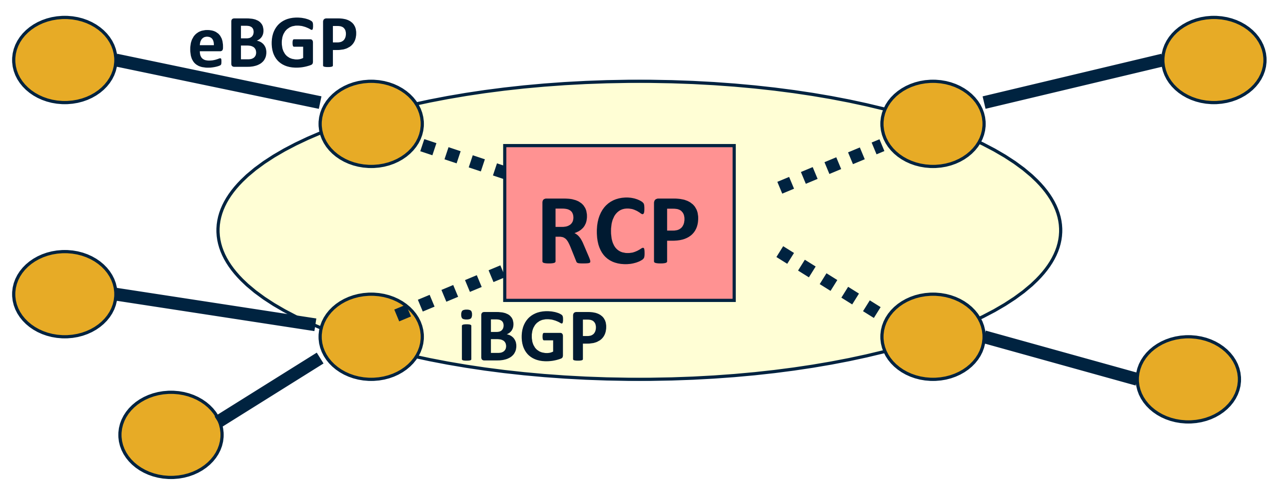


After: RCP gets all eBGP routes from neighbors

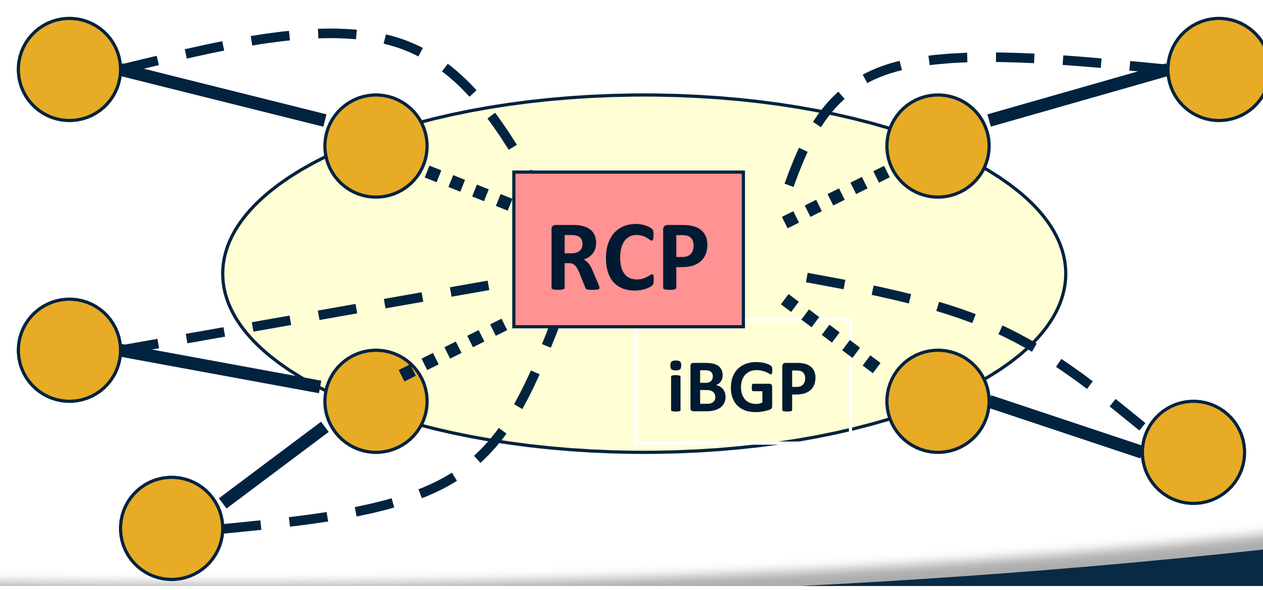


phase 3: AS-wide policy

Before: RCP gets "best" iBGP routes (and IGP topology)

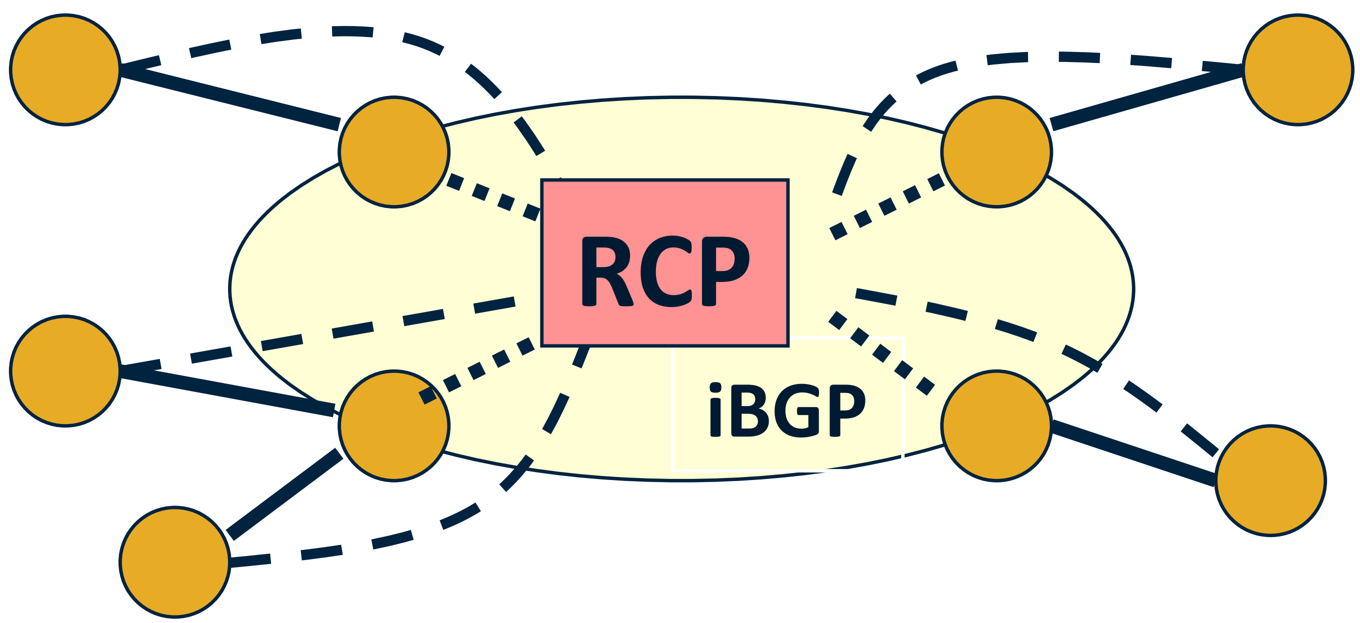


After: RCP gets all eBGP routes from neighbors

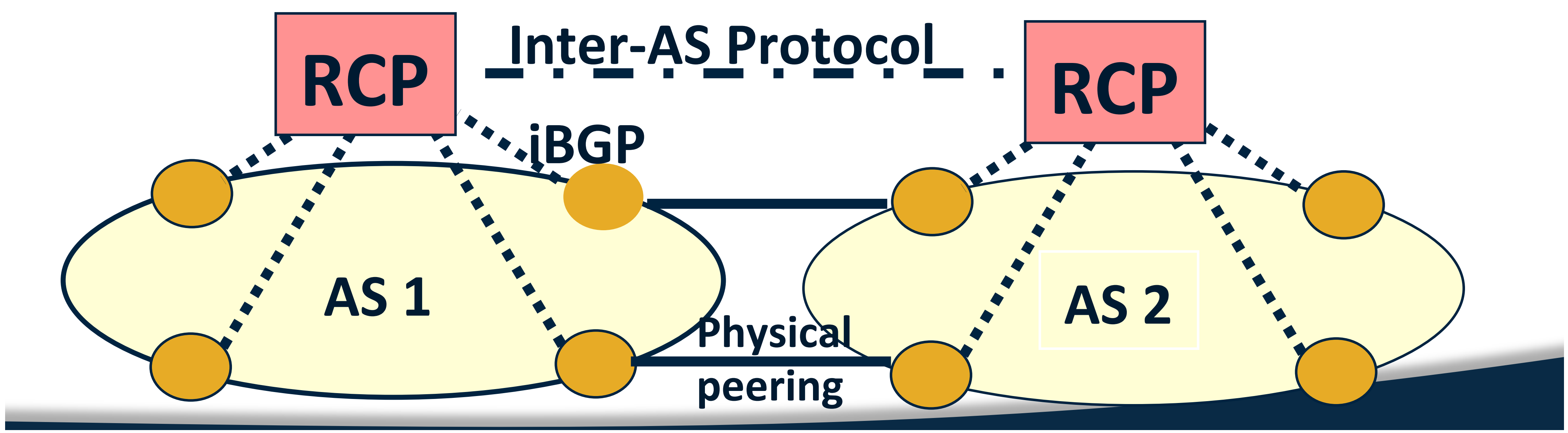


phase 3: all ASes have RCP

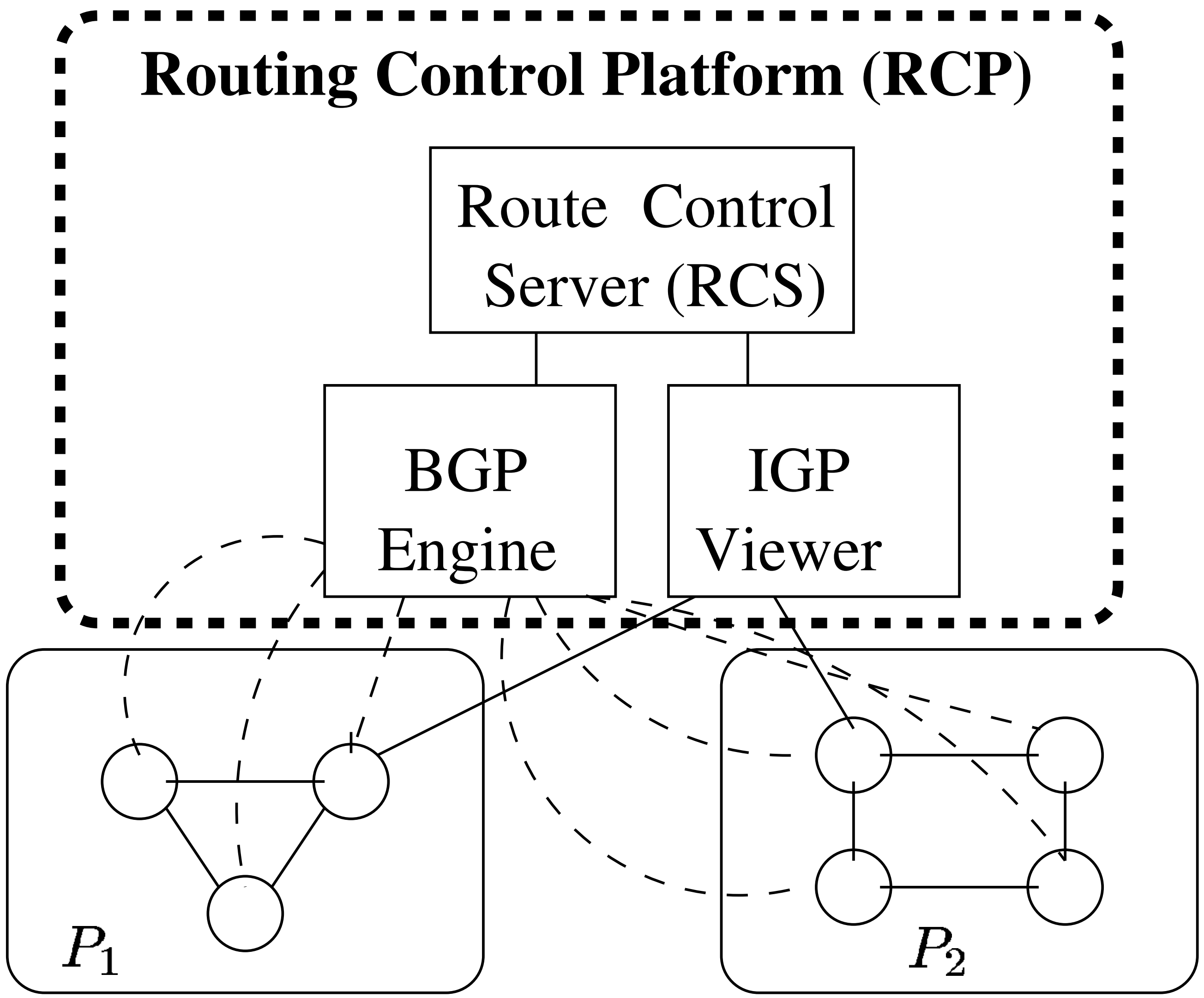
Before: RCP gets all eBGP routes from neighbo



After: ASes exchange routes via RCP



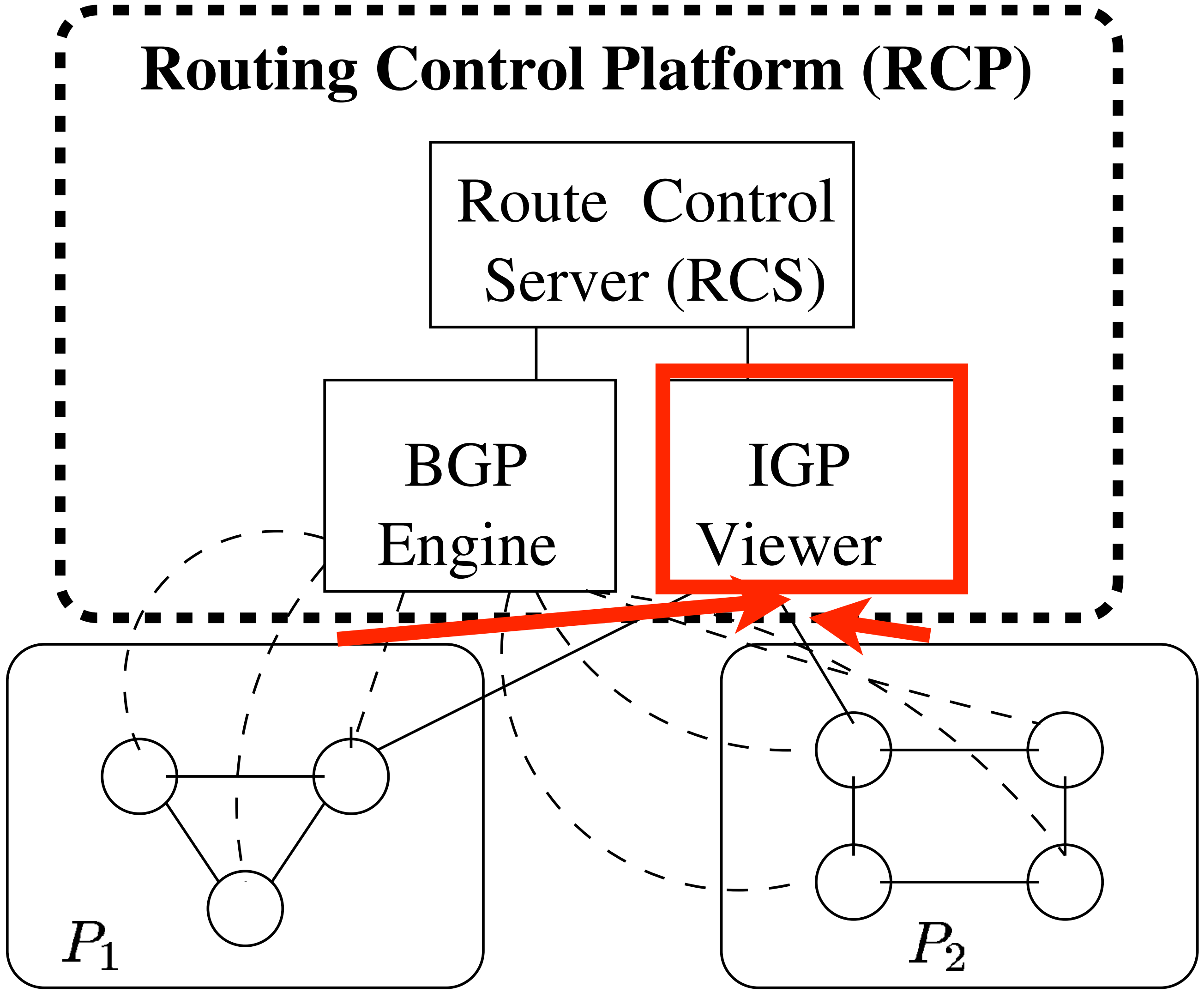
RCP architecture



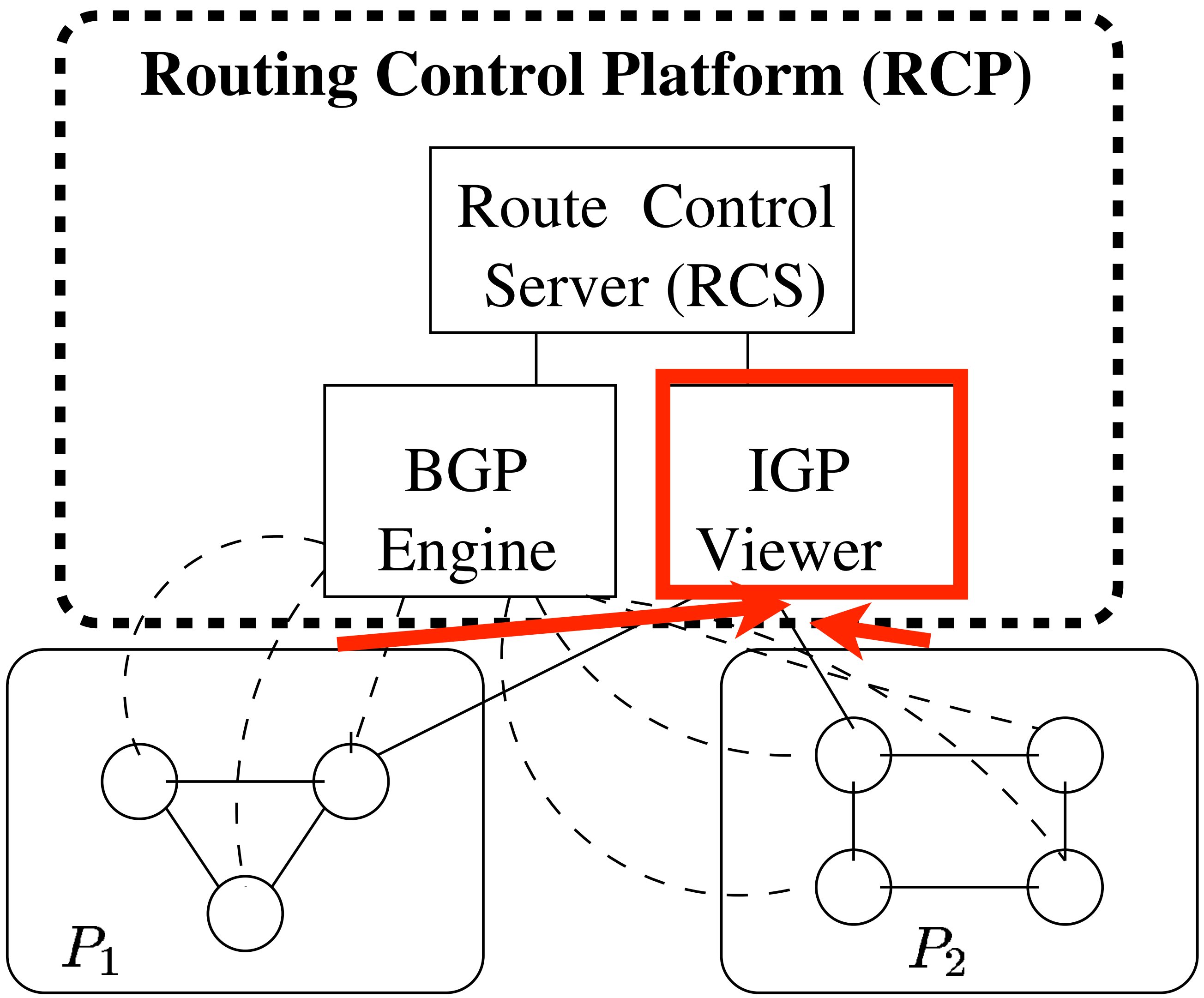
P1, P2

- IGP partitions

RCP architecture



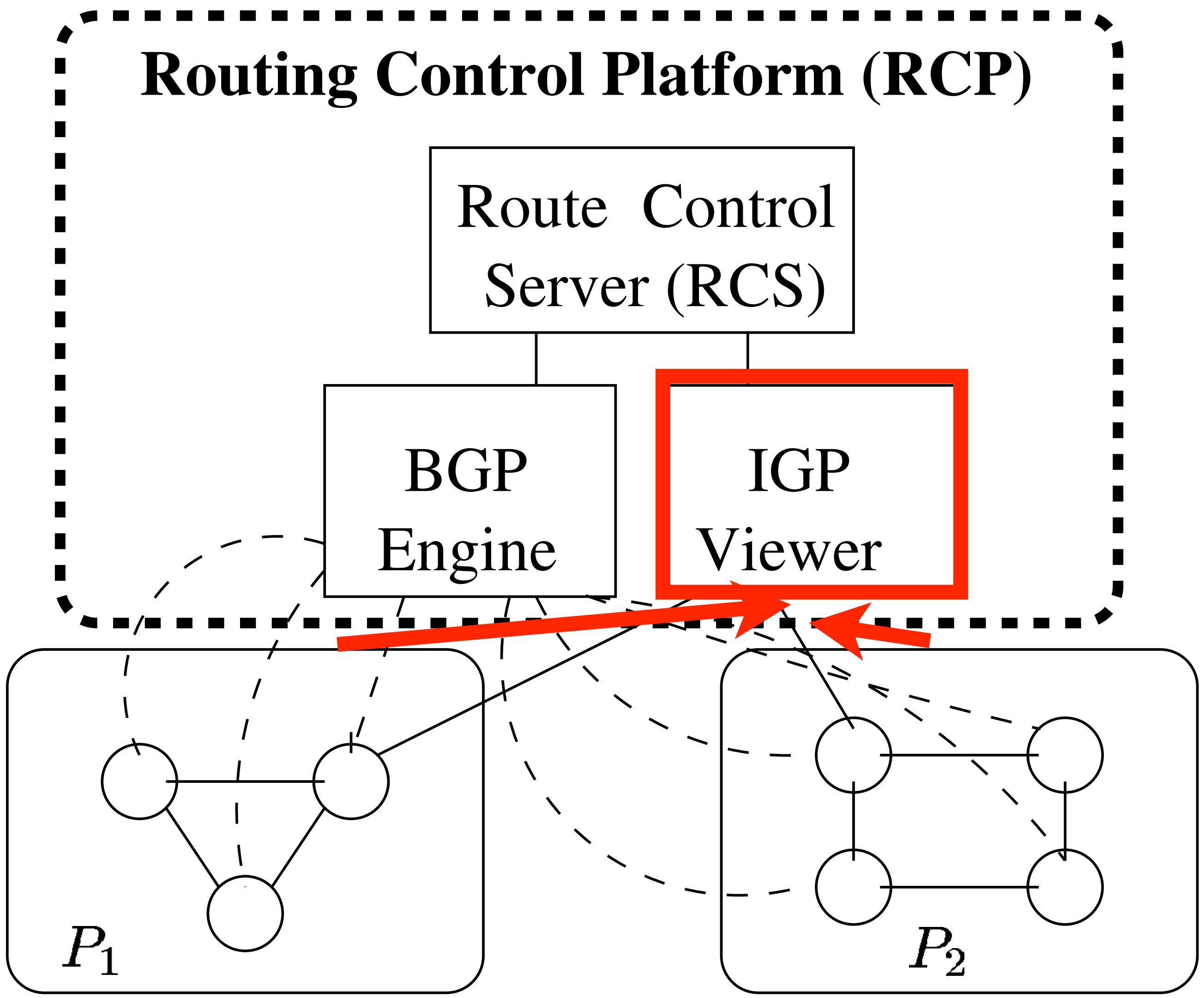
RCP architecture



IGP viewer

- maintains IGP topology
- computes pairwise shortest paths with AS

RCP architecture



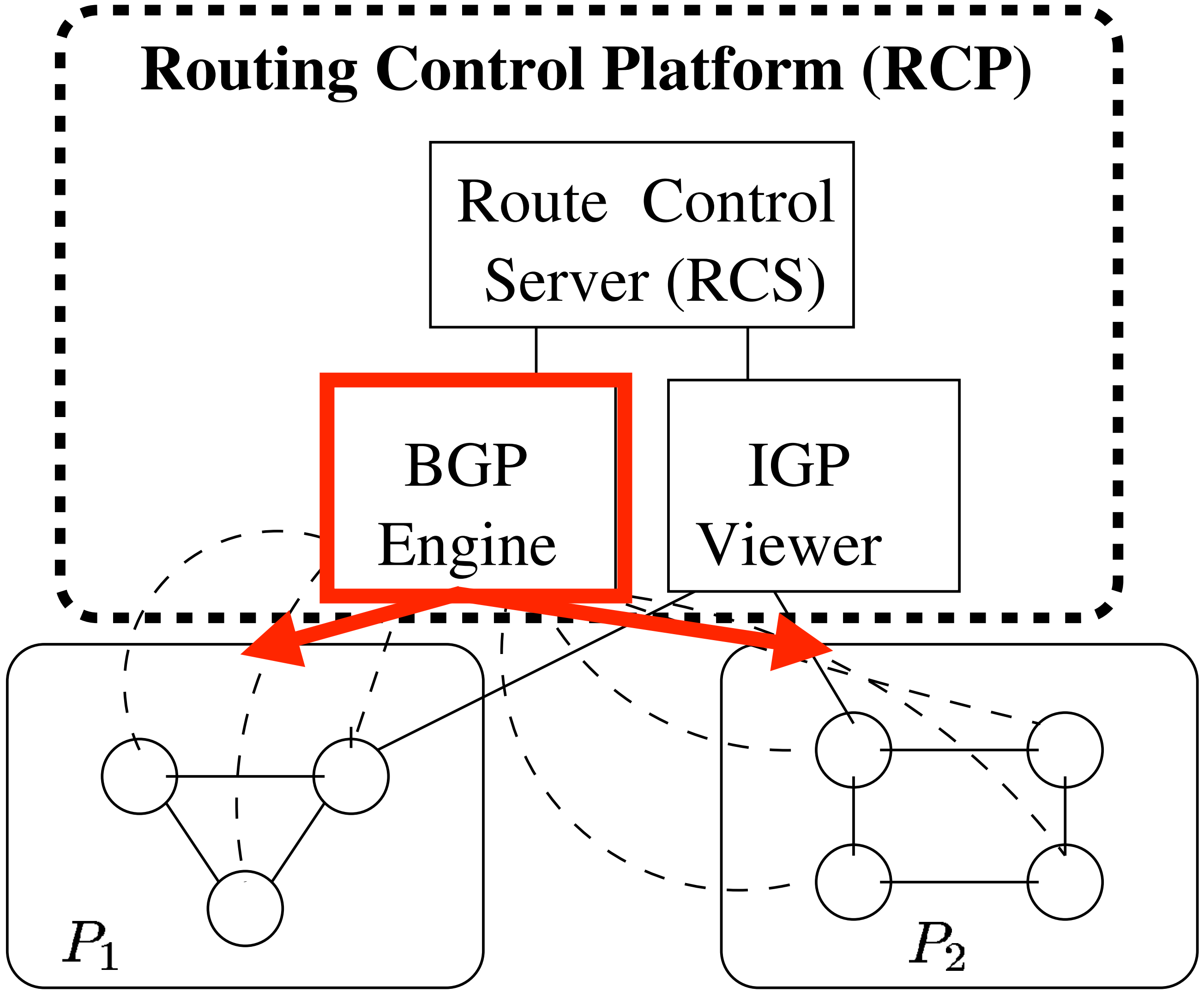
IGP viewer

- maintains IGP topology
- computes pairwise shortest paths with AS

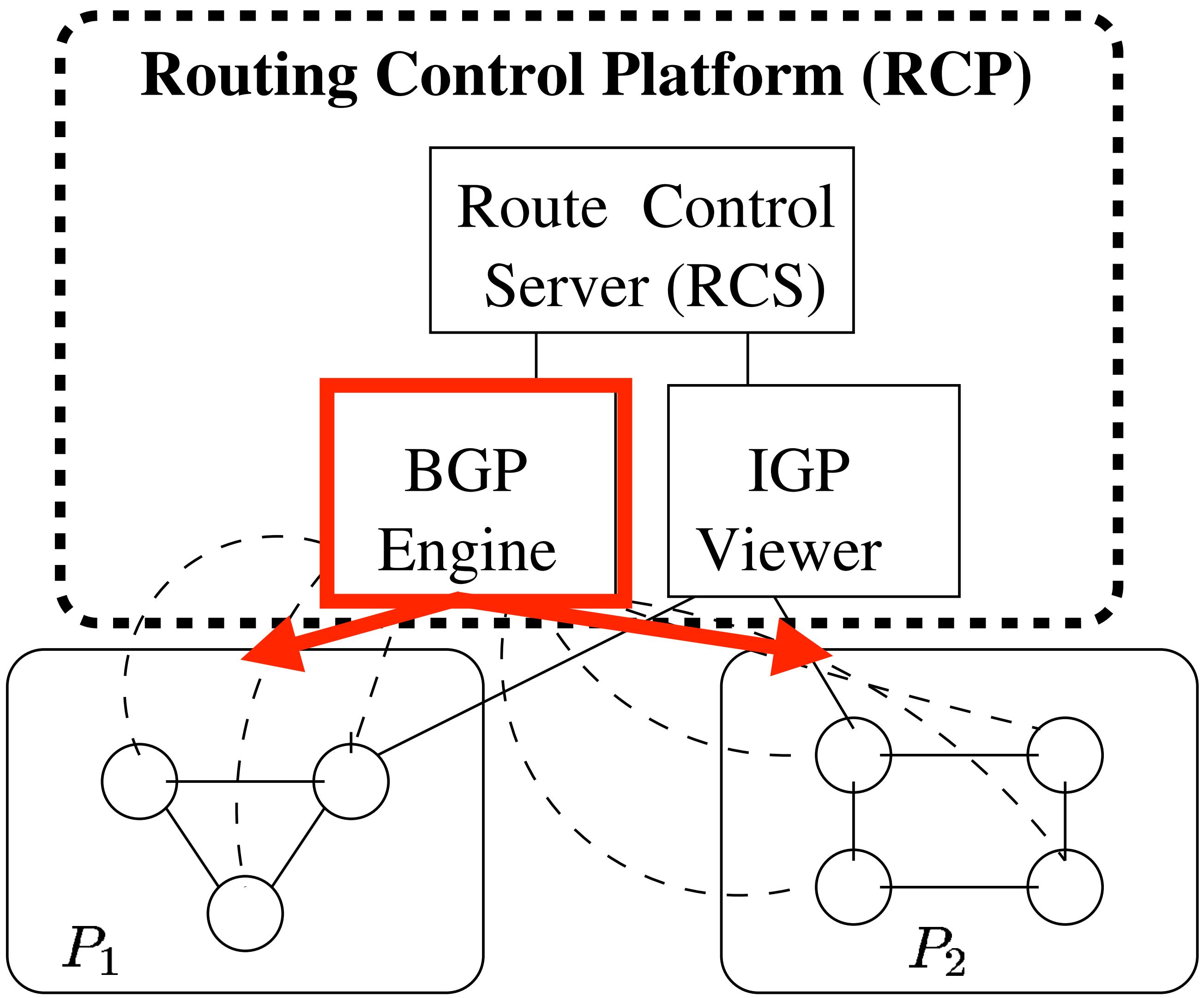
benefit: scalability

- cluster routers
- reduce # independent route computation

RCP architecture



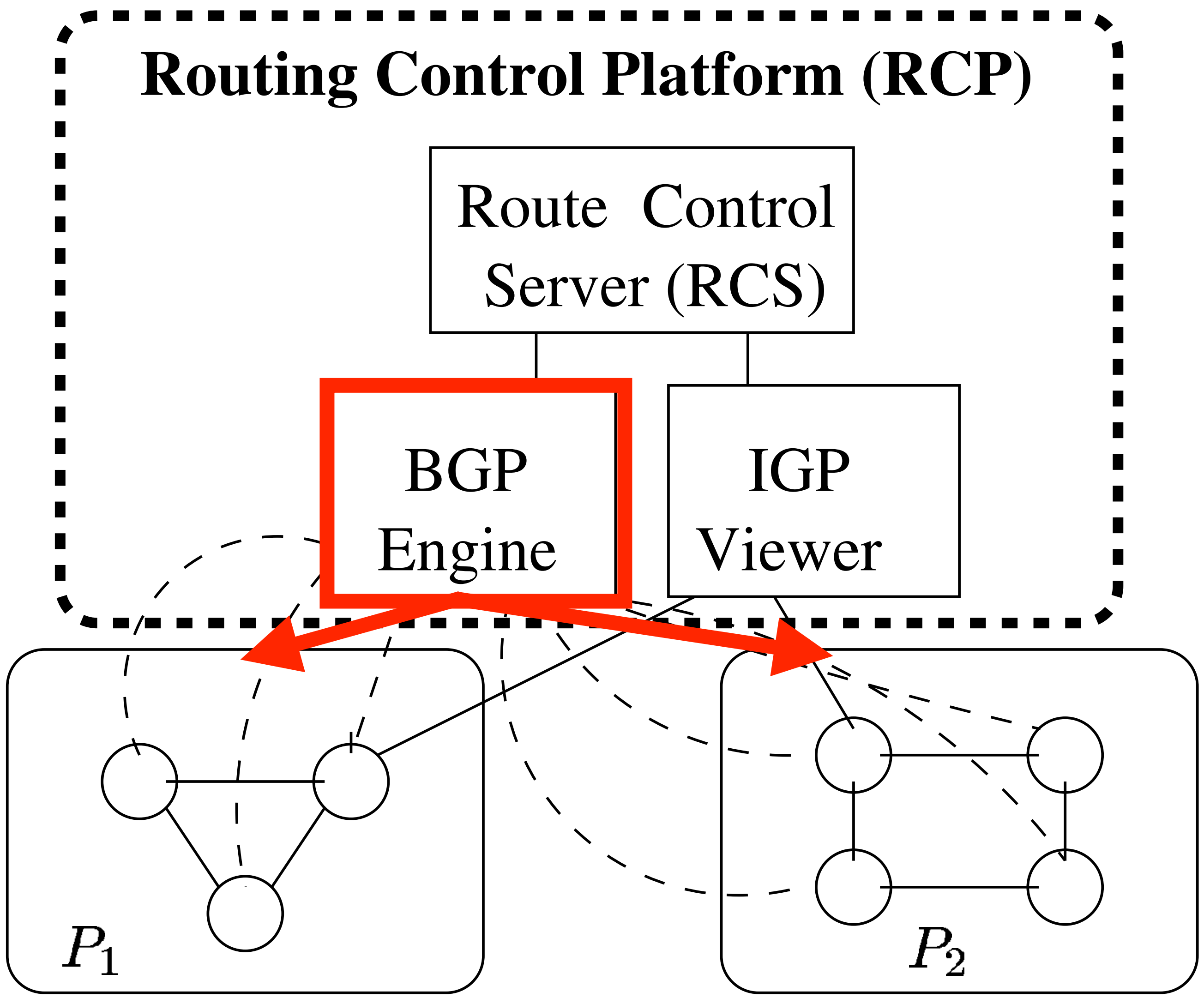
RCP architecture



BGP engine

- communicates RCS decision to routers via iBGP

RCP architecture



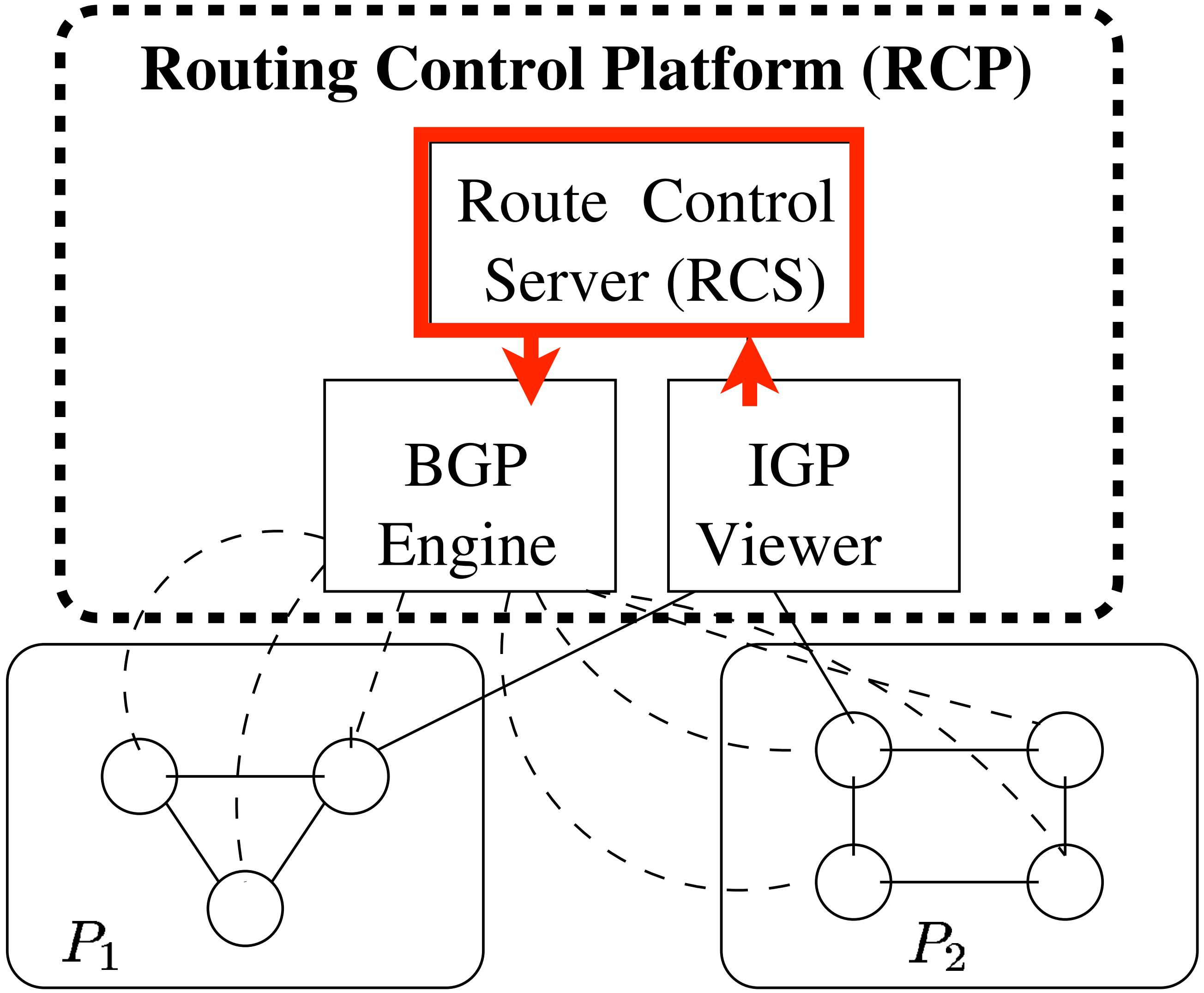
BGP engine

- communicates RCS decision to routers via iBGP

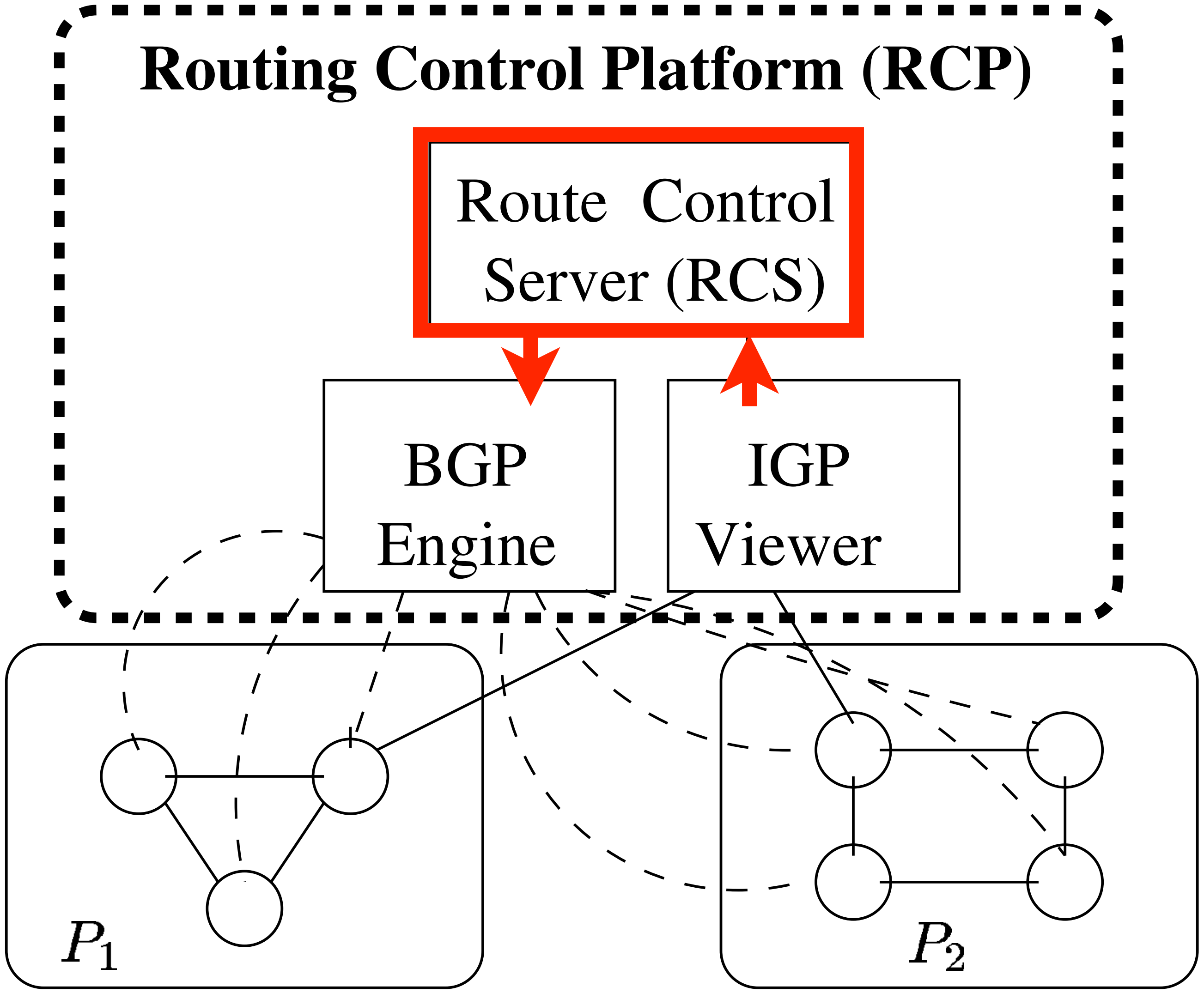
benefit

- backward-compatibility

RCP architecture



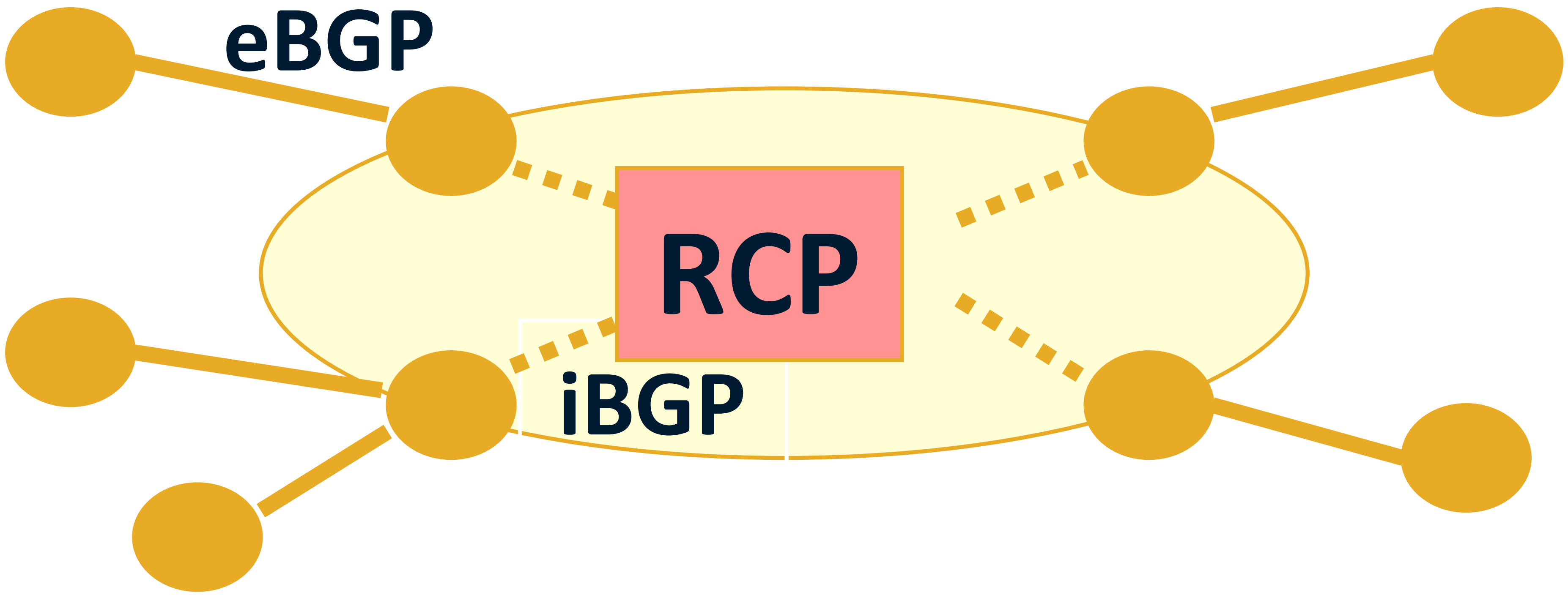
RCP architecture



RCS

- computes BGP route assignments
- obtain topology from IGP
- disseminate decision via BGP engine

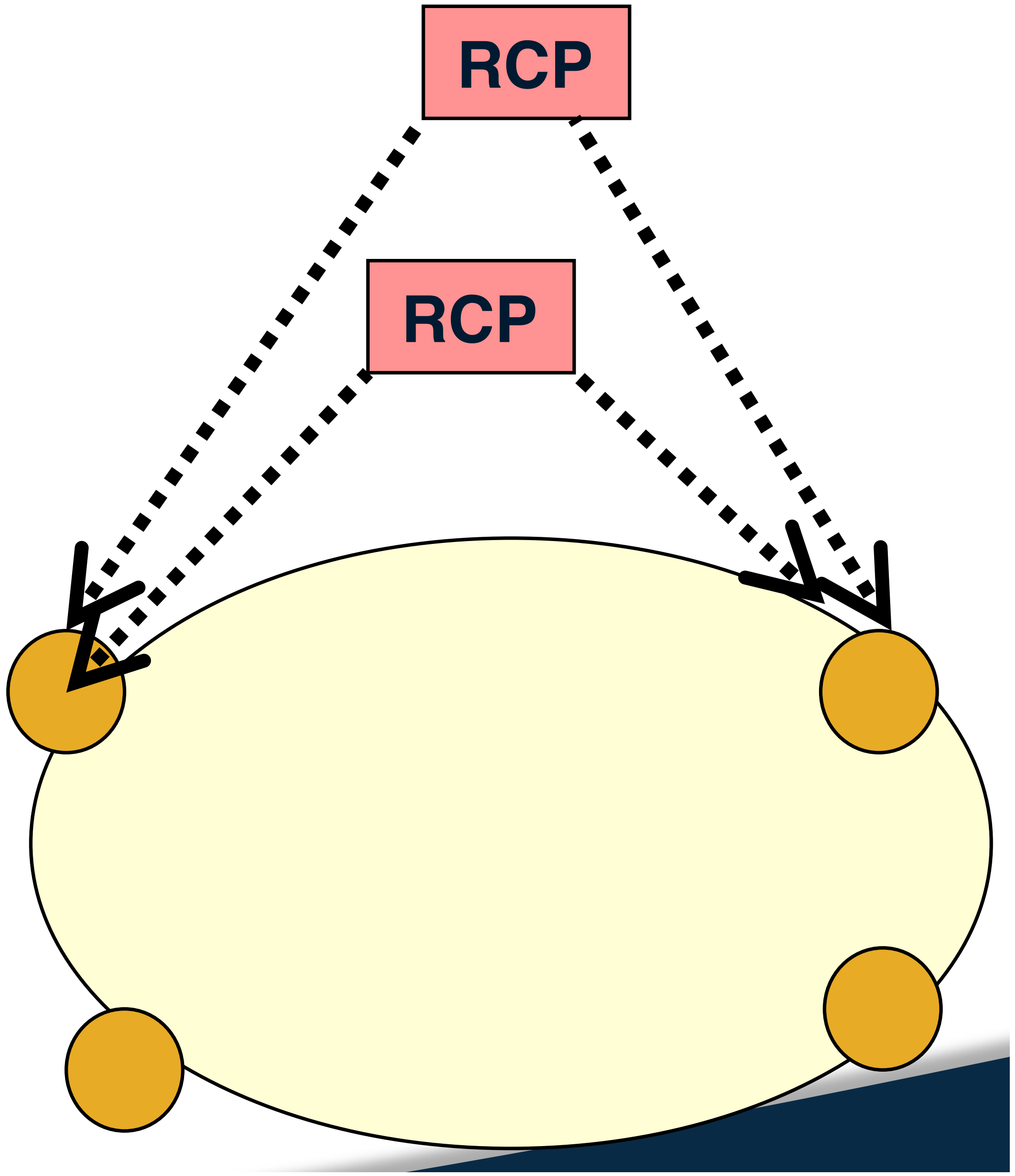
scalability, efficiency, and reliability



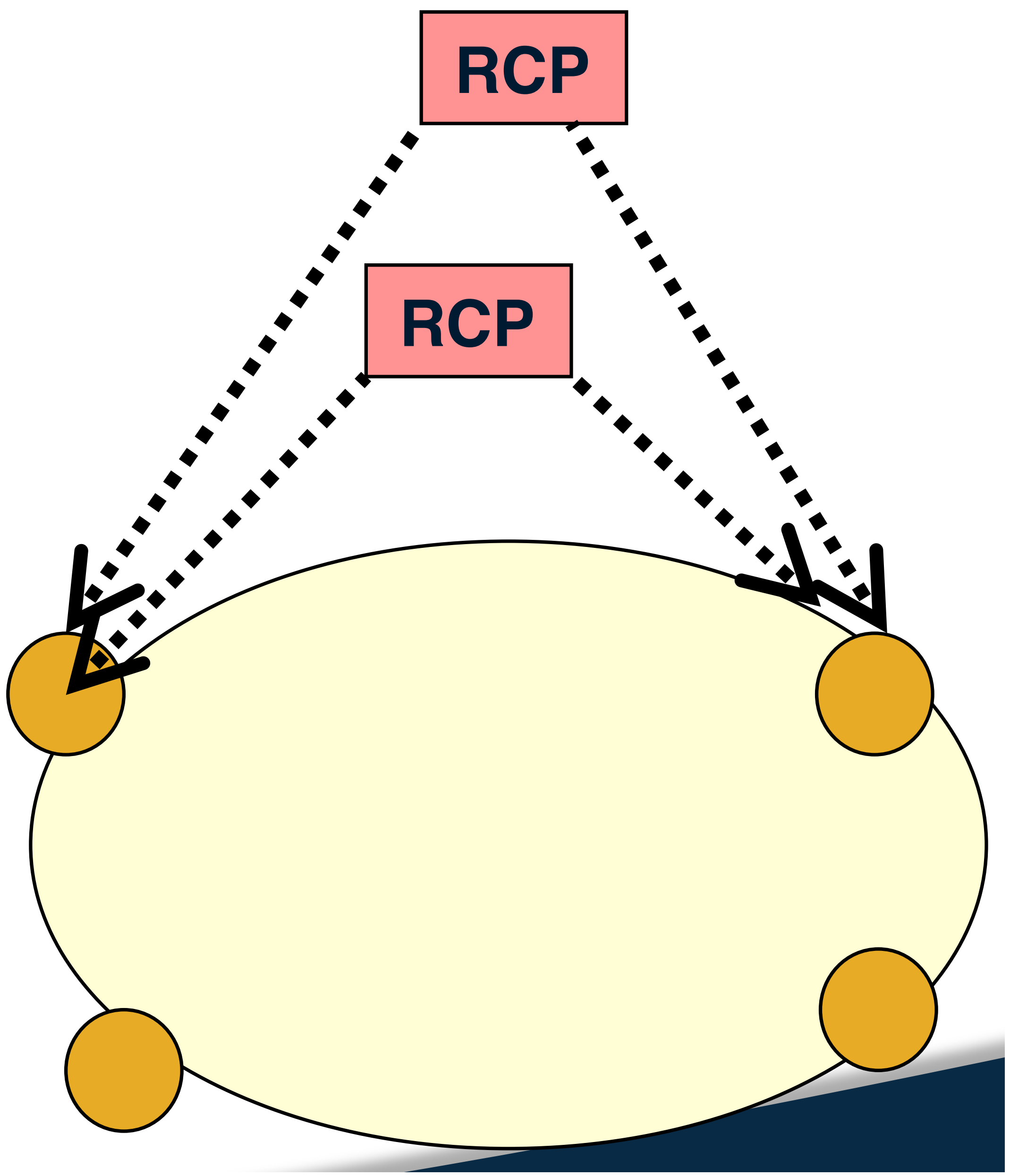
requirements

- many routers (500-1000)
- many destination prefixes (150,000-200,000)
- converge quickly

reliability



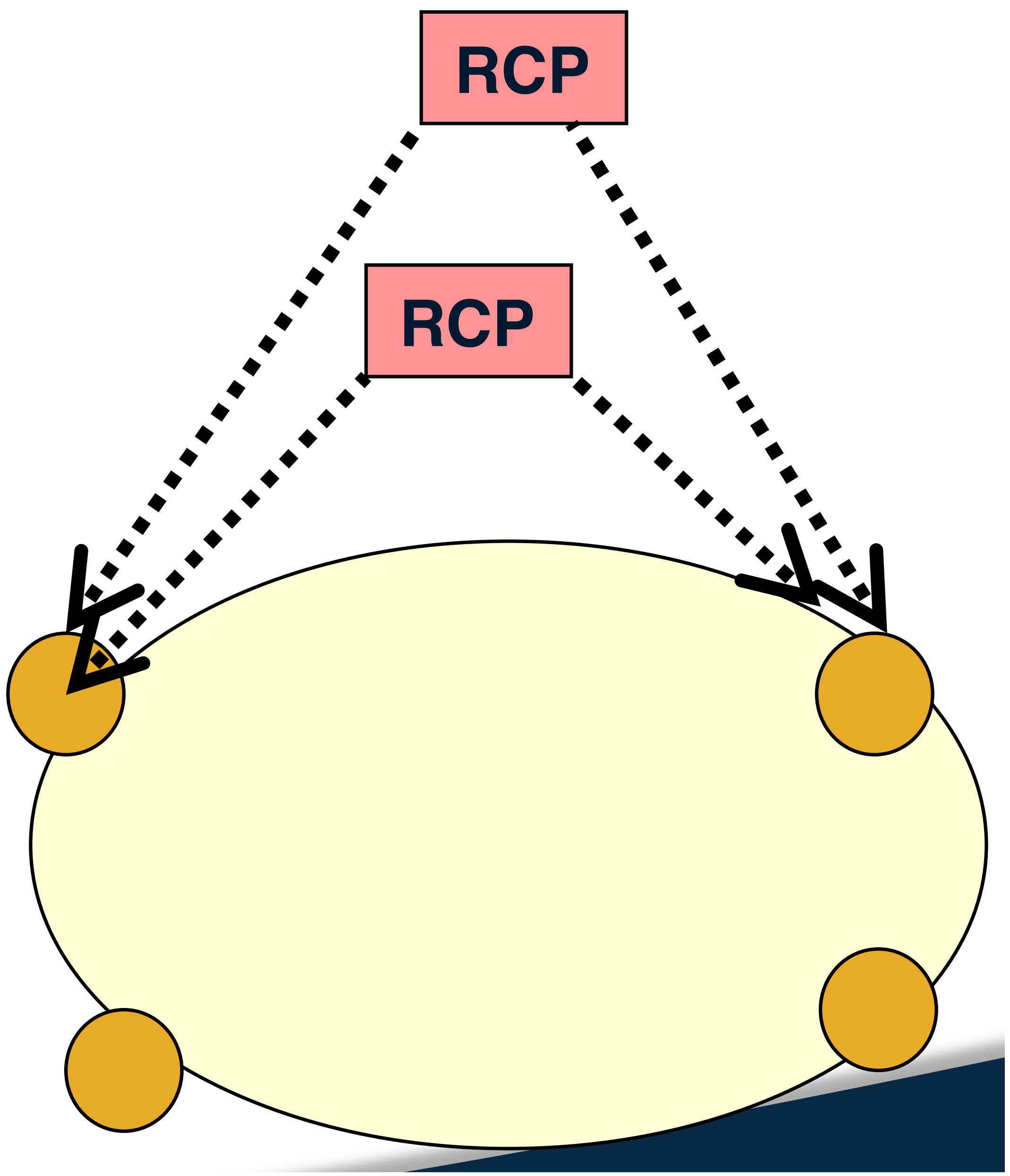
reliability



replicate RCP

- multiple identical servers

reliability



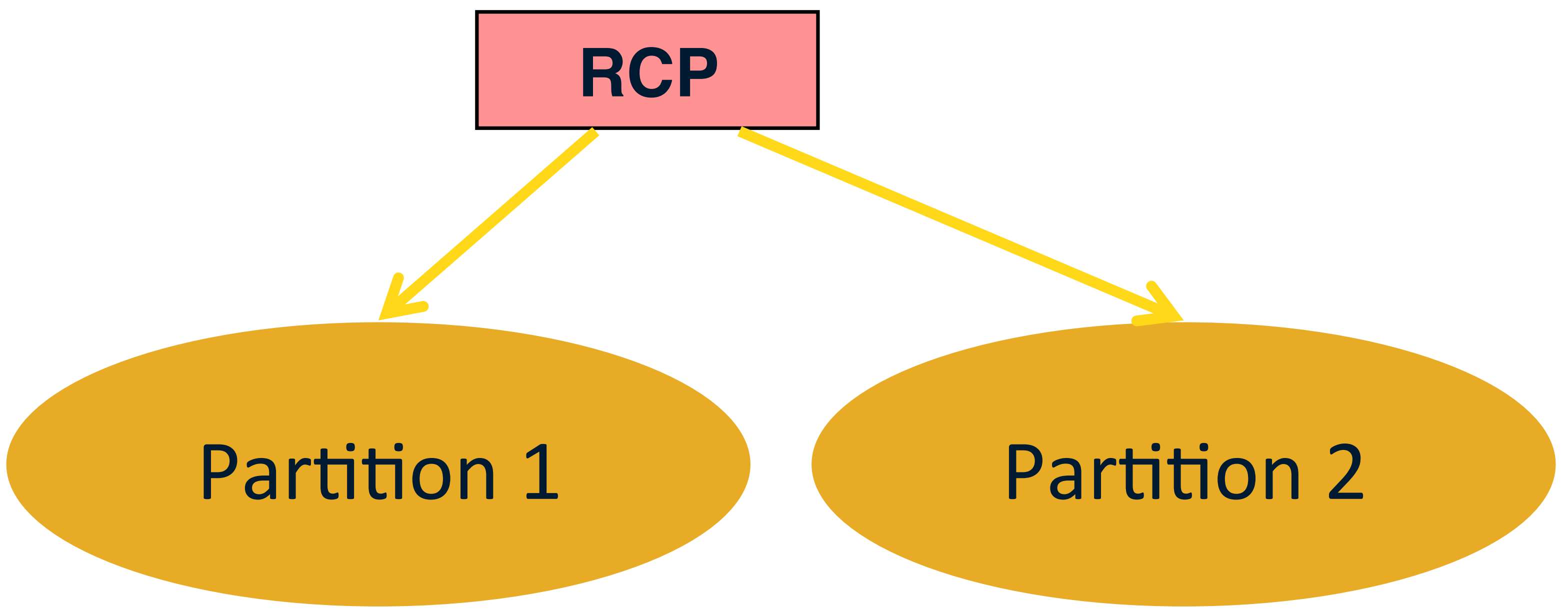
replicate RCP

- multiple identical servers

independent replicas

- each receives same information, running the same routing algorithm
- *NO* need for a consistency protocol if both replicas always see the same information

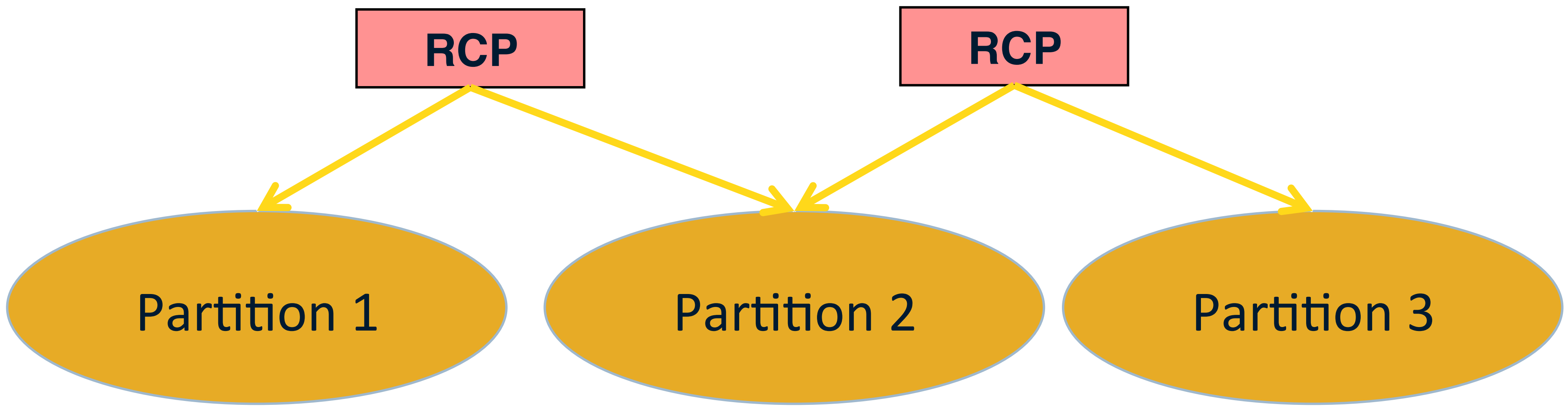
single RCP under partition



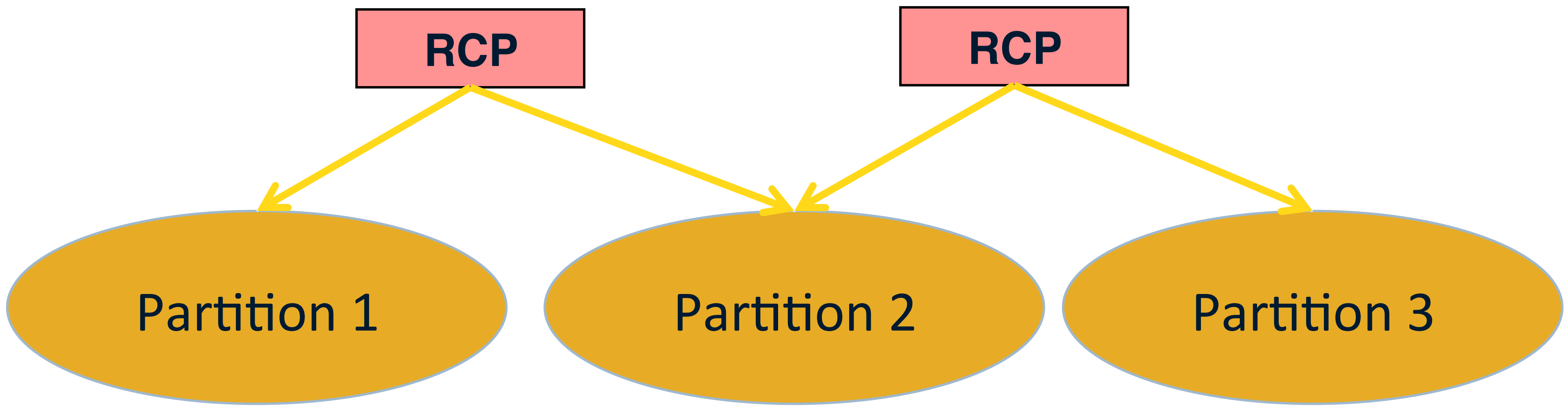
only use state from routers' partition to assign BGP route

- ensure next-hop is reachable

multiple RCPs under partition



multiple RCPs under partition



RCPs receive same state from each reachable partition

- IGP offers complete visibility
- only acts on partition with complete state

three continual challenges

three continual challenges

scalability

- large topology, huge volume of events, flow initiations

three continual challenges

scalability

- large topology, huge volume of events, flow initiations

reliability

- handle equipment (and other) failover gracefully

three continual challenges

scalability

- large topology, huge volume of events, flow initiations

reliability

- handle equipment (and other) failover gracefully

performance

- low control-plane latency