lecture 05: centralized control
—opportunities and challenges

5590: software defined networking

ando 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/](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 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

![Diagram of BGP network showing eBGP and iBGP connections.]

- **eBGP**: External Border Gateway Protocol
- **iBGP**: Internal Border Gateway Protocol
- **Physical Peering**: Physical connectivity between routers
- **Paths**: Various routes and connections within the network

The diagram illustrates the distinction between eBGP and iBGP, showing how they interconnect through physical peering points.
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
Three 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 neighbors

**After:** ASes exchange routes via RCP
RCP architecture

PI, P2
- IGP partitions

Routing Control Platform (RCP)

Route Control Server (RCS)

BGP Engine

IGP Viewer

$P_1$, $P_2$
RCP architecture

Routing Control Platform (RCP)

Route Control Server (RCS)

BGP Engine

IGP Viewer

$P_1$

$P_2$
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

Routing Control Platform (RCP)

Route Control Server (RCS)

BGP Engine

IGP Viewer

$P_1$

$P_2$
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

Routing Control Platform (RCP)

Route Control Server (RCS)

BGP Engine

IGP Viewer

$P_1$

$P_2$
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: RCP

Replicate RCPs ("Hot Spare")

Run multiple identical servers

Run independent replicas

Each replica has its own feed of routes

Each replica receives the same inputs and runs the same routing algorithm

No need for a consistency protocol if both replicas always see the same information
Reliability: RCP

- Replicate RCPs ("Hot Spare")
- Run multiple identical servers

Each replica has its own feed of routes.
Each replica receives the same inputs and runs the same routing algorithm.

No need for a consistency protocol if both replicas always see the same information.
**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

Solution:
- RCPs receive the same state from each partition they can reach.
- IGP provides complete visibility and connectivity.
- Only acts on a partition if it has a complete state.

No consistency protocol needed to guarantee consistency in a steady state.

Diagram:
- RCPs connected to three partitions:
  - Partition 1
  - Partition 2
  - Partition 3
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