Networking: Routing

Dixon Tirtayadi & Winston Jodjana

Multiple Types of Routing

Interior Gateway Protocols
(Within *autonomous systems*)

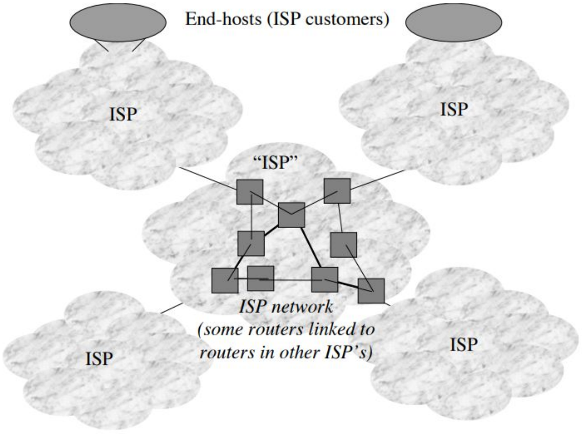
- Type 1, Link-state (e.g. OSPF, IS-IS)
- Type 2, Distance-vector (e.g. RIP, RIPv2, IGRP)

Exterior Gateway Protocols
(Between *autonomous systems*)

- BGP (Path-vector)

Autonomous Systems

- Collections of connected IP routing prefixes
- 2 main types of relationships between ASes
  - Transit (provider-customer)
  - Peering (e.g. 2 ISPs let each other's traffic through)



Relationships between ASes

- Transit (provider-customer)
  - Customer needs to be reachable from and reach to everyone.
  - Provider (ISP) gets money from Customer, Customer gets reachability
  - Example: You pay ISP for internet connectivity
- Peering (peer-peer)
  - 2 ASes let each other's traffic reach only each other's customers
  - Important: traffic needs to avoid being highly asymmetric or it becomes "unfair"
  - Example: 2 ISPs let each other's traffic through

ISPs

- Make up majority of these *autonomous systems* (ASes)
- The main way for people to get internet connectivity
- Are competitive with each other
- 3 Types:
  - Tier 3 ISP (Small)
    - Local scope
  - Tier 2 ISP (Medium)
    - Regional scope (e.g. ISPs that control small countries)
  - Tier 1 ISP (Large)
    - Usually has global connectivity (i.e. routes to the entire internet)
    - Have many ASes

Exporting Routes: Route Filtering

What routes does it want to show to its neighbours?

Neighbour is a:

- Customer (e.g. individuals)
  - Want to show because customers need to be able to receive packets from anywhere
- Provider (e.g. bigger ISPs)
  - Don't want to show unless customers are involved
- Peer (e.g. other ISPs)
  - Want to show only selected routes (especially involving customers)

General Rule: Only the ones they earn revenue from

Importing Routes

ASes receive multiple routes, what routes do ASes want to install in their forwarding table?

Quite complicated but:

General Rule: Customer > Peer > Provider

Border Gateway Protocol: An Overview

Why do we need BGP? Why not just use some shortest-path algorithm for the entirety of Internet?

Design Goals:

- Scalability
  - Many, many ASes will exist
- Policy
  - Independent policy for each AS
- Cooperation under competitive circumstances
  - Needs to be able to route people's traffic even when they are at odds against each other

What are other important **goals** that are not mentioned? Why that goal?

- Security
- Convergence
- Performance

Border Gateway Protocol: How it works

- Application layer protocol: runs over TCP on port 179
- OPEN
  - Initialization, exchange routes
- UPDATE
  - … any updates to routes (changes, deletion, etc.)
- KEEPALIVE
  - … are you alive?
- Misconfiguration by network operators can lead to different kind of issues.

eBGP and iBGP

eBGP (external, the standard)

- BGP between routers in different ASes
- Loop-free forwarding
- Complete visibility

iBGP (internal)

- BGP between routers in same AS
- Only external routes!
- Terrible scalability due to complete mesh requirement
  - Router Reflectors
  - Confederations

iBGP Alternatives

- Route Reflectors
  - Selects a single best route to each destination prefix and announces that route
  -
- Confederations
  - Division to Sub-ASes
  - Full mesh within *confederations*
  - eBGP-like behaviour between *confederations*

BGP Policy Expression: Filters and Rankings

Policy Tasks:

- Ranking of Routes
- Load balancing
- Tagging Routes

Important route **attributes** to enforce policies:

NEXT HOP, ASPATH, LOCAL PREF, MULTIPLE-EXIT DISCRIMINATOR (MED)

NEXT HOP

IP Address of the next-hop router along the path to the destination.

ASPATH

Sequence (vector) of AS identifiers that the route advertisement has traversed.

- Between AS boundaries, each AS adds itself to this attribute
- Loop avoidance
  - Check if AS is already in the ASPATH attribute: if so, drop the route announcement
- Pick a "best" path
  - Usually (but not always) the shortest route i.e. shortest vector

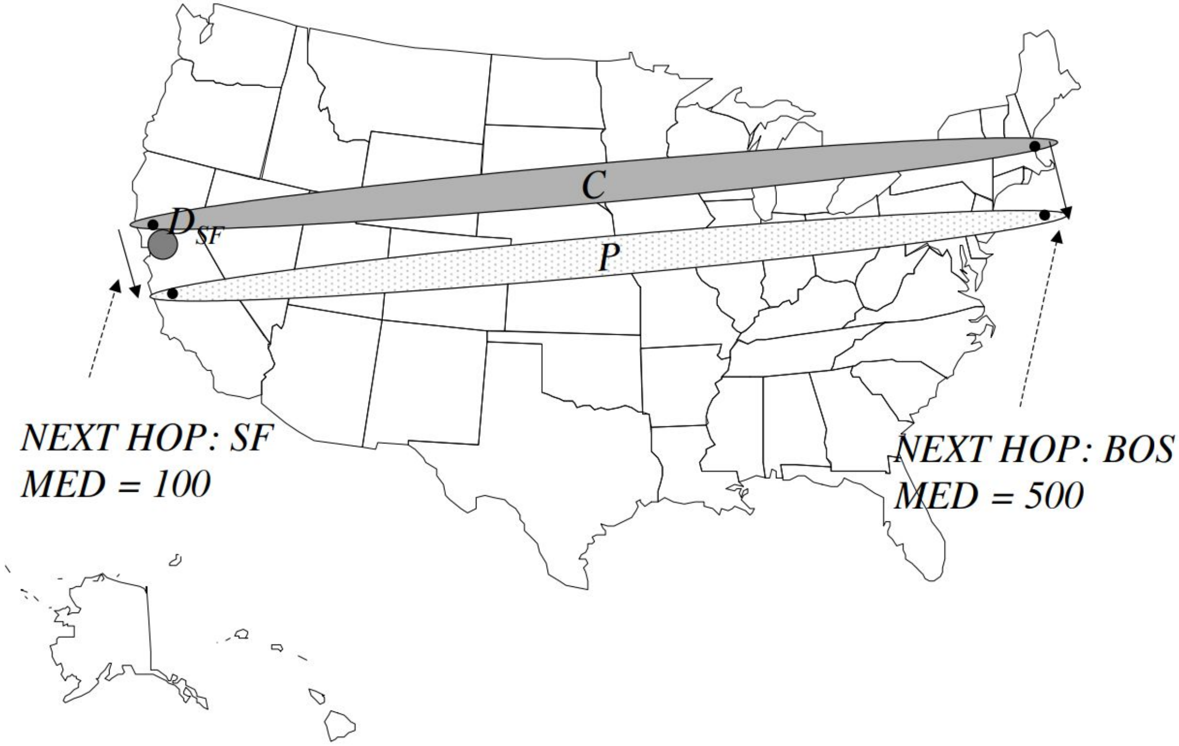
LOCAL PREF

(really only used in iBGP)

This (optional) attribute is the first criteria used to select routes.

It's basically a "filter" to prefer certain routes over others. Very rarely used in practice; usually just use shortest path within an ISP.

MULTIPLE-EXIT DISCRIMINATOR (MED)



ASPATH: Security Issue

Prefix Hijacking with ASPATH

Malicious AS can:

- Advertise itself as the wrong AS (via prefix manipulation)
- Advertise a fake shortest path

e.g.

"Attackers hijacked BGP prefixes that belonged to a South Korean cryptocurrency platform, and then issued a certificate on the domain via ZeroSSL to serve a malicious JavaScript file, stealing $1.9 million worth of cryptocurrency."

What are some possible fixes to these security Issues?

Prefix Hijacking

- Filters
  - Global database about ISP information to authenticate origin
  - Limiting length of prefixes

BGP in Facebook Data Centers

Why BGP in Datacenter?

Before: Uses Layer-2 (Link layer) spanning tree protocol.

- Not scalable as data center grew in size.

Same reasoning as BGP in for inter-domain routing.

- Scalability
- Flexibility of using policy

Network Topology

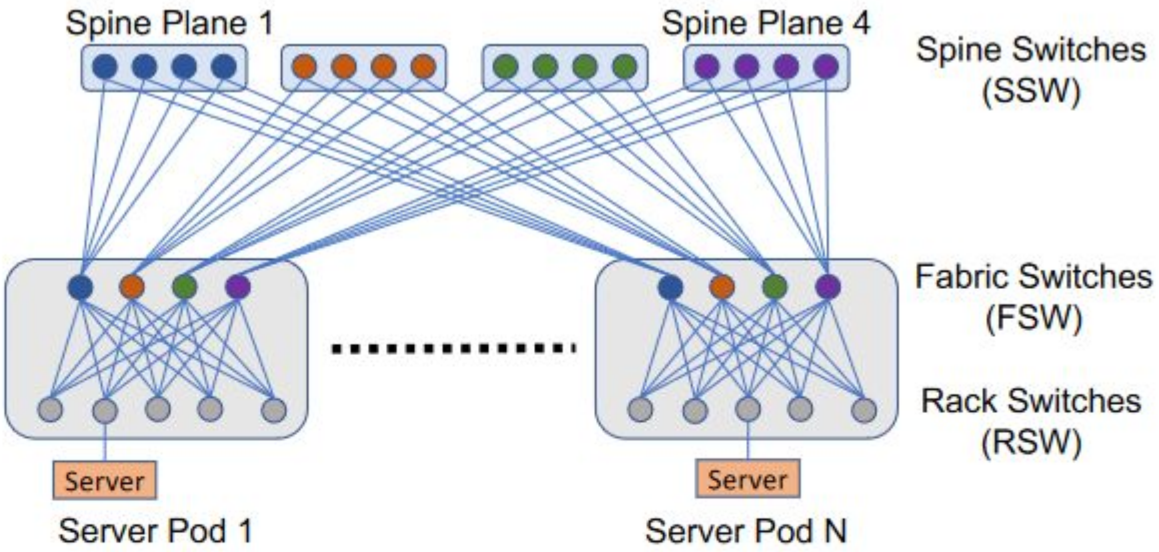


Figure 1: Data Center Fabric Architecture

BGP Cons

Datacenter has lots of failures and maintenance events which can trigger BGP convergence issues. How to provide reliability?

- Uses policy to provide backup paths.
- Does not announce rack-prefixes to outside. (Most reconvergence happens inside pod)

Network Topology

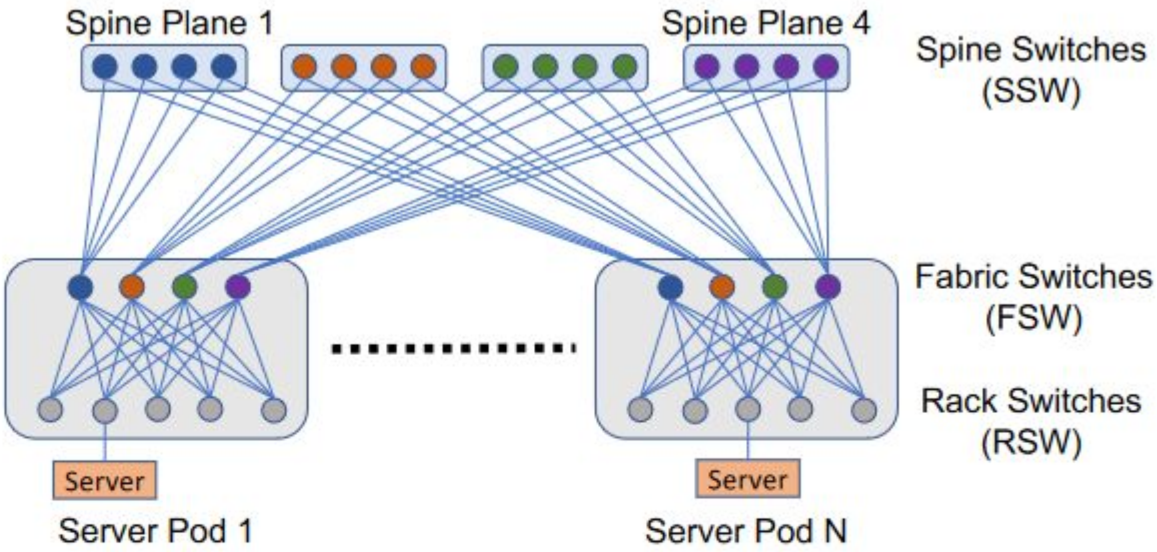


Figure 1: Data Center Fabric Architecture

Uniformity and Simplicity Principle

- Treats every single switch as its own AS.
- All AS-AS relationship are peer.
- BGP Configuration is same across network tier.
- Peering sessions between device tiers (RSW-FSW or FSW-RSW) uses the same features, timers, and parameters. Policy change applies simultaneously to all peers in the group.

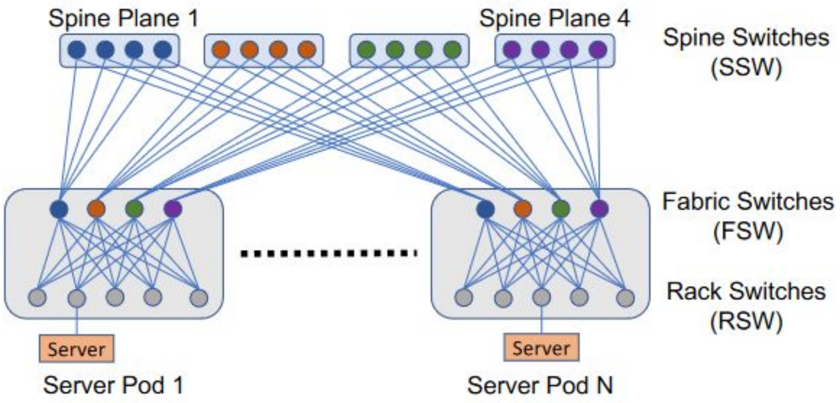


Figure 1: Data Center Fabric Architecture

Uniformity and Simplicity Principle

- All switches in the same spine plane have the same AS number (65001).
- Server pod has a single AS number that is public (65101) (BGP Confederation). Each pod internally has the same numbering with the next pod.
- Re-uses the same exact numbering pattern in all data-center.
- Uses route summarization.

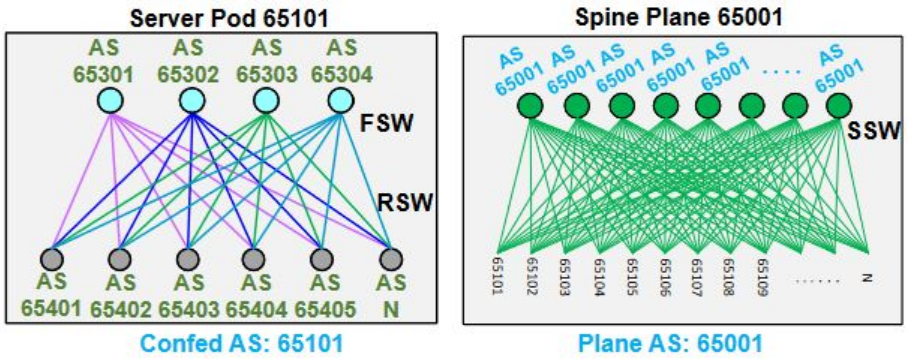
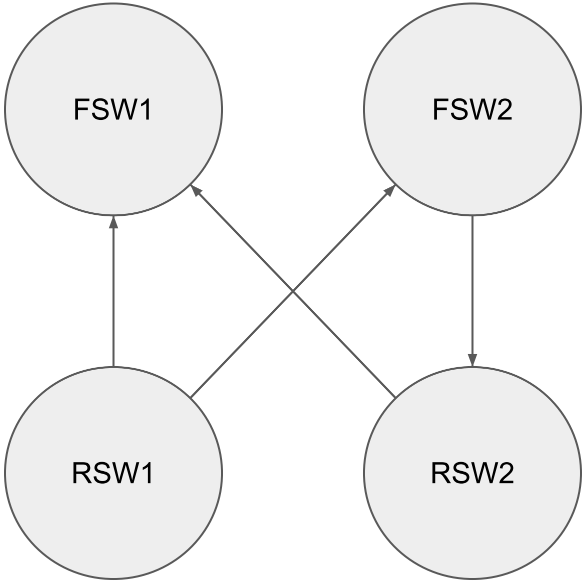


Figure 2: BGP Confederation and AS Numbering scheme for server pods and spine planes in the data center.

Advertisement flow

match: ‘backup_path’
action: add tag
‘completed_backup+path’

match: ‘rack_prefix’
action: add tag ‘backup
path’

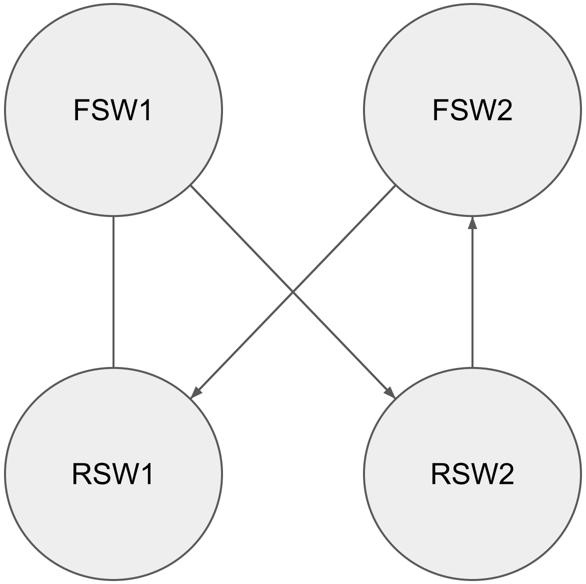


match: ‘backup_path’
action: allow

action: add tag ‘rack_prefix’

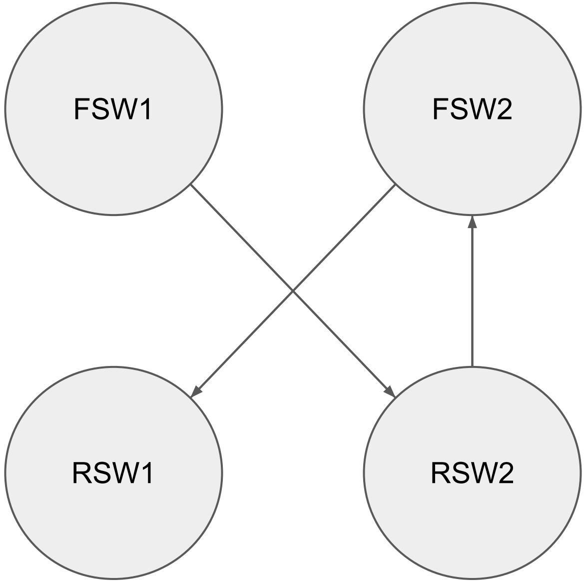
rack_prefix only propagates inside pod.

Traffic flow

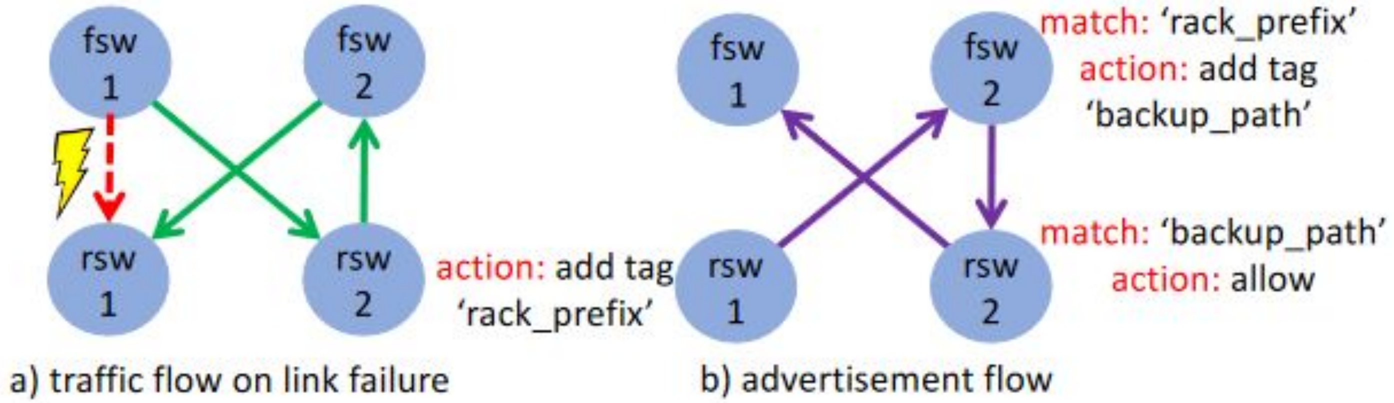


Traffic flow

1. FSW1 Immediately uses backup path.
2. FSW1 does not need to say anything to SSW about the link that is down.
3. FSW will have multiple equal cost backup paths to RSW.



Policies


fsw 1
fsw 2
rsw 1
rsw 2
action: add tag 'rack_prefix'
a) traffic flow on link failure
fsw 1
fsw 2
rsw 1
rsw 2
match: 'rack_prefix'
action: add tag 'backup_path'
match: 'backup_path'
action: allow
b) advertisement flow

BGP vs SDN