facebook nsdi'21

# Running BGP in Data Centers at Scale

**Anubhavnidhi Abhashkumar, Kausik Subramanian, Alexey Andreyev, Hyojeong Kim, Nanda Kishore Salem, Jingyi Yang, Petr Lapukhov, Aditya Akella, Hongyi Zeng**

University of Wisconsin – Madison, Facebook

April 2021

# Motivation

# Facebook Data Center Growth

# Facebook Data Center Growth

Tremendous user growth

500 million in 2010

↓

2.5 billion in 2020

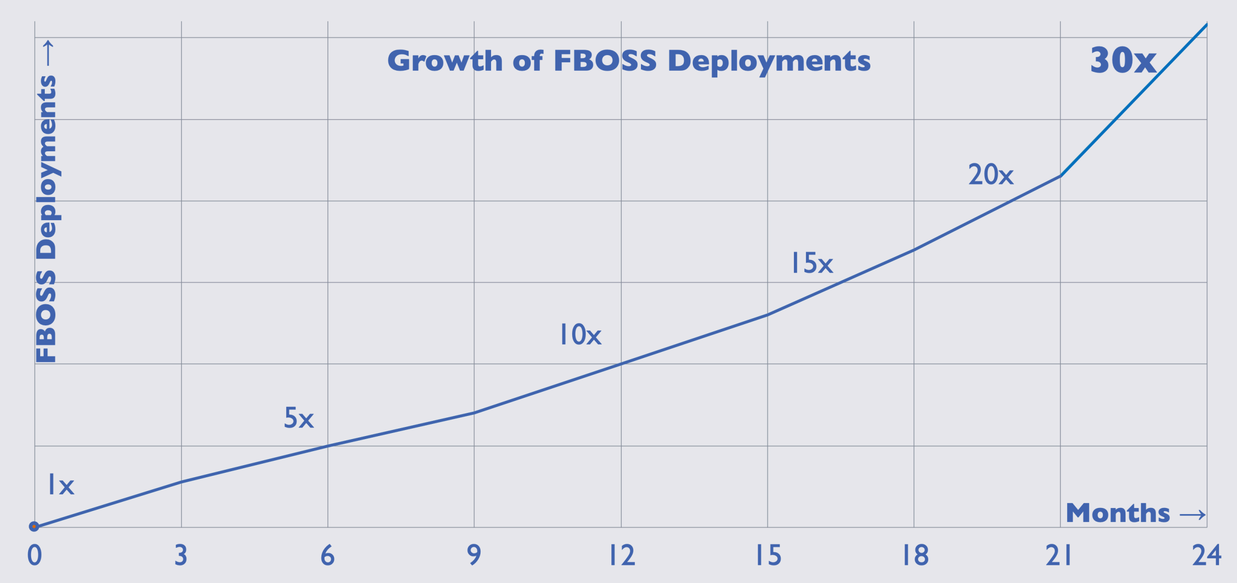
# Facebook Data Center Growth

Tremendous user growth

500 million in 2010

2.5 billion in 2020

Our data center networks have grown to meet with user demand




Source: FBOSS: Building Switch Software at Scale, SIGCOMM 18

# Data Center Routing Requirements

# Data Center Routing Requirements

Scalable design to support quick growth

# Data Center Routing Requirements

Scalable design to support quick growth

Reliability, expecting failures to be common

# Data Center Routing Requirements

Scalable design to support quick growth

Reliability, expecting failures to be common

Operational efficiency to innovate rapidly while providing production reliability

# BGP as the DC routing protocol

# BGP as the DC routing protocol

BGP (Border Gateway Protocol) is chosen for scalability, extensive policy control, and long-term stability.

# BGP as the DC routing protocol

BGP (Border Gateway Protocol) is chosen for scalability, extensive policy control, and long-term stability.

BGP on vendor devices helped grow our networks rapidly initially, compared to developing an SDN routing solution from scratch.

# BGP Challenges

# BGP Challenges

Long convergence, routing instabilities and misconfigurations in the Internet

# BGP Challenges

Long convergence, routing instabilities and misconfigurations in the Internet

Problematic in the DC as well due to frequent failures and maintenance events

# BGP Challenges

Long convergence, routing instabilities and misconfigurations in the Internet

Problematic in the DC as well due to frequent failures and maintenance events

Vendor BGP software not able to support evolving routing requirements fast

# Our Solution

# Our Solution

Scalable BGP routing design adhering to the principles of configuration uniformity and operational simplicity

# Our Solution

Scalable BGP routing design adhering to the principles of configuration uniformity and operational simplicity

Routing policies to achieve reliability of the network under failures

# Our Solution

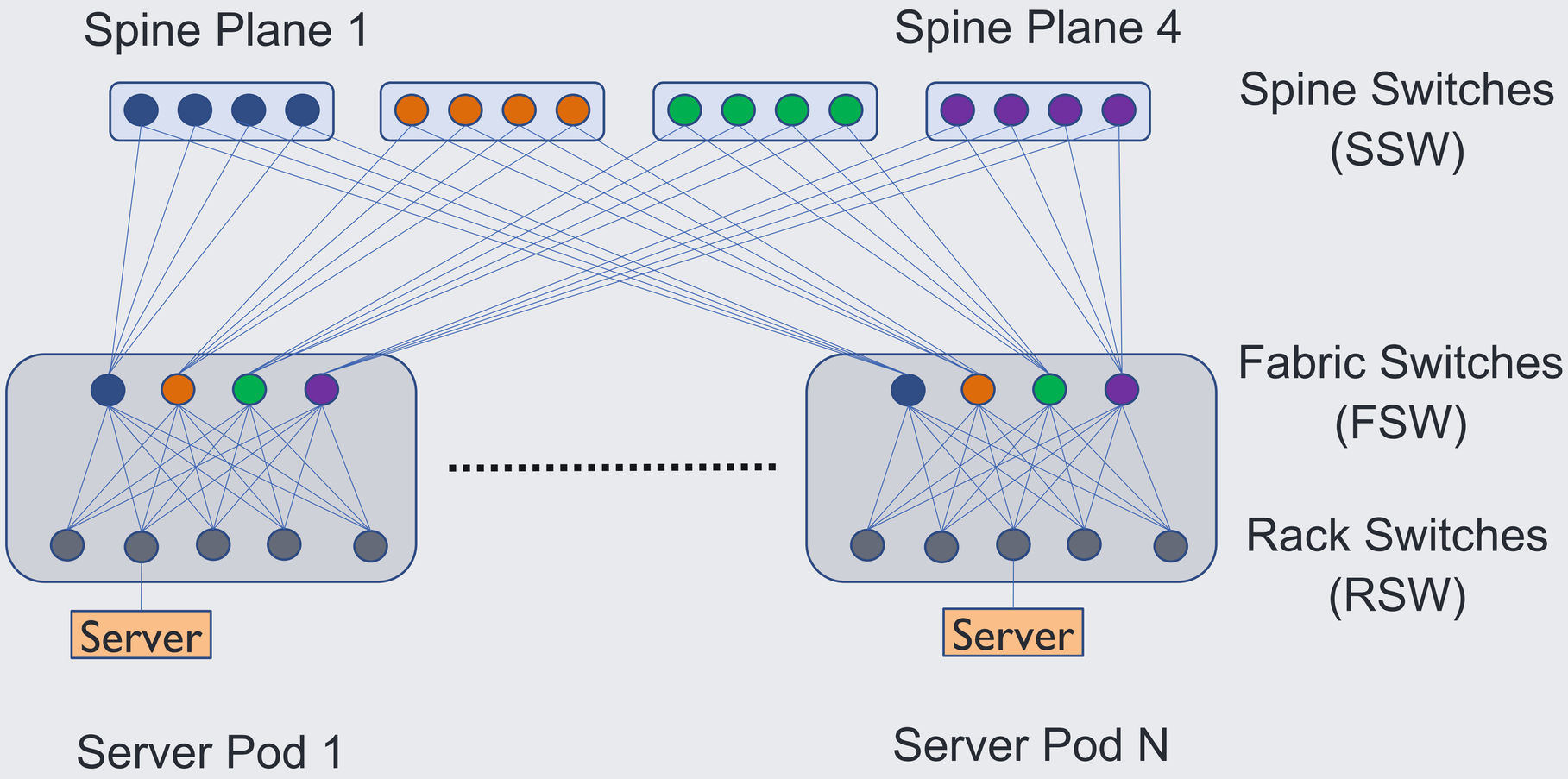
Scalable BGP routing design adhering to the principles of configuration uniformity and operational simplicity

Routing policies to achieve reliability of the network under failures

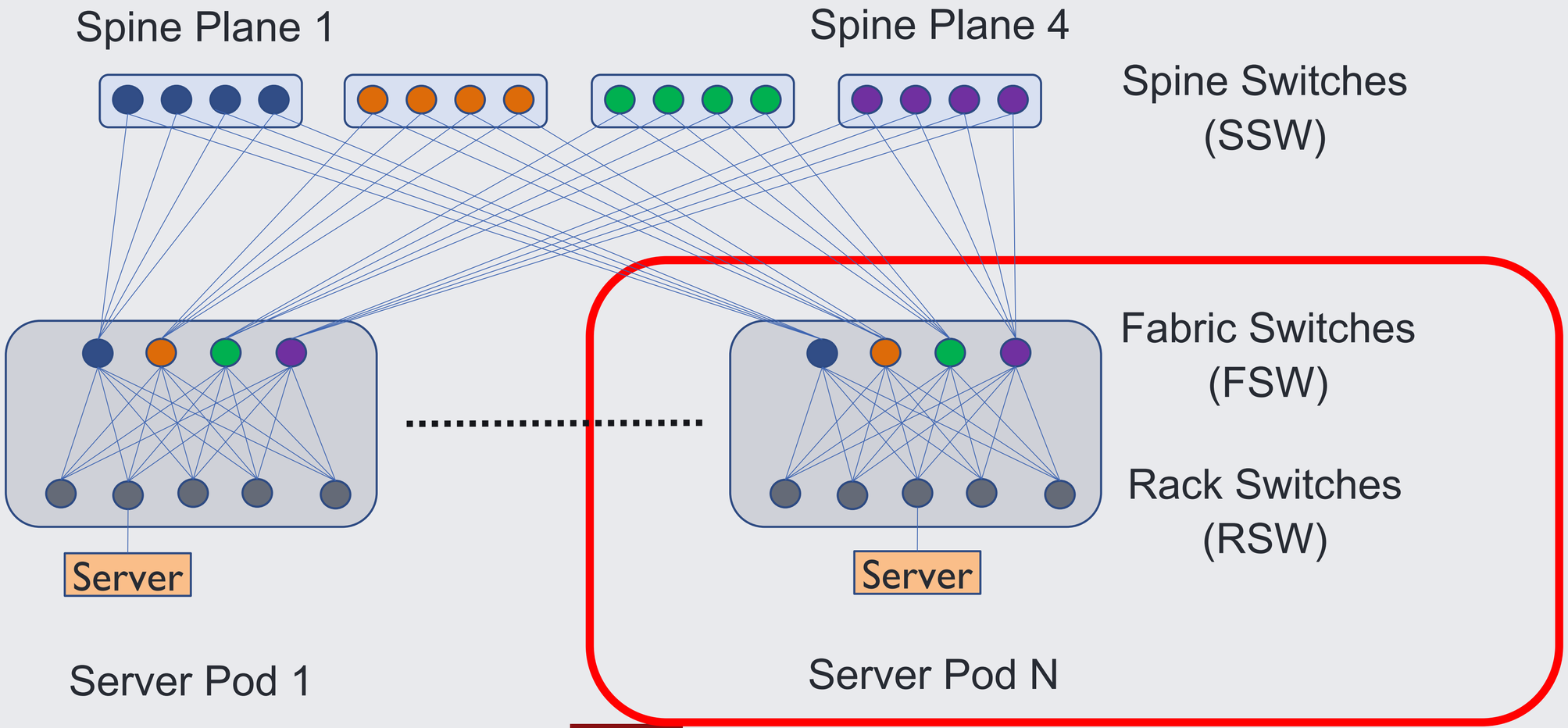
In-house BGP agent coupled with automated testing and incremental deployment framework to achieve operational efficiency

# DC Routing Design with BGP

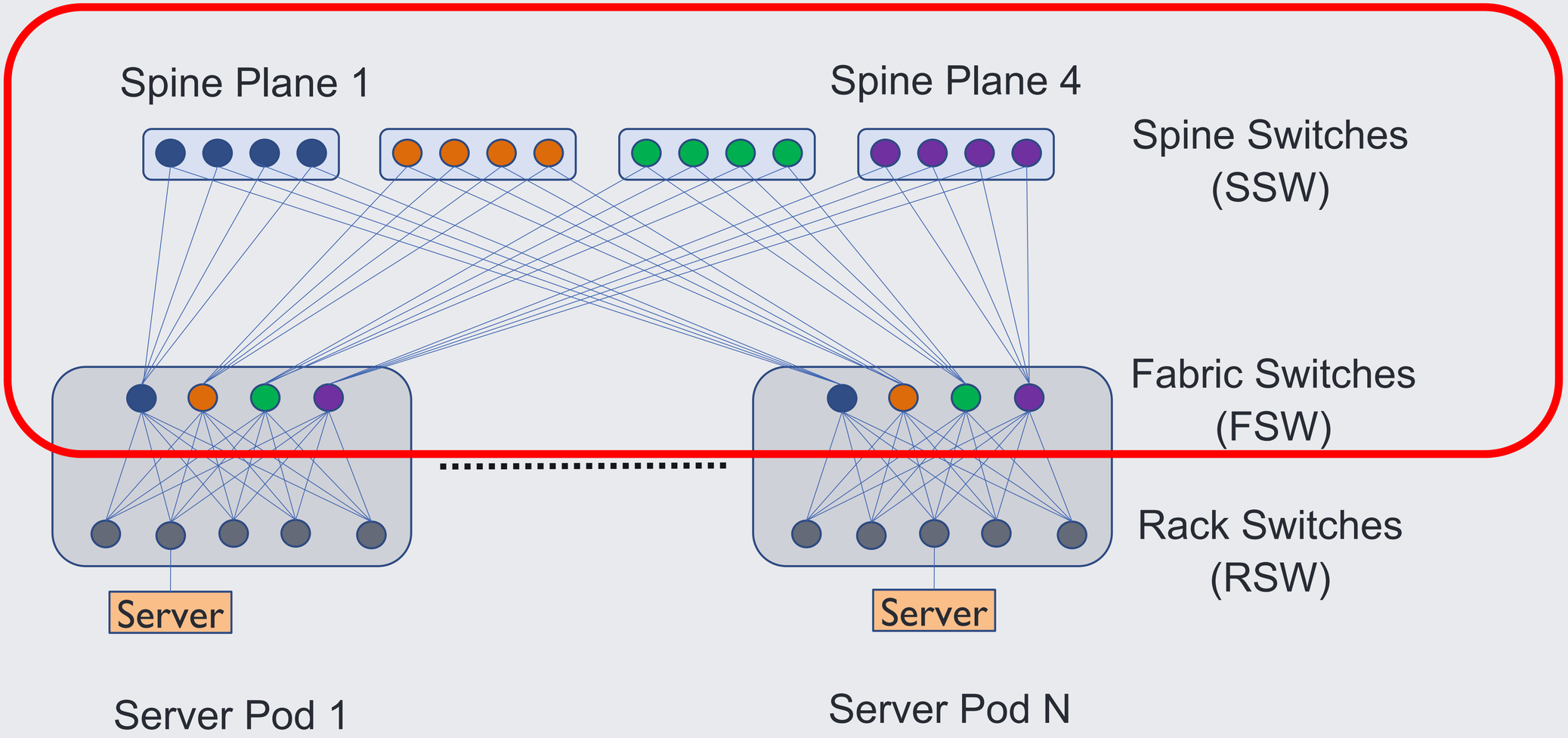
# Scalable Network Topology



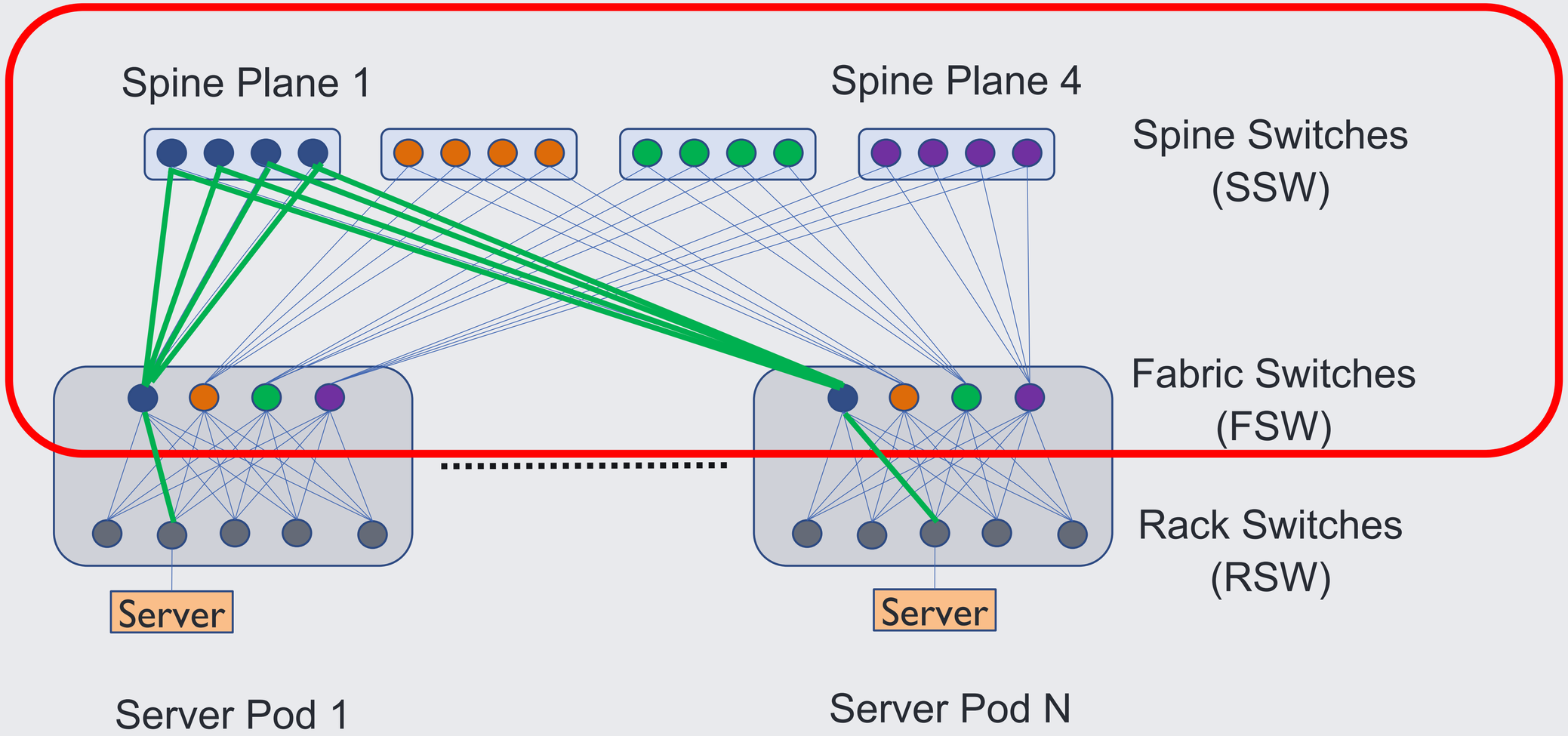
# Scalable Network Topology



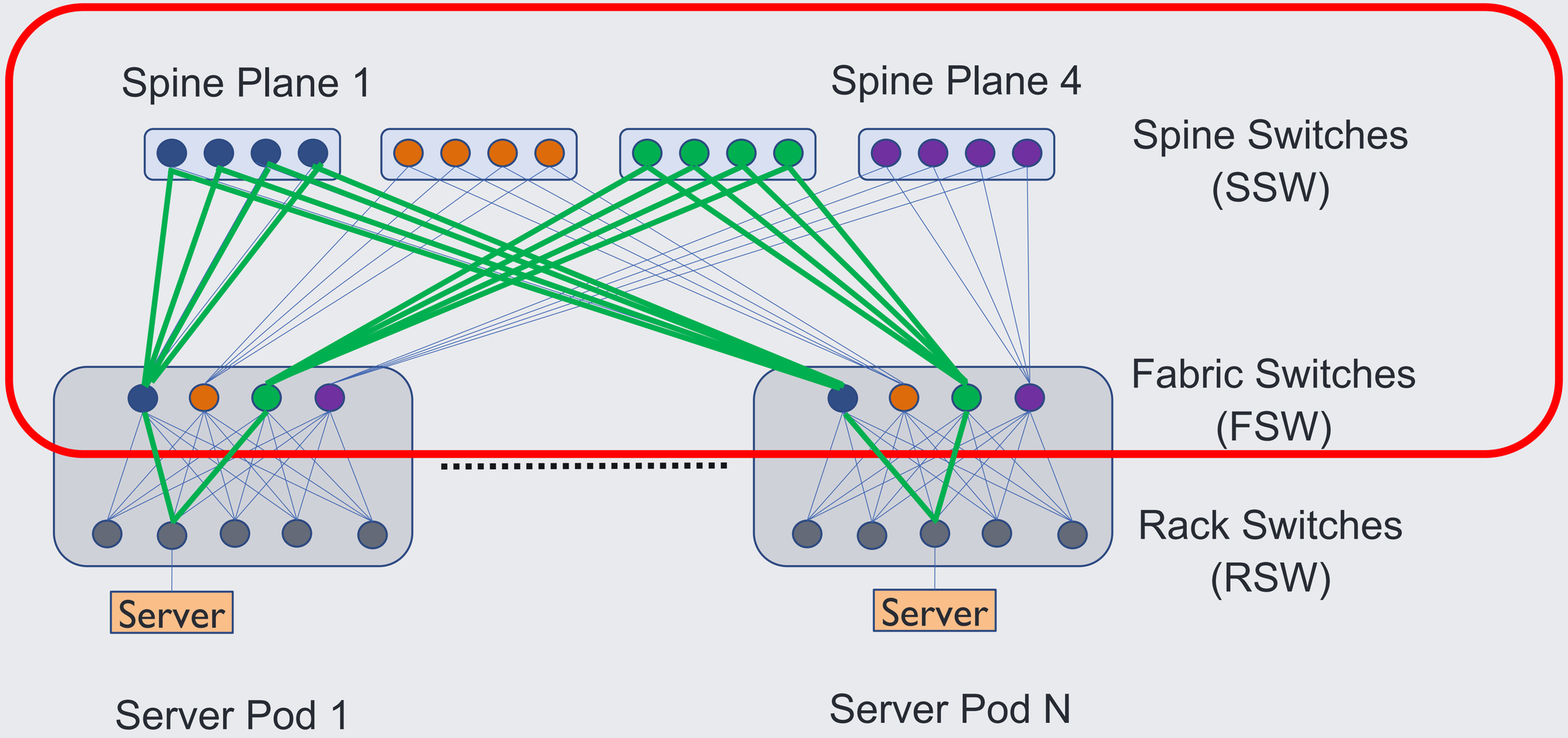
# Scalable Network Topology



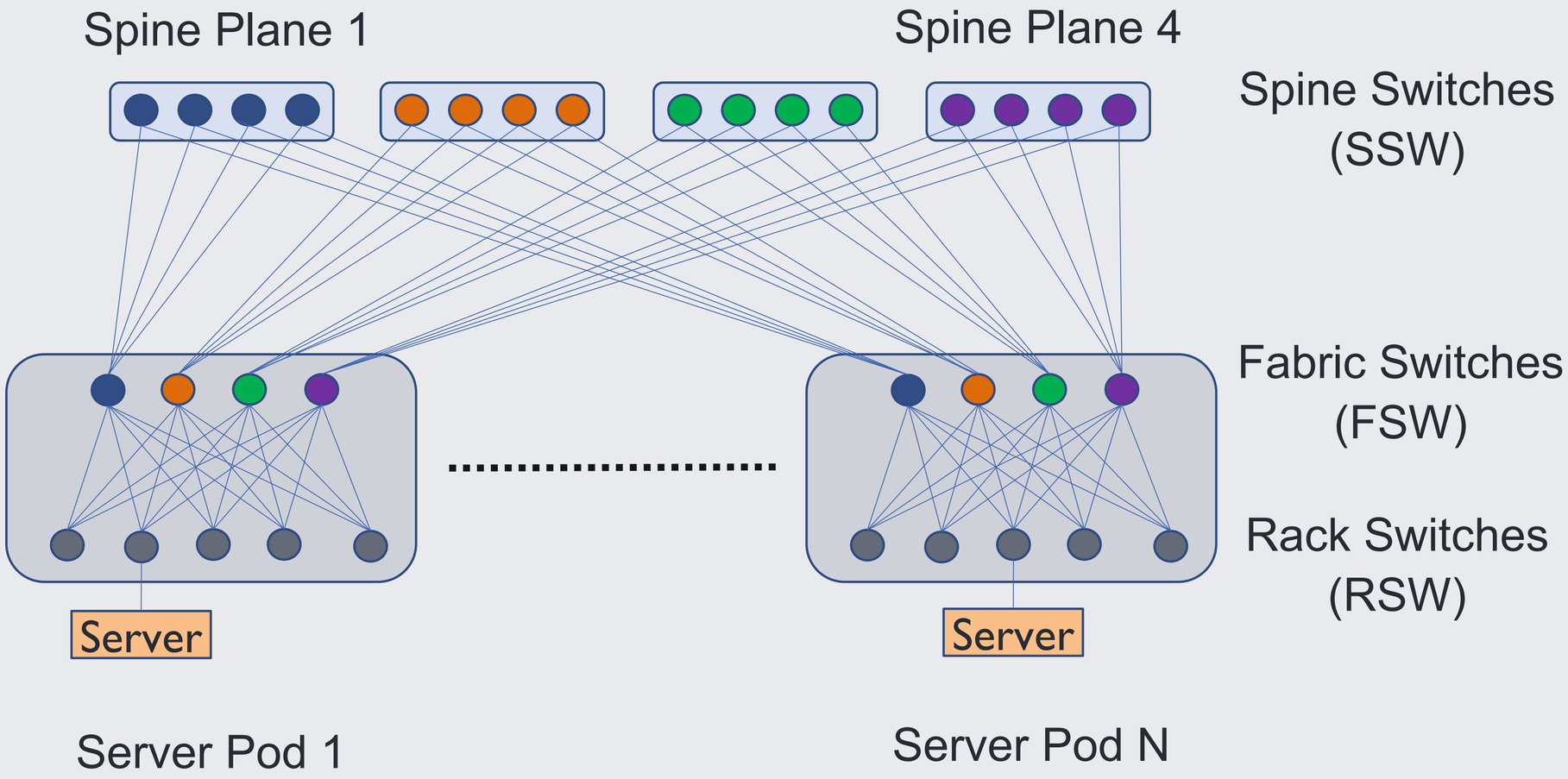
# Scalable Network Topology



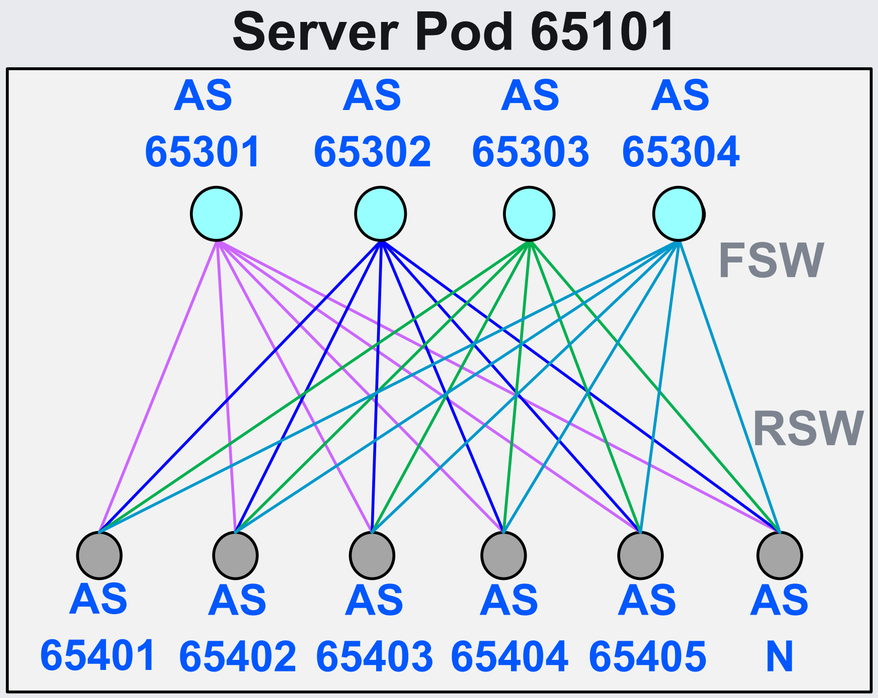
# Scalable Network Topology

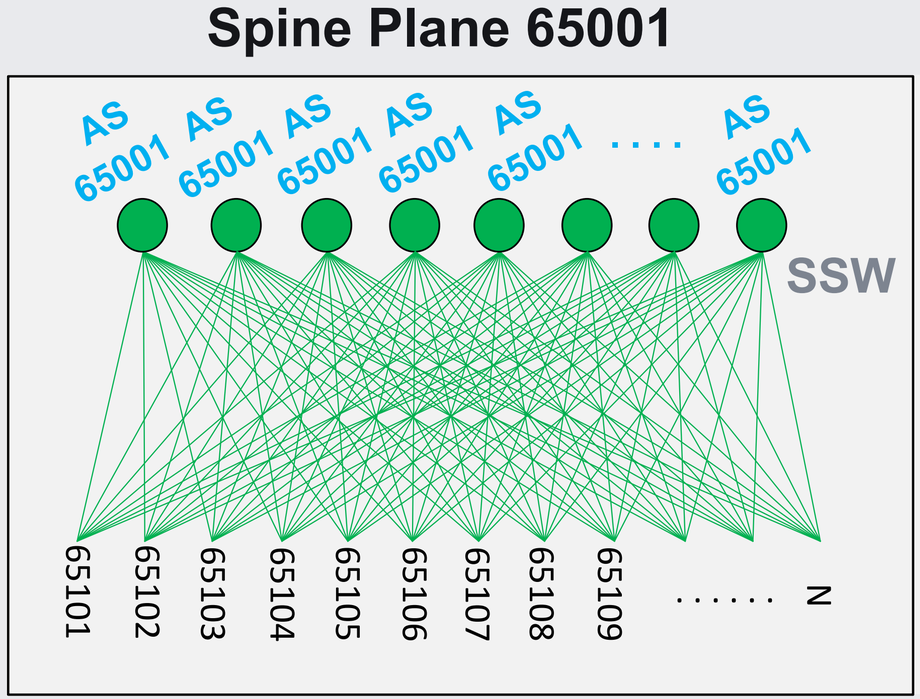


# Scalable Network Topology



# Reusable BGP ASN Allocation



# Route Summarization

## Control route scale

# Route Summarization

## Control route scale

- Minimize HW FIB size requirements

# Route Summarization

## Control route scale

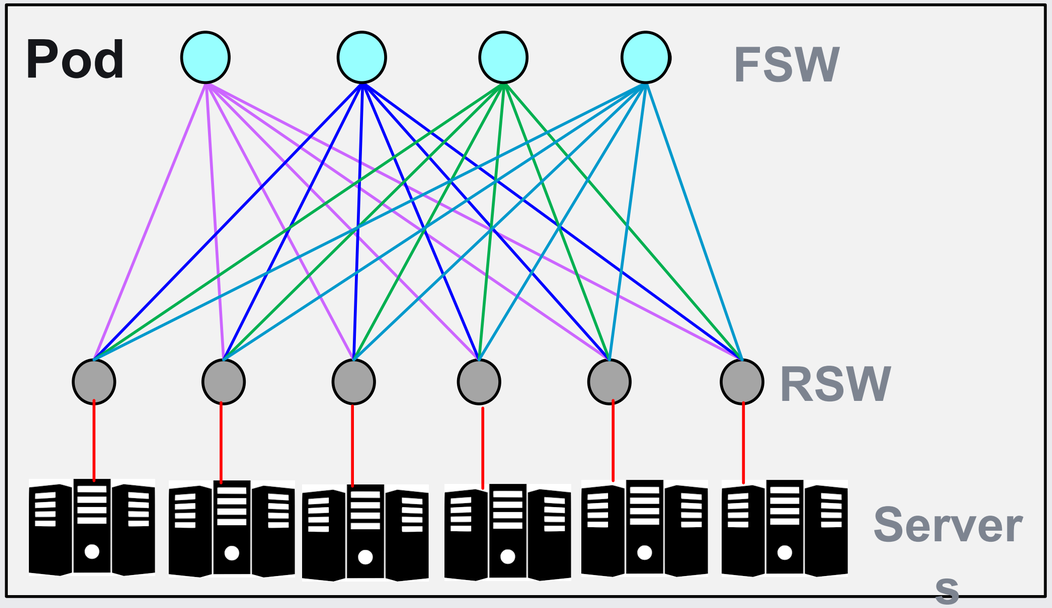
- Minimize HW FIB size requirements
- Reduce control plane processing

# Route Summarization

## Control route scale

- Minimize HW FIB size requirements
- Reduce control plane processing

Example of Production routes, reflecting topology hierarchy

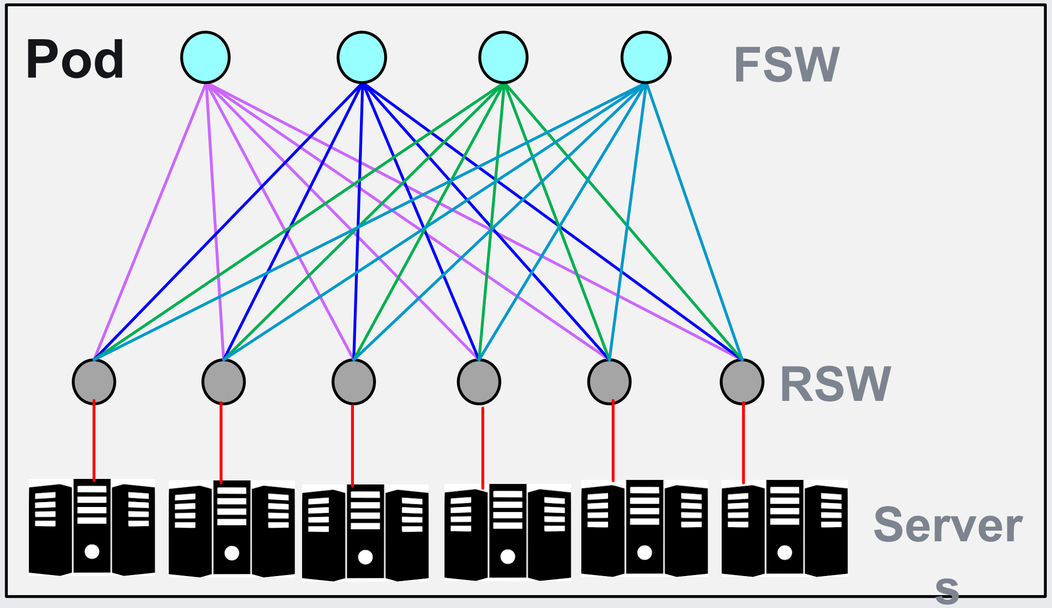


# Route Summarization

## Control route scale

- Minimize HW FIB size requirements
- Reduce control plane processing

Example of Production routes, reflecting topology hierarchy



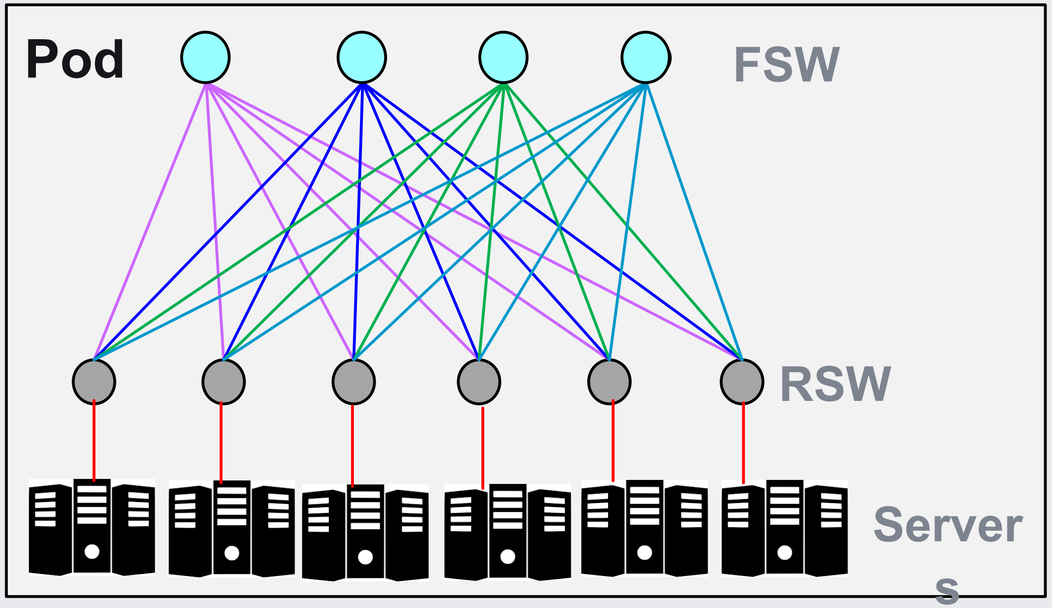

Server IP addresses
face:b00c::1:0/128

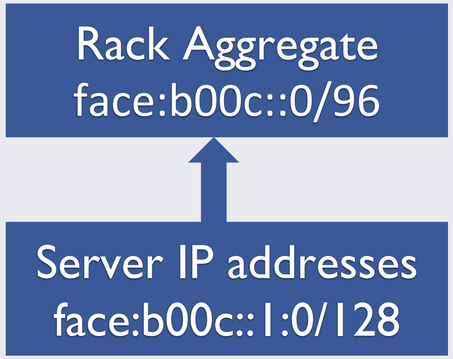
# Route Summarization

## Control route scale

- Minimize HW FIB size requirements
- Reduce control plane processing

Example of Production routes, reflecting topology hierarchy



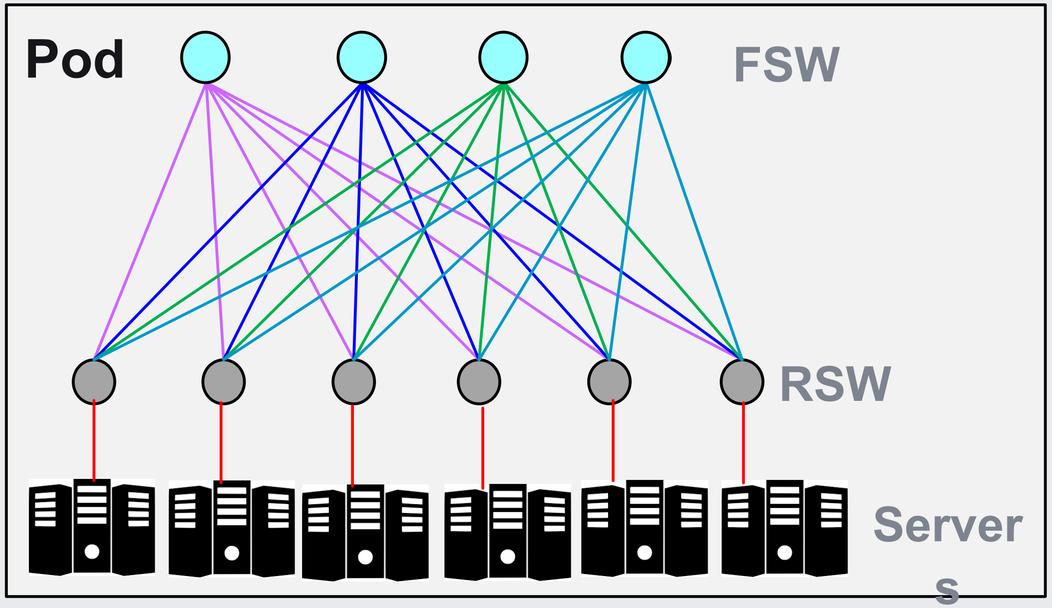



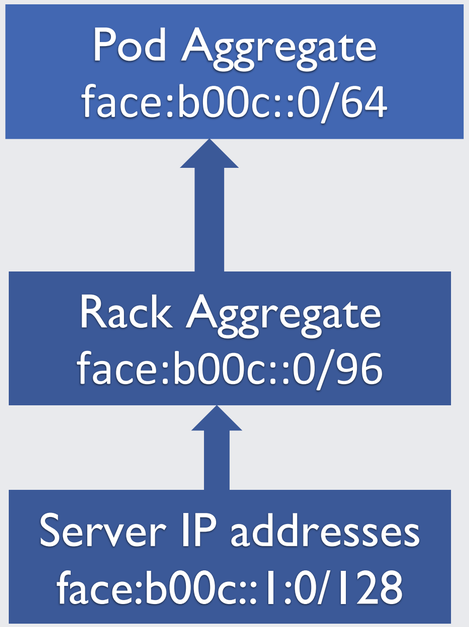
# Route Summarization

## Control route scale

- Minimize HW FIB size requirements
- Reduce control plane processing

Example of Production routes, reflecting topology hierarchy



# DC Policy Goals

In the Internet, BGP polices implement business relationships.

# DC Policy Goals

In the Internet, BGP polices implement business relationships.

| Goal | Description |
| --- | --- |
| Reliability | Enforce route propagation scopes.<br>Predefine backup paths for failure. |

# DC Policy Goals

In the Internet, BGP polices implement business relationships.

| Goal | Description |
| --- | --- |
| Reliability | Enforce route propagation scopes.<br>Predefine backup paths for failure. |
| Scalability | Enforce route summarization<br>Avoid backup path explosion |

# DC Policy Goals

In the Internet, BGP polices implement business relationships.

| Goal | Description |
|---|---|
| Reliability | Enforce route propagation scopes.<br>Predefine backup paths for failure. |
| Scalability | Enforce route summarization<br>Avoid backup path explosion |
| Maintainability | Isolate/remediate problematic nodes<br>without disrupting traffic |

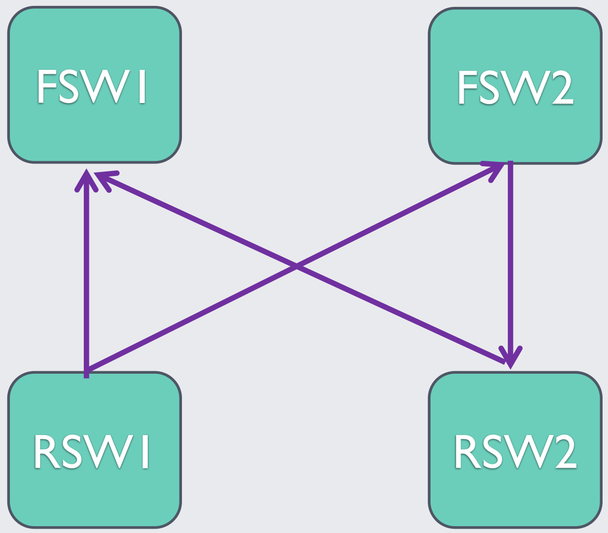
# DC Policy Goals

In the Internet, BGP polices implement business relationships.

| Goal | Description |
|---|---|
| Reliability | Enforce route propagation scopes.<br>Predefine backup paths for failure. |
| Scalability | Enforce route summarization<br>Avoid backup path explosion |
| Maintainability | Isolate/remediate problematic nodes without disrupting traffic |
| Service reachability | Avoid service disruptions when instances of services are added/removed/migrated |

# Reliability

## Predefine backup paths for link failure

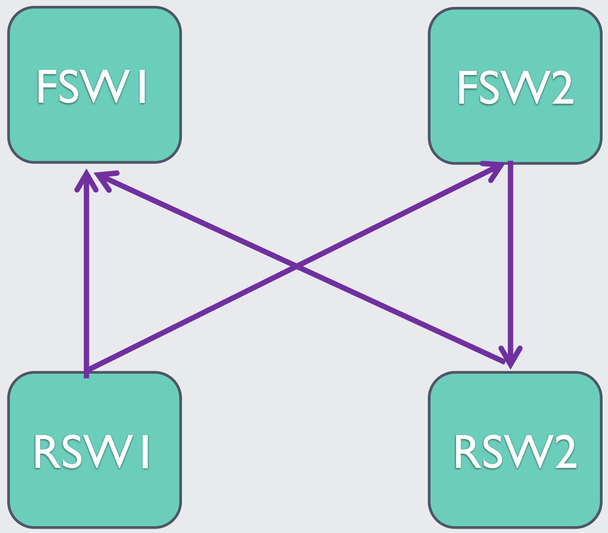



During BGP Route Convergence

# Reliability

## Predefine backup paths for link failure

RSW1's rack_prefix:
-> RSW1 (best)



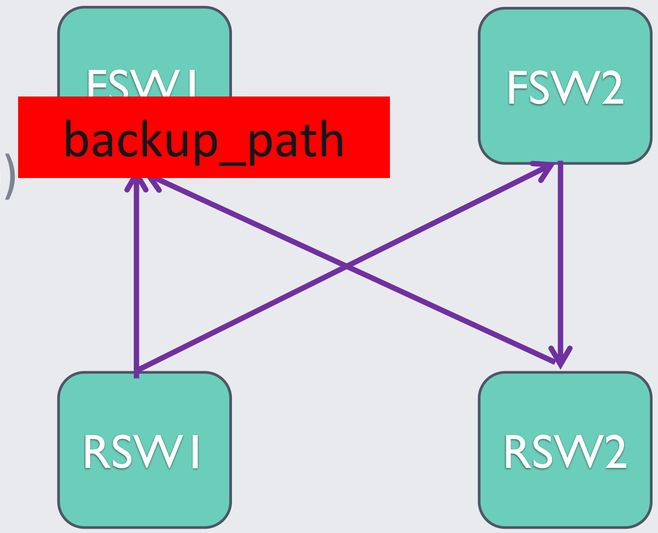

action: add tag
'rack_prefix'

During BGP Route Convergence

# Reliability

## Predefine backup paths for link failure

RSW1's rack_prefix:
-> RSW1 (best)
-> RSW2--FSW2--RSW1 (backup)




**match:** 'rack_prefix'
**action:** add tag 'backup_path'

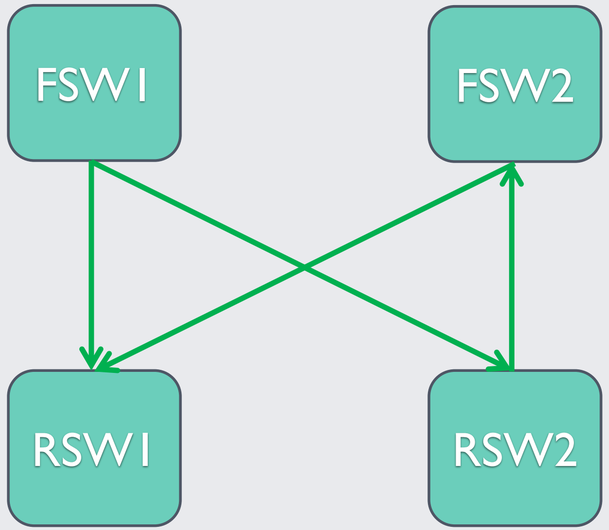
**action:** add tag 'rack_prefix'

**match:** 'backup_path'
action: allow

During BGP Route Convergence

# Reliability

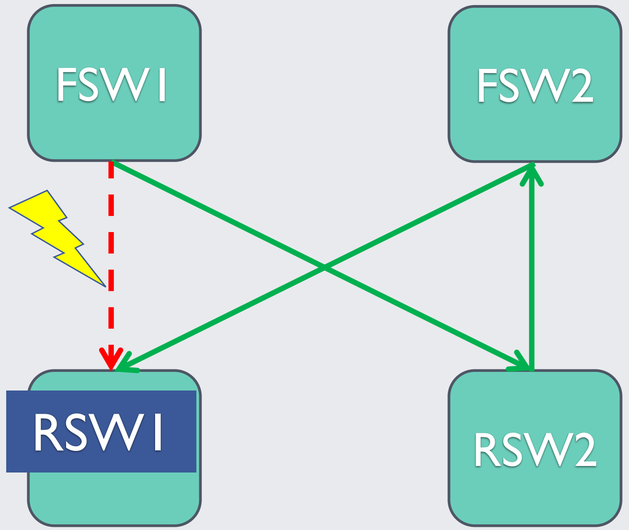
## Predefine backup paths for link failure




Traffic flow on link failure

# Reliability

## Predefine backup paths for link failure

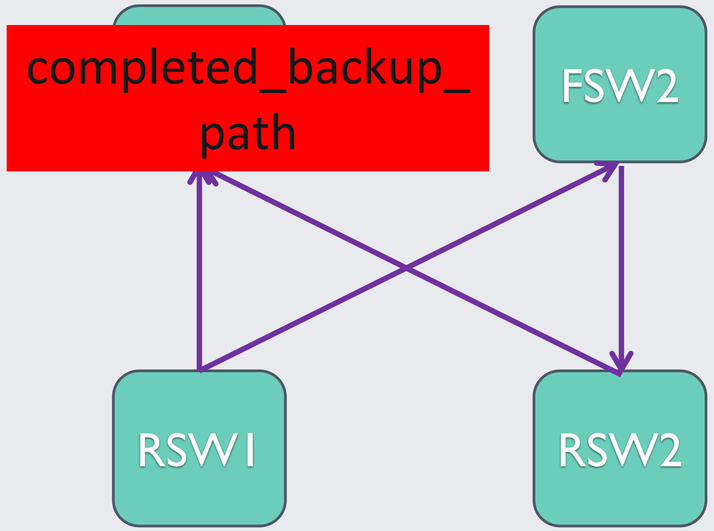



Traffic flow on link failure

# Scalability

## Enforce route summarization and avoid backup path explosion

match: ‘backup_path’
action: add tag ‘completed_backup_path’

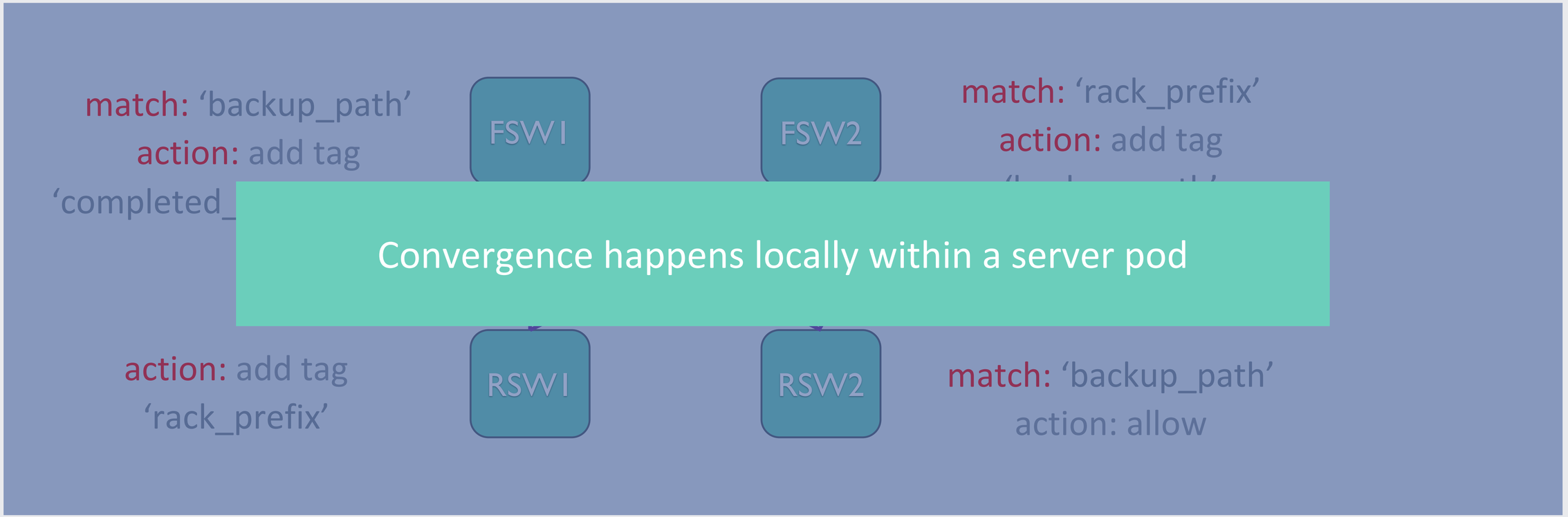



Routes with tag ‘completed_backup_path’ won’t be propagated to other devices.

During BGP Route Convergence

# Scalability

## Enforce route summarization and avoid backup path explosion

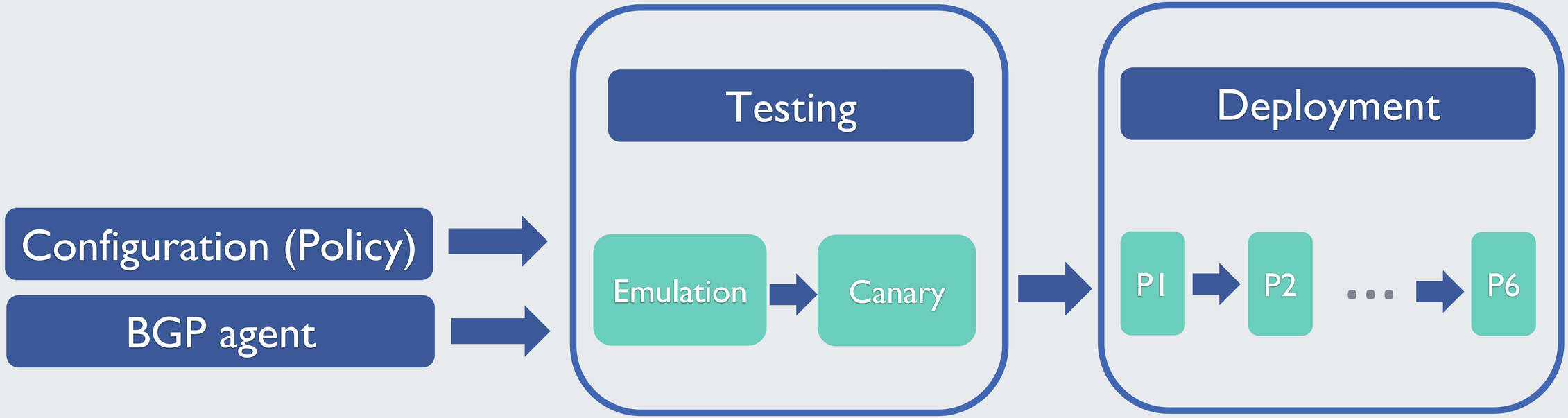



During BGP Route Convergence

# Operational Experience

# Built Operational Pipeline

Support BGP like any other SW, enabling fast incremental updates!



# Conclusion

# Conclusion

Scalable BGP-based routing design: Reusable ASN allocation and route summarization coupled with our scalable topology design

Reliable under failures: Routing policies such as predefined backup path

Operationally Efficient: Automated testing and incremental deployment framework for our in-house BGP implementation

facebook