RIFT-Python Open Source Implementation
Status Update

Version 1, 29-July-2020

RIFT Working Group, IETF 108, Virtual Meeting

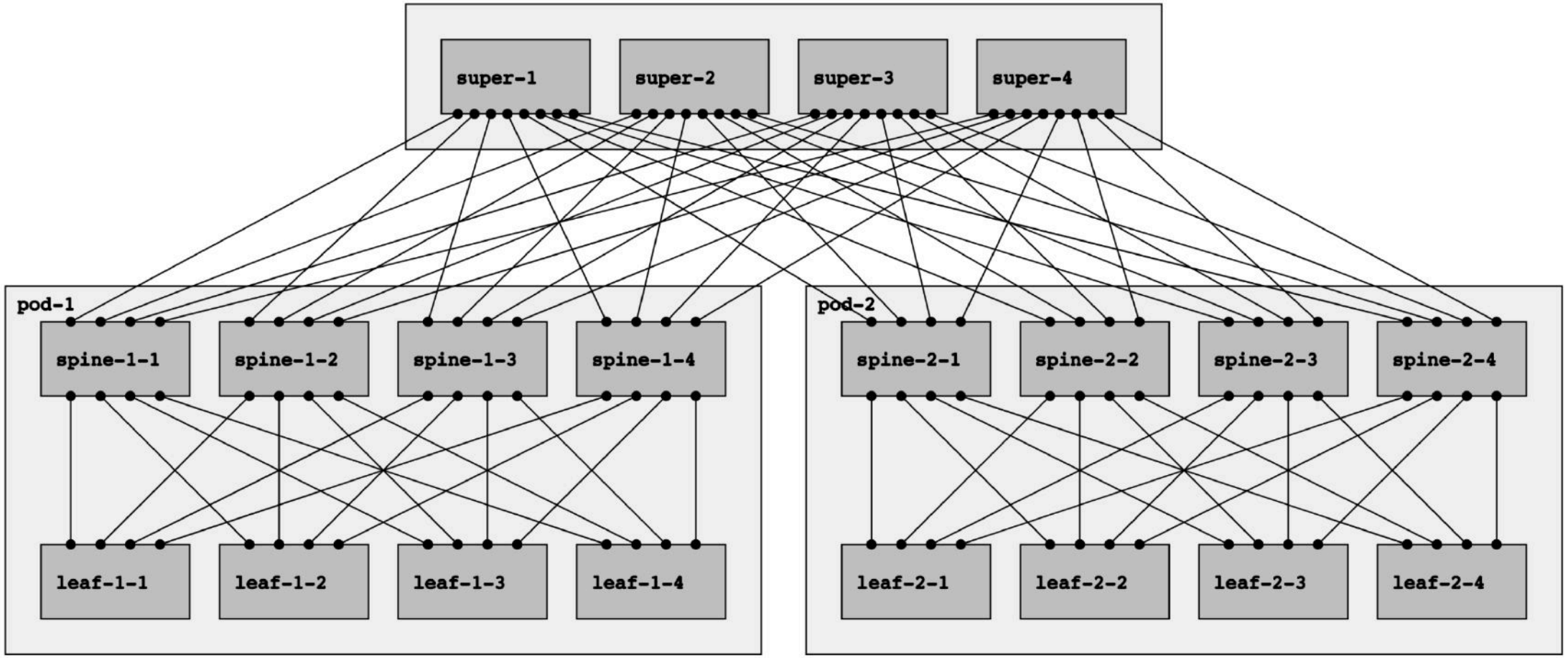
Bruno Rijsman, brunorijsman@gmail.com

New since IETF 105

- Multi-plane with east-west inter-plane loops
- Negative disaggregation
  Implemented by Mariano Scazzariello and Tommaso Caiazzi from Roma University
- Parallel links
- Fabric bandwidth balancing
- Performance monitoring

Multi-plane
with east-west inter-plane loops

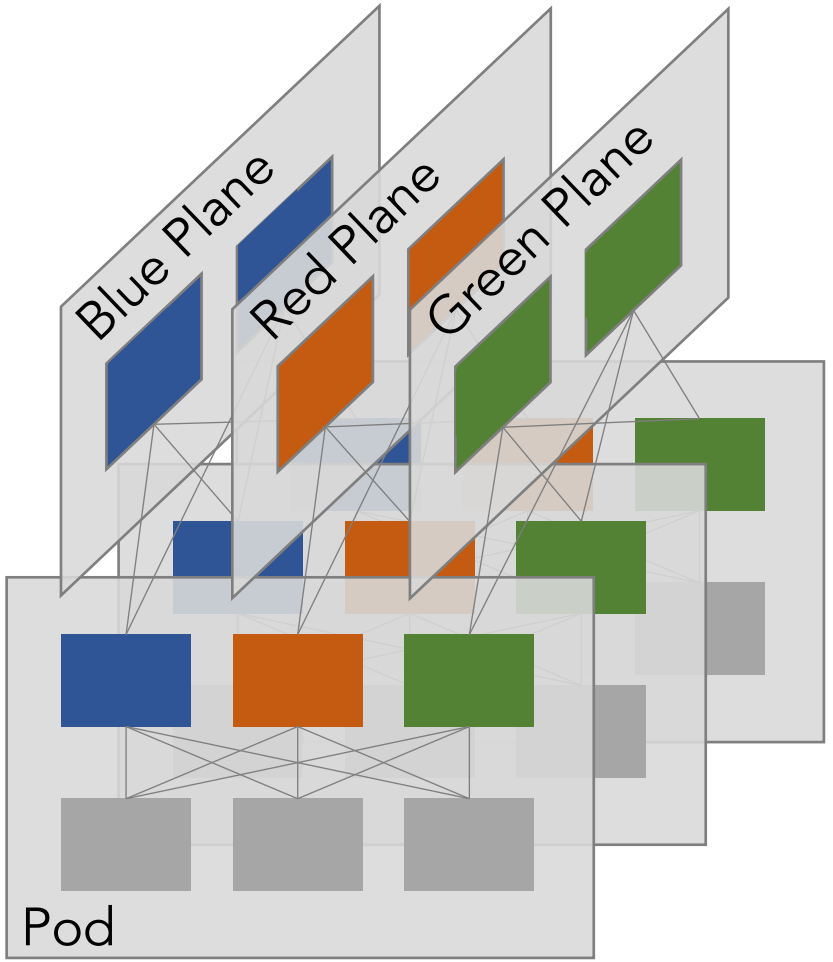
Single-plane



Every super-spine is connected to every spine

In large fabric the super-spines will run out of ports

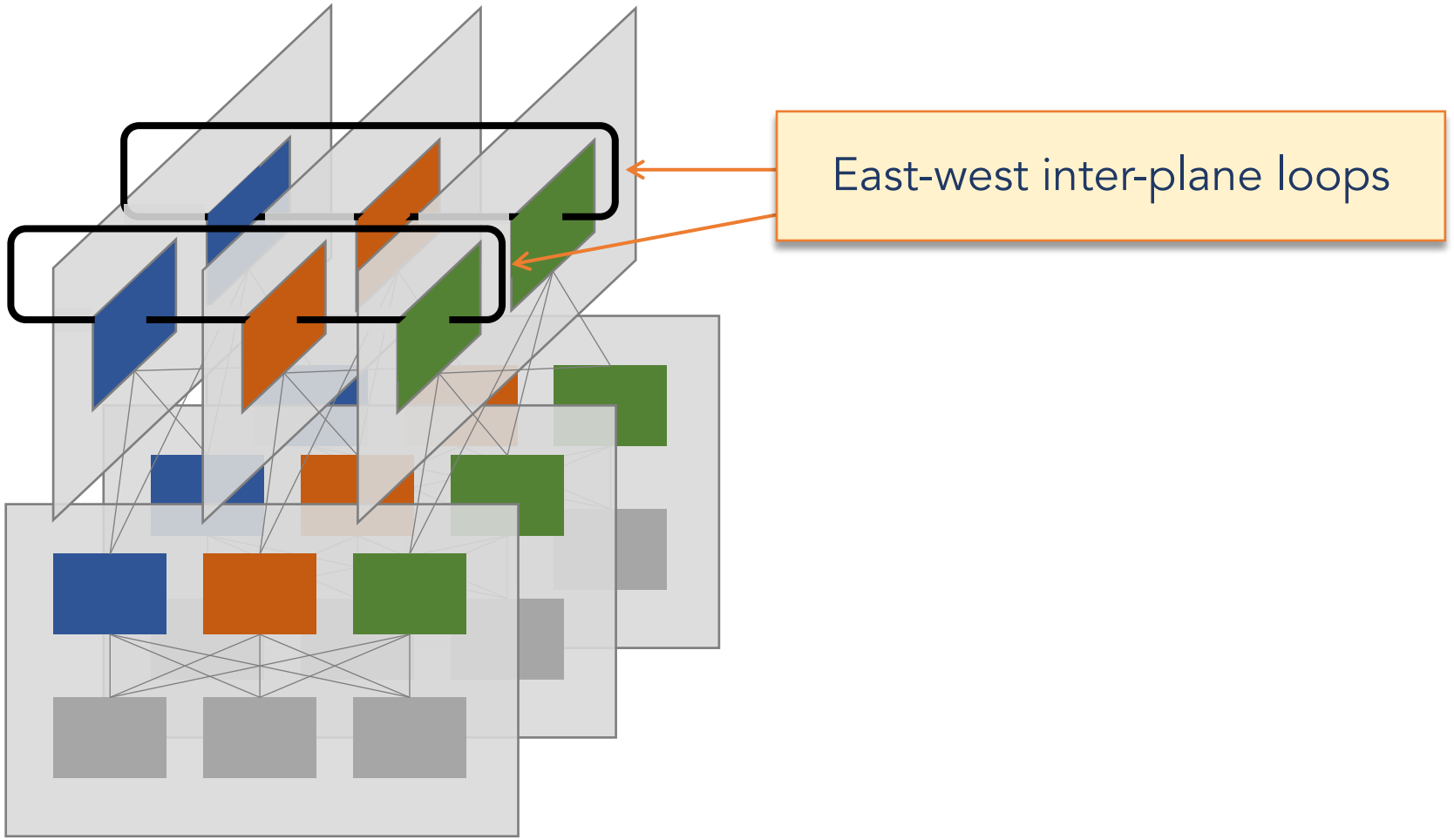
Multi-plane



Each super-spine connects to a subset of spines in each pod

Use different “planes” to connect the pod.

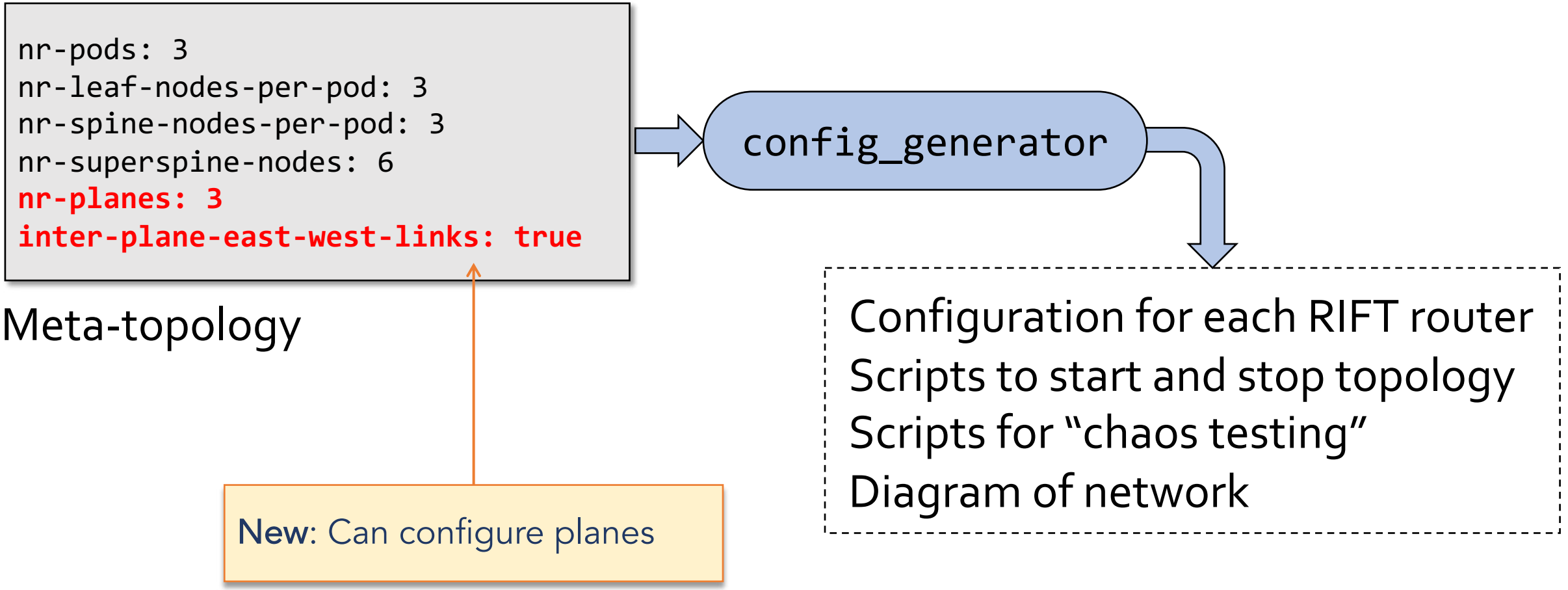
East-west inter-plane loops



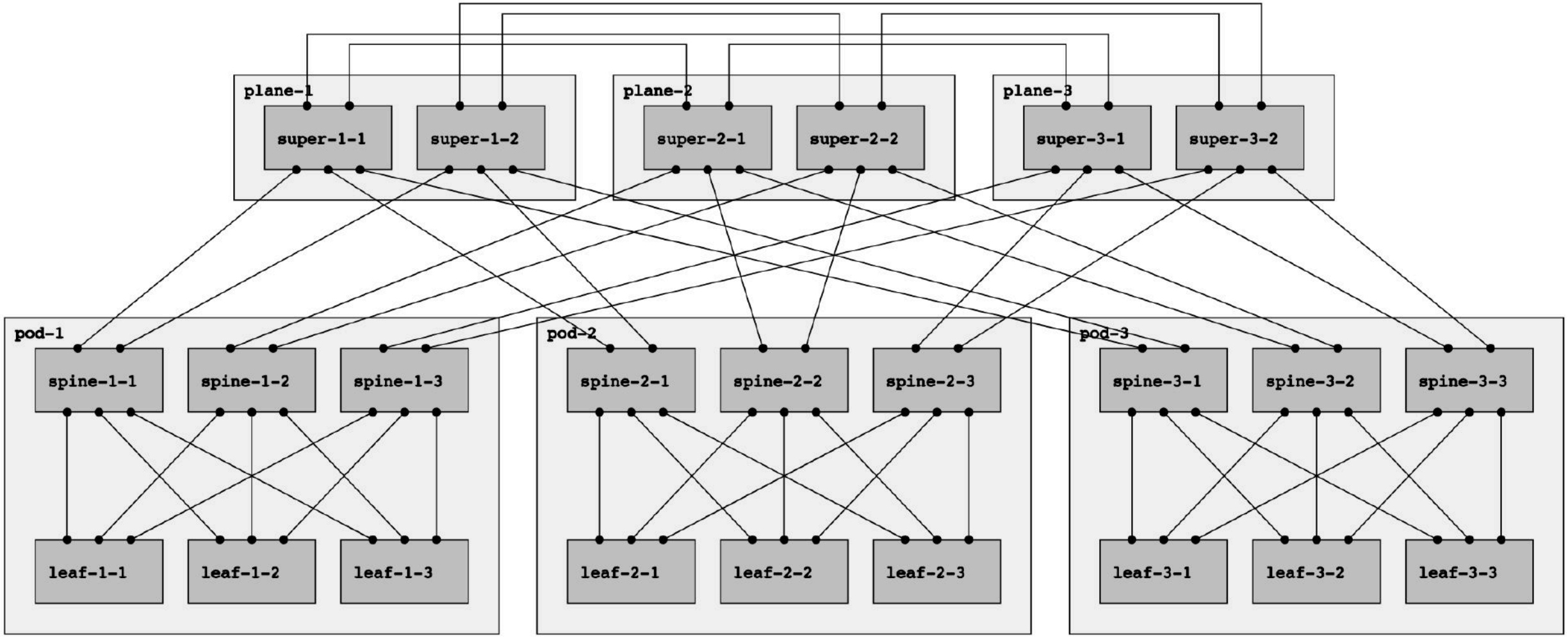
East-west inter-plane links are *only* used for control-plane traffic.

Not used for user data-plane traffic. They can be low-speed links.

Config generator: multi-plane



Example generated multi-plane topology



Negative disaggregation

Positive vs negative disaggregation

- North-bound default route only works if there are no failures.
- RIFT uses disaggregation to route around failures
- Positive disaggregation works for most failures (see slides IETF-105)
- Negative disaggregation is needed in multi-plane topologies.

Concept of a negative prefix advertisement

Positive prefix advertisement:

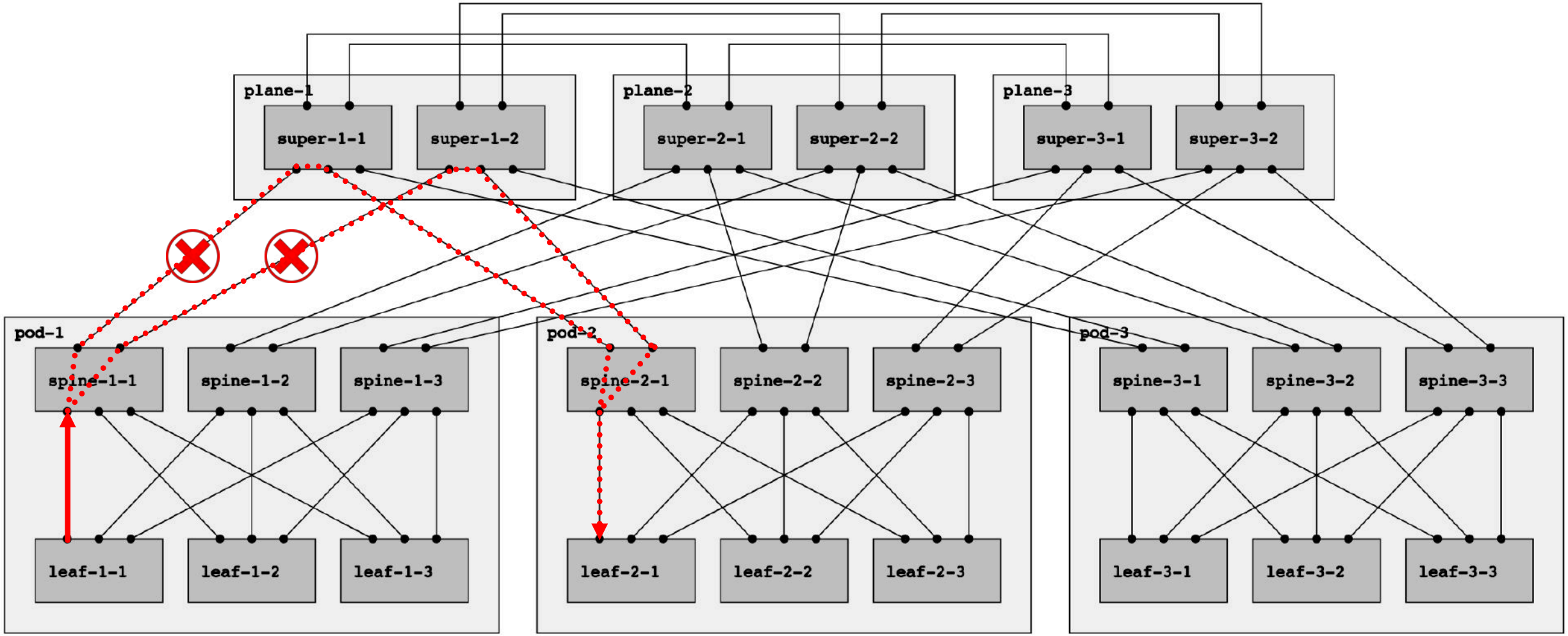
- Has always existed
- "Please send traffic for prefix to me": **attract** traffic for prefix
- Prefer most specific advertisement
- Hardware supports Longest Prefix Match (LPM)

Negative prefix advertisement:

- New concept in RIFT
- "Please **don't** send traffic for prefix to me": **repel** traffic for prefix
- Prefer any other (positive) route, even if it is less specific
- Control-plane concept only
- Negative RIB next-hops translated to positive FIB next-hops

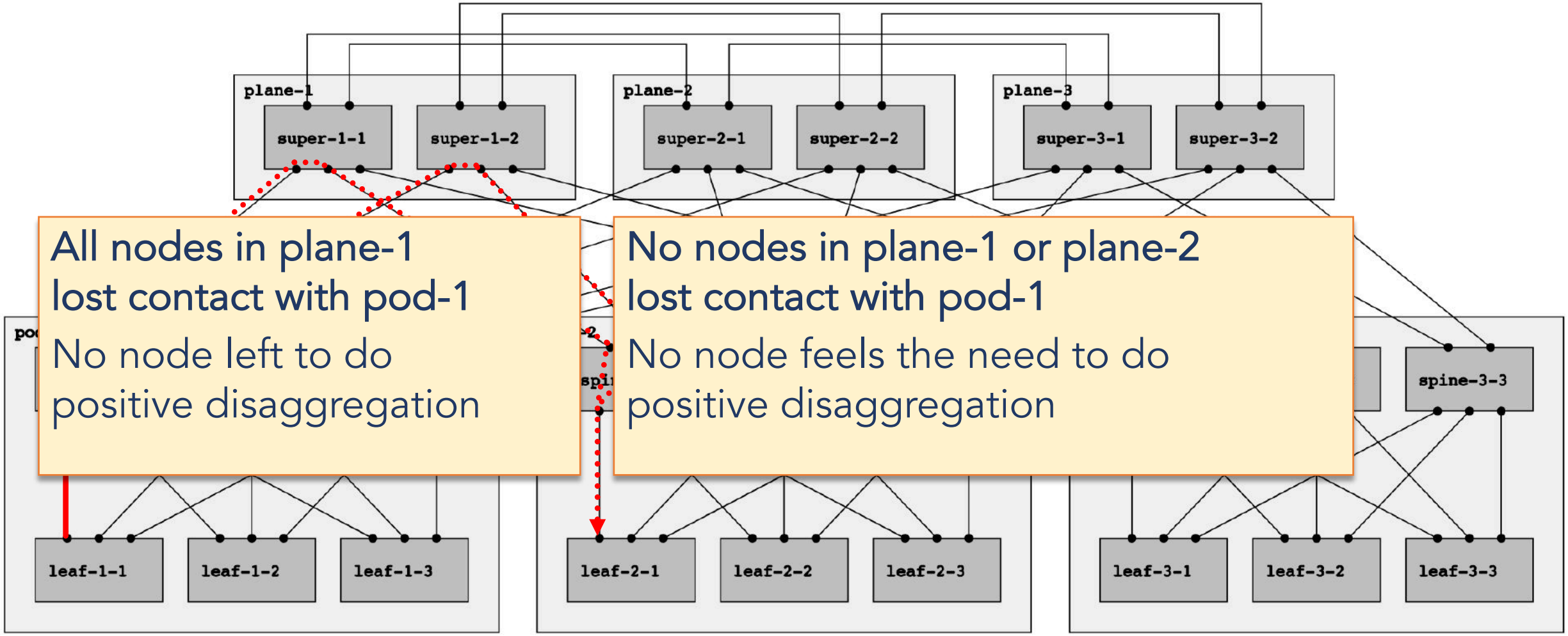
The need for negative disaggregation

Pod-1 is disconnected from plane-1



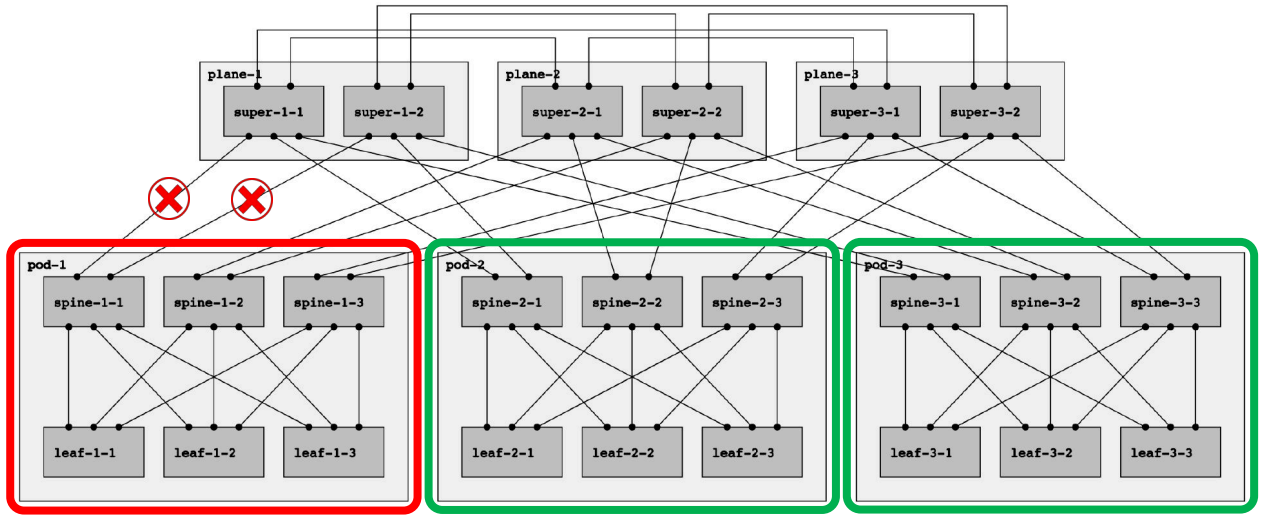
If leaf-1-1 tries to reach leaf-2-1 via spine-1-1 (plane-1) traffic is black-holed

Positive disaggregation does not fix this



If leaf-1-1 tries to reach leaf-2-1 via spine-1-1 (plane-1) traffic is black-holed

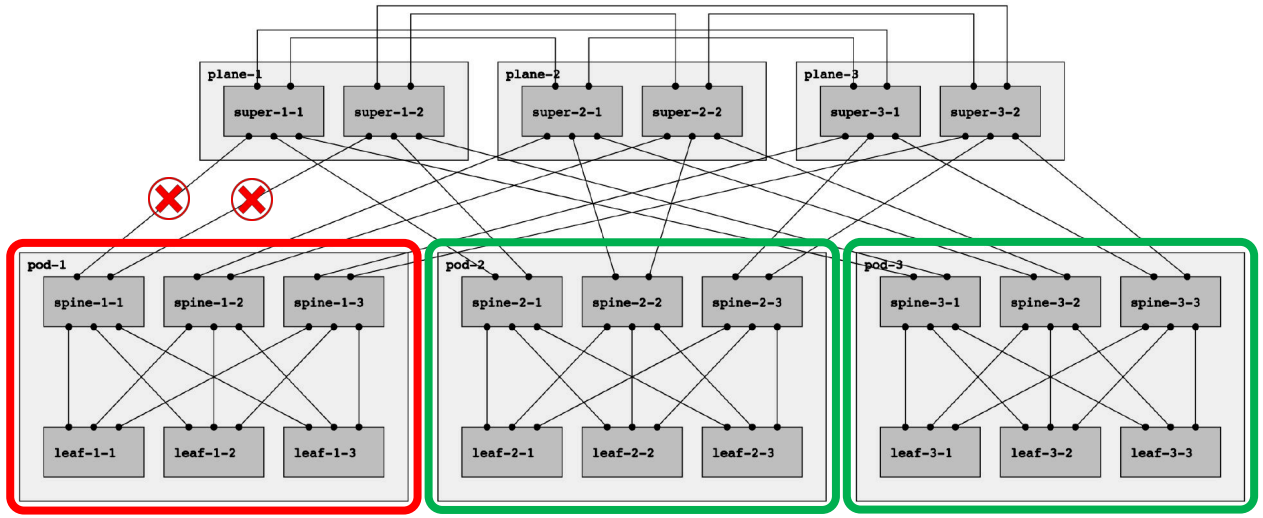
Special SPF detects pod-plane disconnect



Normal South Shortest-Path First (SPF) on superspines in plane-1:

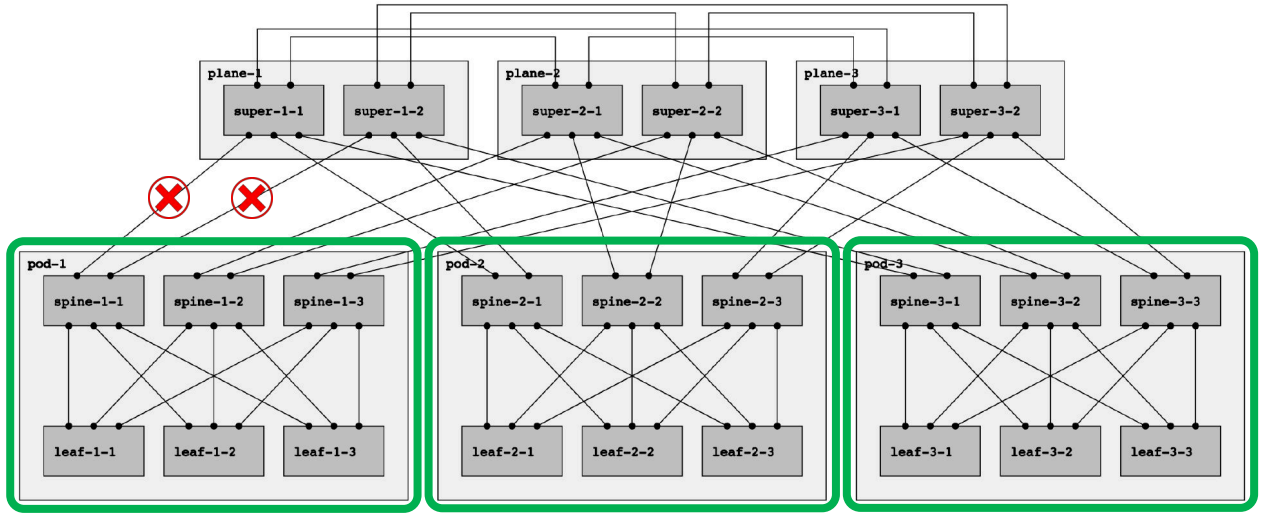
- Do **not** use east-west inter-plane links
- Plane-1 can **not** reach pod-1
- Used to populate RIB and FIB

Special SPF detects pod-plane disconnect



Normal South Shortest-Path First (SPF) on superspines in plane-1:

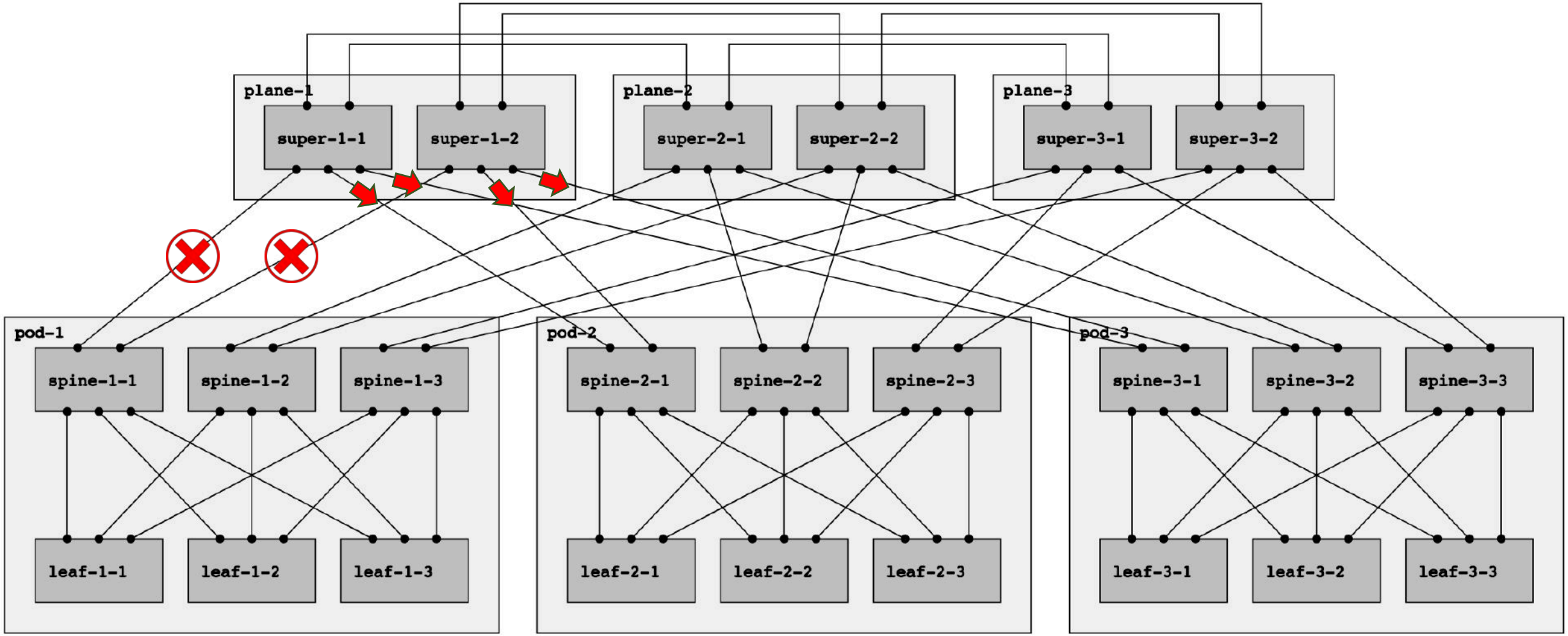
- Do **not** use east-west inter-plane links
- Plane-1 can **not** reach pod-1
- Used to populate RIB and FIB



Special South Shortest-Path First (SPF) on superspines in plane-1:

- **Do** use east-west inter-plane links
- Plane-1 **can** reach pod-1
- **Only** used to trigger negative disagg
- **Not** used to populate RIB and FIB

Super-spines advertise negative disaggregate



Negative disaggregation prefix TIEs for all prefixes originated by leaf-1-1, leaf-1-2, and leaf-1-3

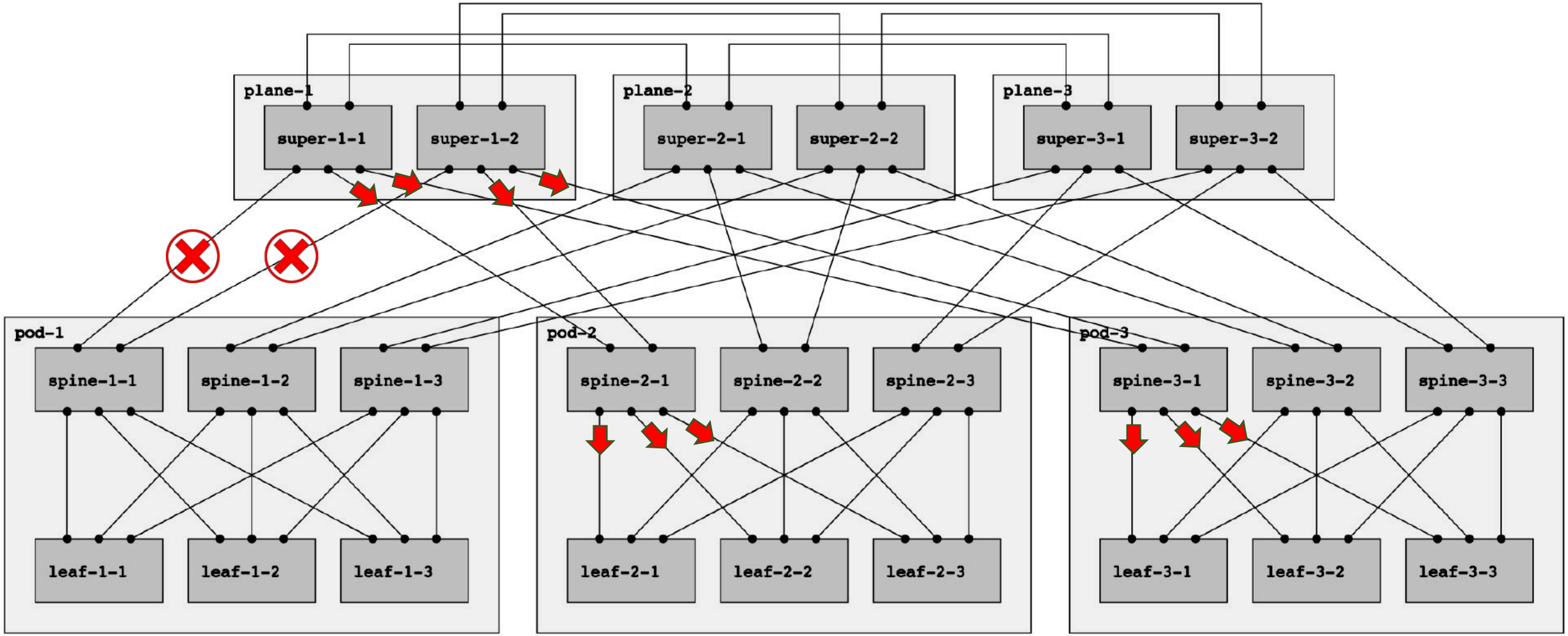
Originated negative prefix advertisement

```
super-1-2> show disaggregation
…

Negative Disaggregation TIEs:
+-----------+------------+-----------------+--------+--------+----------+----------------------------+
| Direction | Originator | Type            | TIE Nr | Seq Nr | Lifetime | Contents                   |
+-----------+------------+-----------------+--------+--------+----------+----------------------------+
| South     | 2          | Neg-Dis-Prefix  | 5      | 1      | 603937   | Neg-Dis-Prefix: 88.0.1.1/32|
|           |            |                 |        |        |          |   Metric: 2147483647       |
|           |            |                 |        |        |          | Neg-Dis-Prefix: 88.0.2.1/32|
|           |            |                 |        |        |          |   Metric: 2147483647       |
|           |            |                 |        |        |          | Neg-Dis-Prefix: 88.0.3.1/32|
|           |            |                 |        |        |          |   Metric: 2147483647       |
+-----------+------------+-----------------+--------+--------+----------+----------------------------+
```

Originated negative disaggregation prefix TIE

Spines propagate negative disaggregate



Negative disaggregation prefix TIEs for all prefixes originated by leaf-1-1, leaf-1-2, and leaf-1-3

Propagated negative prefix advertisement

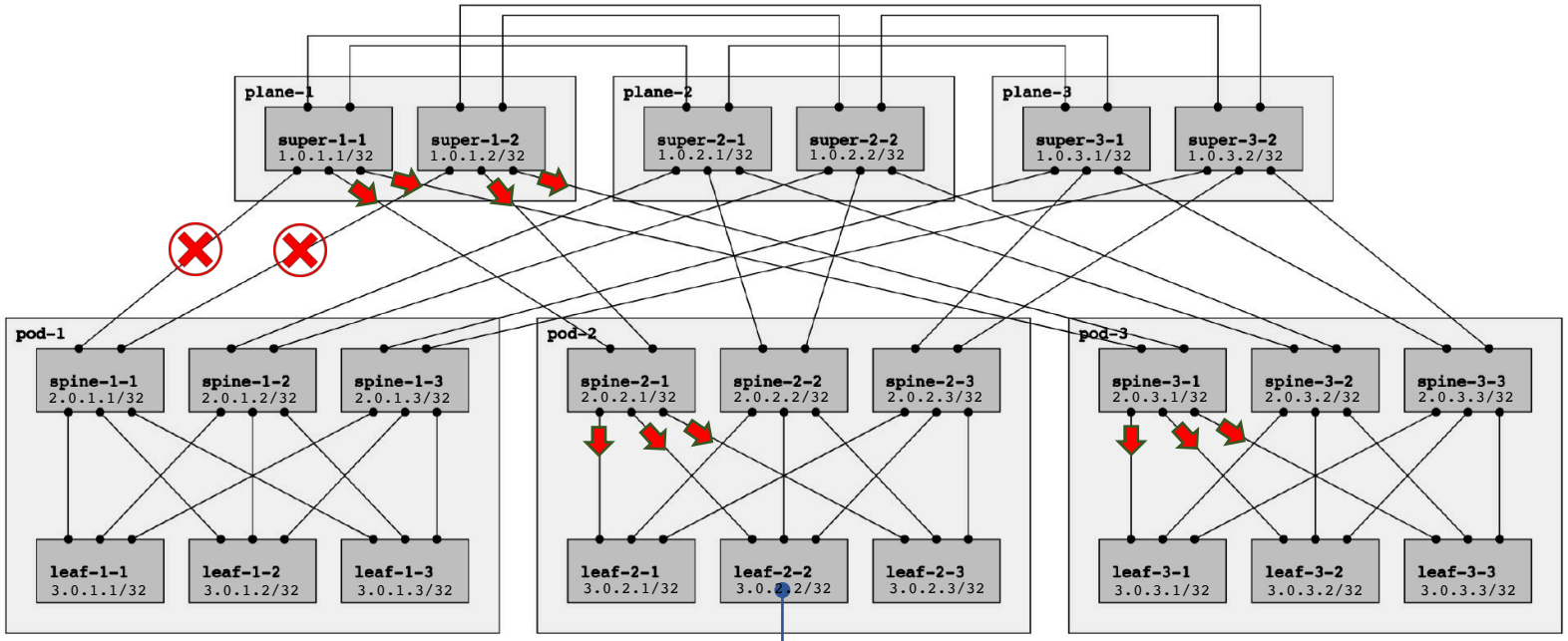
```
spine-3-1> show disaggregation

Negative Disaggregation TIEs:
+-----------+------------+----------------+--------+--------+----------+-----------------------------+
| Direction | Originator | Type           | TIE Nr | Seq Nr | Lifetime | Contents                    |
+-----------+------------+----------------+--------+--------+----------+-----------------------------+
| South     | 1          | Neg-Dis-Prefix | 5      | 1      | 557048   | Neg-Dis-Prefix: 88.0.1.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
|           |            |                |        |        |          | Neg-Dis-Prefix: 88.0.2.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
|           |            |                |        |        |          | Neg-Dis-Prefix: 88.0.3.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
+-----------+------------+----------------+--------+--------+----------+-----------------------------+
| South     | 2          | Neg-Dis-Prefix | 5      | 1      | 601516   | Neg-Dis-Prefix: 88.0.1.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
|           |            |                |        |        |          | Neg-Dis-Prefix: 88.0.2.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
|           |            |                |        |        |          | Neg-Dis-Prefix: 88.0.3.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
+-----------+------------+----------------+--------+--------+----------+-----------------------------+
| South     | 107        | Neg-Dis-Prefix | 5      | 1      | 601517   | Neg-Dis-Prefix: 88.0.1.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
|           |            |                |        |        |          | Neg-Dis-Prefix: 88.0.2.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
|           |            |                |        |        |          | Neg-Dis-Prefix: 88.0.3.1/32 |
|           |            |                |        |        |          |   Metric: 2147483647        |
+-----------+------------+----------------+--------+--------+----------+-----------------------------+
```

Received

Propagated
(re-originated)

Negative next-hops in the RIB



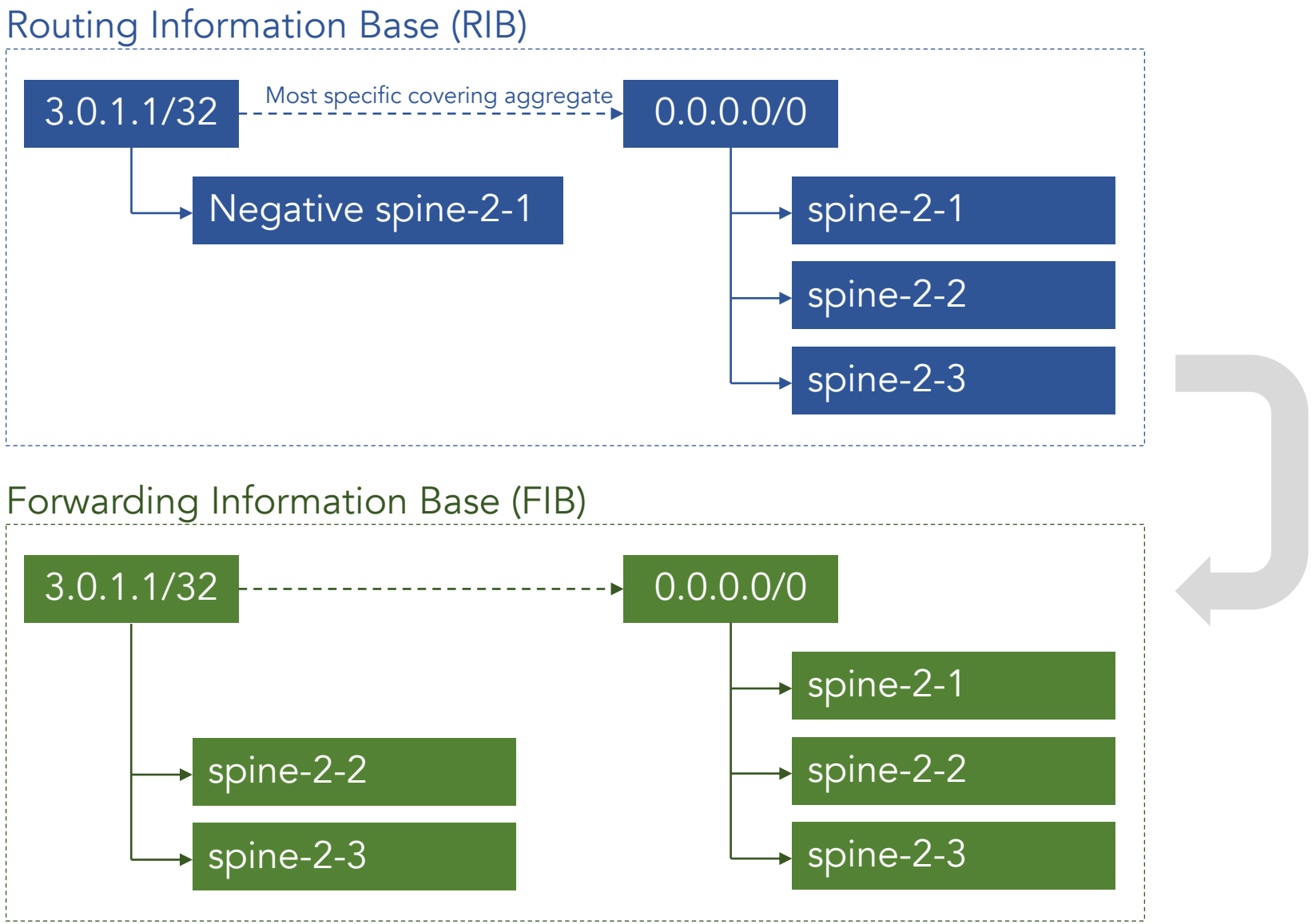
Leaf-2-2 Routing Information Base (RIB)

Destination	ECMP Next-hops
0.0.0.0/0	spine-2-1, spine-2-2, spine-2-3
3.0.1.1/32	Negative spine-2-1
3.0.1.2/32	Negative spine-2-1
3.0.1.3/32	Negative spine-2-1

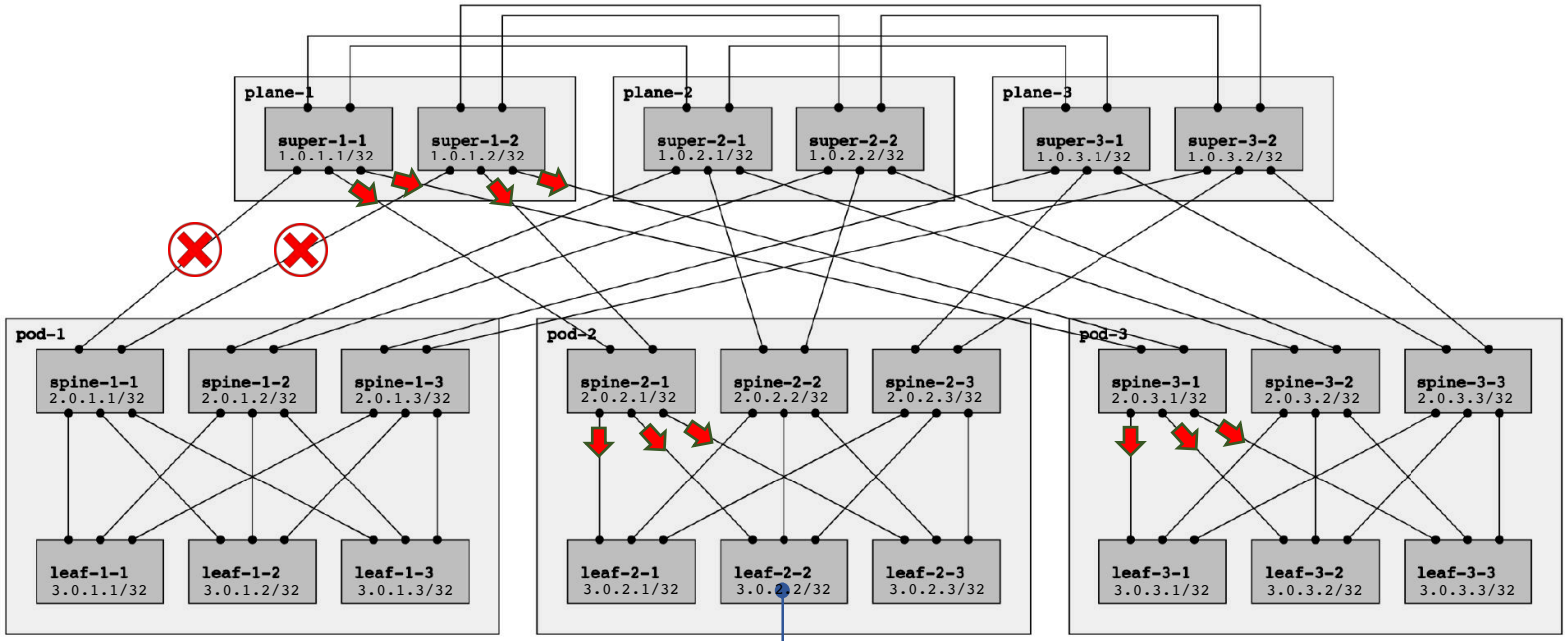
Negative next-hops in the RIB

```
leaf-3-1> show routes
IPv4 Routes:
+-------------+-----------+----------+-----------+---------------+-----------+
| Prefix      | Owner     | Next-hop | Next-hop  | Next-hop      | Next-hop  |
|             |           | Type     | Interface | Address       | Weight    |
+-------------+-----------+----------+-----------+---------------+-----------+
| 0.0.0.0/0   | North SPF | Positive | if-1007a  | 172.31.60.58  |           |
|             |           | Positive | if-1007b  | 172.31.60.58  |           |
|             |           | Positive | if-1007c  | 172.31.60.58  |           |
+-------------+-----------+----------+-----------+---------------+-----------+
| 88.0.1.1/32 | North SPF | Negative | if-1007a  | 172.31.60.58  |           |
+-------------+-----------+----------+-----------+---------------+-----------+
| 88.0.2.1/32 | North SPF | Negative | if-1007a  | 172.31.60.58  |           |
+-------------+-----------+----------+-----------+---------------+-----------+
| 88.0.3.1/32 | North SPF | Negative | if-1007a  | 172.31.60.58  |           |
+-------------+-----------+----------+-----------+---------------+-----------+
```

RIB negative next-hop to FIB positive next-hop



Complementary positive next-hops in FIB



Leaf-2-2 Routing Information Base (RIB)

Destination	ECMP Next-hops
0.0.0.0/0	spine-2-1, spine-2-2, spine-2-3
3.0.1.1/32	Negative spine-2-1
3.0.1.2/32	Negative spine-2-1
3.0.1.3/32	Negative spine-2-1

Leaf-2-2 Forwarding Information Base (FIB)

Destination	ECMP Next-hops
0.0.0.0/0	spine-2-1, spine-2-2, spine-2-3
3.0.1.1/32	spine-2-2, spine-2-3
3.0.1.2/32	spine-2-2, spine-2-3
3.0.1.3/32	spine-2-2, spine-2-3

Negative next-hops in the FIB

```
leaf-3-1> show forwarding
IPv4 Routes:
+-------------+----------+-----------+--------------+-----------+
| Prefix      | Next-hop | Next-hop  | Next-hop     | Next-hop  |
|             | Type     | Interface | Address      | Weight    |
+-------------+----------+-----------+--------------+-----------+
| 0.0.0.0/0   | Positive | if-1007a  | 172.31.60.58 |           |
|             | Positive | if-1007b  | 172.31.60.58 |           |
|             | Positive | if-1007c  | 172.31.60.58 |           |
+-------------+----------+-----------+--------------+-----------+
| 88.0.1.1/32 | Positive | if-1007b  | 172.31.60.58 |           |
|             | Positive | if-1007c  | 172.31.60.58 |           |
+-------------+----------+-----------+--------------+-----------+
| 88.0.2.1/32 | Positive | if-1007b  | 172.31.60.58 |           |
|             | Positive | if-1007c  | 172.31.60.58 |           |
+-------------+----------+-----------+--------------+-----------+
| 88.0.3.1/32 | Positive | if-1007b  | 172.31.60.58 |           |
|             | Positive | if-1007c  | 172.31.60.58 |           |
+-------------+----------+-----------+--------------+-----------+
```

More information

- Blog post on RIFT disaggregation:
  https://hikingandcoding.com/2020/07/22/rift-disaggregation/

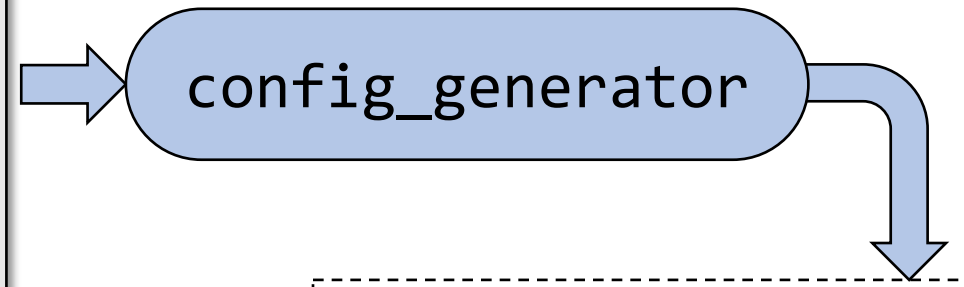
- RIFT-Python disaggregation feature guides:
  https://github.com/brunorijsman/rift-python/blob/master/doc/disaggregation-feature-guide.md
  https://github.com/brunorijsman/rift-python/blob/master/doc/positive-disaggregation-feature-guide.md
  https://github.com/brunorijsman/rift-python/blob/master/doc/negative-disaggregation-feature-guide.md

Parallel links

Config generator: parallel links

```
nr-pods: 3
nr-leaf-nodes-per-pod: 3
nr-spine-nodes-per-pod: 3
nr-superspine-nodes: 6
nr-planes: 3
leaf-spine-links:
    nr-parallel-links: 4
spine-superspine-links:
    nr-parallel-links: 3
inter-plane-links:
    nr-parallel-links: 2
```

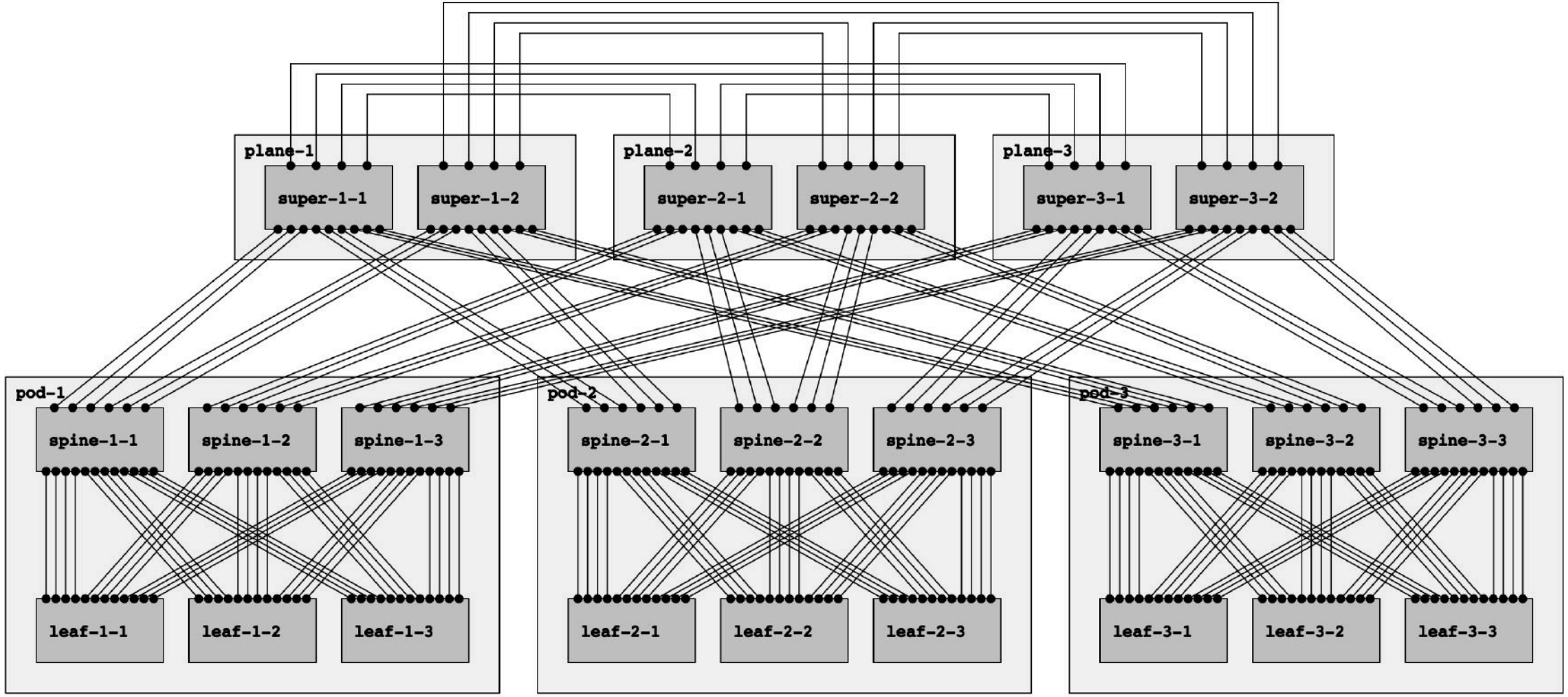
Meta-topology



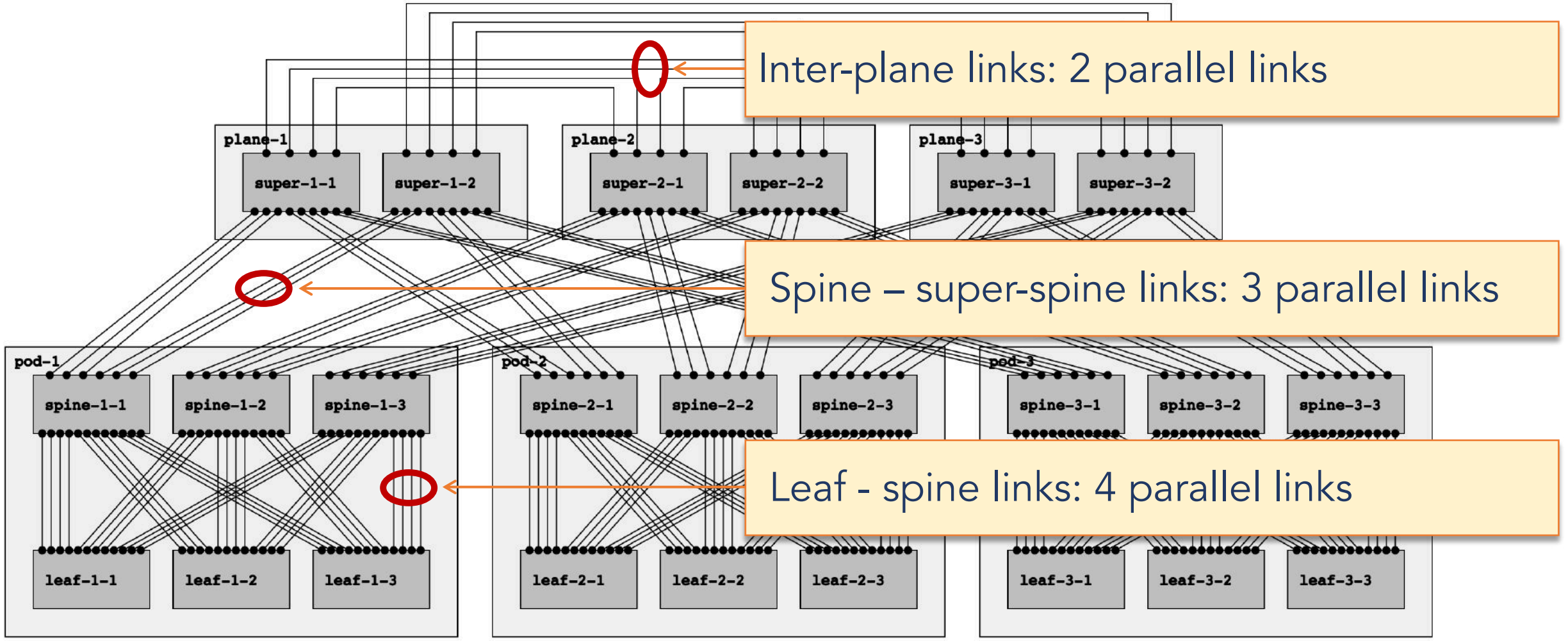
Configuration for each RIFT router
Scripts to start and stop topology
Scripts for "chaos testing"
Diagram of network

New: Can configure parallel links

Topology with parallel links



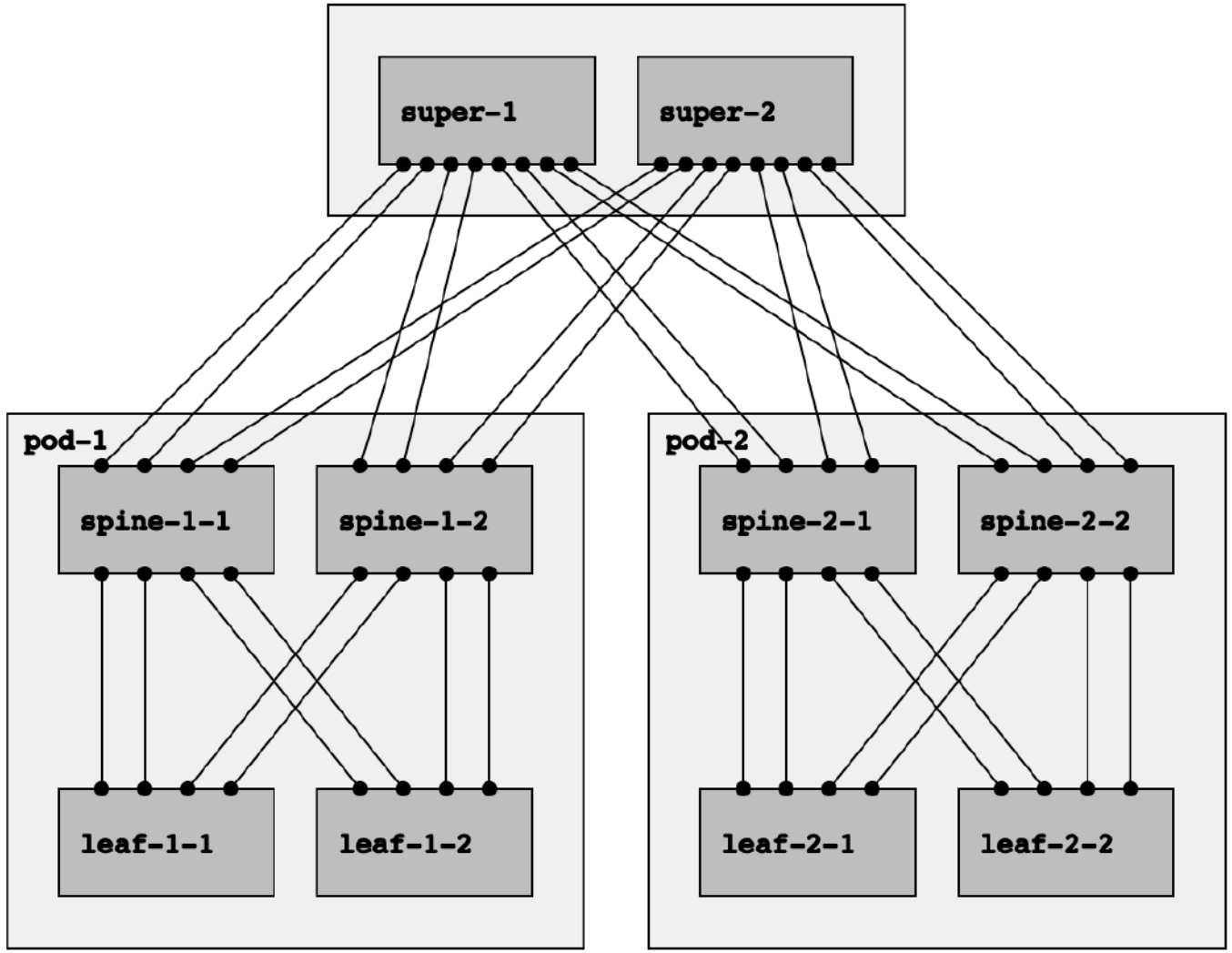
Topology with parallel links



Fabric bandwidth balancing

Scenario 1: no failures

Example topology for this section




```
nr-pods: 2
nr-leaf-nodes-per-pod: 2
nr-spine-nodes-per-pod: 2
nr-superspine-nodes: 2
leaf-spine-links:
    nr-parallel-links: 2
spine-superspine-links:
    nr-parallel-links: 2
```

Concept of neighbor

Multiple parallel links / adjacencies connect to the same neighbor

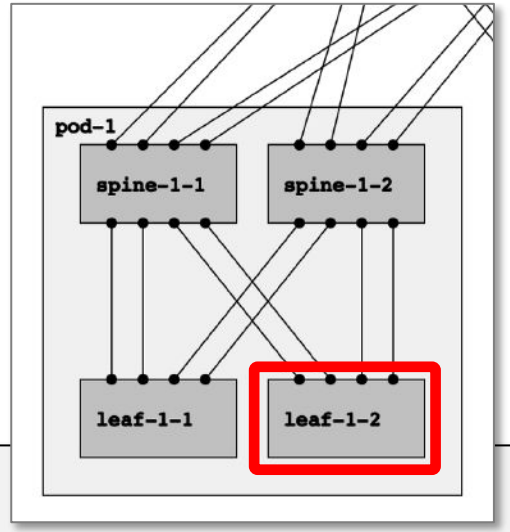
```
spine-1-2> show neighbors
+-----------+-----------+------------------+--------------------------+
| System ID | Direction | Interface        | Adjacency                |
|           |           | Name             | Name                     |
+-----------+-----------+------------------+--------------------------+
| 1         | North     | veth-102e-1c     | super-1:veth-1c-102e     |
|           |           | veth-102f-1d     | super-1:veth-1d-102f     |
+-----------+-----------+------------------+--------------------------+
| 2         | North     | veth-102g-2c     | super-2:veth-2c-102g     |
|           |           | veth-102h-2d     | super-2:veth-2d-102h     |
+-----------+-----------+------------------+--------------------------+
| 1001      | South     | veth-102a-1001c  | leaf-1-1:veth-1001c-102a |
|           |           | veth-102b-1001d  | leaf-1-1:veth-1001d-102b |
+-----------+-----------+------------------+--------------------------+
| 1002      | South     | veth-102c-1002c  | leaf-1-2:veth-1002c-102c |
|           |           | veth-102d-1002d  | leaf-1-2:veth-1002d-102d |
+-----------+-----------+------------------+--------------------------+
```

Fabric bandwidth balancing

Scenario 1: no failures

Scenario 1: no failures

- We are looking at leaf-1-2

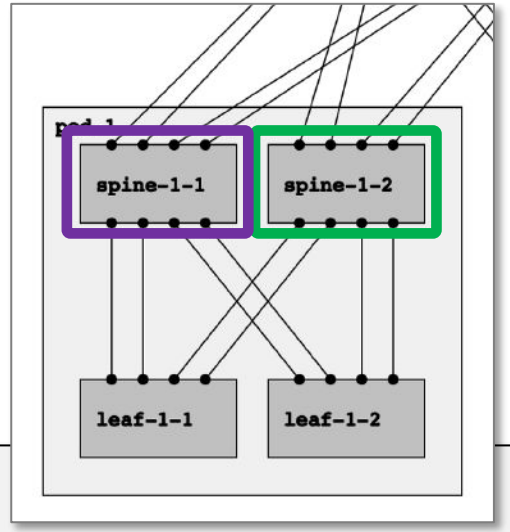



```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002a-101c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002c-102c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 1: no failures

Leaf-1-2 has two neighbors



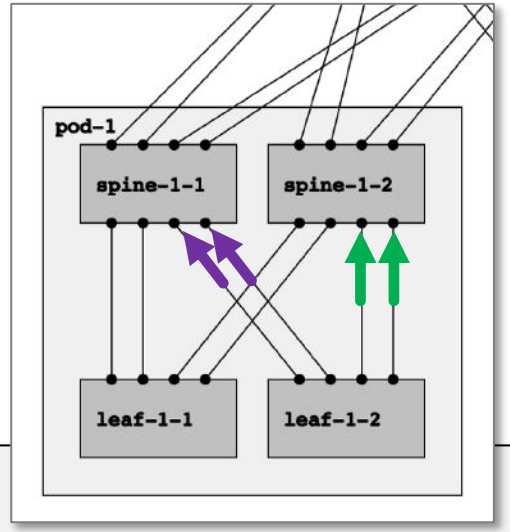

```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002a-101c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002c-102c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 1: no failures

Neighbor ingress bandwidth

- Into neighbor from leaf-1-2



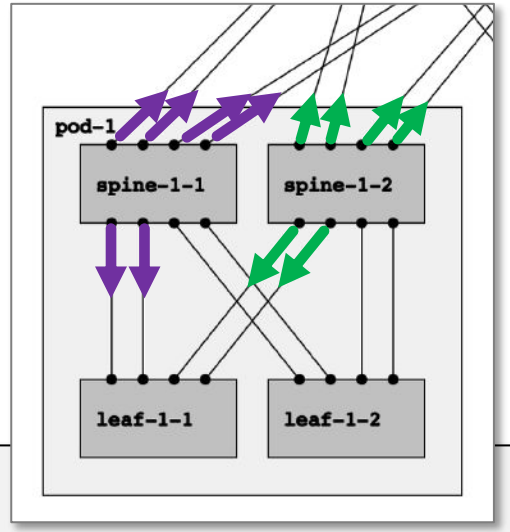

```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002a-101c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002c-102c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 1: no failures

Neighbor egress bandwidth

- Away from neighbor from leaf-1-2
- Different rule than draft-12



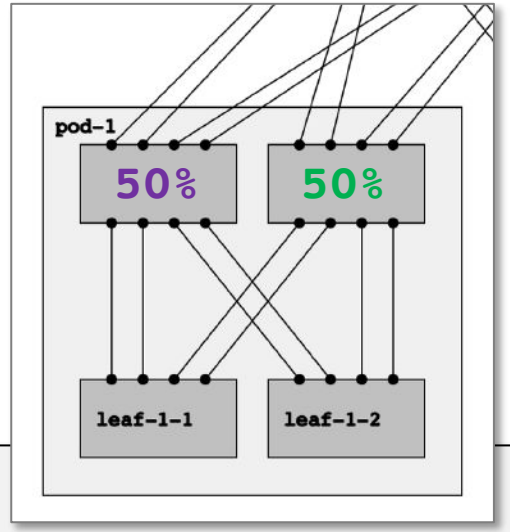

```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002a-101c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002c-102c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 1: no failures

Distribute traffic amongst neighbors

- Relative weight = ingress bw x egress bw
- Different rule than draft-12




```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
```

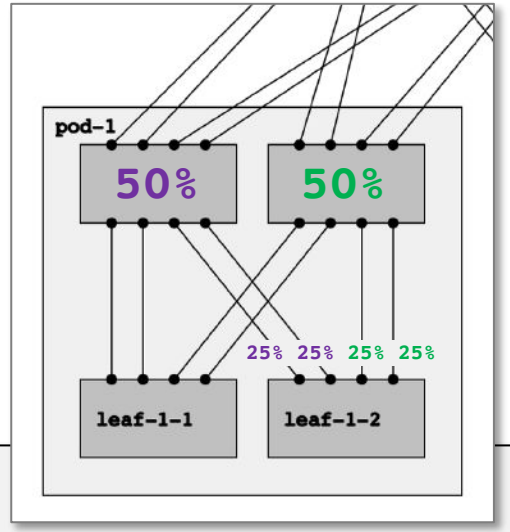
System ID	Neighbor Ingress Bandwidth	Neighbor Egress Bandwidth	Neighbor Traffic Percentage	Interface Name	Interface Bandwidth	Interface Traffic Percentage
101	20000 Mbps	60000 Mbps	50.0 %	veth-1002a-101c	10000 Mbps	25.0 %
				veth-1002b-101d	10000 Mbps	25.0 %
102	20000 Mbps	60000 Mbps	50.0 %	veth-1002c-102c	10000 Mbps	25.0 %
				veth-1002d-102d	10000 Mbps	25.0 %

Fabric bandwidth balancing

Scenario 1: no failures

Within neighbor, distribute traffic across interfaces

- Relative weight = interface bw




```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002a-101c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 50.0 %     | veth-1002c-102c | 10000 Mbps | 25.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 25.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 1: no failures

Final result: north-bound default route uses Equal Cost Multi Path (ECMP)

```
leaf-1-2> show forwarding
IPv4 Routes:
+-----------+----------+------------------+-----------+-----------+
| Prefix    | Next-hop | Next-hop         | Next-hop  | Next-hop  |
|           | Type     | Interface        | Address   | Weight    |
+-----------+----------+------------------+-----------+-----------+
| 0.0.0.0/0 | Positive | veth-1002a-101c | 99.0.10.2 | 25        |
|           | Positive | veth-1002b-101d | 99.0.12.2 | 25        |
|           | Positive | veth-1002c-102c | 99.0.14.2 | 25        |
|           | Positive | veth-1002d-102d | 99.0.16.2 | 25        |
+-----------+----------+------------------+-----------+-----------+
```

Fabric bandwidth balancing

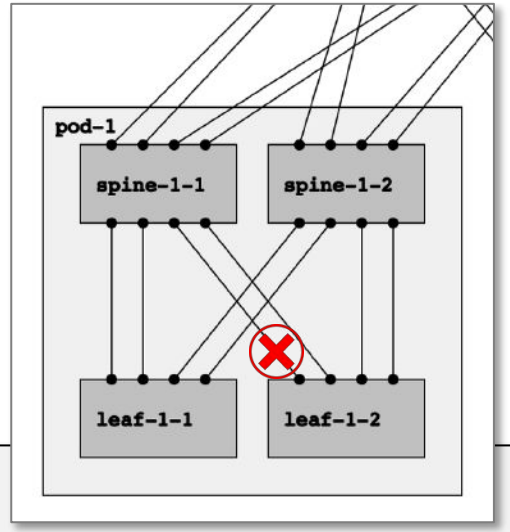
Scenario 2: leaf – spine link failure

Fabric bandwidth balancing

Scenario 2: leaf-spine failure

Scenario 2: leaf – spine link failure

- We are looking at leaf-1-2

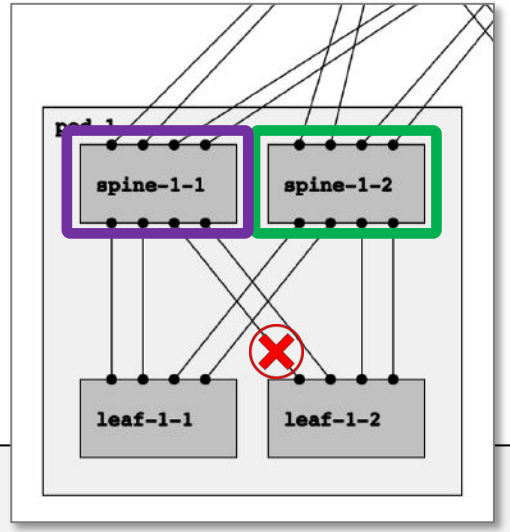



```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 10000 Mbps | 60000 Mbps | 33.3 %     | veth-1002b-101d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 66.7 %     | veth-1002c-102c | 10000 Mbps | 33.3 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 2: leaf-spine failure

Neighbor spine-1-1 is missing an interface

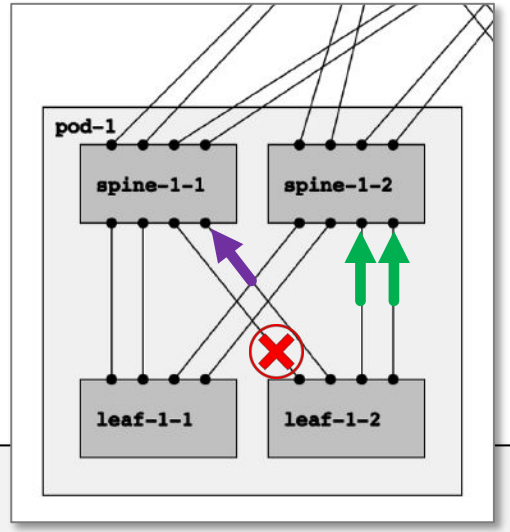



```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 10000 Mbps | 60000 Mbps | 33.3 %     | veth-1002b-101d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 66.7 %     | veth-1002c-102c | 10000 Mbps | 33.3 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 2: leaf-spine failure

Ingress bandwidth for neighbor spine-1-1 reduced from 2 Gbps to 1 Gbps

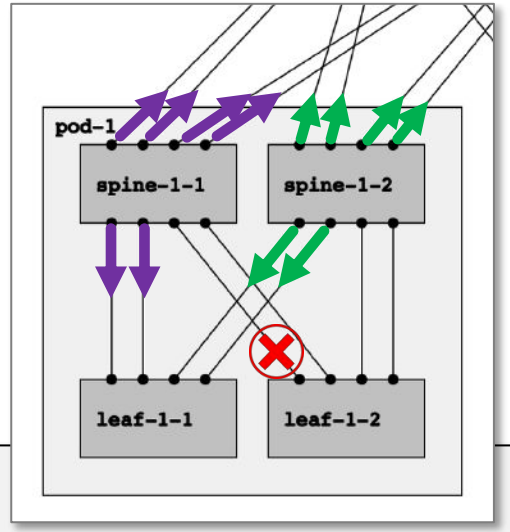



```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 10000 Mbps | 60000 Mbps | 33.3 %     | veth-1002b-101d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 66.7 %     | veth-1002c-102c | 10000 Mbps | 33.3 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 2: leaf-spine failure

Egress bandwidth has not changed



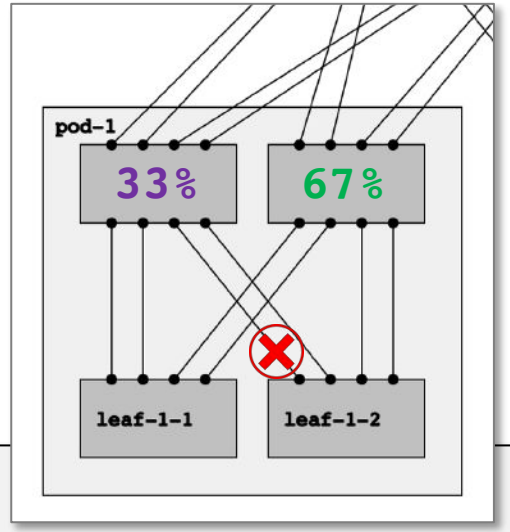

```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+-----------+-----------+-----------+-----------------+-----------+-----------+
| System ID | Neighbor  | Neighbor  | Neighbor  | Interface       | Interface | Interface |
|           | Ingress   | Egress    | Traffic   | Name            | Bandwidth | Traffic   |
|           | Bandwidth | Bandwidth | Percentage|                 |           | Percentage|
+-----------+-----------+-----------+-----------+-----------------+-----------+-----------+
| 101       | 10000 Mbps| 60000 Mbps| 33.3 %    | veth-1002b-101d | 10000 Mbps| 33.3 %    |
+-----------+-----------+-----------+-----------+-----------------+-----------+-----------+
| 102       | 20000 Mbps| 60000 Mbps| 66.7 %    | veth-1002c-102c | 10000 Mbps| 33.3 %    |
|           |           |           |           | veth-1002d-102d | 10000 Mbps| 33.3 %    |
+-----------+-----------+-----------+-----------+-----------------+-----------+-----------+
```

Fabric bandwidth balancing

Scenario 2: leaf-spine failure

Traffic to neighbors is re-distributed

- Neighbor spine-1-1 (101) gets 1/3 (33%)
- Neighbor spine-1-2 (102) gets 2/3 (67%)



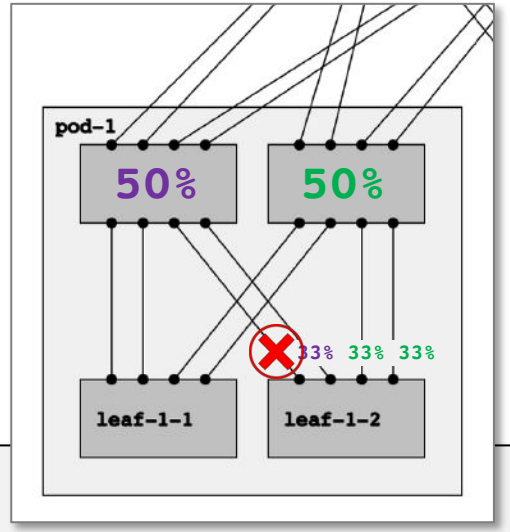

```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 10000 Mbps | 60000 Mbps | 33.3 %     | veth-1002b-101d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 66.7 %     | veth-1002c-102c | 10000 Mbps | 33.3 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 2: leaf-spine failure

Traffic to interfaces is re-distributed

- Each remaing interface gets 1/3 (33%)




```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 10000 Mbps | 60000 Mbps | 33.3 %     | veth-1002b-101d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 66.7 %     | veth-1002c-102c | 10000 Mbps | 33.3 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 33.3 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 2: leaf-spine failure

Final result: north-bound default route still uses Equal Cost Multi Path (ECMP)
Traffic is equally distributed over remaining interfaces
But traffic is not equally distributed over neighbors

```
leaf-1-2> show forwarding
IPv4 Routes:
+-----------+----------+------------------+-----------+-----------+
| Prefix    | Next-hop | Next-hop         | Next-hop  | Next-hop  |
|           | Type     | Interface        | Address   | Weight    |
+-----------+----------+------------------+-----------+-----------+
| 0.0.0.0/0 | Positive | veth-1002b-101d | 99.0.12.2 | 33        |
|           | Positive | veth-1002c-102c | 99.0.14.2 | 33        |
|           | Positive | veth-1002d-102d | 99.0.16.2 | 33        |
+-----------+----------+------------------+-----------+-----------+
```

Fabric bandwidth balancing

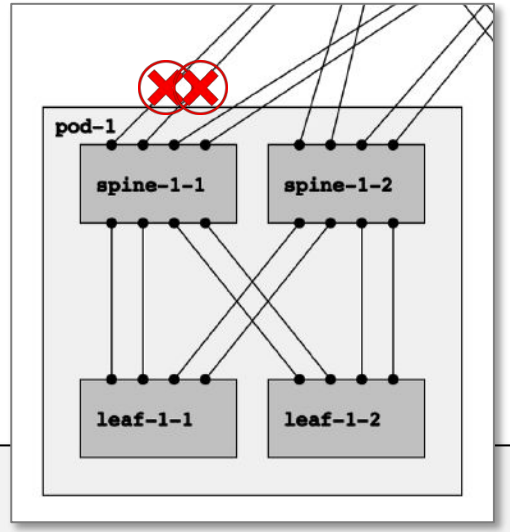
Scenario 3: spine - superspine link failures

Fabric bandwidth balancing

Scenario 3: spine-superspine failures

Scenario 3: spine - superspine link failures

- We are looking at leaf-1-2

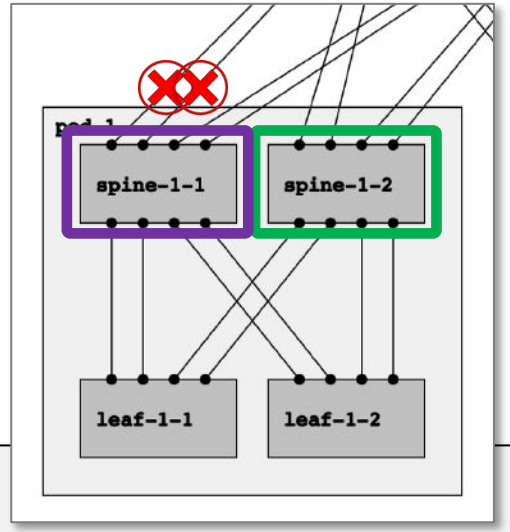



```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 40000 Mbps | 40.0 %     | veth-1002a-101c | 10000 Mbps | 20.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 20.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 60.0 %     | veth-1002c-102c | 10000 Mbps | 30.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 30.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 3: spine-superspine failures

No direct neighbors or interfaces are missing

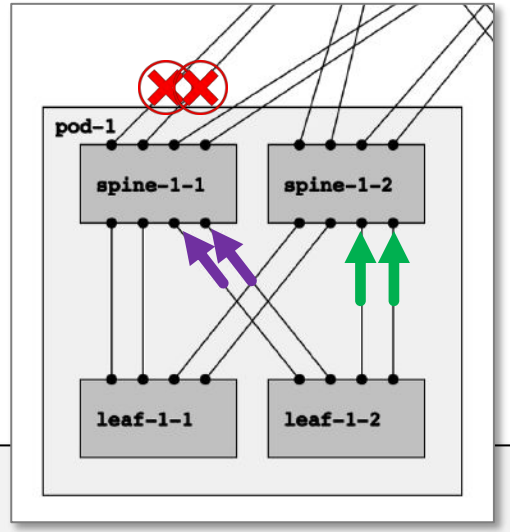



```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 40000 Mbps | 40.0 %     | veth-1002a-101c | 10000 Mbps | 20.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 20.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 60.0 %     | veth-1002c-102c | 10000 Mbps | 30.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 30.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 3: spine-superspine failures

Both neighbors have full ingress bandwidth: 2 Gbps



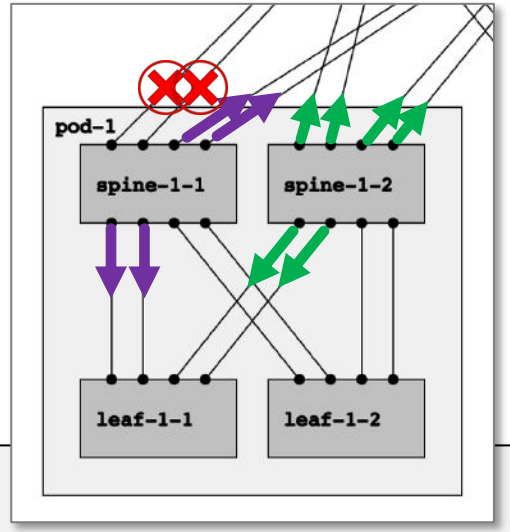

```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 40000 Mbps | 40.0 %     | veth-1002a-101c | 10000 Mbps | 20.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 20.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 60.0 %     | veth-1002c-102c | 10000 Mbps | 30.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 30.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 3: spine-superspine failures

One neighbor has reduced egress bandwidth:

- Neighbor spine-1-1 (101) has 4 Gbps
- Neighbor spine-1-2 (102) has 6 Gbps




```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
```

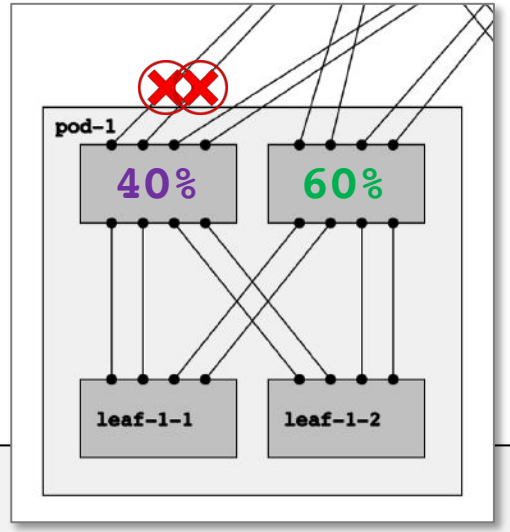
System ID	Neighbor Ingress Bandwidth	Neighbor Egress Bandwidth	Neighbor Traffic Percentage	Interface Name	Interface Bandwidth	Interface Traffic Percentage
101	20000 Mbps	40000 Mbps	40.0 %	veth-1002a-101c	10000 Mbps	20.0 %
				veth-1002b-101d	10000 Mbps	20.0 %
102	20000 Mbps	60000 Mbps	60.0 %	veth-1002c-102c	10000 Mbps	30.0 %
				veth-1002d-102d	10000 Mbps	30.0 %

Fabric bandwidth balancing

Scenario 3: spine-superspine failures

Traffic to neighbors is re-distributed

- Neighbor spine-1-1 (101) gets 40%
- Neighbor spine-1-2 (102) gets 60%



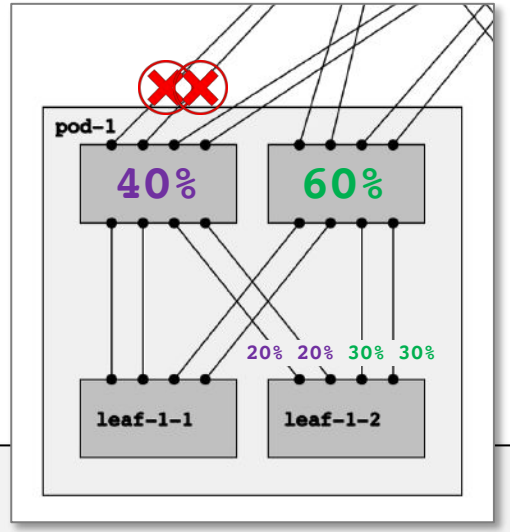

```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 40000 Mbps | 40.0 %     | veth-1002a-101c | 10000 Mbps | 20.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 20.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 60.0 %     | veth-1002c-102c | 10000 Mbps | 30.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 30.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 3: spine-superspine failures

Traffic to interfaces is re-distributed

- Interfaces to spine-1-1 get 20% each
- Interfaces to spine-1-2 get 30% each




```
leaf-1-2> show bandwidth-balancing
North-Bound Neighbors:
+-----------+------------+------------+------------+-----------------+------------+------------+
| System ID | Neighbor   | Neighbor   | Neighbor   | Interface       | Interface  | Interface  |
|           | Ingress    | Egress     | Traffic    | Name            | Bandwidth  | Traffic    |
|           | Bandwidth  | Bandwidth  | Percentage |                 |            | Percentage |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 101       | 20000 Mbps | 40000 Mbps | 40.0 %     | veth-1002a-101c | 10000 Mbps | 20.0 %     |
|           |            |            |            | veth-1002b-101d | 10000 Mbps | 20.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
| 102       | 20000 Mbps | 60000 Mbps | 60.0 %     | veth-1002c-102c | 10000 Mbps | 30.0 %     |
|           |            |            |            | veth-1002d-102d | 10000 Mbps | 30.0 %     |
+-----------+------------+------------+------------+-----------------+------------+------------+
```

Fabric bandwidth balancing

Scenario 3: spine-superspine failures

Final result: north-bound default route uses Non-Equal Cost Multi Path (NECMP)
Two interfaces each get 20% of traffic
The other two interfaces each get 30% of traffic

```
leaf-1-2> show forwarding
IPv4 Routes:
+-----------+----------+------------------+-----------+-----------+
| Prefix    | Next-hop | Next-hop         | Next-hop  | Next-hop  |
|           | Type     | Interface        | Address   | Weight    |
+-----------+----------+------------------+-----------+-----------+
| 0.0.0.0/0 | Positive | veth-1002a-101c | 99.0.10.2 | 20        |
|           | Positive | veth-1002b-101d | 99.0.12.2 | 20        |
|           | Positive | veth-1002c-102c | 99.0.14.2 | 30        |
|           | Positive | veth-1002d-102d | 99.0.16.2 | 30        |
+-----------+----------+------------------+-----------+-----------+
```

Performance monitoring

Processing and queueing time per FSM event

```
leaf-1-2> show interface veth-1002a-101c fsm verbose-history
```

Sequence Nr	Time Since First	Time Since Prev	Queue Time	Processing Time	From State	Event	Actions and Pushed Events	To State	Implicit
108977	24.004663	0.003986	0.000209	0.000070	THREE_WAY	LIE_RECEIVED	process_lie	None	False
108970	24.000677	0.000095	0.000082	0.000712	THREE_WAY	SEND_LIE	send_lie	None	False
108969	24.000582	0.000396	0.000145	0.000011	THREE_WAY	TIMER_TICK	check_hold_time_expired SEND_LIE	None	False
108967	24.000185	0.987224	0.000247	0.000058	THREE_WAY	LIE_RECEIVED	process_lie	None	False
108953	23.012961	0.005232	0.001388	0.000052	THREE_WAY	LIE_RECEIVED	process_lie	None	False
108946	23.007729	0.000169	0.000142	0.001028	THREE_WAY	SEND_LIE	send_lie	None	False
108945	23.007560	0.007361	0.006977	0.000025	THREE_WAY	TIMER_TICK	check_hold_time_expired SEND_LIE	None	False
108943	23.000199	0.999376	0.000314	0.000076	THREE_WAY	LIE_RECEIVED	process_lie	None	False
108929	22.000823	0.000659	0.000124	0.000038	THREE_WAY	LIE_RECEIVED	process_lie	None	False
108927	22.000164	0.022760	0.000259	0.000056	THREE_WAY	LIE_RECEIVED	process_lie	None	False

Extreme processing and queueing times

```
leaf-1-2> show engine
+-----------------------------------+-----------------------------------+
| Stand-alone                       | True                              |
| Interactive                       | False                             |
.                                   .                                   .
.                                   .                                   .
.                                   .                                   .
| Timer slips > 10ms                | 0                                 |
| Timer slips > 100ms               | 0                                 |
| Timer slips > 1000ms              | 0                                 |
| Max pending events processing time | 0.037596                          |
| Max expired timers processing time | 0.077908                          |
| Max select processing time        | 0.969274                          |
| Max ready-to-read processing time | 0.030650                          |
+-----------------------------------+-----------------------------------+
```

Questions?