Aggregating BGP routes in Massive Scale Data Centers
draft-heitz-idr-msdc-bgp-aggregation-00

IETF 103, November 2018

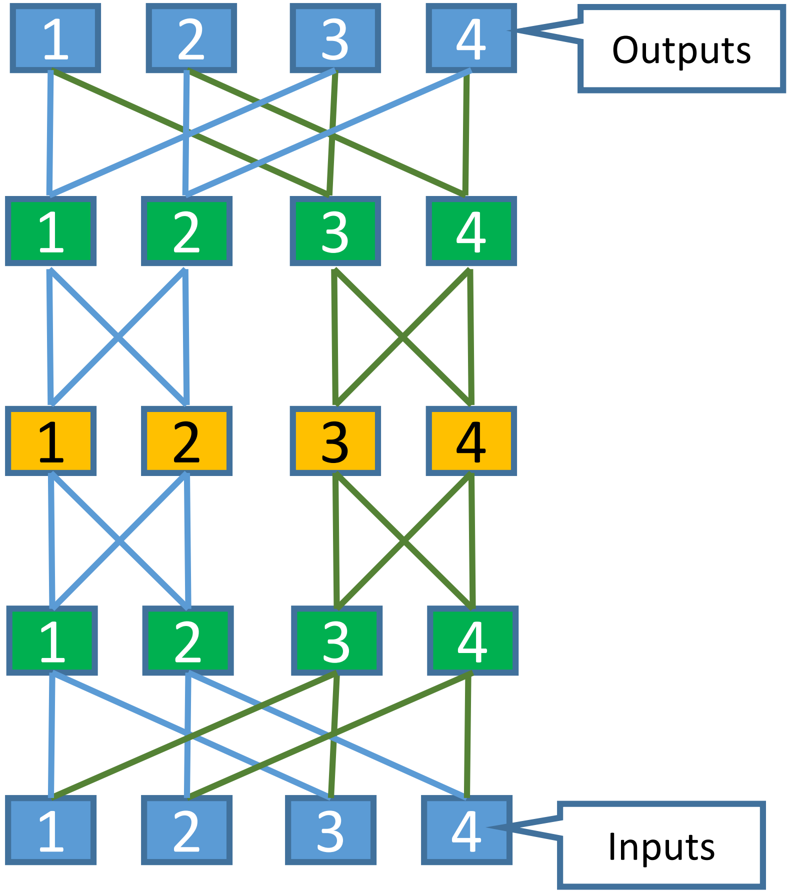
Bangkok

Jakob Heitz (Cisco)

Dhananjaya Rao (Cisco)

Clos Fabric

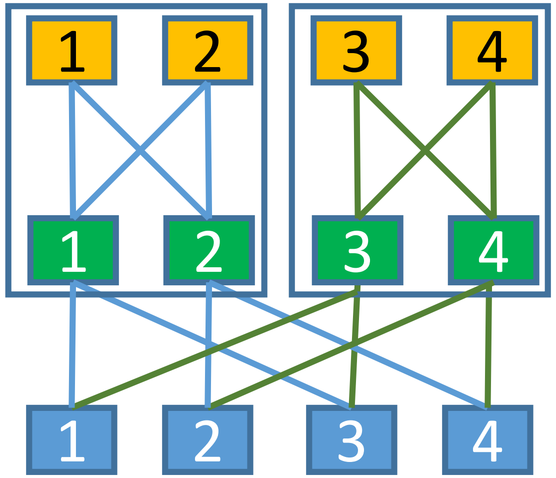
Described in Clos Paper



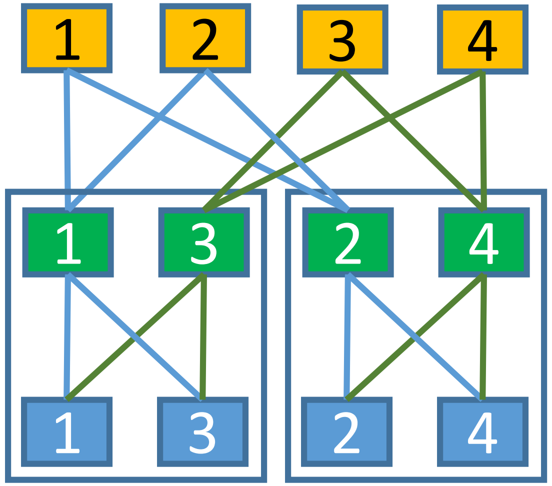
Outputs

Inputs

folded

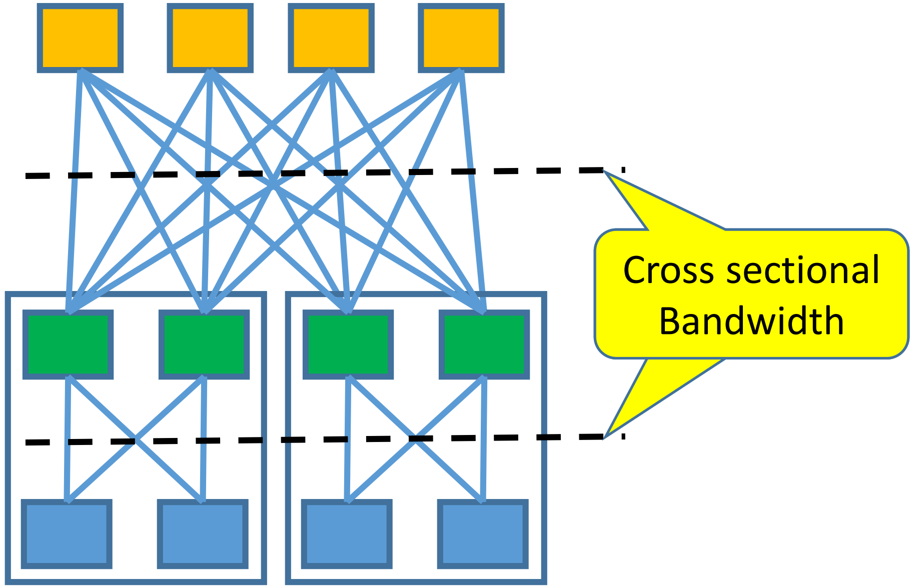


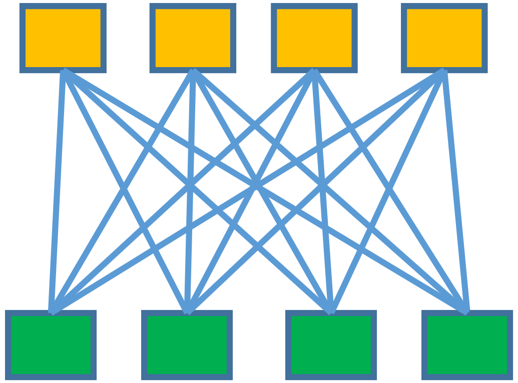
Shift some nodes and
see pods at the bottom



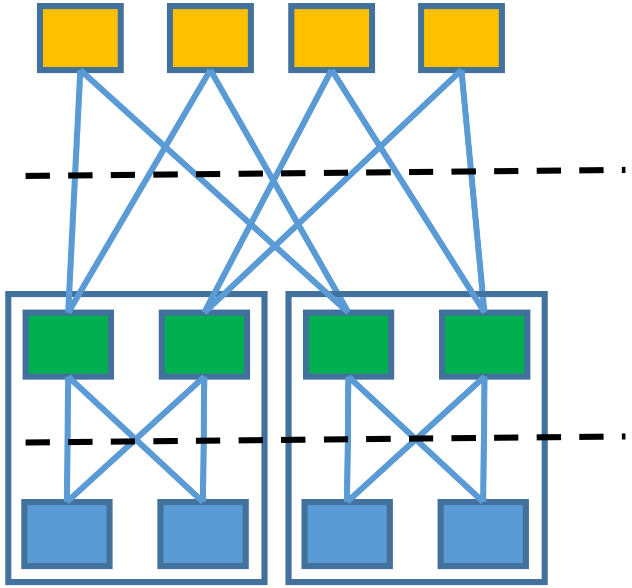
Note the spine planes
No East-West links. They waste ports

Fully Meshed Spines



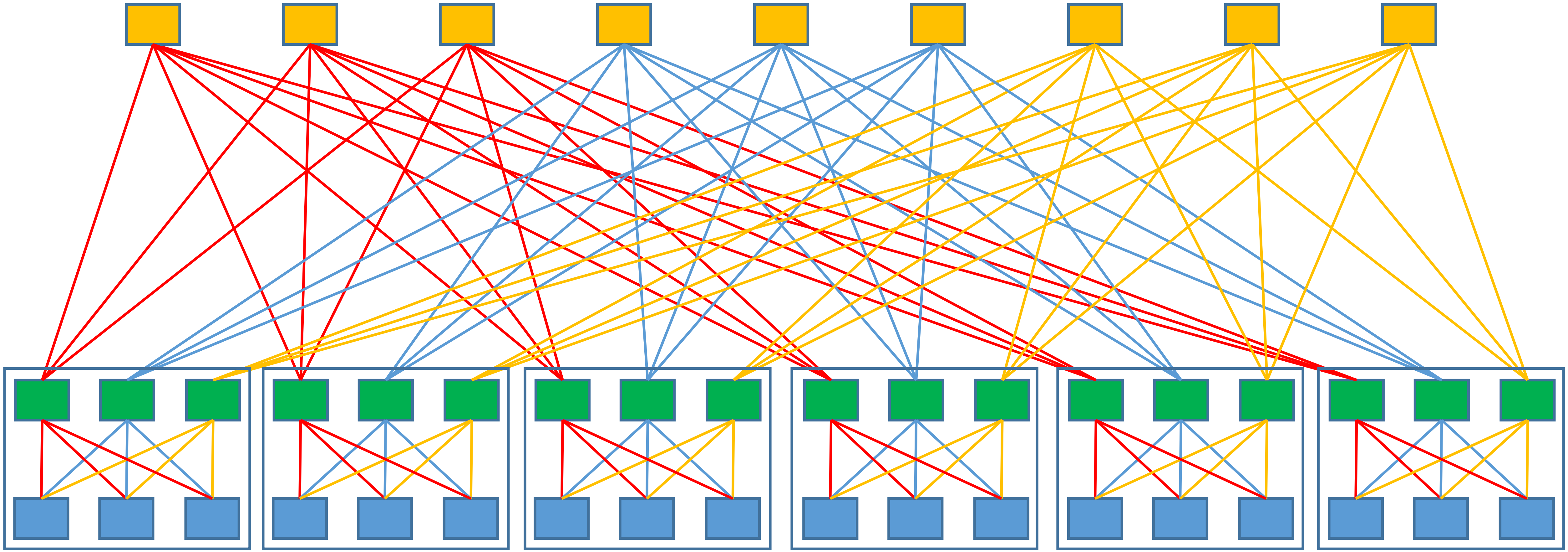


Same connectivity.
Leaves are wasted nodes.



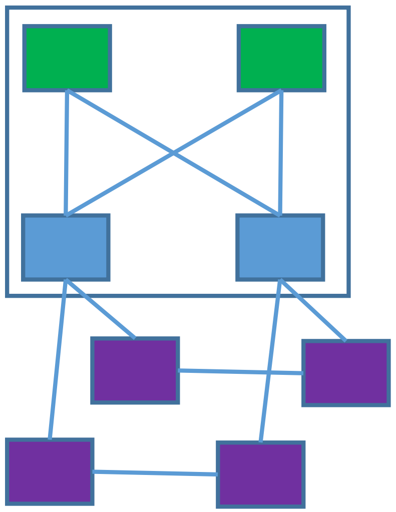
Makes Sense

Spine Planes

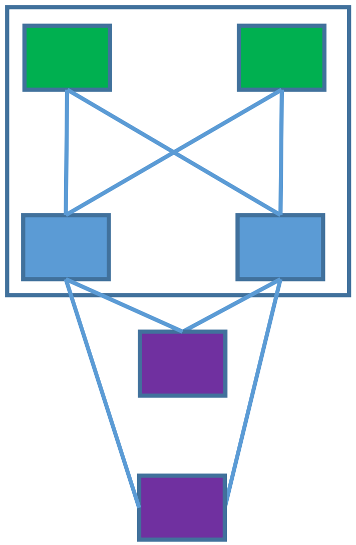


Server Redundancy

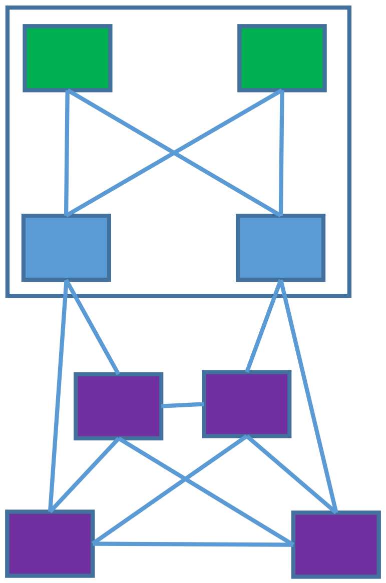
If one northbound port is sufficient



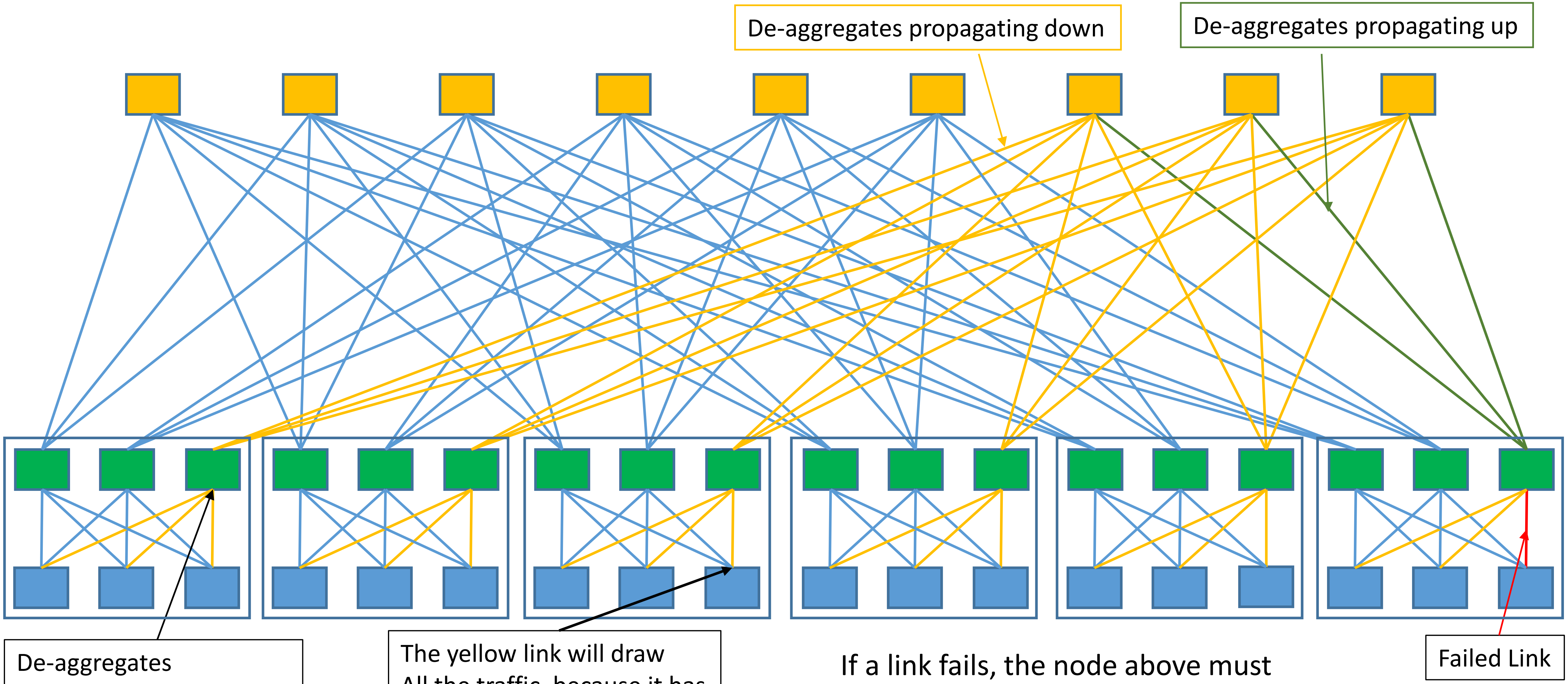
If two northbound ports are needed



Allows any server to burst to 4x link bandwidth.



Why we cannot aggregate



De-aggregates propagating down

De-aggregates propagating up

De-aggregates propagate down into every pod and every TOR

The yellow link will draw All the traffic, because it has A longer netmask

Failed Link

If a link fails, the node above must deaggregate in order to exclude the failed link

Scale

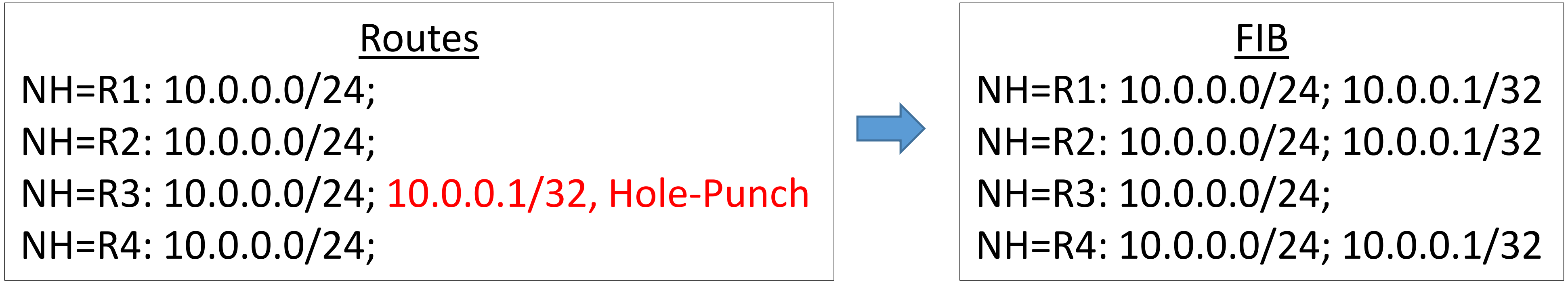
- Number of links in Clos fabric scales in the same order as the number of connected servers.
- SPF computation time scales in a higher order than the number of links.
- Data centers with a few 100,000 servers exist or are planned. A design goal to connect 1 million servers is enough for the foreseeable future
- Maximum requirement is 8 million links and 130,000 switches.
- BGP with route aggregation can do that with only 100’s of routes.
- Every switch aggregates its south side routes and sends one route.
- For each failure, it sends one route: a negative route.

Negative Route Problems

- Massive failures cause many negative routes
  - No better than many positive routes
- Need extra config to know when to send the negative route
  - Or an error prone algorithm to figure it out automatically
- Race condition between overlapping negative and positive routes
- Computation of FIB entries can be CPU intensive in pathological cases

Use of Negative Route

- 3 new BGP communities
  - Hole-Punch: Punch a hole out of an aggregate
  - Punch-Accept: Can take a punch
  - Do-Not-Aggregate
- Example: 4 routers send a route for 10.0.0.0/24, but R3 cannot reach 10.0.0.1/32



Routes

NH=R1: 10.0.0.0/24;
NH=R2: 10.0.0.0/24;
NH=R3: 10.0.0.0/24; 10.0.0.1/32, Hole-Punch
NH=R4: 10.0.0.0/24;

FIB

NH=R1: 10.0.0.0/24; 10.0.0.1/32
NH=R2: 10.0.0.0/24; 10.0.0.1/32
NH=R3: 10.0.0.0/24;
NH=R4: 10.0.0.0/24; 10.0.0.1/32

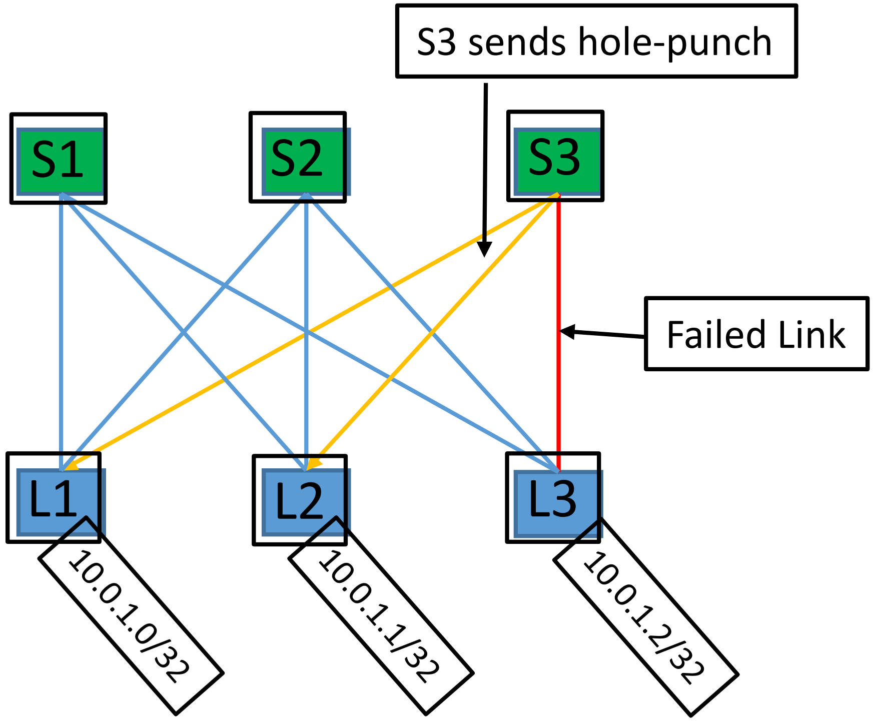
Another Hole-Punch Example

L1 has routes:

- 10.0.1.0/24 with paths:
  - NH=S1, atomic-agg, Punch-Accept, multipath
  - NH=S2, atomic-agg, Punch-Accept, multipath
  - NH=S3, atomic-agg, Punch-Accept, multipath
- 10.0.1.2/32 with paths:
  - NH=S3, Hole-Punch, do-not-aggregate, Low Preference

L1 uses the hole-punch to search up the radix tree to find 10.0.1.0/24 and create chad routes from it. L1 now has routes:

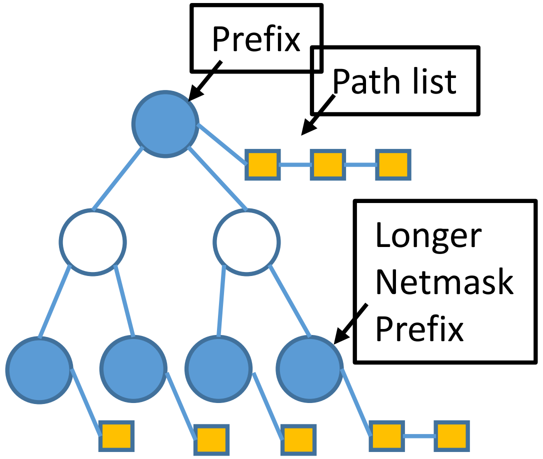
- 10.0.1.0/24 with paths:
  - NH=S1, atomic-agg, Punch-Accept, multipath
  - NH=S2, atomic-agg, Punch-Accept, multipath
  - NH=S3, atomic-agg, Punch-Accept, multipath
- 10.0.1.2/32 with paths:
  - NH=S1, Chad, multipath
  - NH=S2, Chad, multipath
  - NH=S3, Chad, hidden
  - NH=S3, Hole-Punch, do-not-aggregate, Low Preference

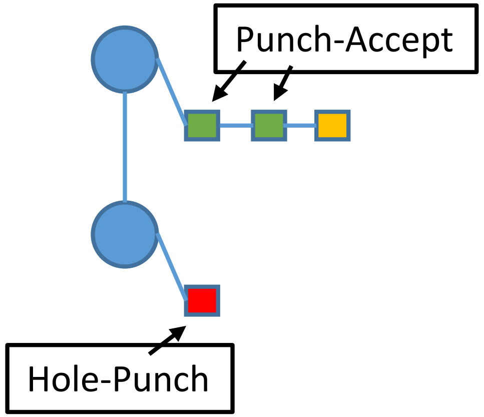


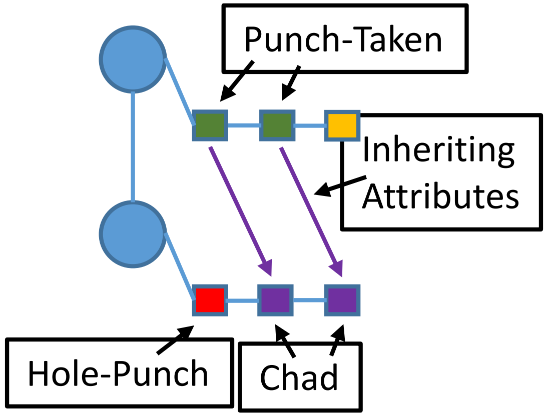
Radix Tree

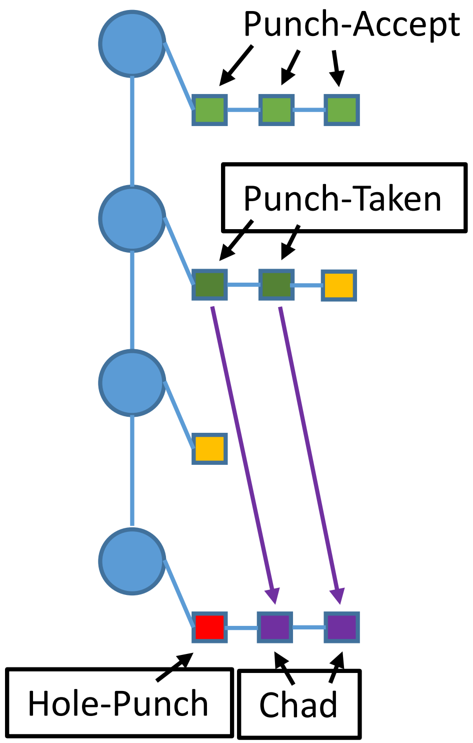
Chads are taken from the closest Punch-Accept routes

A Hole-Punch route finds Punch-Accept routes higher in the radix tree and punches Chad routes from it.

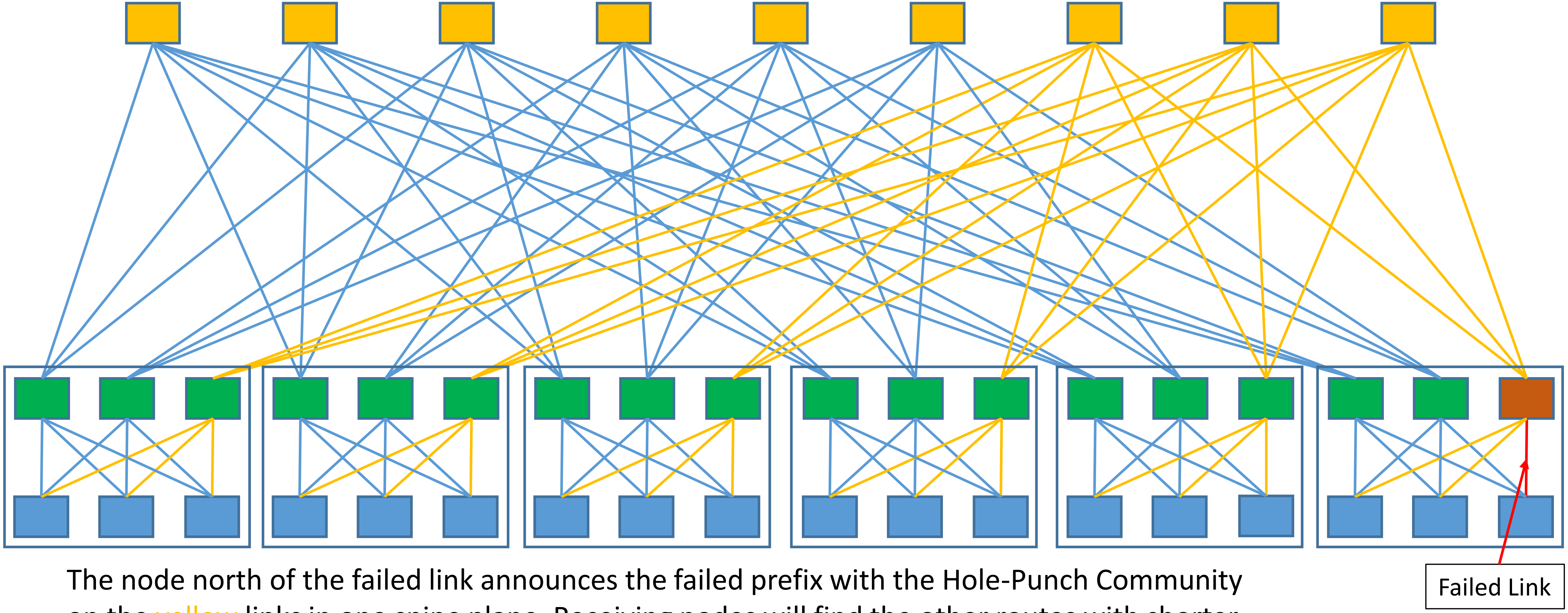






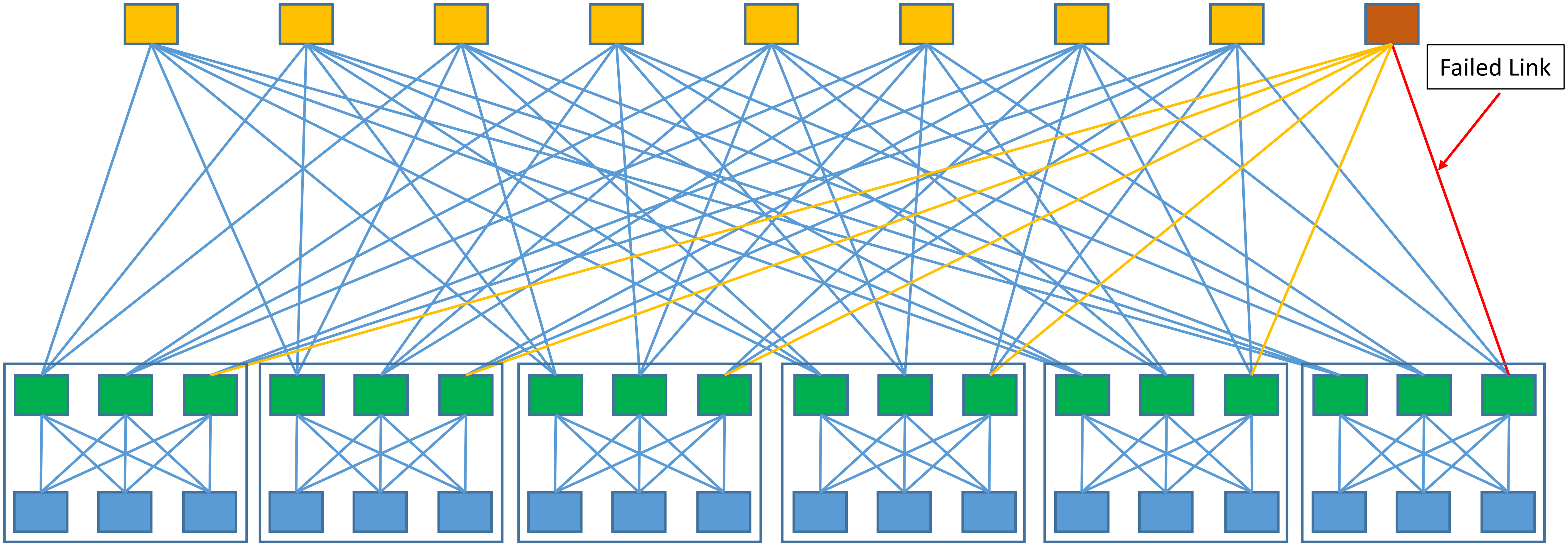


Failure Modes



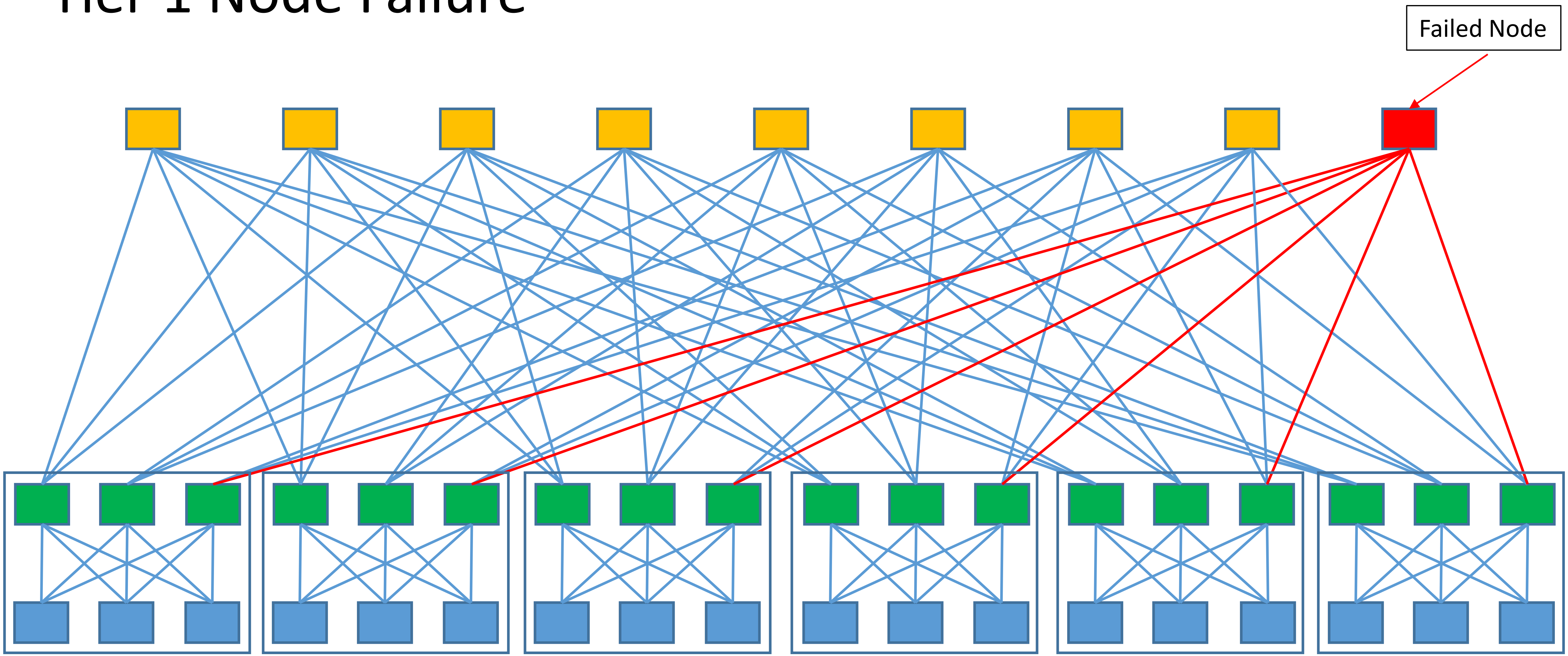
The node north of the failed link announces the failed prefix with the Hole-Punch Community on the yellow links in one spine plane. Receiving nodes will find the other routes with shorter netmask and prefer them instead.

North Side Link Failure



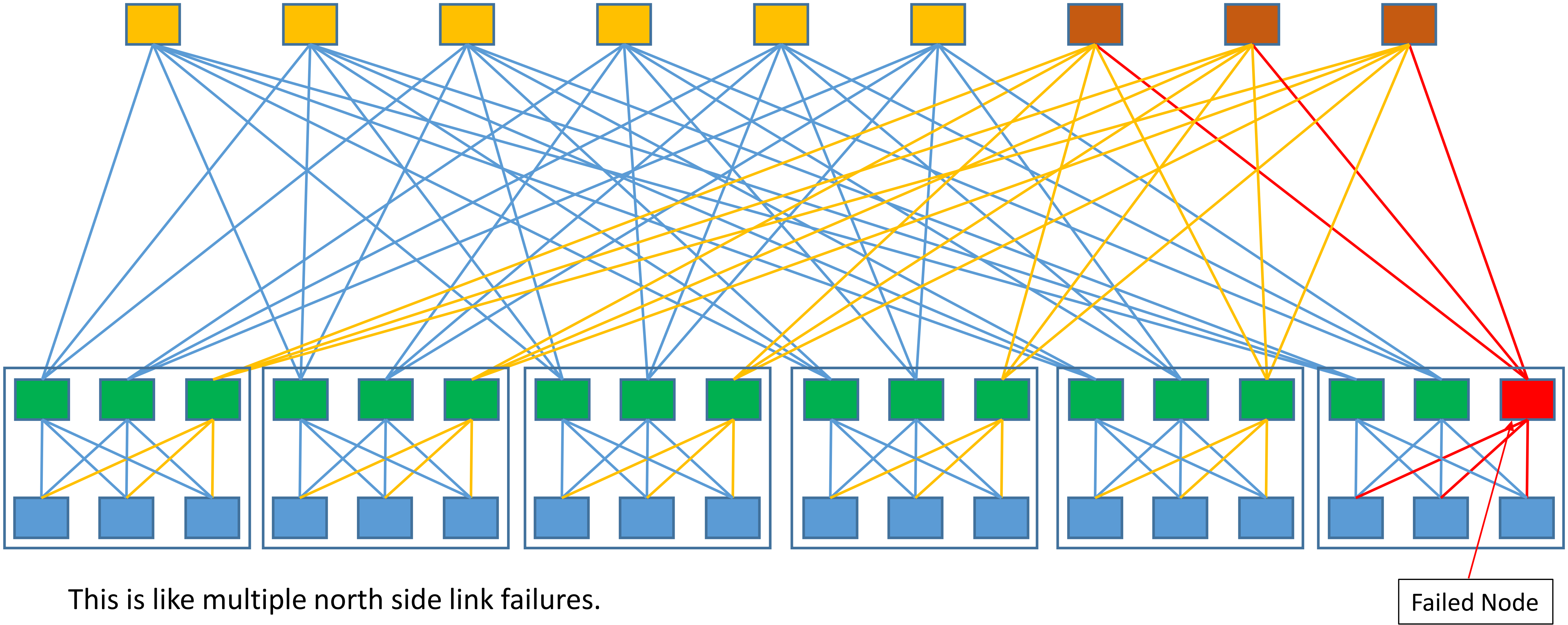
The node north of the failed link announces the failed prefix with the Hole-Punch Community. A single BGP route is sent on the yellow links in the north side of one spine plane.

Tier 1 Node Failure



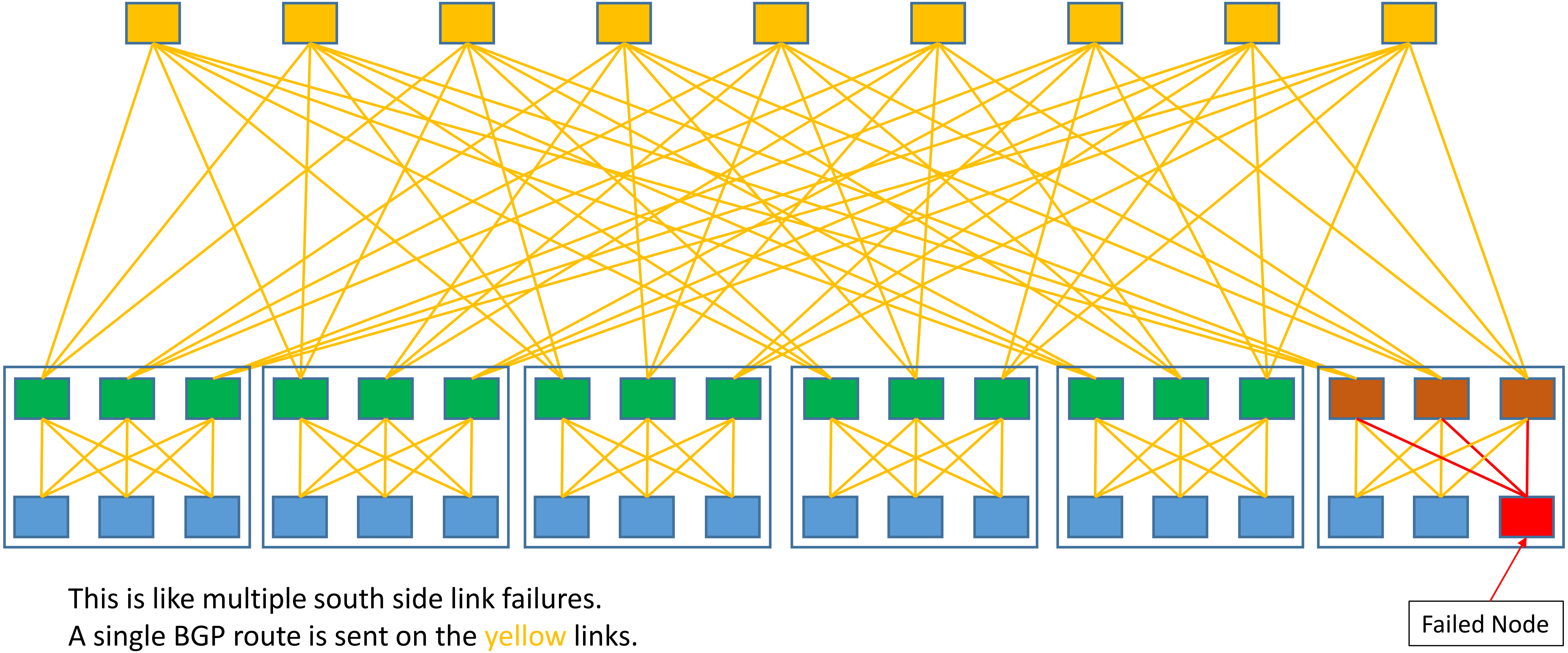
No extra BGP advertisements

Tier 2 Node Failure



This is like multiple north side link failures.
A single BGP route is sent on the yellow links in one spine plane.

Tier 3 Node Failure



This is like multiple south side link failures.
A single BGP route is sent on the yellow links.

Handling Massive Failures

- Reduce switch cost: small FIB

1. Do not install all chad routes
   - Lose bandwidth for just one destination from one switch
2. Do not install an aggregate with too many hole-punches
   - Lose one spine plane from one switch
3. Once no clean spine planes left, install the cleanest and then add individual routes missing on that plane if they are available on other planes.
4. The fabric controller has the overall view and can install the most important routes coordinated across the fabric.

Switch maintenance strategy

- Node failures are most common when they are taken out for software upgrades.
- Best to upgrade many nodes in a single spine plane.
- If servers are connected redundantly, then a TOR outage creates no hole-punch route if its redundant twin remains in service.

Configuration

All the BGP sessions need to be configured on each switch. The BGP sessions need to be configured as northbound or southbound. The routes that are expected to complete an aggregate route must be configured. IP addresses need to be chosen such that they can be aggregated.

https://tools.ietf.org/html/draft-heitz-idr-msdc-fabric-autoconf-00
describes a protocol that can
discover and configure the entire fabric. If that document is used, then no IP addresses or tier designations or any other location dependent configuration is required on the switches.