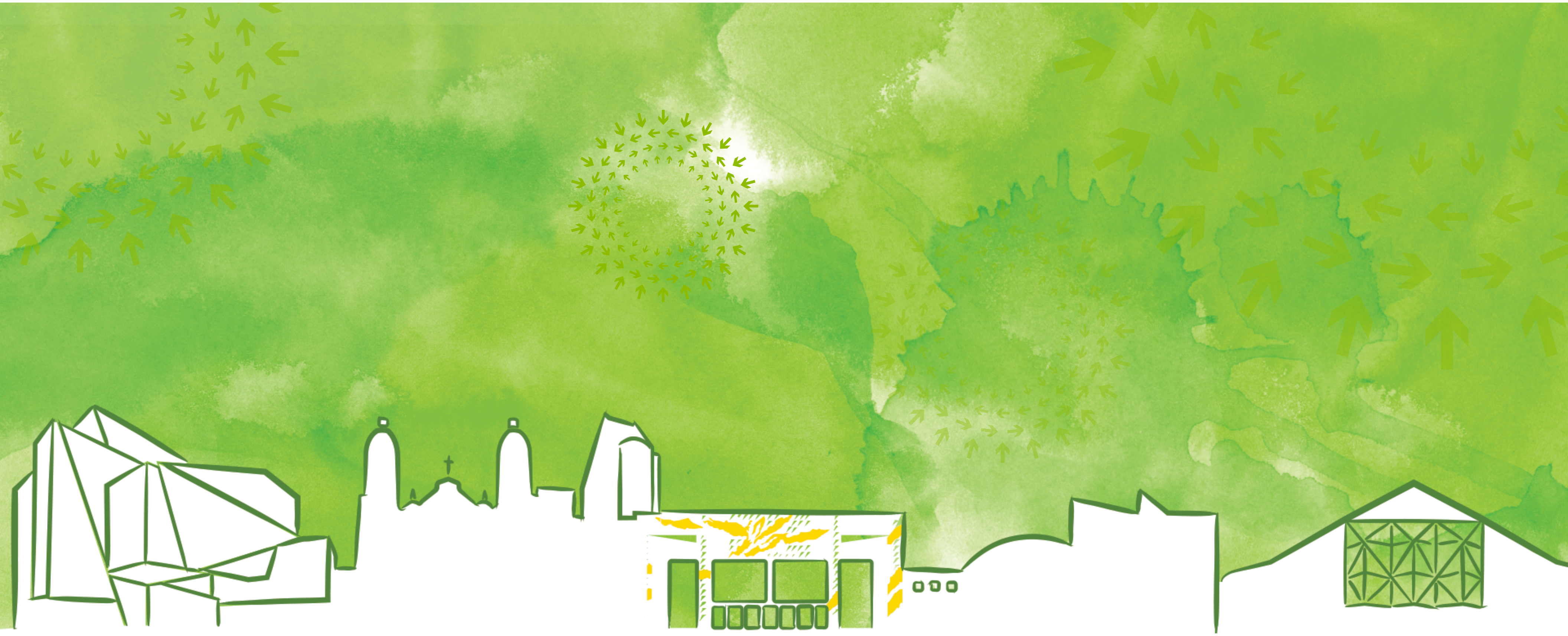
OPEN

Compute Project



OCP U.S. SUMMIT 2016 March 9-10 | San Jose, CA

Platinum Sponsor

OCP U.S. SUMMIT 2016

L3 testing of 6-Pack

Hany Morsy

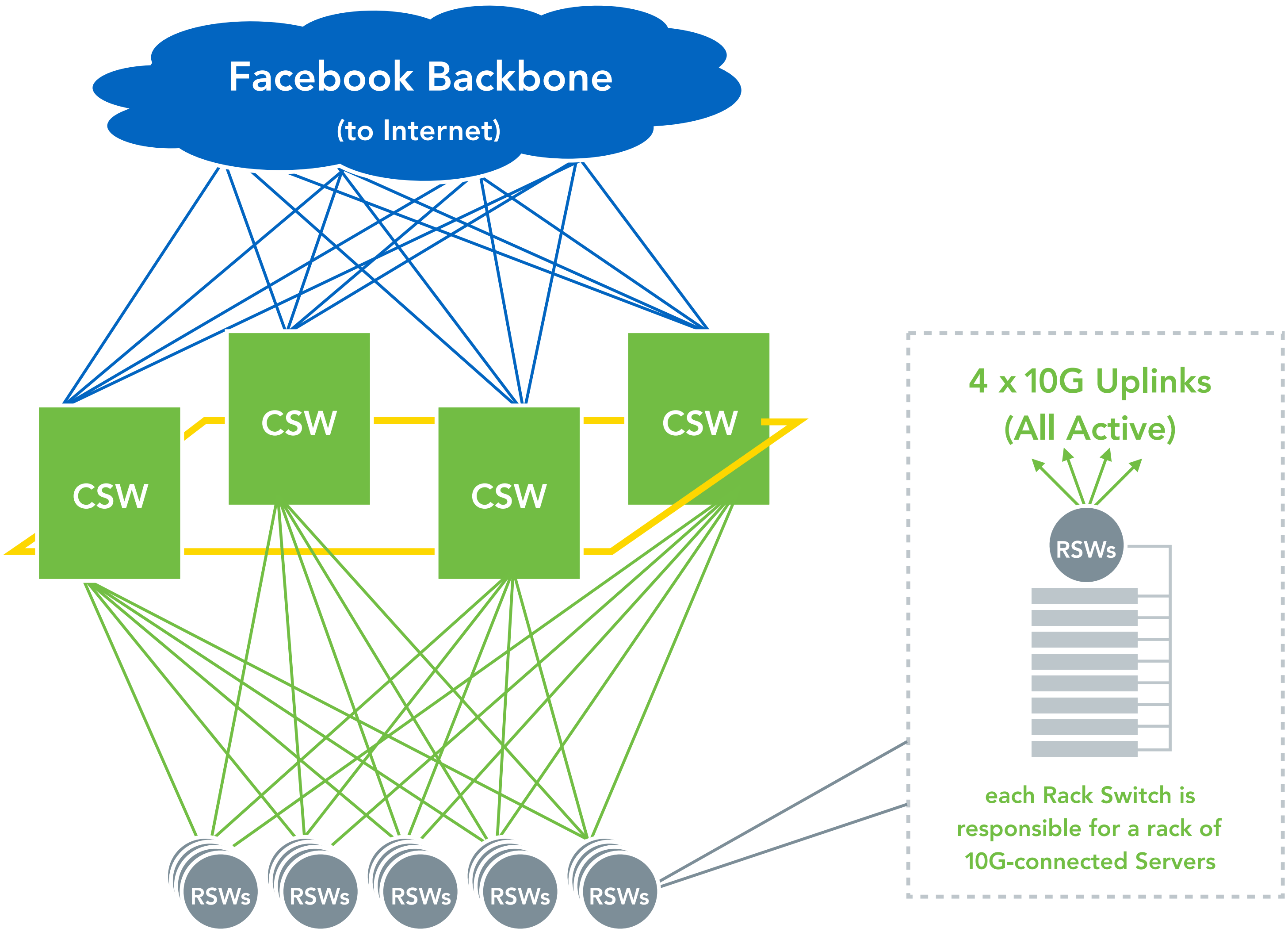
NETWORK ENGINEER

Outline

- Where 6-Pack fits in the DC
- The test topology and tools
- Testing areas and results

Brief history of FB's DC evolution

From the Cluster to the Fabric



4 x 10G Uplinks
(All Active)

each Rack Switch is responsible for a rack of 10G-connected Servers

Disadvantages of Clusters

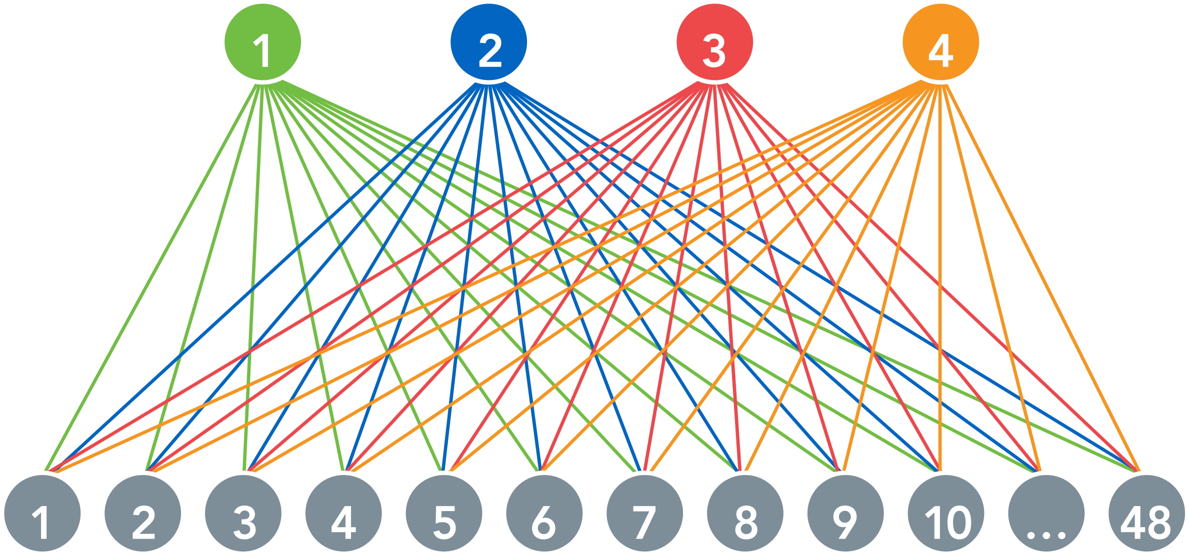
- CSW port capacity versus growth
- CSWs large, complex & few options
- Challenges to achieve higher BW per rack
- 1 CSW failure = 25% less Cluster BW
- Over-subscription inflexible

Introducing the Fabric …

The Server POD

- 4-Post POD
- 48 TORs
- Like a micro-Cluster
- 4x40G uplinks from TOR

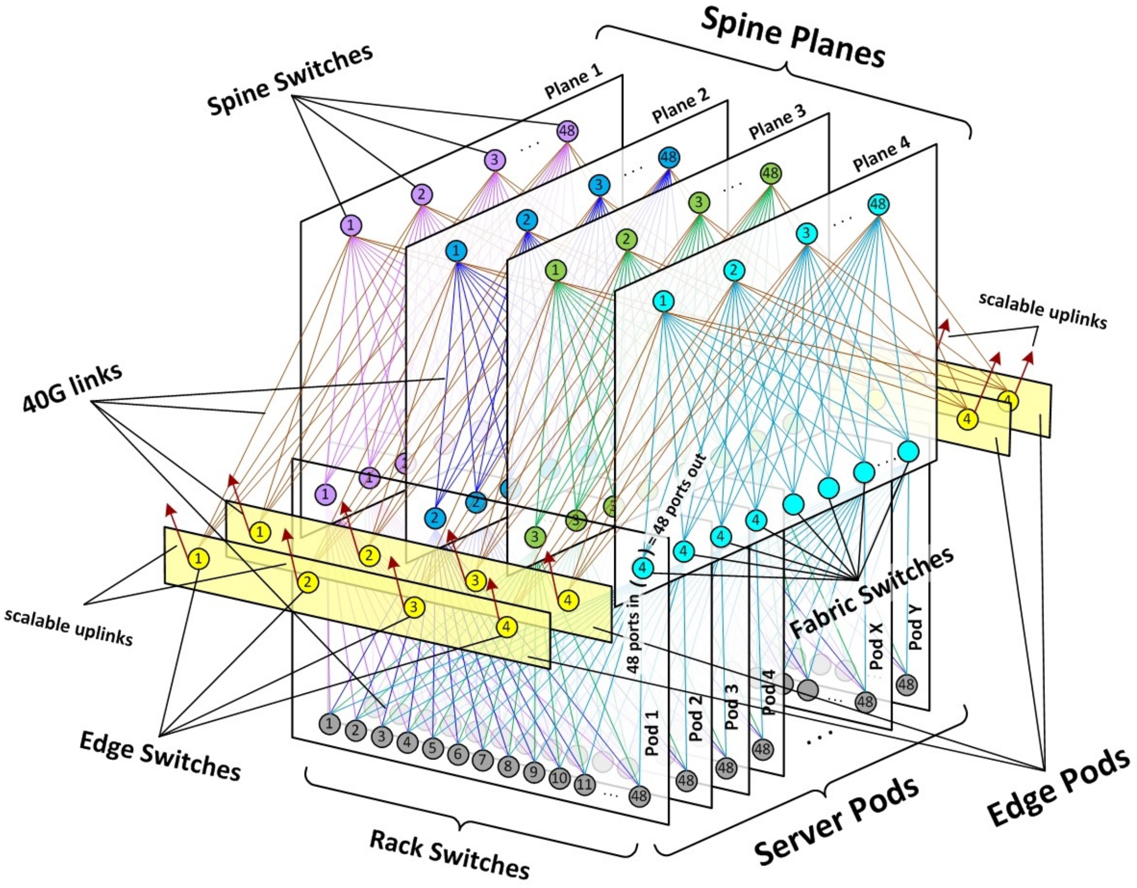
4 fabric switches



48 top of rack switches (TORs)

The whole Fabric

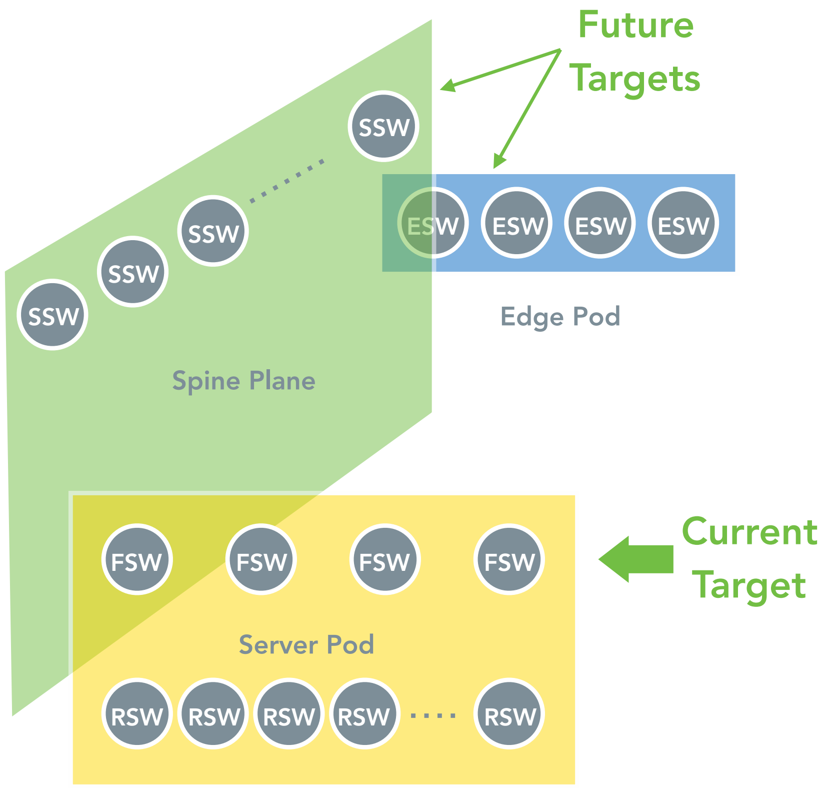
- Server PODs
- Spine Planes
- Edge PODs
- Aggregation



Advantages of the Fabric

- Server POD small repeatable unit
- FSWs small & simple
- Intra-Fabric BW expandability
- External BW expandability
- Maps nicely to data center floor plans

The role of 6-Pack in the Fabric



The FSW Role

- Peers with the RSWs and SSWs
- Controls routes with BGP policies
- Aggregates traffic from RSWs
- Uses ECMP heavily
- Reacts quickly to link failures

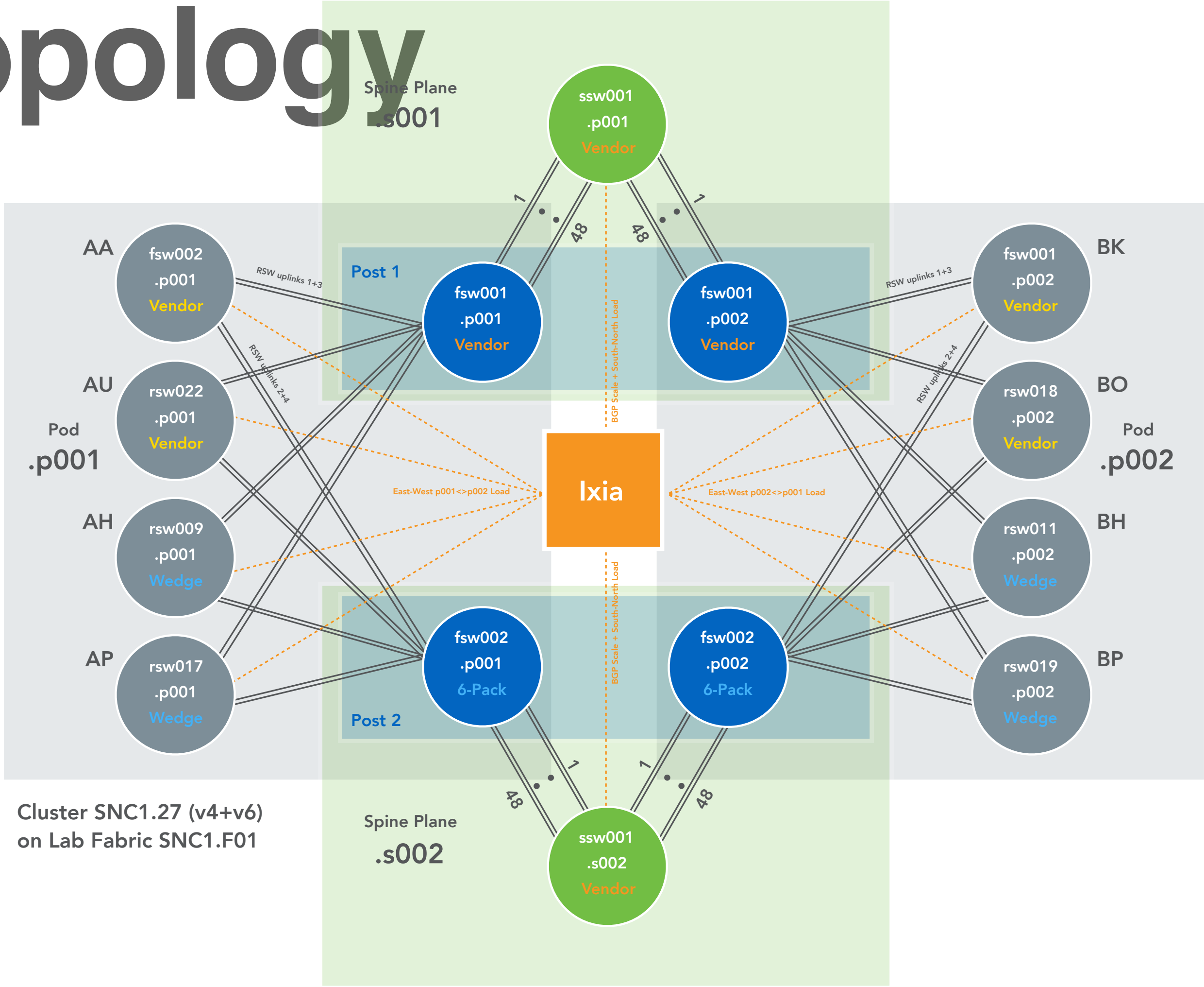
Testing tools for 6-Pack

- Traffic generator
- BGP prefix injector
- Stats collector
- Monitoring dashboards
- Automation scripts
- Benchmark data from vendor platform

The test topology

6-Pack Fabric Dev/Test Lab Topology

- 2 Server PODs
- 2 +2 RSW Uplinks
- 2 Spine Planes
- 48 FSW Uplinks
- Ixia



Cluster SNC1.27 (v4+v6)
on Lab Fabric SNC1.F01

Ixia configuration

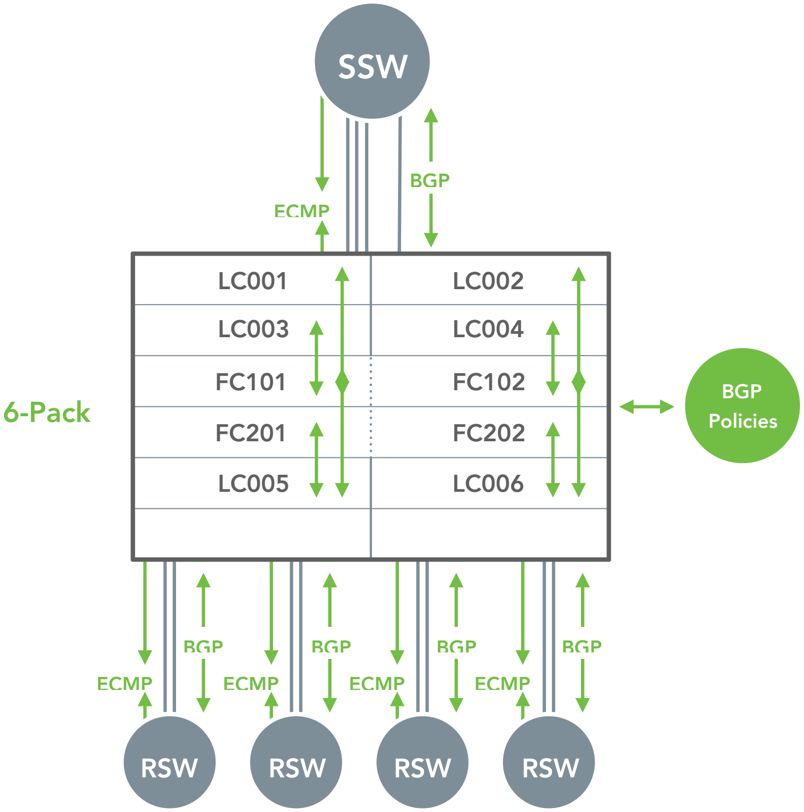
- BGP route scale simulation
  - Total routes 8K V4 + 8K V6 routes in a prefix mix close to our production environments
- Traffic flows
  - 50 simulated servers behind each RSW
  - Traffic Flows are a mesh of 50 Src IPs, 50 Dst IPs with random TCP/UDP ports
  - Both Fixed 400B and Random (74B – 1500B)
  - Bi-directional traffic
  - Intra-Pod traffic (Ixia → RSW → 6-Pack → RSW → Ixia)
  - Inter-Pod traffic (Ixia → RSW → 6-Pack → SSW → 6-Pack → RSW → Ixia)
  - External traffic (Ixia → RSW → 6-Pack → SSW → Ixia)

Testing Areas

- Functionality
- Failure Conditions
- Scale & Stress
- Integration with FB Tools
- Performance
- Longevity

Fuctionality testing

- BGP Peerings between 6-Pack and Wedge and vendor RSWs
- BGP Peerings between 6-Pack and SSW
- BGP Peerings between LCs and FCs
- ECMP over all multi-links at all tiers
- Route policy processing



Sample BGP Output for RSW

Something the technical folks can relate to!

```
fboss -H fsw002-lc001.p001.f01.snc1 bgp neighbors | egrep "Peer|rsw" | grep -v IDLE
```

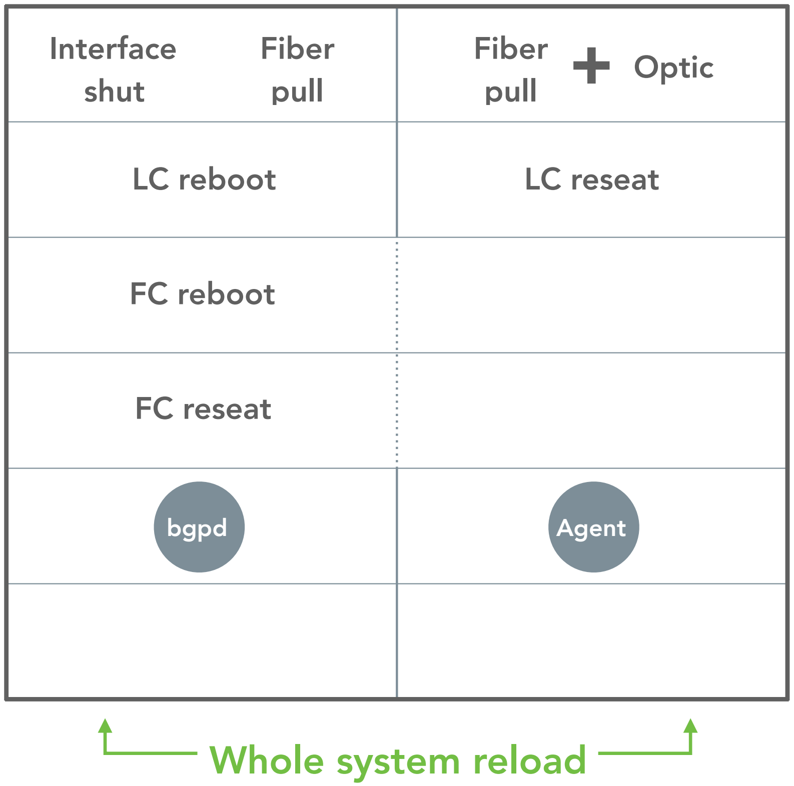
```
Peer IP                       My IP                        Local AS    Remote AS     HoldTime
10.50.44.5                    10.50.44.4                   6002        2002          30
2401:db00:e011:9101:1000::5   2401:db00:e011:9101:1000::4  6002        2002          30

Peer State                    Pfx Rcvd                     Pfx Sent    Desc                   Uptime
ESTABLISHED                   2                            13          rsw002.p001.f01.snc1   33 days, 5:09:28
ESTABLISHED                   2                            12          rsw002.p001.f01.snc1   33 days, 5:09:23
```

Sample BGP output for

```
fboss -H fsw002-lc001.p001.f01.snc1 bgp neighbors | egrep "Peer|FC" | grep -v IDLE

Peer IP              My IP              Local AS   Remote AS   HoldTime  Peer State    Pfx Rcvd  Pfx Sent   Desc       Uptime

169.254.255.101      169.254.255.1      6002       6002        30        ESTABLISHED   6         16         FCv4 1     33 days, 5:10:44
169.254.255.102      169.254.255.1      6002       6002        30        ESTABLISHED   6         16         FCv4 2     33 days, 5:10:44
169.254.255.201      169.254.255.1      6002       6002        30        ESTABLISHED   6         16         FCv4 3     33 days, 5:10:44
169.254.255.202      169.254.255.       6002       6002        30        ESTABLISHED   6         16         FCv4 4     33 days, 5:10:44

fc00::ff:65          fc00::ff:1         6002       6002        30        ESTABLISHED   6         16         FCv6 1     33 days, 5:10:34
fc00::ff:66          fc00::ff:1         6002       6002        30        ESTABLISHED   6         16         FCv6 2     33 days, 5:10:34
fc00::ff:c9          fc00::ff:1         6002       6002        30        ESTABLISHED   6         16         FCv6 3     33 days, 5:10:34
fc00::ff:c           fc00::ff:1         6002       6002        30        ESTABLISHED   6         16         FCv6 4     33 days, 5:10:34
```

Failure conditions, online

- Interface recovery
- LC recovery
- FC recovery
- Processes recovery
- System recovery



Interface shut

Fiber pull

Fiber pull + Optic

LC reboot

LC reseat

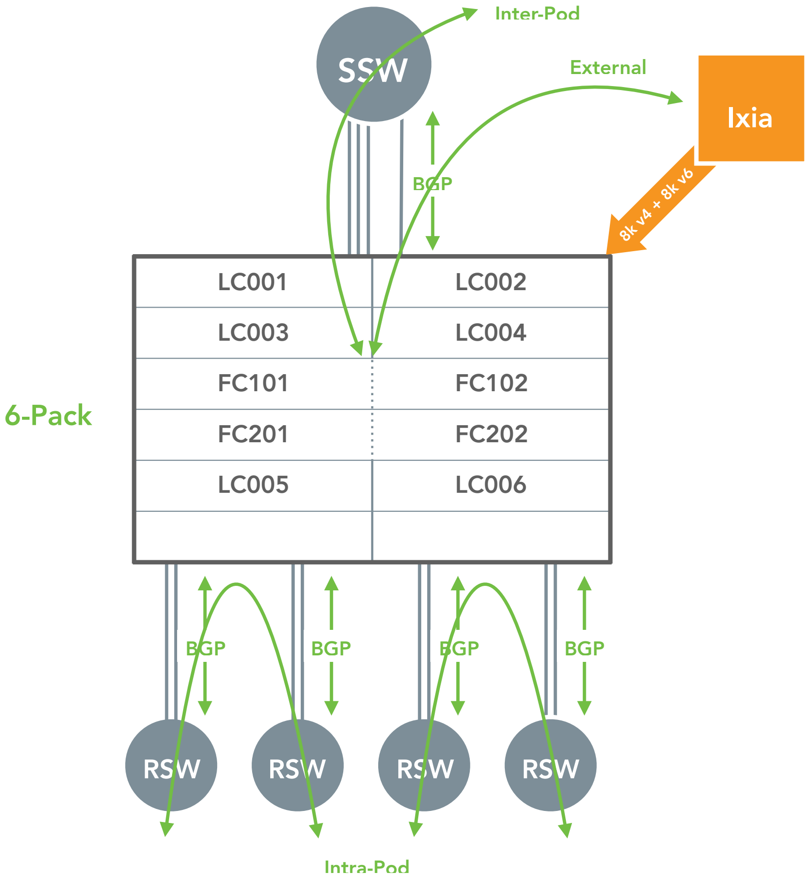
FC reboot

FC reseat

Whole system reload

Scale & stress testing

- Route scale with 8K v4 + 8K v6 prefixes injected from Ixia
- All traffic flows running concurrently
- Flapping all 16K routes every 60s
- High frequency interface shut/unshut
- High frequency BGP neighbor shut/unshut



Integration with FB Tools

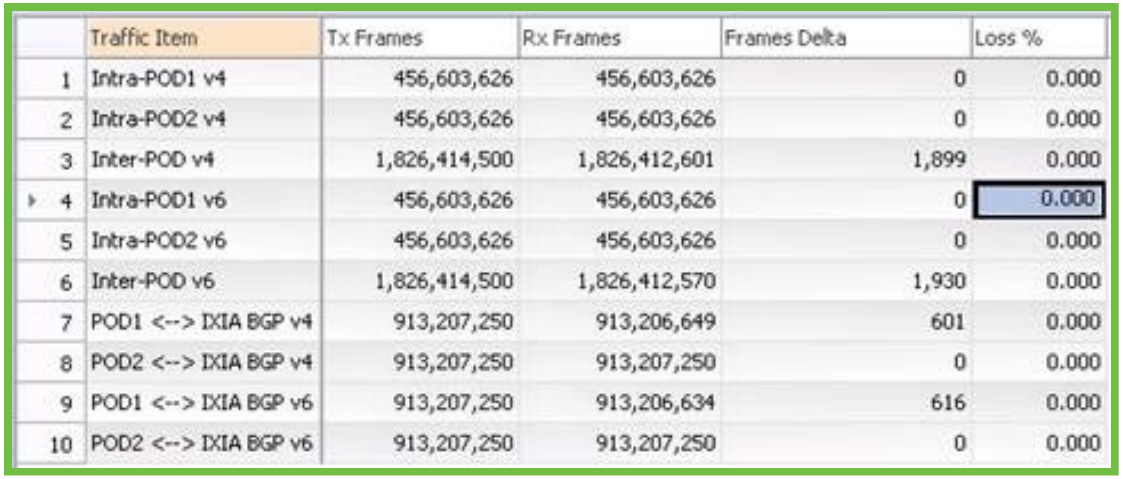
- Config generation
- Auto-provisioning
- Software updates (hitless)
- Monitoring via dashboards
- Auditing and alerting
- Manageability
- Drain/undrain

Performance testing

Measuring the impact of:

- Interface shut on RSW, 6-Pack and SSW
- BGP neighbor shut on RSW and SSW
- BGP failure in indirect connectivity scenario
- FC OIR
- Drain/undrain
- Programmability time of 8K IPv4 + 8K IPv6 routes in the FIB

Sample 1 interface shut



	Traffic Item	Tx Frames	Rx Frames	Frames Delta	Loss %
1	Intra-POD1 v4	456,603,626	456,603,626	0	0.000
2	Intra-POD2 v4	456,603,626	456,603,626	0	0.000
3	Inter-POD v4	1,826,414,500	1,826,412,601	1,899	0.000
4	Intra-POD1 v6	456,603,626	456,603,626	0	0.000
5	Intra-POD2 v6	456,603,626	456,603,626	0	0.000
6	Inter-POD v6	1,826,414,500	1,826,412,570	1,930	0.000
7	POD1 <--> IXIA BGP v4	913,207,250	913,206,649	601	0.000
8	POD2 <--> IXIA BGP v4	913,207,250	913,207,250	0	0.000
9	POD1 <--> IXIA BGP v6	913,207,250	913,206,634	616	0.000
10	POD2 <--> IXIA BGP v6	913,207,250	913,207,250	0	0.000

Calculating the impact in terms of milli-seconds:

With Frame Rate @ 1M FPS, the Inter-POD traffic have suffered

1900 / 1,000,000 ~ **2ms** Awesome!!

Longevity test

- Continuous traffic run with full route scale
- Validates stability & continuity
- Spans weeks

Testing automation

- Overnight scripts to perform the following functions and verify recovery:
  - Shut/unshut interfaces sequentially:
  - Shut/unshut BGP Neighbors
  - Reload LCs
  - Reload FCs
  - Drain/undrain
- Future goal to automate the entire test suite to quickly re-iterate over newer SW images

OPEN
Compute Project