MULTI-STAGE CLOS ARCHITECTURES

Doug Hanks

SR. DATA CENTER ARCHITECT
CDBU
JUNIPER NETWORKS
BIG PICTURE

WAN

Corporate Office

Smart Phone

Cell Tower

Branch Office

Data Centers
DATA CENTER ARCHITECTURE

- Services
- Storage
- Compute
WHY MULTI-STAGE NETWORKS?

- Scale
- No need for Layer 2 between racks
  - Typically non-virtualized and OTT services
- Want to leverage data center overlay technologies
  - VMware NSX
  - Juniper Contrail
LARGE SCALE

Access
Over 10,000 Ports

Layer 3
CHARLES CLOS - 1953

Ingress - Middle - Egress
SPINE AND LEAF

- Spine
- Leaf

Ingress

Middle

Scale

Egress
# CLOS REQUIREMENTS

<table>
<thead>
<tr>
<th>Requirement</th>
<th>OSPF</th>
<th>IS-IS</th>
<th>BGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertise prefixes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Scale</td>
<td>Limited</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td>Traffic Engineering</td>
<td>Limited</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td>Traffic Tagging</td>
<td>Limited</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td>Multi-Vendor Stability</td>
<td>Yes</td>
<td>Yes</td>
<td>Even more so</td>
</tr>
</tbody>
</table>

- OSPF: Yes
- IS-IS: Yes
- BGP: Yes, Even more so
COMPONENTS

- Spine
- Leaf
- ASN 1
- ASN 2
- ASN 3
- ASN 4
- ASN 5
- ASN 6

eBGP connections between the components.
DETAILS

- BGP ASN assignments
- IP address scheme
- P2P network assignments
- P2P address assignments
- Server-facing network assignments
- BGP export policies
EXAMPLE

4 spines + 16 leaves = 64 links
PROBLEM AT SCALE

P2P Link Scale

Number of Leaves

P2P Links

Links
FUNCTIONS

sub next_subnet
{
    my ($network, $subnet) = @_; 
    my $net = unpack 'N', pack 'CCCC', split '/\.', $network; 
    my $step = 1 << 32 - $subnet; 
    my $next = $net + $options{p2p_mask}; 
    $net += $step; 
    my $next_ip = sprintf "%x%d", pack 'N', $net; 
    return $next_ip; 
}

sub generate_topology
{
    my $spine_int = $options{spine_speed} == 10 ? 'xe' : 'et'; 
    my $block = new Net::Netmask($options{base_prefix}); 
    my $block_base = $block->base(); 
    my $sp2p = $block_base; 
    my $as = $options{as}; 
    for(my $s = 0; $s < $options{spines}; $s++)
    {
        $spine{[$s]}{name} = "$s$s"; 
        $spine{[$s]}{as} = $as++; 
        for(my $l = 0; $l < $options{leaves}; $l++)
        {
            my $ip = NetAddr::IP::Lite->new("Sp2p/$options{p2p_mask}"); 
            my $ip_first = sprintf("%s", $ip->first()); 
            my $ip_last = sprintf("%s", $ip->last()); 
        }
    }
}
USAGE


- h, --help
  Display this help menu.

- spines
  Number of spines. This cannot be used with --leaves
  as it will automatically calculate the number of spines
  Example: --spines=4

- leaves
  Number of leaves. This cannot be used with --spines
  as it will automatically calculate the number of leaves
  Example: --leaves=16

- spine-speed
  Speed of network ports in spine (1,10,40,100)
  Example: --spine-speed=40

- spine-ports
  Number of network ports in spine
  Example: --spine-ports=16

- leaf-down-speed
  Speed of server-facing ports on the leaf (1,10,40,100)
  Example: --leaf-down-speed=10

- leaf-down-ports
  Number of server-facing ports
  Example: --leaf-down-ports=48

- leaf-up-speed
  Speed of spine-facing ports on the leaf (1,10,40,100)
  Example: --leaf-up-speed=40

- leaf-up-ports
  Number of spine-facing ports
  Example: --leaf-up-ports=4

- a, --autonomous-system
  Where to begin assigning autonomous system numbers.
  Example: --autonomous-system=64512

- b, --base-prefix
  Base prefix and mask to be used to assign L3 addresses.
  Example --base-prefix=192.168/16

- p, --p2p-mask
  Network mask to use when calculating p2p networks.
  Value must be between 24 and 31.
  Example --p2p-mask=30
EXAMPLE

localhost:SL dhanks$ ./SL --spines=4 \
    --spine-speed=40 \
    --spine-ports=32 \
    --leaf-down-speed=10 \
    --leaf-down-ports=48 \
    --leaf-up-ports=4 \
    --leaf-up-speed=40 \
    --base-prefix=192.168/16 \
    --p2p-mask=31 \
    --autonomous-system=65000
```
"leaf" : {
    "0" : {
        "as" : 65001,
        "name" : "L0",
        "port" : {
            "0" : {
                "address" : "192.168.0.1/31",
                "ifd" : "et-0/0/0"
            },
            "1" : {
                "address" : "192.168.0.65/31",
                "ifd" : "et-0/0/1"
            },
            "2" : {
                "address" : "192.168.0.129/31",
                "ifd" : "et-0/0/2"
            },
            "3" : {
                "address" : "192.168.0.193/31",
                "ifd" : "et-0/0/3"
            }
        }
    },
    "1" : {
        "as" : 65002,
        "name" : "L1",
        "port" : {
            "0" : {
                "address" : "192.168.0.3/31",
                "ifd" : "et-0/0/0"
            },
            "1" : {
                "address" : "192.168.0.67/31",
                "ifd" : "et-0/0/1"
            },
            "2" : {
                "address" : "192.168.0.131/31",
                "ifd" : "et-0/0/2"
            },
            "3" : {
                "address" : "192.168.0.195/31",
                "ifd" : "et-0/0/3"
            }
        }
    }
}
```
DOWNLOAD

https://github.com/dhanil
MULTI-STAGE
FUN WITH
CLOS
MULTI-STAGE CLOS ROLES

Spine
- Backplane of multi-stage CLOS
- Always 1:1 Over-Subscription
- Provide BGP Route Reflection
- Combination of Spine and Leaf
- Acts as a logical switch
- Virtual peering point for Access
- Over-Subscription dependent on the Spine and Leaf roles
- Single BGP Autonomous System Number
- Peers via eBGP to access switches

Leaf
- Peers via iBGP to Leaf nodes of multi-stage CLOS
- Variable Over-Subscription
- Peers via iBGP to Spine nodes
- Peers via eBGP to Access nodes
- Provide access to end-points such as compute and storage
- Typically 3:1 Over-Subscription in ENT and SP environments, and 1:1 for HPC
- Peers via eBGP to vSpine nodes
- Provides L3 gateway services to end-points
- Provides Link Aggregation to end-points

vSpine
- Combination of Spine and Leaf
- Acts as a logical switch
- Virtual peering point for Access
- Over-Subscription dependent on the Spine and Leaf roles
- Single BGP Autonomous System Number
- Peers via eBGP to access switches
MULTI-STAGE CLOS BENEFITS

- Massive scale – over 73,000x10GE access ports
- High performance – variable over-subscription 1:1 to N:1
- Pay as you grow – start small and increment 1U at a time
- Low latency with fixed switches
- Very small “blast radius” upon failures in the network
- Standards based design – supports multiple vendors
- Deterministic latency with a fixed spine and leaf topology
MULTI-STAGE CLOS EXAMPLE
MULTI-STAGE CLOS EXAMPLE

vSpine

vSpine

vSpine

96x10GE
MULTI-STAGE CLOS ARCHITECTURE

- **Spine**: 1:1 OS
  - vSpin e1
  - vSpin e2
  - vSpin e3
  - vSpin e4
  - vSpin e5
  - vSpin e6
  - vSpin e7
  - vSpin e8

- **Leaf**: 1:1 OS
  - Interfaces 512x40GE
  - 1:1 OS

- **Access**: 3:1 OS
  - 96x10G
  - E
MULTI-STAGE CLOS DETAIL

Spine
16 nodes per vSpine

vSpine1 – 512x40GE 1:1 O/S
vSpine2 – 512x40GE 1:1 O/S

Leaf
32 nodes per vSpine

Access
512 nodes
49,152x10GE
### VSPINE OPTIONS

#### 1:1 OVER-SUBSCRIPTION

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Virtual Chassis Fabric</th>
<th>CLOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total 40GE Interfaces</td>
<td>128</td>
<td>512</td>
</tr>
<tr>
<td>Spines</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Leaves</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>Total Switches</td>
<td>12</td>
<td>48</td>
</tr>
<tr>
<td>Point-to-Point Connections</td>
<td>128</td>
<td>512</td>
</tr>
<tr>
<td>Latency</td>
<td>1.6µ</td>
<td>1.6µ</td>
</tr>
<tr>
<td>Control Planes</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>Blast Radius</td>
<td>n</td>
<td>1/n</td>
</tr>
</tbody>
</table>
VSPINE TOPOLOGIES
1:1 OVER-SUBSCRIPTION

Virtual Chassis Fabric vSpine
- 32x40G E
- 4x40GE
- 32x40G E
- 32x40G E
- 32x40G E
- ... 8 Leaves ...
- 16x40GE per Leaf
- 128x40GE per vSpine
- Single Point of Management
- 1:1 Over-Subscription

L3 CLOS vSpine
- 32x40G E
- 1x40GE
- ... 16 Spines ...
- 16x40GE per Leaf
- 512x40GE per vSpine
- 1:1 Over-Subscription
## VSPINE OPTIONS
### 3:1 OVER-SUBSCRIPTION

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Virtual Chassis Fabric</th>
<th>CLOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total 40GE Interfaces</td>
<td>384</td>
<td>768</td>
</tr>
<tr>
<td>Spines</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Leaves</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Total Switches</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Point-to-Point Connections</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>Latency</td>
<td>1.6µ</td>
<td>1.6µ</td>
</tr>
<tr>
<td>Control Planes</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>Blast Radius</td>
<td>n</td>
<td>1/n</td>
</tr>
</tbody>
</table>
VSPINE TOPOLOGIES
3:1 OVER-SUBSCRIPTION

Virtual Chassis Fabric vSpine
- 384x40GE per vSpine
- Single Point of Management
- 3:1 Over-Subscription

L3 CLOS vSpine
- 768x40GE per vSpine
- 3:1 Over-Subscription

24x40GE per Leaf
... 16 Leaves ...
... 8 Spines ...

24x40GE per Leaf
... 32 Leaves ...
### MULTI-STAGE CLOS TOPOLOGY

<table>
<thead>
<tr>
<th>Level</th>
<th>Spine Type</th>
<th>Capacity</th>
<th>OS Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>vSpine VCF</td>
<td>384x40GE</td>
<td>3:1 OS</td>
</tr>
<tr>
<td>2</td>
<td>vSpine CLOS</td>
<td>768x40GE</td>
<td>3:1 OS</td>
</tr>
<tr>
<td>3</td>
<td>vSpine VCF</td>
<td>128x40GE</td>
<td>1:1 OS</td>
</tr>
<tr>
<td>4</td>
<td>vSpine CLOS</td>
<td>512x40GE</td>
<td>1:1 OS</td>
</tr>
</tbody>
</table>

- **Spine**: vSpine
- **Leaf**: vSpine CLOS
- **Access**: vSpine VCF

**OS**: 3:1 OS

**Networks**: 1:1 OS

**Connectivity**: 40GE

- **Access**: 96x10GE
- **Leaf**: 96x10GE
- **Spine**: 96x10GE

**Note**: The table represents the connectivity and network allocation for different stages in a multi-stage CLOS topology.
<table>
<thead>
<tr>
<th></th>
<th>1-Line</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3:1 OS vSpine</td>
<td>384x40GE vSpine</td>
</tr>
<tr>
<td></td>
<td>VCF vSpine</td>
<td>3W/10GE port</td>
</tr>
<tr>
<td>2</td>
<td>3:1 OS vSpine</td>
<td>768x40GE vSpine</td>
</tr>
<tr>
<td></td>
<td>CLOS vSpine</td>
<td>3W/10GE port</td>
</tr>
<tr>
<td>3</td>
<td>1:1 OS vSpine</td>
<td>128x40GE vSpine</td>
</tr>
<tr>
<td></td>
<td>VCF vSpine</td>
<td>3.6W/10GE port</td>
</tr>
<tr>
<td>4</td>
<td>1:1 OS vSpine</td>
<td>512x40GE vSpine</td>
</tr>
<tr>
<td></td>
<td>CLOS vSpine</td>
<td>3.6W/10GE port</td>
</tr>
</tbody>
</table>

- **OS vSpine**: OS vSpine
- **CLOS vSpine**: CLOS vSpine
- **VCF vSpine**: VCF vSpine
- **384x40GE vSpine**: 384x40GE vSpine
- **768x40GE vSpine**: 768x40GE vSpine
- **128x40GE vSpine**: 128x40GE vSpine
- **512x40GE vSpine**: 512x40GE vSpine
- **3W/10GE port**: 3W/10GE port
- **3.6W/10GE port**: 3.6W/10GE port
- **9:1 E2E OS**: 9:1 E2E OS
- **3:1 E2E OS**: 3:1 E2E OS
MULTI-STAGE CLOS BGP OVERVIEW

- Spine
- Leaf
- Access

BGP RR Cluster

BGP ASN 1

BGP ASN 11

vSpine

iBGP

BFD

eBGP

BFD
MULTI-STAGE CLOS BGP DETAIL

Spine
- BGP RR
- iBGP Down

vSpine1 – ASN 1
- BGP RR
- BGP RR

vSpine2 – ASN 2
- BGP RR
- BGP RR

Leaf
- iBGP Up
- eBGP Down

32x40G
- E

BGP
- RR

iBGP

Access
- eBGP Up

96x10G
- E

ASN
- 11 12 13 14 15 16 17 18

E
# VSPINE STRIPING

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SUMMARY

- vSpine = collapsed spine + leaf
- Multiple vSpine options
  - iBGP controlled 3-stage Clos
  - Virtual Chassis Fabric
  - Physical Chassis
- Multi-Stage Clos reduces the blast radius
- BGP glues it all together
- BFD for high availability
LET'S BUILD THE BEST