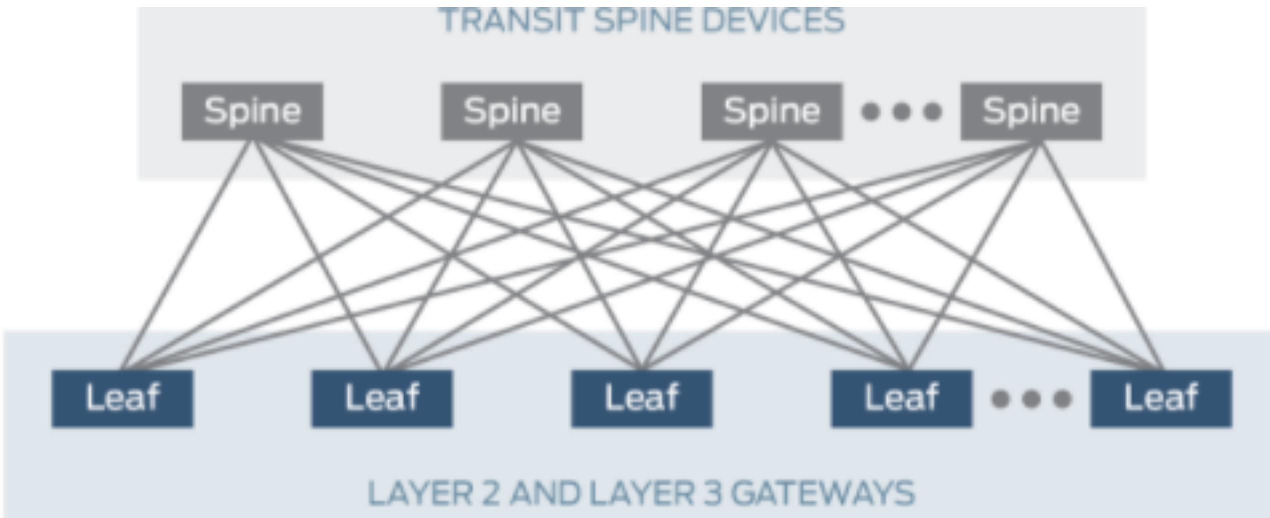
ZCORP

Data Center Design - Project Proposal



Prepared for: Steven Craig
Prepared by: Michael Osman
August 21, 2017

Objective

The objective is to design a data center network that will host the services provided by Zcorp. There are four cabinets available in the datacenter, two upstream internet providers (ISP1 and ISP2) providing 1Gbps of bandwidth each through two Ethernet cables (one for each provider) left at the top of one of the cabinets.

Goals

This document will present the following:

1. An inventory of all networking hardware required, including routers, switches, firewalls, and load balancers (Table 1).
2. A visual diagram of the architecture showing an overview of the network edge (Figure 1).
3. The VLAN configuration for the network indicating membership for each server/group of servers (Table 2).
4. Design of the IP addressing scheme (Table 2) and naming scheme of all the servers/network gear (Table 3); an explanation is also offered regarding routing.
5. An explanation of how the network reacts in different failure scenarios (e.g., core router goes down, aggregation switch goes down, ToR switch goes down, etc.).
6. An explanation of how the network is multi-homed to the two upstream ISPs.

Assumptions

A. Some applications will run in the container cluster and some will run in the Xen virtual machines.

B. The application is a 2-tier app with a pool of web servers connected to the database shards.

C. The Production environment will contain the following:

- 20 bare metal database shards
- 5 bare metal running kafka
- 5 bare metal backup servers
- 5 bare metal file servers running glusterfs
- 10 bare metal servers forming a kubernetes cluster
- 5 bare metal servers running Xen

D. The Staging environment will contain the same server roles as in the Production environment but only 2 servers for each role.

Networking Hardware

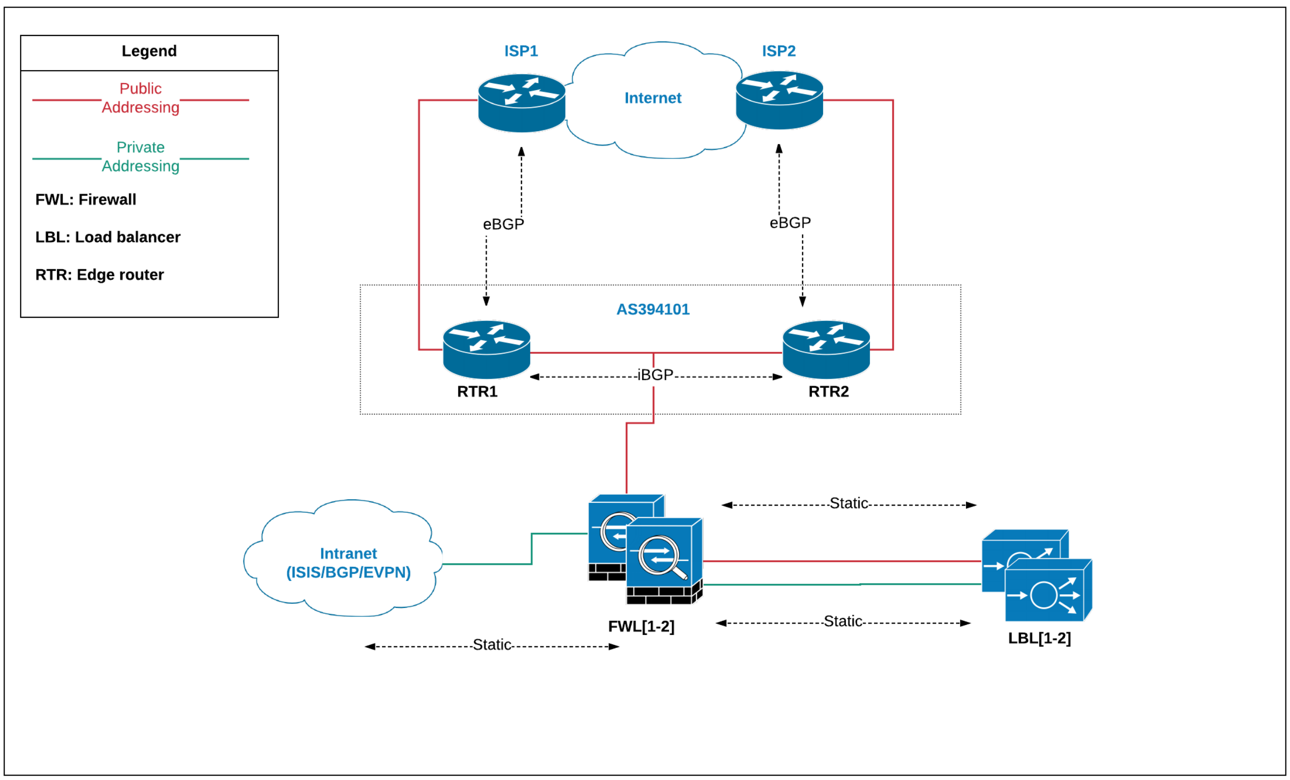
Table 1 - proposed hardware for each type of network device

Role	Mfr	Model	URL	Qty	Location
Edge router	Juniper	MX104	http://www.juniper.net/us/en/products-services/routing/mx-series/mx104/index.page#overview	2	Cab 1
Leaf switch	Juniper	QFX10002-72Q	http://www.juniper.net/us/en/products-services/switching/qfx-series/qfx10000/	8	2 @ Cab 1 - 4
Spine switch	Juniper	QFX10002-72Q	http://www.juniper.net/us/en/products-services/switching/qfx-series/qfx10000/	2	Cab 1
Firewall	Juniper	SRX4100	http://www.juniper.net/us/en/products-services/security/srx-series/srx4000/#overview	2	Cab 1

Role	Mfr	Model	URL	Qty	Location
Load balancer	F5	i2800	https://www.f5.com/pdf/products/big-ip-platforms-datasheet.pdf	2	Cab 1

Networking Diagram

Figure 1 - Network edge overview



VLAN and Subnet Assignment

Table 2 - VLAN and subnet assignment for each server type

VLAN #	VLAN name	Membership	Environment	IP Subnet
100	PROD_KAFKA	Kafka servers	PROD	10.0.100.0/24
101	PROD_BACKUP	Backup servers	PROD	10.0.101.0/24
102	PROD_GLUSTERFS	Glusterfs servers	PROD	10.0.102.0/24
103	PROD_KUBERNETES	Kubernetes hosts	PROD	10.0.103.0/24
104	PROD_XEN	Xen hosts	PROD	10.0.104.0/24
200	PROD_WEB	Web servers	PROD	10.0.200.0/24
210	PROD_DB	Database shards	PROD	10.0.210.0/24
1100	STG_KAFKA	Kafka servers	STG	10.1.100.0/24
1101	STG_BACKUP	Backup servers	STG	10.1.101.0/24
1102	STG_GLUSTERFS	Glusterfs servers	STG	10.1.102.0/24
1103	STG_KUBERNETES	Kubernetes hosts	STG	10.1.103.0/24
1104	STG_XEN	Xen hosts	STG	10.1.104.0/24
1200	STG_WEB	Web servers	STG	10.1.200.0/24
1210	STG_DB	Database shards	STG	10.1.210.0/24

Naming Scheme

We propose the following naming convention:

XXX-XXX[X]-XXX-X[X]

XXX: 3-letter data center designation (e.g. DC1).

XXX[X]: 3- or 4- letter environment designation (e.g. PROD or STG).

XXX: 3-letter role designation (e.g. WEB).

X[X]: A number valued 1-99 indicating the host iteration (e.g. 1).

Table 3 - Naming scheme

Datacenter	Environment	Role	Role code	Example
DC1	PROD	Backup server	BAC	DC1-PROD-BAC-1
DC1	PROD	Database shard	DBS	DC1-PROD-DBS-1
DC1	PROD	Firewall	FWL	DC1-PROD-FWL-1
DC1	PROD	Glusterfs server	GLU	DC1-PROD-GLU-1
DC1	PROD	Kafka server	KAF	DC1-PROD-KAF-1
DC1	PROD	Kubernetes host	KUB	DC1-PROD-KUB-1
DC1	PROD	Load balancer	LBL	DC1-PROD-LBL-1
DC1	PROD	Network edge router	RTR	DC1-PROD-RTR-1
DC1	PROD	Spine switch	SPN	DC1-PROD-SPN-1
DC1	PROD	Leaf/TOR switch	TOR	DC1-PROD-TOR-1
DC1	PROD	Web server	WEB	DC1-PROD-WEB-1
DC1	PROD	Xen host	XEN	DC1-PROD-XEN-1
DC1	STG	Backup server	BAC	DC1-STG-BAC-1
DC1	STG	Database shard	DBS	DC1-STG-DBS-1
DC1	STG	Firewall	FWL	DC1-STG-FWL-1

Datacenter	Environment	Role	Role code	Example
DC1	STG	Glusterfs server	GLU	DC1-STG-GLU-1
DC1	STG	Kafka server	KAF	DC1-STG-KAF-1
DC1	STG	Kubernetes host	KUB	DC1-STG-KUB-1
DC1	STG	Load balancer	LBL	DC1-STG-LBL-1
DC1	STG	Network edge router	RTR	DC1-STG-RTR-1
DC1	STG	Spine switch	SPN	DC1-STG-SPN-1
DC1	STG	Leaf/TOR switch	TOR	DC1-STG-TOR-1
DC1	STG	Web server	WEB	DC1-STG-WEB-1
DC1	STG	Xen host	XEN	DC1-STG-XEN-1

Routing

At the network edge, we propose to use BGP for routing. An eBGP peering session will be established with each of the two internet providers, and an iBGP peering session will be established between the two internet edge routers. We will instruct our providers to advertise their full internet routing tables to us, while we will advertise to them our public prefixes. To prevent our autonomous system from carrying transit traffic, we will not re-advertise any prefix we receive from one provider to the other provider. We will not accept a default route from either of the internet providers.

Within the data center, we will run a Layer-3 fabric, using the Clos (leaf-spine) architecture. IS-IS will be used as the underlay routing protocol, while iBGP will be used on the overlay network. Each leaf switch will act as a virtual tunnel endpoint (VTEP). EVPN will be used as a control plane protocol to permit reachability of hosts between leaf switches, and VXLAN will be used for data plane encapsulation.

The firewalls and load-balancers will strictly rely on static routing. This is in an effort to reduce complexity on those appliances, and hence ease troubleshooting. The two edge routers will each have a static route toward 10.0.0.0/16 via the virtual address of the firewall cluster. The firewall cluster will have a default route toward a virtual address that is configured on the two edge routers (note: by configuration, we make edge router 1 the holder of this virtual address, i.e. VRRP master). Router 1 (RTR1 in Figure 1) will forward internet-bound traffic (received from the firewall) using the BGP best path algorithm - i.e. router 1 will decide to either forward the traffic to ISP1 or indirectly to ISP2 via router 2 (RTR2 in Figure 1).

Failure Scenarios

If edge router 1 fails, then edge router 2, in less than 1 second, will become VRRP master; the firewall will still be able to forward internet-bound traffic via its default gateway (the VRRP master's ip address).

If a spine switch fails, reachability between the VTEP's is not lost, by virtue of the full-mesh tree in the underlay. If we configure BFD between the switches, then a spine switch's failure will be detected by its downstream leaf switches in less than 1 second. IS-IS will immediately remove from the routing table the failed switch as a transit node.

We assume that each host is dual-homed to its two top-of-rack switches in an active/backup configuration (Mode 1). If the leaf switch connected to the host's active interface fails, then the host immediately detects this failure and enables the backup interface.

Multi-homing to Two ISP's

As shown in Figure 1, each edge router connects to one upstream internet provider. We rely on BGP to decide on the best path toward any particular internet destination, using either ISP1 or ISP2. Further, due to the dynamic nature of BGP, a failure of an internet circuit or an edge router is seamless to the user.