

Refresh DNS Infrastructure with Modern Datacenter Network

KAWAKAMI KENTO, VERDA NETWORK DEVELOPMENT TEAM, LINE CORPORATION

Agenda

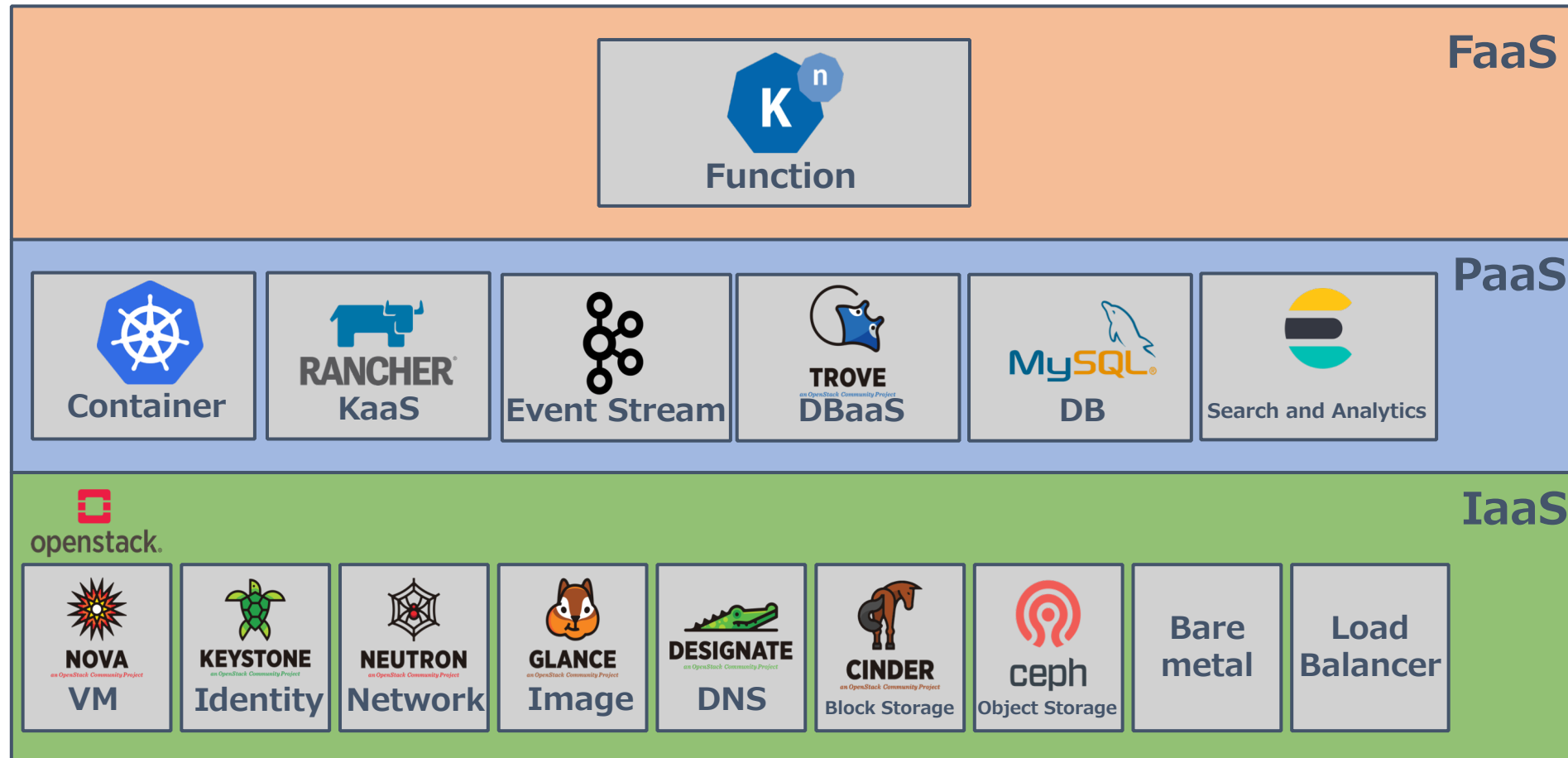
- LINE Private Cloud
- Previous DNS redundancy methods
- IP ANYCAST with CLOS Network

- Kawakami Kento
- LINE Corporation
 - Network Development Team
- LINE New grad in 2019
- Network software engineer
 - DNS
 - OpenStack

LINE Private Cloud



Our Services



DNS as a Service

 **LINE Verda** Prod

Project Verda DNS

Product DNS



 **DNS**

My Recordset

Search

No-owner Recordset

Approval

Support

 Notice

 API Doc

 Documents

OpenStack v3 RC

Go to Help Verda

Verda DNS / DNS / My Recordset

My Recordset

Approval Policy Documents

Create Recordset +

Zone

Name

Type

Data

TTL

Status

Created at

☐

naver.jp.

ns2.naver.jp.

A

147.92.162.216

600

Active

2017-05-11 15:13:12 (JST)

☐

naver.jp.

ns1.naver.jp.

A

147.92.152.224

600

Active

2017-05-11 15:06:10 (JST)

☐

linecorp.com.

devns2.linecorp.com.

A

147.92.162.217

600

Active

2017-05-11 15:04:49 (JST)

☐

linecorp.com.

devns1.linecorp.com.

A

147.92.152.225

600

Active

2017-05-11 15:04:47 (JST)

50 / page

< 1 >

LINE

5

DNS Architecture

Designate

- ユーザからのレコード操作を受け付ける
- ゾーンのマスターデータを保持してる

DNS Relay

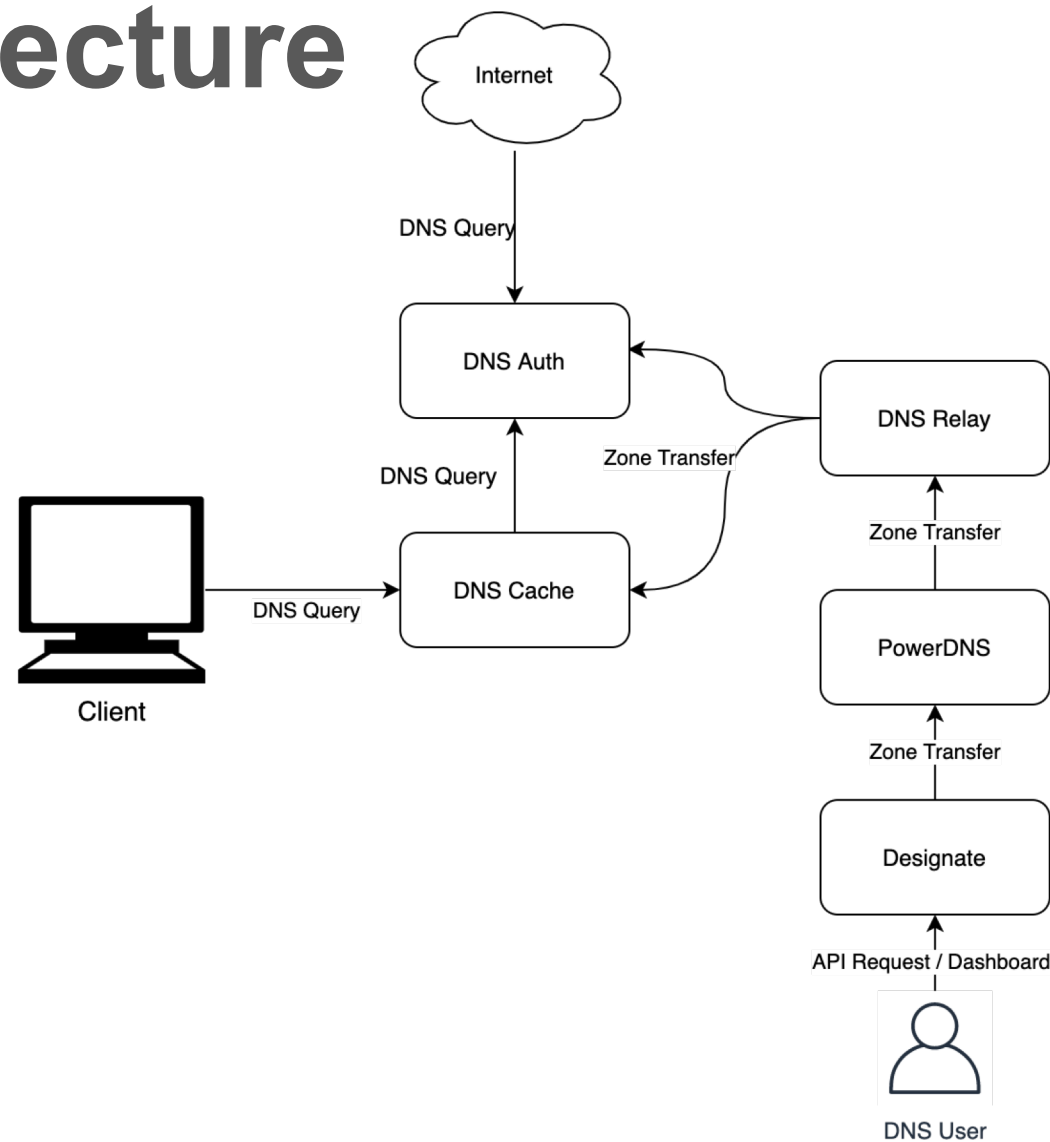
- DesignateからDNS CacheやDNS Authへのゾーン転送の仲介を行う

DNS Auth

- 外部からの権威サーバへのリクエストに応答する

DNS Cache

- Full resolverとして動作しており、一部のホスト名など内部でのみ名前解決を行いたいゾーンをホストしている



DNS Architecture

Designate

- ユーザからのレコード操作を受け付ける
- ゾーンのマスターデータを保持してる

DNS Relay

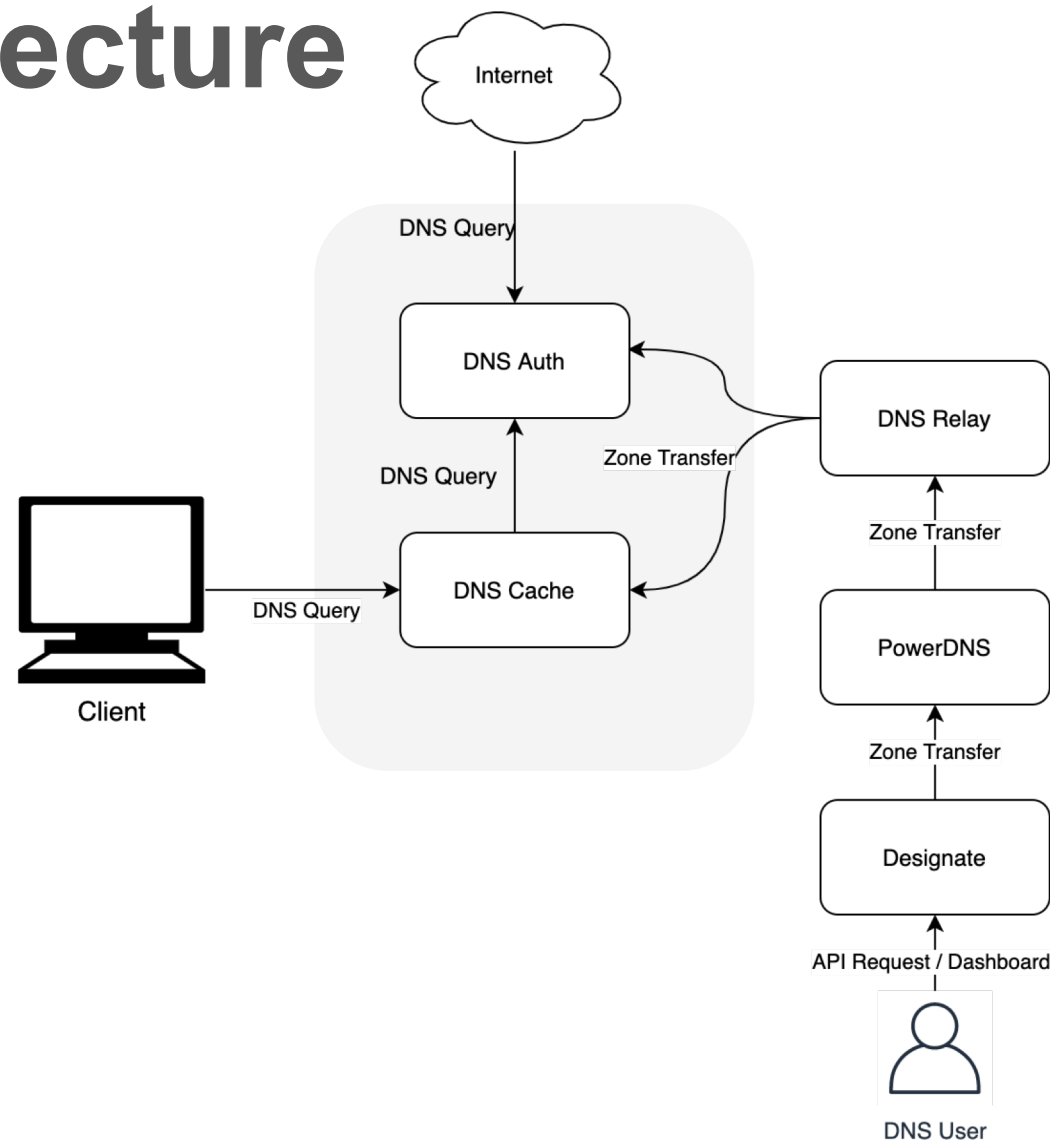
- DesignateからDNS CacheやDNS Authへのゾーン転送の仲介を行う

DNS Auth (VRRP => IP ANYCAST)

- 外部からの権威サーバへのリクエストに応答する

DNS Cache (HWLB => IP ANYCAST)

- Full resolverとして動作しており、一部のホスト名など内部でのみ名前解決を行いたいゾーンをホストしている

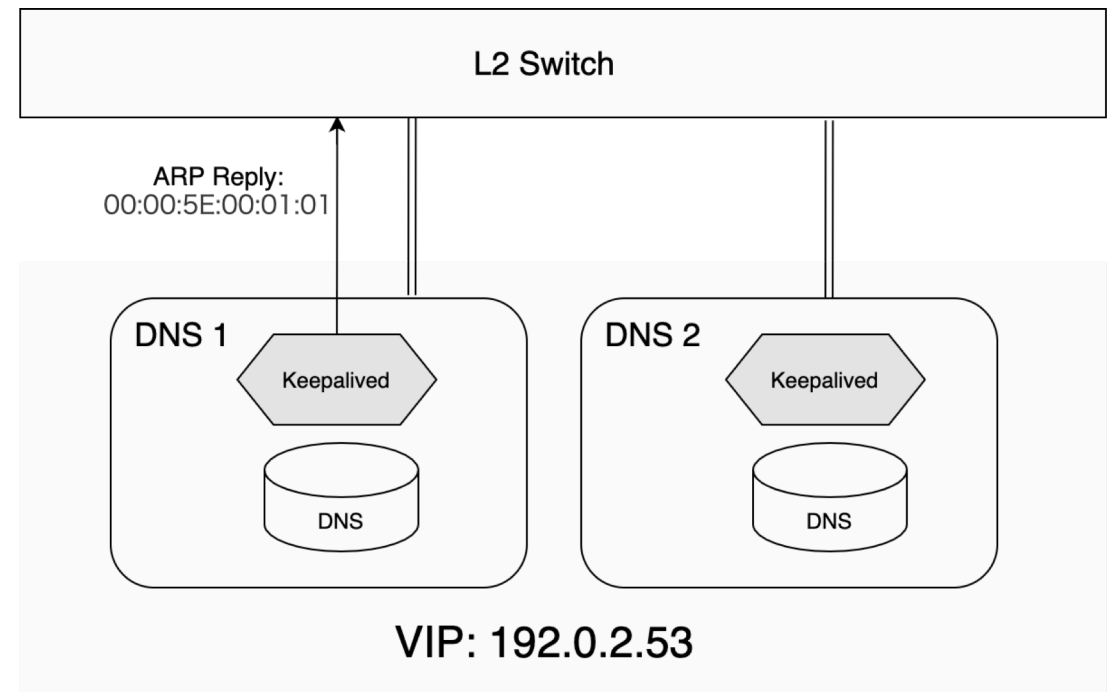


Background of the replacement

- マニュアルオペレーションの排除
 - ゾーンなどの追加の度にDNSサーバに手動でログインしnamed.confの修正を行っていた
 - 構成のコード化したい
- Physical Machineの排除
 - PMはサーバのリプレイス時のリードタイムが大きい
 - Verda上のVMとして構築したい
- DNS Auth/Cacheの冗長化手法の統一
 - メンテナンスの手順などを個別に作成する必要がある
 - 冗長化手法の統一を行いたい
- 新しいインフラの利用
 - CLOS Network上に構築したい
 - CLOS Networkのメリットを享受できる構成にしたい

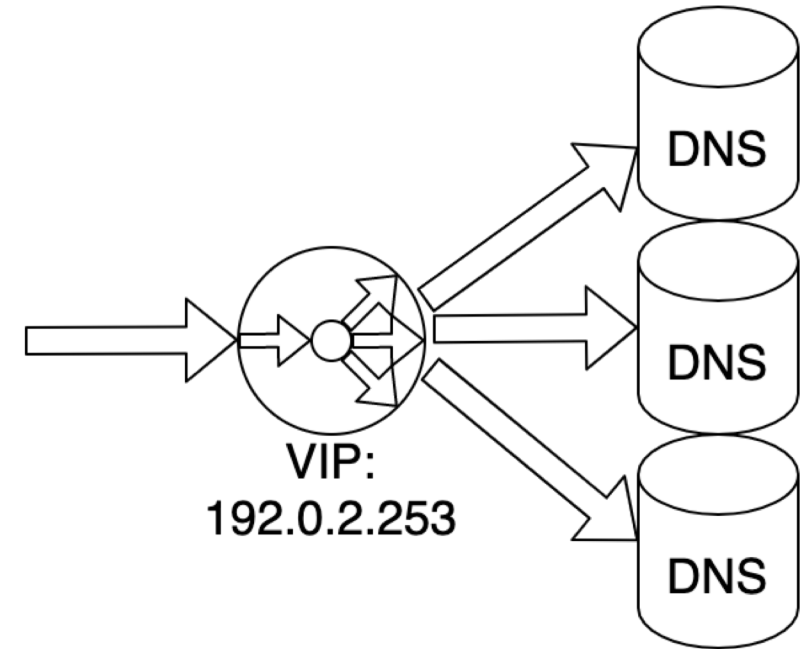
VRRP(Virtual Router Redundancy Protocol)

- VIPに対するAPR Requestに対して Primaryとなっているサーバが応答を行う事で、冗長化を行う
- DNS Authにて冗長化方法として採用
 - DNS Authは攻撃を受けやすい Service
 - DoS攻撃時に他のServiceへ影響が発生しにくい
- 問題
 - Act-Stb構成
 - サーバのスケールを行えない
 - L2構成が必須
 - Act-Stbの切り替えパケットロスが発生する可能性がある



HWLB(Hardware Load Balancer)

- HWLBを経由してバックエンドのサーバにトラフィックの分散を行う
- DNS Cacheにて冗長化方法として採用
 - DNS Authと異なり外部から攻撃される可能性が低い場所での利用
 - スケールが容易
 - ヘルスチェックをHWLBで行える
- 問題
 - サーバ単体でDNSのメンテナンスが難しい場合がある
 - VIPのライフサイクルがHWLBに依存する
 - EoL
 - EoS



VRRP

HWLB

Pros

- トラフィックの集約点がない

- Act-Act構成がとれ公平にトラフィックをバランス
- ヘルスチェック

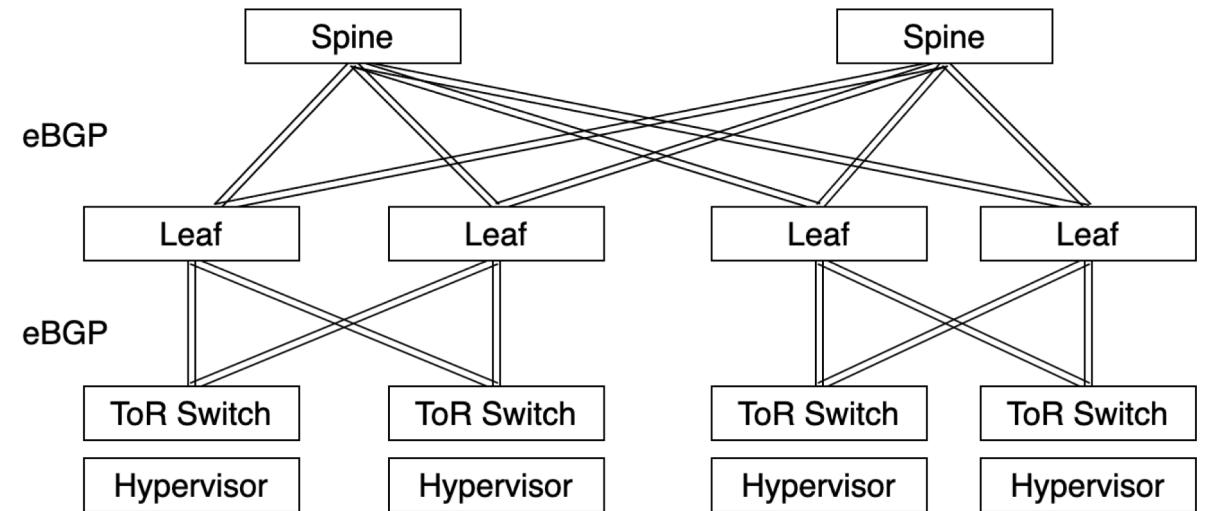
Cons

- Act-Act構成が行えない
- L2ネットワークが必要
- Act-Stbの切り替えにパケットロスが発生する

- トラフィックに集約点がある

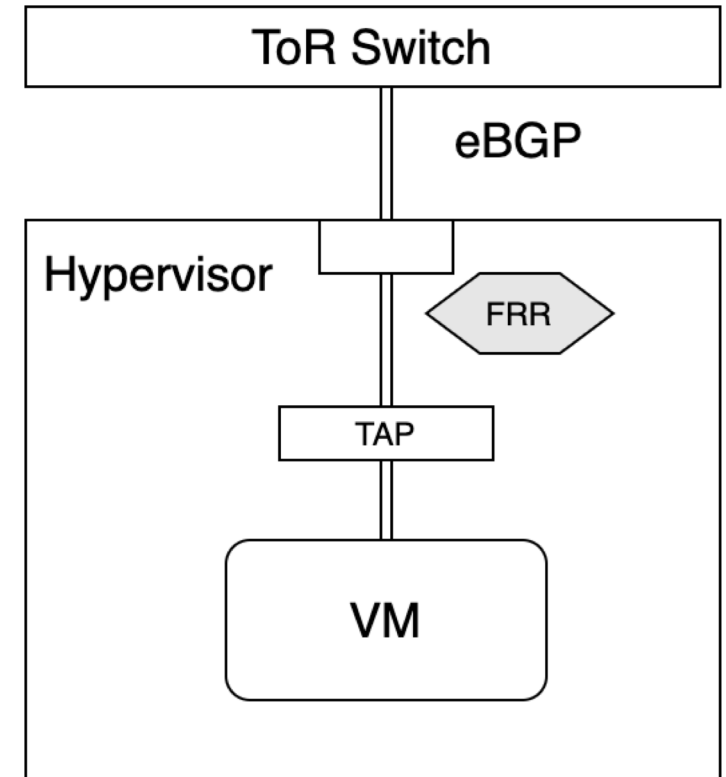
CLOS Network

- LeafやSpineといった階層的なメッシュ構造で構築されたネットワーク
- ネットワークの拡張が行いやすい
 - Spineの追加でLeaf間の帯域の増加
 - Spineの上にSuper-Spineを追加しより大規模なネットワークに
- East-Westトラフィックに強い
 - ECMP(Equal Cost Multi Path)で経路を分散し帯域をスケールさせる事が出来る



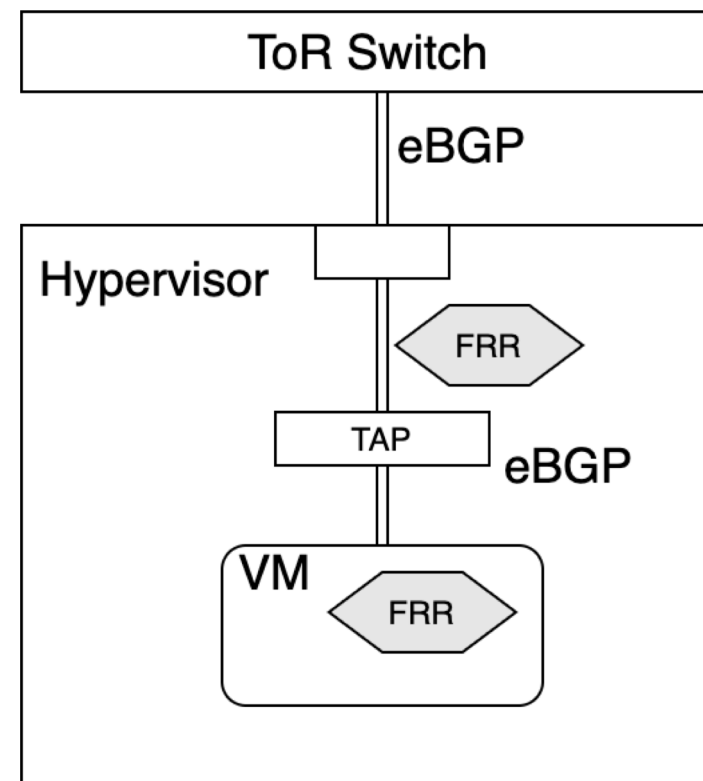
Full L3 CLOS Network

- ToR SwitchとHypervisorがeBGPでPeering
- Hypervisor上のFRRがCLOSに向けてVMの経路を広報
- HypervisorがホストしているVMのIPアドレスを/32で広報



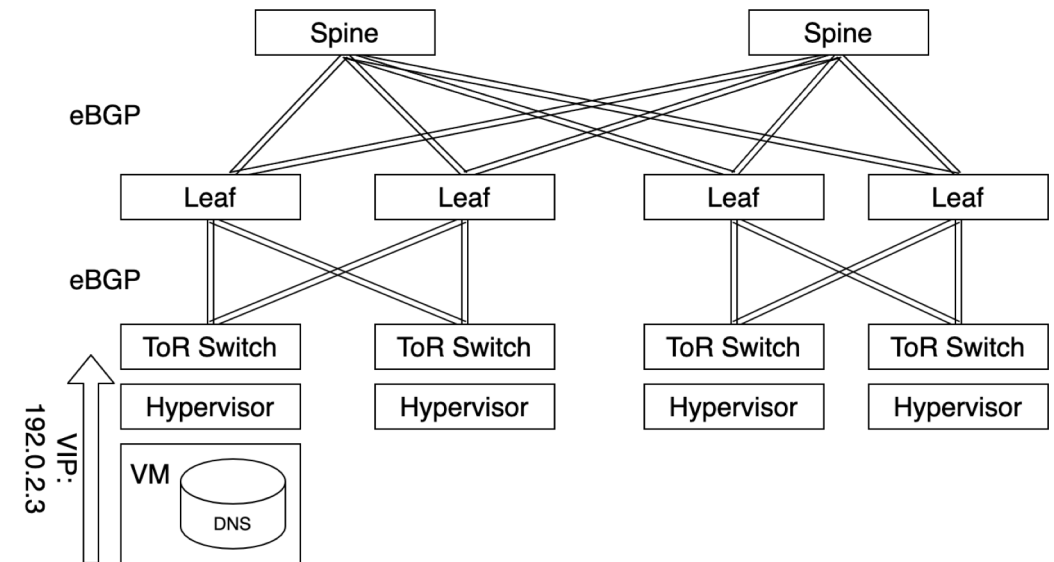
BGP advertisement from VM

- VM上のFRRとHypervisor上のFRRが Peering
- Hypervisor上のFRRが経路をToRに再広報する



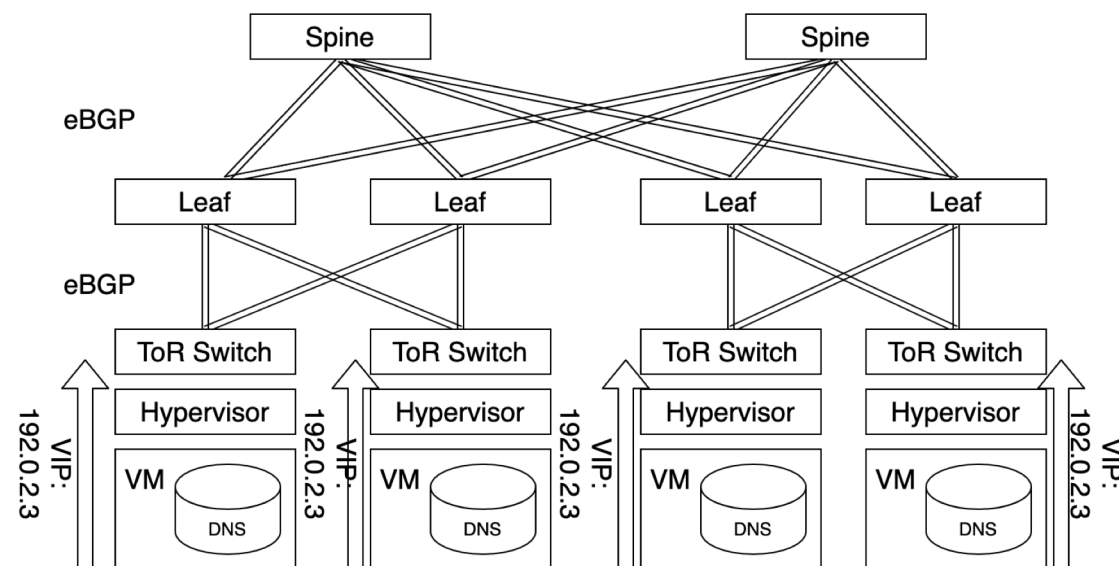
Advertise DNS VIP(1/3)

- BGPの経路広報をVMから行える機能を利用し、VMからDNSのVIPを広報



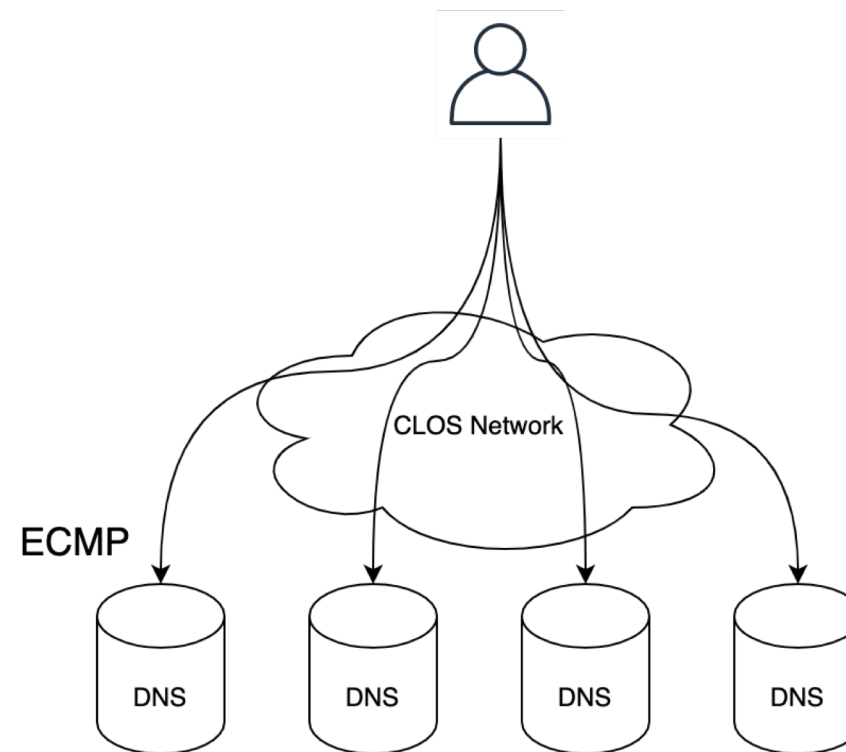
Advertise DNS VIP(2/3)

- BGPの経路広報をVMから行える機能を利用し、VMからDNSのVIPを広報
- DNSのVMをHypevisor上にそれぞれ個別に作成
- それぞれのDNSのVMからVIPを広報



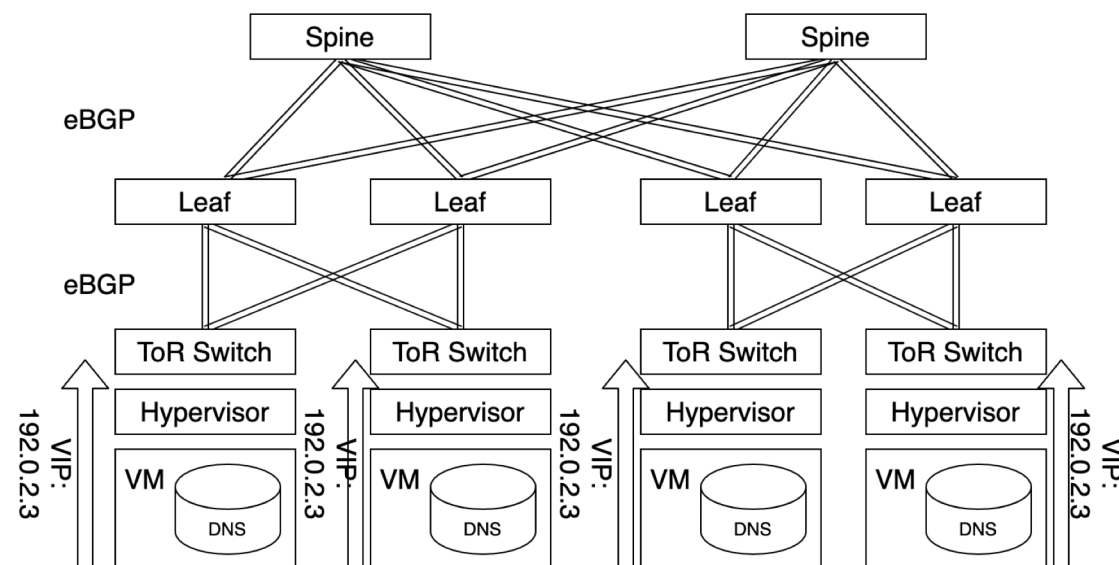
Advertise DNS VIP(3/3)

- BGPの経路広報をVMから行える機能を利用し、VMからDNSのVIPを広報
- DNSのVMをHypevisor上にそれぞれ個別に作成
- それぞれのDNSのVMからVIPを広報
- ECMPを用いて分散
 - CLOS Network上でBest Pathが選択され経路が決定される



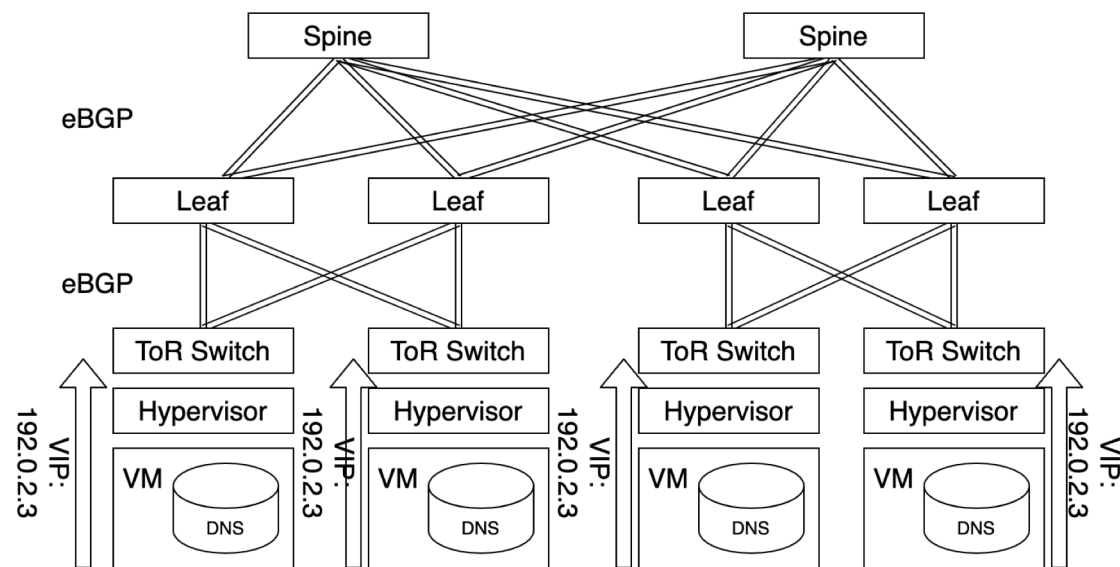
Benefit of IP ANYCAST for DNS(1/4)

- トラフィックの集約点がない
 - ECMP
 - L3 Switchが自動的に経路の選択を行う
 - CLOS Networkで許容されるだけトラフィックを分散する事が出来る
- L2 networkが不要
 - BGPを用いているのでFull L3のNetworkで構築が可能



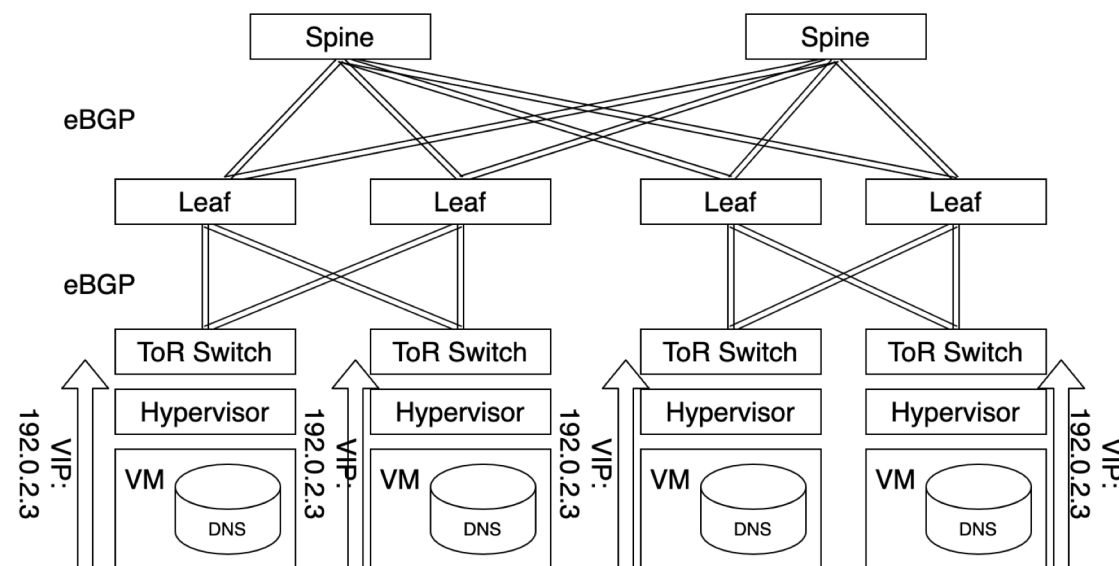
Benefit of IP ANYCAST for DNS(2/4)

- Act-Act構成がとれ公平にトラフィックをバランスする事が出来る
 - CLOS上でIP ANYCSTを行っているので、ECMPを用いてトラフィックと経路が分散される
- サーバのメンテナンス処理をBGPを用いて行う事が出来る
 - 通常のNW機器のトラフィックの迂回と同様にパケットロス無しにオペレーションを行う事が出来る



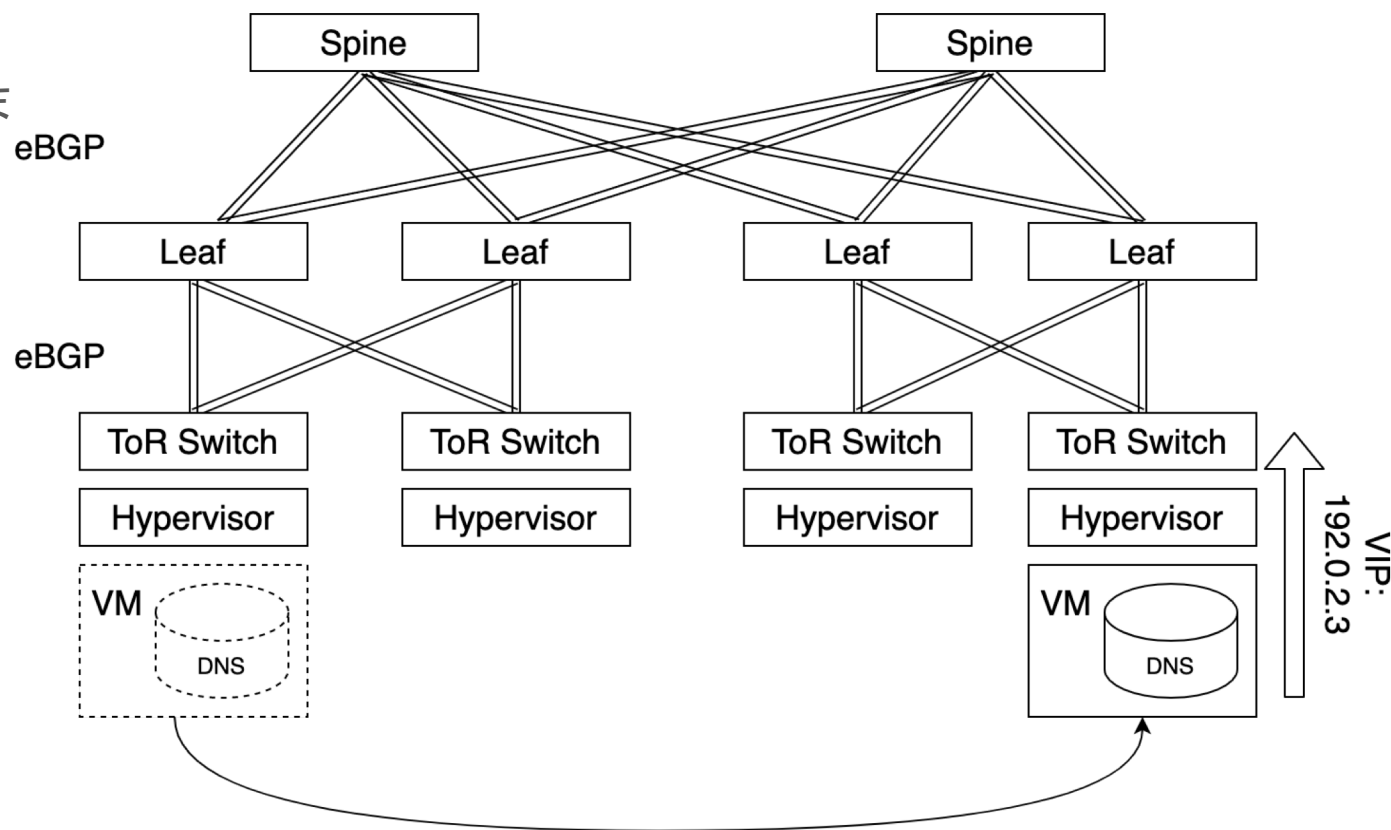
Benefit of IP ANYCAST for DNS(3/4)

- DNS AuthとDNS Cacheにて全く同じメンテナンスオペレーションが可能
 - VRRPとHWLBではメンテナンス手順が異なっていた
 - IP ANYCASTではDNS Auth/Cache共に同様の冗長化手法が利用出来る



Benefit of IP ANYCAST for DNS(4/4)

- VIPのポータビリティが高い
 - IPアドレスの経路を広報出来れば
様々な場所にDNS Cacheをそのま
まの構成で配置出来る



VRRP

HWLB

Pros

- トラフィックの集約点がない

- Act-Act構成がとれ公平にトラフィックをバランス
- ヘルスチェック

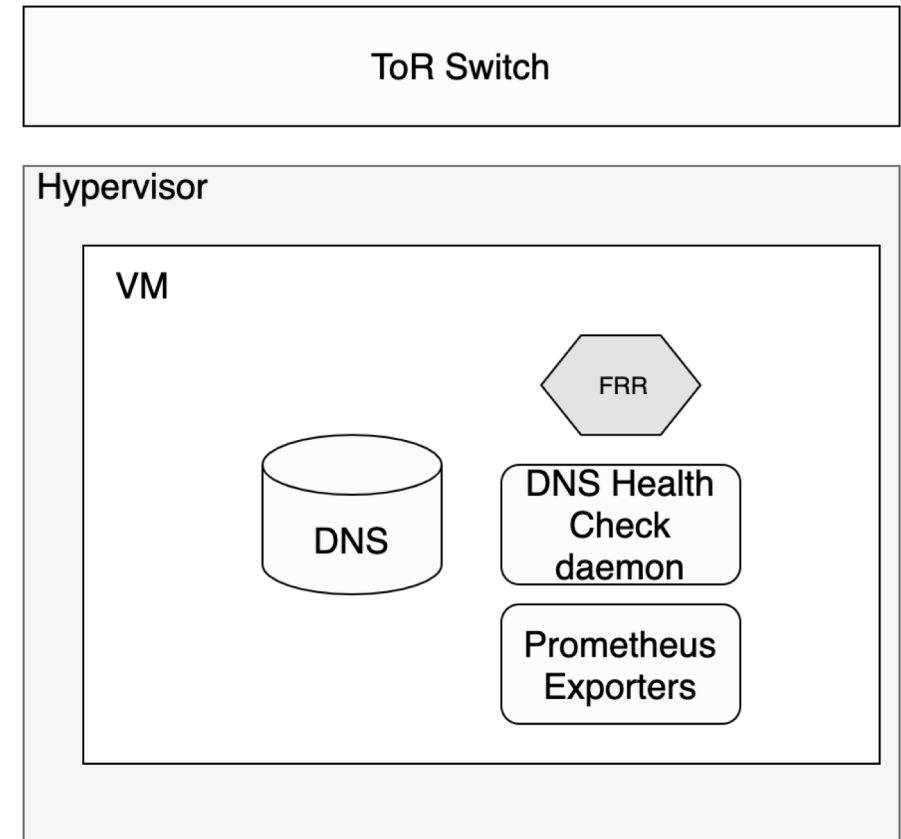
Cons

- Act-Act構成が行えない
- L2ネットワークが必要
- Act-Stbの切り替えにパケットロスが発生する

- トラフィックに集約点がある

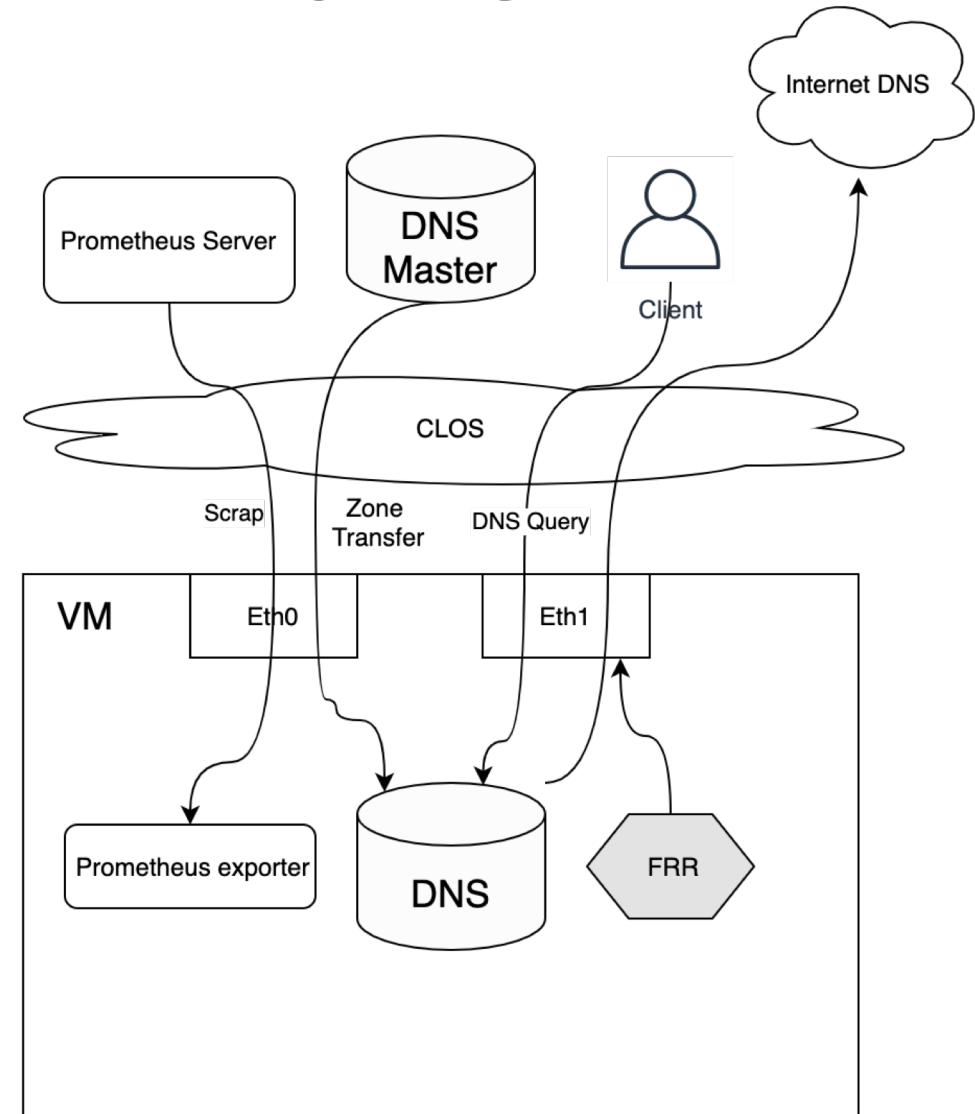
DNS Server VM

- DNSが動作しているVMでは以下のプロセスが動作している
 - FRR
 - オープンソースのルーティングソフトウェア
 - DNS Server
 - Bind
 - NSD
 - Prometheus Exporters
 - node_exporter
 - bind_exporter
 - nsd_exporter
 - Health Check Daemon
 - Next Page=>



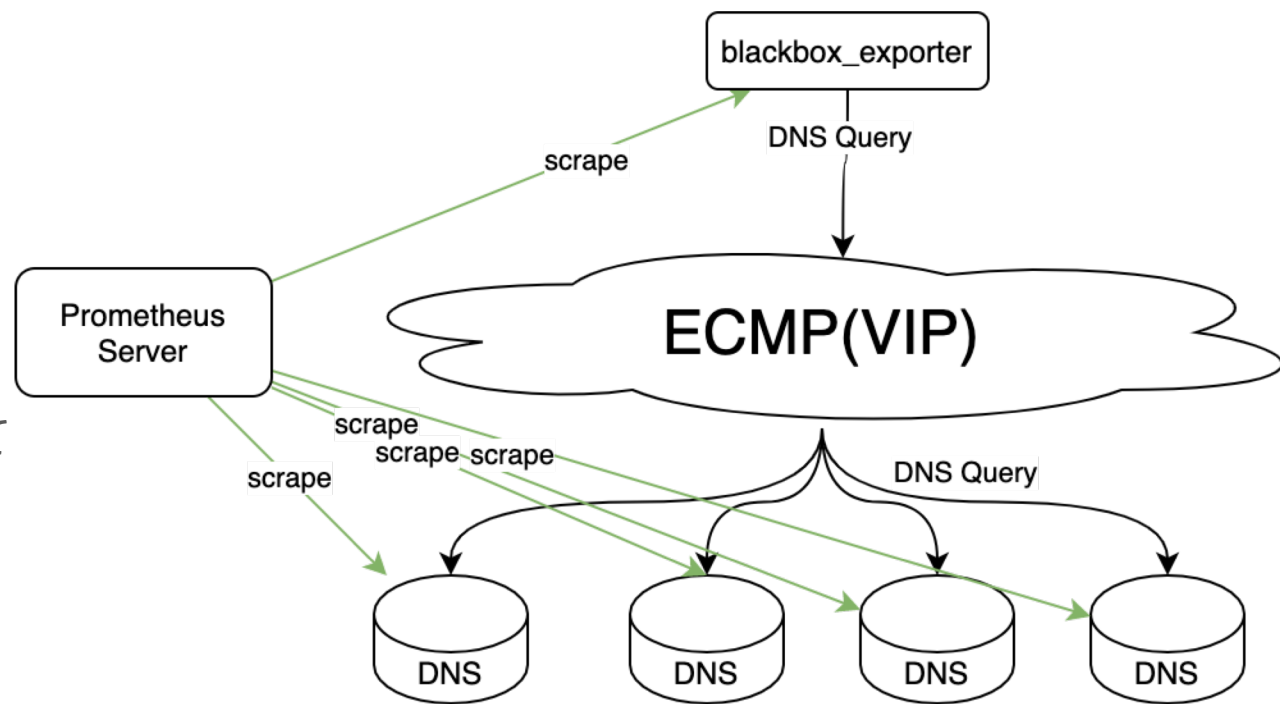
DNS Server VM Traffic

- DNS用のVMではInterfaceが2つ
 - Mgmt Interface
 - Prometheus scrape
 - ssh
 - ゾーン転送
 - Service Interface
 - ユーザからのDNS Queryの処理
 - BGP



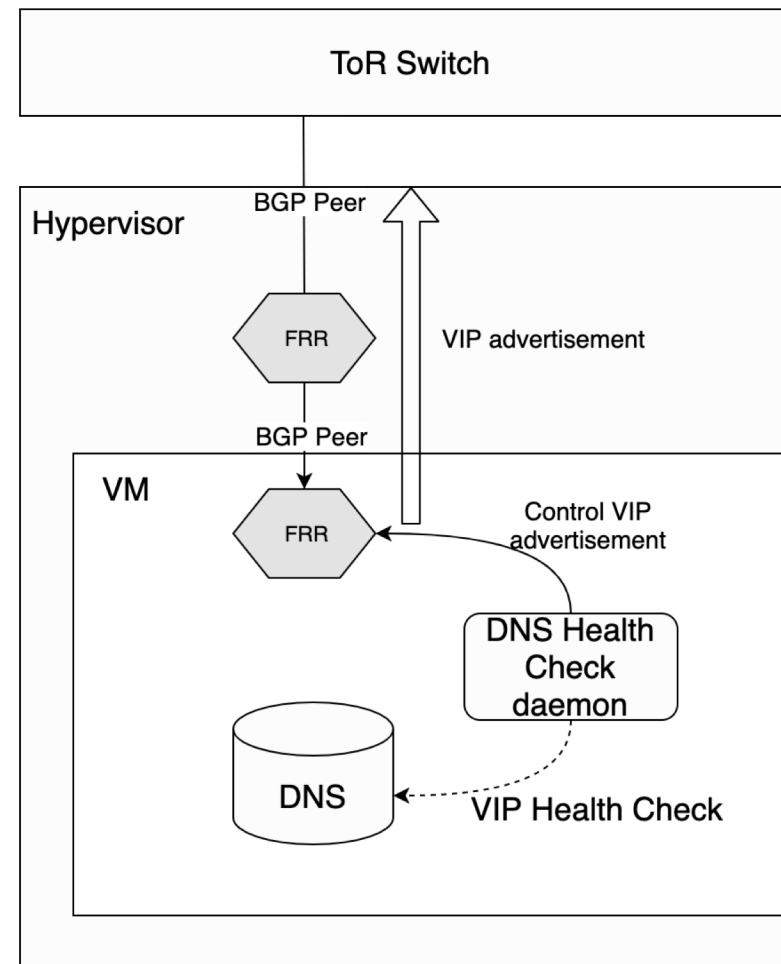
Monitoring

- Prometheus
 - Scrape exporters
 - DNSサーバのExporterに直接
 - blackbox_exporter
 - 外形監視をExporter化する事が出来る
 - DNS QueryをVIPに対して送信してユーザからのDNSの状態をモニタリング
 - 外形監視では、特定のDNSサーバにクエリが偏る可能性がある



Health Check Daemon

- Health Check DaemonはDNSのVIPに対して随時ヘルスチェックを行っている
- VMやHVの問題に関してはBGPがDownする事でService Outを行う
- DaemonはDNSの動作のチェックを行う
- ローカルのDNSサーバに対してDNS Queryを送信
 - TTL: 1
 - Destination: VIP
 - ヘルスチェック用のレコード
 - 返答をチェック
- ヘルスチェックがDownしたことを検知し経路広報の停止



VRRP

HWLB

Pros

- トラフィックの集約点がない

- Act-Act構成がとれ公平にトラフィックをバランス
- ヘルスチェック

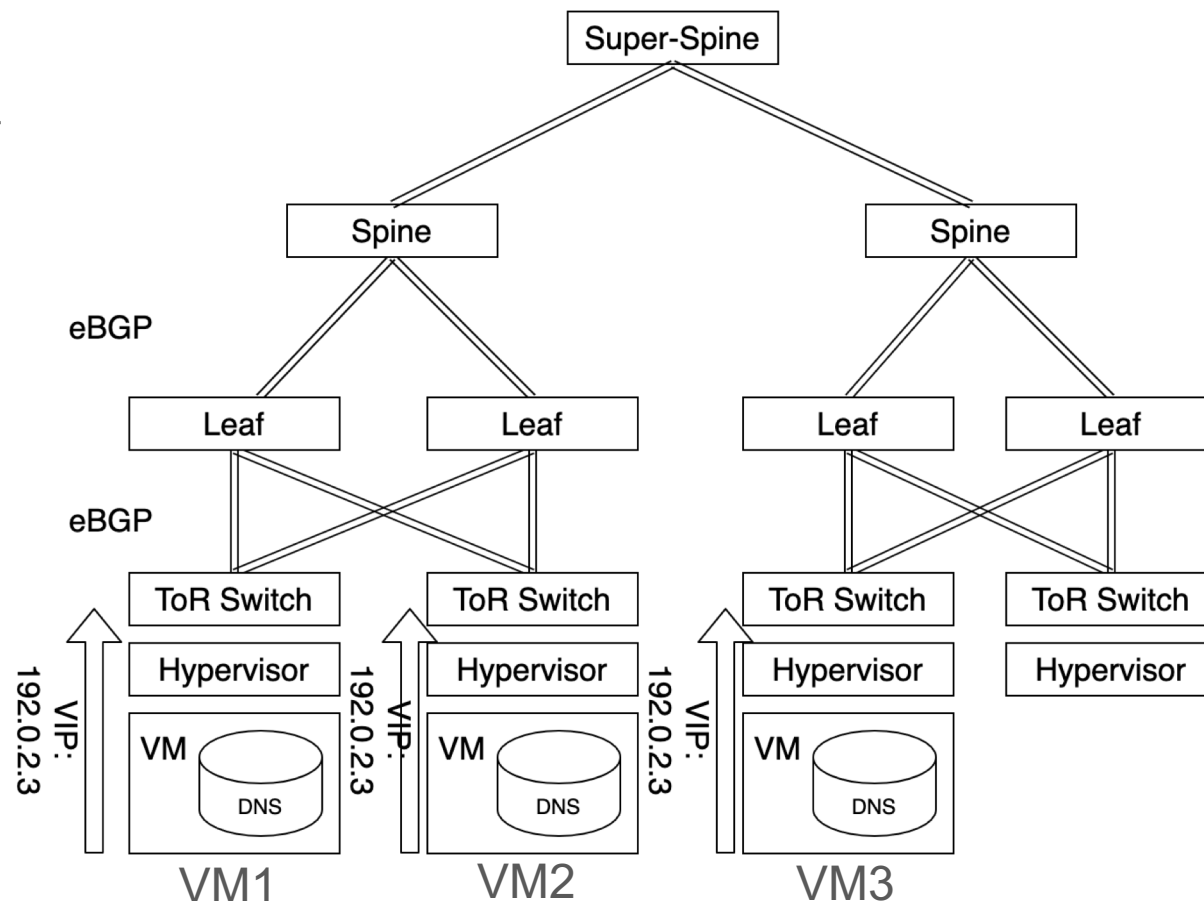
Cons

- Act-Act構成が行えない
- L2ネットワークが必要
- Act-Stbの切り替えにパケットロスが発生する

- トラフィックに集約点がある

Disadvantage of IP ANYCAST

- DNSのトラフィックを正しく分散させるためにVMの配置に検討が必要
 - BGPではBest Pathで経路が選択される
 - 例えば右の図の場合ではVM3に50%のトラフィックが流れる事になる
 - NWの構成を意識してVMの作成を行う必要がある
- 最初の経路広告のトラフィックが分散出来ない
 - BGPの経路の集約に時間差があるので例えば経路広告のタイミングにずれがあるとVM1に100%のトラフィックが流れる事になる



Conclusion

- DNSの冗長化にIP ANYCASTを用いる事で多くのメリットが存在する
 - スケーラビリティの向上
 - CLOS上にVMの追加を行うだけでキャパシティが増強可能
 - メンテナンス性の向上
 - VMから経路のコントロールが可能
 - VIPのポータビリティが向上
- IP ANYCASTの冗長化のデメリット
 - BGPはBest Pathで経路選択されるため、NWの構成を意識してバランスングする必要がある
- IP ANYCASTにて対応出来ない問題について対処する必要がある
 - Health Check Daemonを実装

Discussion

- DC内のDNSの冗長化について
- DNSのリプレイスについて
- CLOS NWの活用について

Related Documents

1. LINEのネットワークをゼロから再設計した話
<https://www.janog.gr.jp/meeting/janog43/program/line/>
2. フルサービスリゾルバのロードバランサーなし構成について
<https://dnsops.jp/event/20210625/13-kosaka.pdf>