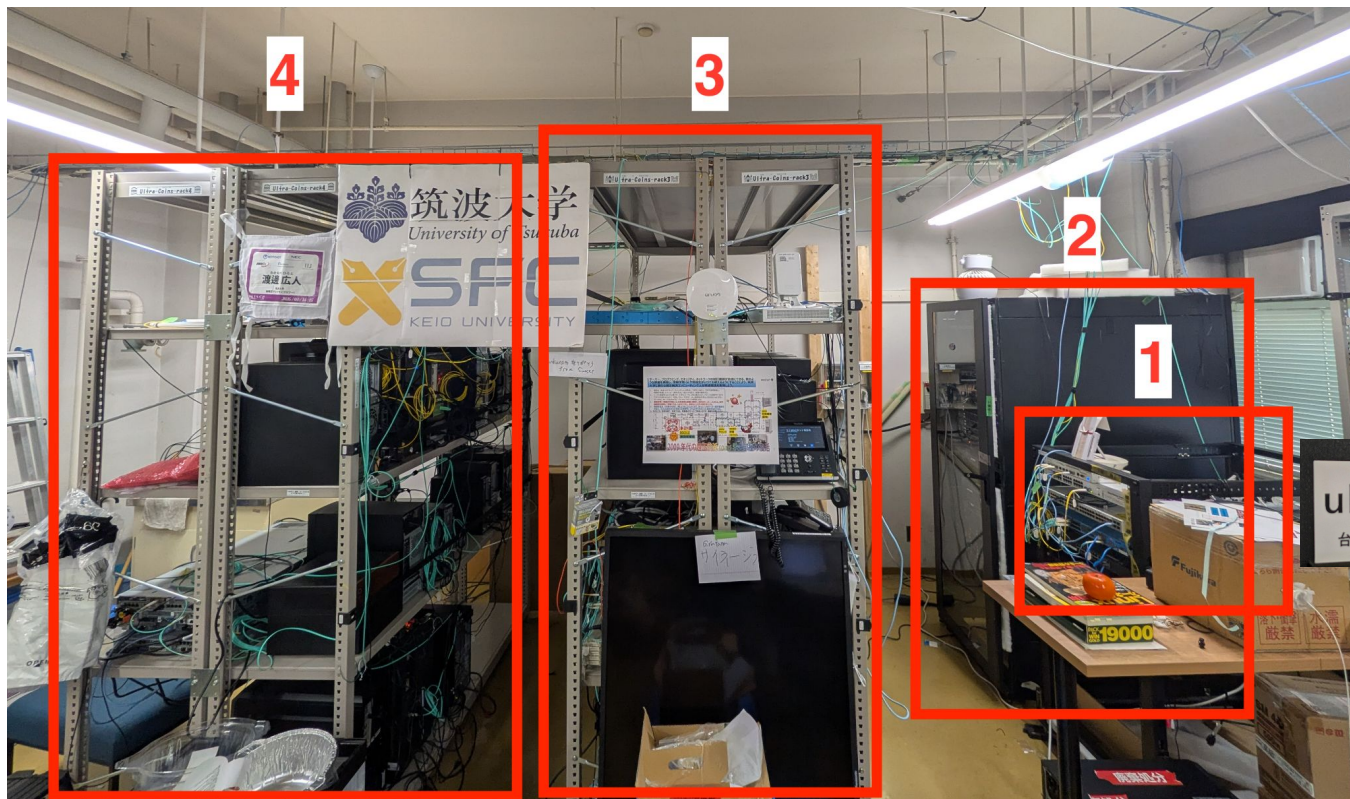
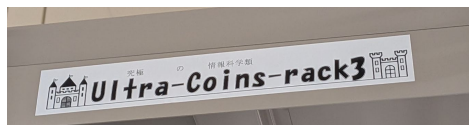


Kubernetesクラスターを ラック間で冗長化する

Ultra-Coins 中村天晴

背景: Ultra-Coinsの物理構成

- Ultra-Coins(3D212)にはラックが4つある
- rack1がネットワーク機器
- rack2~4がサーバ類

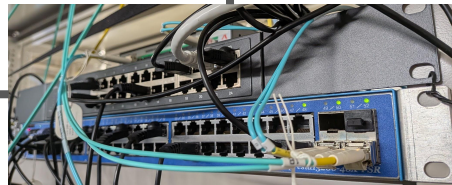


3D212には4つのラックがあります

rack4

rack3

rack1



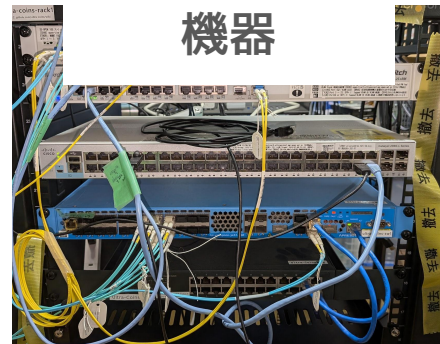
パソコン



パソコン



ネットワーク
機器



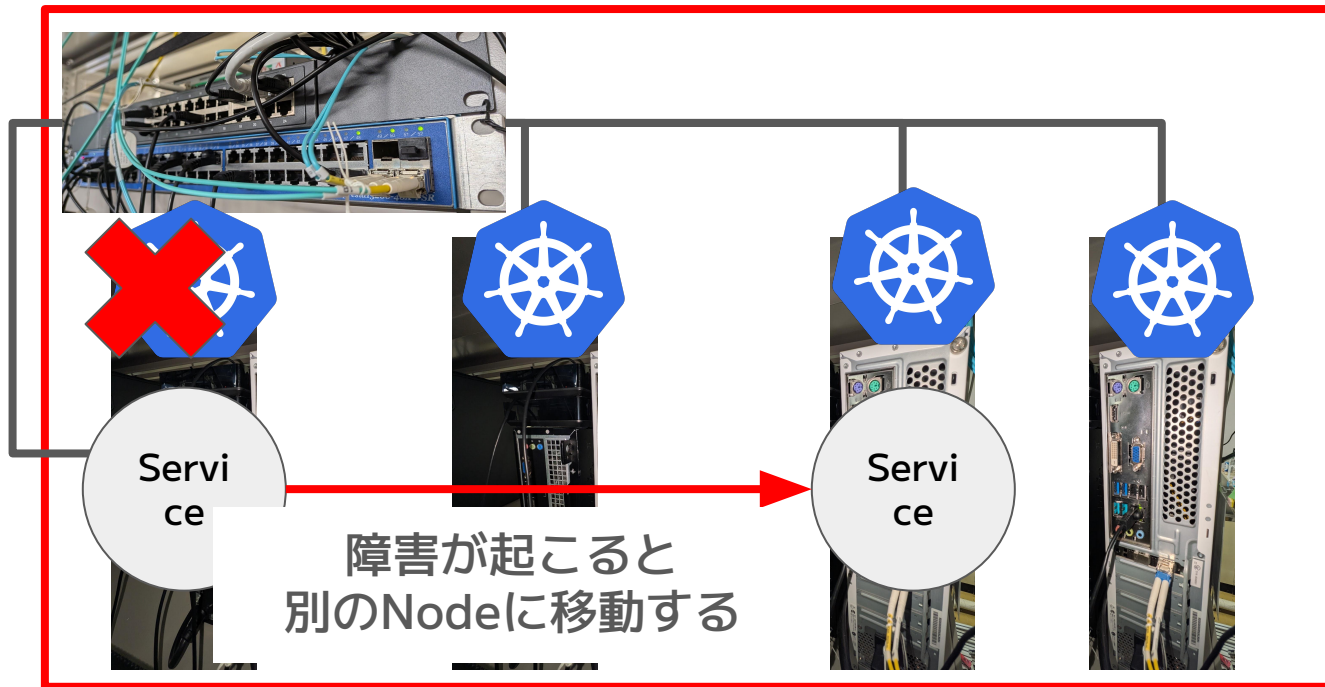
インター
ネットへ

rack2/3/4はパソコン・rack1はネットワーク機器

背景: Kubernetesクラスタ

- Kubernetesというコンテナ管理ツールを利用してWebサイトなどを公開していた
- Kubernetesを使うことで
 - 特定のホストマシン（ノード）が利用できなくなると、自動で他のホストにサービスを移動する
 - いい感じにスケジューリングしてくれる
 - 耐障害性

rack4



Kubernetesがスケジューリングする(reconcile)

課題: ラック単位での停電

- ラック単位での停電が複数回発生
 - (最近は解消しつつある)
- すべてのノードが単一ラックにいるためラック単位で停電するとKubernetesによる障害回復の恩恵を得られない



中村 雄一 (2023-04-14) 00:33

@channel

rack4に関して、GPUマシンでの実験中にブレーカが落ちてしまいました。可能な限り早く復旧します。



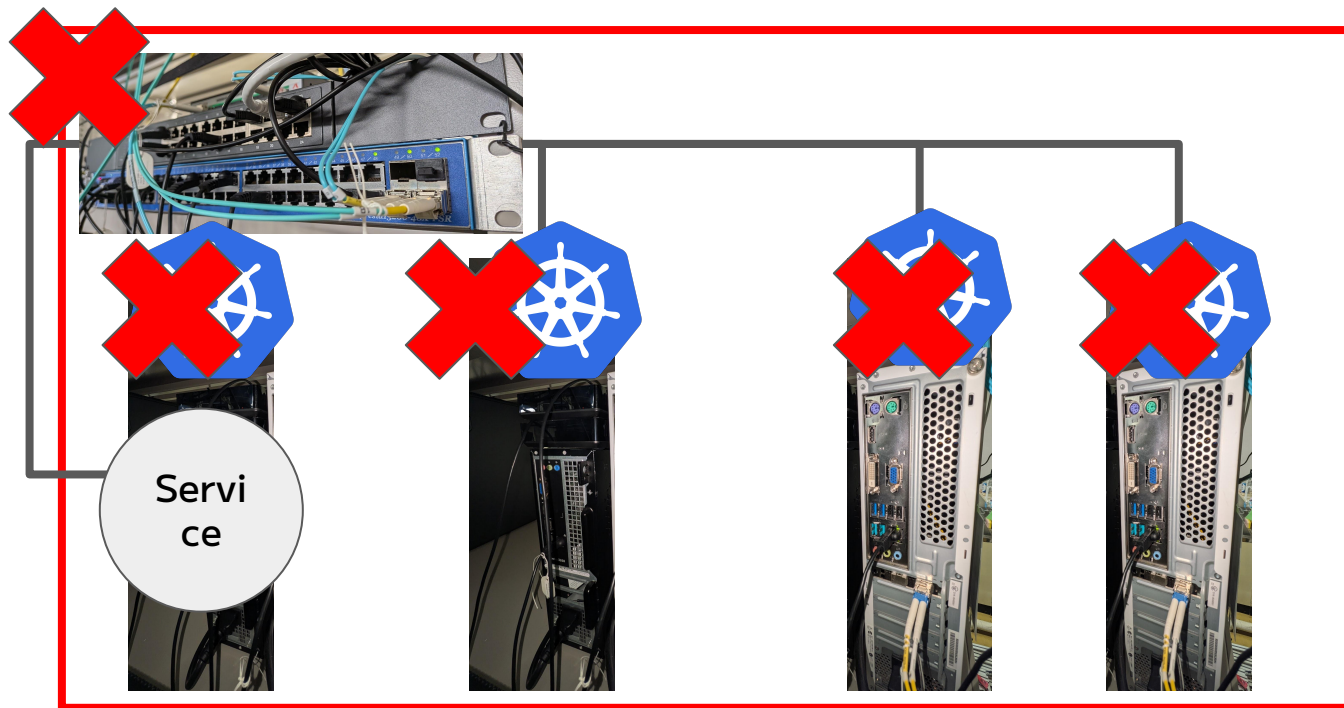
中村 雄一 (2023-04-14) 21:59

@channel

rack4に関して、またもGPUマシンでの作業中にブレーカが落ちてしまいました。復旧次でご連絡します。



rack4



ラックがとまるとすべてのNodeが停止する

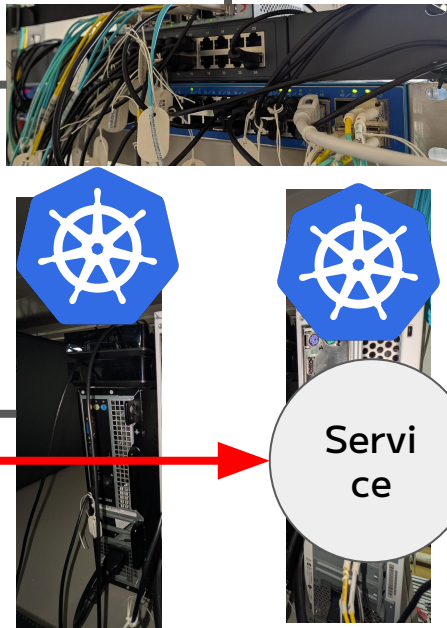
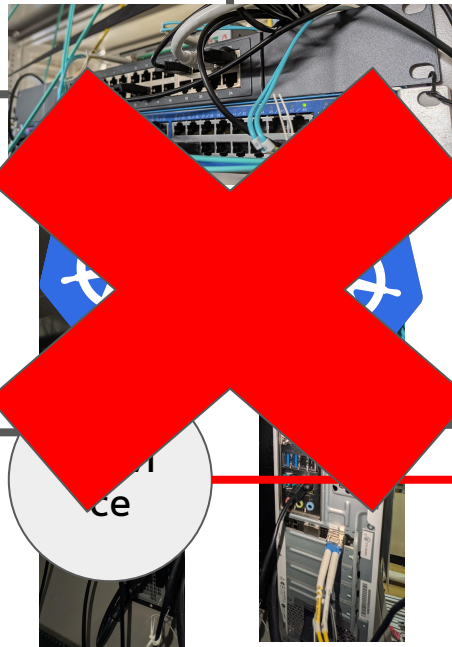
解決策: ラック間で冗長化する

- 別のラックにKubernetesが動作するNodeを設置して冗長化する
- サービスはすべてのNodeで動作させる

rack4

rack3

rack1



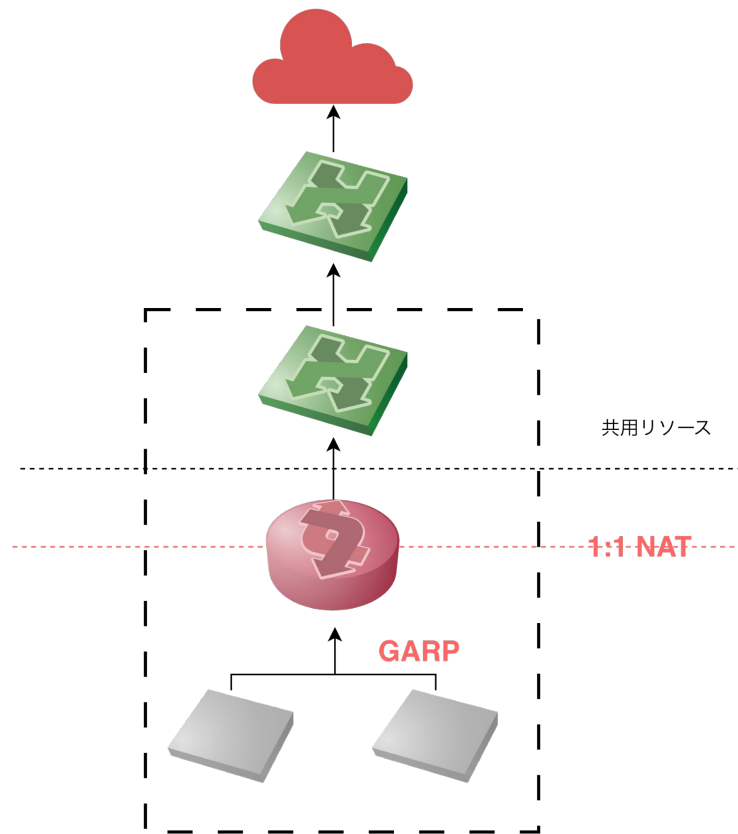
インター
ネットへ

困難な点

- 複数のレプリカに対して外部から**1つのIPアドレス**でアクセスできるようにしたい

1ラックの場合の経路

- グローバルIPとプライベートIP(VIP)の間で
1:1NATを行うルータを配置
 - ホストをグローバルに晒さない
 - 集約によりIPアドレスを節約
- LAN側ではGARPを使い生きているノードがVIPを宣言

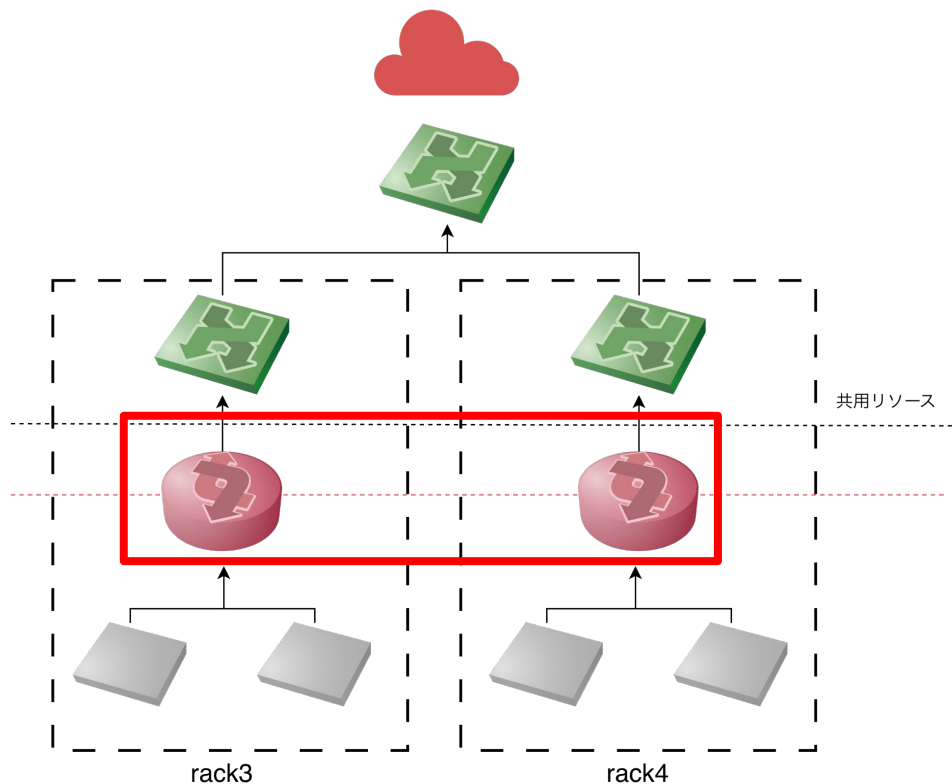


2ラック特有の問題

- WANとLANの間に位置するルータが2カ所に増加

問題

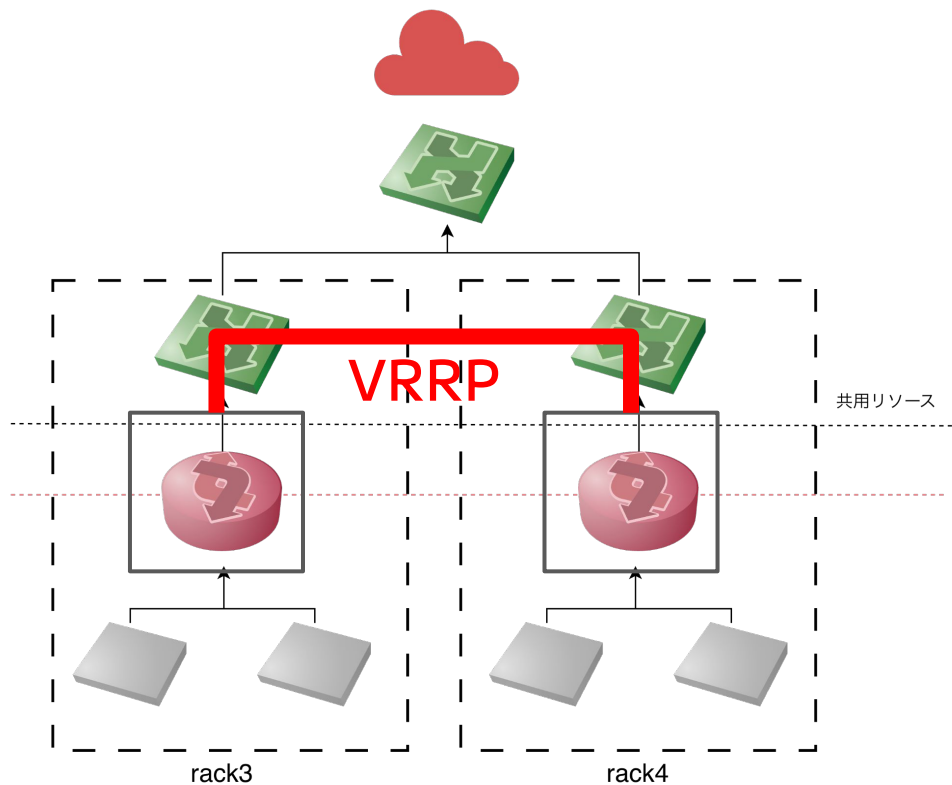
- 両方で1:1NATしたいが、**外側のIPは1つ**に固定したい
- 内側で動作するサービスがどこにいるか分からない
- GARPが他のラックに到達しない



2ラックの場合のGlobal IP

両方で1:1NATしたいが、外側のIPは1つに固定したい

- **VRRP**でActive-StandbyによりGlobal IPもVIPとして扱う
- 1:1 NATは両方のルータに設定しておく



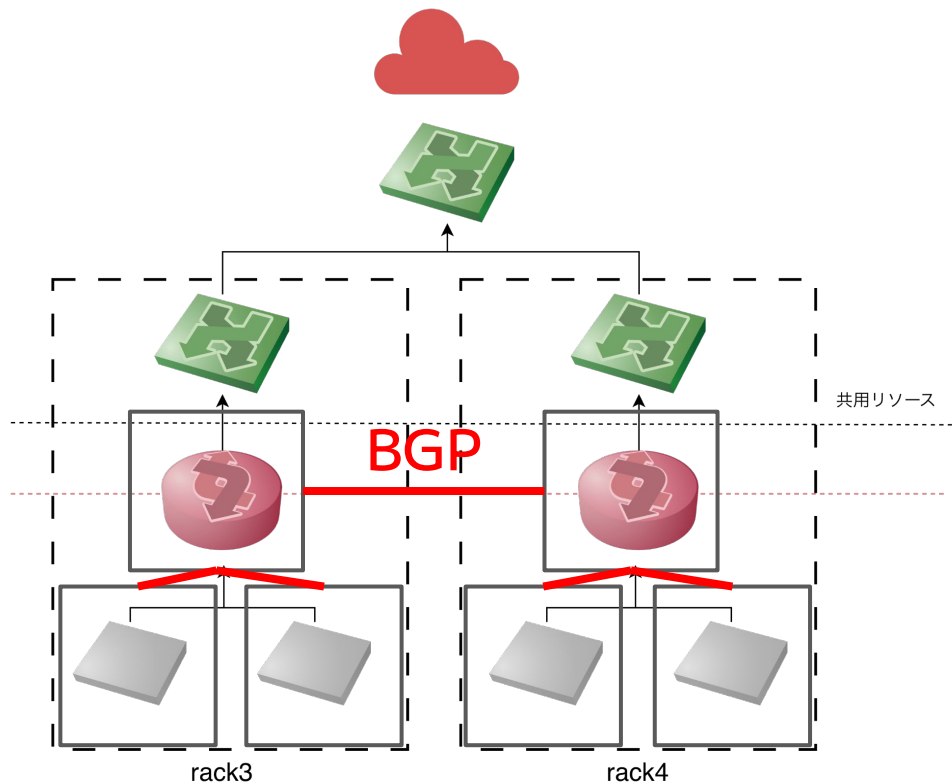
2ラックの場合のLocal IP

内側で動作するサービスがどこにいるか
分からない

GARPが他のラックに到達しない

- **BGP**による経路交換
- 各サービスが1つのVIPをルータに
広報
- ルータはパス長が最も短いサー
ビスにパケットを転送する

K8sのWorker-Control Planeの接続に
もBGPを用いたLoadBalancerを利用



```
🖥️ ping 163.220.236.82
Every 1.0s: curl -v http://163.220.2... TakaharunoMacBook-Air-10.local:

  % Total    % Received % Xferd  Average Speed   Time    Time     Tim
             Dload  Upload  Total      Spent    Lef
  0      0     0     0     0     0      0      0  --:--:--  --:--:--  --:--
20.236.23:80...
* Connected to 163.220.236.23 (163.220.236.23) port 80
> GET / HTTP/1.1
> Host: 163.220.236.23
> User-Agent: curl/8.7.1
> Accept: */*
>
* Request completely sent off
< HTTP/1.1 200 OK
< Server: nginx/1.30.4
< Date: Mon, 20 Jul 2026 13:55:55 GMT
< Content-Type: text/html
< Content-Length: 896
< Last-Modified: Wed, 15 Jul 2026 17:20:38 GMT
< Connection: keep-alive
< ETag: "6a57c166-380"
< Accept-Ranges: bytes
<
{ [896 bytes data]
100 896 100 896 0 0 47553 0 --:--:-- --:--:-- --:--
* Connection #0 to host 163.220.236.23 left intact
<!DOCTYPE html>
<html>
<head>
<title>Welcome to nginx!</title>
<style>
html { color-scheme: light dark; }
body { width: 35em; margin: 0 auto;
font-family: Tahoma, Verdana, Arial, sans-serif; }
</style>

64 bytes from 163.220.236.82: icmp_s
eq=177 ttl=63 time=4.039 ms
64 bytes from 163.220.236.82: icmp_s
eq=178 ttl=63 time=6.573 ms
64 bytes from 163.220.236.82: icmp_s
eq=179 ttl=63 time=3.410 ms
64 bytes from 163.220.236.82: icmp_s
eq=180 ttl=63 time=4.686 ms
64 bytes from 163.220.236.82: icmp_s
eq=181 ttl=63 time=3.492 ms
64 bytes from 163.220.236.82: icmp_s
eq=182 ttl=63 time=4.249 ms
64 bytes from 163.220.236.82: icmp_s
eq=183 ttl=63 time=7.795 ms
64 bytes from 163.220.236.82: icmp_s
eq=184 ttl=63 time=4.658 ms
64 bytes from 163.220.236.82: icmp_s
eq=185 ttl=63 time=5.539 ms
64 bytes from 163.220.236.82: icmp_s
eq=186 ttl=63 time=3.946 ms
64 bytes from 163.220.236.82: icmp_s
eq=187 ttl=63 time=4.054 ms
64 bytes from 163.220.236.82: icmp_s
eq=188 ttl=63 time=7.255 ms
Request timeout for icmp_seq 189
Request timeout for icmp_seq 190
Request timeout for icmp_seq 191
Request timeout for icmp_seq 192
Request timeout for icmp_seq 193
```

現状と課題

- ネットワーク的な冗長化は完了した
 - VRRPとBGPによるFail Overにより継続的にサービスを提供し続けることができる
- Kubernetesでは過半数のNodeが生きている必要がある
 - 現状ではラックを2つしか使っていないので、ノードが多いラックで障害が起こると継続動作できなくなる
- グローバルIPがActive/Standbyなので片側のノードに負荷が集中する
 - BGPのPATH長がルータと同一ラックのノードのほうが短くなるため

今後の展望

- 他のラックへのノードの追加
- ネットワーク以外の問題
 - 特にストレージを分散配置する必要がある
- Global IPのActive/Active構成
 - 上流とのBGP接続
- マルチリージョン対応

まとめ

- Ultra-Coinsでは複数のラックを利用して可用性を向上させることが出来る
- そのためにはネットワーク面でVRRP・BGPの利用が必要
- 今後はラック追加・グローバルIPもActive/Active構成を目指す