



BIGLOBEのキャッシュDNS基盤を OpenShiftに移行できるか検証してみた

ビッグロブ株式会社 プロダクト技術本部
ネットワーク技術部 開発G
島袋 健司

DNS Summer Day 2024
2024年6月21日

自己紹介

氏名

島袋 健司

経歴

- 2021年4月 BIGLOBEに新卒で入社
- ～2024年3月 バックボーンNWの設計・運用チームに所属
→ ルータ/スイッチ/NW構成変更等を経験
- 2024年4月～ 開発Gに異動
→ DNSキャッシュサーバのコンテナ化を担当
- ※ 気づいたらアサインされていた...



出身地

神奈川県座間市 ※ 沖縄出身ではありません。夏は苦手ですw

趣味

旅行、あんまん、日本酒



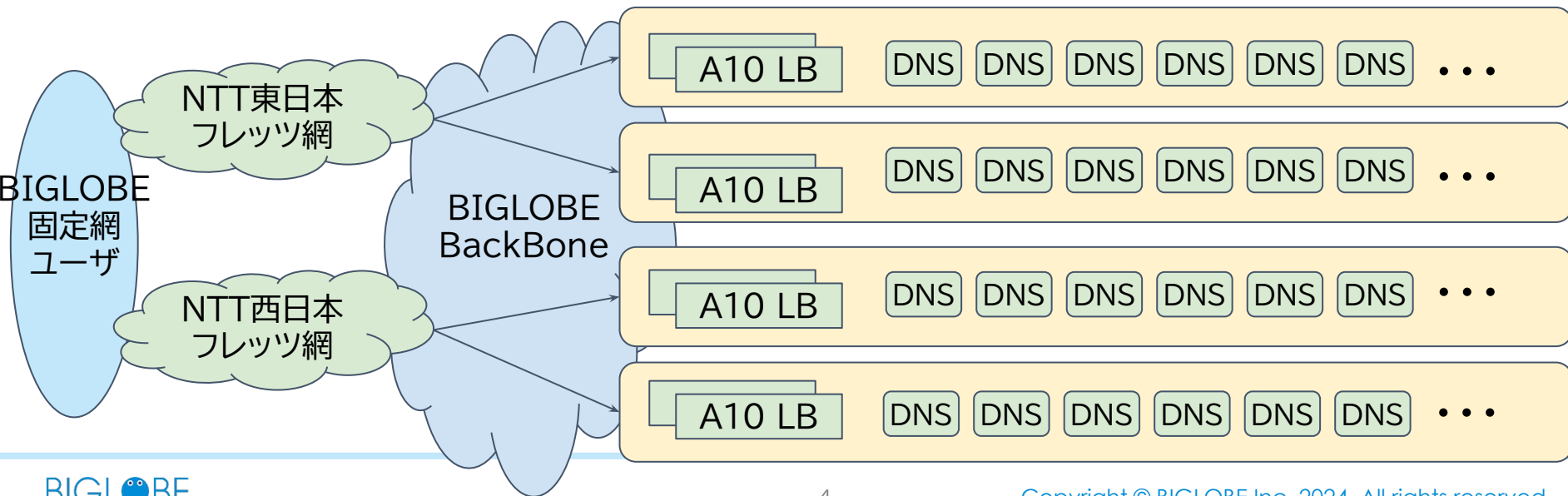
背景

BIGLOBEのDNS基盤の現状

BIGLOBEのキャッシュDNS構成

- 拠点:4拠点
- 台数:200VM以上
- BIND/Unboundの2種のOSSを利用

⇒10年以上も変わらず昔ながらのVM基盤の構成で運用中



現行キャッシュDNS基盤の問題

- ・脆弱性対応:

BIND/Unboundで年に数回緊急度の高い脆弱性がでる。
1週間以内にバッチ適応Verを評価して、人カリリース

- ・OS更改:

数年に一度OSの更改で数百台の更新作業

- ・拠点追加:

CDNトラフィックを各拠点でオフロードするにはキャッシュDNS基盤も置きたい

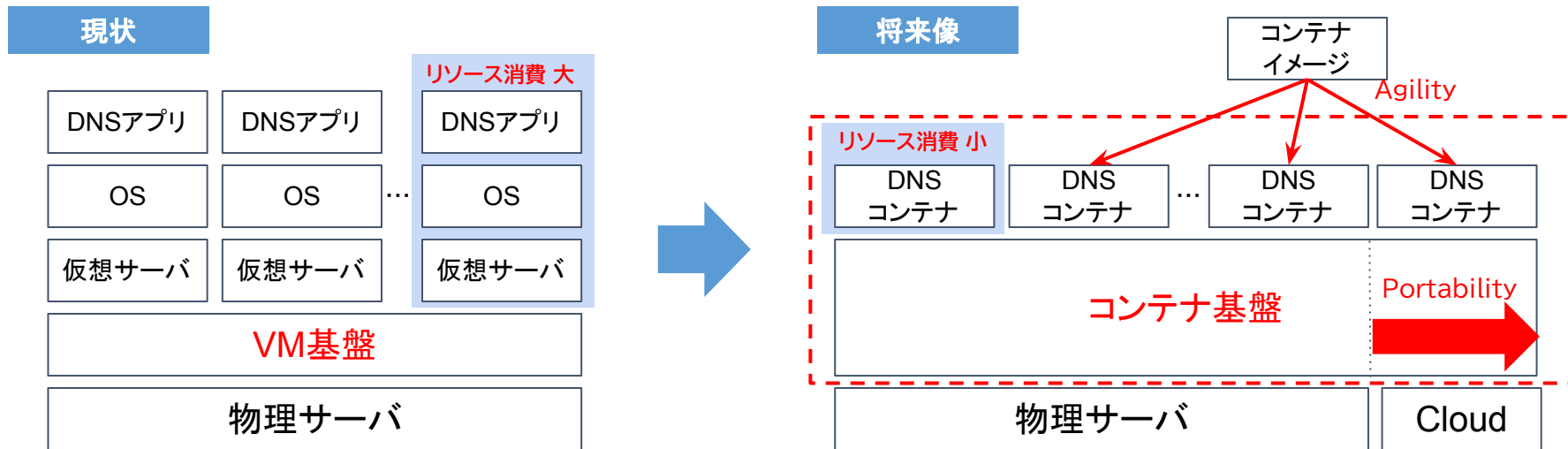
- ・手動なスケールアウト

突発的なクエリ増加には対応できず、
数時間かけてVMを足して手動でLBへ組み込み

⇒基盤のスケールやセキュリティ対応の緊急化に伴い、現場は保守・運用に疲弊。。。

コンテナ技術導入によるDNS基盤のSLA劇的向上

コンテナ基盤化することで、
可搬性・リソースの効率化・OSやDNSアプリの更新自動化を実現してSLAを向上させる



VMのOS単位での構築・運用・保守となり、
DNSの柔軟な運用が困難

軽量のコンテナ化により
DNSの柔軟な運用及び自動化が可能



Kubernetes®について

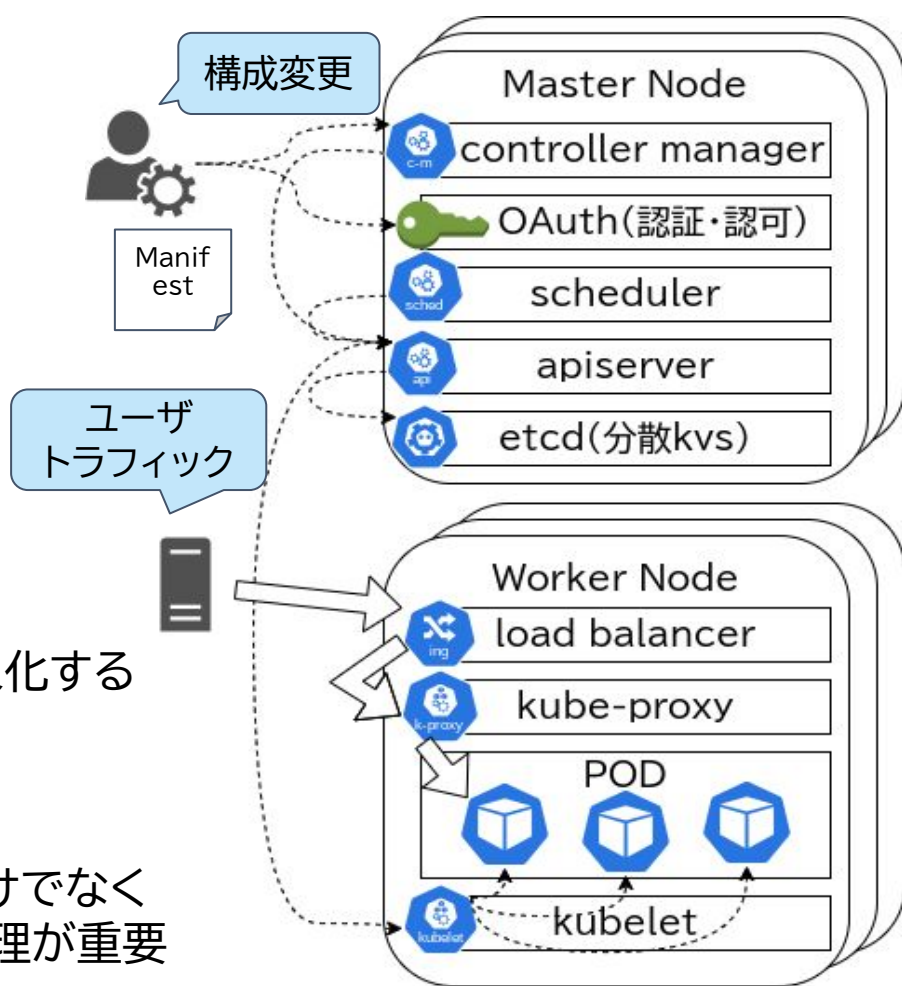
・Kubernetesとは

Googleのクラスタ管理システムBorgをベースにデファクトスタンダードとなったコンテナオーケストレーション

・コンテナオーケストレーションの主要機能

- スケジューリング:
コンテナの最適なリソース配置を行う
- 抽象化:
各クラウド・オンプレミスのリソースを抽象化する
- セルフヒーリング:
障害が起こっても自動で修復する

⇒ユーザトラフィックを捌くWorker Nodeだけでなく
沢山機能を持っているMaster Nodeの管理が重要



BIGLOBEのコンテナキャッシュDNS基盤向けの4つの必須条件

- ・オンプレミス

キャッシュDNSは低レイテンシーが必須。またDNS用の自前グローバルIPもあり、クラウド上に持っていくのも難しい。自前の各拠点上で基盤を実装する必要がある

- ・Master Nodeの統合管理、自動構築

各拠点にクラスターを配置するので、
各拠点のMaster Nodeを楽に構築、統合管理したい

- ・物理LBとの連携

物理LBとのデュアルスタック構成が必須。また、LBの性能要件が高く求められるため、既存のVLAN構成からは変えたくない

- ・コンテナの自動展開(CI/CD)

各拠点へ同じコンテナイメージを横断で自動デプロイ・正常性確認も自動化したい



Red Hat®社様に相談したところ
Red Hat社のKubernetesディストリビューションの
OpenShift®とACMならできる！！
とのことで検証スタート

Red Hat OpenShiftのACMとは

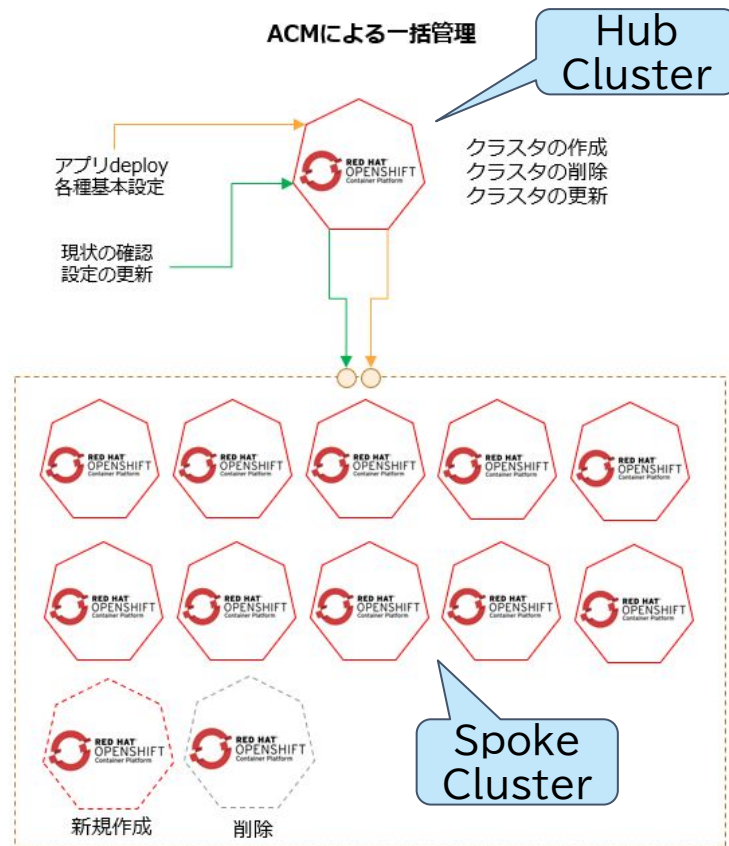
ACM(Red Hat® Advanced Cluster Management for Kubernetes)とは
OpenShift®をHub&Spoke構成にして、
Spokeとなる各クラスターを集中管理する
仕組み

主に以下の4つの機能を提供

- ・管理対象クラスターの可視化
- ・クラスターのライフサイクル管理
- ・クラスターをまたがったアプリのデプロイ
- ・クラスターに対する統一ポリシーの適応

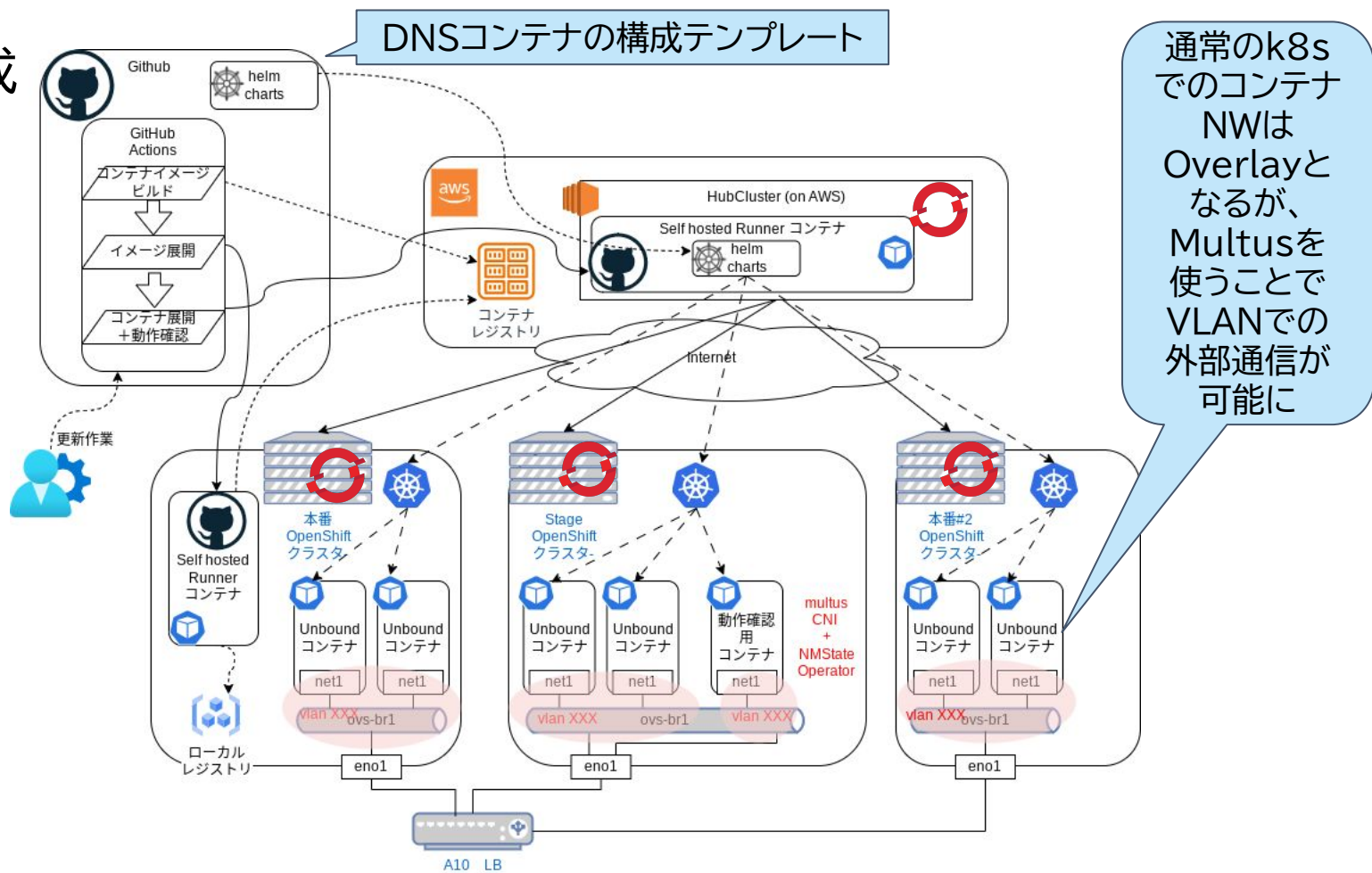
参考

HP: https://rheb.hatenablog.com/entry/2020/12/21/%E3%81%9F%E3%81%8F%E3%81%95%E3%82%93%E3%81%AE_Kubernetes_%E3%82%AF%E3%83%A9%E3%82%B9%E3%82%BF%E3%83%BC%E3%82%92%E7%AE%A1%E7%90%86%E3%81%97%E3%81%9F%E3%81%84_%EF%BD%9E_ACM_2.1%E3%81%AE%E3%83%AA%E3%83%AA

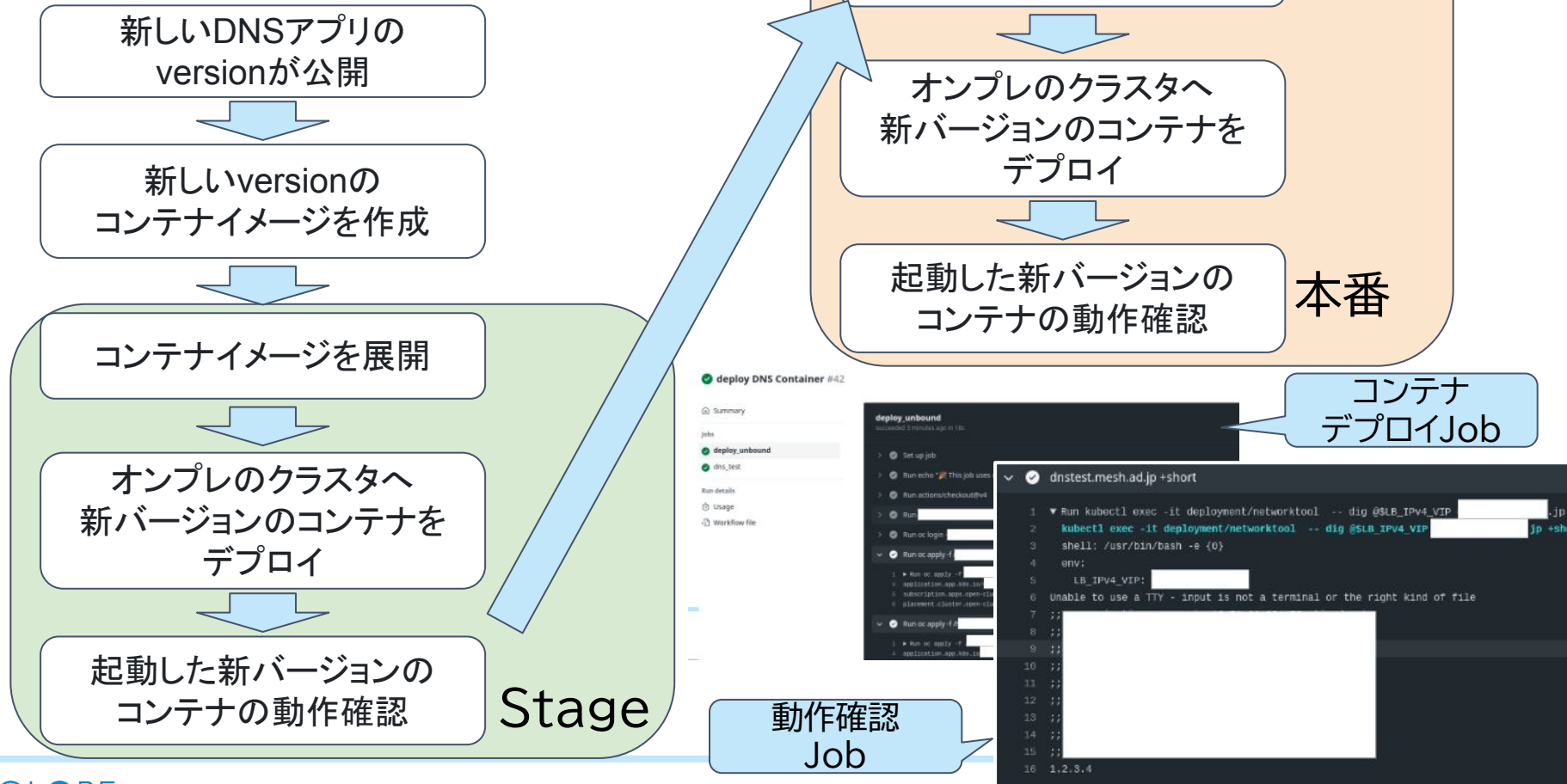


PoCやってみた

PoC構成



PoCのCI/CDの流れ(抜粋)



PoCで出たトラブルについて

①A10 LBとの連携について

ベンダーのA10社が公式にKubernetes/OpenShiftとのLB連携機能(Ingress Controller)として「TKC」※を提供しているが、、

※<https://www.a10networks.co.jp/news/blog/load-balancing-traffic-to-applications-in-kubernetes-cluster.html>

⇒DualStack(IPv4+IPv6)構成(DNSでは必須)に対応していなかった。
現状はIPv4のみの動作サポート

⇒KubernetesのCNI(NWプラグイン)はCalicoのIPIPのみサポート。
(Overlayはスループット劣化があるので使いたくない。VLANでLBと連携したい)

**Multus(VLAN)およびIPv6対応を実現するLB連携機能を
BIGLOBEにて簡易的な連携ツールを内製開発中です
また、A10社側にも対応の要望を依頼しております**

②IPv6のアドレス払い出し周りについて

②-1 IPv6のRA機能とMultusのIPAMの機能が競合

Multusが右の設定で指定された**アドレスプール**
(**赤太字部分**)から払い出しするが、
PODが起動した際にRA機能が働き、
別のIPv6アドレスが自動についてしまった。
その自動付与されたIPv6アドレスが優先され、
外部通信できない問題が発生！！

⇒IPv6のRA機能を止めるには
カーネルパラメータを設定が必須だったが
コンテナのセキュリティ設計的に弄りたくなかった
のでワークアラウンドを検討

ワークアラウンド

InitContainerで自動付与された
アドレスを消す処理とした

※InitContainerとは、PODにてメイン処理をするコンテナが
起動する前に前処理を行うコンテナのこと

Multusの
設定

```
namespace: default
rawCNICConfig: |-
{
  "name": "tagvlanXXX",
  "cniVersion": "0.3.1",
  "type": "ovn-k8s-cni-overlay",
  "topology": "localnet",
  "mtu": 1500,
  "vlanID": XXX,
  "netAttachDefName": "default/tagvlanXXX",
  "subnets":
    "10.XX.ZZ.0/24,2XXX:YYY:31f:20::/64"
}
type: Raw
```

3: net1@if13872:
<省略>

コンテナ内のアドレス状態

```
inet6 2XXX:YYY:31f:20:858:aff:fe33:bf0/64 scope global dynamic
mngtmpaddr
  valid_lft 2425274sec preferred_lft 438074sec
inet6 2XXX:YYY:31f:20::f0/64 scope global
  valid_lft forever preferred_lft forever
```

赤枠がRA機能で自動付与されたアドレス。こちらの
アドレスが優先されてしまう。
青枠のアドレスだけ付与したい。。

②-2 既存払い出しアドレスのexclude機能がおかしい

LBなどの物理NW機器で使っているアドレスをMultusで使わないようにしたい

⇒excludeSubnetsでCIDR単位で除外できる！！！！

しかし、、、IPv6CIDRの指定の仕方でPODが起動しない・・・

CIDR範囲を大きい範囲にしないのがおすすめ。

```
namespace: default
rawCNConfig: |-
{
  "name": "tagvlanXXX",
  "cniVersion": "0.3.1",
  "type": "ovn-k8s-cni-overlay",
  "topology": "localnet",
  "mtu": 1500,
  "vlanID": XXX,
  "netAttachDefName": "default/tagvlanXXX",
  "subnets": "10.XX.ZZ.0/24,2XXX:YYY:31f:20::/64"
  ,"excludeSubnets": "2XXX:YYY31f:20::/65"
}
type: Raw
```

範囲が大きいと
以下のPOD起動
エラーが起こる

```
namespace: default
rawCNConfig: |-
{
  "name": "tagvlanXXX",
  "cniVersion": "0.3.1",
  "type": "ovn-k8s-cni-overlay",
  "topology": "localnet",
  "mtu": 1500,
  "vlanID": XXX,
  "netAttachDefName": "default/tagvlanXXX",
  "subnets": "10.XX.ZZ.0/24,2XXX:YYY:31f:20::/64"
  ,"excludeSubnets": "2XXX:YYY31f:20::/121"
}
type: Raw
```

こちらは
起動OK

Events:

Type	Reason	Age	From	Message
------	--------	-----	------	---------

Warning FailedCreatePodSandBox <中略>

: error adding container to network "tagvlanXXX": CNI request failed with status 400: '[default/unbound3 <中略>

network tagvlanXXX NAD default/tagvlanXXX] failed to get pod annotation: timed out waiting for annotations: context deadline exceeded

この挙動はサポート
問い合わせ中

③Multus + OVN-Kubernetesでの Static Routingの実装

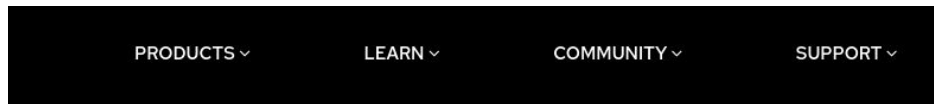
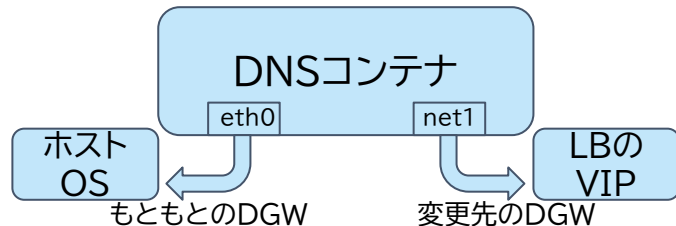
VLANネットワーク内のルータへDGWを向けたい

Multusが自体はIPAMの設定で指定できるが、、、
OVN-Kubernetesと連携させるやり方だと指定できない。。。

(egressの設定を入れればDGWの設定は入るが、本構成ではいらない。。。)

⇒InitContainerで別途下記処理を実装

- ・自動DGW設定削除
- ・指定したDGW設定追加



ner Platform 4.15 / Networking / Multiple networks / Configuring an additional network

Table 12. `ipam.routes[]` array

Field	Type	Description
<code>dst</code>	string	The IP address range in CIDR format, such as <code>192.168.17.0/24</code> or <code>0.0.0.0/0</code> for the default route.
<code>gw</code>	string	The gateway where network traffic is routed.

ipamフィールドは
ovn-kubernetes向けには指定できなかった

まとめ

LB連携機能やIPv6周りの実装が足りていない面がまだありますが、ACMおよびOpenShiftのアーキテクチャ構成であれば、BIGLOBEがめざしているDNSコンテナ基盤が実現できるとわかるPoCとなったので、足りない部分は開発して補っていき、DNSコンテナ基盤のリリースに向けて頑張っていきます。

OpenShift/Kubernetes/DNSの概要は知っているがほとんど触ったことがないため、まずは機能を覚えるところから頑張ります。慣れてきたら、皆様と技術力を高め合いながらこの業界を盛り上げます。



商標について

※ NTTは、日本電信電話株式会社の登録商標です。

※ フレッツは、東日本電信電話株式会社および西日本電信電話株式会社の商標または登録商標です。

※ Kubernetesは、米国およびその他の各国における The Linux Foundationの商標または登録商標です。

※ Googleは、Google LLC の商標です。

※ Red Hat 、Red Hat ロゴ および OpenShiftは、米国およびその他の国における Red Hat, Inc. またはその子会社の商標または登録商標です。

※ AWSおよびAWSロゴは、[Amazon.com](https://www.amazon.com), Inc.またはその関連会社の商標です。

※ 記載している企業、団体、製品、サービス等の名称は各社またはその関連会社の商標または登録商標です。

BIGLOBE

Appendix

ROSA(ACM有効化)の管理画面

☰

Red Hat
OpenShift Service on AWS

すべてのクラスター ▾

ホーム ▸

インフラストラクチャー ▾

クラスター

自動化

ホストインベントリー

アプリケーション

ガバナンス

認証情報

クラスター ①

クラスターリスト クラスターセット クラスタープール 検出されたクラスター

□ ▾

🔍 検索

🔼 フィルター ▾

クラスターの作成

クラスターのインポート

アクション

	名前 ↑ ②	Namespace ⓘ	ステータス ⓘ	インフラストラクチャー ⓘ	コントロールプレーンタイプ ⓘ
☐	local-cluster	local-cluster	🟢 準備完了	aws Amazon Web Services	ハブ
☐	rosa-client01	rosa-client01	🟢 準備完了	☁ Other	スタンドアロン
☐	rosa-client02	rosa-client02	🟢 準備完了	☁ Other	スタンドアロン
☐	rosa-serverclient	rosa-serverclient	🟢 準備完了	☁ Other	スタンドアロン

Demo内容(Stageへの展開準備)

新しいversionのunboundのソースが公開



新しいversionの
Unboundコンテナイメージを作成
Job名 : Create DNS Container Image



無事イメージビルドできたら

オンプレのStage環境にて新しいversionの
Unboundコンテナイメージを展開
Job名 : Actions Runner Controller Demo

← Create DNS Container Image
✓ Create DNS Container Image #35

Summary

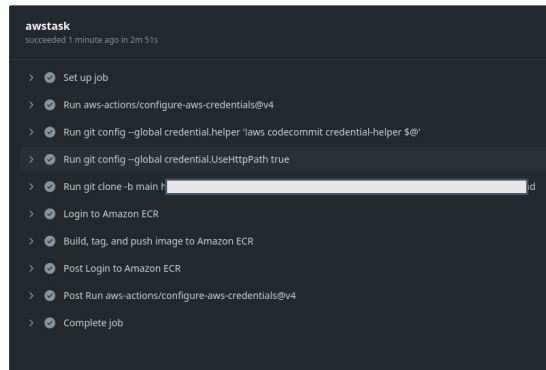
Jobs

awtask

Run details

Usage

Workflow file



← Actions Runner Controller Demo

✓ Actions Runner Controller Demo #52

Summary

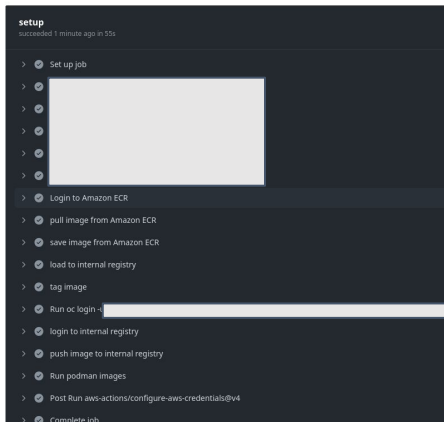
Jobs

setup

Run details

Usage

Workflow file



Demo内容(Stageへ展開、動作確認)

ACMによる各環境のサブスクリプションの
YAML情報に指定されているコンテナの
バージョンを更新(プルリク▶承認)



Stageのクラスタへ新バージョンのunbound
のコンテナイメージをデプロイ



起動した新バージョンのunboundに
動作確認

Stage unbound1.19.1 #16

[Open](#) bgl-tkgh wants to merge 2 commits into [stage](#) from [stage-unbound1.19.1](#)

Conversation 0 Commits 2 Checks 0 Files changed 2

Changes from all commits File filter Conversations

Filter changed files

stage
unbound01.yaml
unbound02.yaml

```
@@ -35,7 +35,7 @@ spec:
  35 35      image:
  36 36        pullPolicy: Always
  37 37        repository: image-registry.openshift-image-registry.svc:5000/default/biglobe-unbo
  38 -      tag: 1.19.0
  38 +      tag: 1.19.1
  39 39      podAnnotations:
  40 40        k8s.v1.cni.cncf.io/networks: '[{"name": "tagvia", "ips": [ "
  41 41      securityContext:
```

✓ deploy DNS Container #42

Summary

Jobs

✓ deploy_unbound

✓ dns_test

Run details

Usage

deploy_unbound

succeeded 3 minutes ago in 18s

> ✓ Set up job

> ✓ Run echo "This job uses runner scale set runners!"

> ✓ Run actions/checkout@v4

> ✓ Run

> ✓ Run oc login

> ✓ Run oc apply -f /home/runner/work/unbound01.yaml

```
1 ▶ Run oc apply -f /home/runner/work/unbound01.yaml
4 application.app.k8s.io/
5 subscription.apps.open-cluster-management.io/
6 placement.cluster.open-cluster-management.io/
```

> ✓ Run oc apply -f /home/runner/work/unbound02.yaml

```
1 ▶ Run oc apply -f /home/runner/work/unbound02.yaml
4 application.app.k8s.io/
5 subscription.apps.open-cluster-management.io/
6 placement.cluster.open-cluster-management.io/
```

✓ dnstest.mesh.ad.jp +short

```
1 ▼ Run kubectl exec -it deployment/networktool -- dig @$LB_IPv4_VIP dnstest.mesh.ad.jp +short +retry=10
2 kubectl exec -it deployment/networktool -- dig @$LB_IPv4_VIP dnstest.mesh.ad.jp +short +retry=10
3 shell: /usr/bin/bash -e {0}
4 env:
5   LB_IPv4_VIP: 10.51.11.233
6 [I0318 21:19:01.585906 1999401 request.go:655] Throttling request took 1.18408648s, request:
   timeout=32s
7 Unable to use a TTY - input is not a terminal or the right kind of file
8 1.2.3.4
```

Demo内容(本番へ展開)

オンプレの本番環境にて新しいversionの
Unboundコンテナイメージを展開
Job名: Actions Runner Controller Demo



ACMによる各環境のサブスクリプションの
YAML情報に指定されているコンテナの
バージョンを更新(プルリク▶承認)



Prdのクラスタへ新バージョンのunboundの
コンテナイメージをデプロイ

Prd unbound 1.19.1 #18

Open bgl-tkgh wants to merge 2 commits into product from prd-unbound-1.19.1

Conversation 0

Commits 2

Checks 0

Files changed 2

Changes from all commits File filter Conversations

Filter changed files

product
unbound01.yaml
unbound02.yaml

product/unbound01.yaml

	@@ -35,7 +35,7 @@ spec:
35	image:
36	pullPolicy: Always
37	repository: image-registry.openshift-image-registry.svc:5000/default/biglobe-unbound
38	tag: 1.19.0
38	+ tag: 1.19.1
39	podAnnotations:
40	k8s.v1.cncf.io/networks: '[{"name": "tagv1", "ips": [REDACTED]}
41	securityContext:

product/unbound02.yaml

	@@ -35,7 +35,7 @@ spec:
35	image:
36	pullPolicy: Always
37	repository: image-registry.openshift-image-registry.svc:5000/default/biglobe-unbound
38	tag: 1.19.0
38	+ tag: 1.19.1

```
> Run oc login -p "" --username=[REDACTED] k
v Run oc apply -f /home/runner/_work/[REDACTED]/unbound01.yaml
1 Run oc apply -f /home/runner/_work/[REDACTED]/unbound01.yaml
4 application.app.k8s.io/[REDACTED]-product01 unchanged
5 subscription.apps.open-cluster-management.io/[REDACTED]-product01-subscription-1 configured
6 placement.cluster.open-cluster-management.io/[REDACTED]-product01-placement-1 unchanged
v Run oc apply -f /home/runner/_work/[REDACTED]/unbound02.yaml
1 Run oc apply -f /home/runner/_work/big4110-cicd/big4110-cicd/product/unbound02.yaml
4 application.app.k8s.io/rosa-unbound-product02 unchanged
5 subscription.apps.open-cluster-management.io/rosa-unbound-product02-subscription-1 configured
6 placement.cluster.open-cluster-management.io/rosa-unbound-product02-placement-1 unchanged
```

Demo内容(本番へ展開)



起動した新バージョンのunboundに
動作確認(DNSクエリ応答)



リリース完了！！

```
▼ ✓ dnstest.mesh.ad.jp +short
1 ▼ Run kubectl exec -it deployment/networktool -- dig @$LB_IPv4_VIP [redacted].jp +short +retry=10
2 kubectl exec -it deployment/networktool -- dig @$LB_IPv4_VIP [redacted].jp +short +retry=10
3 shell: /usr/bin/bash -e {0}
4 env:
5   LB_IPv4_VIP: [redacted]
6 Unable to use a TTY - input is not a terminal or the right kind of file
7 ;;
8 ;;
9 ;;
10 ;;
11 ;;
12 ;;
13 ;;
14 ;;
15 ;;
16 1.2.3.4

> ✓ Complete job
```

ROSA(ACM)でのUIからも各環境のデプロイ状況が見れる

アプリケーション

概要 詳細設定

Q unbound

4 / 172 x

▼ フィルター ▼

アプリケーションの作成 ▼

名前	タイプ	Namespace
rosa-unbound-stage01	サブスクリプション	default
rosa-unbound-stage02	サブスクリプション	default
rosa-unbound-product01	サブスクリプション	default
rosa-unbound-product02	サブスクリプション	default



ACMでどうやって各クラスタ(環境)に分けてデプロイしているか

名前	Namespace	ステータス	インフラストラクチャー	コントロールプレーンタイプ	ディストリビューションバージョン	ラベル	Node
☐ rosa-serverclient	rosa-serverclient	準備完了	Other	スタンドアロン	OpenShift 4.14.12 利用可能なアップグレード	env=tester あと 14 個 openshiftVersion-major=4 openshiftVersion-major-minor=4.14	☒
☐ rosa-client01	rosa-client01	準備完了	Other	スタンドアロン	OpenShift 4.14.12 利用可能なアップグレード	datacenter=omori device=rosa-client01 env=product openshiftVersion-major=4 openshiftVersion-major-minor=4.14 あと 14 個	☒
☐ rosa-client02	rosa-client02	準備完了	Other	スタンドアロン	OpenShift 4.14.7 利用可能なアップグレード	env=stage あと 14 個 openshiftVersion-major=4 openshiftVersion-major-minor=4.14	☒

クラスター単位でenvのラベルを付与。
ACMからデプロイする
サブスクリプションの定義のときに
Placementで指定する
(GitOpsOperatorで更に効率化できますが、
まずはシンプルな管理方法を試しました)

```
---
apiVersion: cluster.open-cluster-management.io/v1beta1
kind: Placement
metadata:
  labels:
    app: rosa-unbound-stage01
  name: rosa-unbound-stage01-placement-1
  namespace: default
spec:
  predicates:
    - requiredClusterSelector:
        labelSelector:
          matchExpressions:
            - key: 'env'
              operator: 'In'
              values:
                - 'stage'
```