急速に進化を続ける CNI プラグイン「Antrea」

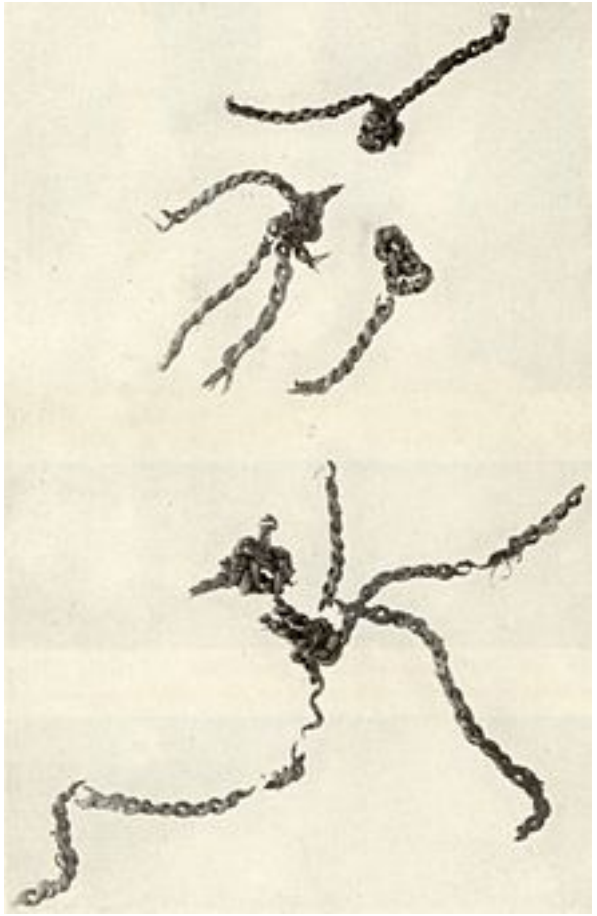
Oct. 16, 2020

CTO, North Asia (Japan, Korea and Greater China)

Motonori Shindo / motonori_shindo

vmware®

Antrea の由来



kubernetes

HELM

VELERO

ANTREA

Source: https://en.wikipedia.org/wiki/Antrea_Net

Project Antrea

NSX による管理を Kubernetes の実行環境に拡張

ANTREA = CNI + OvS Open vSwitch + kubernetes

K8S の動く環境ならどこでも動作

簡単に始められる – kubectl コマンド一発でインストール

K8S が動くあらゆる OS、プラットフォーム、クラウド、擬似環境などで動作

パブリッククラウドでも動作 – DIY or マネージド K8S

コミュニティ主導

オープンソース、誰でも利用可

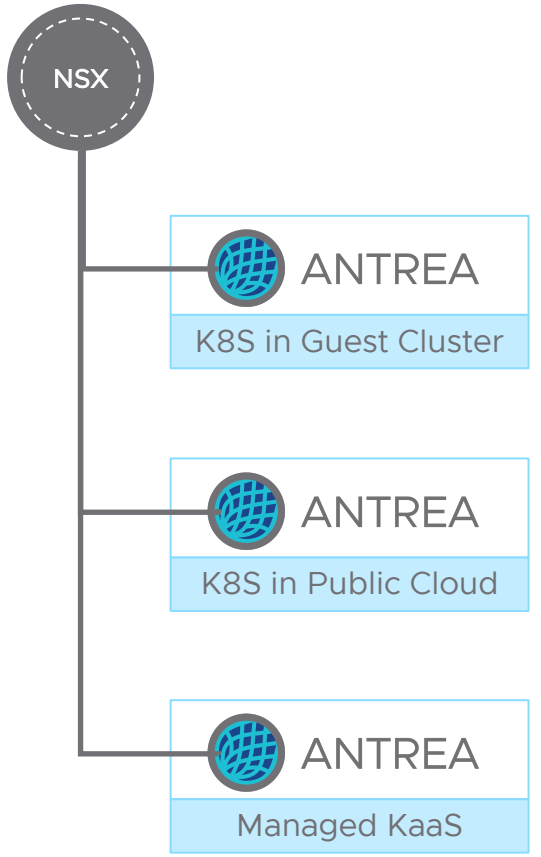
CNCF および K8S Network SIG に参加しているコントリビュータたちによるアクティブなコミュニティ

拡張性とスケーラビリティ

新機能の追加が容易で拡張しやすい

K8S の大規模環境にも適用可能なスケーラビリティ

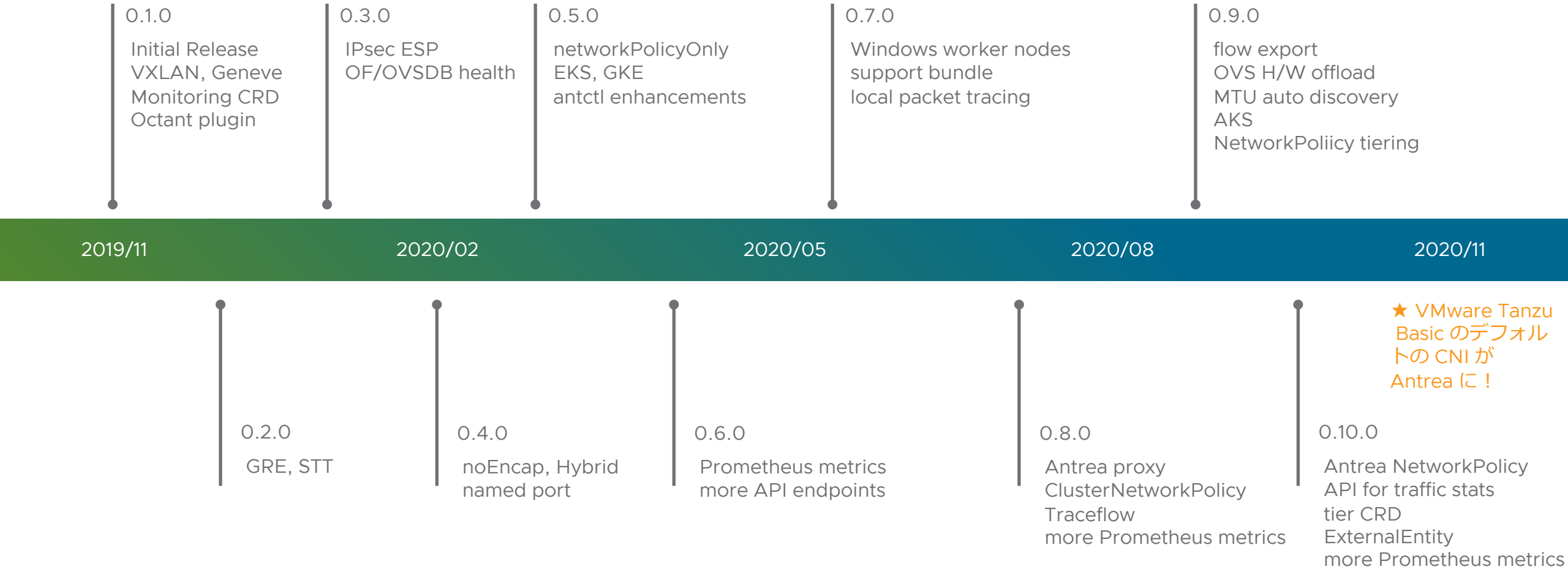
可視化とグローバルなポリシー配信に関して NSX と連携




```
kubectl apply -f https://github.com/vmware-tanzu/antrea/releases/download/v0.10.1/antrea.yml
```

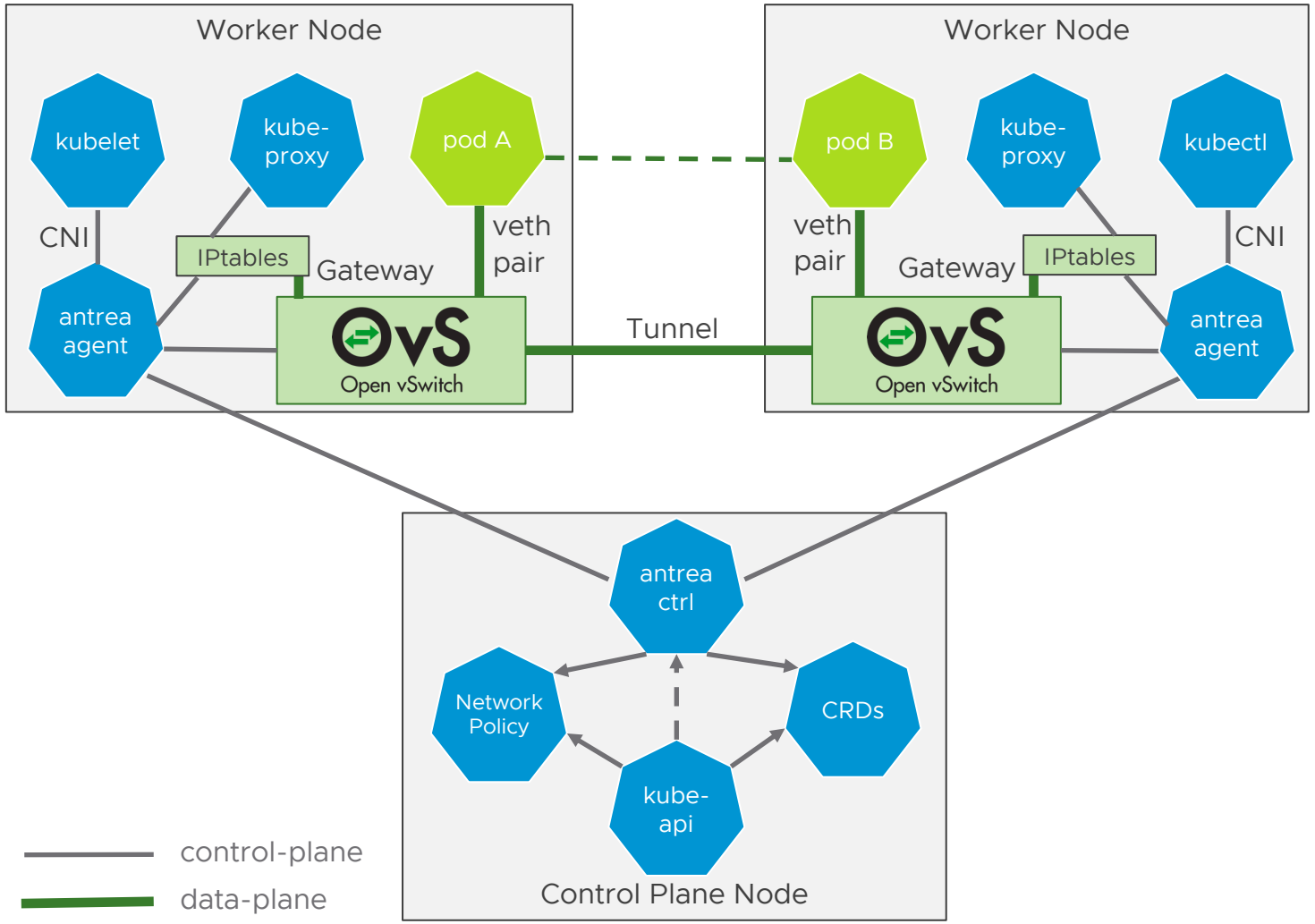
Antrea の進化

Every 4-6 weeks Release Cadence



Antrea アーキテクチャ

Open vSwitch が柔軟性と優れたパフォーマンスを実現



K8S クラスタネットワークをサポート

Antrea Agent

- Pod ネットワークインターフェースと OVS ブリッジの管
- ノード間のオーバーレイ トンネルの作成
- OVS へのネットワークポリシーの設定

Antrea コントローラ

- K8S ネットワークポリシーの計算し、結果を Antrea Agent への投入

Open vSwitch をデータパスに使用

- Antrea Agent が Open vSwitch に OpenFlow フローテーブルを設定
- Geneve、VXLAN、GRE または STT トンネルをノード間に設定
- Policy-only および no-encap モードをサポート

K8S 技術を使って構築

- API、UI、デプロイメント、コントロールプレーン、CLI などについて、K8S および K8S ソリューションを活用
- Antrea Controller と Agent は K8S コントローラと apiserver ライブラリを使用

Open vSwitch (OVS)

分散仮想スイッチのオープンソース実装

データプレーンに OpenFlow を使用

幅広いプラットフォーム(OS やハイパーバイザ）をサポート（Linux、 Windows、FreeBSD、NetBSD, など)

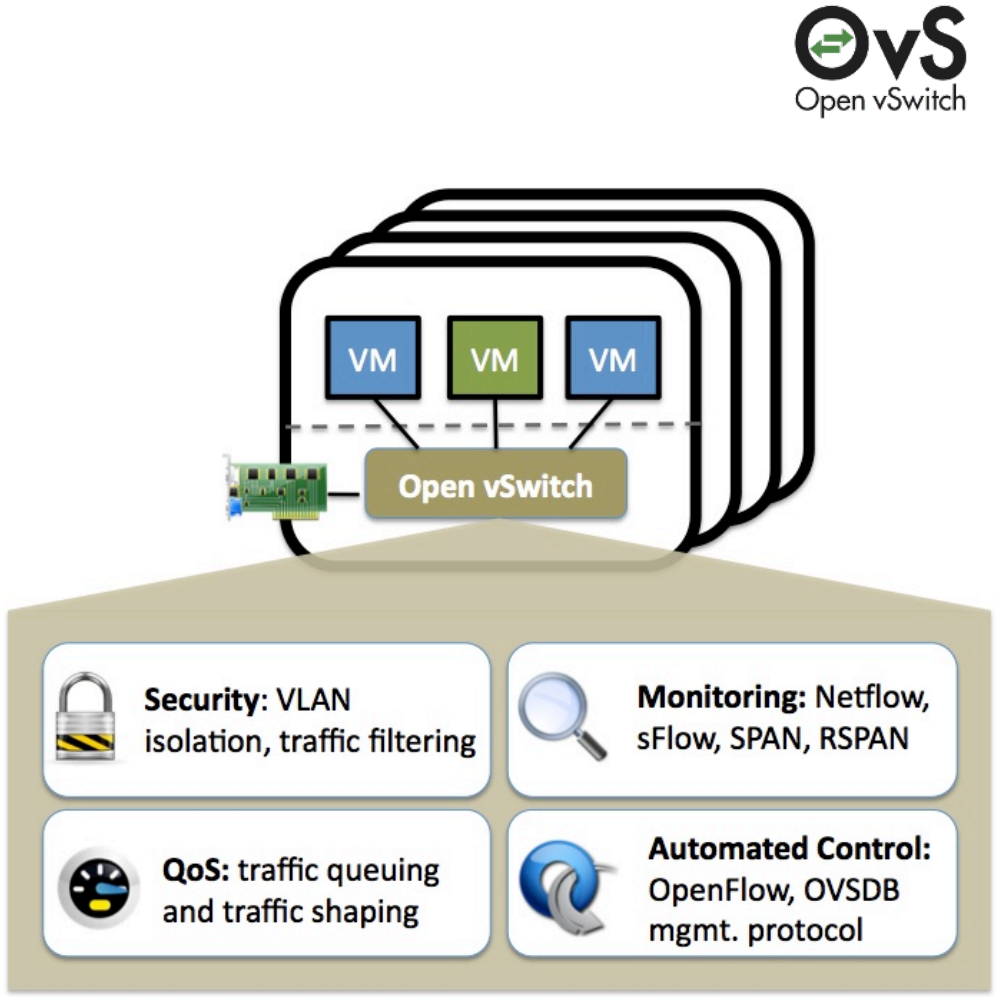
リッチな機能セット

- xFlow、RSPAN、LACP、802.1ag、など

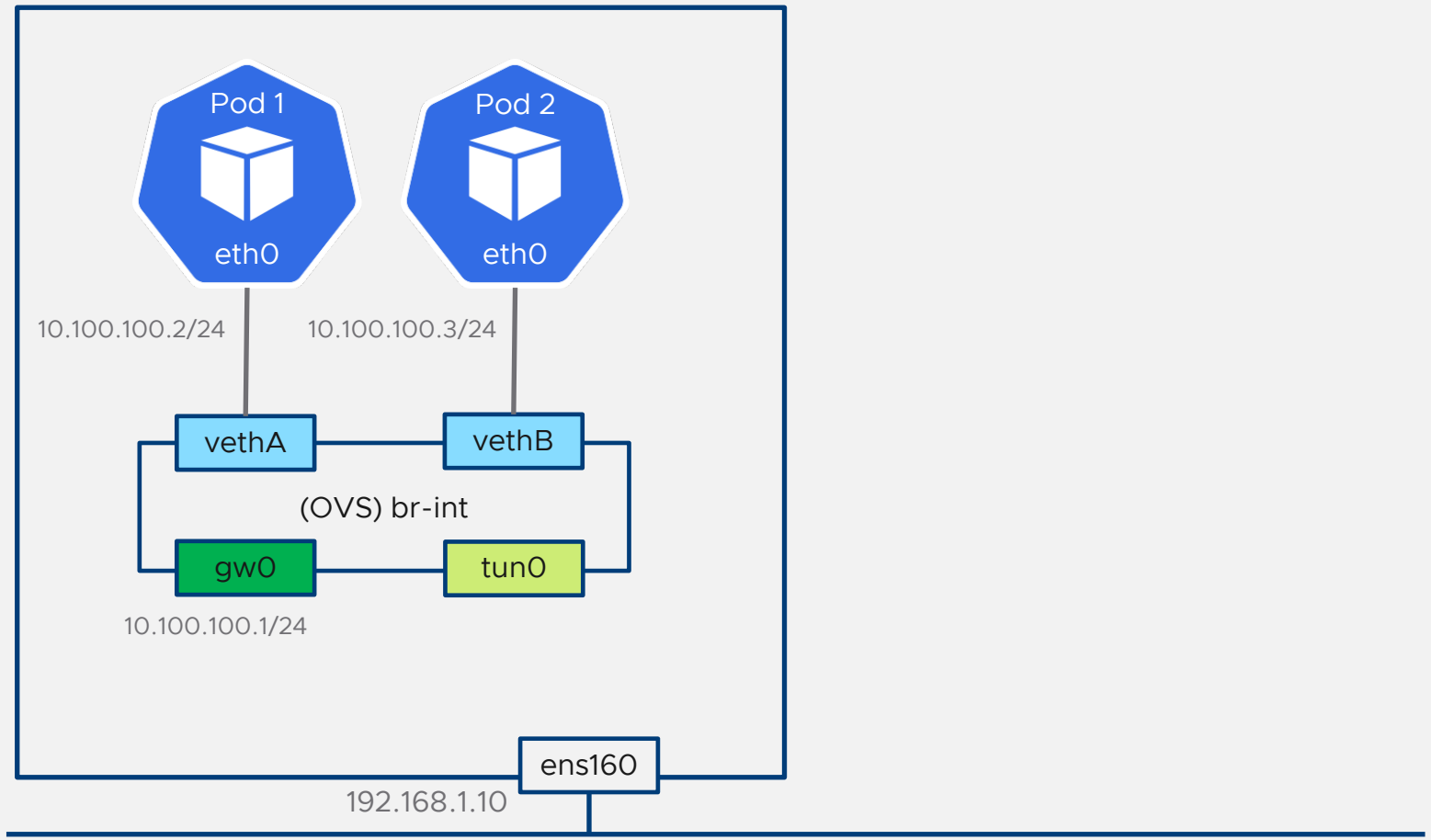
DPDK, AF_XDP 対応

豊富なエコシステム

- オフロード機能



Pod ネットワーク



OVS カーネルモジュールがインストールされている

Veth ペアが各 Pod ネットワーク namespace を OVS ブリッジに接続

K8S NodeIPAM コントローラが各ノードに一つのサブネットを割り当てる

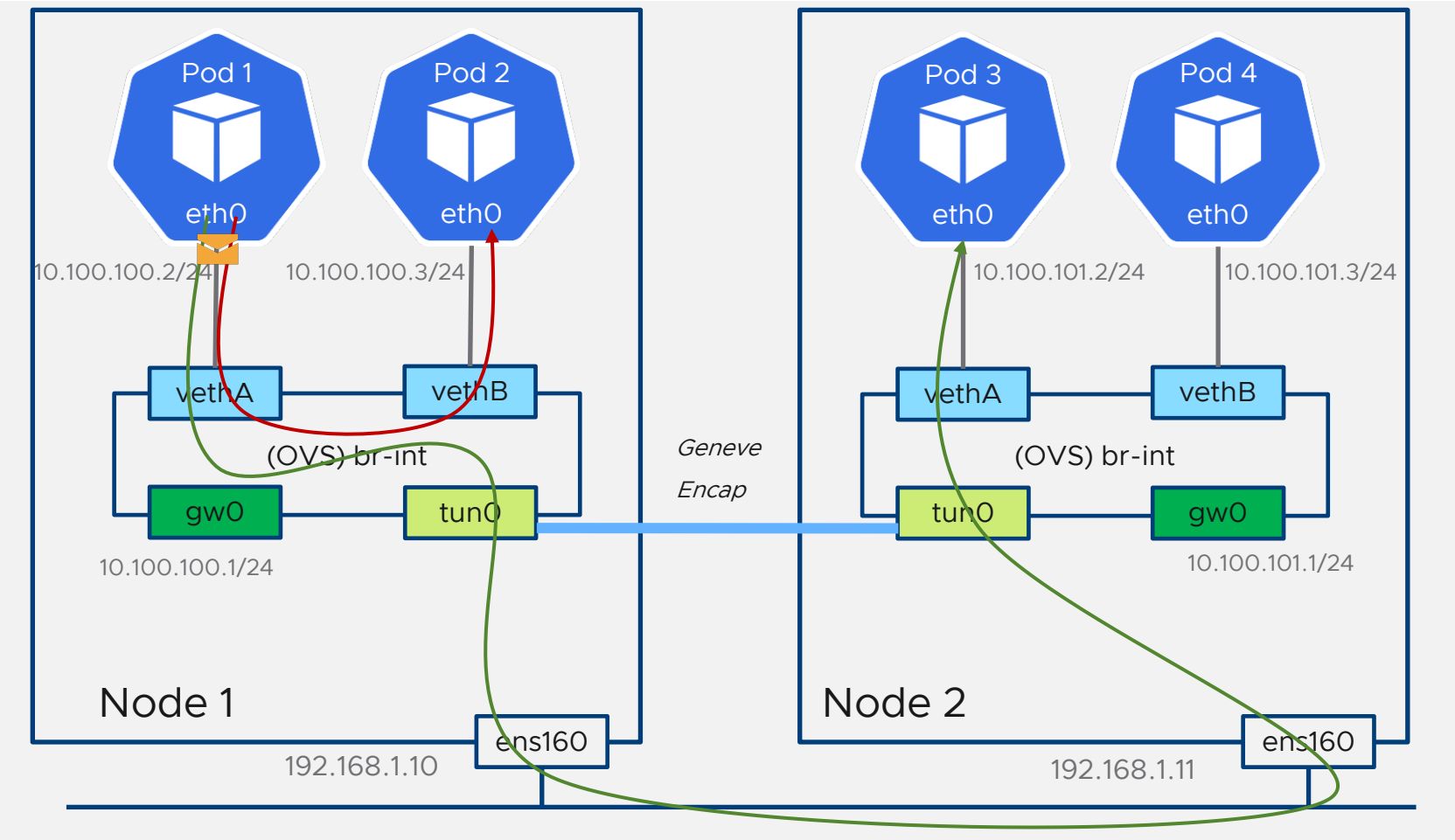
各 Pod の IP アドレスはノードのサブネットから割り振られる

サブネットの gateway IP アドレスが ‘gw0’ インターフェースに設定される

エンキャップされるトラフィック のためのトンネルインターフェース ‘tun0’ がOVS ブリッジに作成される

Pod から出るパケットの流れ

Pod 間の通信 (Encap モード)



ノード内

- 同じノード上の Pod 間の通信は、単純に OVS ブリッジ経由でパケットが届く。

ノード間

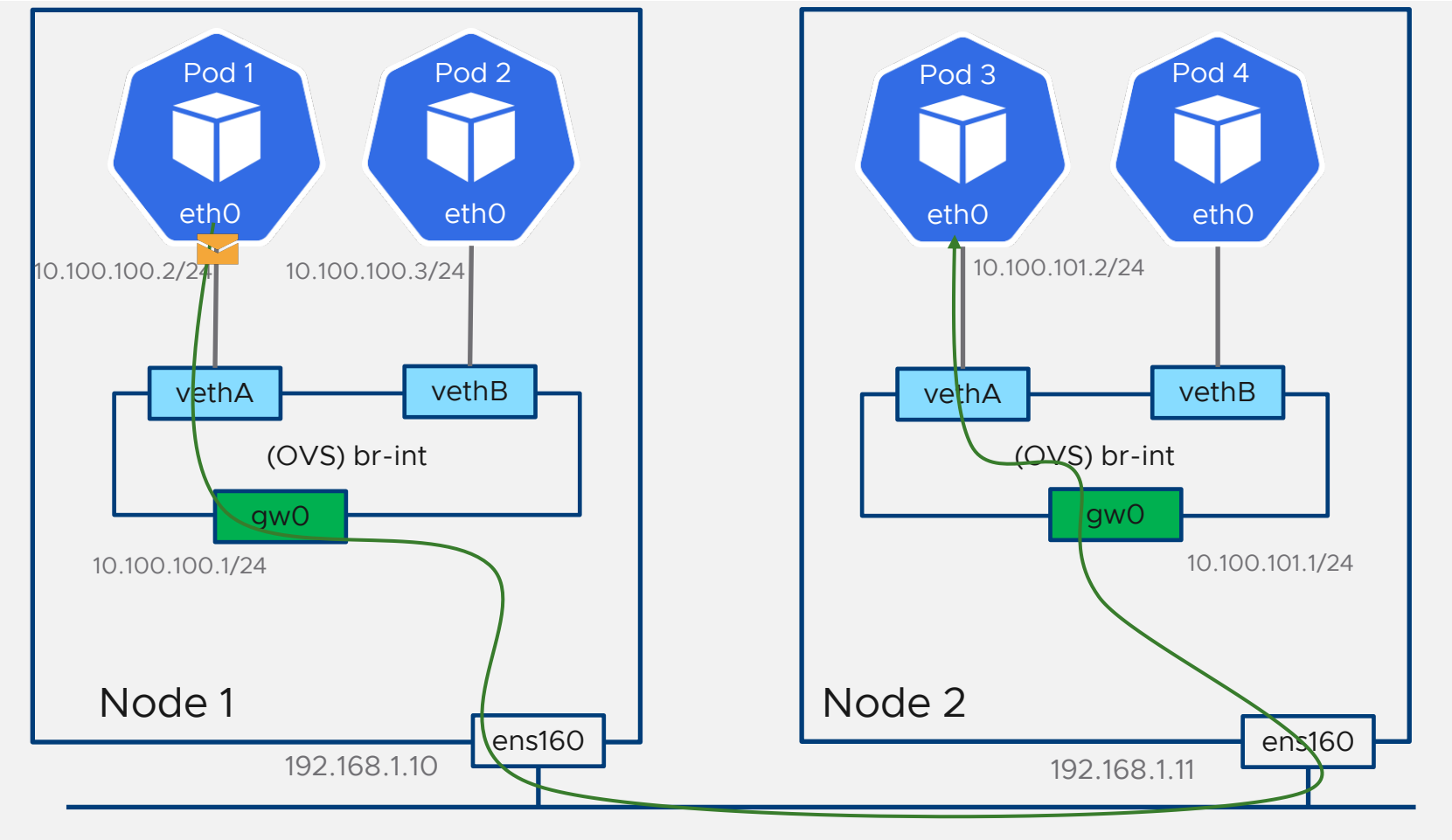
- 異なるノード上の Pod 間の通信の場合は、物理ネットワークにパケットが出る前にエンキャップされてから物理ネットワーク側に出る

ノード内 Pod 間通信

ノード間 Pod 通信

Pod から出るパケットの流れ

Pod 間の通信 (NoEncap モード)



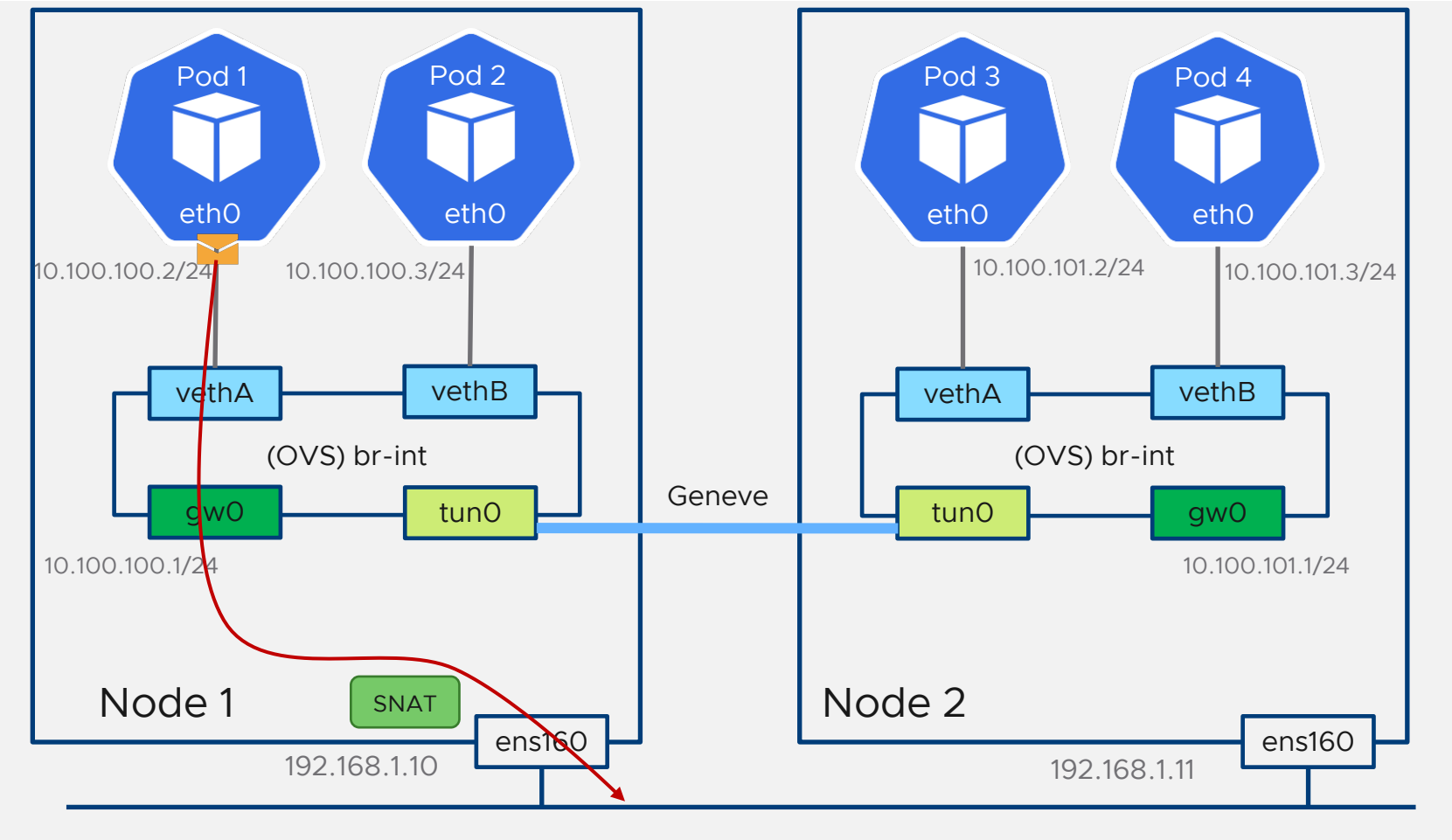
ノード間

- Pod から gw0 を経由して外に出るパケットのソース IP アドレスは変わらない
- パケットは送信先 next hop に送られる
- ノードが L2 隣接でない場合は、pod の経路は Antrea と物理ネットワークの間で交換する必要がある。

ノード間 Pod 通信

Pod から出るパケットの流れ

Pod から外部に出る通信



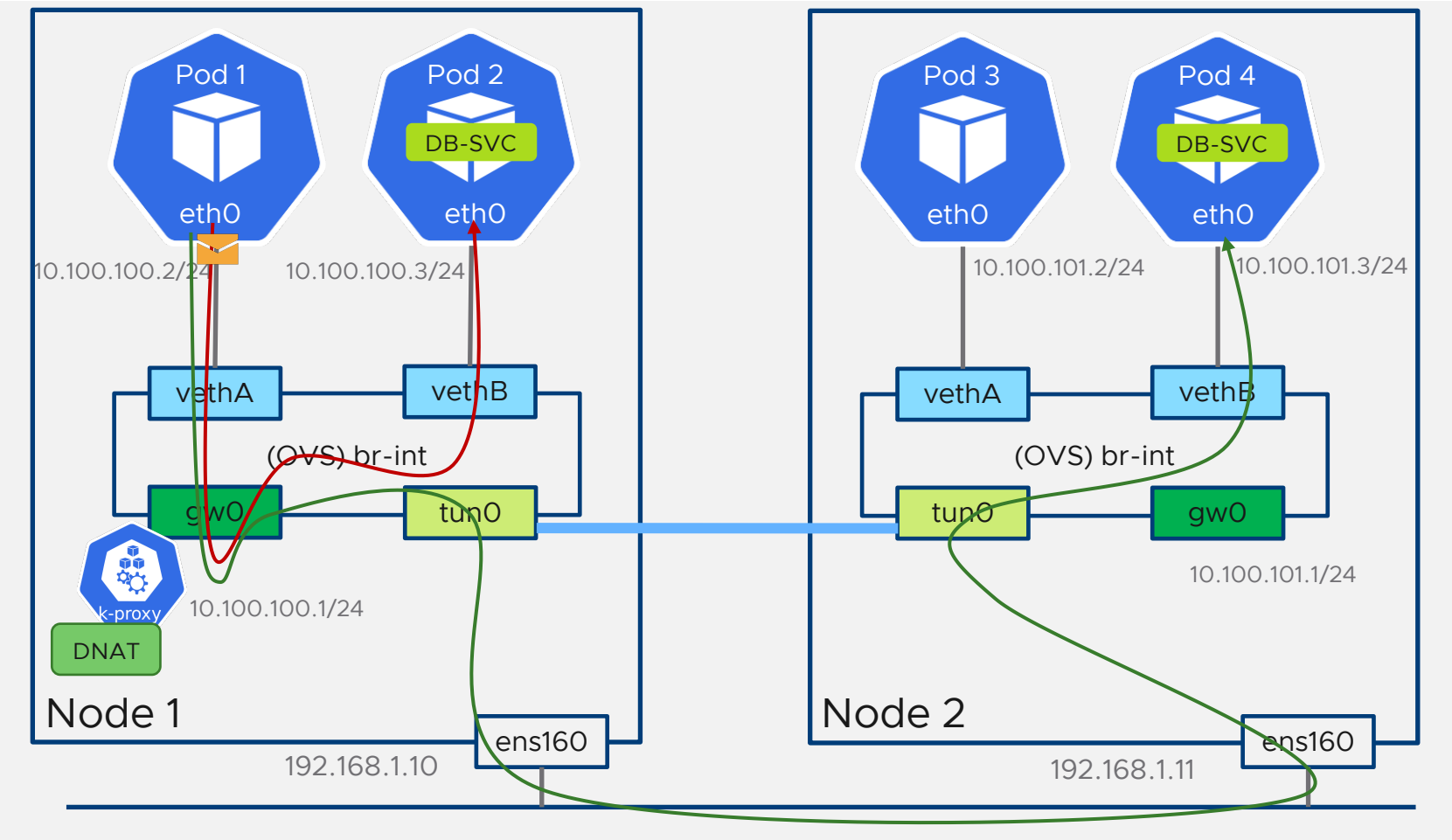
外部へのトラフィック

- Antrea Agent が Pod から外部に出るトラフィックを SNAT するための iptables (MASQUERADE) ルールを作成
- Pod から外部に出るトラフィックは Node IP に SNAT される
- SNAT はノードの iptables で処理される

Pod から外部に出る通信

Pod から出るパケットの流れ

Pod から Service への通信 (デフォルトの kube-proxy での動作)



ノード内の Pod から Service へ

- トラフィックは gw0 に送られ、kube-proxy によって DNAT されてから Service の IP アドレスに送られる

ノード間の Pod から Service へ

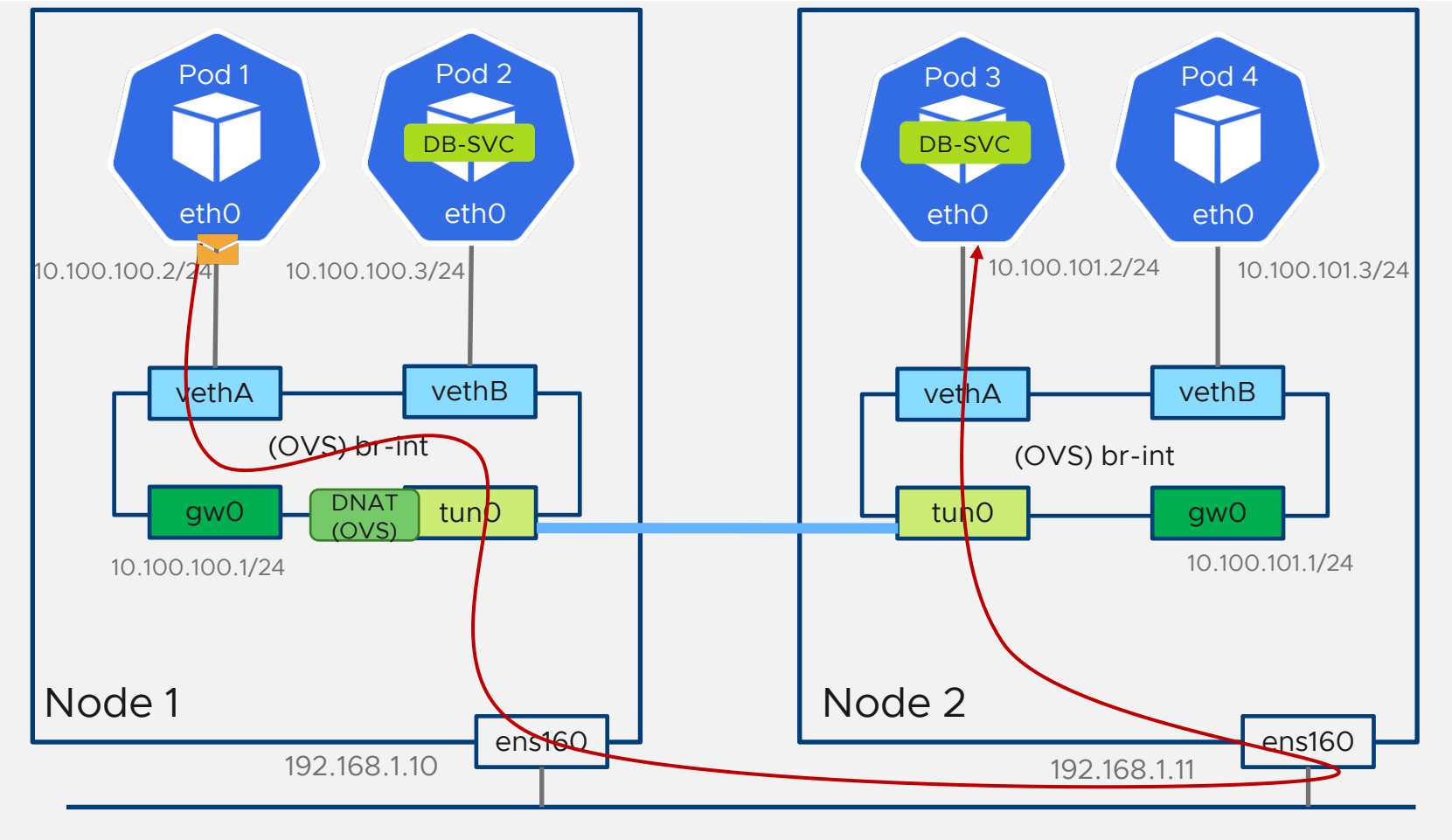
- トラフィックは gw0 に送られ、kube-proxy の iptables によって DNAT されて、送信先ノードの Service IP アドレスにトンネルされて送られる

ノード内 Pod から Service への通信

ノード間 Pod から Service への通信

Pod から出るパケットの流れ

Antrea Proxy による Service ロードバランシング



ノード間の Service トラフィック

Pod から出るトラフィック の DNAT は OVS で実行される

iptables によるコンテクストスイッチ処理を回避

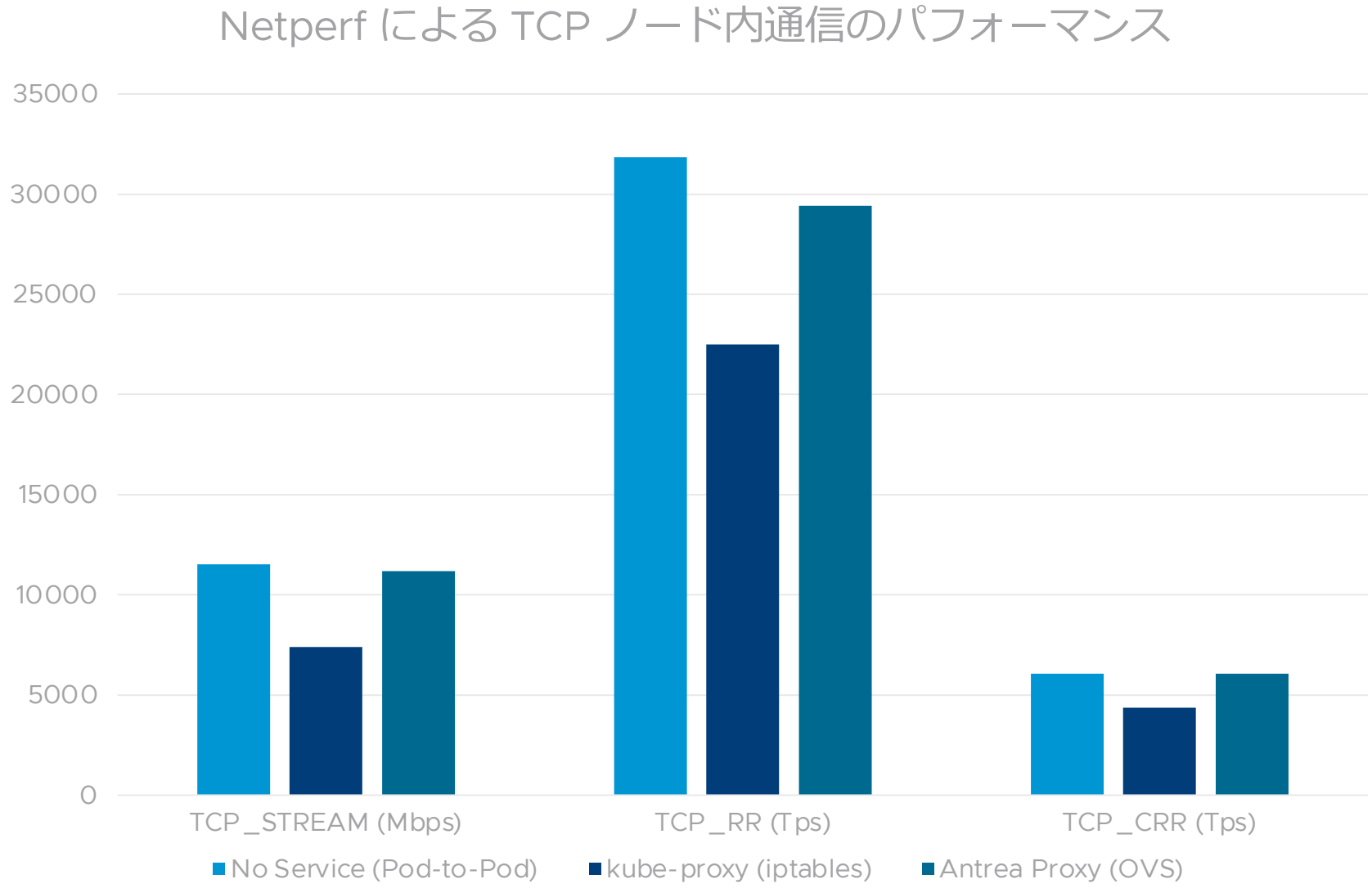
利点

トンネルポートへパケットを直接出せるため、パフォーマンスが向上

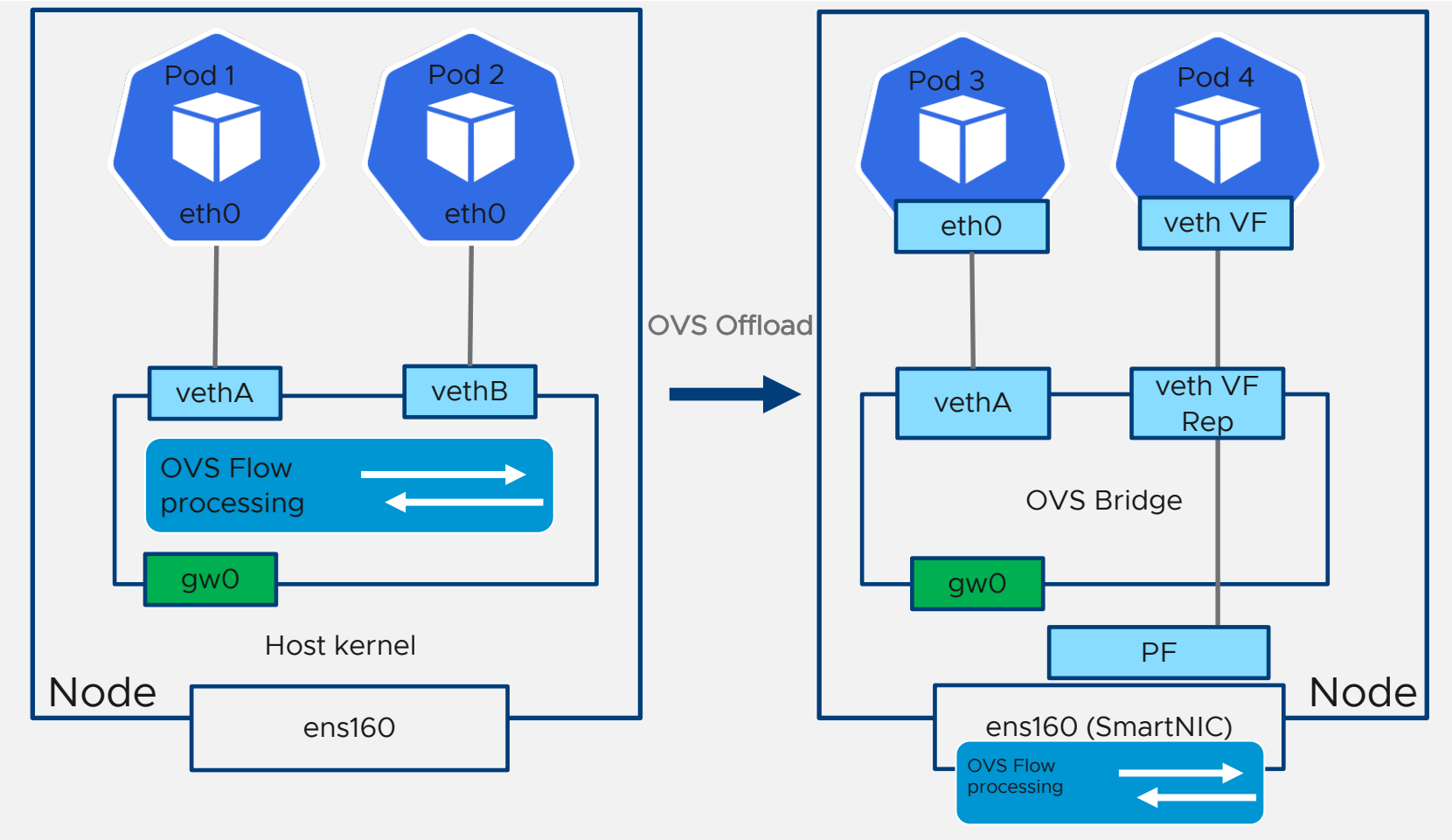
ノード間 Pod から Serrvie への通信

Antrea Proxy のパフォーマンス

Antrea Proxy v.s. kube-proxy (IPtables/IPVS)



OVS の H/W によるオフロード



OVS オフロードは SR-IOV と Multus CNI を使い、フロー処理を NIC にオフロード

各 Pod には VF (Virtual Function) が割り振られる

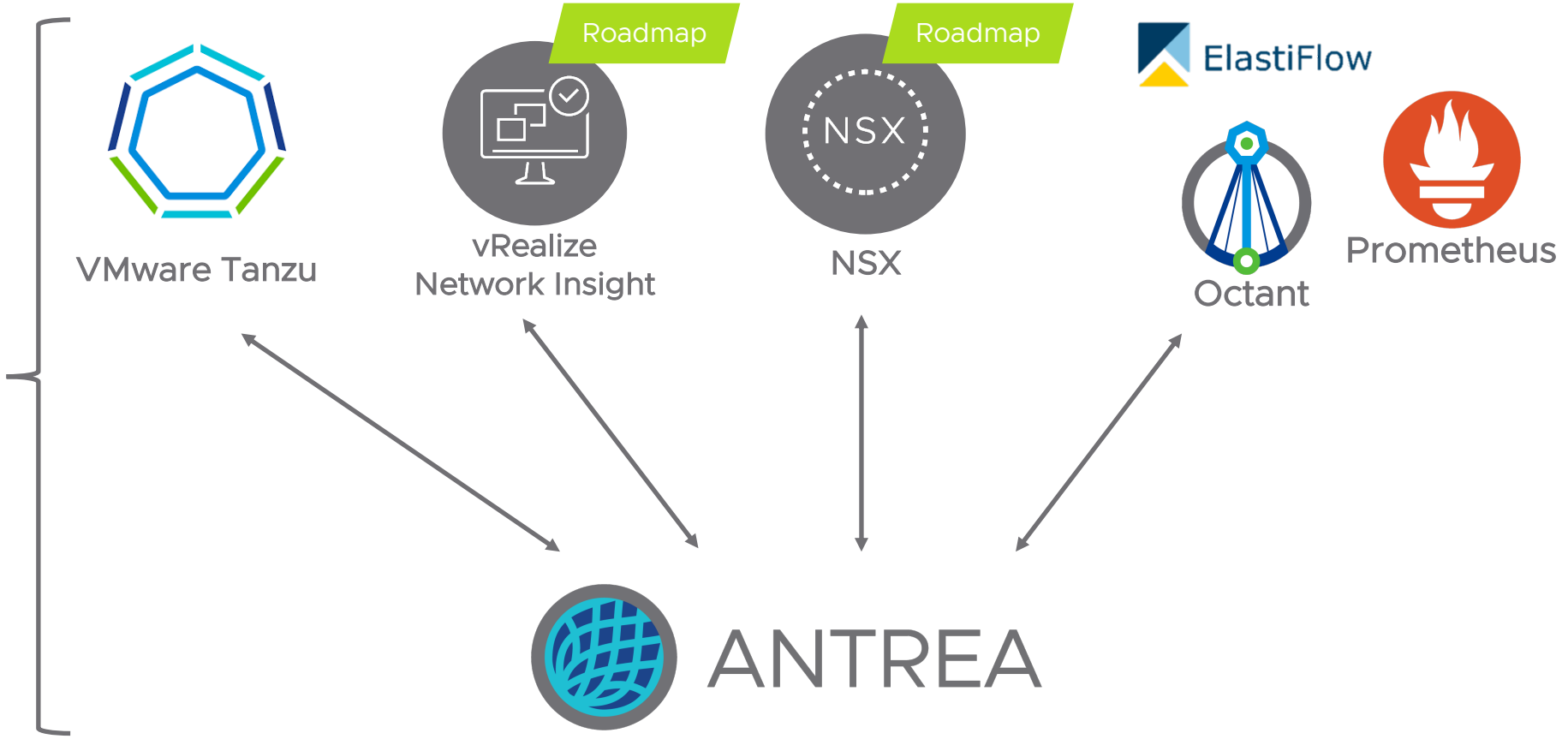
Pod の VF は OVS ブリッジ上の VF の representor に接続される

PF (Physical Function)。SR_IOV をサポートする物理 NIC

オペレーションの容易性とトラブルシュート

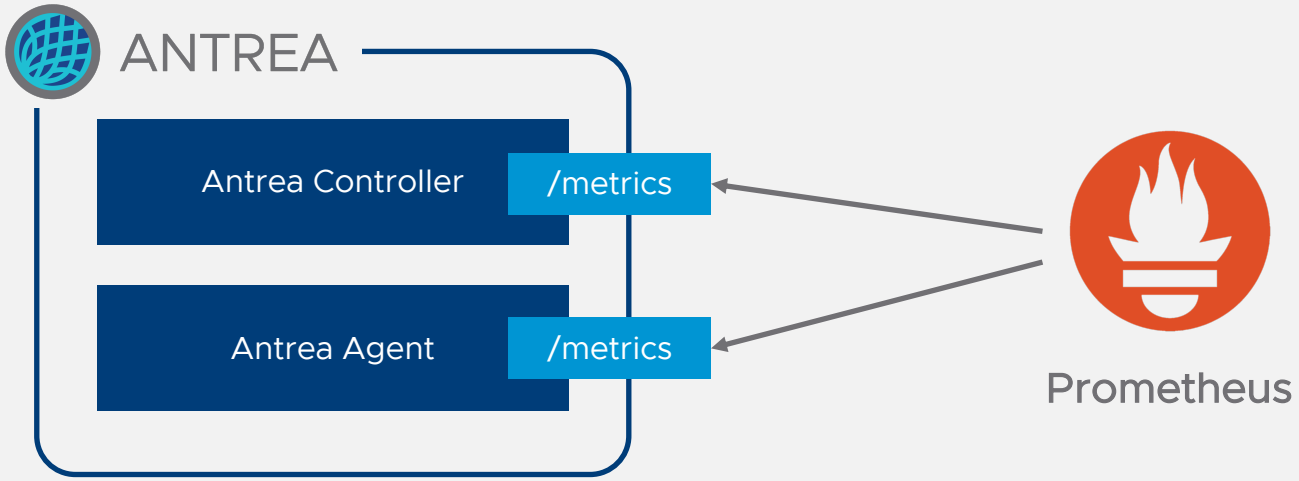
ポリシー配布*
可視化
- IPFIX Flow ログ
- メトリック
- インベントリ*

トラブルシュート
- TraceFlow
- サポートバンドル



メトリック

Antrea エージェントおよびコントローラが Prometheus エンドポイントにメトリック提供



機能

コントローラとエージェントがネイティブに Prometeus にメトリックを公開可能

- ノードあたりのネットワークポリシー数
- テーブルあたりの OVS フロー数
- OVS flow 操作の遅延
- ネットワークポリシーの計算による遅延
- …

利点

Antrea のコントール&データプレーンの正常性監視

ネットワークトラフィックのメトリックを把握

オペレーションに関する通知

Grafana で統計情報を可視化

メトリック

Prometheus メトリックを Grafana などのツールで可視化

機能

コントローラとエージェントがネイティブに Prometeus にメトリックを公開可能

- ノードあたりのネットワークポリシー数
- テーブルあたりの OVS フロー数
- OVS flow 操作の遅延
- ネットワークポリシーの計算による遅延
- …

利点

Antrea のコントール&データプレーンの正常性監視

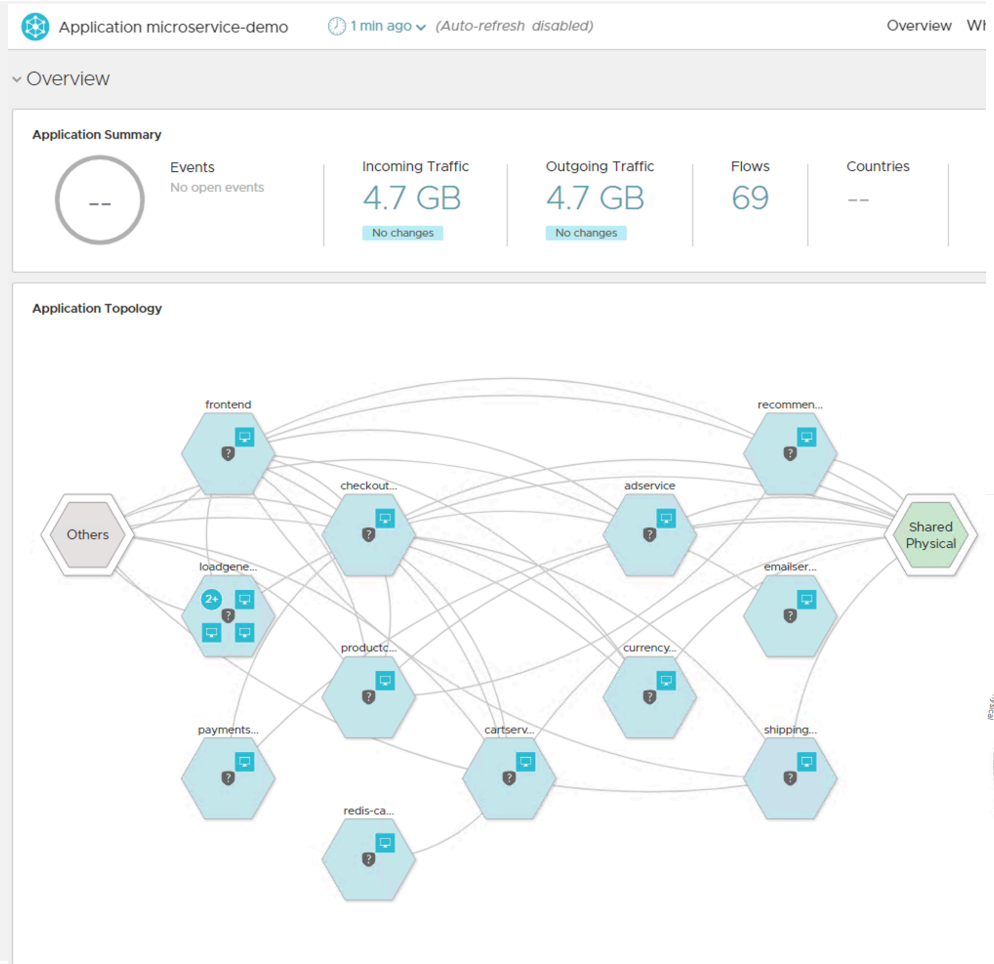
ネットワークトラフィックのメトリックを把握

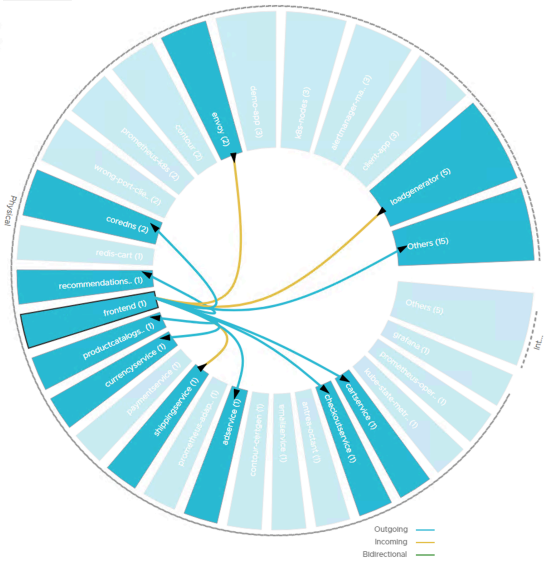
オペレーションに関する通知

Grafana で統計情報を可視化

ネットワーク・フローの監査

エクスポートされた IPFIX フローを記録・可視化し、クラスタの状態を分析





機能

すべてのクラスタのトラフィックを IPFIX でエクスポート

- コネクション数
- 各コネクションの使用帯域
- ノード間の使用帯域
- Service の合計使用帯域

Prometheus のメトリックを補完

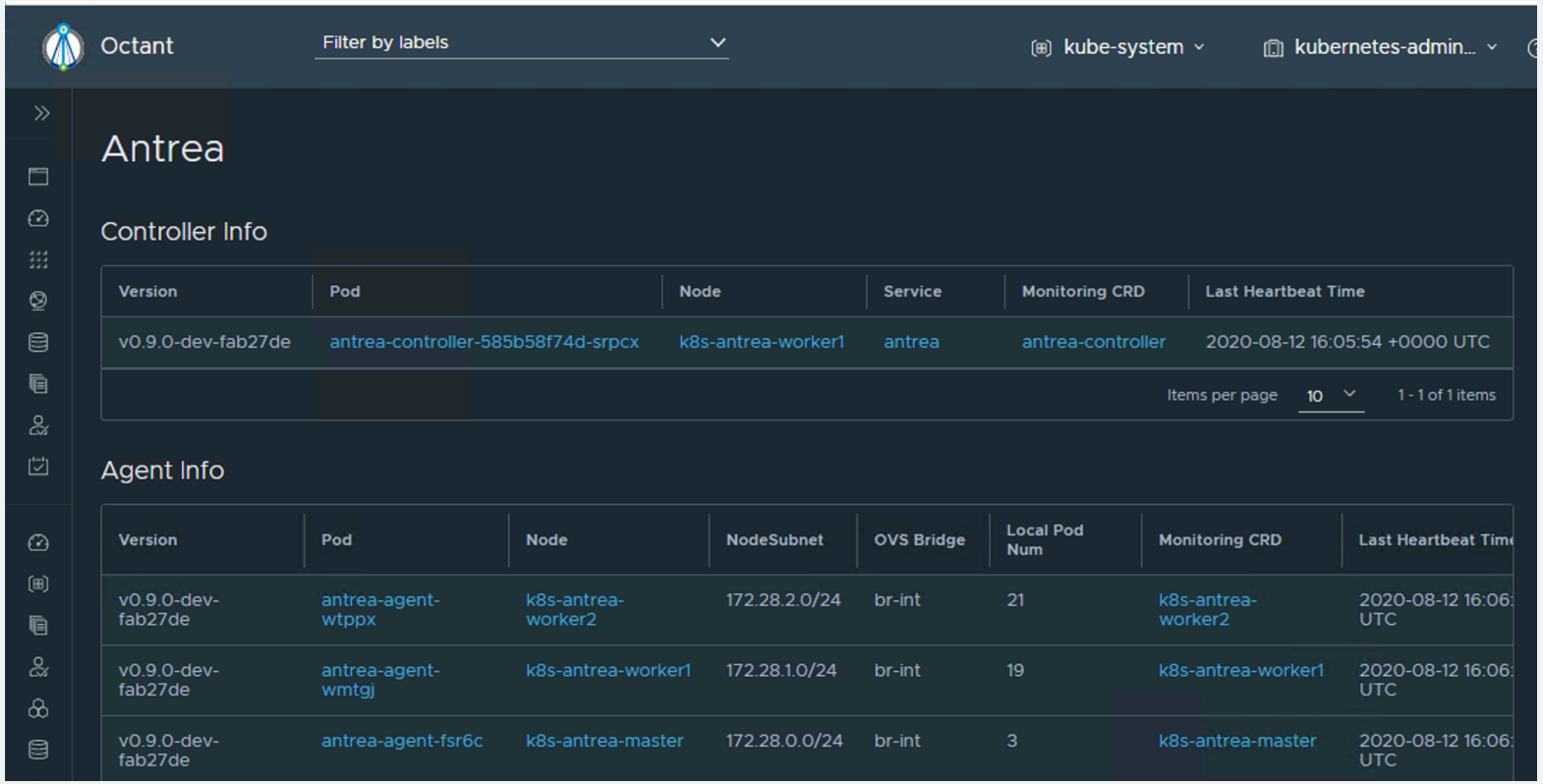
利点

IPFIX レコードが Kubernetes のコンテクストを含んでいる (Namespace, Name, Labels, ...)

クラスタトラフィックの可視化

コントロールプレーンの正常性確認とステータス

Octant プラグインによる Antrea のモニタとトラブルシュート



機能

インベントリ情報を CRD として提供

- コントローラ
- エージェント

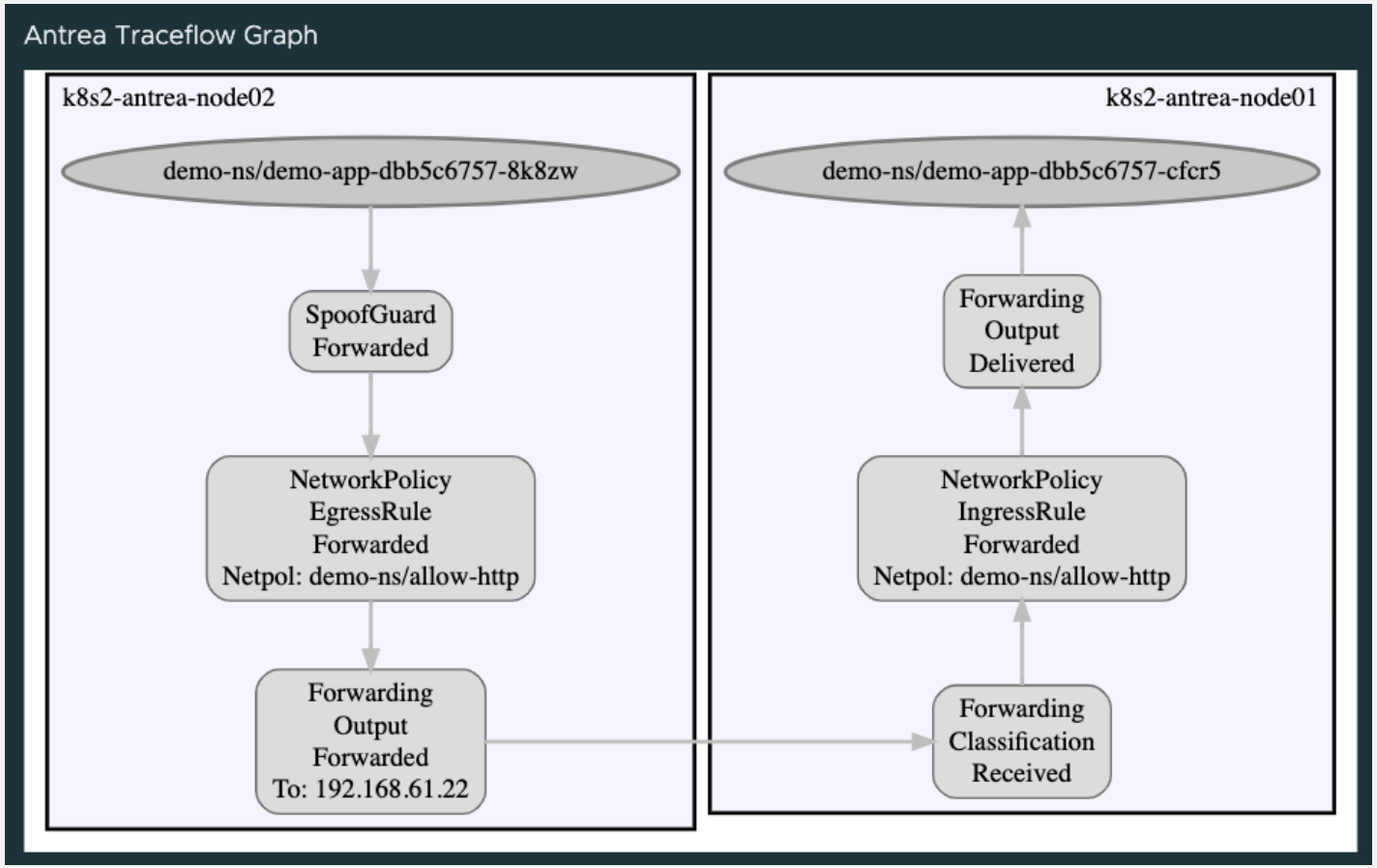
Octant プラグインはこれらのインベントリのための UI を提供

利点

Antrea ステータスダッシュボード

Antrea Traceflow

Ttraceflow パケットのインジェクションによる Pod トラフィックの調査



機能

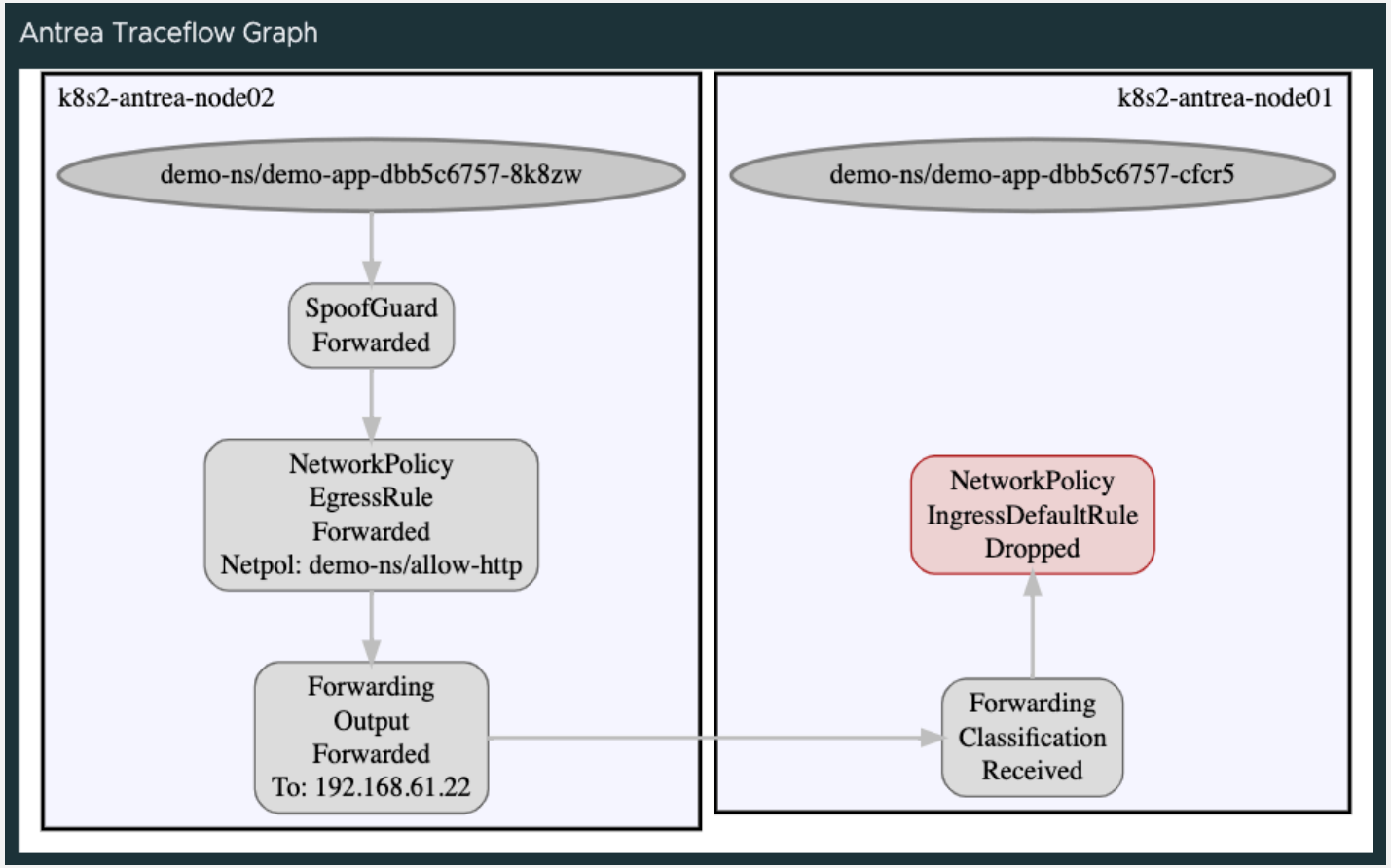
Traceflow CRD が OVS のパケットインジェクションを設定し、ネットワークのエンドポイント間をトレース

利点

ポリシー適用の影響を確認でき、ネットワークの問題をいちはやく見つけることができる

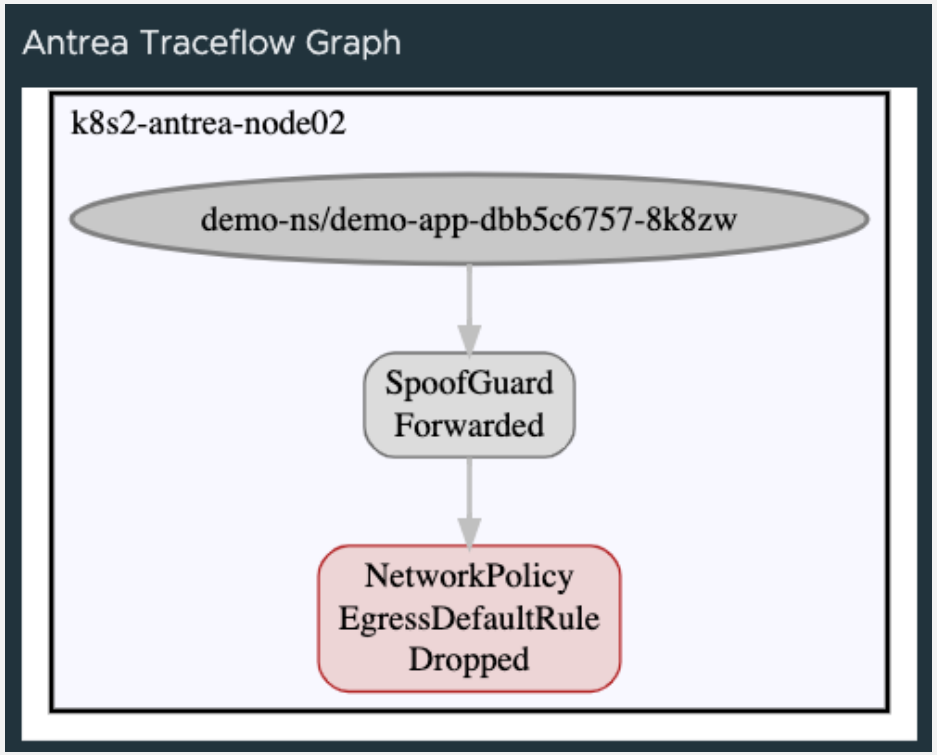
- 模擬トラフィックを作成する必要なし
- パケットドロップ等の networkpoliicy による影響をわかりやすく表示する
- 間欠接続障害を発見することができる

Antrea Traceflow によるトラブルシュート



この例では、Ingress 方向のネットワークポリシーがリクエストパケットをドロップし、不達になっていることが分かる。

Antrea Traceflow によるトラブルシュート



この例では、egress 方向のポリシーによってパケットがドロップしていることが分かる。

Antrea Demo

Antrea ロードマップ

https://antrea.io/docs/master/roadmap/

Windows サポートの拡充

IPv6 Pod ネットワークサポート

Antrea Network Policy の拡充

ネットワーク診断および可視化機能の拡充

より柔軟な IPAM 機能

Egress / SNAT ポリシー

NFV / Telco ユースケース（Multus、SRIOV、Network Service Chain、など）

NetworkPolicy のスケール&パフォーマンステスト

DPDK or AF_XDP の対応

各種 CNI 比較

	Flannel	Calico	Cilium	Antrea
Datapath	N/A	iptables / eBPF	eBPF	OVS
Network Policy	No	Yes	Yes	Yes
Policy Tiering	N/A	No	No	Yes
Encryption	No	Yes	Yes	Yes
Metric Export	No	Yes	Yes	Yes
Flow Export	No	No (Log only)	No	Yes
Windows support	No	No	No	Yes
K8S only	Yes	No (K8S, OS)	Yes	Yes
Resource Consumption	Low	Low	High	Low
Maturity	Mature	Mature	New	Very New

References

Project Antrea Official Docs

- https://antrea.io/

Antrea GitHub repo

- https://github.com/vmware-tanzu/antrea

CNCF Webinar: “Securing and Accelerating Kubernetes CNI Data Plane with Project Antrea and NVIDIA Mellanox ConnectX SmartNICs”

- https://www.cncf.io/webinars/securing-and-accelerating-the-kubernetes-cni-data-plane-with-project-antrea-and-nvidia-mellanox-connectx-smartnics/

Blog: “Antrea – Yet Another CNI Plug-in for Kubernetes” (日本語)

- https://blog.shin.do/2020/01/antrea-yet-another-cni-plugin-for-kubernetes/

Thank You

vmware