DNSサーバ運用の苦悩 ～さくらインターネット編

さくらインターネット(株) http://www.sakura.ad.jp/
技術部　大久保修一　ohkubo@sakura.ad.jp

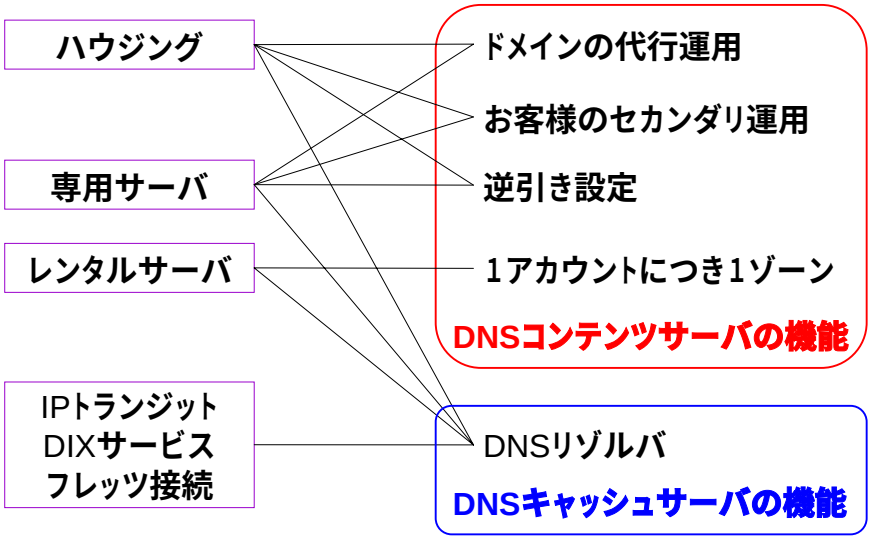
自己紹介

- さくらインターネットにてネットワークの仕事を担当
  - 実はDNSの主担当ではありません・・・
- さくらインターネット SAKURA Internet とは？
  - 東証Mothersに上場(証券コード：3778)しています
  - AS9370(東京), AS9371(大阪), AS7684(IPv6), AS2.8
  - いわゆるiDC(インターネットデータセンター)です。
  - ハウジングサービス
  - 専用サーバサービス
  - レンタルサーバサービス
  - フレッツ接続/IPトランジット/DIXサービス
  - オンラインゲーム

今日のAgenda

- 弊社サービスとDNSの関係
- DNSコンテンツサーバの運用について
- DNSキャッシュサーバの運用について
- まとめ

弊社サービスとDNSの機能



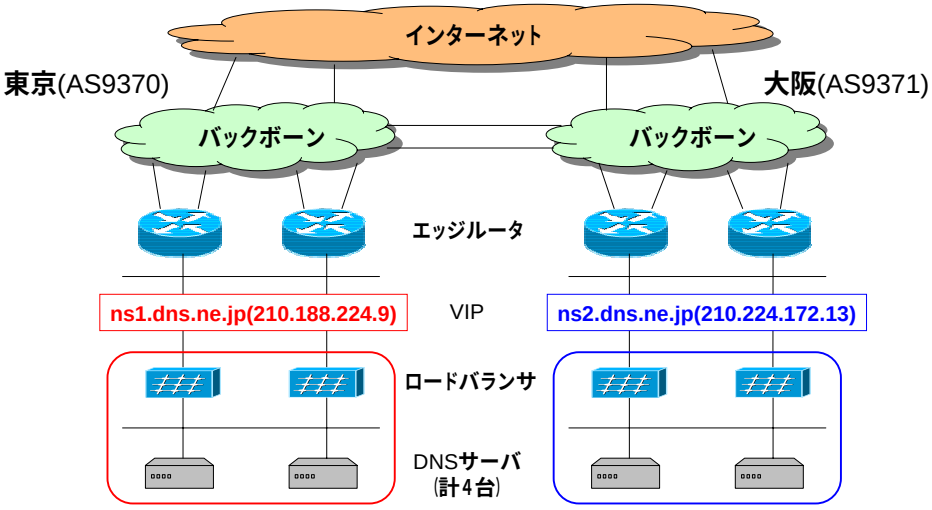
DNSコンテンツサーバの運用課題

- サービスが停止すると・・・
  - iDCの機能が停止する
  - 会社の存続にもかかわる (汗
  - 重要なインフラ！
- 運用上の課題
  - DNSサーバの冗長化
  - 大量のゾーン数

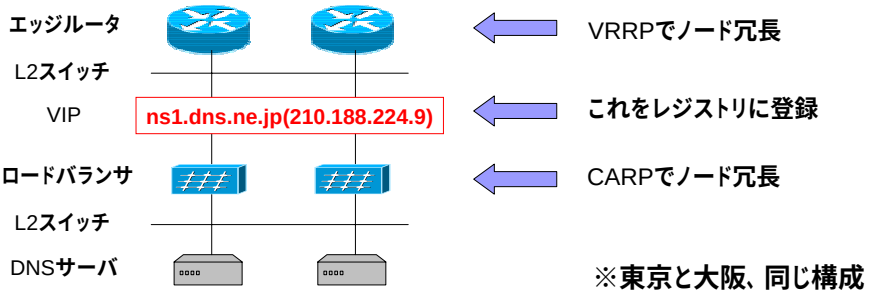
冗長化の設計方針

- 弊社が運用するドメイン
  - ホスティング等で用意するドメイン
  - お客様ドメインの代行運用
  - お客様のセカンダリ
  - →弊社のコンテンツサーバ2台をレジストリに登録
- 2台のコンテンツサーバの拠点冗長
  - 電源断、ファイバ断などにより、同時にダウンしないように
- 各コンテンツサーバのサーバ冗長
  - 基本的に各コンテンツサーバを落とさない
  - 片方が生きていればDNSの解決はできるが、アクセス遅延（品質低下）が発生する。

2台のコンテンツサーバの拠点冗長



コンテンツサーバの拠点内冗長化



- ロードバランサ
  - FreeBSD 5.x + ipnatベース
  - ヘルスチェックプログラムは独自実装(perl + digコマンド)
  - DNS notifyパケットを両方のDNSサーバに送信するプログラム (キャプチャにbpf、送信にrawsocketを利用)

日々増大するゾーンに対応するには？

- ゾーン数の増大
  - さくらのレンタルサーバ開始時より
  - → 1アカウントにつき1ゾーン
  - 2007年11月現在約40万ゾーン
  - 毎月10,700ゾーン程度増加
- BIND9の問題点
  - rndc reconfigコマンド実行時、5分間程度停止する
  - デーモンが落ちると、起動に2時間程度かかる
  - メモリ消費量が大きい (約1.8GB)

BIND9では扱えなくなる

ANSの導入

- Nominum社製ANSの導入 http://www.nominum.com/
- 起動時間：1分以下
- ゾーンの追加削除時：クエリロスなし
- 消費メモリ：約600MB
- 24時間中、パケットを落とすのは1〜2秒程度
  - ガベージコレクション処理中
  - FreeBSDのマルチスレッドの仕様らしい。。
  - Linuxだと大丈夫らしい。

非常に安定するようになった！

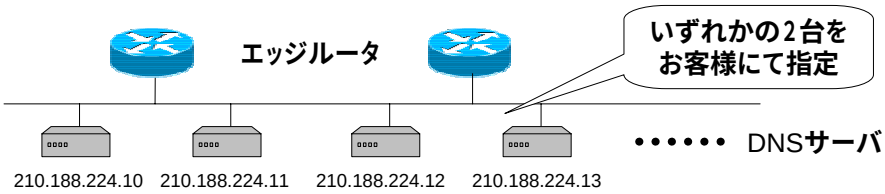
DNSキャッシュサーバの運用課題

- サービスが停止すると・・・
  - iDCの機能が停止する
  - 会社の存続にもかかわる
  - 重要なインフラ！

DNSコンテンツサーバと同じ

- 課題
  - DNSサーバの冗長化
  - 多量のクエリをどう捌くか

2004年9月以前：単体DNSサーバを並べる



問題点

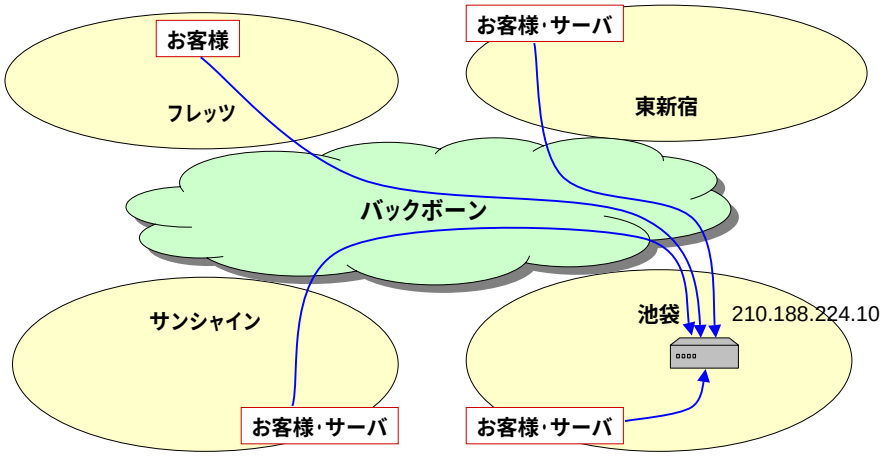
- 冗長化ができていない
  - 2台指定いただくが、メイン側が落ちるとアクセス遅延発生
- 提供IPアドレスの種類が増える
  - 東京では6種類、大阪では2種類
- スケーラビリティの問題
  - 1台で捌ける能力を超えるとサーバのスペックアップしか対応方法がない

2004年9月～ Anycastによる分散冗長化

- IGP(OSPF) Anycastを利用
- 同一IPアドレスを振ったサーバを分散配置
  - IPアドレスはループバックインターフェイスに
- キャッシュサーバのIPアドレス、/32の経路を各サーバからバックボーンに広報
- そのIPアドレス宛のパケット
  - ネットワーク的に近くで処理
- サーバの故障時
  - 経路広報がストップ
  - ルーティングが別のサーバに向く
  - 切り替る

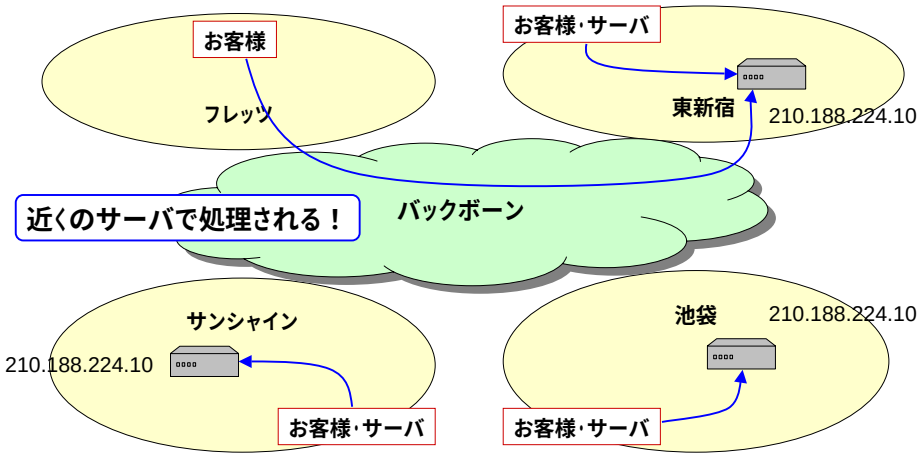
従来の構成：Anycastしない場合

クエリは一台の物理サーバで処理される



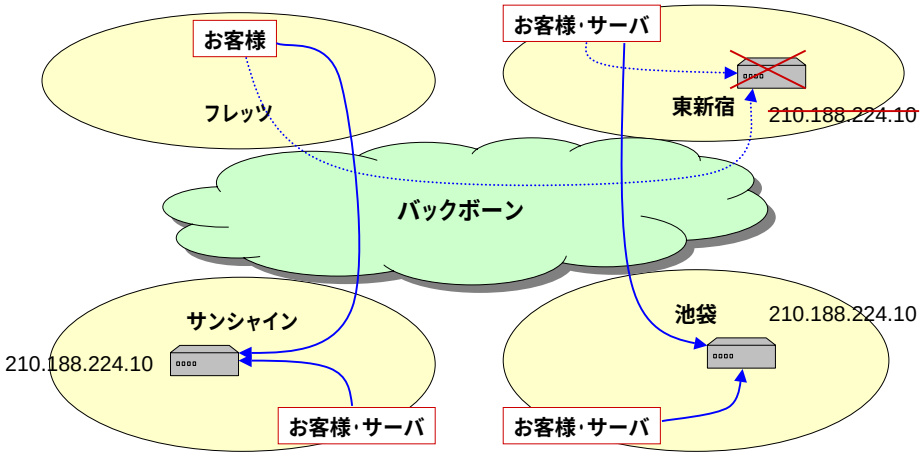
Anycast化後の動作

各データセンターに同一IPアドレスのサーバを配置



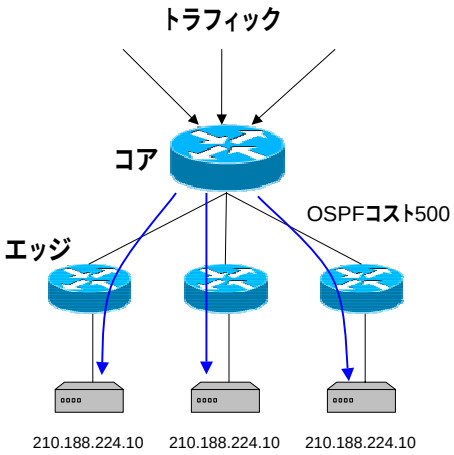
障害発生時の切り替り

近くの他拠点のサーバに自動的に切り替る



Anycastの応用：ECMPを使った分散

ECMP = Equal Cost Multi Path



- コアから各DNSサーバまでのOSPFコストが同じになるように接続
- トポロジーに強く依存
  - エッジルータのアップリンク故障、迂回
  - バックボーンの構成変更
  - →トラフィックが偏る
  - →最悪、アクセス遅延発生
- 運用が難しい。。。

ロードバランサの導入

- AnycastやECMPによる分散に限界

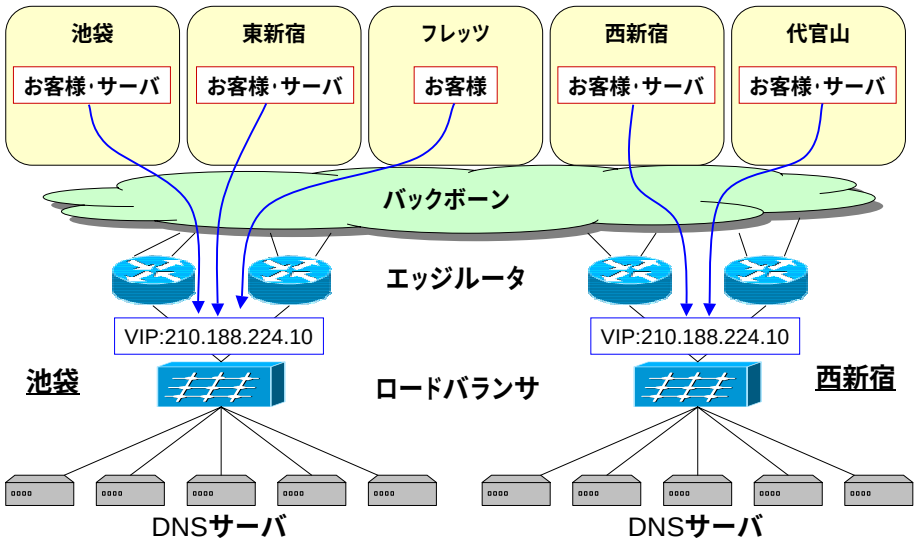
↓

- Anycastはあくまでも拠点冗長目的で利用
- 負荷分散はロードバランサを使う

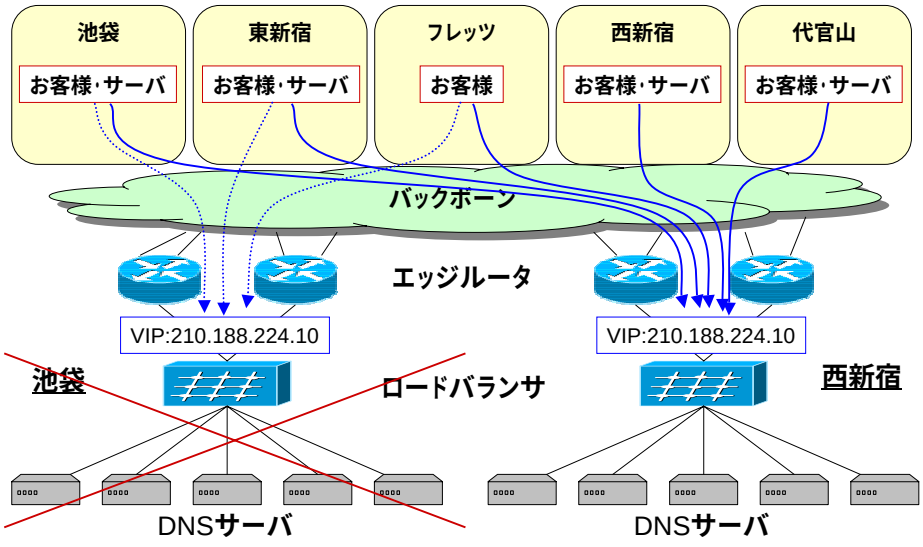
↓

- 東京都内2箇所にクラスタを配置（1+1冗長）
- 各クラスタは、全拠点のクエリを捌ける能力

2006年9月～ ロードバランサ+Anycast



障害発生時の迂回動作



各機器の仕様

- ロードバランサ：Foundry ServerIron GT-C
- ロードバランサの構成：シングルアーム
- 負荷分散アルゴリズム：Least Connection
  - サーバのslow downに敏感に反応 ☺
- キャッシュサーバ：FreeBSD6+BIND9
- キャッシュサーバの台数：各拠点5台ずつ

- 現在元気に稼働中！

まとめ

- コンテンツサーバ
  - 東京、大阪2拠点で冗長
  - 各拠点内のサーバもロードバランサで冗長
  - ロードバランサは負荷分散目的ではなく冗長目的
  - ゾーン数が多くなると、BINDでは扱えない
  - 現在のところANSが唯一の解？
- キャッシュサーバ
  - Anycastは拠点冗長に利用
  - → 負荷分散目的で使わない方がよさそう
  - クラスタ内の負荷分散はロードバランサを利用