



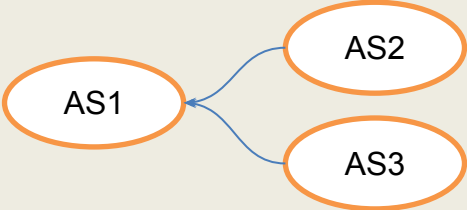
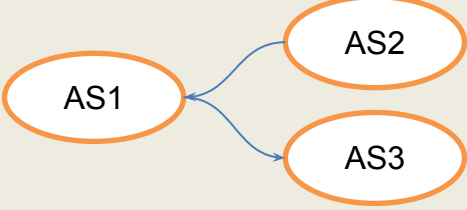
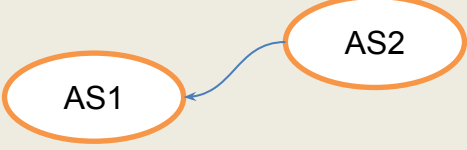
HighLoad++

**Некоторые аспекты влияния сходимости
протокола BGP на доступность
сетевых ресурсов**

Александр Азимов

aa@highloadlab.com

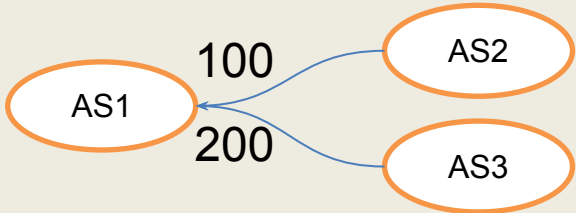
АВТОНОМНЫЕ СИСТЕМЫ

| | |
|-------------------|--|
| Многоинтерфейсные |  <pre>graph LR; AS2((AS2)) --> AS1((AS1)); AS3((AS3)) --> AS1</pre> |
| Транзитные |  <pre>graph LR; AS2((AS2)) --> AS1((AS1)); AS3((AS3)) --> AS1</pre> |
| Ограниченные |  <pre>graph LR; AS2((AS2)) --> AS1((AS1))</pre> |

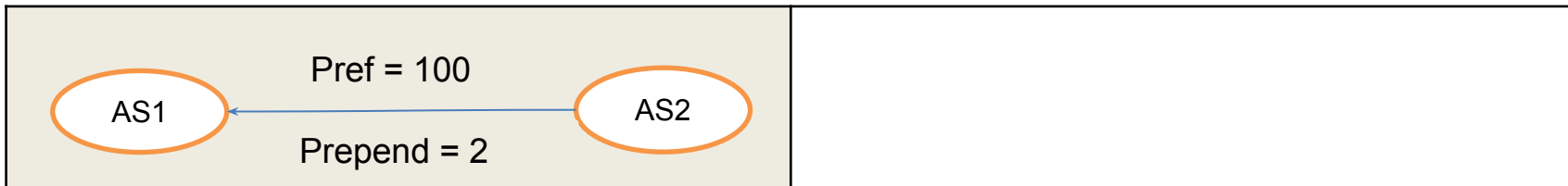
Протокол BGP

- Border Gateway Protocol (RFC 1771, 4271)
- Де-факто стандартный протокол внешней маршрутизации
- Дистанционно-векторный протокол
- Политика маршрутизации
- Набор атрибутов маршрутов
 - LOCAL_PREF
 - AS_PATH
 - ...

Политика маршрутизации

| | |
|-----------------|---|
| LOCAL_PREF |  <p>The diagram illustrates a routing policy where AS1 is the source. It has two outgoing paths: one to AS2 with a local preference of 100, and another to AS3 with a local preference of 200. Both AS2 and AS3 are circled in orange.</p> |
| AS_PATH prepend | |

Формальная модель



Представим модель сети автономных систем, как ориентированный граф $G(V, E)$:

- V – автономные системы
- $(as_1, as_2) \in E$ тогда и только тогда, когда АС as_1 будет анонсировать маршрут к префиксу I BGP АС as_2
- Вес дуги состоит из двух частей:
 - Prepend – политика AS_PATH АС as_1
 - Pref – политика LOCAL_PREF as_2

Транзитные АС

Пусть $ASPath$ – множество, состоящие из всех существующих значений атрибута маршрута BGP AS_PATH в текущий момент времени.

Тогда признаком транзитной автономной системы будет:

$$x \in ASTransit \leftrightarrow \exists path \in ASPath, \exists i < |path|: x = path_i$$

«Ядро» Интернета

Множество транзитных АС неоднородно:

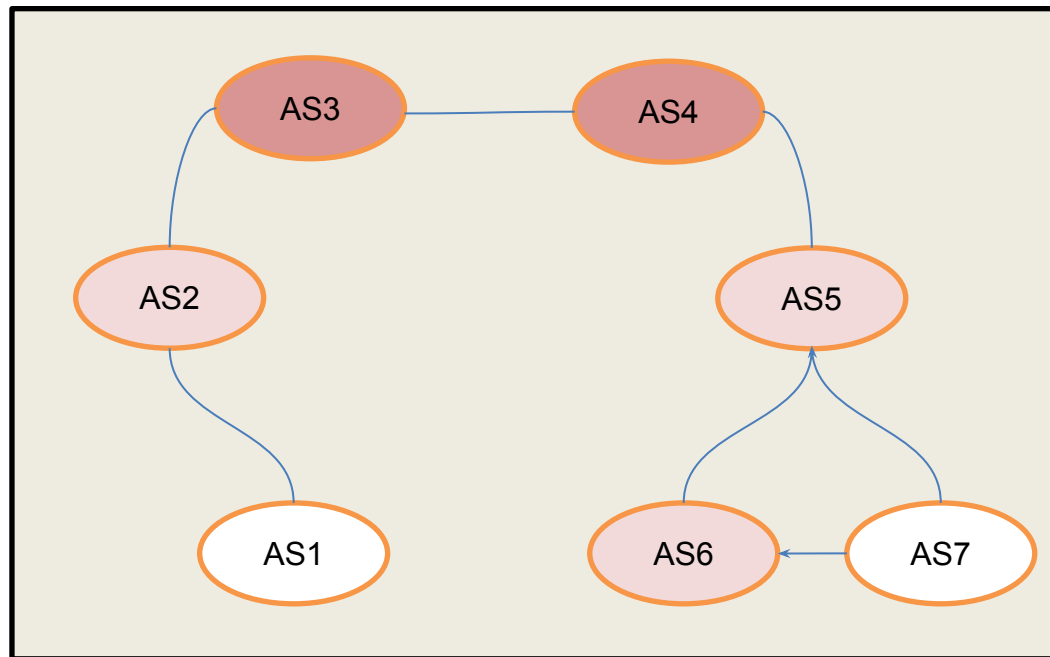
- Пропускают только трафик клиентов
- Имеют пиринговые отношения с

соседями

1. $Peers(0) = Transit$
2. $Peers(i + 1) = \{x \in Peers(i) \mid \exists path \in ASPath, \exists i < j < |path|: x = path_i \& path_j \in Peers(i)\}$
3. $\lim_{i \rightarrow \infty} Peers(i) = CORE$

Пример нахождения CORE

- $Transit = \{AS_2, AS_3, AS_4, AS_5, AS_6\}$
- $Peer(1) = \{AS_3, AS_4, AS_5\}$
- $Peer(2) = \{AS_3, AS_4\} = CORE$



Свойства CORE

Лемма

Предел $\lim_{i \rightarrow \infty} Peers(i)$ существует и отличен от пустого множества.

- $G'(E', V') \leq G(E, V)$:
- $\forall v \in V' \rightarrow v \in Core$,
- $\forall (v_1, v_2) \in E' \rightarrow \max(pref(v_i, v_2)) = pref(v_1, v_2)$

Теорема

Прикладное применение

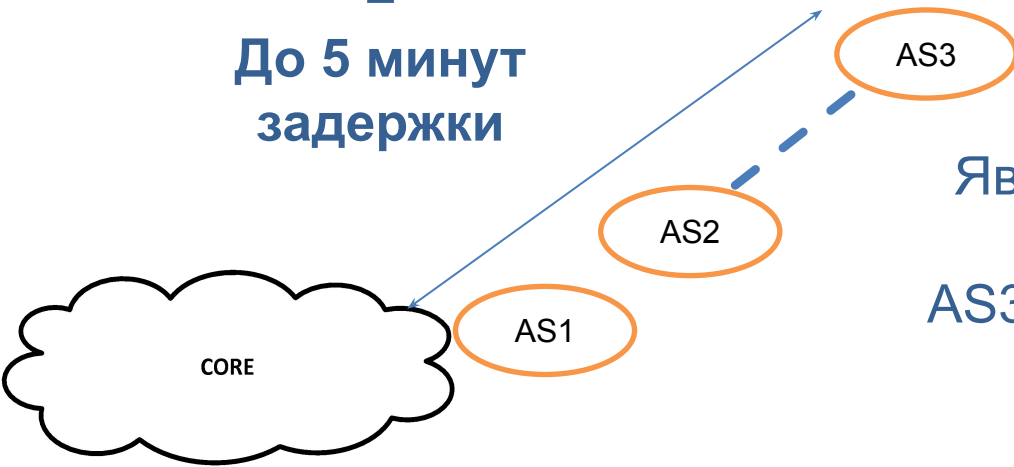
- Хостингам: выбор сервис провайдера
- Reverse Traceroute
- Обнаружение LOCAL_PREF циклов
- Определение времени сходимости
- Моделирование механизмов самого BGP

Время сходимости

10 АС до CORE

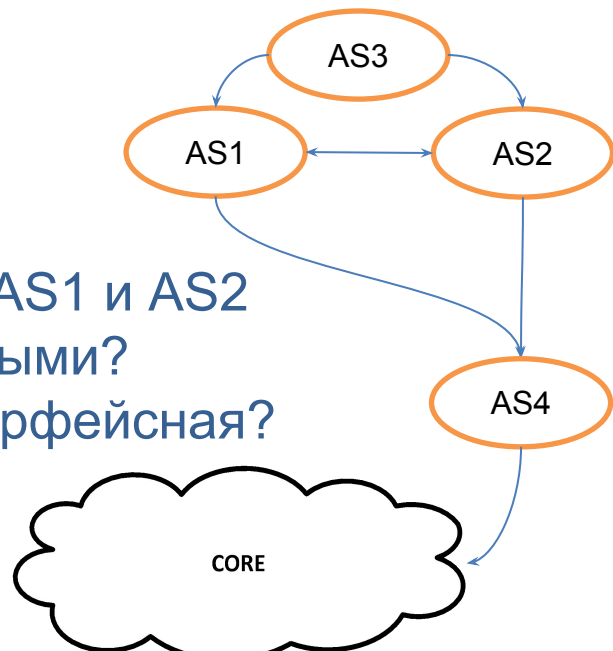
=

До 5 минут
задержки



Псевдо-транзитные АС

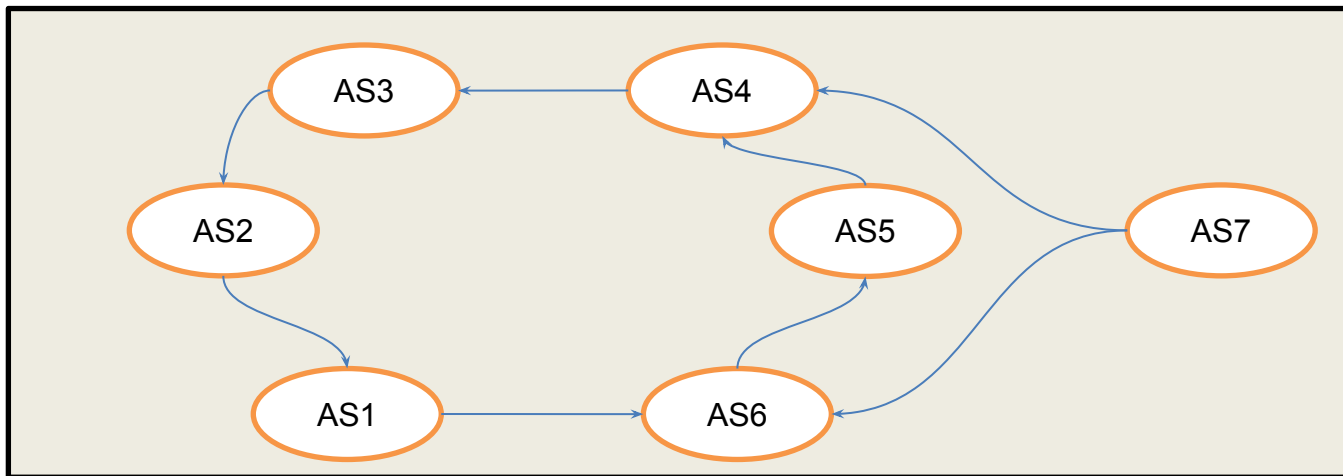
Являются ли AS1 и AS2
транзитными?
AS3 многоинтерфейсная?



Reverse Traceroute

- AS – единая политика маршрутизации
- Reverse Traceroute – знание, как к тебе идет трафик от других AS
- Если знать, как идет трафик, можно его балансировать – profit!

LOCAL_PREF ЦИКЛЫ



- Причина сетевой нестабильности для целевого префикса
- Создание постоянного «шума» из BGP сообщений
- Замедление времени сходимости по всей сети AS

Время сходимости

Рассматриваемые события:

- Объявление маршрута к префиксу l автономной системой X

~~• Удаление маршрута к префиксу l автономной системой~~

- $ET_{up} = \theta(d \times Et_{wait})$, где d – диаметр относительно вершины X в графе G
- $ET_{down} = \theta(D \times Et_{wait})$, где D – длина гамильтонова пути относительно вершины X в графе G

Система моделирования

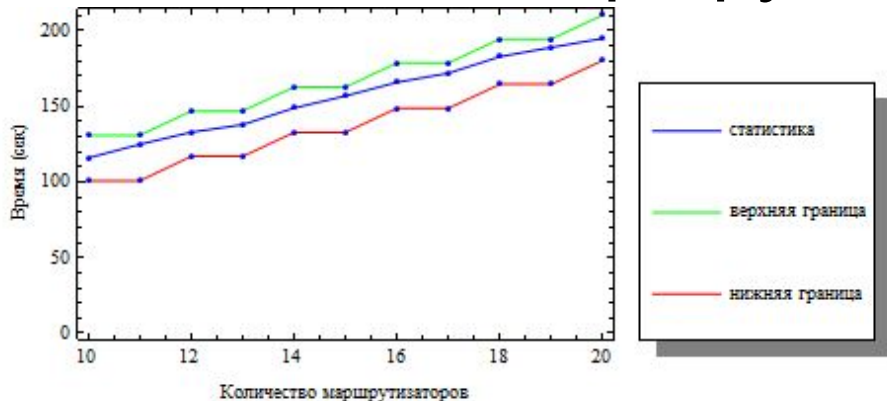
| Критерии сравнения | SSFNet | NS-2 | PRIME | CBGP | Разработанная система |
|--|------------------|------------------|------------------|-----------------|-----------------------|
| Реализация политик маршрутизации | - | - | + | + | + |
| Полнота стека BGP-решений | +/- | +/- | + | + | + |
| Модель передачи BGP сообщений | Пакетный уровень | Пакетный уровень | Пакетный уровень | В виде объектов | В виде объектов |
| Возможность моделирования сходимости BGP маршрутизации | + | + | + | - | + |
| Возможность распределенных вычислений | + | - | + | - | + |
| Масштабируемость | + | + | - | + | + |

Проверка распределений

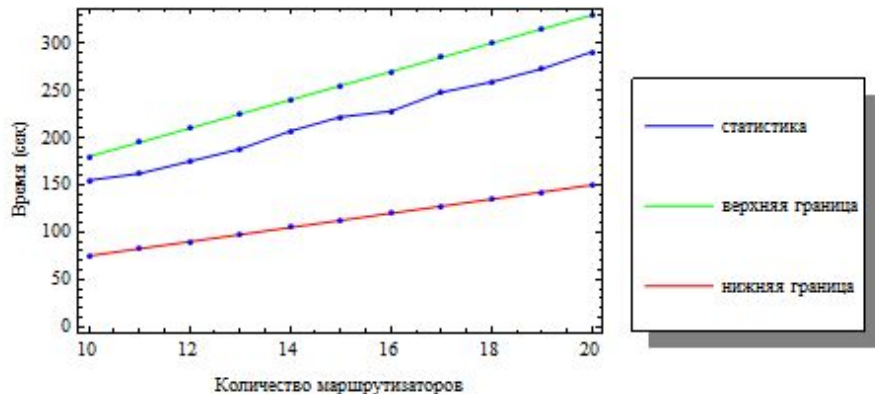
$$d^{\max}(X) \times E(t_{wait}) \leq ET_{up} \leq d^{\max}(X) \times E(t_{wait}) + s$$

$$D_{\frac{1}{2}}^{\max}(x) \times E(t_{wait}) \leq ET_{down} \leq D^{\max}(x) \times E(t_{wait}) + s$$

Объявление маршрута



Удаление

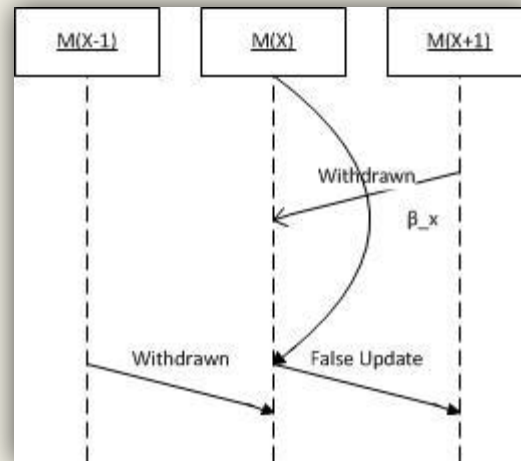
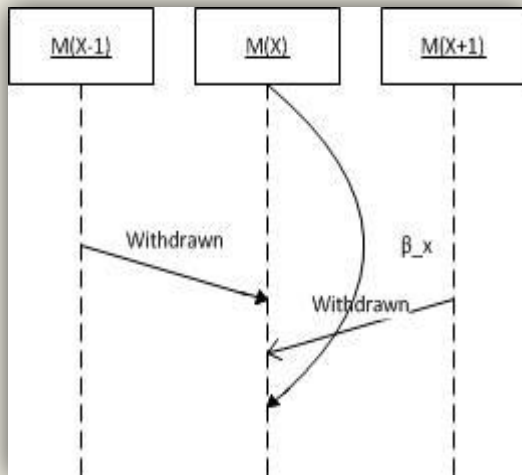
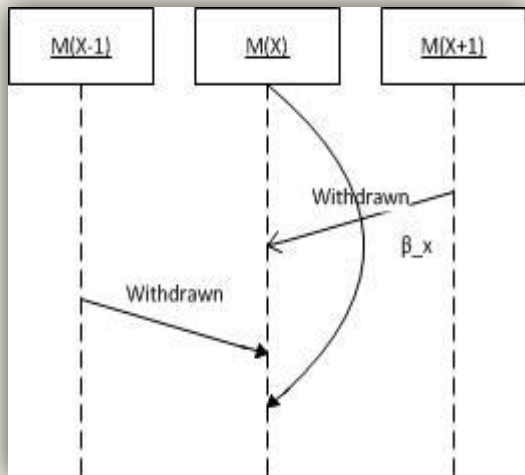


Flap Damping

Зачем?

Снижение нагрузки на сеть BGP маршрутизаторов от нестабильных маршрутов, не оказывая влияния на время сходимости в стабильных участках сети.

Flap Damping

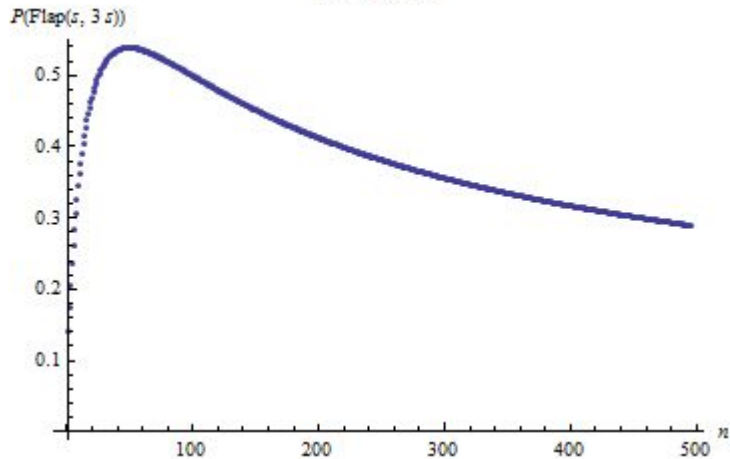


Flap Damping

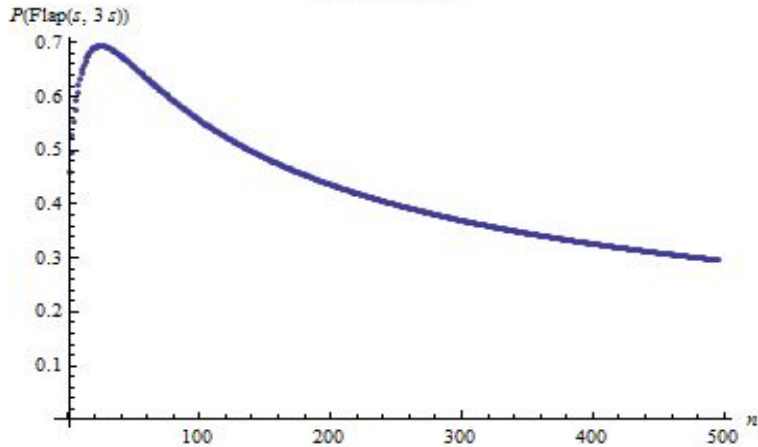
Вероятность появления мигающего маршрута в кольце маршрутизаторов длины n :

$$2 \times \left(\Phi \left(7 \times \sqrt{\frac{3}{n+1}} \right) - \Phi \left(2 \times \sqrt{\frac{3}{n+1}} \right) \right), \text{ при } n \text{ четном, } 2 \times \left(\Phi \left(6 \times \sqrt{\frac{3}{n+1}} \right) - \Phi \left(\sqrt{\frac{3}{n+1}} \right) \right), \text{ при } n \text{ нечетном}$$

n – четное



n – нечетное



Flap Damping DoS?

- Yes!
- Не блокируется ни одним существующим расширением BGP S-BGP

Данные для экспериментов

- List of router registries
- <http://www.irr.net/docs/list.html>
- BGP dumps
- <http://www.ripe.net/projects/ris/rawdata.html>

Автономных систем: 36200

Транзитных систем: 5876

CORE: 1757

Prepends:

Результаты

- Разработана модель BGP маршрутизации
- Сформулированы оценки времени сходимости протокола BGP
- Разработана система моделирования для проверки теоретических оценок
- Рассмотрены механизмы BGP LOCAL_PREF и Flap Damping и их влияние на доступность сетевых ресурсов
- Предложен метод для построения Reverse Traceroute

hl⁺⁺

HighLoad⁺⁺

Спасибо за внимание!

Вопросы?