

На краю пиринга (методы оптимизации AS-Path)

Дисклаймер

Написать данную статью побудило общение в кулуарах [КРОС-2015](#), в котором обсуждались механизмы "захвата" трафика на свою сеть для его последующей продажи клиентам. Одна из идей, а именно вырезание своей AS из транзитной BGP-информации, показалась не очень реалистичной, но все же требовала осмысления и проверки. Проверка подтвердила, что такой способ реально существует и практически уже используется.

Данная статья является техническим взглядом на методы, используемые операторами для оптимизации AS-Path с целью увеличения доли продаваемого ими трафика.

Вся информация в статье взята из открытых источников и не является инсайдом ни в какой ее части. При проведении анализа использована информация со следующих ресурсов:

- [radar.qrator.net](#)
- [ripe.net](#)
- [cisco.com](#) и [juniper.net](#)
- RFC 4271
- открытого листинга участников локальных и распределенных пиринговых систем
- cli собственного бордера
- открытых looking glass и route-server операторов

Pipe-IX

В борьбе за клиентов операторами сломано много копий, отгремели пиринговые войны 00-х, наступила эра "pipe-ix". Последние, построив инфраструктуру на "неожиданно" появившемся рынке оптического транспорта, продают недорогую распределенную, в отличие от классических IX, пиринговую связность на всей им доступной территории РФ. Распределенность pipe-ix делает их продукт более привлекательным, чем точечная пиринговая связность классических IX'ов. Благодаря этому они уже стали самостоятельным классом игроков на межоператорском рынке интернет провайдеров.

Пайпиксы, такие как W-IX, Cloud-IX и Data-IX за последние годы подключили основную массу тех клиентов, которых могли. При этом они набрали хорошую скорость и их потенциал, по крайней мере, внутренний, далеко не исчерпан. Однако емкость их рынка не так велика и при отсутствии свежих идей они неизбежно попадают на конкурентное поле настоящих магистралей, где продемонстрировать дальнейший рост будет не так просто, так как в отличие от классических магистралей их канальная емкость практически полностью является арендованной.

Свежая идея нашлась в использовании уже построенной сети для предоставления клиентам распределенных IX'ов не только пиринговой, но и полной связности, не уходя при этом из поля Distributed-IX. Упрощенно говоря, надо решить задачу: как приобрести и продать full-view, оставаясь при этом "pipe-ix".

Для этого Pipe-IX нужно добавить в пиринговую сеть апстримовую связность и остаться при этом похожим на IX, то есть прозрачным в смысле отсутствия лишнего (своего) as-hop'a между клиентами распределенной сети обмена. Сама идея несложная, но ее корректное технологическое решение делает из нее новый продукт – распределенный IX с полной связностью. И эта задача решена – транзитная сеть со своей автономной системой, полноценной маршрутизацией пакетов внутри и с взаимодействием с клиентами по BGP является прозрачной для остального мира.

Зачем это нужно тоже ясно - желание оператора сократить AS-PATH и стать более привлекательным в борьбе за клиента вполне понятно. Профиты, которые даст короткий AS-Path - это видимость более короткого маршрута к себе и своим клиентам тоже очевидны: чем короче AS-PATH, тем больше шансов при равном LocalPref получить трафик на себя и продать его клиенту.

Как "выпилить" свою AS из AS-Path

Итак, давайте попробуем разобраться, как это делается. Начать нужно с технического понимания, когда в BGP добавляется своя автономная система. Я думаю, что многие знают, что в BGP своя AS добавляется к исходящему анонсу, вернее должна добавляться, чтобы взаимодействующие операторы по принимаемым от сопредельных сетей AS-Path'ам могли по полным данным оценить выгодность и целесообразность тех или иных маршрутов для выбора лучших путей передачи пакетов

данных.

Информация о своей AS должна добавляться в исходящие анонсы. Для сомневающих ниже приведена пара выдержек из RFC 4271 (первая о транзите анонса и вторая об оригинации оного) для EBGP.

If the first path segment of the AS_PATH is of type AS_SEQUENCE, the local system prepends its own AS number as the last element of the sequence (put it in the leftmost position with respect to the position of octets in the protocol message).

When a BGP speaker originates a route then: the originating speaker includes its own AS number in a pathsegment, of type AS_SEQUENCE, in the AS_PATH attribute of all UPDATE messages sent to an external peer. In this case, the AS number of the originating speaker's autonomous system will be the only entry the path segment, and this path segment will be the only segment in the AS_PATH attribute.

Кстати тут есть забавный хинт. Если смотреть на Cisco или Juniper-анонсы, которые мы объявляем, то там мы свою AS не увидим, хотя она есть. Это может вызывать небольшое недоумение, если не знать данный факт.

```
SIS-CISCO7606#sh ip bgp neighbors 213.248.0.9 advertised-routes
```

```
BGP table version is 12952075, local router ID is 172.31.255.1
```

```
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
```

```
                r RIB-failure, S Stale, m multipath, b backup-path, x best-external, f RT-Filter
```

```
Origin codes: i - IGP, e - EGP, ? - incomplete
```

Network	Next Hop	Metric	LocPrf	Weight	Path
*>i172.31.0.0/21	172.31.255.2			100	0 i

Итак, с этим понятно. Но это еще не все. Правило приема анонса с первой AS отличной от AS пира, не регламентированы в RFC, то есть отданы на усмотрение вендорам. Cisco не примет его, если первой в AS-PATH будет не AS пира, а какая-то другая. Об этом нам может сказать cisco.com

By default, the software requires the first autonomous system (in the AS path) of a route received from an eBGP peer to be the same as the remote autonomous system configured.

Вот здесь уже без удаленного пира не обойтись. Для того чтобы обойти это правило на Cisco (XR) есть команда. Впрочем, на IOS она тоже есть с немного другим синтаксисом:

```
enforce-first-as disable
```

Для Juniper же наоборот дефолтным является поведение, при котором такая проверка отсутствует. Включить ее можно командой:

```
bgp enforce-first-as
```

Данная опция не является новым словом в BGP и уже давно успешно применяется в классических IX, для которых она и была создана.

Пиринг без границ

Получается, что технически "выпиливание" из AS Path информации о своей автономной системе возможно при согласии взаимодействующих сторон.

Со стороны клиентов распределенных IX'ов это не вызывает вопросов и проблем – IX анонсирует им всех своих участников "напрямую". Так и должен поступать IX, чтобы быть привлекательным.

Для того чтобы в распределенный IX были подмешаны апстримы с их полной связностью, апстрим-оператор должен стать соучастником процесса и снять проверку принимаемых им от ріре-IX анонсов на принадлежность префиксов оператору, их анонсирующих.

Отсутствие такой проверки означает, что анонс клиентов ріре-IX появится в мире без AS транзитного провайдера и условный Distributed-IX может остаться таковым, но при этом анонсировать своего клиента в мир от имени своего аплинка и получить существенный прирост трафика через себя.

Если распределенный IX сумеет убедить классического магистрала, продать ему недорогого Full-View для обслуживания "не своих" анонсов, то путь к новому транзитному мироустройству открыт.

Возникает резонный вопрос, что выигрывает классический магистрал от такой сделки с распределенным IX'ом, а что проигрывает.

Выигрыш прост и сразу виден – Ріrex новый клиент с новыми доходами, пусть и не самый выгодный с точки зрения места подключения, но все же.

Риски не так очевидны потому, что они отложены во времени. Я вижу два:

1. Возрастание вероятности потерять текущих собственных клиентов и не включить новых потенциальных;
2. Усиление конкурентного давления на цены в регионах.

Первый и главный риск - потерять прямой трафик своих собственных клиентов. Если клиент включен одновременно и в магистрала и в ріrex, то при равном AS-PATH, а в этом случае он будет именно таким, в игру вступают другие атрибуты. В этом случае даже не обязательно бить сеть на спецификации. Достаточно задать origin code – incomplete для менее предпочтительного маршрута, чтобы получить трафик через более дешевую пиринговую связность. Как защитится от этого? Естественно через LocalPref, сделав его для ріrex меньше, чем для всех остальных клиентов, но больше пирингового. В этом случае трафик будет отправлен клиенту сразу, если есть прямая сессия. Если прямой сессии нет, то трафик уйдет без проблем в сторону ріrex, которую последний и оплатит.

Однако защита своих коммерческих интересов с помощью LocalPref не столь крепка, ведь против LP есть спецификации (More specific routes) и с ними магистрал уже ничего сделать не сможет. Конечный оператор получит искомый трафик от ріrex, если захочет. И по минимальной цене, оставив магистралу роль дорогого бэкапа, что при наличии тарификации берст сделать весьма не сложно ☺.

Второй риск – ценовая конкуренция за клиента. Ведь конечный оператор может использовать возможность получить трафик от ріre-ix, как инструмент давления на своего аплинка. Впрочем, этот риск я пока воздержусь комментировать. Станет ли повсеместное внедрение новой услуги фактором изменения цены на региональном операторском рынке? Этот вопрос я оставляю для обсуждения с читателями.

Итак, техническая возможность нового BGP мироустройства есть, драйверы его внедрения понятны, кандидаты намечены. Попробуем посмотреть внутрь сетей.

О чем говорит BGP

Для анализа я взял трех кандидатов (W-IX, Data-IX, Cloud-IX) и попробовал проанализировать анонсы их участников. Методика будет понятна из первого примера.

Берем список участников одного из ріrex'ов - W-IX. Берем первого участника, кто попадется, и изучаем его апстримы:

ООО "Квик Ком"	57662	qik
--------------------------------	-----------------------	---------------------

Через куратор проверим, кто анонсит его в мир:

CURRENT (5)			
NEW (5)			
LEFT (1)			
RETURNING (3)			
AS Number	Name	Description	History
AS9002	RETN-AS	RETN Limited	
AS20485	TRANSTELECOM	TransTelecom	
AS1299	TELIANET	TeliaSonera International Carrier	
AS20764	RASCOM-AS	CJSC Rascom, St.Petersburg, Russia	
AS42632	MNOGOBYTE-AS	MnogoByte LLC	

Берем AS I-home – A25478 и проверяем там же его аплинков:

CURRENT (4)			
NEW (1)			
LEFT (0)			
RETURNING (0)			
AS Number	Name	Description	History
AS20485	TRANSTELECOM	TransTelecom	
AS8359	MTS	MTS OJSC	
AS1299	TELIANET	TeliaSonera International Carrier	
AS20764	RASCOM-AS	CJSC Rascom, St.Petersburg, Russia	

Видим пересечение по 3-м из них: ТТК, Раском и Телия. Теперь на lg.rascom.ru проверяем префикс данного клиента:

```
show bgp ipv4 unicast 37.26.232.1
```

Thu Jun 4 12:58:14.923 MSK

BGP routing table entry for 37.26.232.0/21

Versions:

Process bRIB/RIB SendTblVer

Speaker 568048833 568048833

Last Modified: Nov 8 13:07:25.422 for 29w4d

Paths: (10 available, best #2)

Path #2: Received by speaker 0

57662

85.112.122.43 from 85.112.122.1 (85.112.122.1)

Origin IGP, metric 0, localpref 99, valid, external, best, group-best, multipath, import-candidate

Received Path ID 0, Local Path ID 1, version 568048833

Community: 20764:1440 20764:1460 20764:1470 20764:1500 20764:1530 20764:1550 20764:3002

20764:3011 20764:3021 20764:3136 25478:1000

Origin-AS validity: not-found

Что видим? Клиентская AS анонсируется с некой пиринговой сети, Next-hop там сразу клиентский стоит. Local Pref – 99 (Customer 100 – 1). Наличие Community с AS I-Home. Чтобы окончательно убедиться, что Раском отдает эту сеть своему аплинку Level3 – смотрим на его LG:

BGP Routing table entry for 37.26.232.0/21

Paths: (2 available, best #1)

20764 57662

AS-path translation: 37.26.232.0/21

None (metric 0)

Community: Lclprf_100 Backbone_2 Level3:10725 Frankfurt Germany Europe Level3_Customer
20764:1440 20764:1460 20764:1470 20764:1500 20764:1530 20764:1550 20764:3002 20764:3011
20764:3021 20764:3136 25478:1000 Cluster : No

Origin: IGP, metric 0, localpref 100, Used Valid Best IGP Group-Best

Originator: None

Думаю, можно не комментировать. Level3 принял клиентский анонс и сделал его бэстом с достаточно коротким AS-Path. Ну, разве, что community по дороге стоило зачистить. ;)

Давайте посмотрим на ТТК. Вот с их LG.

Thu Jun 4 13:23:36.556 UTC

BGP routing table entry for 37.26.232.0/21, Route Distinguisher: 20485:1

Versions:

Process bRIB/RIB SendTblVer

Speaker 950080862 950080862

Local Label: 16028

Last Modified: May 21 13:22:39.804 for 2w0d

Paths: (6 available, best #1)

Advertised to update-groups (with more than one peer):

0.2 0.4

Path #1: Received by speaker 0

Advertised to update-groups (with more than one peer):

0.2 0.4

[57662](#)

85.112.122.43 from 85.112.122.1 (85.112.122.1)

Origin IGP, metric 1000, localpref 99, valid, external, best, group-best, import-candidate

Received Path ID 0, Local Path ID 1, version 950080862

Community: 20485:11799

Extended community: RT:20485:1

Картина та же: с того же RS анонс, да и NEXT-HOP совпадает. Опять LocalPref 100-1. Проверяем этот

анонс на пире ТТК Мегафоне. Он там есть и он там бэстом.

inet.0: 545696 destinations, 2219897 routes (543847 active, 4 holddown, 4234 hidden)

37.26.232.0/21 (6 entries, 1 announced)

*BGP Preference: 170/-281

Next hop type: Indirect

Address: 0x4b998b58

Next-hop reference count: 43860

Source: 10.222.253.253

Next hop type: Router, Next hop index: 2097311

Next hop: 83.169.204.25 via ae11.0, selected

Label operation: Push 4097

Label TTL action: no-prop-ttl

Session Id: 0x4cd

Next hop: 83.169.204.33 via ae22.0

Label operation: Push 4321

Label TTL action: no-prop-ttl

Session Id: 0x4ce

Protocol next hop: 10.222.253.177

Indirect next hop: 26c34938 2098630 INH Session ID: 0x242

State: <Active Int Ext>

Local AS: [31133](#) Peer AS: [31133](#)

Age: 2w0d 0:03:45 Metric2: 200

Validation State: unverified

Task: BGP_31133.10.222.253.253+???

Announcement bits (5): 0-KRT 2-RT 8-BGP_RT_Background 9-Resolve tree 4 10-Resolve tree 6

AS path: [20485](#) [57662](#) I (Originator)

Cluster list: 10.222.253.253

Originator ID: 10.222.253.177

AS path: Recorded

Communities: 20485:11799 31133:200 31133:300 31133:46090 target:20485:1

Accepted Multipath

Localpref: 280

Router ID: 10.222.253.253

Ну, и Телия в Москве.

inet.0: 598479 destinations, 2406022 routes (598463 active, 172 holddown, 90 hidden)

+ = Active Route, - = Last Active, * = Both

[37.26.232.0/21](#) *[BGP/170] 3w4d 13:12:43, MED 0, localpref 200, from [85.112.122.1](#)

AS path: 57662 I, validation-state: unverified

> to 85.112.122.43 via ae5.888

[BGP/170] 3w4d 13:12:34, MED 0, localpref 200, from [85.112.122.2](#)

AS path: 57662 I, validation-state: unverified

> to 85.112.122.43 via ae5.888

О чем это свидетельствует? О том, что новое BGP-мироустройство уже действует. Telia, ТТК и Раском – формально апстримы I-HOME, но фактически их участие прозрачно с точки зрения BGP.

Transit-IX

Можно сделать вывод, что схема не такая уж тайная, коль 3 аплинка согласились сделать нетрадиционные настройки для своего клиента. И то, что реальный трафик с сети наших тирванов идет через W-IX, это уже серьезный бизнес-драйвер для привлечения/удержания клиентов и конкурировать с таким провайдером классическому агрегатору будет совсем непросто. Более того, в том, что данные изыскания являются секретом Полишинеля, говорит тот факт, что на сайте i-home.ru данная услуга анонсирована как вполне полноценный дополнительный сервис к w-ix. Вот выдержка оттуда:

Услуга предоставляет собой аналог пиринговой площадки (L2 транспорт), участники которой находятся в общем VLAN. Обмен маршрутной информацией осуществляется через RS, что позволяет не устанавливать индивидуальные BGP сессии между отдельными участниками. На одной площадке присутствуют как операторы-клиенты, так и провайдеры верхнего уровня, предоставляющие полную таблицу маршрутизации (full view).

К сожалению, ссылка на подробное описание услуги недоступна, но работоспособность данной схемы взаимодействия не вызывает никаких сомнений, и ТТК с Телией можно смело записать в участники Inet2. Раском же заявлен как участник W-IX, что подразумевает более сложные отношения с пиринговой сетью.

А что другие? А ничего.:) Я не нашел явных доказательств наличия такой схемы у Cloud-IX и Data-IX для анонсирования своим аплинкам, хотя слышал, что они также могут ее реализовать, может мало примеров посмотрел. Что я увидел, так это то, что Cloud-IX анонсирует свою пиринговую сеть как клиентский префикс в сторону Cogent, что еще не дает возможность включать аплинка в свою пиринговую сеть, но теоритически это возможно.

Если же конкуренты действительно не играют в эти игры, (если вы знаете, кто играет, напишите на форуме), то остается поздравить W-IX с хорошим конкурентным преимуществом. Остальным же можно посоветовать спросить у своих аплинков "А что! Так можно было?"