Selectel

IPv4 prefix “hijack”

Кирилл Малеванов, Selectel

ENOG 14, 2017

ENOG RIPE NCC Regional Meeting
Eurasia Network Operators Group

Схема сети

- 4 пограничных маршрутизатора – 2 в Москве, 2 в СПб
- AS49505 анонсирует около 300 IPv4 префиксов
- Превалирование исходящего трафика над входящим (ЦОД)
- Аплинки - ТТК, Rascom, RETN
- Пиринги - MSKIX, DATA-IX, Cloud-IX, DE-CIX, LINX etc

Схема сети

- На IX используются роут-серверы, прямых сессий минимум
- На некоторых IX анонсировались more-specific маршруты для увеличения входящего трафика

Anycast

- Anycast используется для DNS

  ns1.selectel.org, ns2.selectel.org –

  31.131.254.0/24, 31.131.255.0/24

- DNS-серверы размещены во внешних датацентрах

- Для коннективити с миром используется AS49505

IX

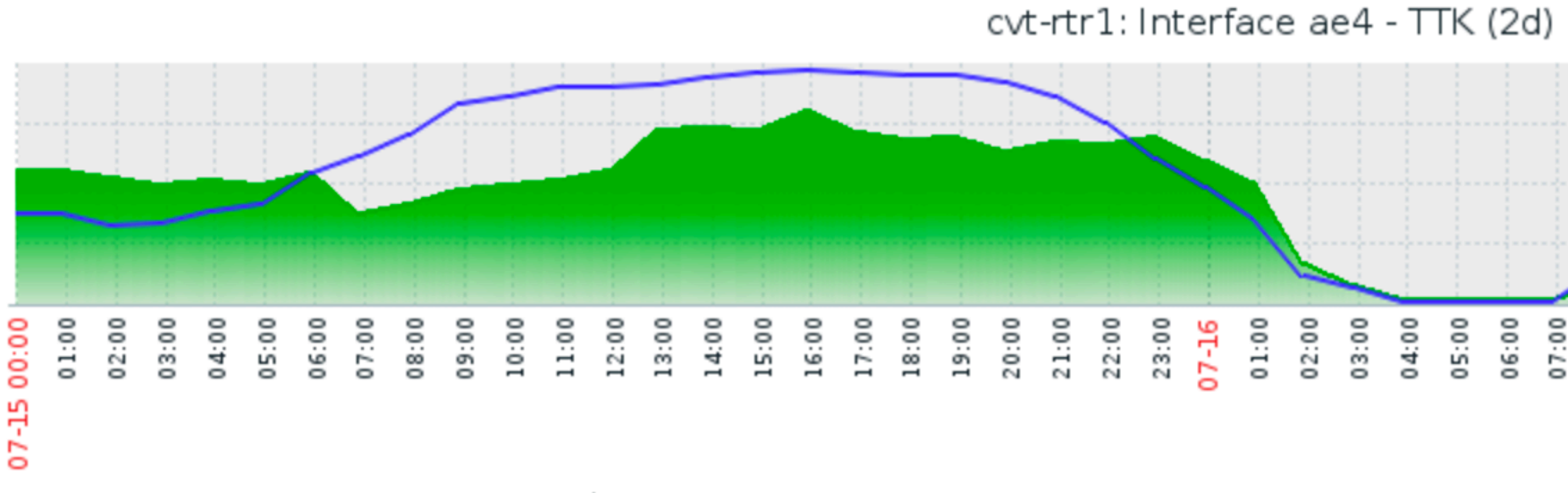
- IX – это большой L2 сегмент, где любой участник видит любого другого по L2
- Информация о маршрутах, как правило, имеется на RS
- Роут-серверов обычно два.
- Blackhole-коммунити на IX – специальная коммунити для IX, которую участники могут анонсировать на RS
- Маршрут с blackhole-community или идет в Null0, или есть специальный маршрутизатор на IX, который дропает весь трафик

Первая кровь

- Ночь с пятницы на субботу, 0:30
- Часть клиентов стала жаловаться на отсутствие связности. С чем, как, почему – непонятно, так как запрос маршрутов и т.п. – занимает время.

Первая кровь

- Из признаков – техподдержка жалуется, что перестал работать Slack
- Падение трафика на ТТК
- На других аплинках трафик не вырос



Первая кровь - реакция

«Наверное, что-то сломалось в ТТК»

deactivate protocols bgp group Uplinks neighbor <ТТК>

- Slack заработал
- Клиентские ресурсы, на отсутствие связности с которыми они жаловались, стали доступны

Основная часть

- В районе 02:00 опять перестал работать Slack
- Исследование проблем показало, что частично нет связности именно с зарубежными ресурсами
- Массовые звонки клиентов

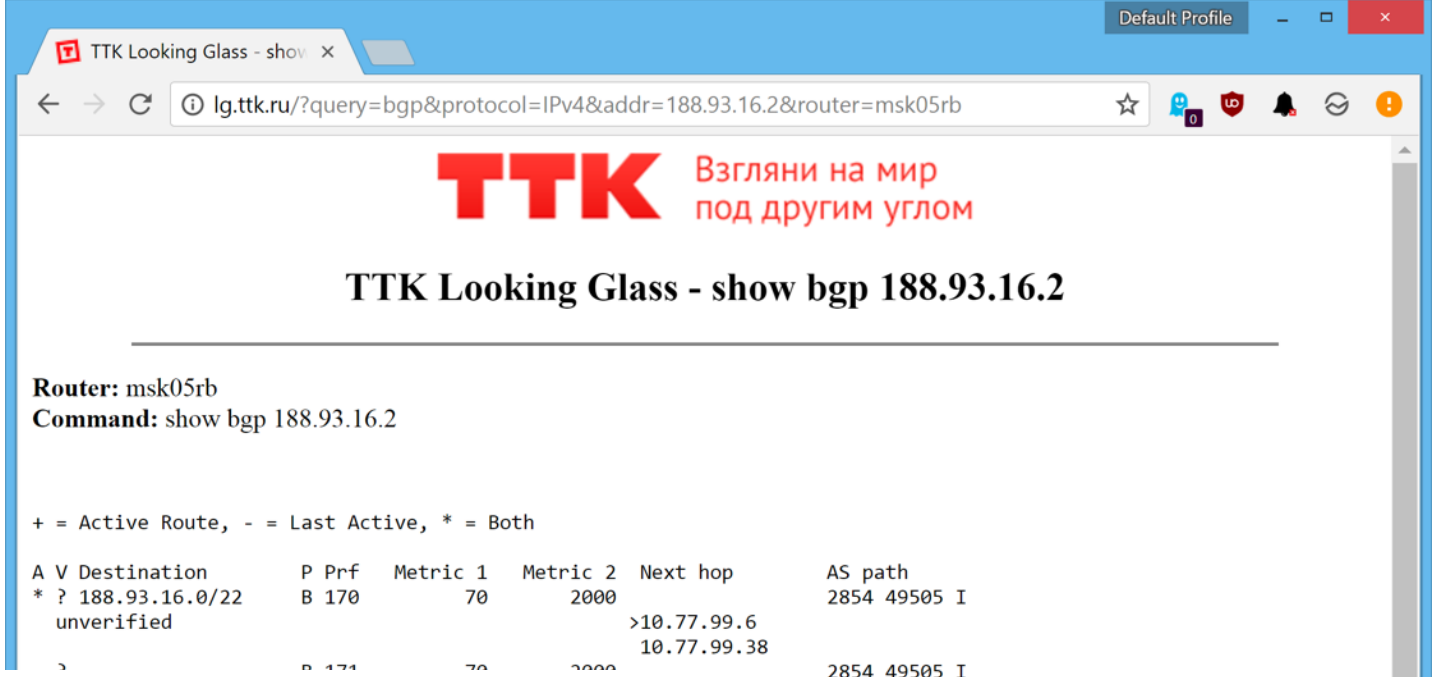
Диагностика

- Основной инструмент – это traceroute.
  Именно с его помощью обнаружили,
  что проблема в районе DE-CIX.

- Асимметрия трафика – если трафик идет от нас через DE-CIX, все ОК.
  Если трафик идет к нам через DE-CIX, то все плохо

- Смотрим Looking glass по всему маршруту трафика.

ТТК. Почему так?

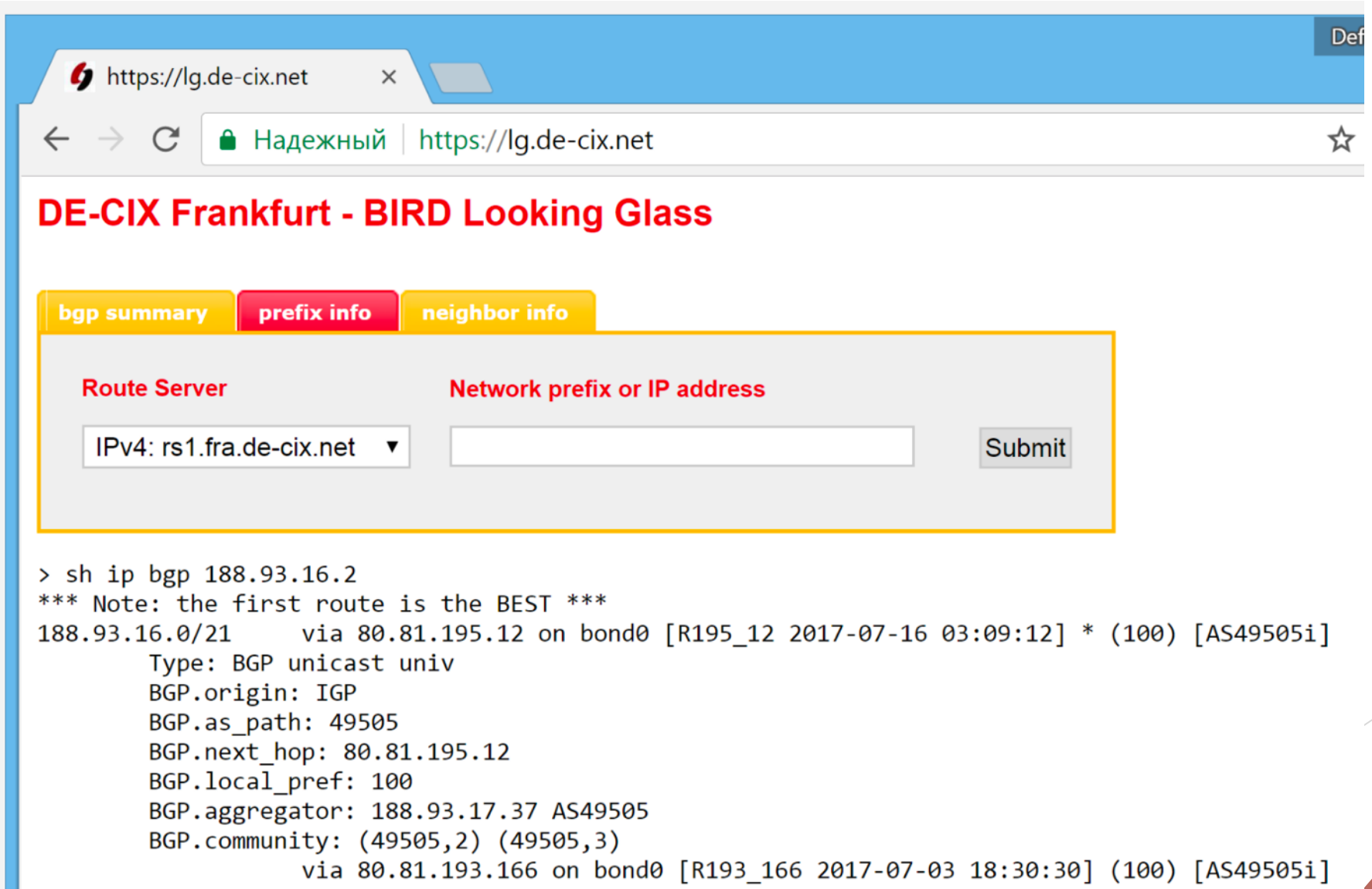


Router: msk05rb
Command: show bgp 188.93.16.2

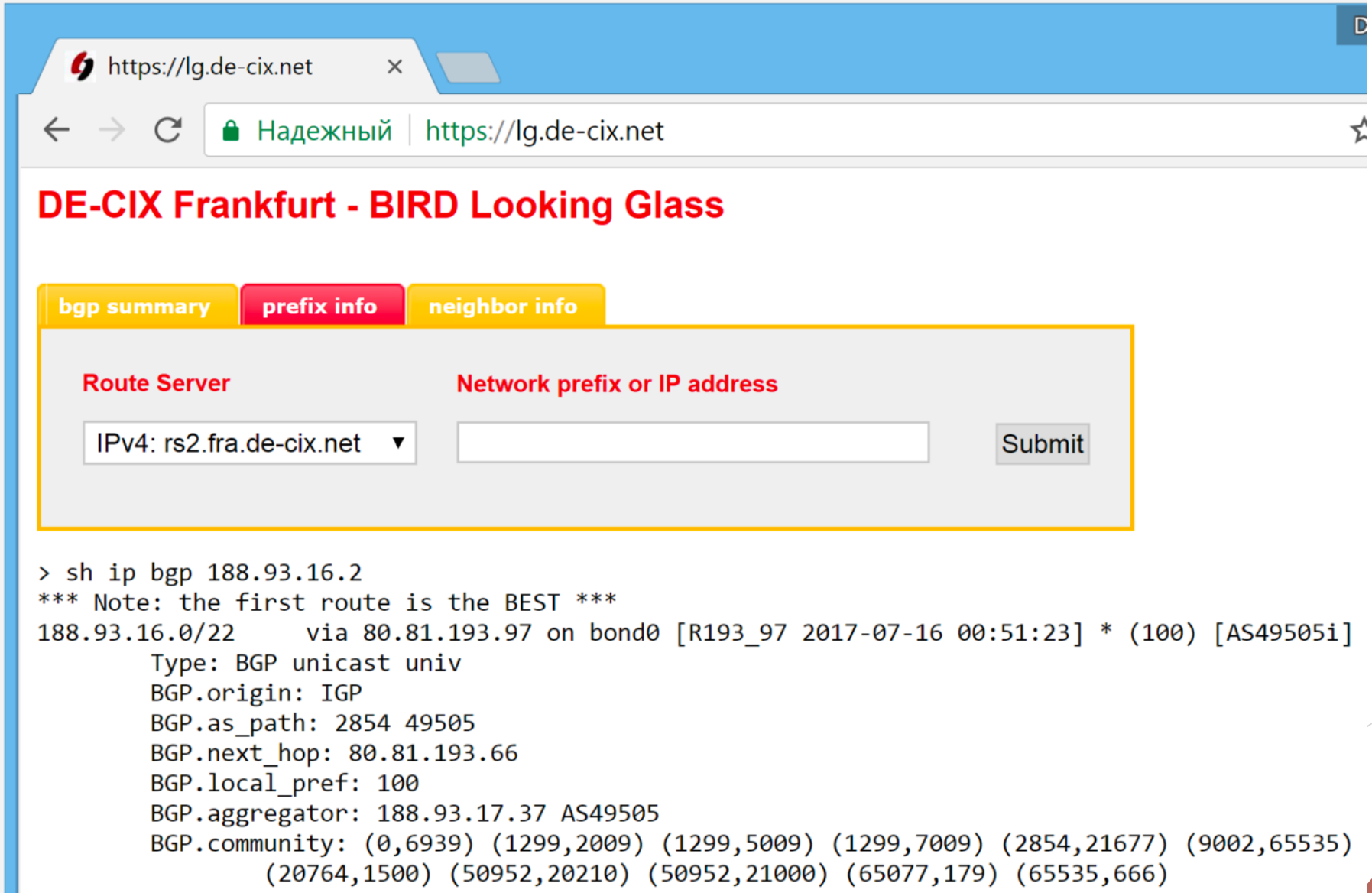
```
+ = Active Route, - = Last Active, * = Both

A V Destination        P Prf   Metric 1   Metric 2  Next hop        AS path
* ? 188.93.16.0/22     B 170         70       2000                  2854 49505 I
  unverified                                          >10.77.99.6
                                                       10.77.99.38
```

DE-CIX. RS1


https://lg.de-cix.net
Надежный
https://lg.de-cix.net
DE-CIX Frankfurt - BIRD Looking Glass
bgp summary
prefix info
neighbor info
Route Server
Network prefix or IP address
IPv4: rs1.fra.de-cix.net
Submit
> sh ip bgp 188.93.16.2
*** Note: the first route is the BEST ***
188.93.16.0/21 via 80.81.195.12 on bond0 [R195_12 2017-07-16 03:09:12] * (100) [AS49505i]
Type: BGP unicast univ
BGP.origin: IGP
BGP.as_path: 49505
BGP.next_hop: 80.81.195.12
BGP.local_pref: 100
BGP.aggregator: 188.93.17.37 AS49505
BGP.community: (49505,2) (49505,3)
via 80.81.193.166 on bond0 [R193_166 2017-07-03 18:30:30] (100) [AS49505i]

DE-CIX. RS2


https://lg.de-cix.net
Надежный
https://lg.de-cix.net
DE-CIX Frankfurt - BIRD Looking Glass
bgp summary
prefix info
neighbor info
Route Server
Network prefix or IP address
IPv4: rs2.fra.de-cix.net
Submit
> sh ip bgp 188.93.16.2
*** Note: the first route is the BEST ***
188.93.16.0/22 via 80.81.193.97 on bond0 [R193_97 2017-07-16 00:51:23] * (100) [AS49505i]
Type: BGP unicast univ
BGP.origin: IGP
BGP.as_path: 2854 49505
BGP.next_hop: 80.81.193.66
BGP.local_pref: 100
BGP.aggregator: 188.93.17.37 AS49505
BGP.community: (0,6939) (1299,2009) (1299,5009) (1299,7009) (2854,21677) (9002,65535)
(20764,1500) (50952,20210) (50952,21000) (65077,179) (65535,666)

DE-CIX

- RS1

```
> sh ip bgp 188.93.16.2
*** Note: the first route is the BEST ***
188.93.16.0/21     via 80.81.195.12 on bond0
        Type: BGP unicast univ
        BGP.origin: IGP
        BGP.as_path: 49505
        BGP.next_hop: 80.81.195.12
        BGP.local_pref: 100
```

- RS2

```
> sh ip bgp 188.93.16.2
*** Note: the first route is the BEST ***
188.93.16.0/22     via 80.81.193.97 on bond0
        Type: BGP unicast univ
        BGP.origin: IGP
        BGP.as_path: 2854 49505
        BGP.next_hop: 80.81.193.66
```

Кто виноват?

```
role:           EQUANT RUSSIA Network Operation Center
address:        LLC Equant/Russia
address:        1-st Krasnogvardeyskiy proezd, build. 15
address:        Moscow, 123100, Russia
phone:          +7 495 705 9229
phone:          +7 495 929 9500
fax-no:         +7 495 929 9420
e-mail:         noc@rosprint.net
admin-c:        YVM-RIPE
tech-c:         YVM-RIPE
tech-c:         DRP23-RIPE
tech-c:         YA589-RIPE
tech-c:         OA2285-RIPE
tech-c:         VP12948-RIPE
nic-hdl:        ERNO3-RIPE
tech-c:         PD7507-RIPE
remarks:        trouble: ==============================
remarks:        trouble: General questions: noc@rosprint.net
remarks:        trouble: Spam and abuse: abuse@rosprint.net
remarks:        trouble: Routing: noc@rosprint.net
```

Решение проблемы

- Убрать more-specific!
- Частично связность восстановилась.

- На DE-CIX RS2 перестали быть more-specific маршруты, которые 100% становились best для участников IX

- Но – обычные IPv4 aggregates были на RS2 с блекхол-коммунити

Решение проблемы

Попытка связи с NOC Rosprint (Orange, Equant)

- На городской телефон почти не дозвониться
- Один из сотрудников NOC найден на мобильном, но он в отпуске
- Реакция сотрудников Orange – пишите на noc@, в понедельник разберемся.
- Повторный звонок – «вам сказали писать email, до свидания»

Решение проблемы

- Письмо и звонок в DE-CIX

Ahmer Baig via RT <support@de-cix.net>
to sizovkirill, me, noc, support

Dear Kirill,

Thanks for the calling. Please note that we are Layer 2 platform we do not manage customer prefix routing.

Решение проблемы

- Письма в MSK-IX и DATA-IX с просьбой отключить Rosprint.
  Диагностика route-leaking
- 11:26 – проблема решена

aleksandr.komarov@orange.com

to me

Коллеги, прошу проверить работоспособность сервиса.

С уважением,

orange Business Services

Александр Комаров
Инженер, Сектор экспертного управления решениями по передаче данных

тел. +7 (495) 777 0800 доб. 4095

Уроки на будущее

- Наличие more-specific упростило задачу!
- Мониторинг
  должен быть мониторинг связности своей сети
  На IX должно быть какое-то API для RS
- Связность надо проверять из разных точек,
  прямым и обратным трафиком.

Уроки на будущее

У ТТК есть нюанс в маршрутизации внутри сети

- 188.93.16.0/21 анонсируется как глобальный префикс
- 188.93.16.0/22 и 188.93.20.0/22 анонсируются как more-specific на IX
- More-specific анонсируются в ТТК с no-export
- Но московский маршрутизатор ТТК «не видит» этих маршрутов из СПб
- То-есть, no-export в ТТК работает внутри маршрутизатора, а не внутри AS ?

Уроки на будущее

- На IX надо всегда смотреть на оба RS
- Почему IX принимает маршруты с blackhole, отличные от /32?

Уроки на будущее

- Anycast имеет смысл отделять по AS от основной сети

As you are the member of AS-URAL that is a member in the AS set of AS-ROSPRINT (AS2854),
http://irrexplorer.nlnog.net/search/AS-URAL
so routes working accordingly.

For me, it looks like a human mistake by AS2854. The most faster resolution for such problem will be to contact them directly and ask them the reason of blackholes routes.

But now the route 5.178.85.90 looks fine and I do not see for Blackholing (at the moment):

Вопросы