

Маршрутен протокол BGP

Понятие за автономна система (AS)

Автономната система (**autonomous system - AS**) е сбор от свързани помежду си IP мрежи (префикси), които са под административното и техническо управление на мрежов оператор или др. организация (напр. СУ).

В рамките на AS е възможно да работят различни вътрешни протоколи за маршрутизация (IGP).

AS поддържат строго дефинирана политика за маршрутизация в Internet (вж. **RFC 1930**).

Понятие за AS

AS трябва да има глобален уникален номер
(ASN - Autonomous System Number)

Този номер се използва при обмен на маршрутизираща информация със съседни AS-и и като идентификатор на самата AS.

Кога ни трябва AS (ASN)

AS са задължителни при обмен на външни маршрути с други ASs с помощта на протоколи за външна маршрутизация.

В момента такъв е BGP Border Gateway Protocol.

Но това не е достатъчно условие, за да искаме да имаме AS.

Кога да, кога не - AS

AS ни е необходим единствено и само тогава, когато имаме политика за маршрутизация (**routing policy**), различна от тази на други партньори - съседни (**peers**).

routing policy – как останалата част от Internet взима решения за маршрутизация на базата на информация от нашата AS.

Кога да, кога не - AS

Single-homed site, единствен или множество префикси, свързан към един единствен доставчик (т.е. една AS).

Не ни е необходима AS. Префиксът/те се поставя в AS на провайдера.

Multi-homed site. Необходима е AS.

multi-homed означава префикс или група от префикси, която се свързва към **повече от един доставчик** (т.е. повече от една AS, всяка със своя политика).

AS 5421 (CY) - Multi-homed



ASNs

До 2007 г. ASNs бяха единствено 16-битови, максимален брой 65536.

IANA е резервирала следния блок от AS номера за частно ползване (да не се анонсират в глобалния Internet):

64512 - 65535

ASNs 0, 54272–64511 и 65535 са резервирани от IANA. С ASN 0 обозначават немаршрутизирани мрежи.

Раздаване на ASNs

Всички останали номера (1–54271) се раздават от IANA.

IANA ги разпределя на съответните RIR, които от своя страна присвояват AS номера на организации в тяхната област, които отговарят на дадените по-горе критерии.

(RIR за територията на България е RIPE, <http://ripe.net>).

Разпределението на ASN ресурса от страна на IANA до момента, можете да видите на :

<http://iana.org/assignments/as-numbers/as-numbers.xml>

32-битови AS номера

Поради големия брой IP оператори адресното пространство на 16-битовите автономни системи застрашително се запълва.

Към м. септември 2008 г. са свободни само 49152–54271

Затова се въведоха 32-битови AS номера, RFC 4893, които вече се раздават от IANA.

Border Gateway Protocol (BGP)

Border Gateway Protocol (BGP) е основният протокол за маршрутизация в Internet.

Поддържа таблица от IP мрежи (префикси), които определят достижимостта на мрежите между автономните системи.

BGP е протокол с вектор на пътищата, **path vector protocol**.

Вектор на пътищата

BGP не използва метриката на вътрешните протоколи, а взема решения за определяне на маршрути на база на пътя между ASs, мрежови политики и/или правила.

От 1994 г. насам се използва версия 4 на протокола, която поддържа CIDR и обединяване (агрегация) на маршрути, с което се намалява размера на маршрутните таблици.

От януари 2006 г. **версия 4** е стандартизирана в **RFC 4271**.

Принцип на действие

BGP съседите (**neighbors** или **peers**) - маршрутизатори, се формират, след като ръчно са зададени.

Между тях се установява **TCP** сесия по **порт 179**.

Всеки BGP възел периодично изпраща 19-байтови “keep-alive” съобщения за поддържане на връзката.

BGP единствен от маршрутизиращите протоколи използва TCP за транспорт, което го прави приложен протокол до известна степен.

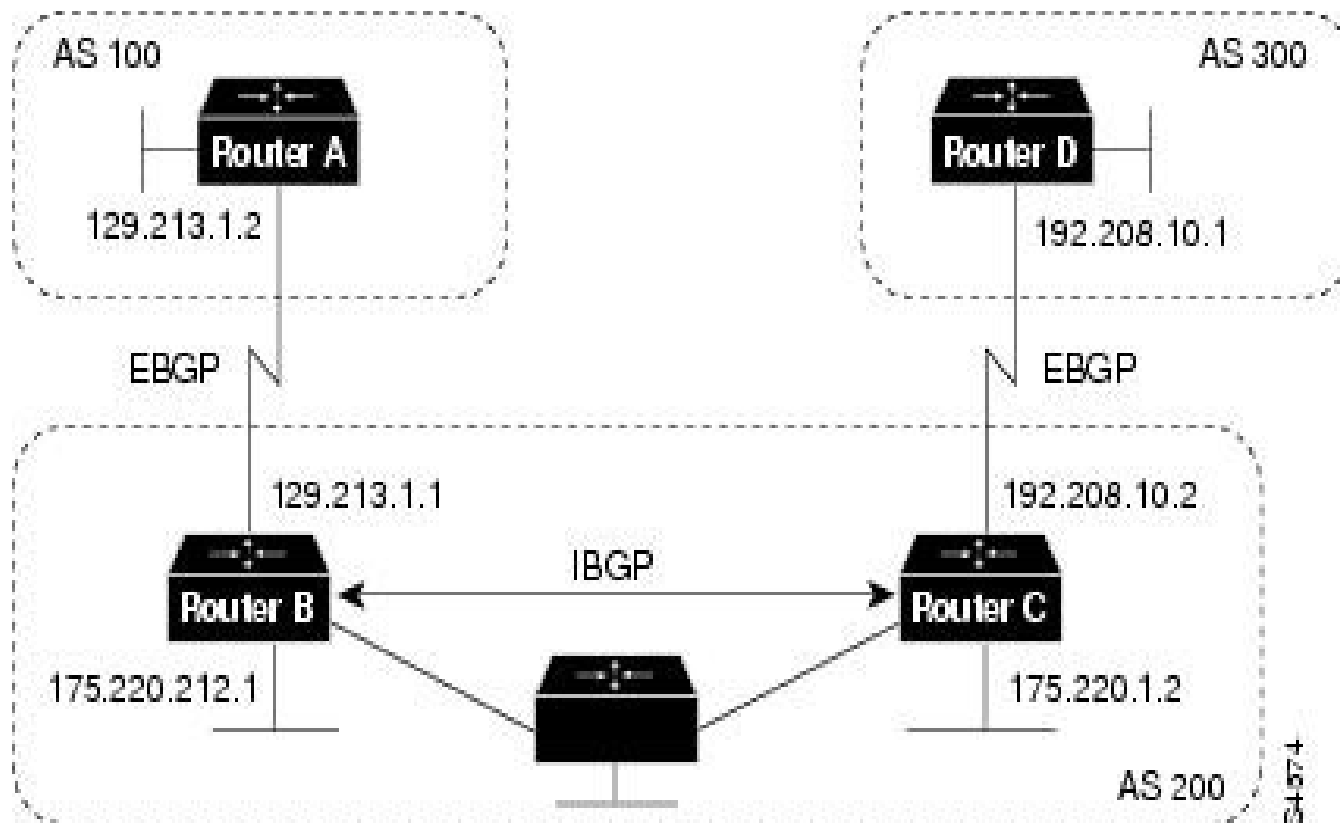
iBGP и eBGP

Когато BGP работи в рамките на AS, третира се като вътрешен (*iBGP Interior Border Gateway Protocol*).

Когато работи между ASs, нарича се външен (*eBGP Exterior Border Gateway Protocol*).

Маршрутизаторите на границата на дадена AS, които обменят информация с друга AS, се наричат *гранични* (*border* или *edge*).

iBGP и eBGP



iBGP и eBGP. Конфигурации.

Router B:

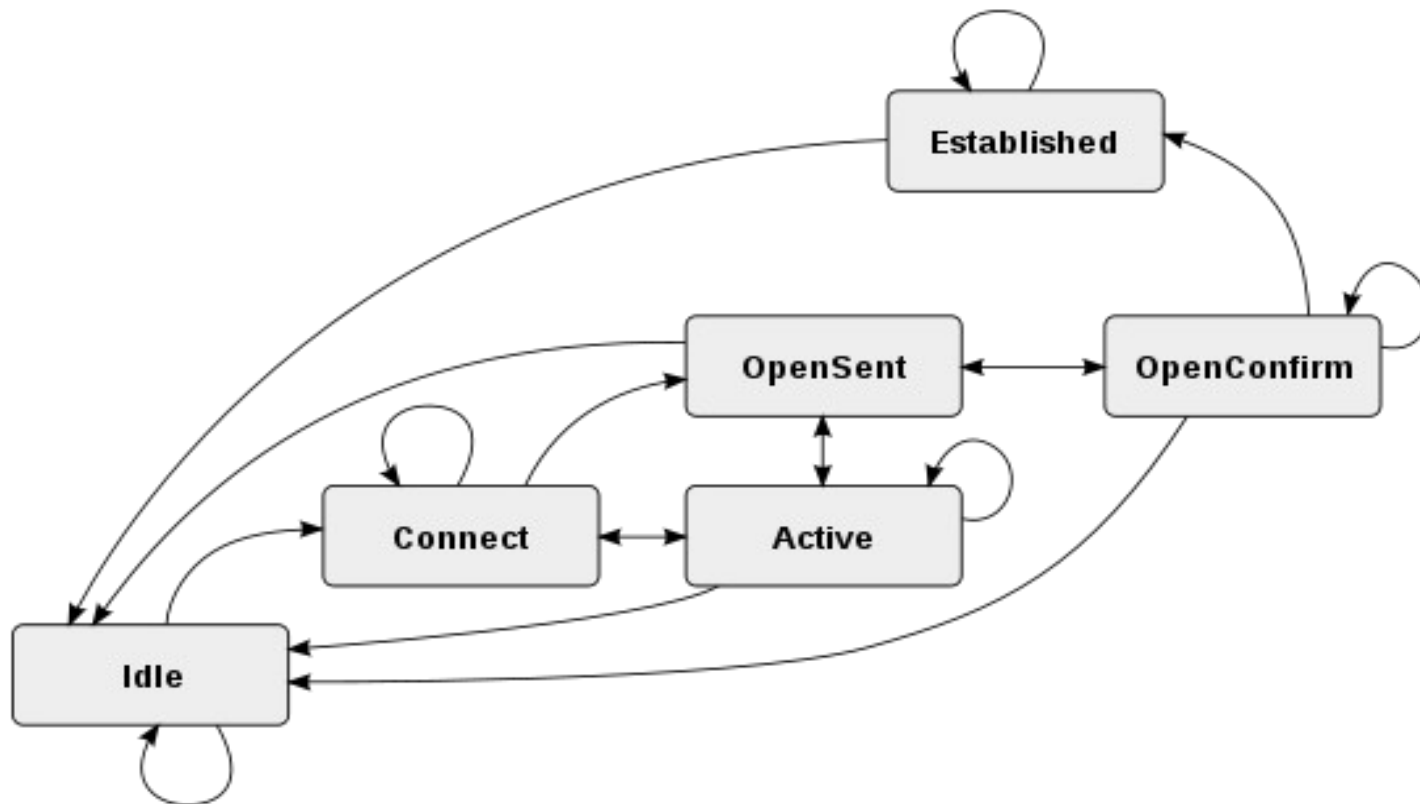
```
router bgp 200  
neighbor 129.213.1.2 remote-as 100  
neighbor 175.220.1.2 remote-as 200
```

Router C:

```
router bgp 200  
neighbor 175.220.212.1 remote-as 200  
neighbor 192.208.10.1 remote-as 300
```

Принцип на работа на BGP.

Схема на състоянията.



Състояния при установяване на BGP сесия

За да установи сесия с партньор (peer), BGP преминава през 6 състояния, описани с краен автомат (*finite state machine - FSM*).

Това са: *Idle, Connect, Active, OpenSent, OpenConfirm и Established*.

В BGP е дефинирана променлива на състоянието, която определя в кое от шестте състояния се намира сесията.

При преход от едно състояние в друго се генерират стандартни съобщения.

Idle, Connect и ...

Първоначално BGP маршрутизаторът е в състояние “Idle”. Инициализира всички ресурси, отказва всички входящи опити за установяване на BGP свързаност и инициира TCP сесия със съседа си.

Второто състояние е “Connect”.

Маршрутизаторът изчаква да се установи TCP сесия.

Ако е успешно, преминава в “OpenSent”.

Ако не, преминава в състояние “Active”, докато се нулира таймера ConnectRetry.

Active, OpenSent, Established

В състояние "Active" маршрутизаторът нулира таймера ConnectRetry, след което се връща в състояние "Connect".

След "OpenSent" маршрутизаторът изпраща съобщение Open и чака за подобно в отговор.

Разменят се съобщения Keepalive и след успех рутерът влиза в състояние "Established".

Готов е да изпраща и получава от съседа си съобщения Keepalive, Update и Notification.

BGP updates

С установяване на BGP сесия маршрутизаторите (BGP speakers - говорители) започват да си обменят UPDATE-и.

Те включват дестинациите, до които говорителят предлага свързаност.

В BGP описанието на маршрут до дадена дестинация се нарича **Network Layer Reachability Information (NLRI)**.

NLRI

NLRI включва префикса на дестинацията и дължината му, пътят през автономните системи и следващия възел, както и допълнителна информация - атрибути.

BGP говорителите съобщават за нови NLRI, до които предлагат достижимост, а също така оттеглят префикси, до които вече нямат свързаност.

Свързаност и научаване на маршрути в BGP

По принцип всички BGP маршрутизатори в дадена AS трябва да бъдат конфигурирани да “говорят” всеки с всеки (**full mesh**).

При това положение броят на връзките нараства квадратично с увеличаване на броя на рутерите.

BGP има две решения на това неудобство: рефлекторна схема (**route reflectors - RFC 4456**) и конфедерации (**confederations - RFC 5065**).

iBGP mesh

BGP рутер приема NLRI в UPDATE-те от множество съседни и анонсира (advertise - рекламира) NLRI към същите или други съседни.

BGP поддържа своя собствена "master" база от данни с маршрути, **Loc-RIB** (Local Routing Information Base). Тя е отделена от главната маршрутна таблица на рутера.

iBGP mesh

За всеки процес BGP поддържа “съседска” база с приетите от съсед NLRI-та - **Adj-RIB-In** (Adjacent Routing Information Base, Incoming).

Adj-RIB-Out (Outgoing) съдържа NLRI, които следва да бъдат изпратени на съсед.

Как ще бъдат съхранени и структурирани тези таблици (бази) зависи от конкретната реализация на BGP.

Легитимни маршрути

Възможно е Adj-RIBs и Loc-RIB да бъдат в една структура.

Допълнителни полета показват на кой съсед принадлежат дадени редове, дали те удовлетворяват “политиката” и дали са “легитимни” да попълнят маршрутната таблица.

Легитимни са най-добрите. Не е задължително BGP маршрут да бъде избран.

Например, директно свързан префикс, научен от хардуера, е най-предпочитания.

Легитимни маршрути

BGP не носи със себе си “политики”, а по-скоро информация, с чиято помощ BGP рутерите вземат “политически” решения, съгласно наложени правила.

Такава информация например са **Local Preference** и **multi-exit discriminators (MED)**.

Избор на маршрут

BGP се ръководи от различни фактори при взимане на решения даден NLRI да влезе в Loc-RIB.

На първо място е следващия възел по пътя (**next-hop** attribute) да бъде **достижим**.

Казано с други думи, в главната таблица с маршрутите **да има активен маршрут до префикса, към който принадлежи този “next-hop”**.

Избор на маршрут

- По-нататък за всеки съсед BGP прилага различни стандарти и критерии в зависимост от конкретната реализация, за да реши кои маршрути да се вмъкнат в Adj-RIB-In.
- Съседът може да изпрати множество маршрути до дадена дестинация, но само най-предпочитания ще влезе в Adj-RIB-In.
- Маршрути, които са оттеглени от съсед, ще бъдат изтрети от Adj-RIB-In.

Избор на маршрут

При промяна в Adj-RIB-In BGP процесът решава дали някой от новите маршрути са за предпочитане пред наличните в Loc-RIB.

Ако да, ги заменя. Ако даден маршрут е оттеглен от съсед и няма друг маршрут до тази дестинация, маршрутът се премахва от Loc-RIB и повече не се препраща към управлението на главната маршрутна таблица.

Ако не съществува алтернативен маршрут от друг не-BGP протокол, оттегленият маршрут се премахва от маршрутната таблица.

Решения на ниво съсед

1. Маршрут с **най-къс AS_PATH** - поредицата от номера на AS, които трябва да бъдат прекосени, за да се стигне до дестинацията. AS1-AS2-AS3 е по-къс от AS4-AS5-AS6-AS7.
2. Маршрути с **най-ниска стойност** на атрибута **ORIGIN**.
 1. Маршрути с **най-висока стойност** на **LocPref**.
 1. Маршрути с **най-ниска стойност** на **MED**.

Решения на ниво Loc-RIB

След като кандидат маршрутите са получени от съседите, Loc-RIB софтуерът прилага допълнителни критерии:

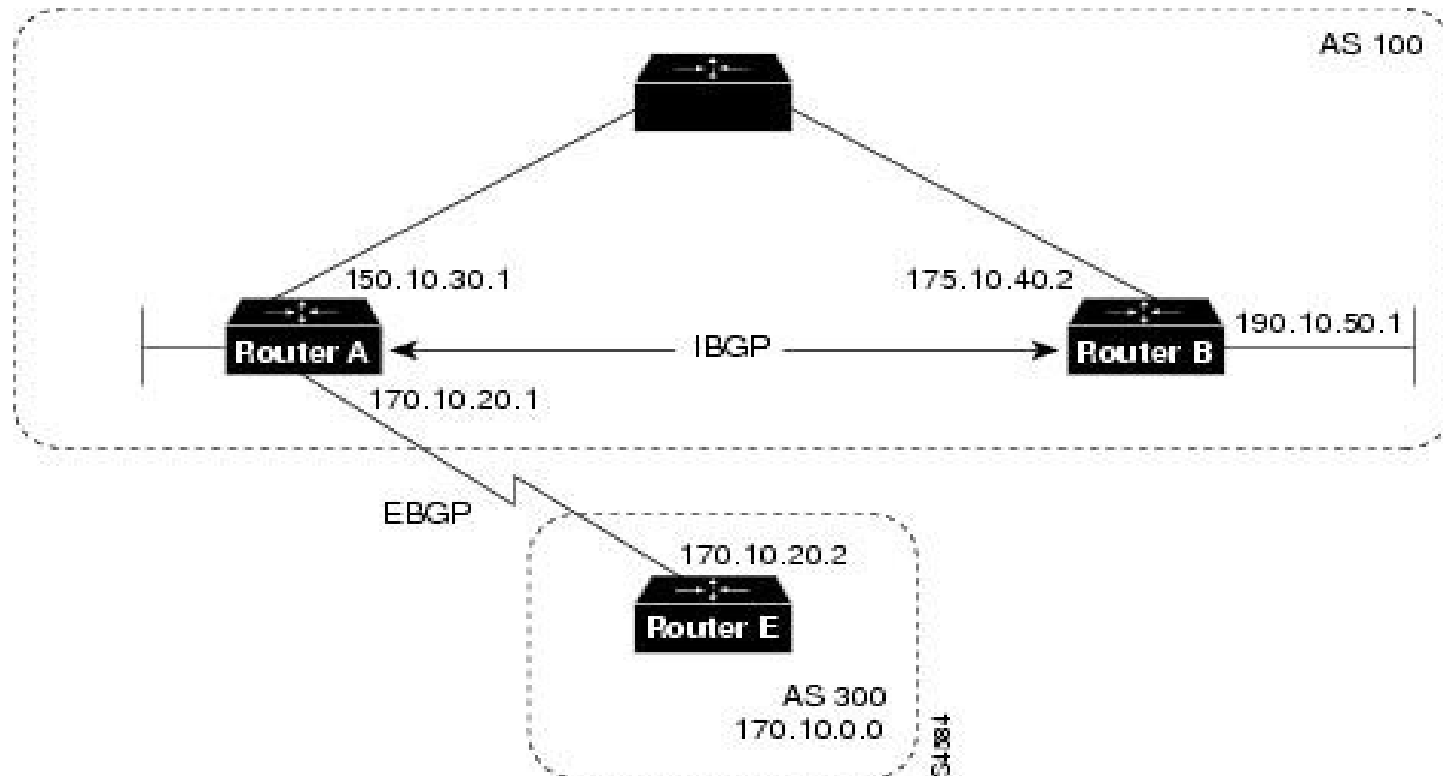
1. Ако **поне един маршрут** е научен от външен съсед (т.е. **по eBGP**), всички маршрути, научени от iBGP, се изхвърлят.

Решения на ниво Loc-RIB

2. Предпочита се маршрута с най-ниска вътрешна стойност до NEXT_HOP според главната маршрутна таблица.

Ако един и същ маршрут е рекламиран от два съседа, но единият съсед е достижим по по-бавна връзка от другия и вътрешният протокол изчислява по-ниска стойност на базата на по-високата скорост, ще бъде предпочетен маршрутът по по-бързия канал.

BGP. Origin. Next Hop.



Origin

- **IGP (i)**—маршрутът е вътрешен за AS, от която идва, създаден чрез команда **network**.
- **EGP (e)**—Маршрутът е научен от EGP.
- **Incomplete (?)**—Произходът на маршрута не е известен

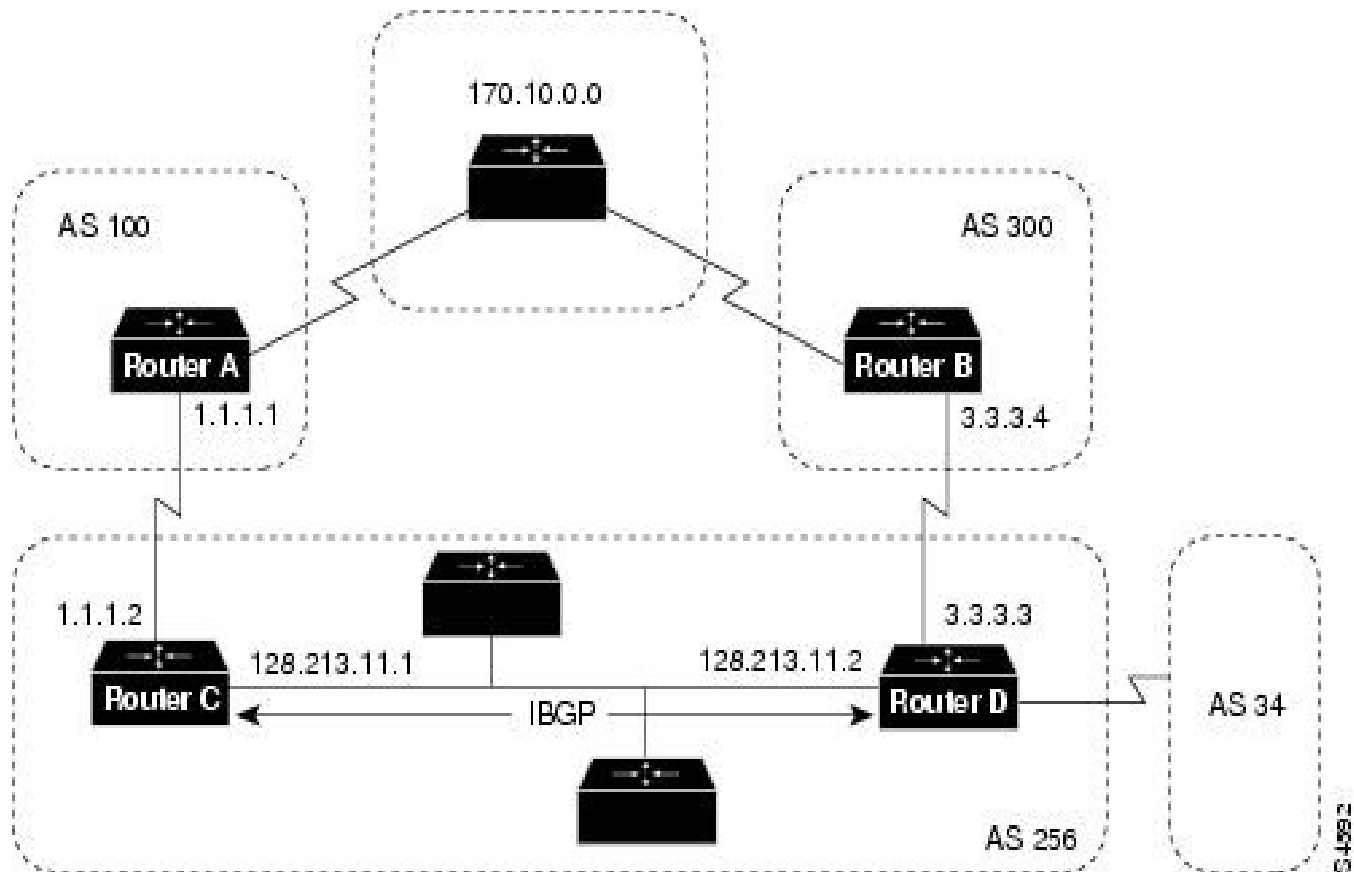
Local Preference

Когато имаме множество пътища до една и съща дестинация, пътят с най-висока стойност на атрибута **Local Preference** (**LocPref**) е предпочитан.

По подразбиране **LocPref** е 100.

LocPref се обменя между рутерите в една AS заедно с обновяването на маршрутите.

Local Preference



Local Preference

!Router C

```
router bgp 256
```

```
neighbor 1.1.1.1 remote-as 100
```

```
neighbor 128.213.11.2 remote-as 256
```

```
bgp default local-preference 150
```

!Router D

```
router bgp 256
```

```
neighbor 3.3.3.4 remote-as 300
```

```
neighbor 128.213.11.1 remote-as 256
```

```
bgp default local-preference 200
```

Local Preference

!Router D

```
router bgp 256
```

```
neighbor 3.3.3.4 remote-as 300
```

```
route-map SETLOCALIN in
```

```
neighbor 128.213.11.1 remote-as 256
```

```
!
```

```
ip as-path 7 permit ^300$
```

```
route-map SETLOCALIN permit 10
```

```
match as-path 7
```

```
set local-preference 200
```

```
route-map SETLOCALIN permit 20
```

Показване на Origin, Next Hop...

```
bgpd@border-lozenetz# sh ip bgp 2.0.0.0
```

BGP routing table entry for 2.0.0.0/16

Paths: (1 available, best #1, table Default-IP-Routing-Table)

Advertised to non peer-group peers:

```
62.44.127.2 62.44.127.11 62.44.127.15 62.44.127.16  
 62.44.127.19 62.44.127.23 62.44.127.43 62.44.127.51  
 62.44.127.52 62.44.127.61 62.44.127.70 62.44.127.71  
 62.44.127.72 62.44.127.73
```

```
6802 20965 559 30132 12654
```

```
194.141.252.21 from 194.141.252.21 (194.141.252.13)
```

Origin IGP, localpref 100, valid, external, best

Community: 6802:1

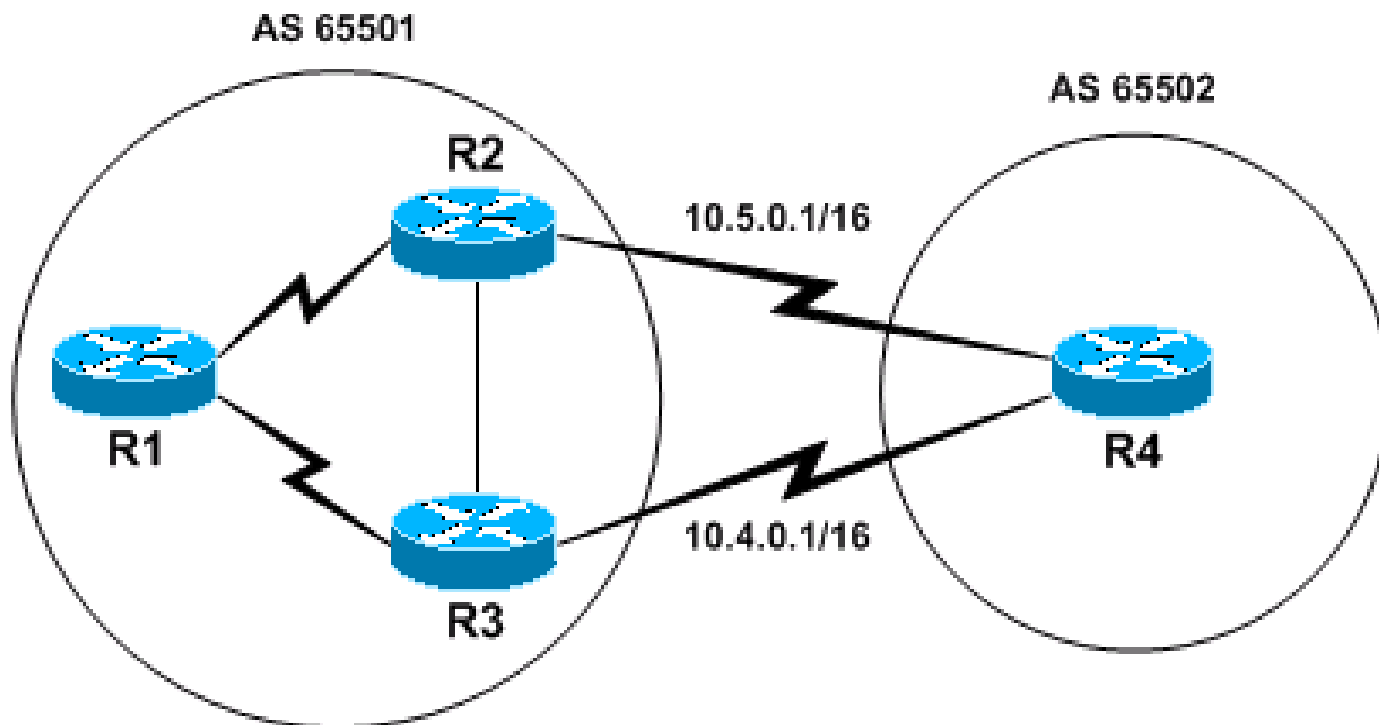
Last update: Sun Dec 13

MED

MED (multi-exit discriminator), подсказва на съседите предпочитания път в AS, която има множество входни точки.

MED е известна и като външна метрика на маршрут. MED с по-ниска стойност е предпочитана пред по-високата.

MED



MED

```
router bgp 65502
```

```
....
```

```
neighbor 10.5.0.2 remote-as 65501
```

```
neighbor 10.5.0.2 route-map setMED-R2 out
```

```
neighbor 10.4.0.2 remote-as 65501
```

```
neighbor 10.4.0.2 route-map setMED-R3 out
```

```
!
```

```
route-map setMED-R3 permit 10
```

```
match ip address 1
```

```
set metric 200
```

```
!
```

```
route-map setMED-R3 permit 20
```

```
match ip address 2
```

```
set metric 100
```

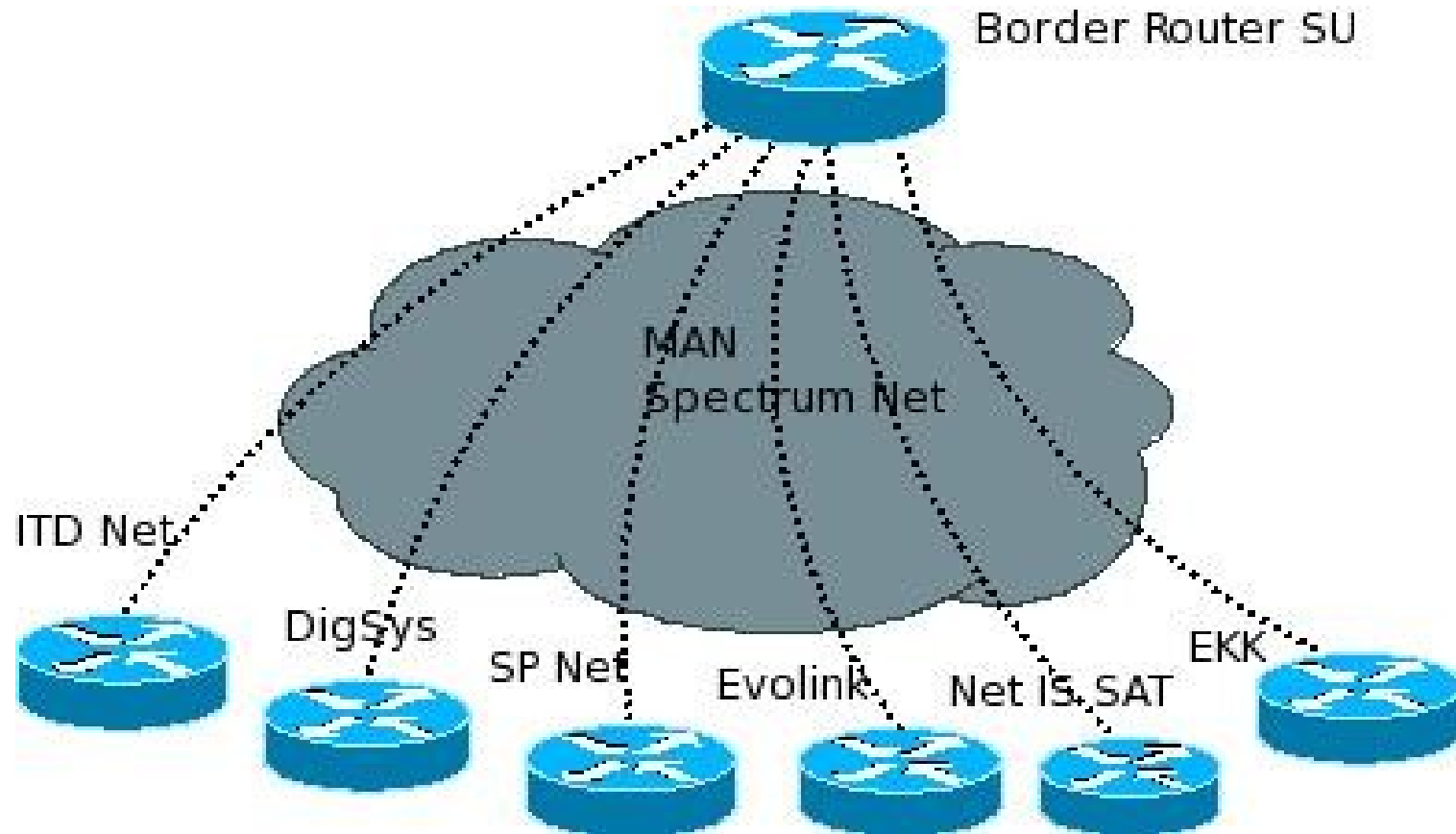
BGP Peer Groups

BGP peer group представлява група от BGP съседни, които споделят обща политика, определена от маршрутни карти и филтри - **route maps, distribution lists**.

Вместо политиката да се прилага на всеки съсед поотделно, тя се прилага върху цялата група.

Членовете на групата наследяват всички конфигурации на групата.

Peering партньори на СУ (AS 5421)



BGP Peer Groups. Конфигурация.

```
neighbor PEERING_SINGLE_IPV4 peer-group
neighbor PEERING_SINGLE_IPV4 activate
neighbor PEERING_SINGLE_IPV4 soft-
  reconfiguration inbound
neighbor PEERING_SINGLE_IPV4 maximum-
  prefix 50000
neighbor PEERING_SINGLE_IPV4 route-map
  PEERING_SINGLE_IMPORT_IPV4 in
neighbor PEERING_SINGLE_IPV4 route-map
  PEERING_SINGLE_EXPORT_IPV4 out
```

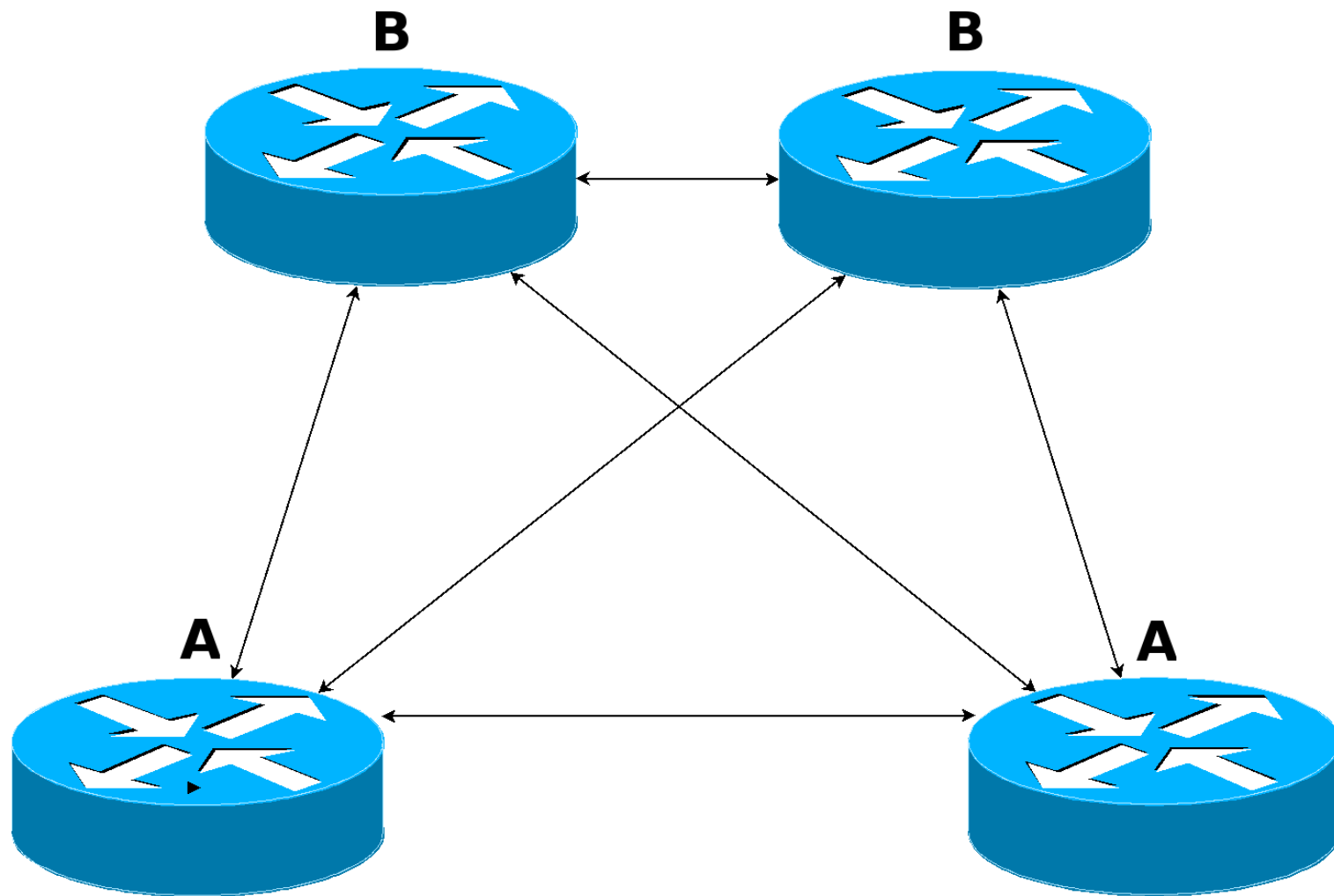
BGP Peer Groups. Конфигурация.

```
neighbor 62.44.96.250 remote-as 8717
neighbor 62.44.96.250 peer-group
    PEERING_SINGLE_IPV4
neighbor 62.44.96.250 description
    SPNET_PEERING_IPV4
```

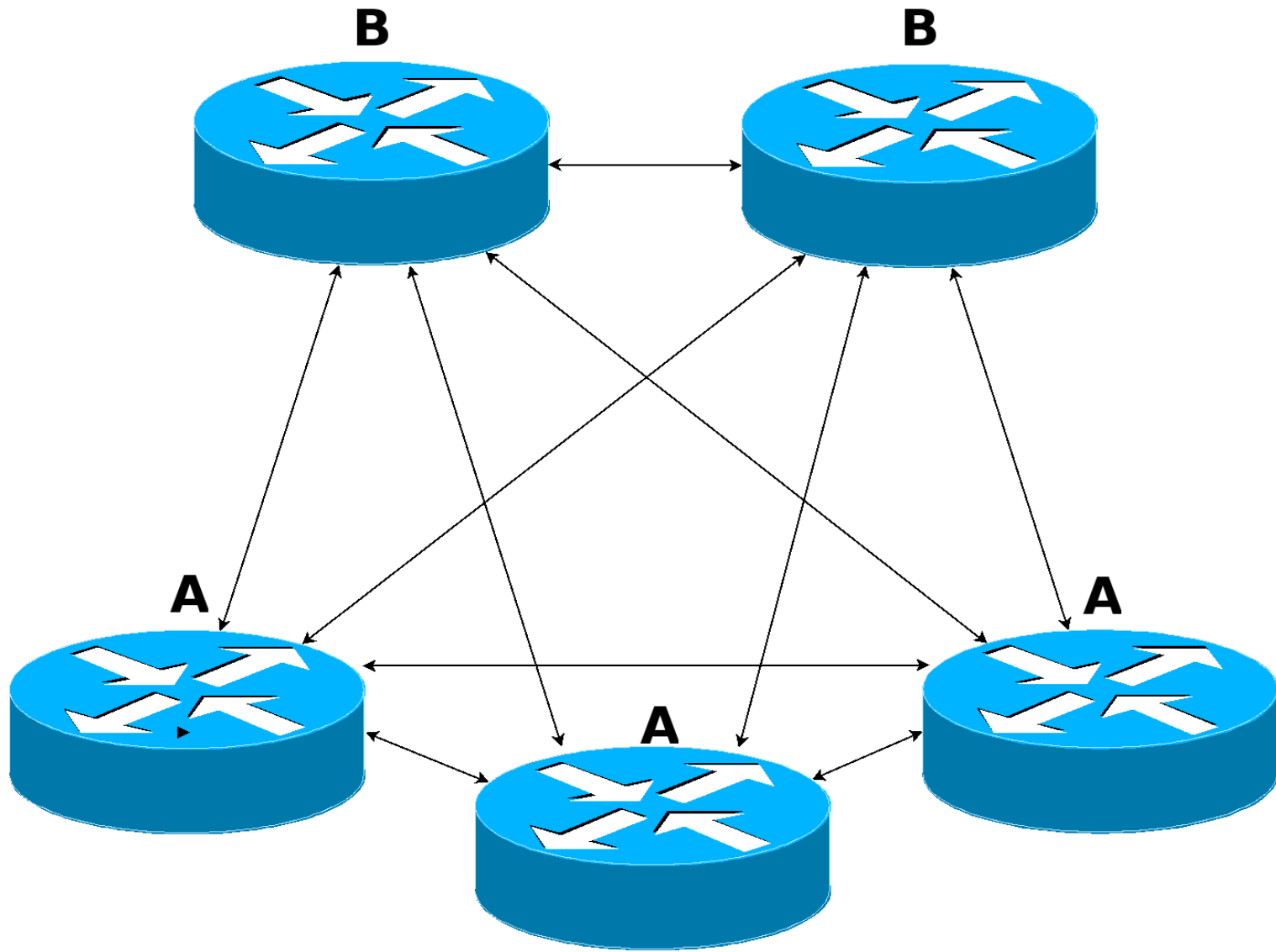
Рефлектори

- В автономна система с iBGP трябва да има свързаност към всички iBGP peers (съседи), т.е “всеки с всеки” - **full mesh**.
- С помощта на рефлекторна схема се редуцира броя на iBGP съседите и от там натоварването на процесори и комуникационни канали.
- Един рутер (или два за резервираност) става рефлекторен сървър, а другите – рефлекторни клиенти.

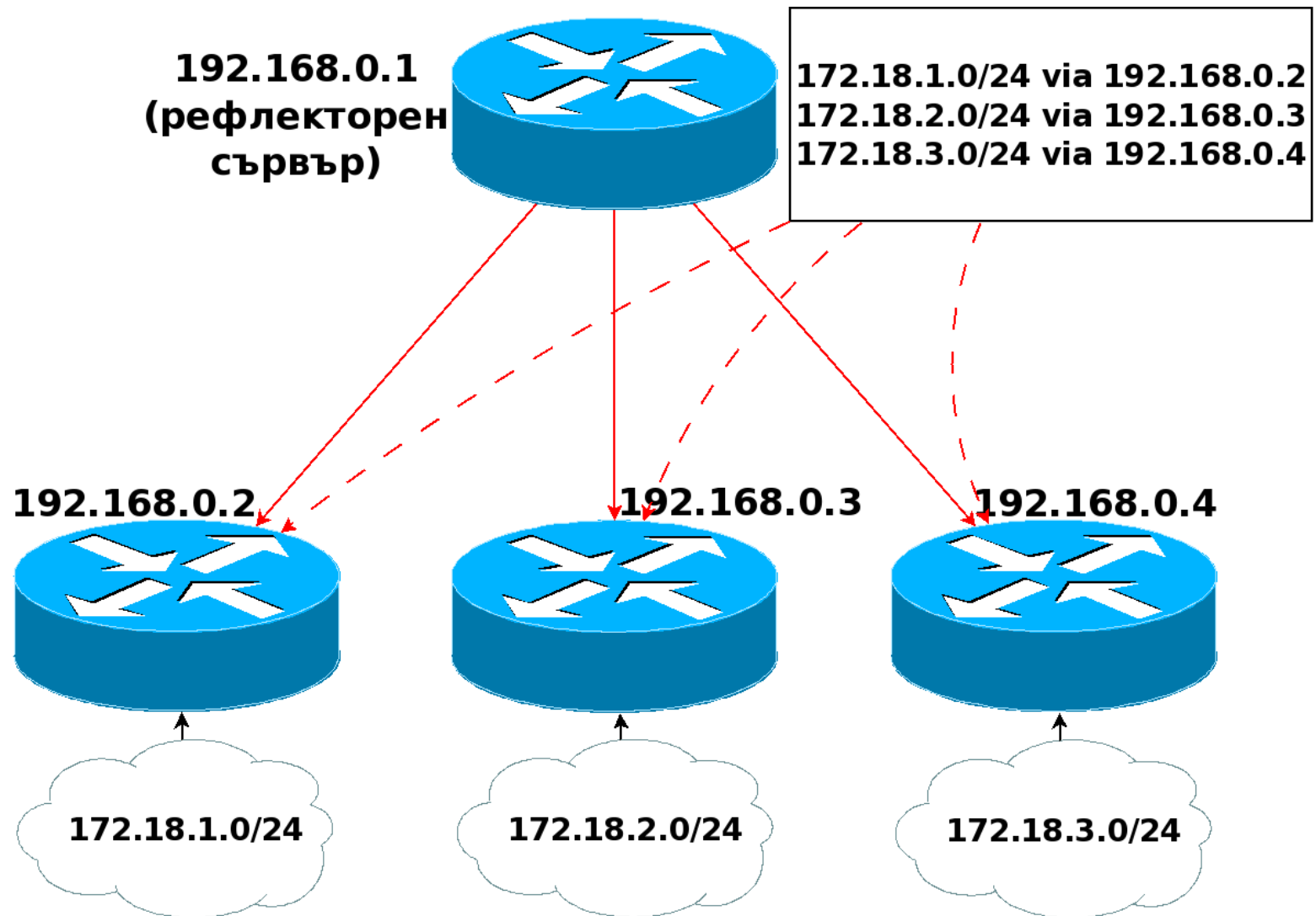
**Full-mesh: 4 маршрутизатора =>
6 вътрешни BGP сесии**



**5 маршрутизатора => 10
вътрешни BGP сесии**



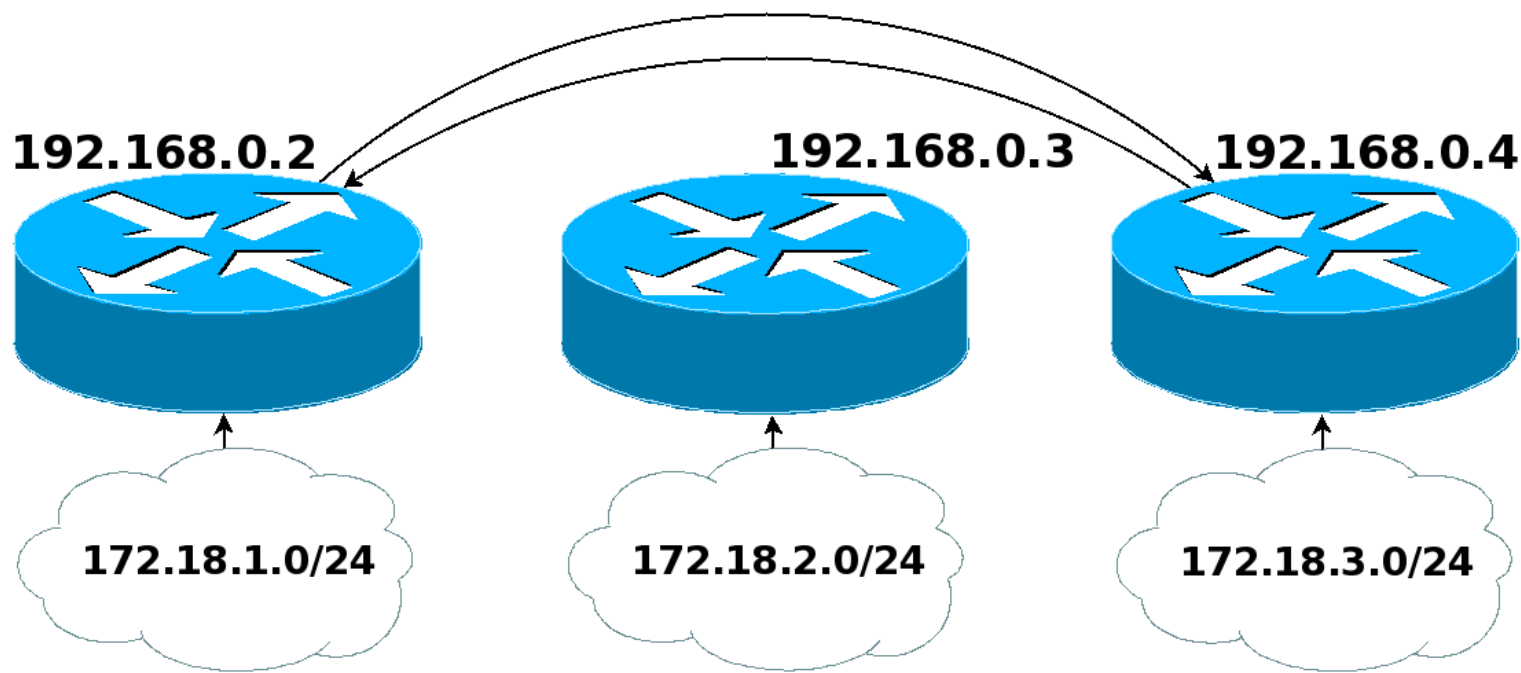
Рефлекторният сървър при решаване на задачата



Обмен на трафик между клиентите

172.18.1.0/24 via 192.168.0.2
172.18.2.0/24 via 192.168.0.3
172.18.3.0/24 via 192.168.0.4

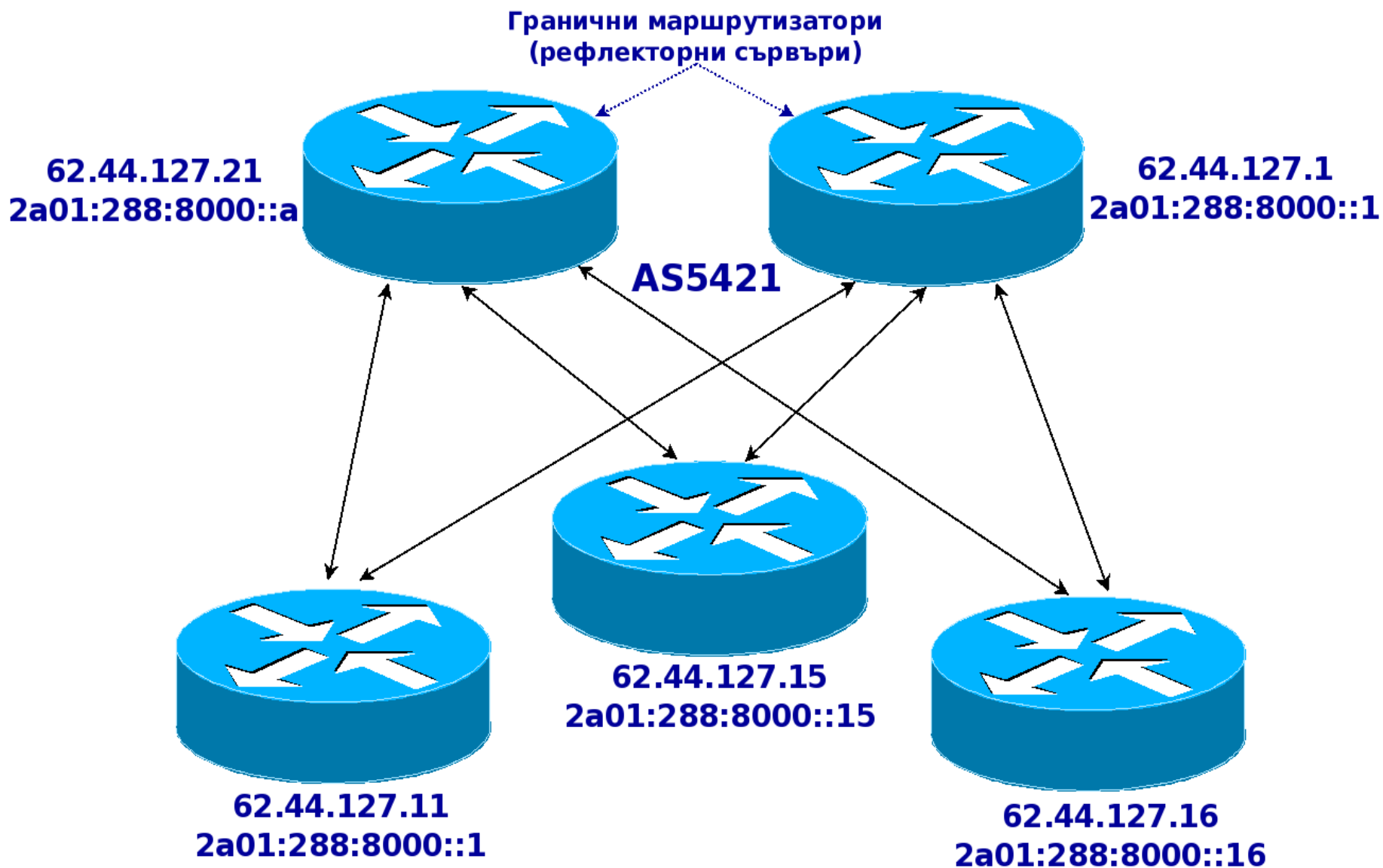
**ОБМЕН НА ТРАФИК ПО НАЙ-КРАТКИЯ ПЪТ -
ДИРЕКТНО МЕЖДУ МАРШРУТИЗАТОРИТЕ В ЕТЕРНЕТ СЕГМЕНТА,
БЕЗ УЧАСТИЕ НА РЕФЛЕКТОРНИЯ СЪРВЪР ПРЯКО В
МАРШРУТИЗАЦИЯТА**



Рефлекторна схема в СУ



Рефлекторна схема в СУ



Големина на маршрутната таблица

Един от основните проблеми пред BGP, респ. Internet, е растежа на глобалната таблица с маршрутите.

Не всички рутери са в състояние да я поемат (RAM, CPU) и ефективно да обработват трафика.

И, още по-важно, голямата, колкото е по-голяма таблицата, толкова по-бавно се стабилизира (конвергира).

В момента броят на префиксите в Глобалната мрежа е малко над 300 000.

Префиксите подавани от GEANT (194.141.252.21) на СУ

```
bgpd@border-lozenetz# sh ip bgp summary
```

```
BGP router identifier 62.44.127.21, local AS number  
5421
```

```
RIB entries 563365, using 34 MiB of memory
```

Neighbor	V	AS	MsgRcvd	MsgSent	TblVer
InQ	OutQ	Up/Down	State/PfxRcd		
194.141.252.21	4	6802	71565	1015	0 0
0 20:55:05	302776				

Решение: CIDR и агрегиране (summary) на маршрути

Да приемем, че на AS1 е присвоено адресно пространство 172.16.0.0/16,

AS1 иска да анонсира по-специфични маршрути: 172.16.0.0/18, 172.16.64.0/18 и 172.16.128.0/18.

Префиксът 172.16.192.0/18 не съдържа никакви хостове и AS1 не го анонсира.

При това положение AS1 ще анонсира 4 маршрута: 172.16.0.0/16 и 172.16.0.0/18, 172.16.64.0/18, 172.16.128.0/18.

Решение: CIDR и агрегиране на маршрути

Тези 4 маршрута ще бъдат видяни от AS2.

Въпрос на политика е дали да ги копира 4-те или или да запише само обобщения, 172.16.0.0/16.

Ако AS2 иска да изпрати данни към 172.16.192.0/18, те ще се отправят по маршрут 172.16.0.0/16.

Граничният маршрутизатор на AS1 или ще изхвърли пакета, или ще го върне като “unreachable” в зависимост от конфигурацията.

Решение: CIDR и агрегиране на маршрути

Ако AS1 реши да изхвърли маршрут 172.16.0.0/16 и остави 172.16.0.0/18, 172.16.64.0/18 и 172.16.128.0/18, в таблицата ѝ ще има три маршрута.

AS2 ще вижда тези три маршрута в зависимост от политиката си или ще запише в паметта и трите, или ще агрегира префиксите 172.16.0.0/18 и 172.16.64.0/18 на 172.16.0.0/17.

Тогава в паметта на граничния маршрутизатор на AS2 ще се съхраняват само два маршрута: 172.16.0.0/17 и 172.16.128.0/18.

Решение: CIDR и агрегиране на маршрути

Ако AS2 иска да изпрати данни към 172.16.192.0/18, те ще бъдат изхвърлени или към маршрутизаторите в AS2 ще бъде изпратено съобщение “unreachable” (а не към AS1), защото 172.16.192.0/18 няма да е в маршрутната таблица.