Starting and running an IXP

All that stuff around the switch
Some guidelines

Matjaž Straus Istenič, SIX/ARNES

Agenda

- all that stuff around the switch
- practical examples
  - addressing
  - configuration examples
  - guidelines and hints for members

Stuff around the switch

- proper location with many fibre providers
  - a building with one single provider is a bad idea
- different fibre paths inside of the building
- power supplies and grounding
- cooling system
- physical security
- staff, support, remote hands
- good and accurate documentation

Stuff around the switch (cont.)

- monitoring and alarming
- ticketing system
- mailing lists
- web portal
- best current practices and knowledge base
- contracts, SLAs, billing, ...

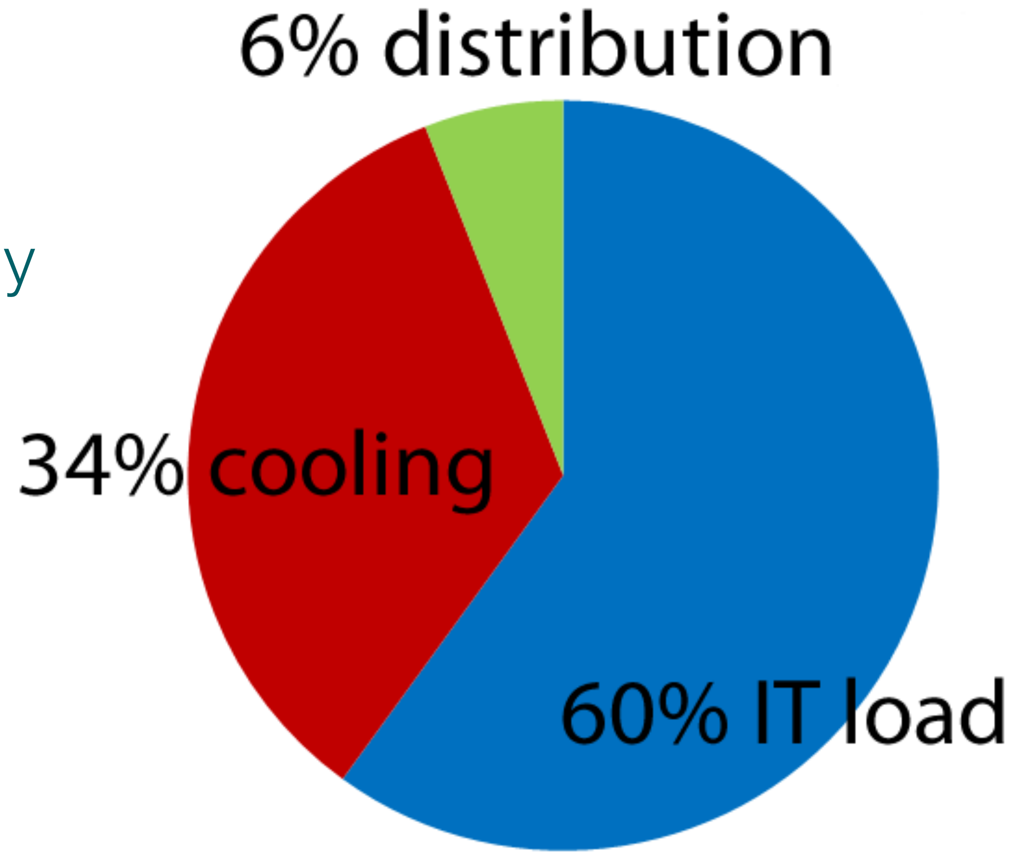
- planning for a collocation/datacenter

The power

©Frank Sibayan

The power

- allocate up to 20 kW per rack
- actual usage 5 kW - 10 kW per rack
- dual separate circuit breaker for each rack
- power supply redundancy
  - dual feed from electrical distribution company
  - separate dual UPS system N+1 and PDU
  - diesel generator
- cooling equipment is independently dual powered, including chillers
- how much power does datacenter use
  - monitoring on UPS, on PDU
  - monitoring total on main branch circuit
- typicaly the load will double in 5 years



Cooling

- full redundancy of cooling system
  - two different power grids
  - separate piping
  - chiller redundancy
  - room units redundancy
- hot/cold isle
  - reduce air mixing
  - cold aisle with barriers made of metal, plastic or fiberglass
  - use blanking panels on the cabinets without servers
- no need for double floor
  - run network cabling over the top of the cabinets
  - "in row" cooling
- recommended temperature in cold isle is between 23 - 25 °C
- cooling system rating must be 1.3 x IT load rating
- make sure that the space will allow for future growth
  - for more cooling capacity and redundancy if required
- Power usage effectiveness (PUE = Total Facility Power/IT Equipment Power)
  - typical PUE is 2.0 or higher

Fire protection

- sensing the smoke/fire

type	✔	✗
aspiration sensor	• very sensitive • early warning • single point of electrical instalation • targeted sensing is possible	• more expensive • plastic air ducting under the ceiling must be installed
optical sensor	• cheaper • can be used as confirmation for fast aspiration sensors	• less sensitive • each sensor needs its own cable

Fire protection

- extinguishing fire

Gaseous fire extinguishing system

All are considered safe for breathing after release, although, products of burning plastics are always dangerous!

type	active substance	✔	✗
displacement of air	Inergen - mixture of gases, displaces air with “air” with less oxygen	• totally natural • environmentaly neutral	• big storage requirements • high pressure (200 or 300 bar) • computer room needs big exhaust vents • bug rush of gas at release causes dust and objects to lift
chemical action	Novec 1230 - chemical bonding, cooling	• small storage area • stored as fluid • very small greenhouse gas footprint	• has some effect on environment • expensive • stored under pressure (40/50 bar)
	FM200 (phasing out) - chemical bonding	• small storage area • small greenhouse gas footprint	• being phased out • has some ozone depletion impact • stored under pressure (40/50 bar)
cooling	water mist	• totally natural • environmentaly neutral	• water in computer room is not a good idea ;-) • possible condensation on cold surfaces

Examples and guidelines

- addressing
- port configuration
- guidelines for members

Examples: addressing

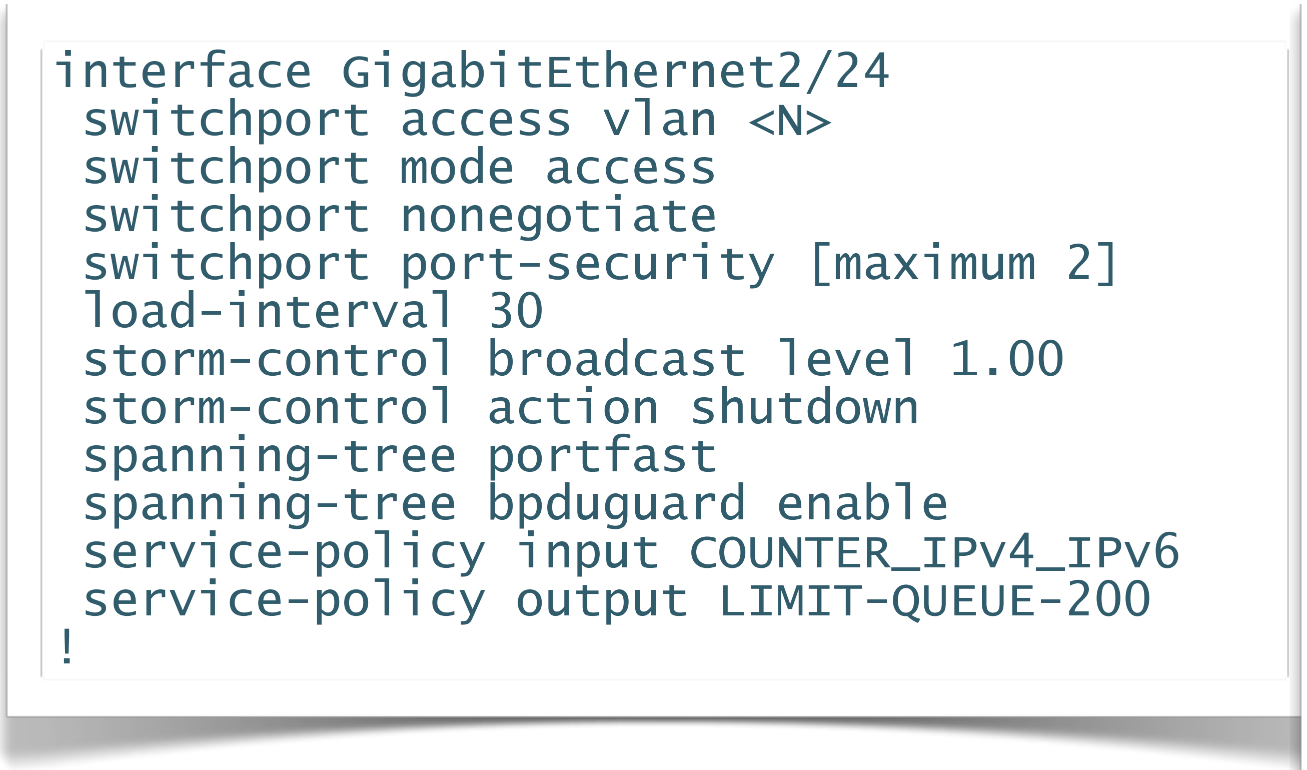
- a single subnet taken from independent address space
  - member address is assigned per location
- address schema at SIX

91.220.194.n/24
$n = n_1 = 2..99$ at location 1
$n = n_1 + 100 = 102..199$
at location 2
n = 1, 101 for route-reflectors

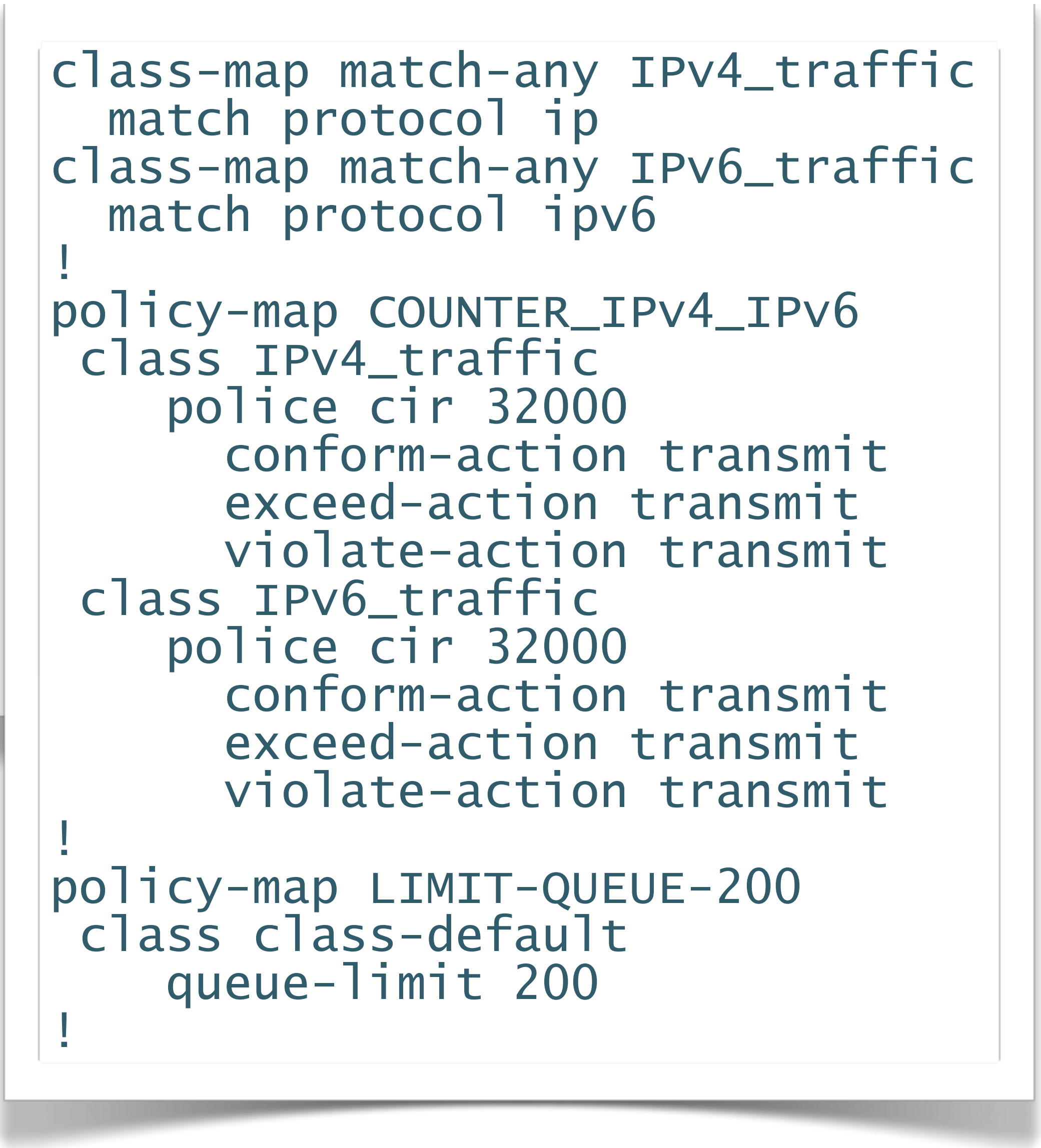
2001:7f8:46:0:L:N::<AS>/64
L = 0 at location 1
L = 1 at location 2
N = 0 for a single router,
otherwise N = 1, 2, ...
AS = member AS in decimal
AS = 51988 for RRs
- diverse lower 24 bits which form solicited-node mcast address

Examples: port configuration

- access port on Cisco 4900M



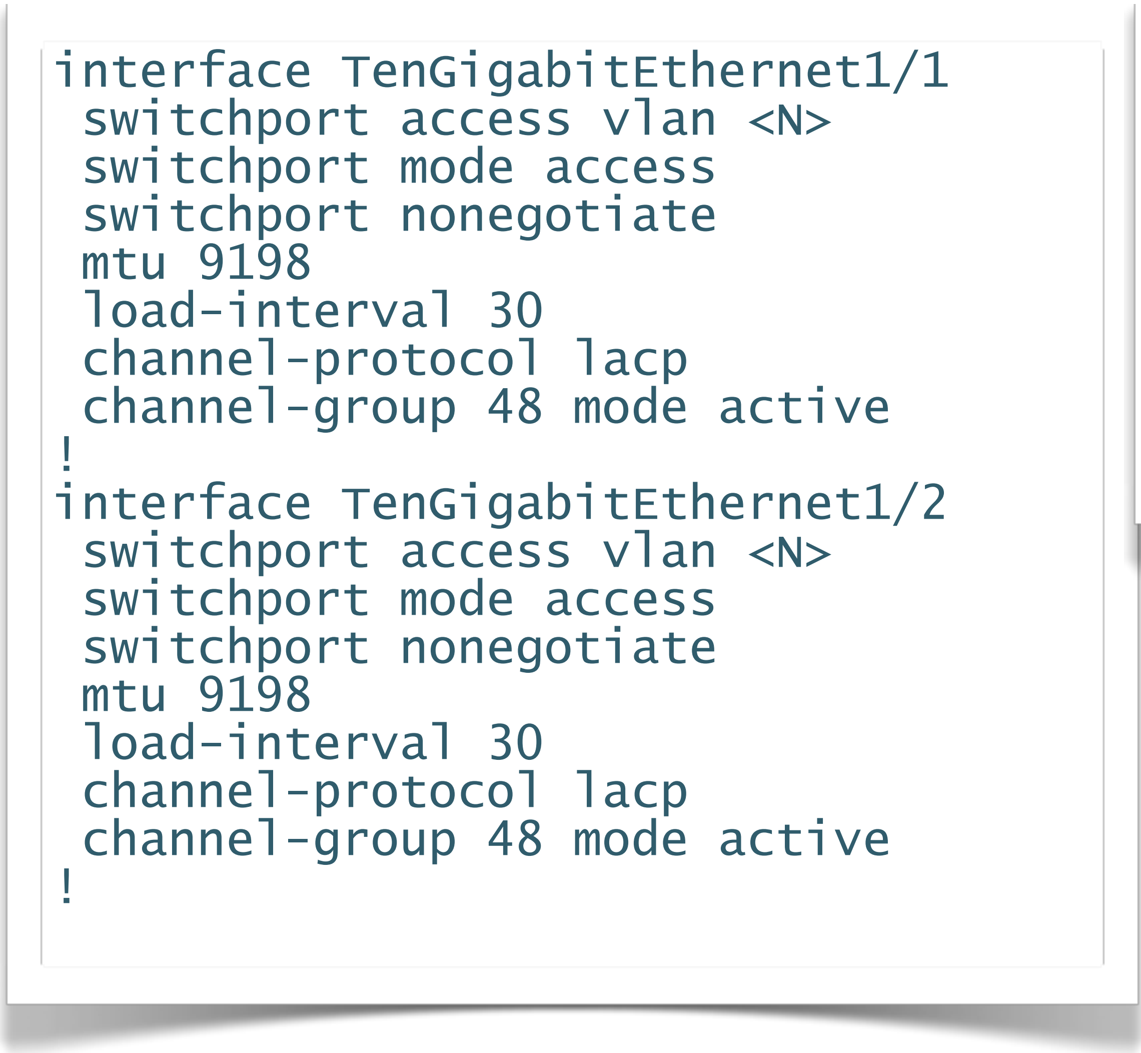
```
interface GigabitEthernet2/24
 switchport access vlan <N>
 switchport mode access
 switchport nonegotiate
 switchport port-security [maximum 2]
 load-interval 30
 storm-control broadcast level 1.00
 storm-control action shutdown
 spanning-tree portfast
 spanning-tree bpduguard enable
 service-policy input COUNTER_IPv4_IPv6
 service-policy output LIMIT-QUEUE-200
!
```



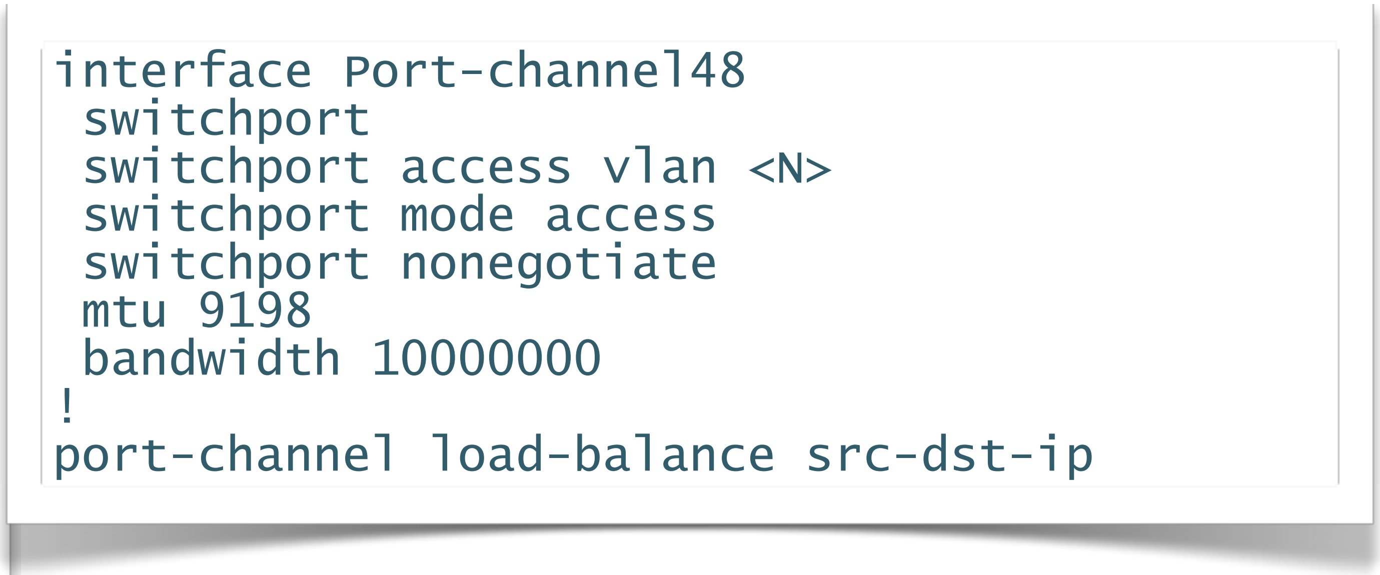
```
class-map match-any IPv4_traffic
  match protocol ip
class-map match-any IPv6_traffic
  match protocol ipv6
!
policy-map COUNTER_IPv4_IPv6
 class IPv4_traffic
    police cir 32000
      conform-action transmit
      exceed-action transmit
      violate-action transmit
 class IPv6_traffic
    police cir 32000
      conform-action transmit
      exceed-action transmit
      violate-action transmit
!
policy-map LIMIT-QUEUE-200
 class class-default
    queue-limit 200
!
```

Examples: port configuration

- interconnecting ports
  - aggregated to EtherChannel with LACP
  - maximal MTU



```
interface TenGigabitEthernet1/1
 switchport access vlan <N>
 switchport mode access
 switchport nonegotiate
 mtu 9198
 load-interval 30
 channel-protocol lacp
 channel-group 48 mode active
!
interface TenGigabitEthernet1/2
 switchport access vlan <N>
 switchport mode access
 switchport nonegotiate
 mtu 9198
 load-interval 30
 channel-protocol lacp
 channel-group 48 mode active
!
```



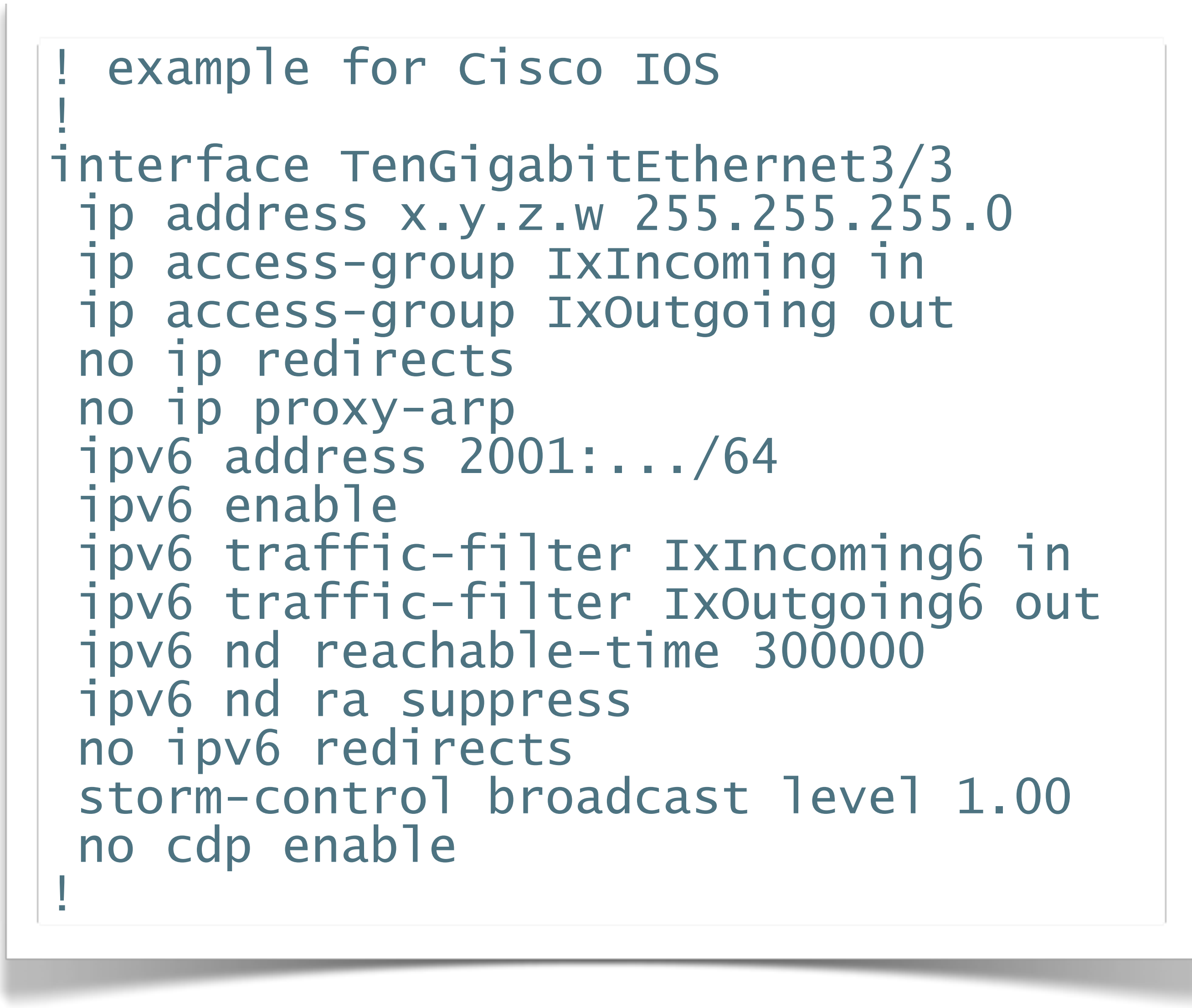
```
interface Port-channel48
 switchport
 switchport access vlan <N>
 switchport mode access
 switchport nonegotiate
 mtu 9198
 bandwidth 10000000
!
port-channel load-balance src-dst-ip
```

Guidelines for members

- access port configuration
- BGP
  - routing considerations
  - MD5 authentication
  - filtering announcements
    - control received prefixes
    - control advertised prefixes

Example: access port configuration

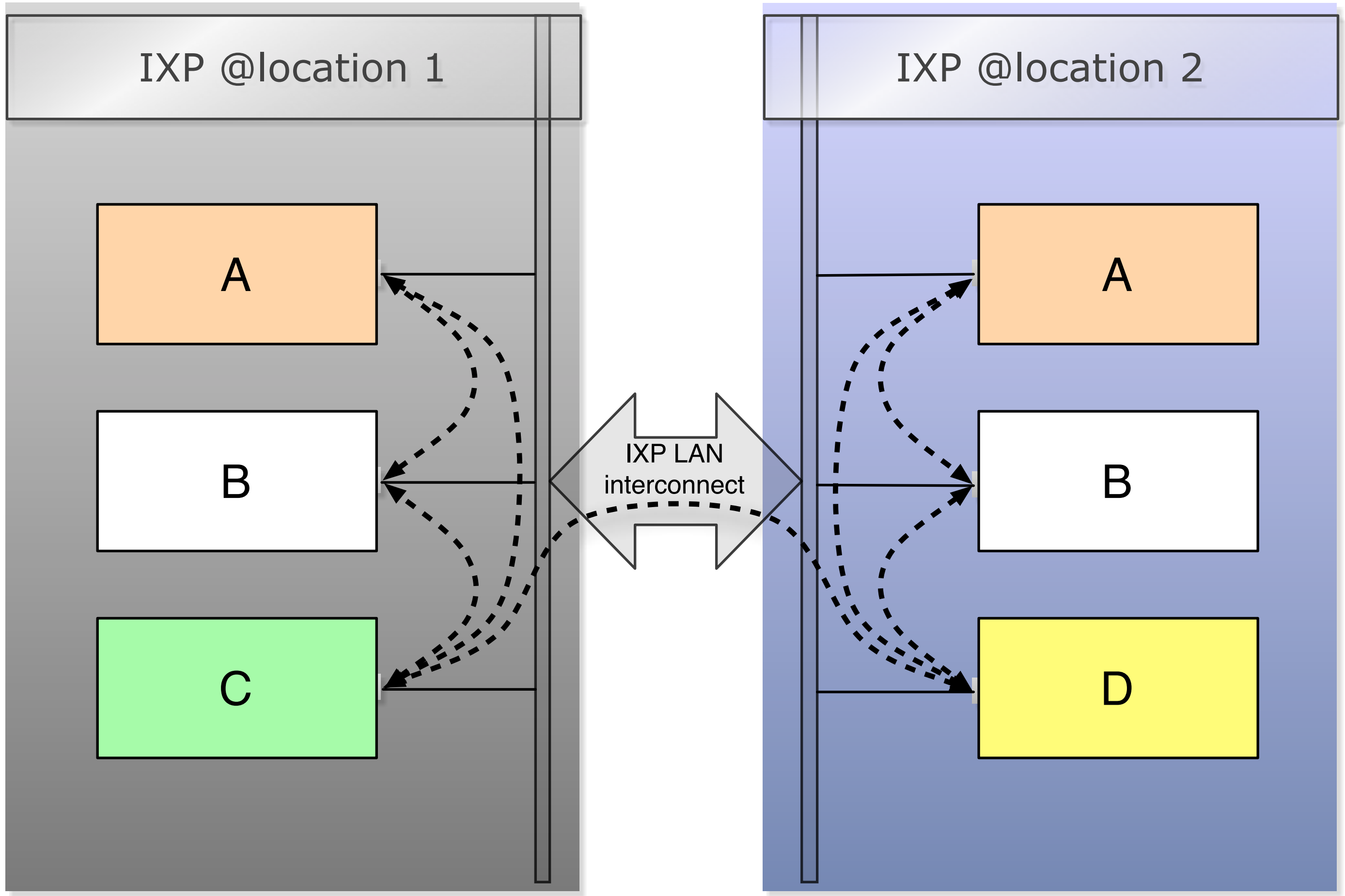
- turn off anything but IP and ARP
  - no redirects
  - no vendor proprietary protocols like CDP
  - no broadcast
  - no IPv6 RA
  - ! ICMP unreachables are used in PMTU discovery



```
! example for Cisco IOS
!
interface TenGigabitEthernet3/3
 ip address x.y.z.w 255.255.255.0
 ip access-group IxIncoming in
 ip access-group IxOutgoing out
 no ip redirects
 no ip proxy-arp
 ipv6 address 2001:.../64
 ipv6 enable
 ipv6 traffic-filter IxIncoming6 in
 ipv6 traffic-filter IxOutgoing6 out
 ipv6 nd reachable-time 300000
 ipv6 nd ra suppress
 no ipv6 redirects
 storm-control broadcast level 1.00
 no cdp enable
!
```

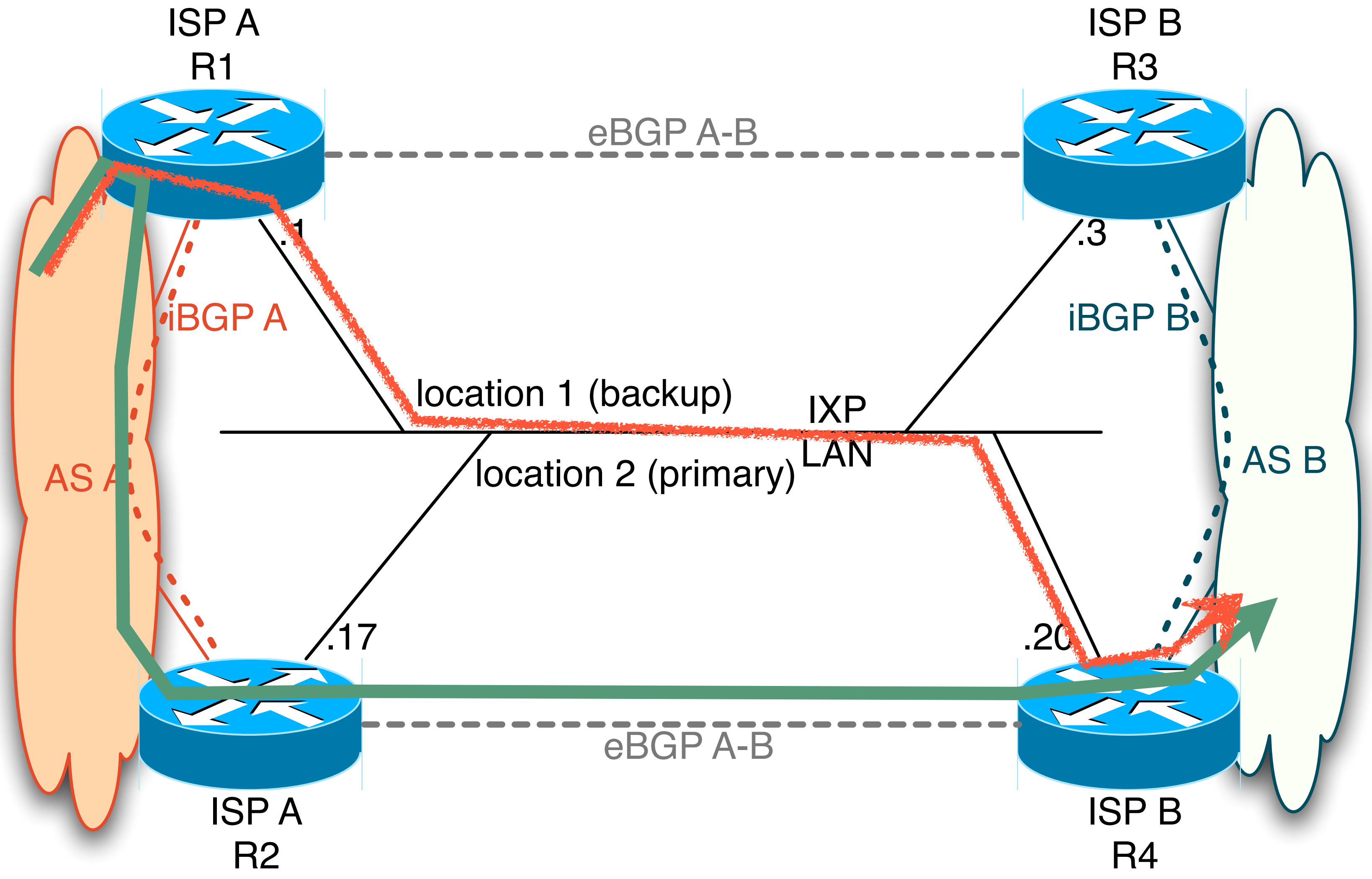
Multiple locations

- routing considerations
    - localize traffic
    - minimize traffic between locations



Examples: two members prefer one location

- the importance of next-hop self in iBGP
  - a member should use next-hop self in his iBGP sessions to avoid using the IX interconnect link

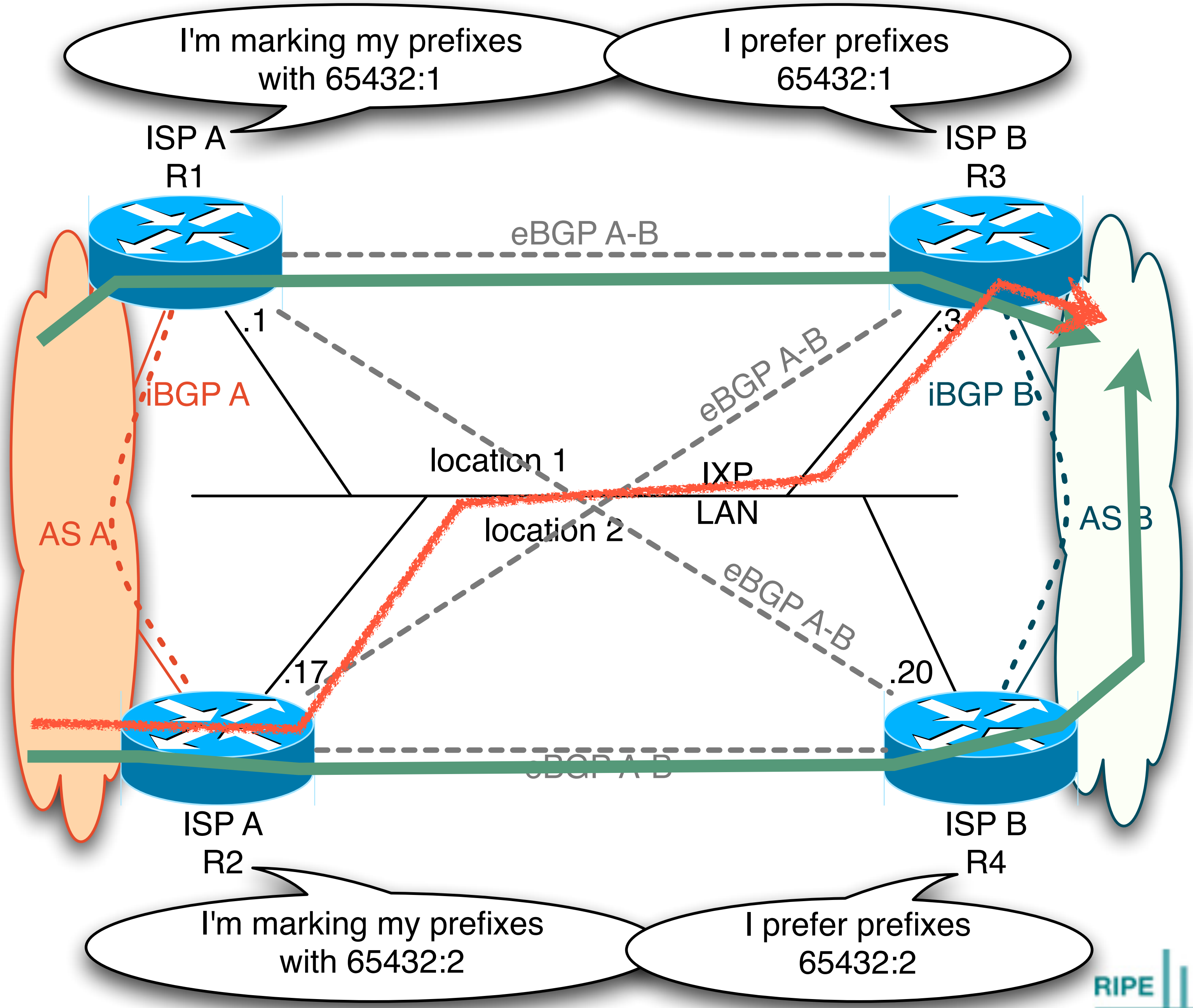


✔ preffered path

✗ avoid this path

Examples: members on both locations

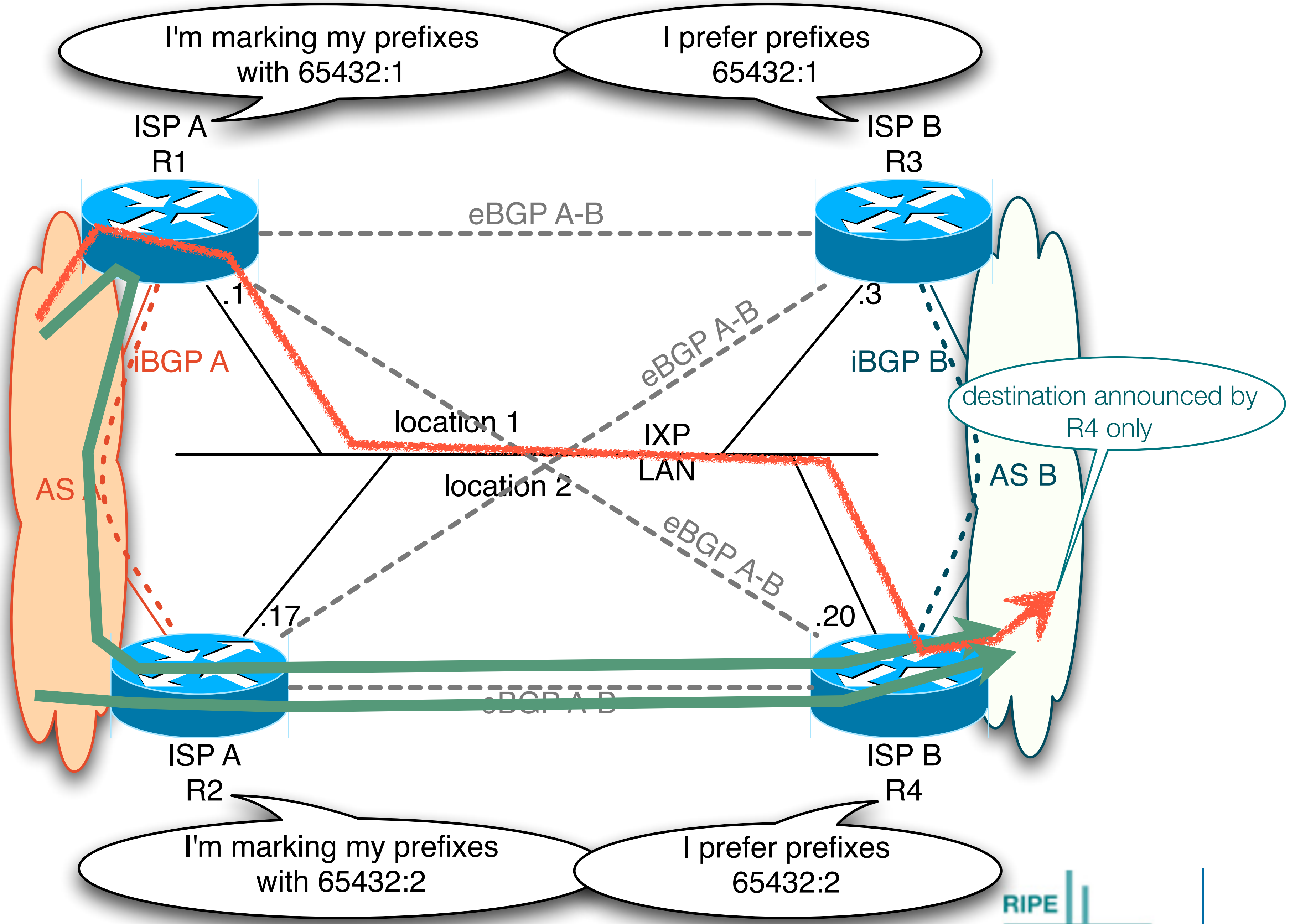
- prefixes are marked according to the location where they are being announced
- adjusting the metric
- next-hop self in iBGP


I'm marking my prefixes with 65432:1
I prefer prefixes 65432:1
ISP A R1
ISP B R3
eBGP A-B
.1
.3
iBGP A
eBGP A-B
iBGP B
location 1
IXP LAN
location 2
AS A
AS B
eBGP A-B
.17
.20
eBGP A-B
ISP A R2
ISP B R4
I'm marking my prefixes with 65432:2
I prefer prefixes 65432:2

✔ preffered path
✗ avoid this path

Examples: members on both locations

- prefixes are marked according to the location where they are being announced
- adjusting the metric
- next-hop self in iBGP


I'm marking my prefixes with 65432:1
I prefer prefixes 65432:1
ISP A R1
ISP B R3
eBGP A-B
.1
.3
iBGP A
eBGP A-B
iBGP B
destination announced by R4 only
location 1
IXP LAN
location 2
AS
AS B
eBGP A-B
.17
.20
eBGP A-B
ISP A R2
ISP B R4
I'm marking my prefixes with 65432:2
I prefer prefixes 65432:2

✔ preffered path

✗ avoid this path

Examples: localization

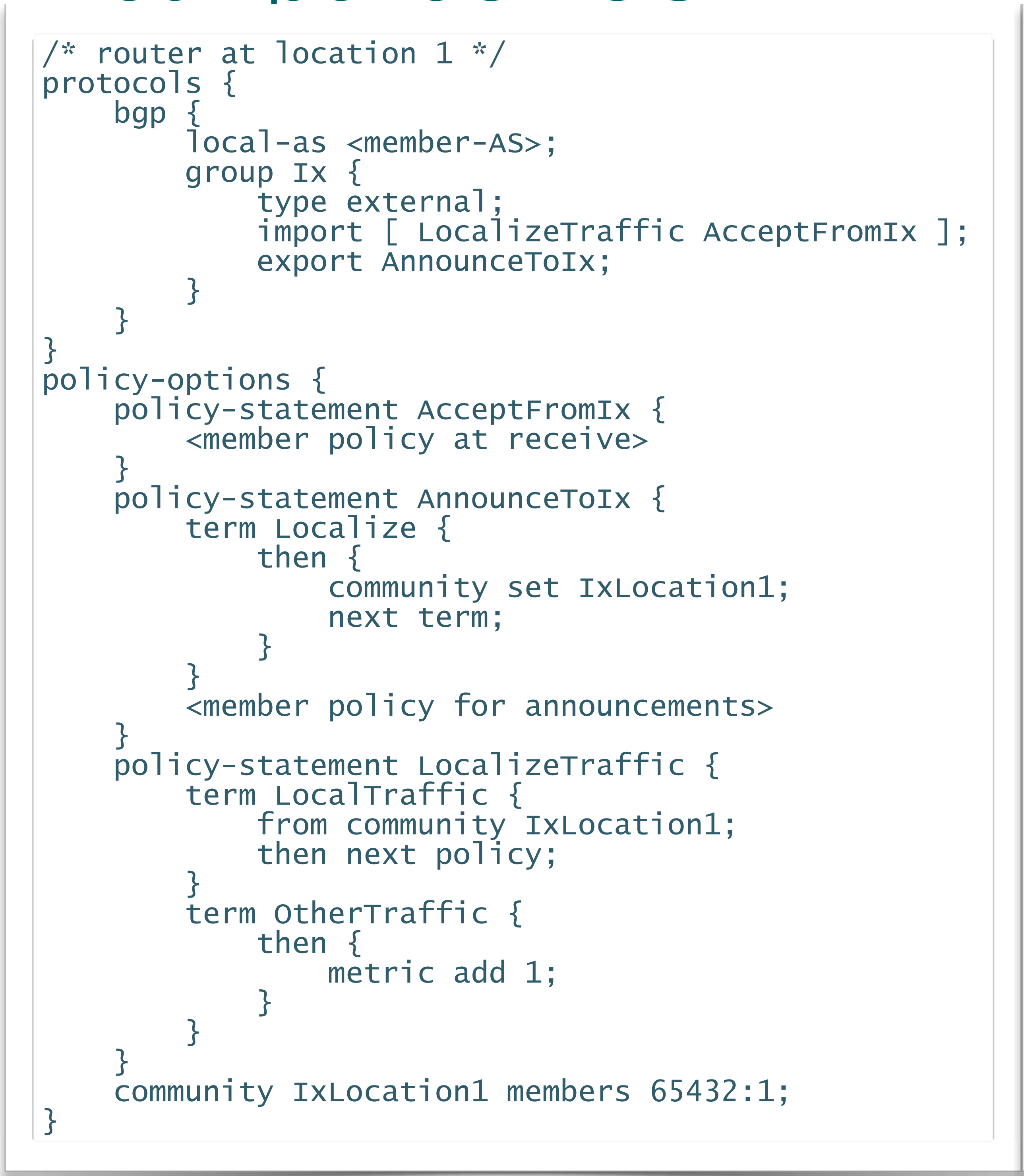
- Cisco IOS

```
! router R3 at location 1
ip community-list 61 permit 65432:1
!
route-map AnnounceToIX permit 10
 set community 65432:1
!
route-map AcceptFromIX permit 10
 ! this location
 match community 61
route-map AcceptFromIX permit 20
 ! other location - worse metric
 set metric +1
!
router bgp <member-AS>
 template peer-policy IX
  route-map AcceptFromIX in
  route-map AnnounceToIX out
  next-hop-self
  send-community
!
 address-family ipv4|6
 neighbor <R1> inherit peer-policy IX
 neighbor <R2> inherit peer-policy IX
!
```

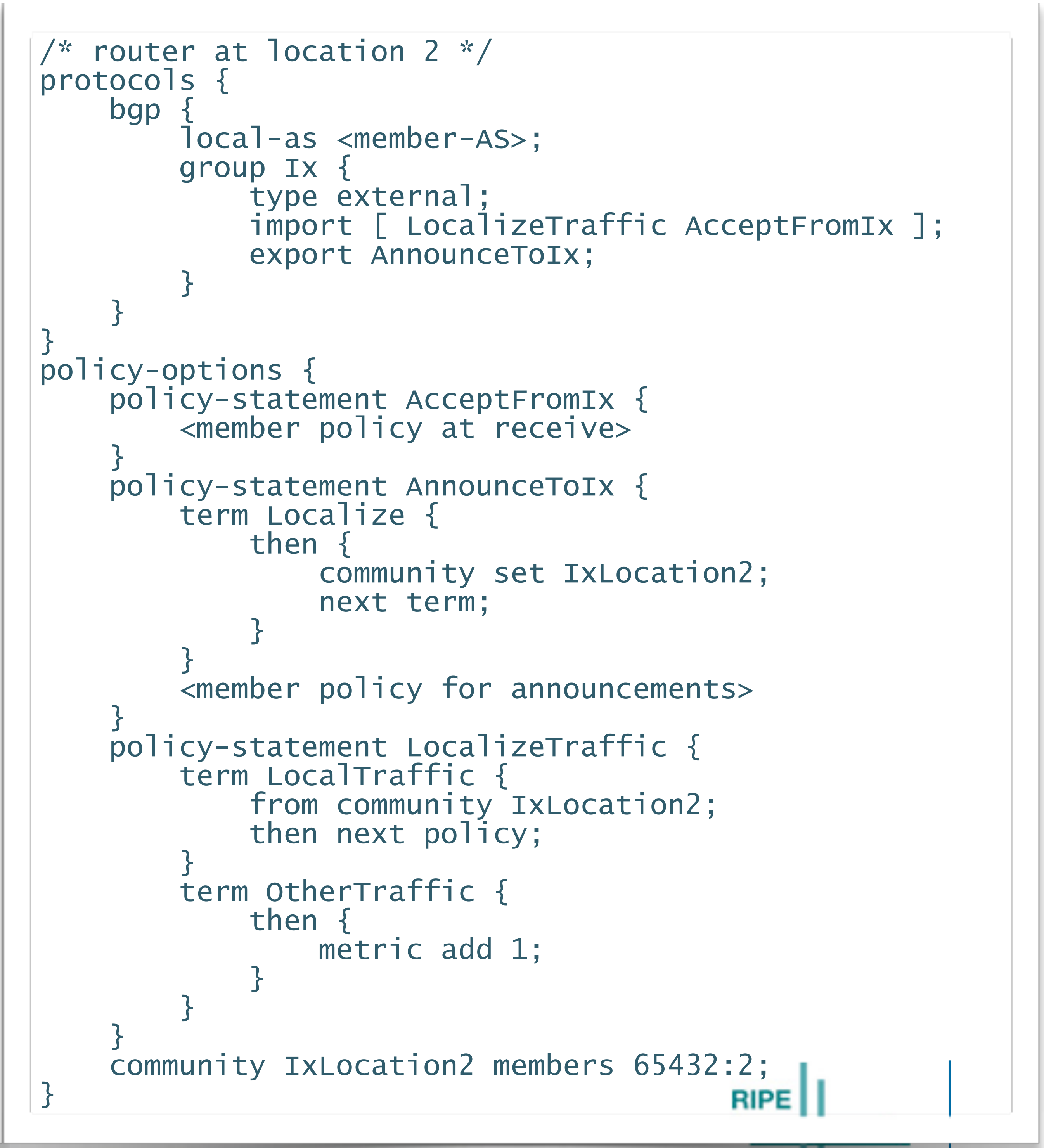
```
! router R4 at location 2
ip community-list 62 permit 65432:2
!
route-map AnnounceToIX permit 10
 set community 65432:2
!
route-map AcceptFromIX permit 10
 ! this location
 match community 62
route-map AcceptFromIX permit 20
 ! other location - worse metric
 set metric +1
!
router bgp <member-AS>
 template peer-policy IX
  route-map AcceptFromIX in
  route-map AnnounceToIX out
  next-hop-self
  send-community
!
 address-family ipv4|6
 neighbor <R1> inherit peer-policy IX
 neighbor <R2> inherit peer-policy IX
!
```

Examples: localization

- Juniper JUNOS



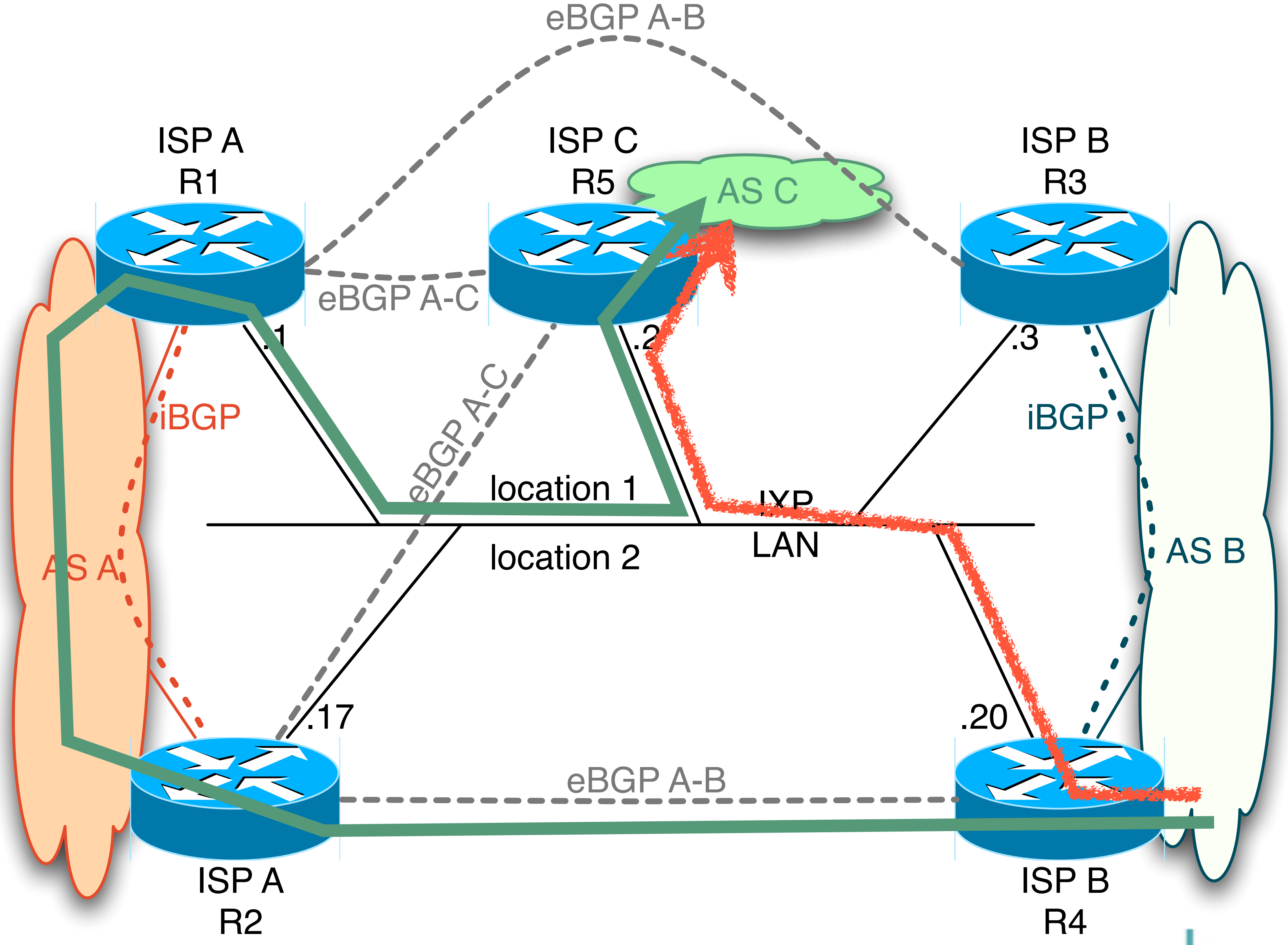
```
/* router at location 1 */
protocols {
    bgp {
        local-as <member-AS>;
        group Ix {
            type external;
            import [ LocalizeTraffic AcceptFromIx ];
            export AnnounceToIx;
        }
    }
}
policy-options {
    policy-statement AcceptFromIx {
        <member policy at receive>
    }
    policy-statement AnnounceToIx {
        term Localize {
            then {
                community set IxLocation1;
                next term;
            }
        }
        <member policy for announcements>
    }
    policy-statement LocalizeTraffic {
        term LocalTraffic {
            from community IxLocation1;
            then next policy;
        }
        term OtherTraffic {
            then {
                metric add 1;
            }
        }
    }
    community IxLocation1 members 65432:1;
}
```



```
/* router at location 2 */
protocols {
    bgp {
        local-as <member-AS>;
        group Ix {
            type external;
            import [ LocalizeTraffic AcceptFromIx ];
            export AnnounceToIx;
        }
    }
}
policy-options {
    policy-statement AcceptFromIx {
        <member policy at receive>
    }
    policy-statement AnnounceToIx {
        term Localize {
            then {
                community set IxLocation2;
                next term;
            }
        }
        <member policy for announcements>
    }
    policy-statement LocalizeTraffic {
        term LocalTraffic {
            from community IxLocation2;
            then next policy;
        }
        term OtherTraffic {
            then {
                metric add 1;
            }
        }
    }
    community IxLocation2 members 65432:2;
}
```

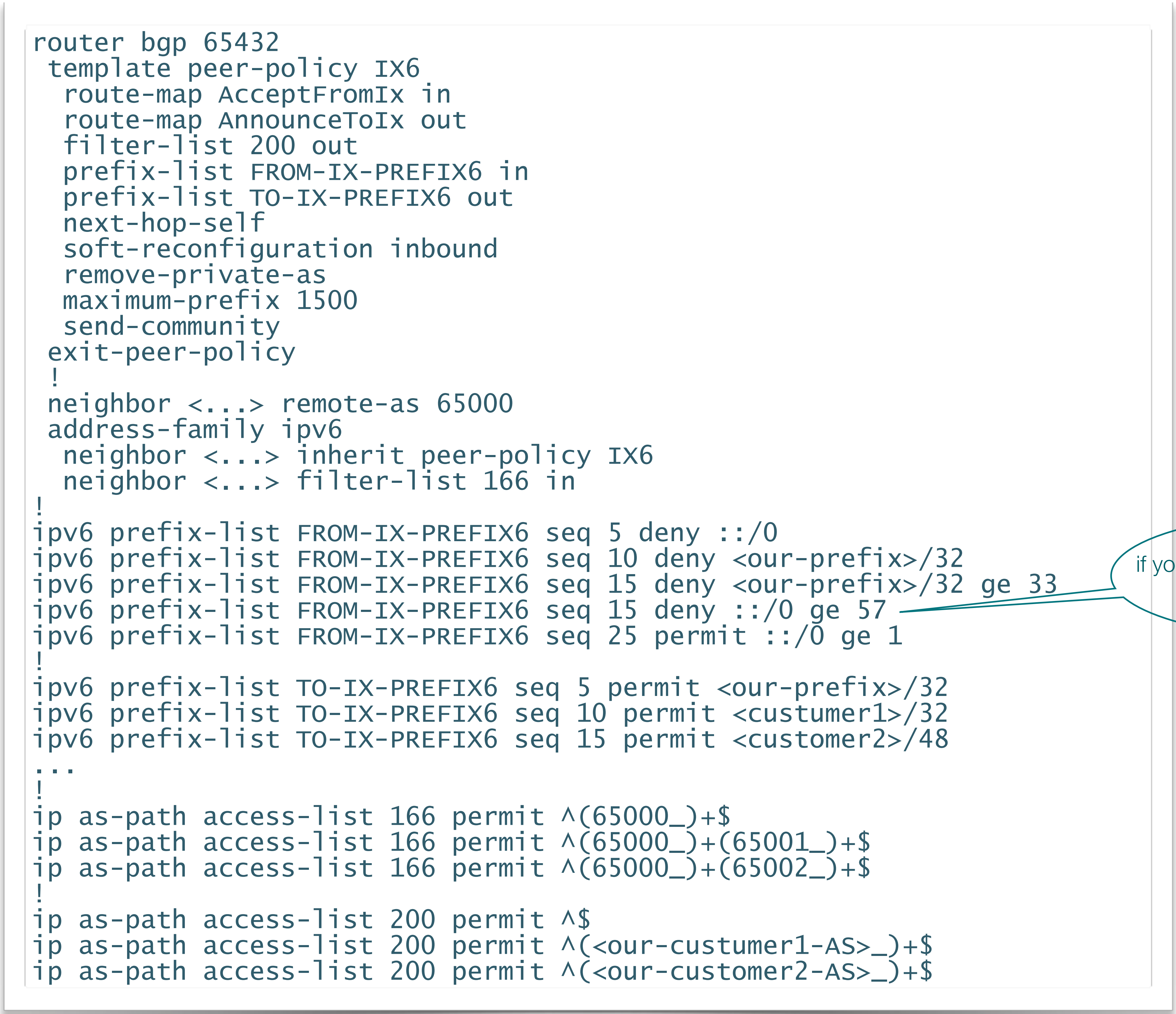
Examples: next-hop self, no redirects

- A wants to get the traffic from B and send it to C
- B should not send traffic directly to C
  - next-hop self in eBGP also
  - no ICMP redirects



✔ preffered path
✗ avoid this path

Example: BGP filters



```
router bgp 65432
 template peer-policy IX6
  route-map AcceptFromIx in
  route-map AnnounceToIx out
  filter-list 200 out
  prefix-list FROM-IX-PREFIX6 in
  prefix-list TO-IX-PREFIX6 out
  next-hop-self
  soft-reconfiguration inbound
  remove-private-as
  maximum-prefix 1500
  send-community
 exit-peer-policy
 !
 neighbor <...> remote-as 65000
 address-family ipv6
  neighbor <...> inherit peer-policy IX6
  neighbor <...> filter-list 166 in
!
ipv6 prefix-list FROM-IX-PREFIX6 seq 5 deny ::/0
ipv6 prefix-list FROM-IX-PREFIX6 seq 10 deny <our-prefix>/32
ipv6 prefix-list FROM-IX-PREFIX6 seq 15 deny <our-prefix>/32 ge 33
ipv6 prefix-list FROM-IX-PREFIX6 seq 15 deny ::/0 ge 57
ipv6 prefix-list FROM-IX-PREFIX6 seq 25 permit ::/0 ge 1
!
ipv6 prefix-list TO-IX-PREFIX6 seq 5 permit <our-prefix>/32
ipv6 prefix-list TO-IX-PREFIX6 seq 10 permit <custumer1>/32
ipv6 prefix-list TO-IX-PREFIX6 seq 15 permit <customer2>/48
...
!
ip as-path access-list 166 permit ^(65000_)+$
ip as-path access-list 166 permit ^(65000_)+(65001_)+$
ip as-path access-list 166 permit ^(65000_)+(65002_)+$
!
ip as-path access-list 200 permit ^$
ip as-path access-list 200 permit ^(<our-custumer1-AS>_)+$
ip as-path access-list 200 permit ^(<our-customer2-AS>_)+$
```

if you decide to block small prefixes, for example, less than /56

Goodies

- looking-glass router
- route-server (reflector)
- graphs
  - public
  - or members only
  - or private
- meetings :-)