IIJ

Internet Initiative Japan

BGP is Chattier than we Think

In Fact, it can be Ugly

APNIC / Seoul / 2003.08.21

Tim Griffin <tim.griffin@intel.com>

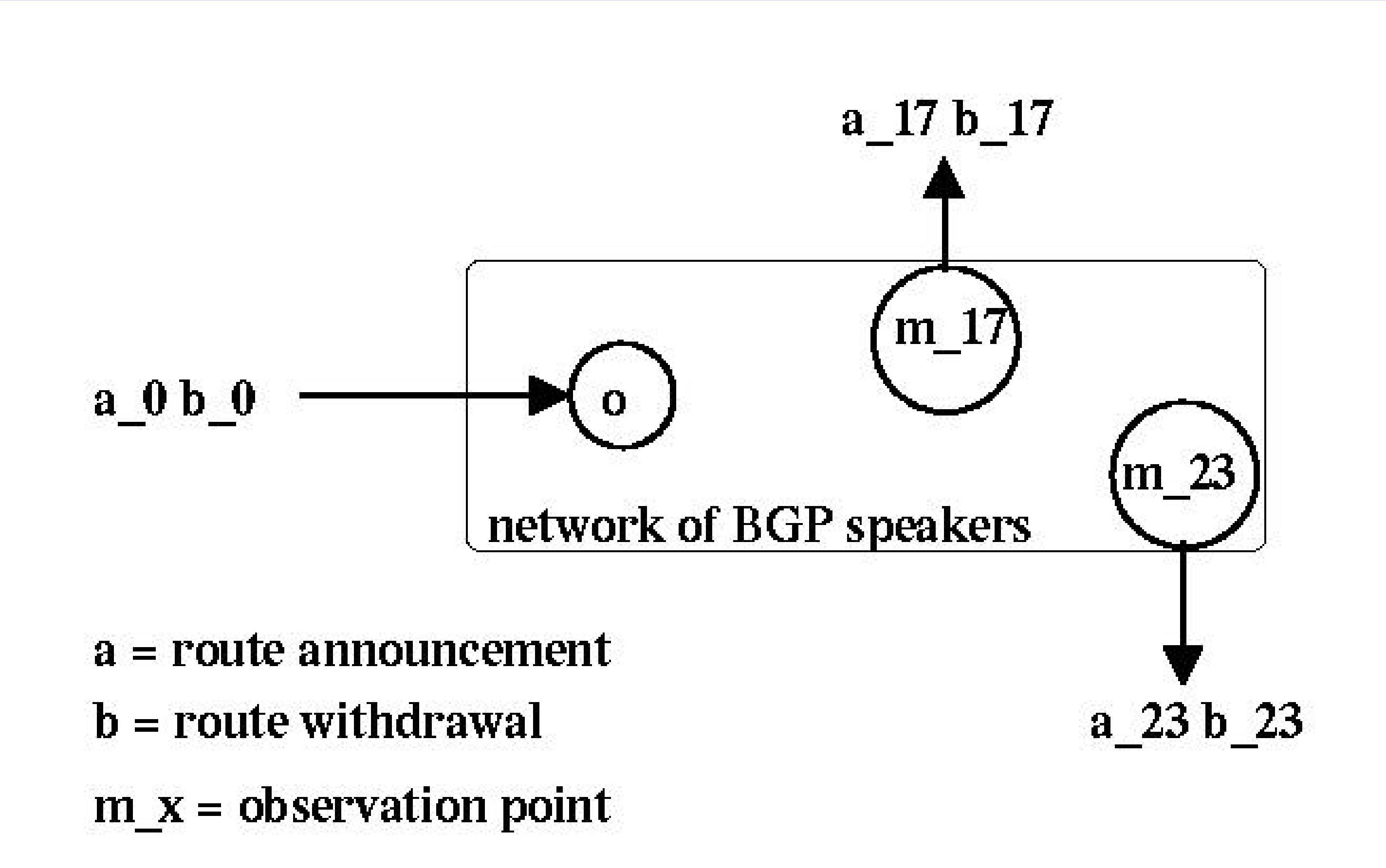
Randy Bush <randy@iij.com>

Z. Morley Mao <zmao@cs.berkeley.edu>

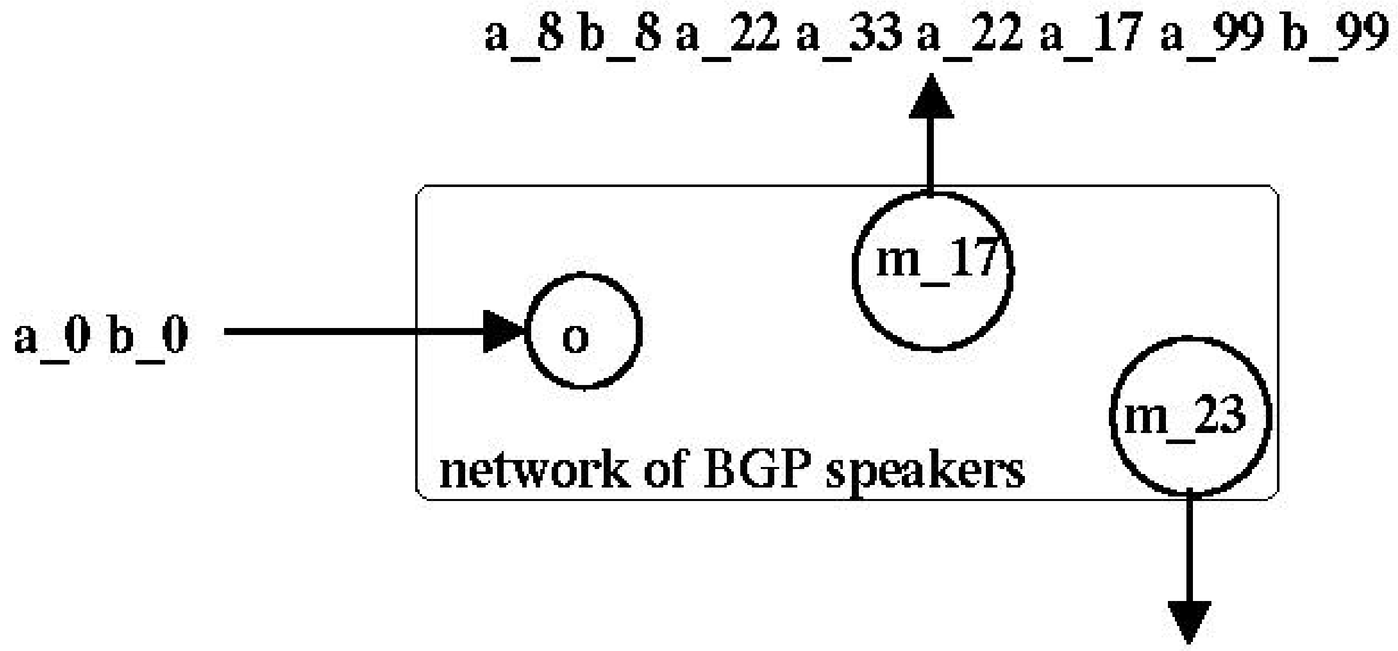
John Heasley <heas@shrubbery.net>

<http://psg.com/~randy/030821.apnic-bbgp.pdf>

The Naïve View



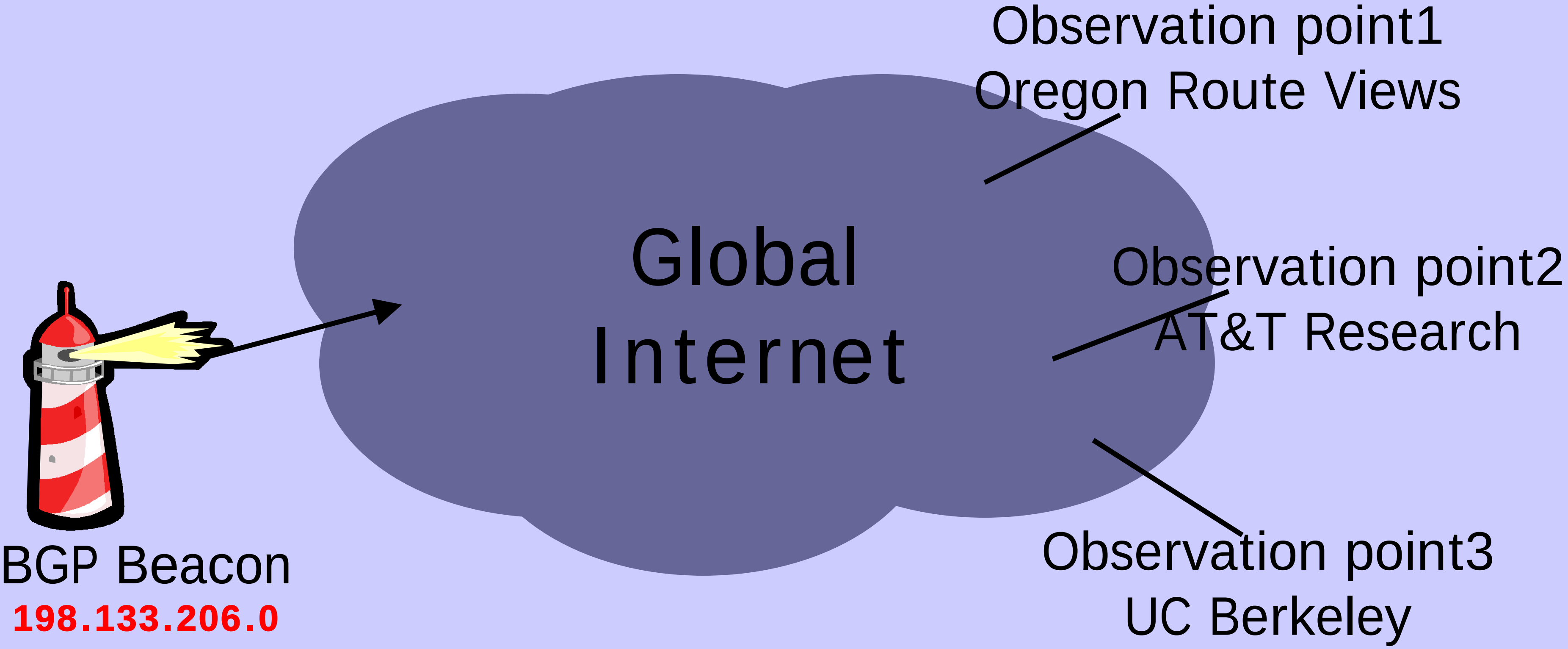
Reality


a_8 b_8 a_22 a_33 a_22 a_17 a_99 b_99
a_0 b_0
o
m_17
m_23
network of BGP speakers

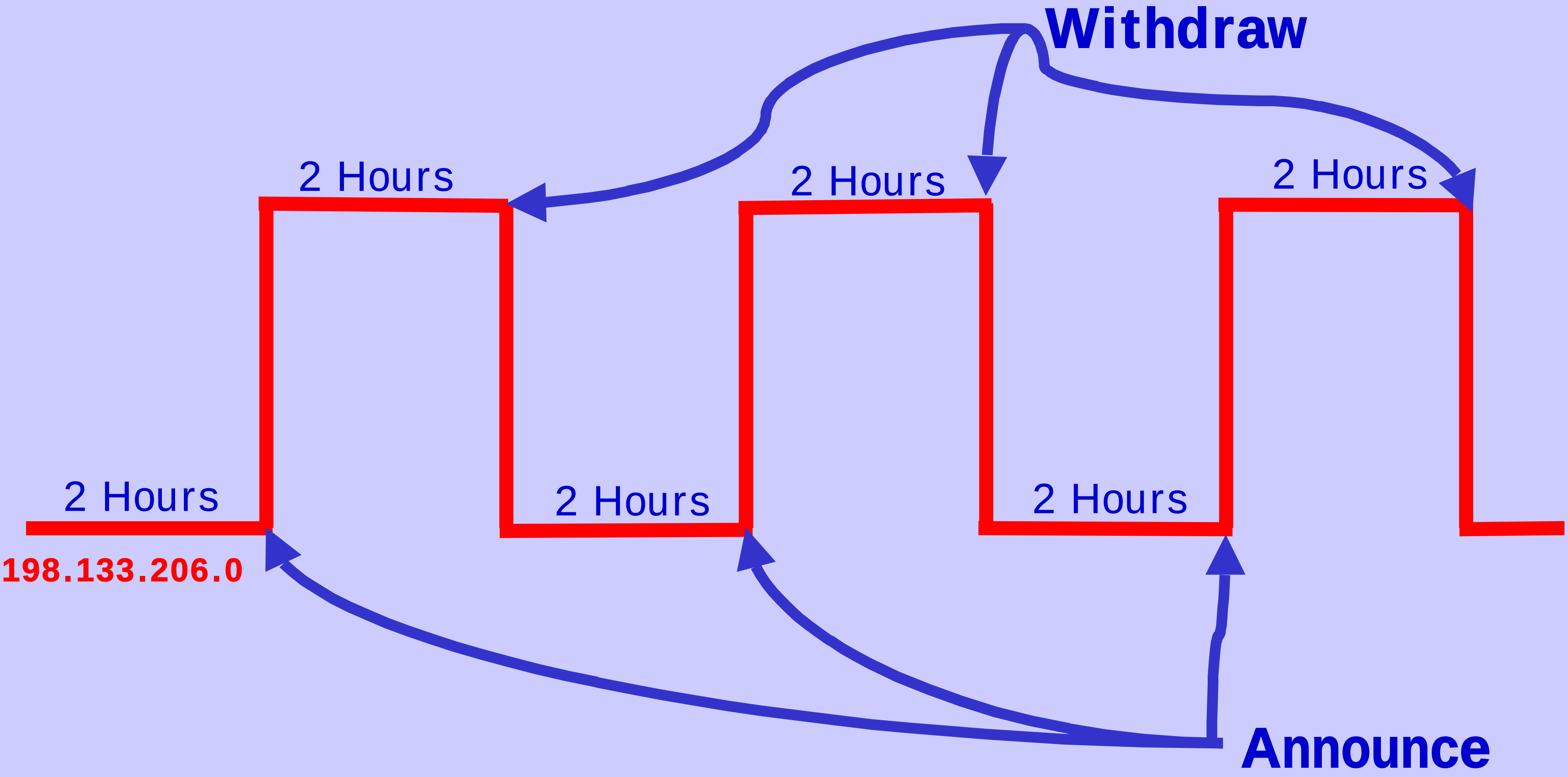
BGP Beacon

BGP Beacon:

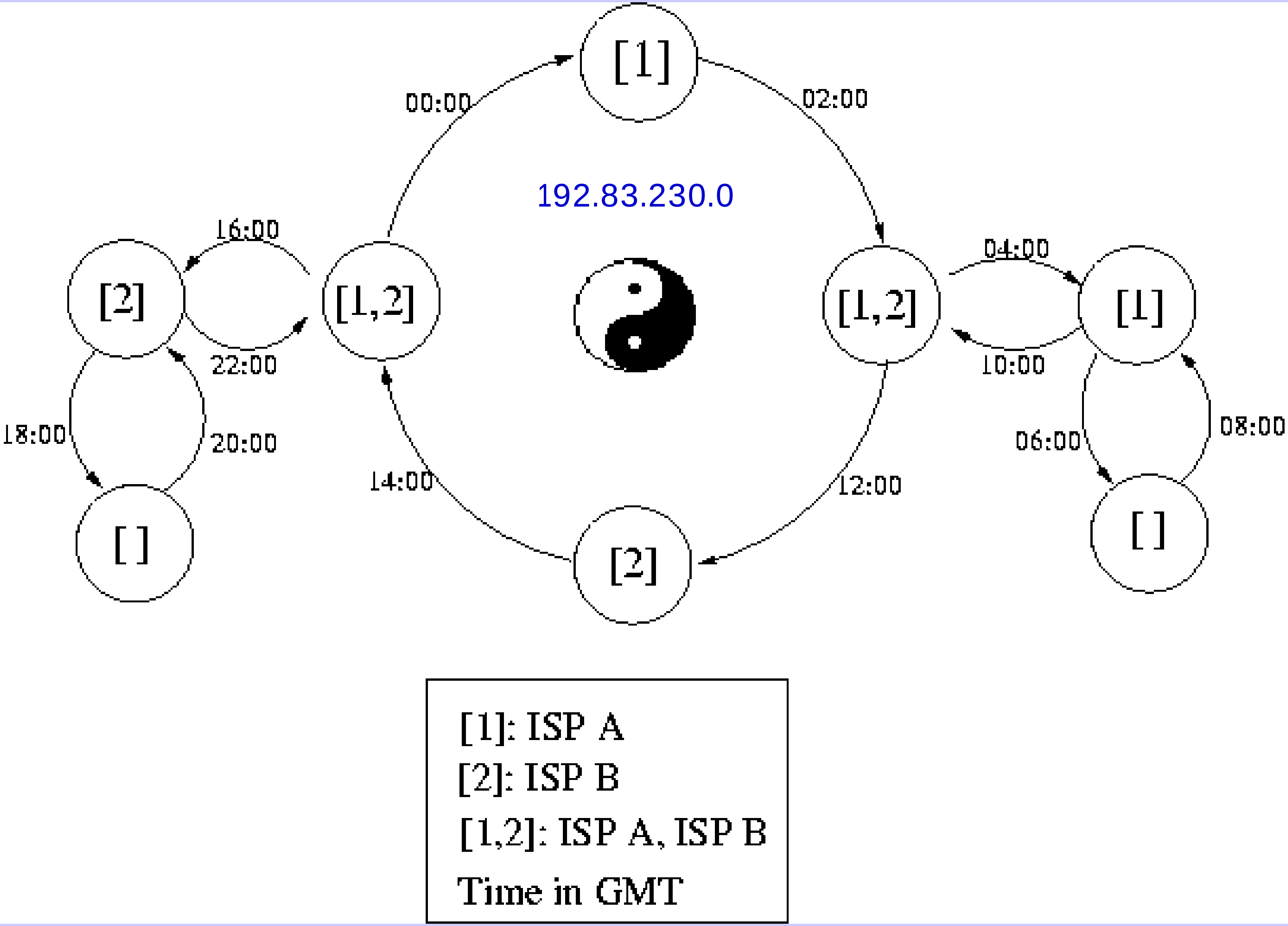
A prefix that is Announced and Withdrawn at **well-known** times



BGP Beacons
Announce & Withdraw



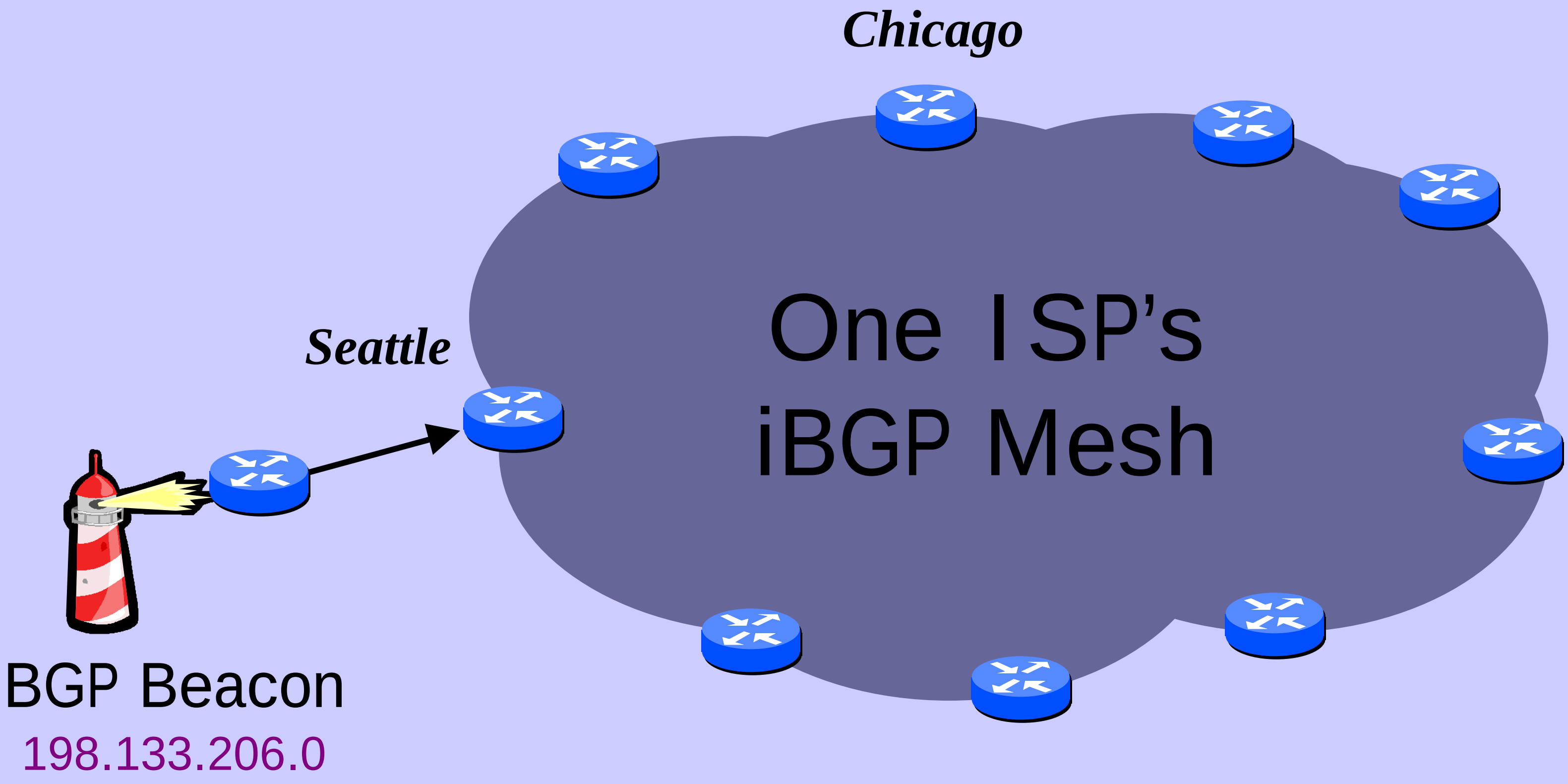
Multi-Homed Second Beacon



[1]: ISP A
[2]: ISP B
[1,2]: ISP A, ISP B
Time in GMT

Measurement within One ISP

- Measure at peering edges of a global ISP
- Archived (and anonymized)
- Multi-month study



The Sound of Discovery is not “Eureka!”.

It is “Oh $#@%$!!!”

Life Looks Simple in Seattle

```
2003-07-01:ANN-event, updateCnt=1, []->[ispA]
20:00:13 A asA 3130
```

Even if Beacon is Multi-Homed

```
2003-07-02:ANN-event, updateCnt=2, [asA,]->[asB]
00:00:31 W
00:00:31 A asB 3130
```

Chicago Sees More Complexity

```
2003-07-08:ANN-event, updateCnt=4, []->[ispA]
20:00:24 A asA 3130  34
20:00:24 A asA 3130  33
20:00:28 A asA 3130  34
20:00:34 A asA 3130  33
```

Route
Oscillation

Much More!

```
2003-06-11:ANN-event, updateCnt=41, []->[ispA,ispB].
13:00:08 A asA asB 3130  33
13:00:10 A asA asB 3130  30
13:00:17 A asA asB 3130   1
13:00:18 A asA 3130      34
13:00:18 A asA 3130      33
13:00:18 W
13:00:19 A asA asB 3130  30
13:00:19 A asA 3130      33
13:00:19 A asA 3130      34
13:00:19 A asA 3130      33
13:00:19 A asA asB 3130  37
13:00:19 A asA asB 3130  30
13:00:19 A asA 3130      33
13:00:19 A asA 3130      34
13:00:19 A asA 3130      33
13:00:19 A asA 3130      34
13:00:19 A asA 3130      33
13:00:19 A asA asB 3130  37
13:00:19 W
13:00:19 A asA 3130      34
13:00:19 A asA 3130      33
13:00:20 A asA 3130      34
13:00:20 A asA 3130      33
13:00:20 A asA asB 3130   1
13:00:20 A asA 3130      33
13:00:20 A asA 3130      34
13:00:22 A asA 3130      33
13:00:23 A asA 3130      34
13:00:24 A asA asB 3130   1
13:00:24 A asA 3130      34
13:00:24 A asA asB 3130  27
13:00:24 A asA asB 3130  42
13:00:24 A asA 3130      33
13:00:24 A asA 3130      34
13:00:24 A asA asB 3130  27
13:00:24 A asA asB 3130  30
13:00:24 A asA 3130      34
13:00:24 A asA asB 3130  27
13:00:25 A asA asB 3130  30
13:00:25 A asA 3130      34
13:00:26 A asA 3130      33
```

41 Events
39 Announcements
  2 Withdraws!

In 26 seconds (and that’s fast!)

And we don’t even charge extra

And the feature-rich vendors tell us that BGP is “Rock solid stable”

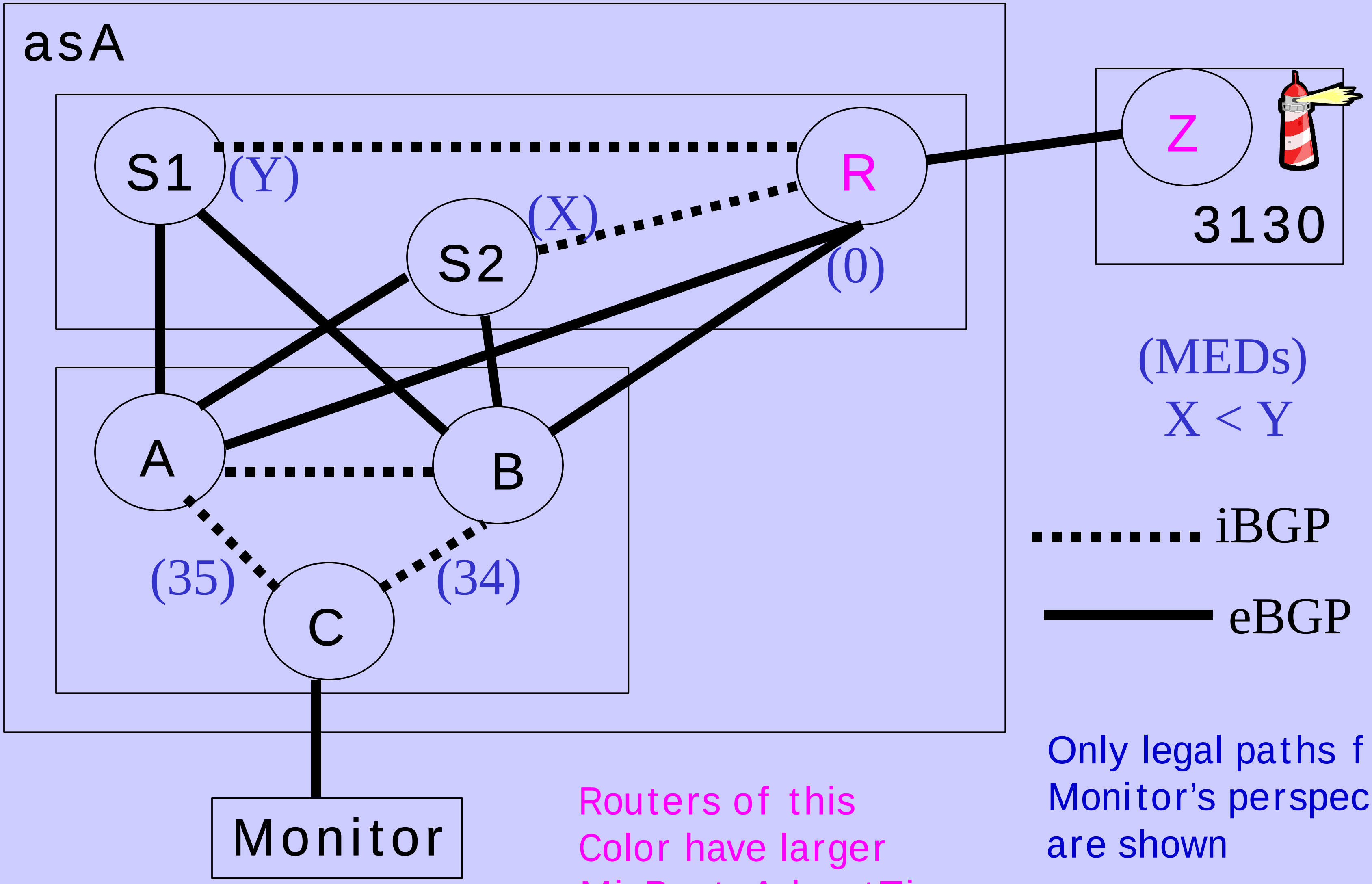
Why?

- BGP - Path Vector protocol (remember RIP?)
- Distributed Computation in Time and Delay
- Made worse by MinRouteAdvertTimer implementation differences between vendors
  - MRAI is the Delay before Propagation of a Route
  - 30 seconds is advised
  - Implementations vary, and some do zero
- Seattle is much nicer than Chicago :-)

Notes

- Examples are simplified for clarity
- Messages in transit or queued up are not shown
- MEDs, IGPs, ... are not always shown
- One possible sequence among many is explored --- the goal is to explain how some bad cases happen

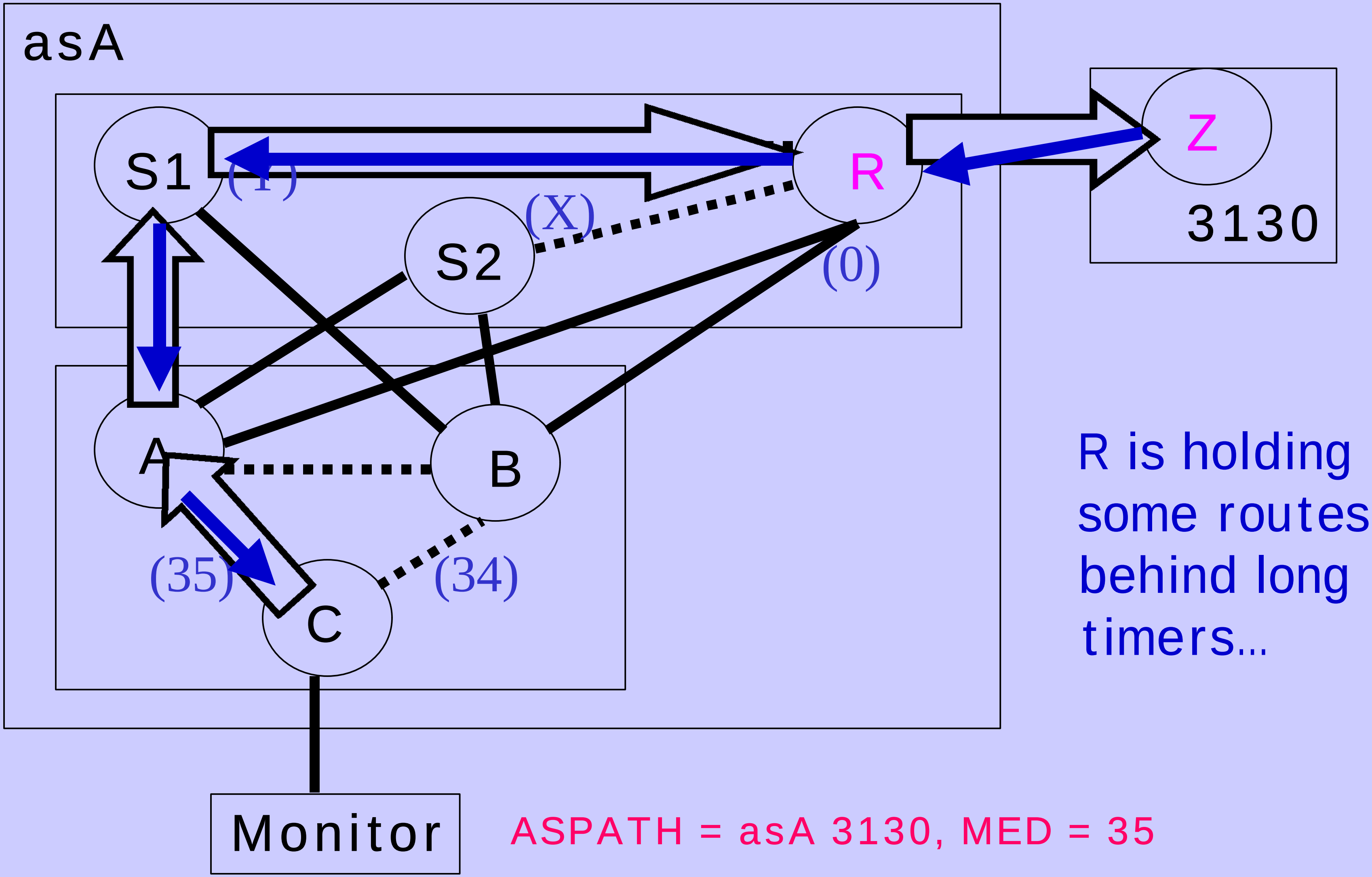
Simple Set-up 1



Routers of this Color have larger MinRouteAdvertTimer

Only legal paths from Monitor's perspective are shown

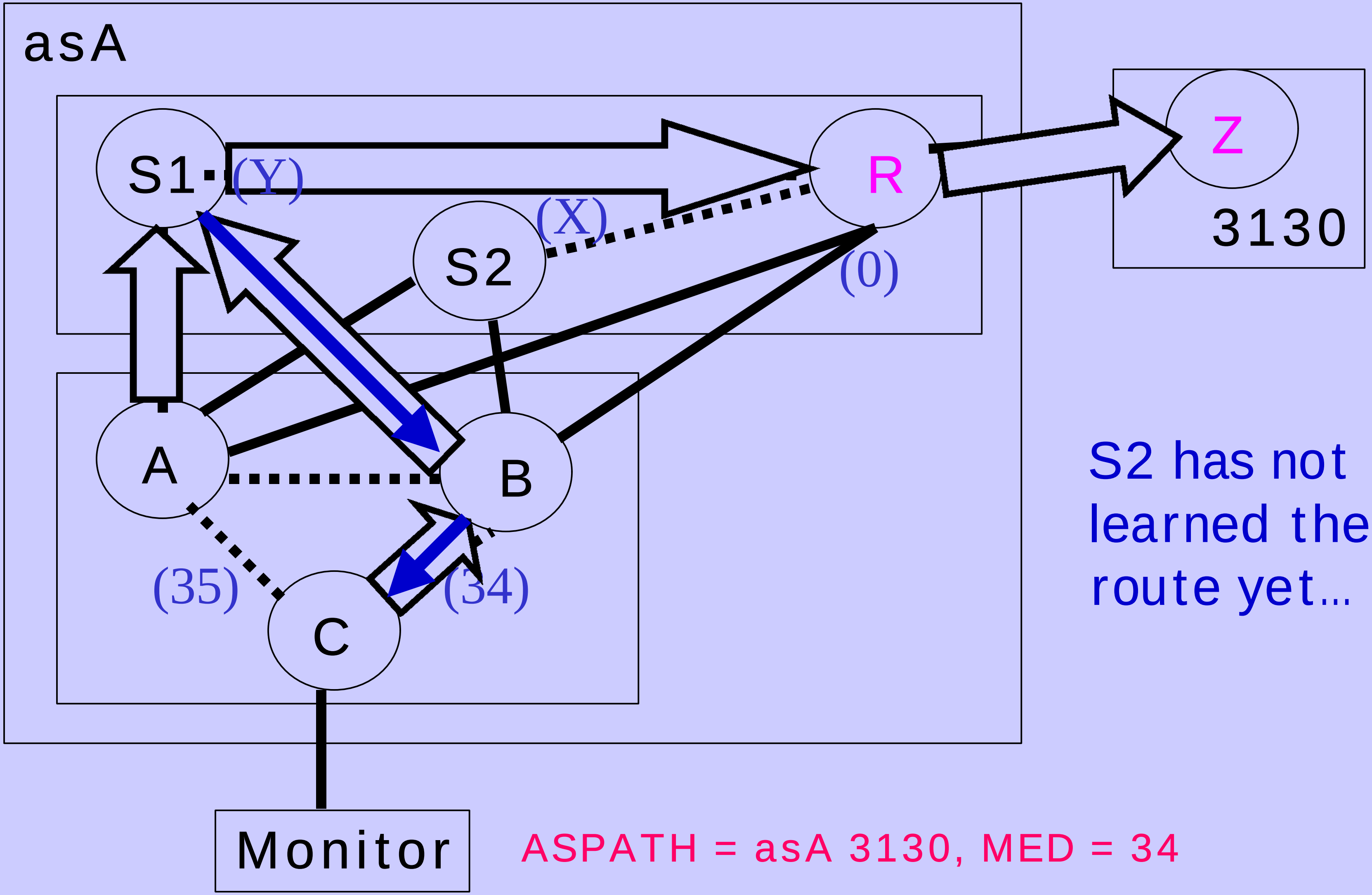
State 1



R is holding some routes behind long timers…

ASPATH = asA 3130, MED = 35

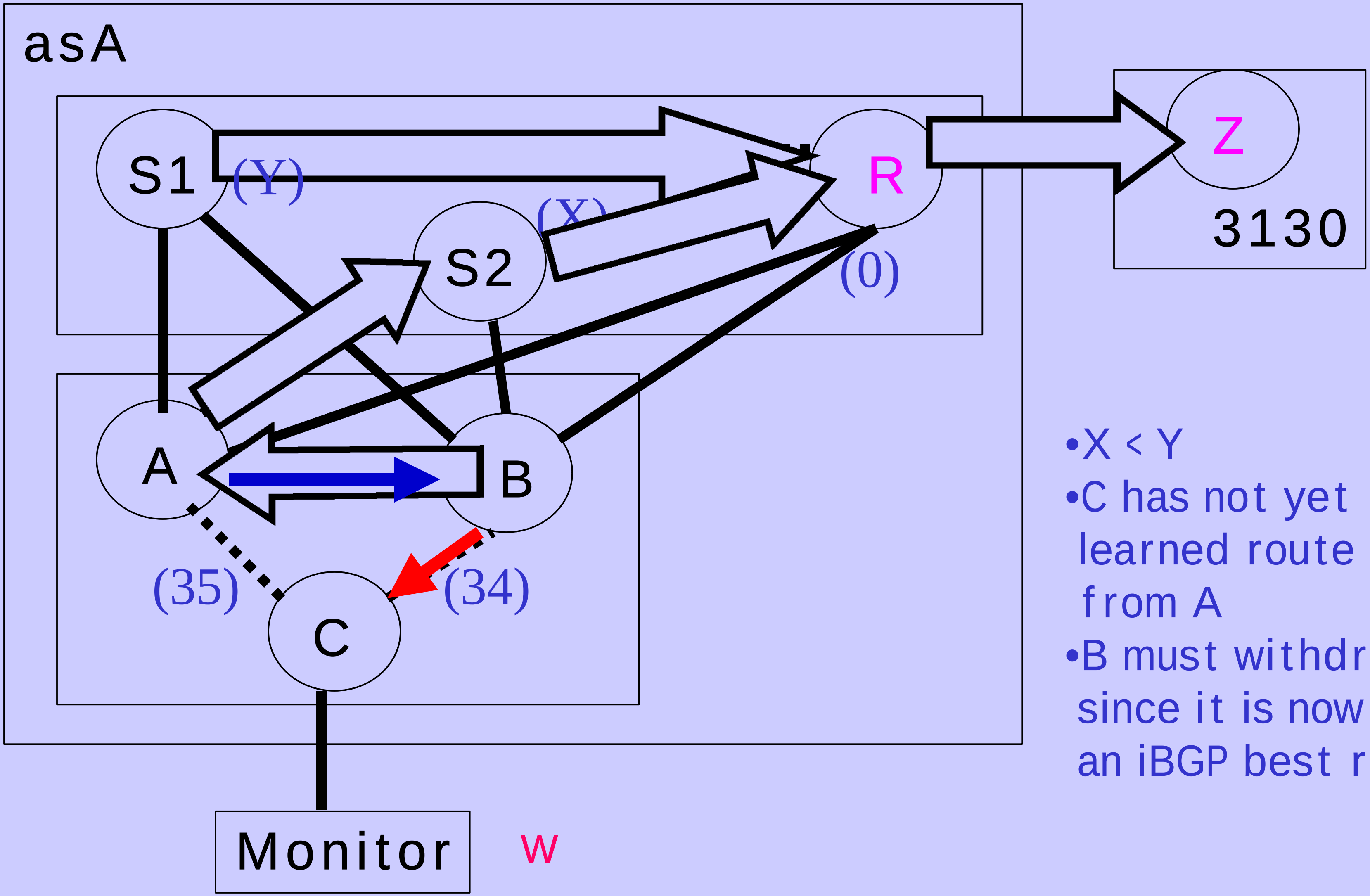
State 2



S2 has not learned the route yet…

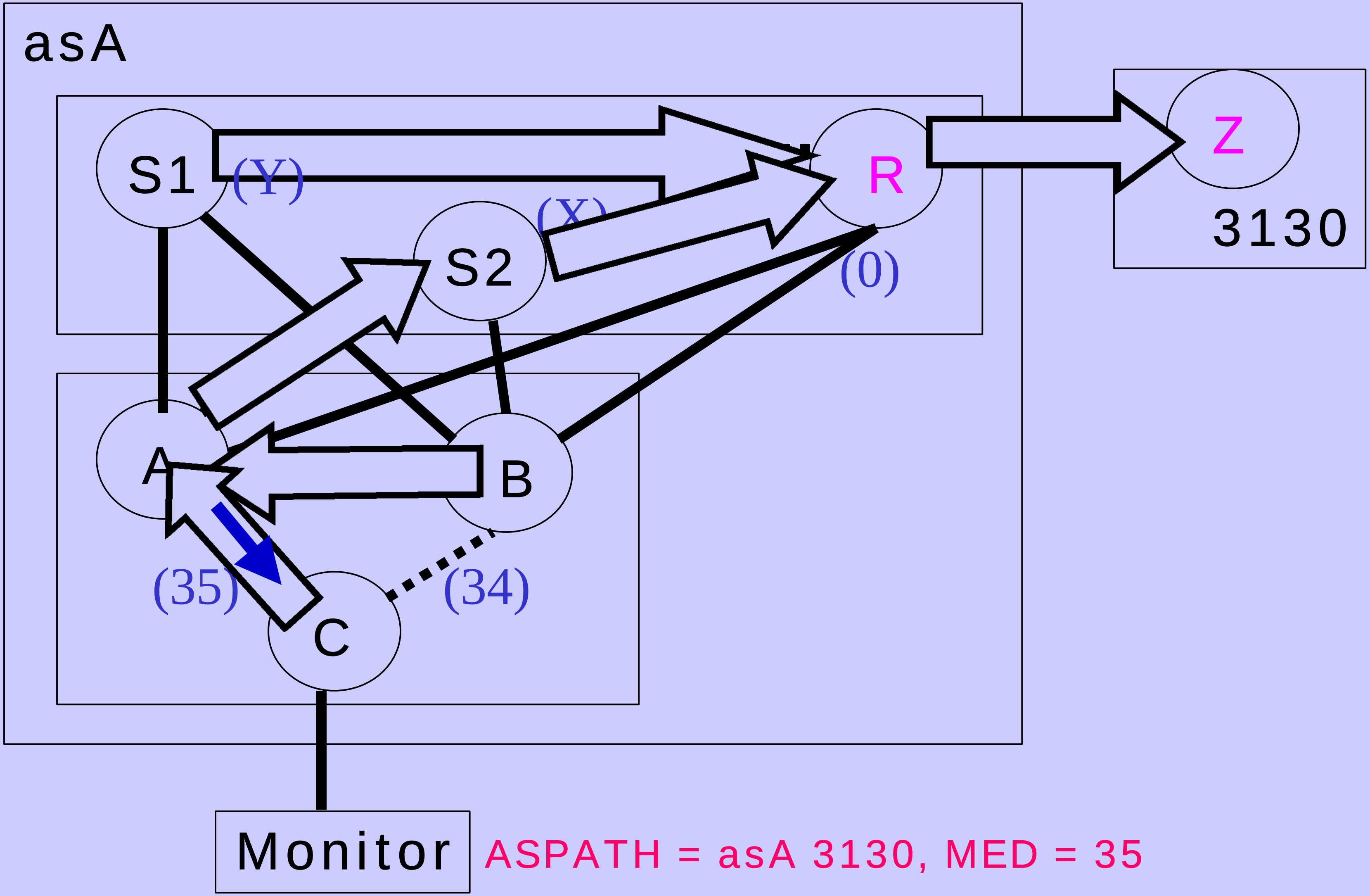
ASPATH = asA 3130, MED = 34

State 3



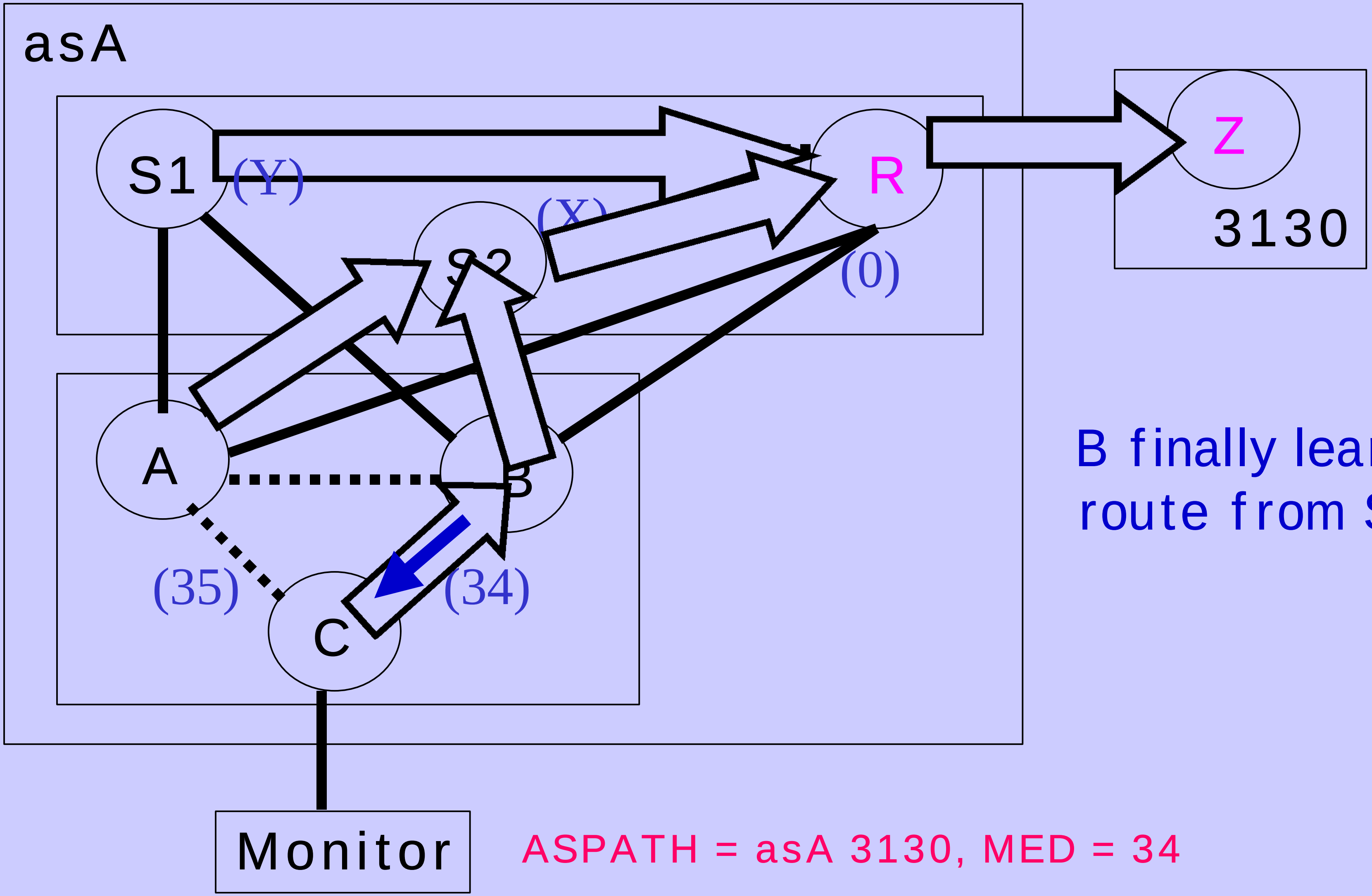
- X < Y
- C has not yet learned route from A
- B must withdraw, since it is now has an iBGP best route

State 4



ASPATH = asA 3130, MED = 35

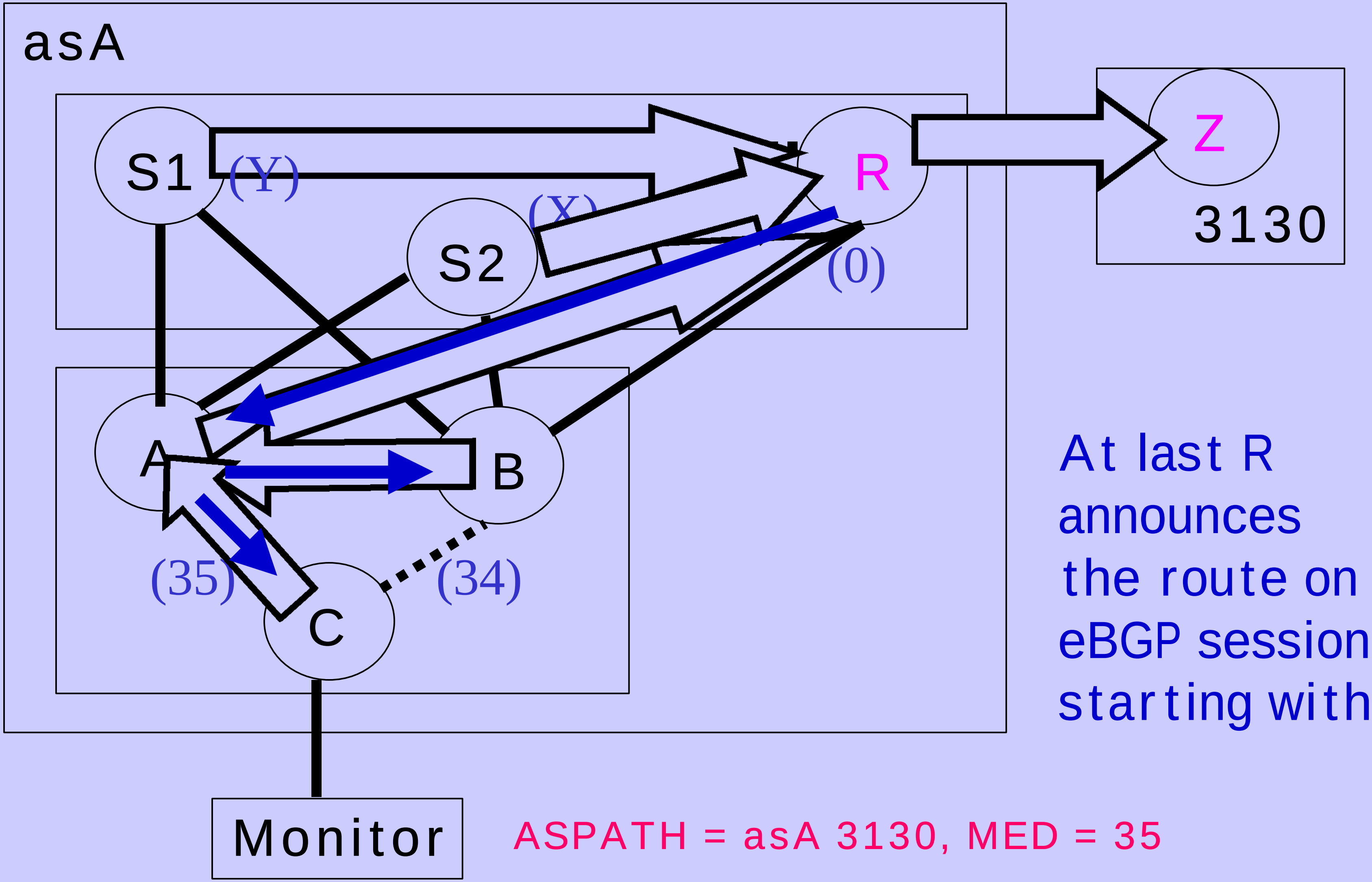
State 5



B finally learns route from S2

ASPATH = asA 3130, MED = 34

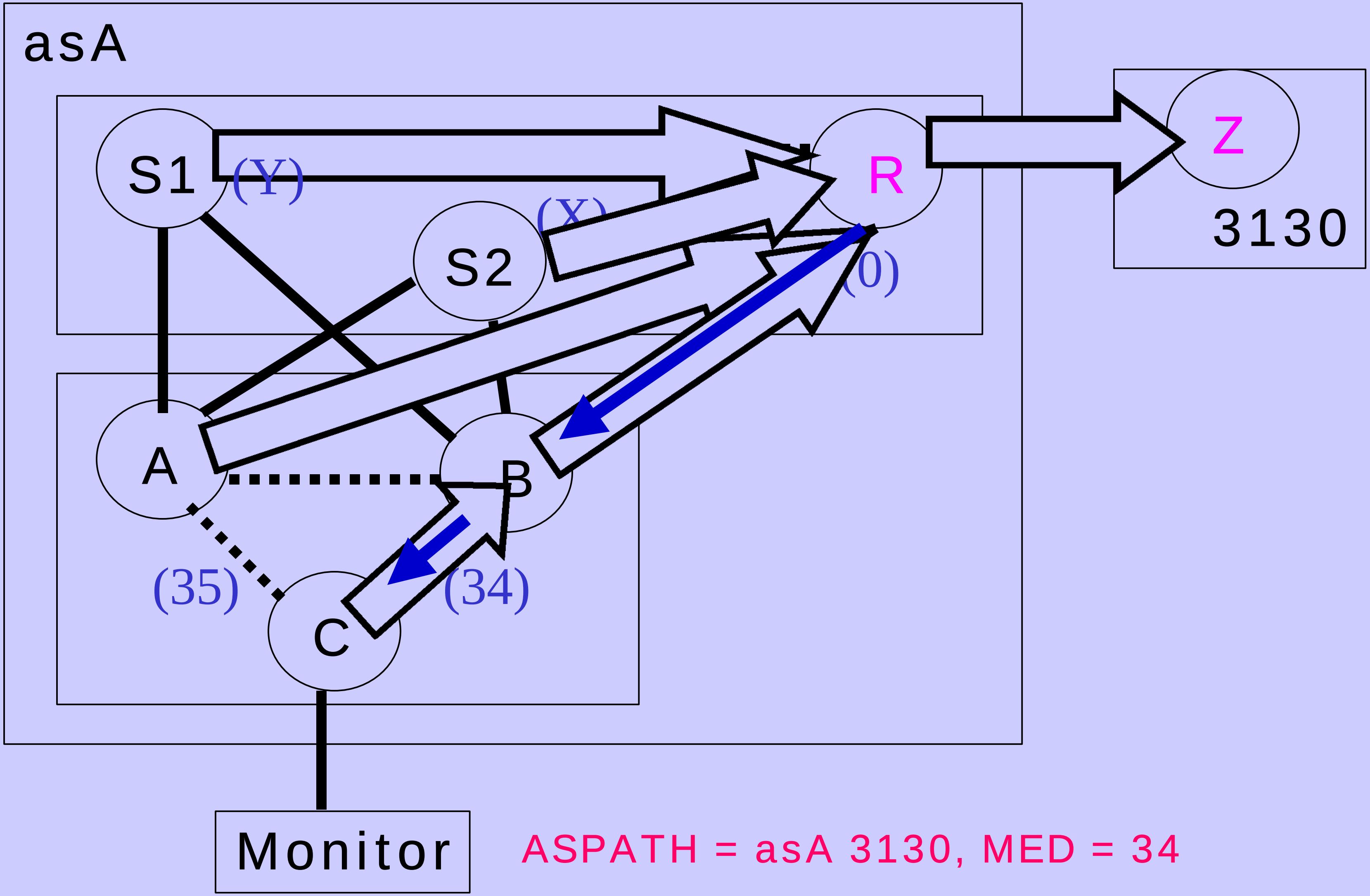
State 6



At last R announces the route on its eBGP sessions, starting with A

ASPATH = asA 3130, MED = 35

State 7



ASPATH = asA 3130, MED = 34

Signals Seen by the Monitor

ASPATH = asA 3130, MED = 35

ASPATH = asA 3130, MED = 34

W

ASPATH = asA 3130, MED = 35

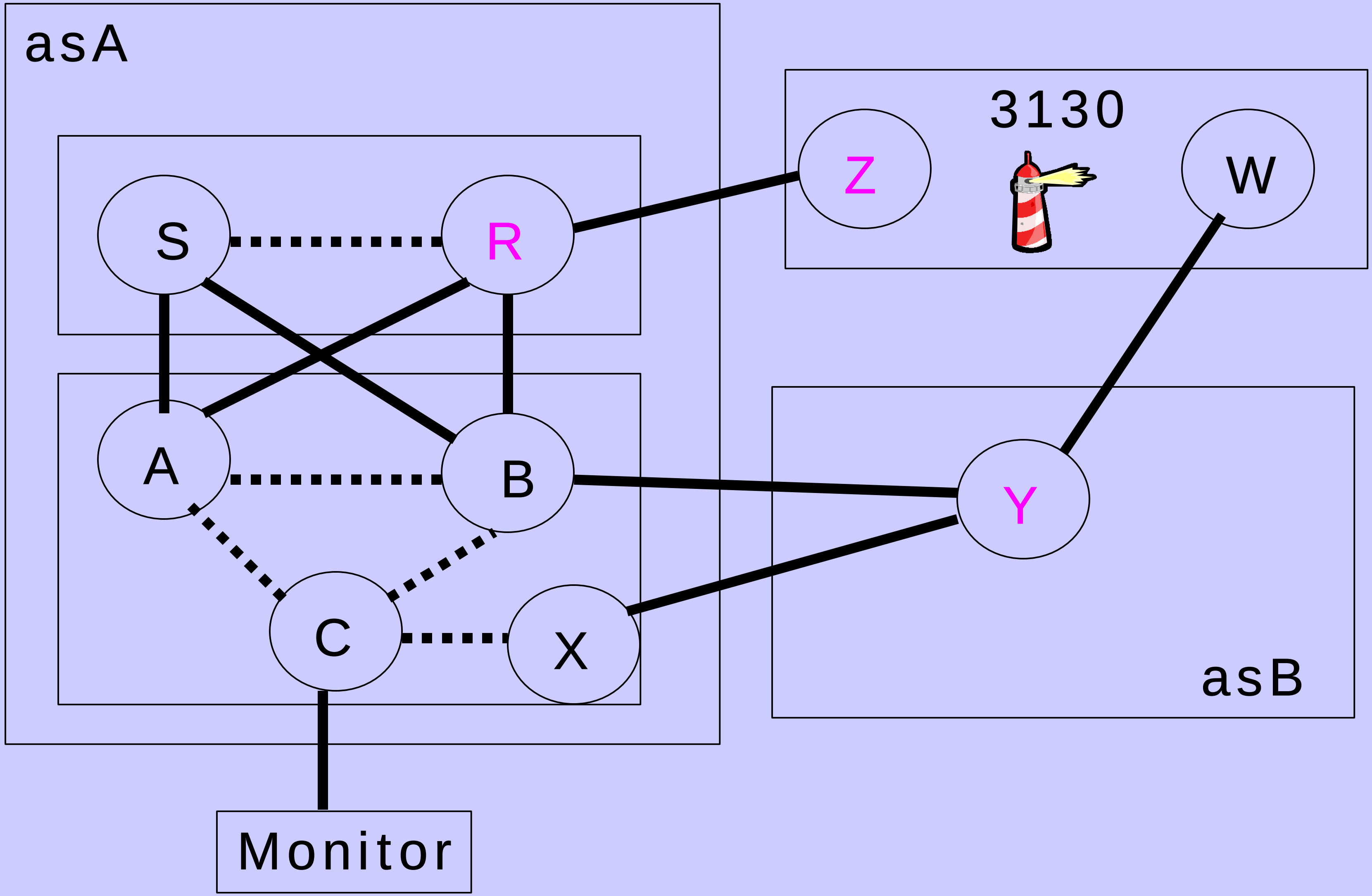
ASPATH = asA 3130, MED = 34

ASPATH = asA 3130, MED = 35

ASPATH = asA 3130, MED = 34

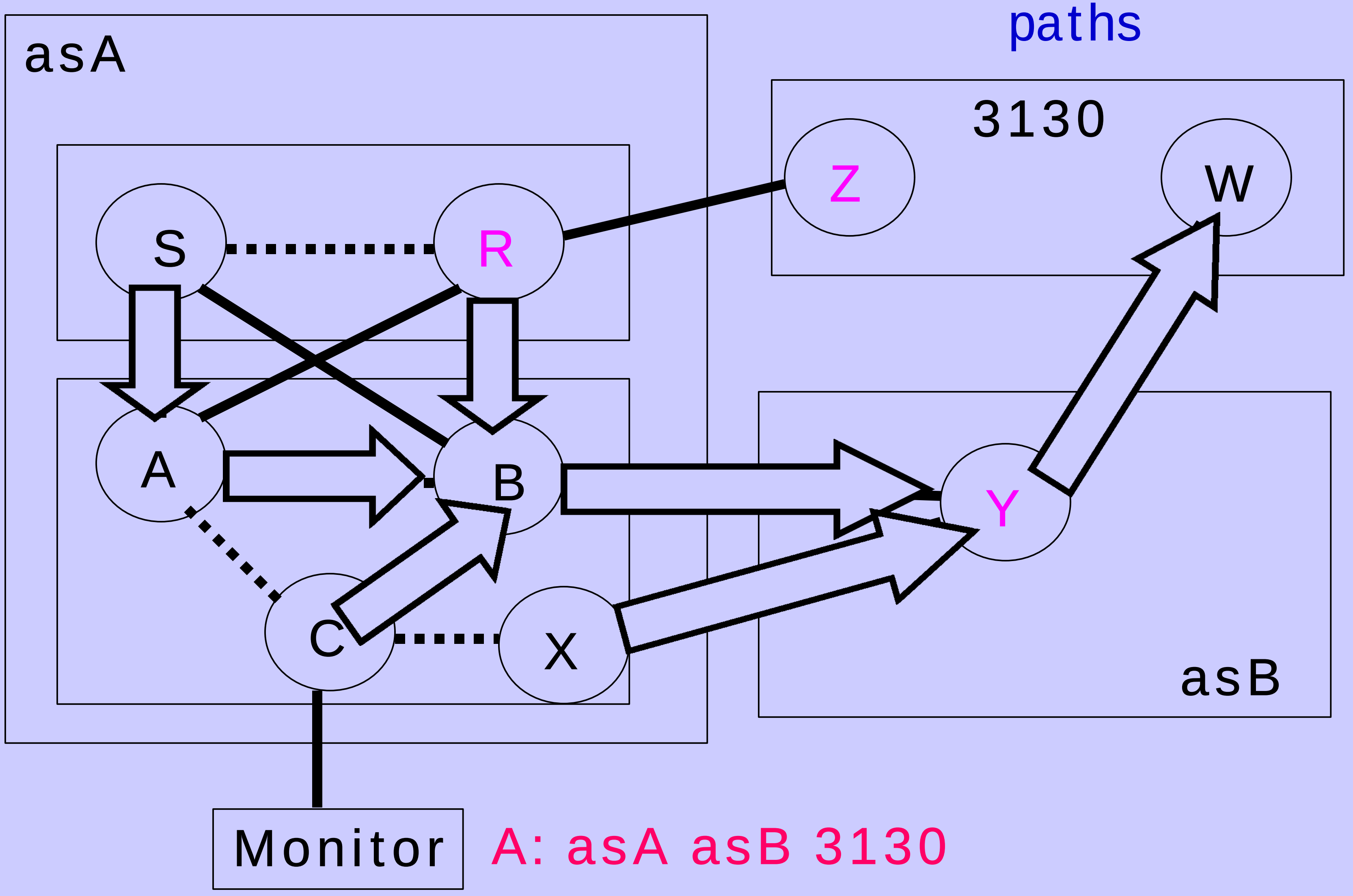
Conclusion : simple Announcements can be very noisy.

Simple Set up II - MultiHomed



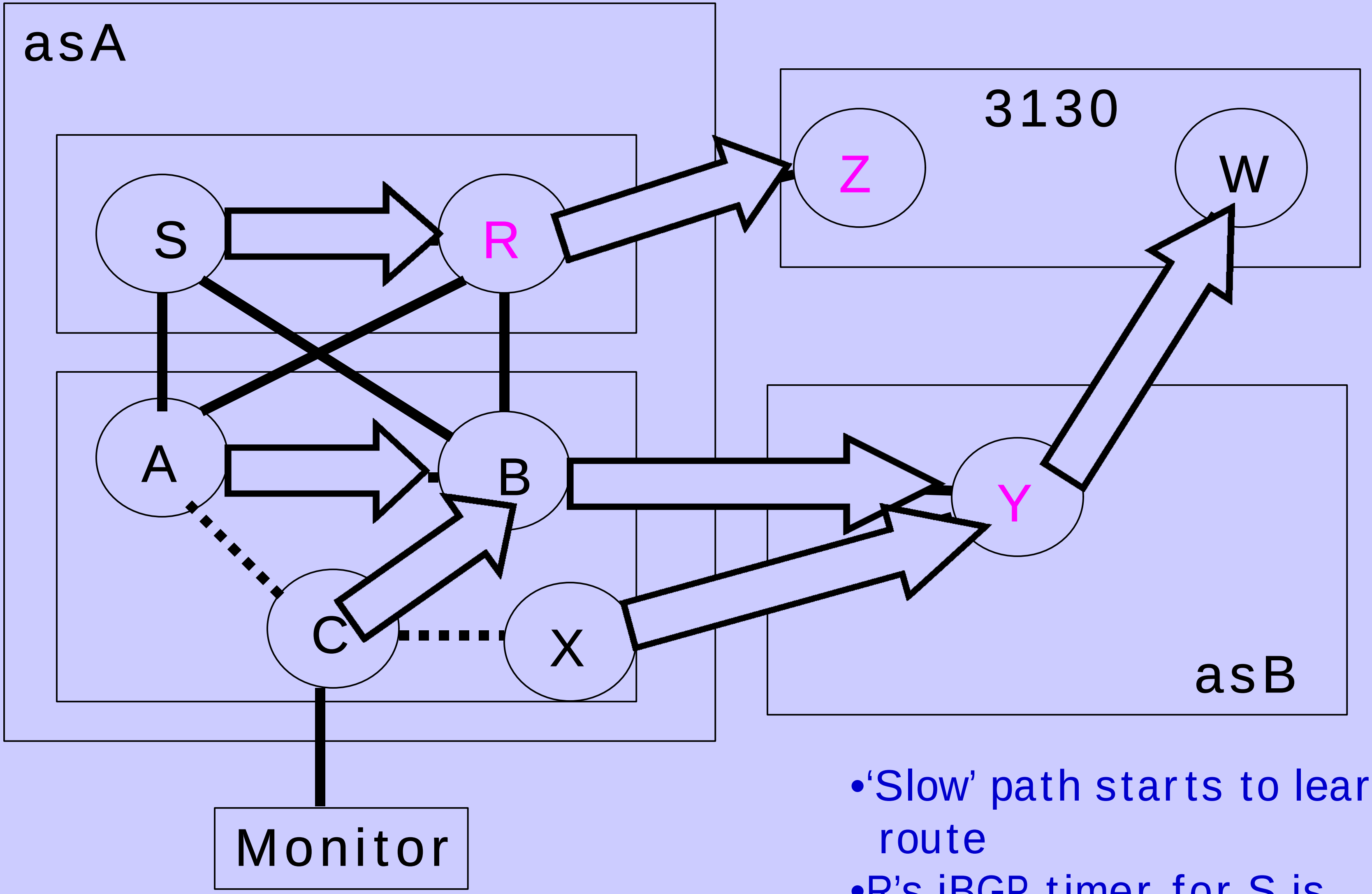
State 1

routers hear route on ‘fast’ paths



A: asA asB 3130

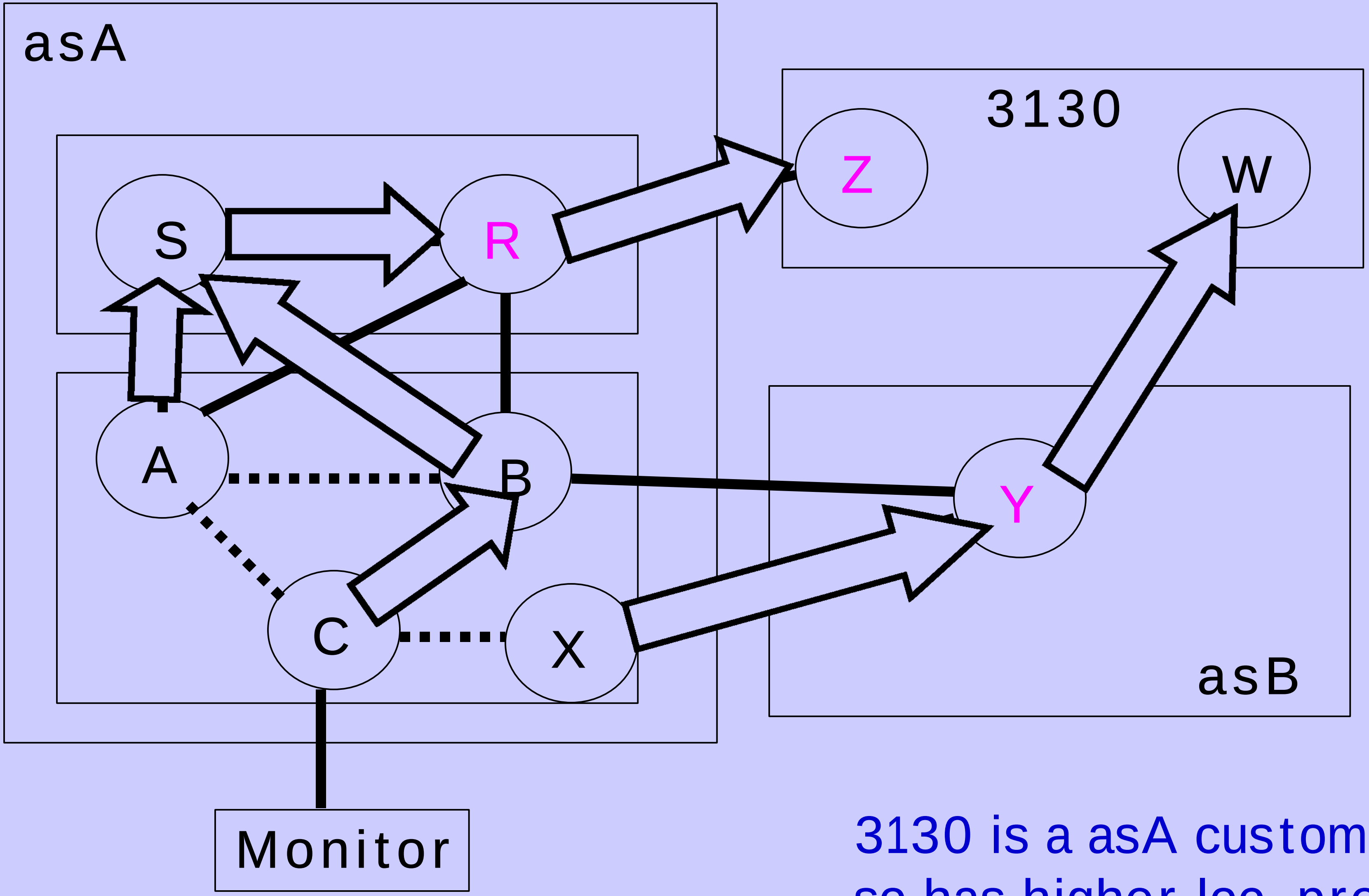
State 2



- ‘Slow’ path starts to learn route
- R’s iBGP timer for S is smaller than it’s eBGP timers for A and B

A: asB 3130

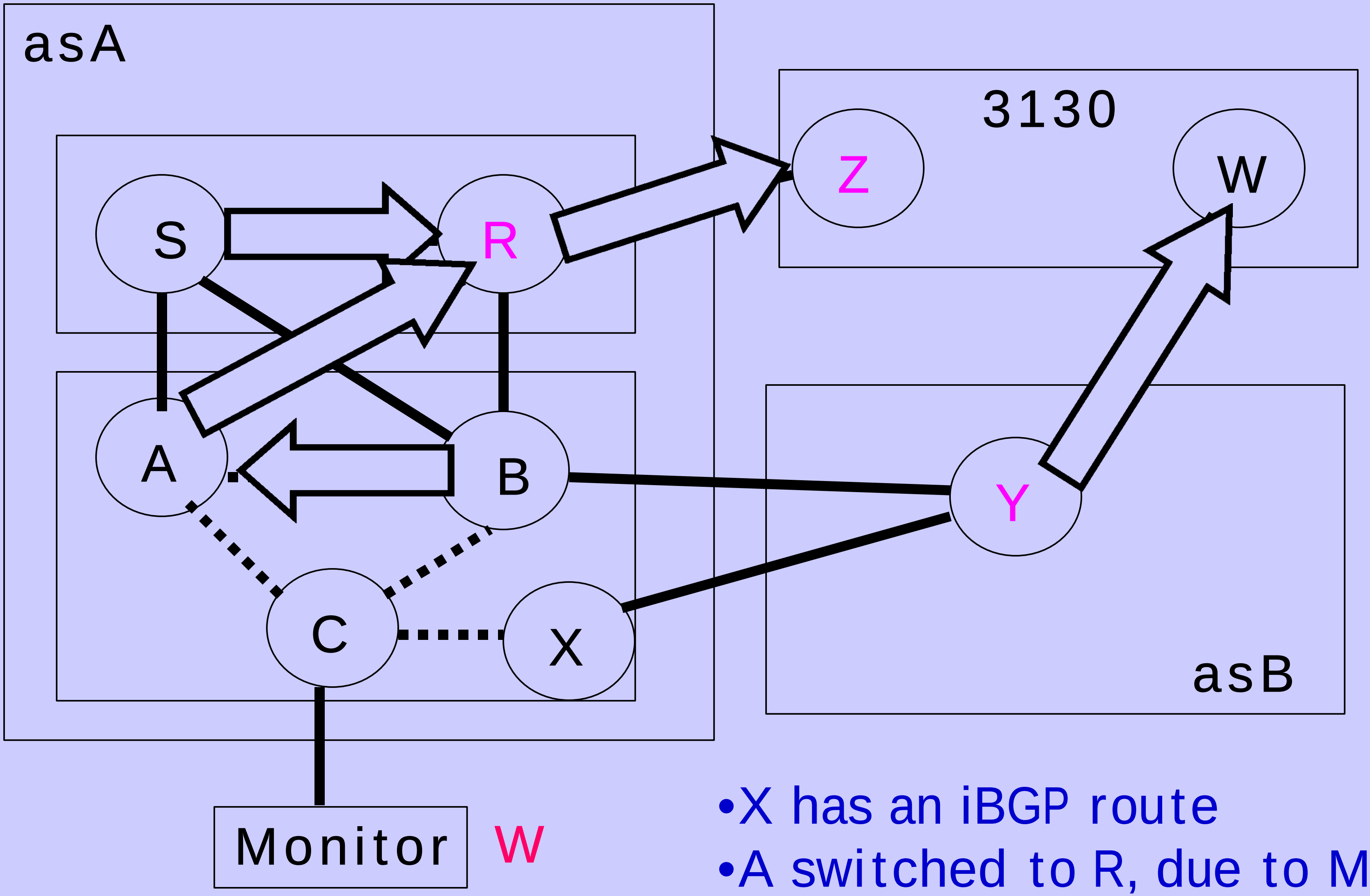
State 3



A: asA 3130

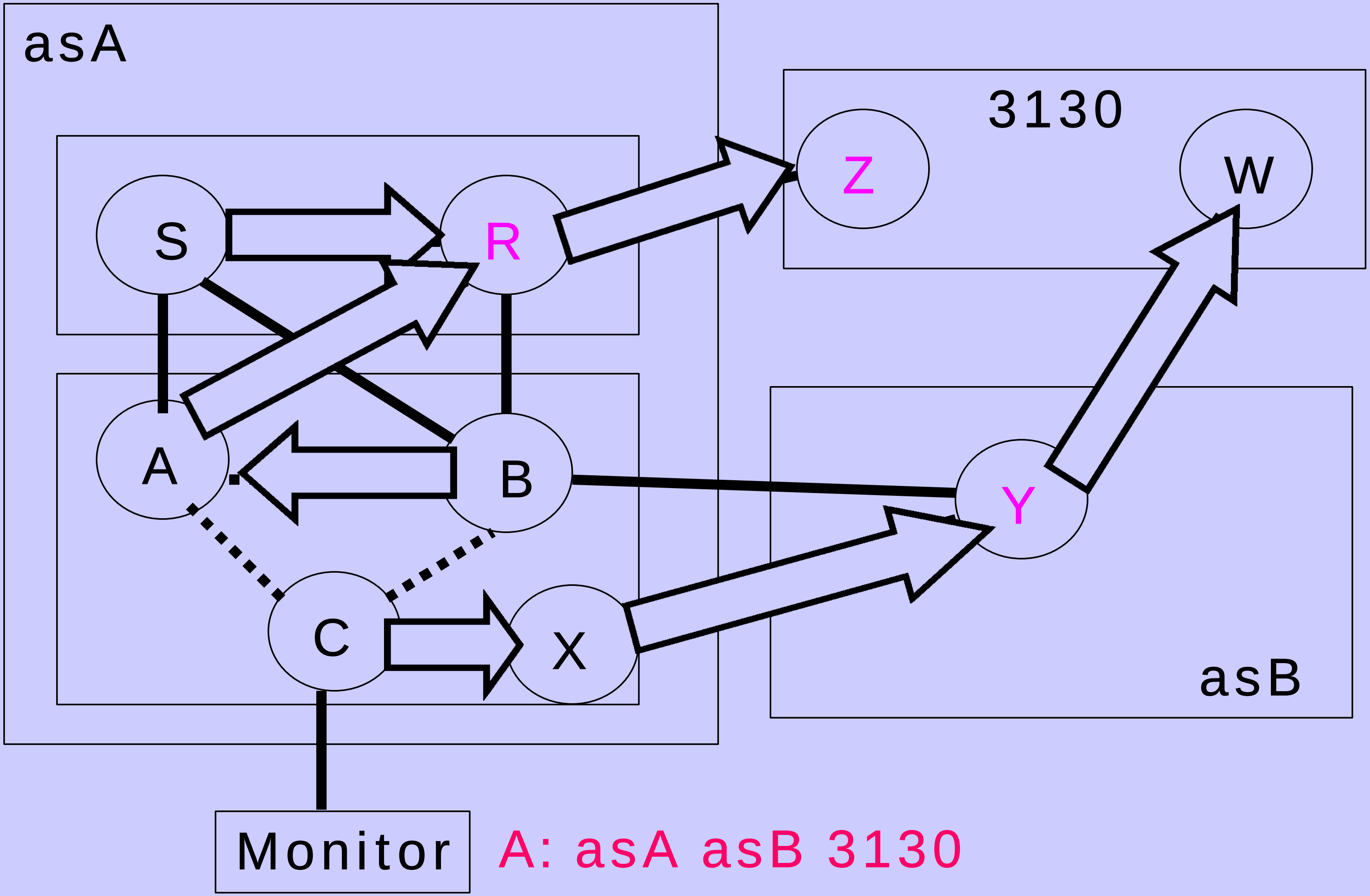
3130 is a asA customer, so has higher loc_pref, and A and B like it…

State 4



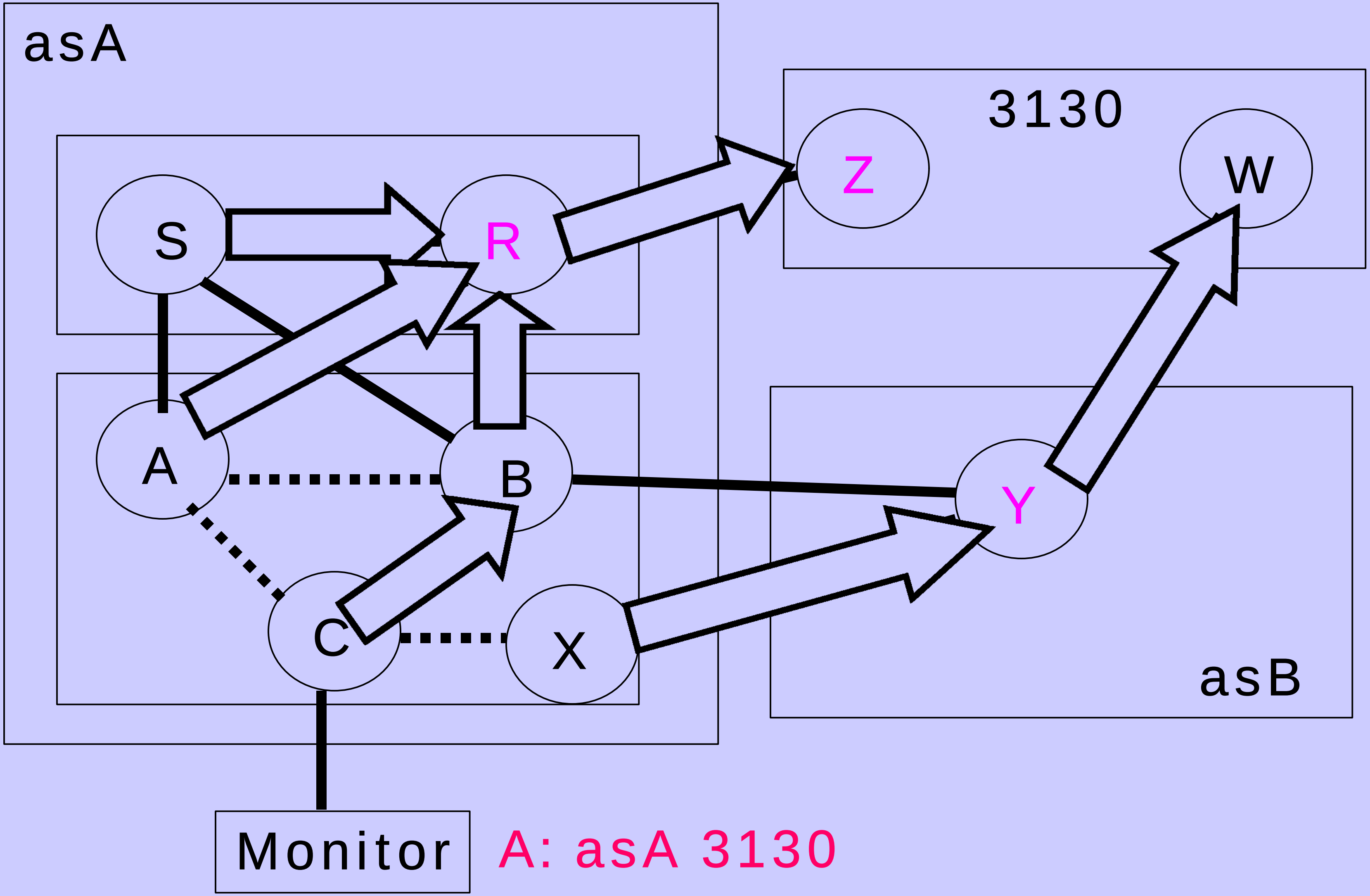
- X has an iBGP route
- A switched to R, due to MED
- B switches to A, also due to MED

State 5



A: asA asB 3130

State 6



A: asA 3130

Signals Seen by the Monitor

A: asA asB 3130
A: asA 3130
W
A: asA asB 3130
A: asA 3130

- With multiple S nodes, and multiple X nodes, it is possible to explain multiple Withdraws
- It has been shown in the lab that there are reasonable configurations which **never settle**
- Also see Griffin on iBGP configuration issues

`http://www.acm.org/sigcomm/sigcomm2002/papers/ibgp.html`

Idealism

If route withdraws are treated immediately (or at least quickly) and changes propogated more slowly, then route withdraw is order(1). A route addition is order(1), the addition of a better route is order(1) and a route change where the better route is removed is order(1).

-- Curtis Villamizar (router vendor)

Fri, 01 Aug 2003 16:35:08 -0400

routing-discussion@ietf.org

This Talk was about Observed Reality