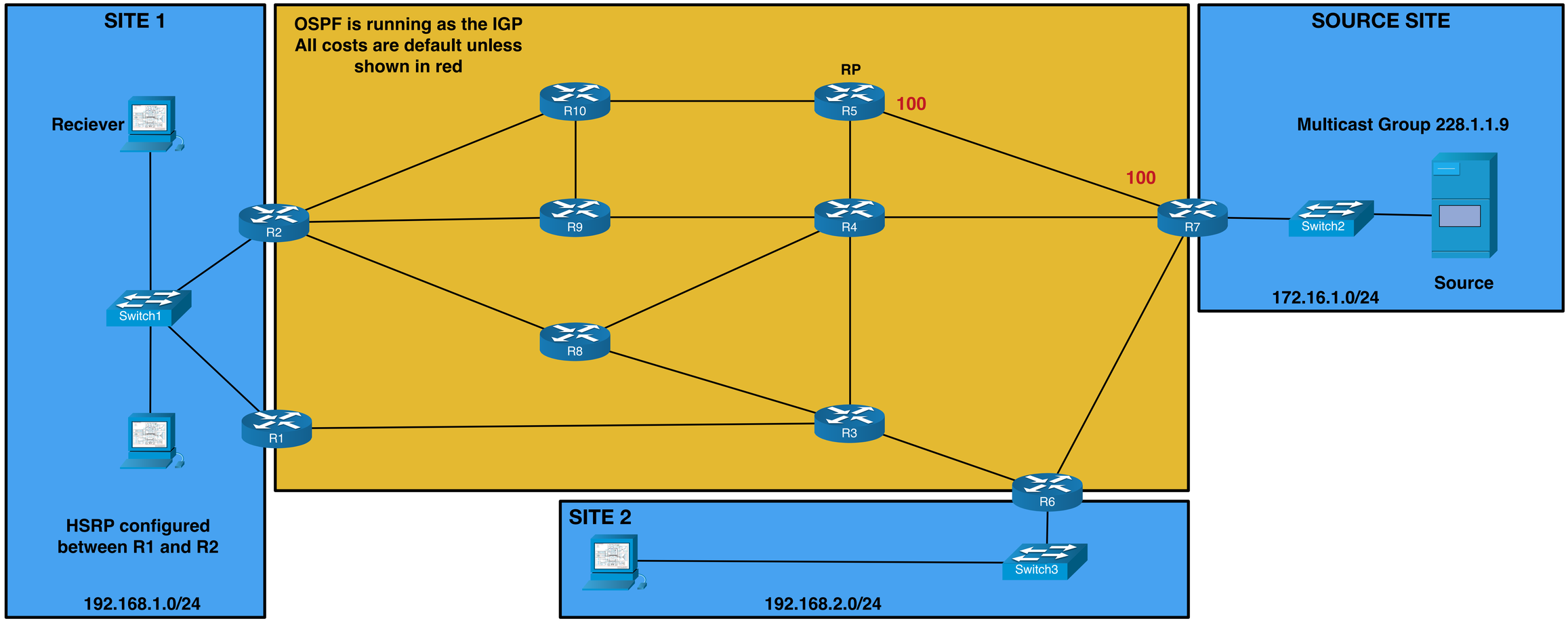
This document will outline the process of how a multicast source on the right will send to a receiver at site one on the left using PIM-SM configured in the core.

The following will be explored:
> RP elections - using both Bootstrap and AutoRP
> The receiver joining the multicast group and connecting to the shared tree
> The multicast source sending traffic the multicast group and how it registers to the RP
> SPT switchover once multicast traffic is flowing.

The diagram below shows the network that will be demonstrated. This topology is downloadable as an EVE-NG lab from the labs page on netquirks.

Multicast
PIM Sparse-Mode



PIM Sparse-Mode

RP election

A rendezvous point (or RP) in a multicast network is a common root of a shared distribution tree. In this infrastructure, traffic will first flow to the RP before flowing downstream to the clients/recievers. RPs can be statically configured or elected. There are two common election protocols - Bootstrap (non-proprietary) or Auto-RP (Cisco proprietary). This page shows Bootstrap Protocol.

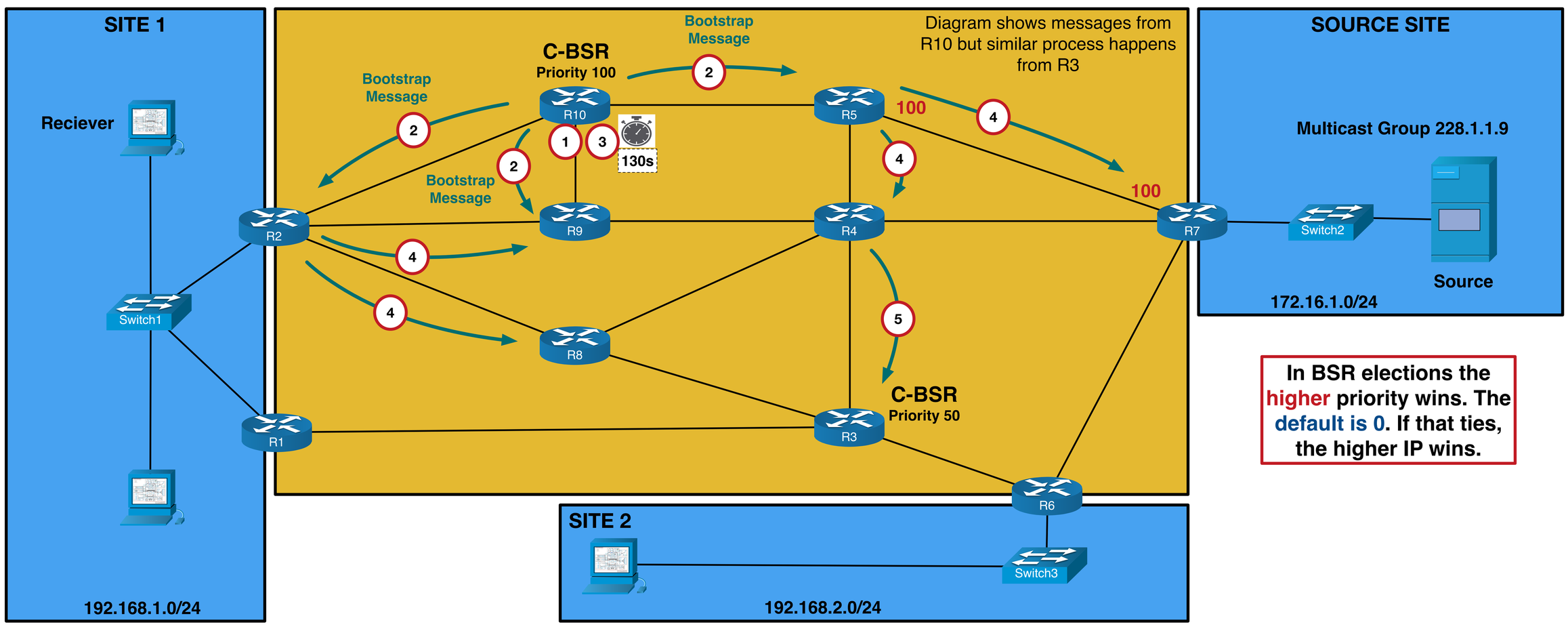
Bootstrap Protocol

Bootstrap protocol defines two different router roles:
> Candidate Boot Strap Routers - C-BSR
> Candidate Rendezvous Points - C-RP

C-BSRs will communicate with one another to elect a single BSR who will then listen to messages from the C-RPs. C-RPs will advertise the multicast groups for which they would like to be the RP along with a priority. The BSR, upon hearing these messages, will compile an RP-SET. This RP-SET is then advertised throughout the network and used by each router to decide on a common RP for each multicast group.

The downloadable EVE-NG lab on the netquirks lab page is configured with a BSR election.

BSR Election

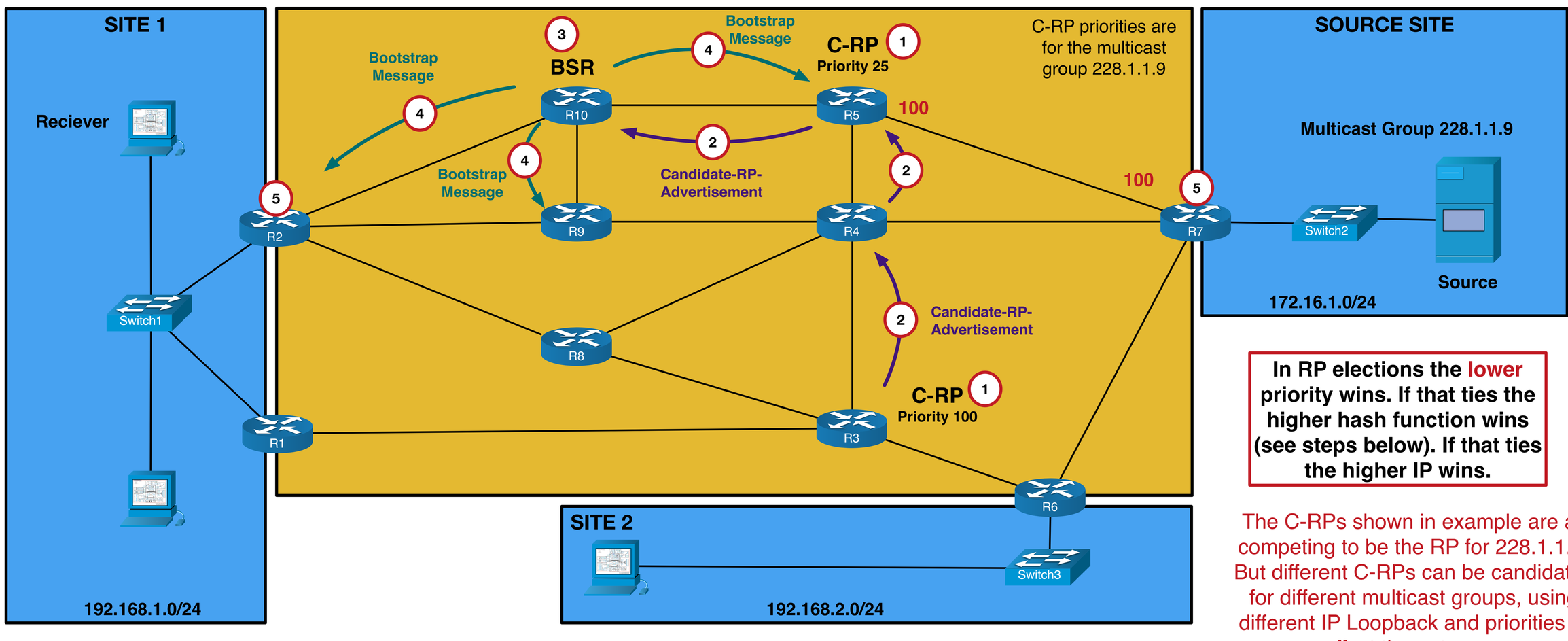


In BSR elections the higher priority wins. The default is 0. If that ties, the higher IP wins.

Step	Description
1	Each BSR is assigned an IP and a Priority (0-255, default is 0). In the above diagram R10 gets priority 100.
2	Once configured, each C-BSR assumes itself to be the election winner and final BSR. It starts to send Bootstrap Messages to 224.0.0.13 with a TTL of 1 out of all interfaces with multicast enabled.
3	At the same time the C-BSR will start a Bootstrap Timer of 130s and listens for Bootstrap Messages from other C-BSRs.
4	If a non-BSR receives the bootstrap message it will send it out all of it's interfaces except the one it received it on (showing only for R2 and R5 in the diagram above)
5	When a C-BSR gets a Bootstrap Message (showing R3 receiving R10s in the above diagram)… > If the Messages priority is higher than its own, it will reset its Bootstrap Timer and will stop sending messages of its own (in this way this highest priority will win). If R3's Bootstrap Timer expires due to lack of Bootstrap Messages with a higher priority, it will again assume that it is the BSR and start to send Bootstrap Messages again. > If the message priority is lower, the router will continue sending Bootstrap Messages every 60 seconds - essentially announcing itself as the BSR. If both BSR priorities are the same the Higher BSR IP wins. Since R10 has a higher priority than R3, R3 stops sending Bootstrap Messages.

R10 will win the BSR election and be the only one sending Bootstrap Messages. Next the RP-SET is generated…

RP-SET Generation and RP Election



In RP elections the lower priority wins. If that ties the higher hash function wins (see steps below). If that ties the higher IP wins.

The C-RPs shown in example are all competing to be the RP for 228.1.1.9. But different C-RPs can be candidates for different multicast groups, using different IP Loopback and priorities to affect the outcome.

Step	Description
1	Each C-RP is assigned an IP, a priority (0-255) and a list of the multicast groups for which it would like to become the RP (this is typically done with an ACL in the config).
2	Once the BSR is known, each C-RP will send unicast Candidate-RP-Advertisements to the BSR. These advertisements include the IP, priority and multicast group list.
3	The BSR compiles all of the Candidate-RP-Advertisements it receives and creates an RP-SET. The RP-SET contains the C-RPs for each multicast group, their priorities and IPs, as well as an 8-bit hash-mask.
4	The RP-SET is advertised throughout the domain in Bootstrap Messages to 224.0.0.13.
5	Each router (R2 and R7 highlighted above), receives the RP-SET in the Bootstrap Messages and elects an RP for each multicast group based on the following: a. C-RP with lower priority wins b. If that ties, a hash function between the group prefix, 8-bit hash-mask and C-RP address is performed. The higher hash wins. c. If that ties, the C-RP with the higher IP wins The above diagram we will see that R5 and R7 both elect R5 as the RP.

BSR is non-proprietary, unlike Auto-RP…

PIM Sparse-Mode

RP election

An alternative to Bootstrap Protocol
is Cisco's Proprietary Auto-RP

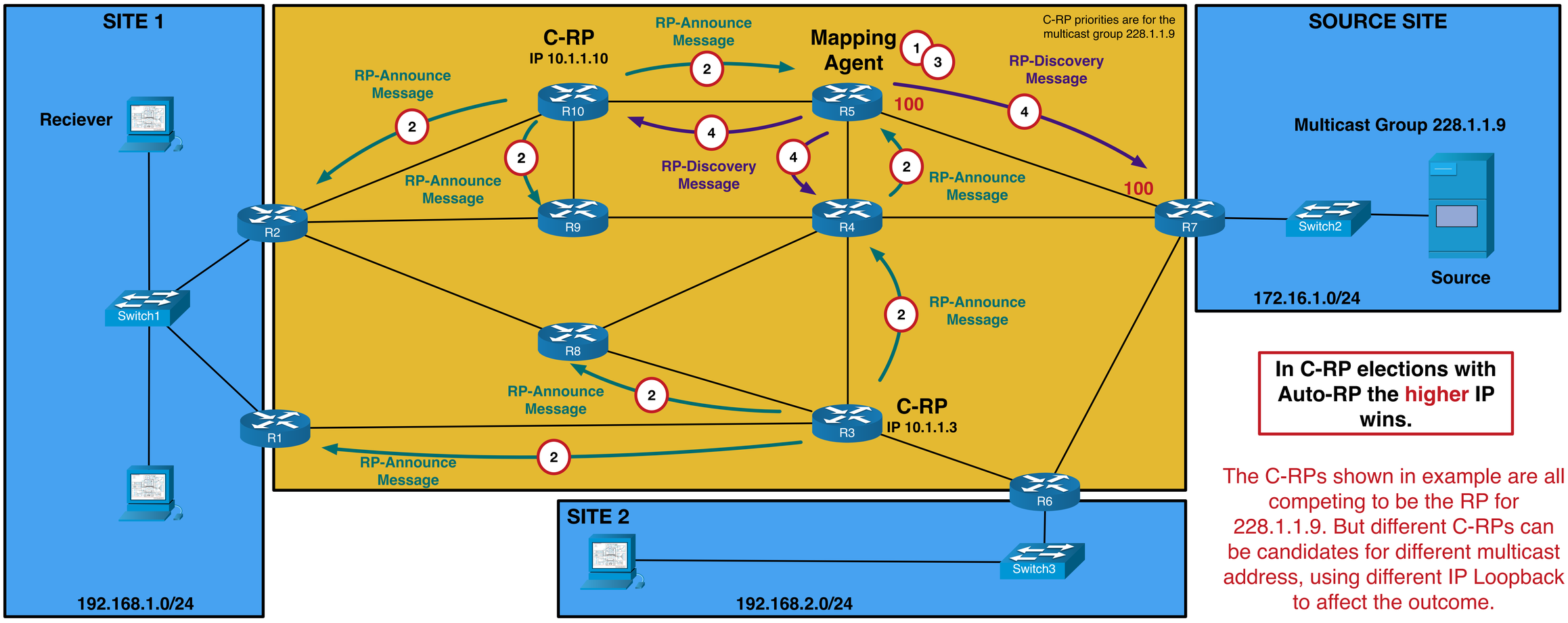
Note that in Auto-RP the Mapping Agents work similar to the BSR in Bootstrap Protocol. The significant difference is that the Mapping Agent decides which RP is mapped to which multicast group. In Bootstrap Protocol, the BSR simply sends out the RP-SET. It is actually each receiving router that performs the elections based on information in the RP-SET.

Auto-RP Protocol

Auto-RP defines two different router roles:
> Mapping Agents
> Candidate Rendezvous Points - C-RP

Mapping agents are manually assigned, unlike BSRs in Bootstrap Protocol. Candidate RPs advertise themselves to the Mapping Agents who then build a table containing RP to multicast group mappings.

RP Election



In C-RP elections with Auto-RP the higher IP wins.

The C-RPs shown in example are all competing to be the RP for 228.1.1.9. But different C-RPs can be candidates for different multicast address, using different IP Loopback to affect the outcome.

Step	Description
1	Mapping Agents are configured manually (in this case R5). They listen for RP-Annouce Messages from C-RPs
2	C-RPs send RP-Announce messages to 224.0.1.39. These include the routers IP and the multicast groups for which they want to be the RP.
3	Mapping agents will receive these RP-Announce messages and build an RP to multicast group mapping table. If multiple C-RPs compete for the same multicast group, the C-RP with the higher IP will win. Here R10 wins with the higher IP.
4	Mapping agents will send out the mapping table to all routers using RP-Discovery messages. These messages are sent to 224.0.1.40.

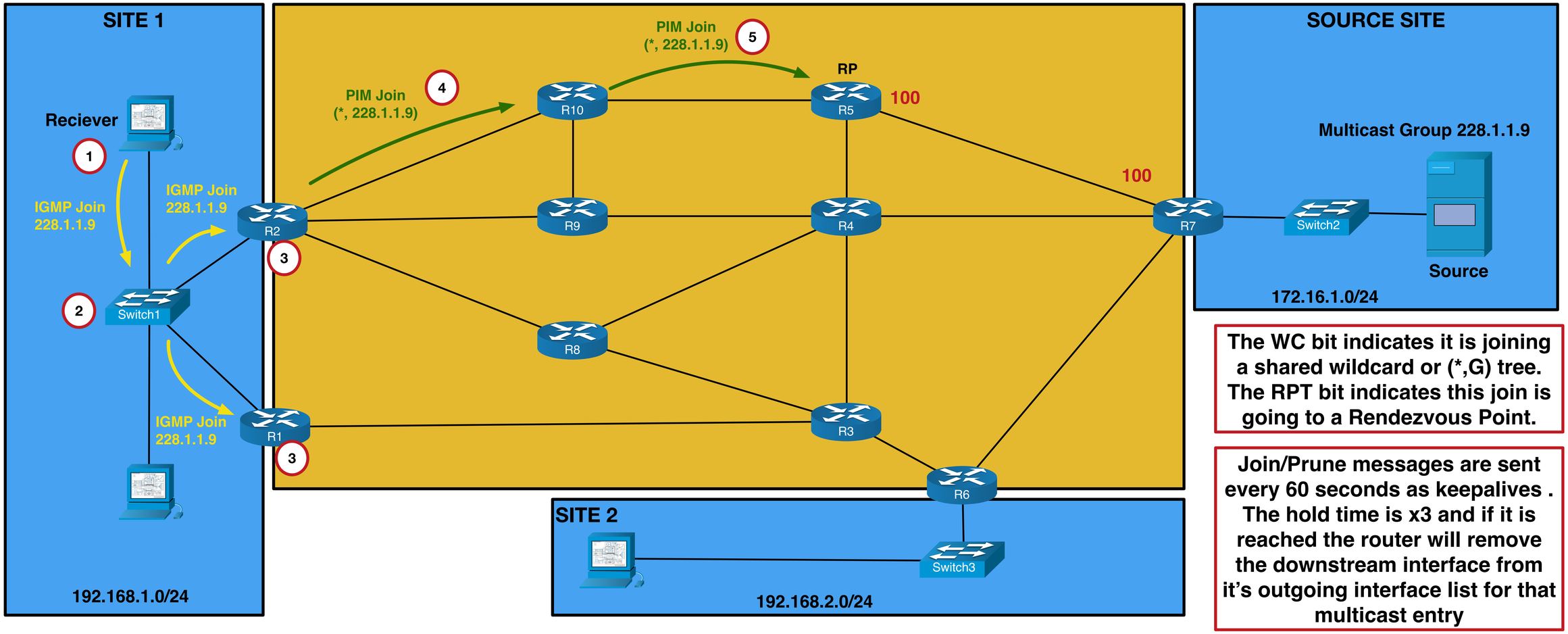
Once the RP is known, receivers can request to join the multicast group…

by Steven Crutchley

PIM Sparse-Mode

Receiver Joining Tree

Now that the RP election has been completed, we will look at what happens when a multicast receiver wants to join a multicast group and how its local Designated Router (DR) will join the shared tree. This example uses R5 as the RP.



The WC bit indicates it is joining a shared wildcard or (*,G) tree. The RPT bit indicates this join is going to a Rendezvous Point.

Join/Prune messages are sent every 60 seconds as keepalives . The hold time is x3 and if it is reached the router will remove the downstream interface from it's outgoing interface list for that multicast entry

Step	Description
1	Host sends an IGMP Join. The destination address is the multicast group but this is also included in the IGMP packets payload.
2	The destination MAC address is comprised of 0100:5E00:0000 combined with the lower 23 bits of the multicast address. In this case 0100:5E00:0009. The switch will forward it onto both routers.
3	R2, having the higher IP address is the Designated Forwarder for this LAN segment. So it will send an a PIM (*,G) join towards the RP after adding an entry into its multicast routing table with it's LAN interface as an outgoing interface. R1, having a lower IP, will be the IGMP querier, responsible for monitoring and communicating Group Memberships.
4	The (*,G) join is sent on the IGPs best path to the RP. The WC and RPT bits are set.
5	As each router receives the PIM join it does the following: > If the router is already part of the shared tree it just adds the receiving interface to the OIL (outing interface list) and does nothing else. > If it isn't part of the shared tree, and isn't the RP, it adds the receiving interface to the OIL, creates a (*,G) entry in the multicast routing table and sends a join along its way upstream to the RP. > If the router is the RP it will create a (*,G) entry if it doesn't have one.

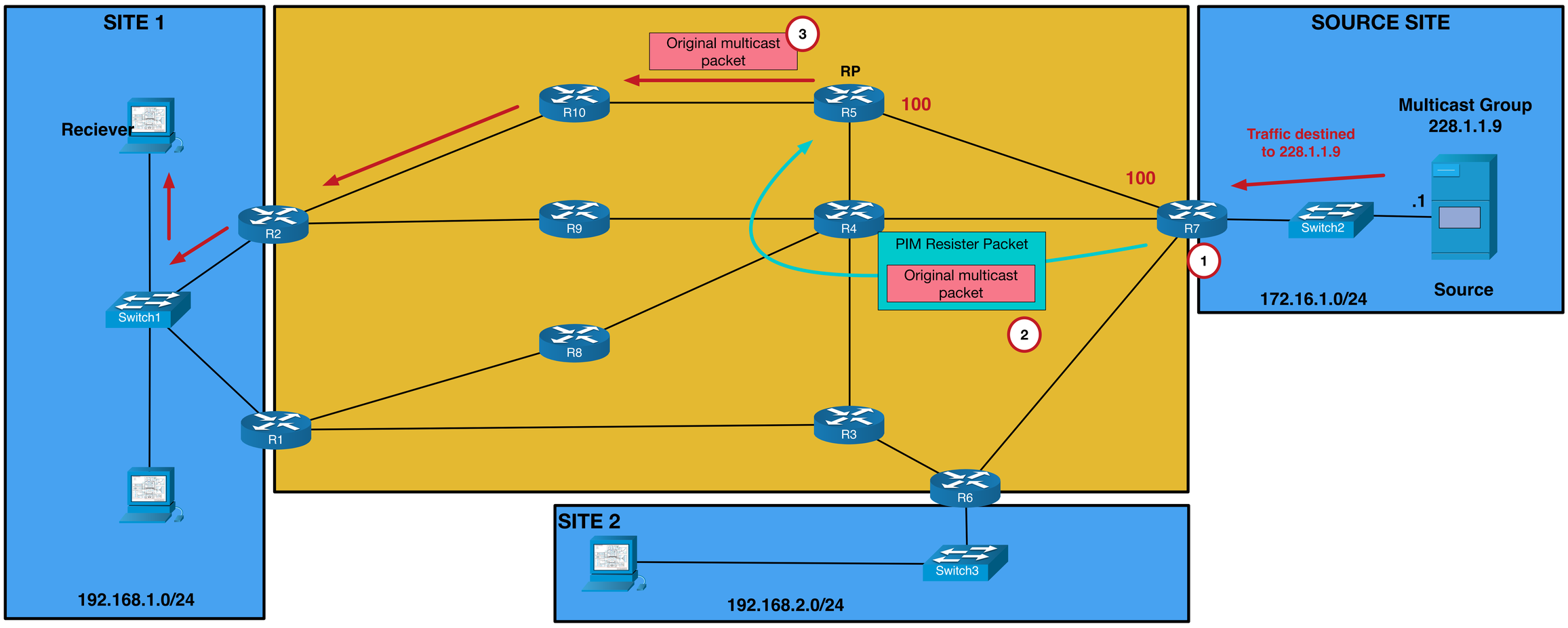
Once this connection to the RP is complete, the next phase is to see what happens when the source starts sending…

PIM Sparse-Mode

Source starts sending

When the multicast source starts to send traffic it must initially go through the RP before heading downstream to any receivers. The multicast traffic first reaches the RP, from the DR, by being encapsulated in unicast register packets before moving to native multicast…

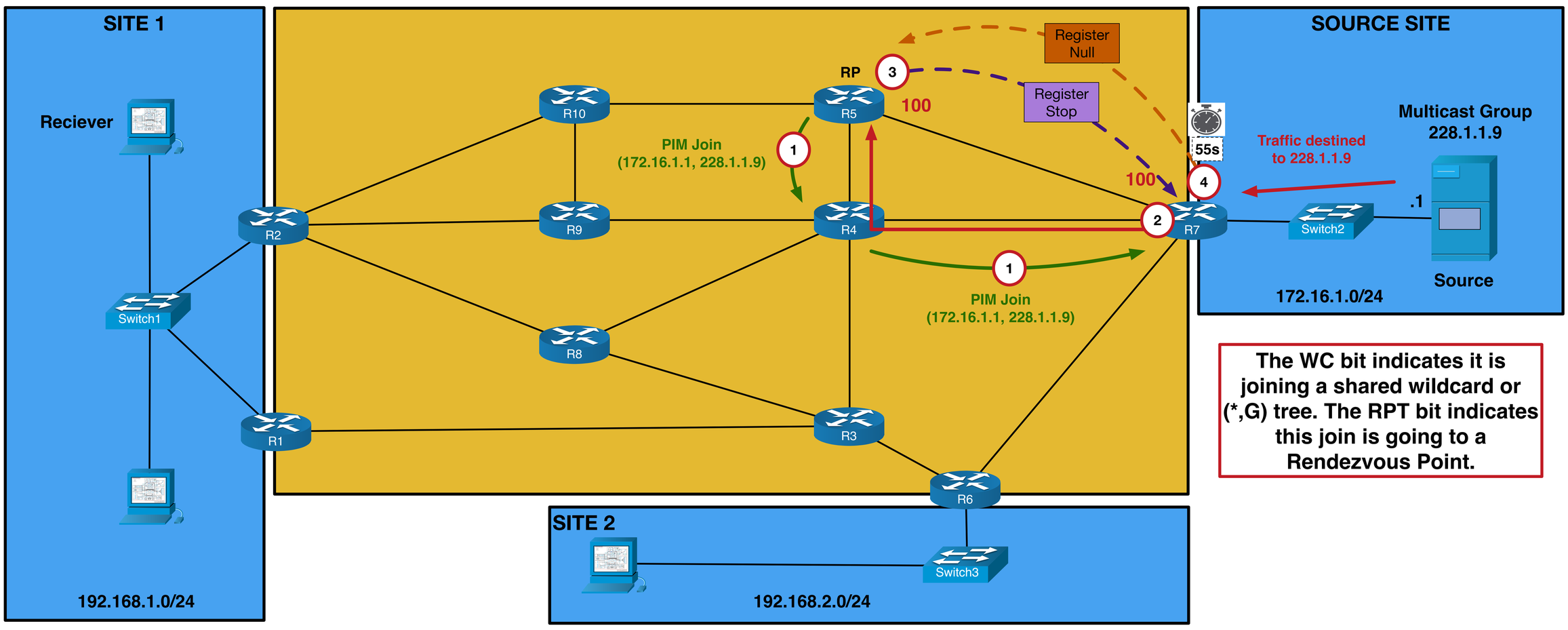
Initial PIM Register



Step	Description
1	The Designated Router, R7 in this case, receives packets from the directly connected source. It looks up its RP for the multicast group (see Bootstrap Protocol and Auto-RP on previous pages).
2	R7 then encapsulates the original multicast packet into a unicast Register Packet and forwards it on to the RP.
3	The RP will decapsulate the Register Packet and send the original multicast packet down the shared tree that the receiver has already joined (see previous page). It does this for all multicast traffic to begin with.

It is inefficient to keep encapsulating traffic from the DR to the RP in unicast packets, so the RP builds an multicast tree to the source DR…

RP building tree to DR



The WC bit indicates it is joining a shared wildcard or (*,G) tree. The RPT bit indicates this join is going to a Rendezvous Point.

Step	Description
1	The RP creates an (S,G) entry. Where S is the source (in our case 172.16.1.1) and G is the multicast group. It sends a join upstream towards the DR (R7) which will be the root of this SPT. WC and RPT bits are 0 since this is not a shared tree.
2	Once the SPT (shortest path tree) is built, R7 will sent the multicast traffic directly down this tree.
3*	As soon as the RP starts to get multicast traffic over the SPT, it will send a Register Stop message to the DR to stop it from sending encapsulated packets
4*	When R7 gets the Register Stop message it will stop ending encapsulated traffic to the RP. R7 will then keep two timers: **Register Suppression Timer (60s) -** if this expires R7 will start sending encapsulated packets again. **Register Null Message Timer (55s) -** This is an empty Register Message sent to the RP, which will prompt the RP to send a Register Stop message back the DR, resetting the Register Suppression Timer. In the event that all group members leave the group and the RP has no downstream recievers, it will simply send a Register Stop to the source DR and prune (or not build) the SPT.

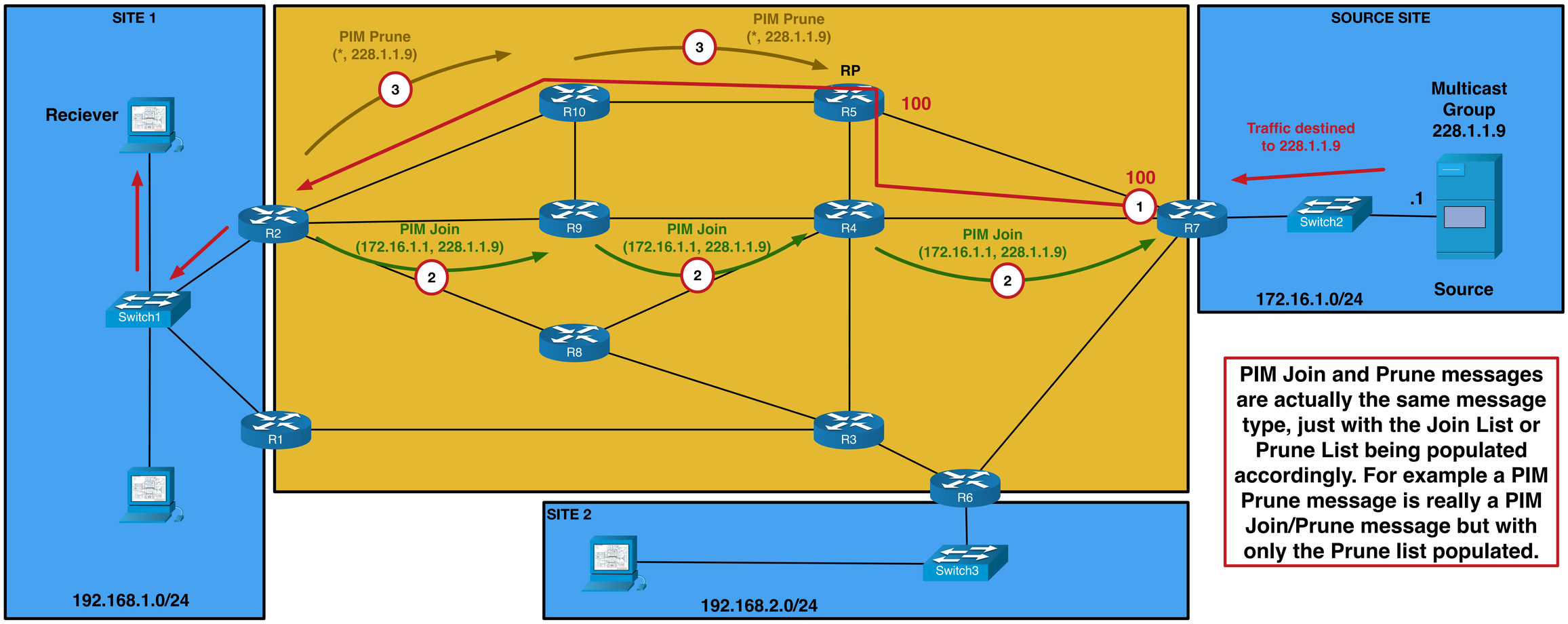
*In the diagram above, Register Stop and Null messages will follow the best path to reach they respective destinations. This detail has been omitted for clarity.

At this point traffic is flowing down the shared tree via the RP, but this isn't the optimal path from R7 to R2…

PIM Sparse-Mode

SPT Switchover

Once R2 learns the source of the multicast traffic it will build a direct SPT to it. This process is called SPT switchover.



PIM Join and Prune messages are actually the same message type, just with the Join List or Prune List being populated accordingly. For example a PIM Prune message is really a PIM Join/Prune message but with only the Prune list populated.

Step	Description
1	Traffic will originally be going via the RP to reach the receivers.
2	R2 will learn of the multicast source. In this case it's best unicast route to the source is out of a different interface, than the one to reach the RP. It builds an SPT directly to the source - in this case the source is 172.16.1.1, the server on the right.
3	Once the SPT has been built, R2 will send a prune up to the RP to remove itself from the shared tree.

The SPT switchover happens after a bandwidth threshold is met. The Cisco default threshold is 0. This basically means it will switch to the SPT tree immediately.

The final traffic flow looks as follows: