Bandwidth Aware Multicast Load Balancing

Liming Wei, Vincent Ng – Cisco Systems

Apricot 2013

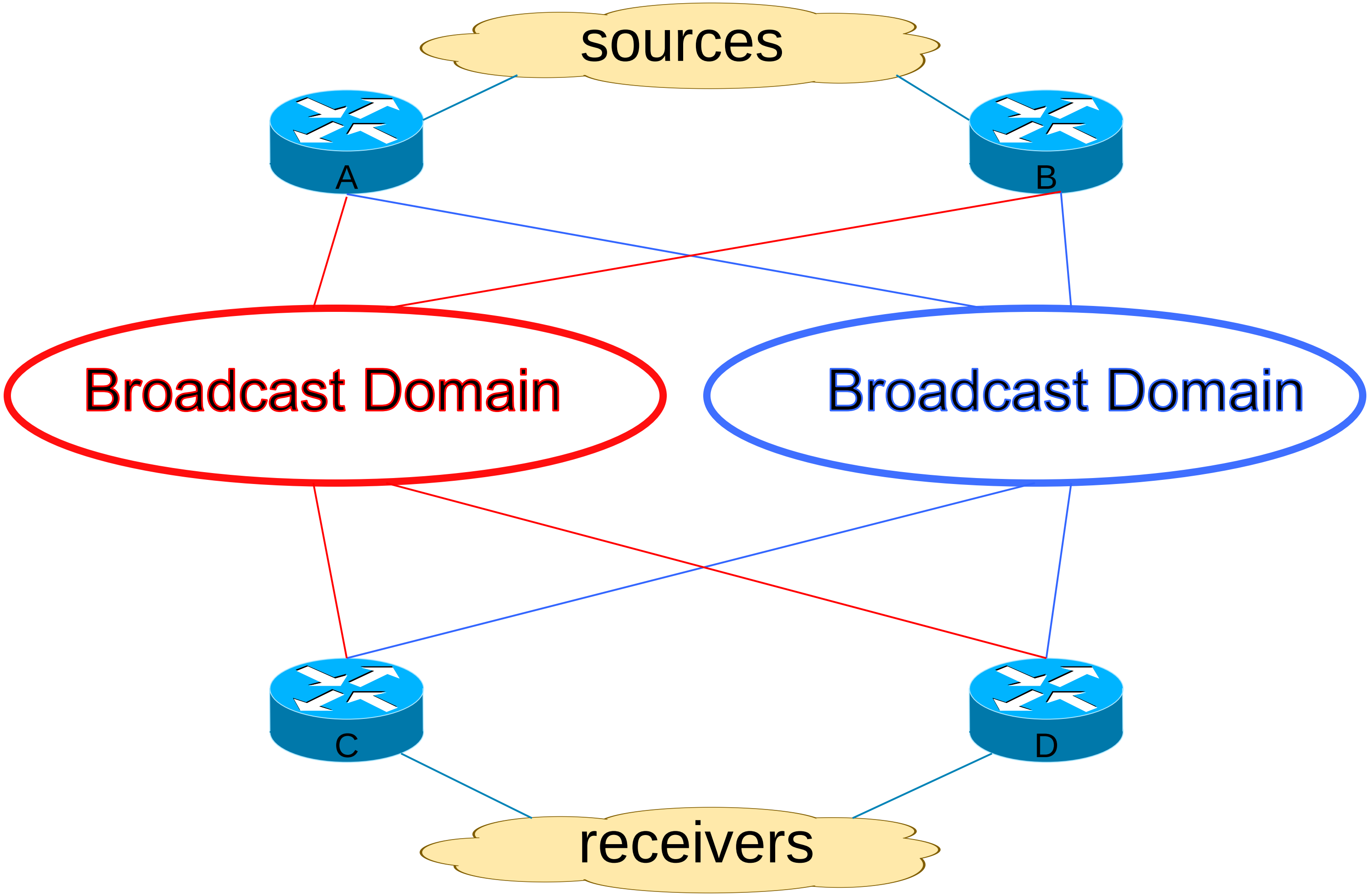
History of RFC6754 PIM ECMP Redirect

- Problem statement started to form in early-mid 2010
  - Look for simple and effect solution while protecting existing PIM based deployment.
- Feature development
  - Brainstorm and discussions on operations and objectives.
  - First design concept near end of 2010.
  - First presentation to IETF in Prague Mar/April 2011.
  - ID adopted by IETF PIM WG July 2011
  - Standard Track RFC6754 issued October 2012.
- Feature Implementation December 2012

Use Case for Multicast Load Balancing

- Core of network has multiple equal cost paths
- User want to load balance and also limit the proportion of multicast on each path so unicast traffic is not affected by multicast loads
- Previous multicast ECMP algorithm has disadvantages
  - It is not bandwidth aware of each flow
  - It is not aware of the bandwidth limit of each path
  - It can't protect against oversubscription of each path
- Manual allocation of multicast traffic and/or multiple IGP topologies tuning cannot scale

Reference Topology



Existing ECMP/RPF Issues

- ECMP RPF selection is downstream driven only
- Two ways to choose an RPF path if ECMP present:
  - Choose PIM neighbor with higher IP address
  - Use a hash algorithm on S,G addresses
- Main issues:
  - Based on IP addresses only, not bandwidth aware of links and flows
  - Cannot avoid under-utilization or oversubscription of links
  - Flows may be sent over multiple links causing waste of bandwidth

Solution Overview

- Need an automated mechanism to:
  - put multicast traffic into links with available multicast bandwidth.
  - avoid oversubscription of links into the set of available paths.
- In summary, the solution has two parts
  - Downstream nodes steer multicast traffic by policy based RPF selection.
  - Upstream nodes steer multicast traffic by triggering PIM ECMP Redirect messages to downstream nodes.

Assumptions and Goals

- Key Assumptions
  - Multicast traffic is significant on the paths
  - The paths form ECMP from IGP point of view
  - The paths may have different physical bandwidths to facilitate migration and/or addition of links of new technologies
  - The paths may have different allocated bandwidths for use with multicast
- The goals are to
  - Run a single instance of IGP and support ECMP
  - Choose the RPF from ECMP based on bandwidth
  - Avoid using different paths for the same (S,G)

New PIM Concepts

- PIM Joins are only sent to paths with the most available multicast bandwidth.

  Hash is not used to select RPF interface/neighbor

- New PIM ECMP Redirect are used to solve the following problems
  - The same (S,G) is forwarded on to two different paths.
  - One path has exceeded bandwidth threshold and another hasn’t.
- Two new concepts introduced
  - ECMP bundle
  - PIM ECMP Redirect

ECMP Bundle

- An ECMP Bundle is configured to have multiple, independent L3 interfaces
  - Red and Blue form a bundle in the reference topology
  - Created on both upstream (A and B) and downstream (C and D) routers
- IGP and PIM are run on each individual interfaces
  - Hence creating ECMP between upstream and downstream routers
- Used by downstream routers to load balance multicast traffic and by upstream routers to send PIM ECMP Redirect

Bandwidth Policy Definition

- A policy database is created to match multicast flows and map to appropriate per flow bandwidths
- Configure two link multicast bandwidths: X(threshold) and Y(max)
  - Per interface configuration X <= Y <= Interface Bandwidth
  - Downstream router will attempt to use links with most available multicast bandwidth
  - When X is reached, a downstream router always attempts to use a different RPF interface for new (S,G)
  - When Y is reached on all interfaces, then new flows will not be established

PIM ECMP Redirect

- New PIM ECMP Redirect Hello option and ECMP Redirect message defined in RFC6754
- Sent by upstream routers to tell downstream routers to join another “desired” interface
- Triggered by receiving PIM Joins from “non-desired” outgoing interfaces, for example
  - If the upstream router is forwarding out to another interface within the same ECMP bundle
  - If the upstream believes Y has occurred on that interface

Packet Format

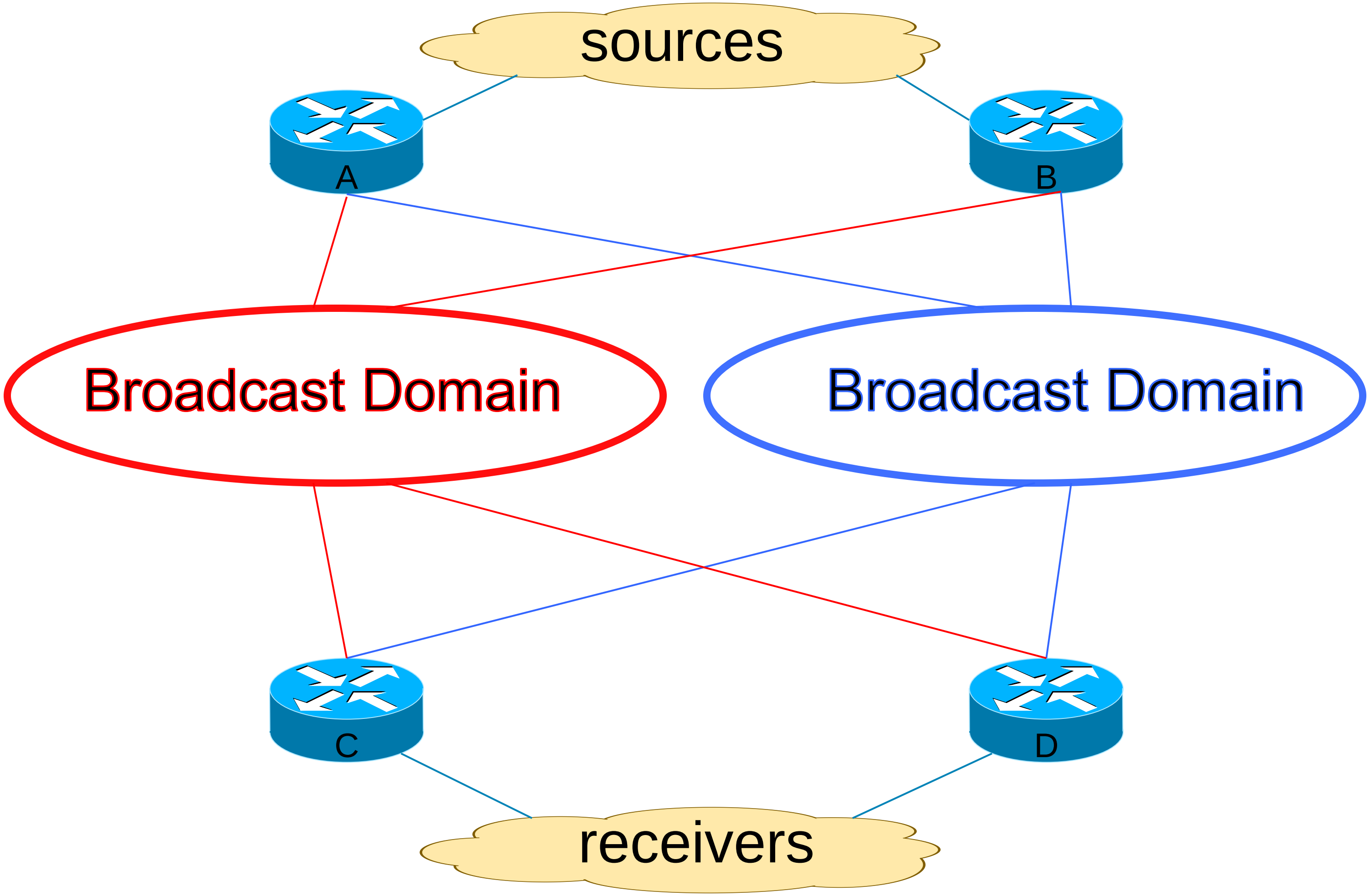
- PIM Hello Option

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type = 32           |        Length = 0             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

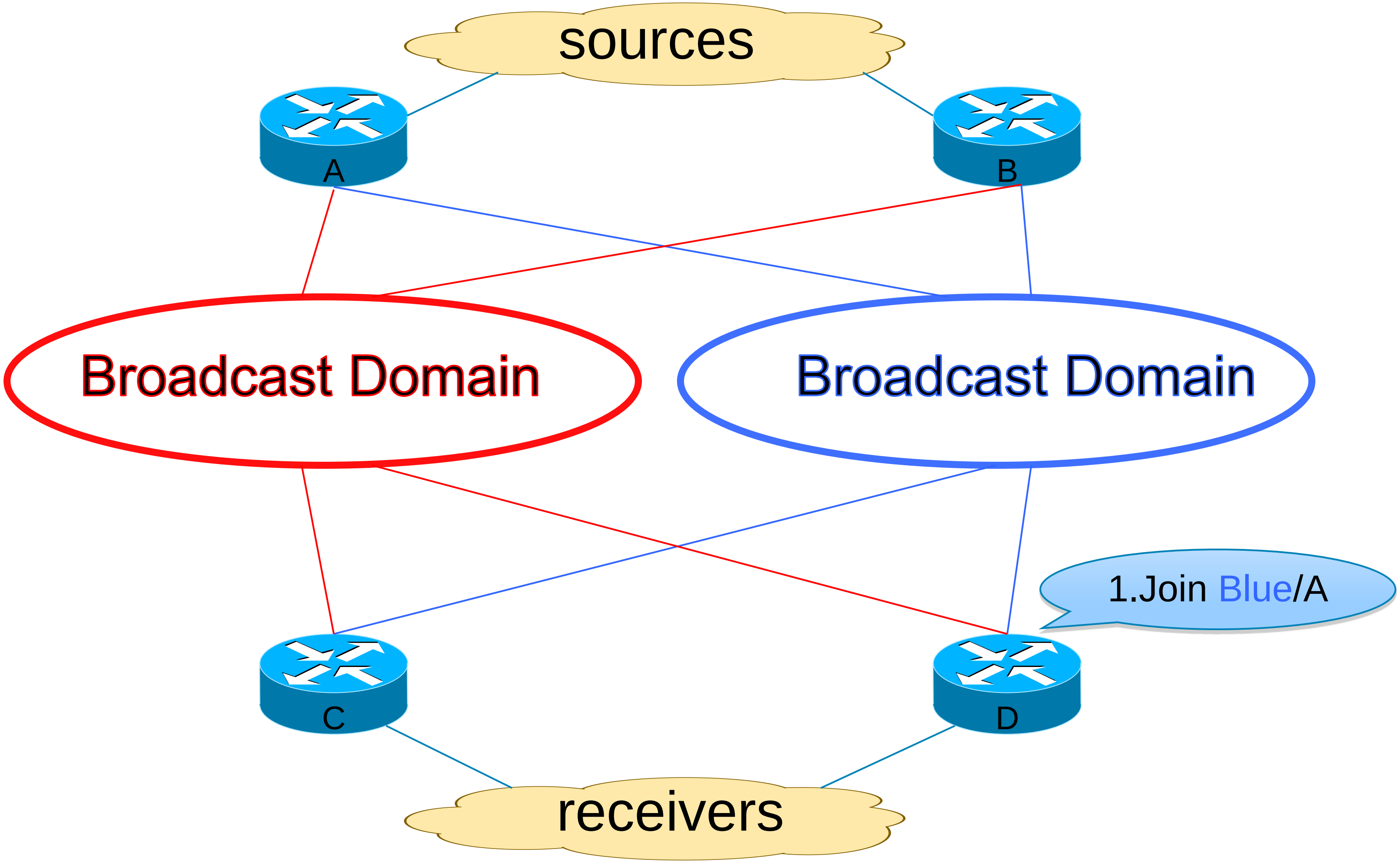
- PIM ECMP Redirect Format

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver| Type  |   Reserved    |          Checksum             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Group Address (Encoded-Group format)             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Source Address (Encoded-Unicast format)            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Neighbor Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+-+-+-+-+-+- ............ Interface ID ........... -+-+-+-+-+-+-+
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Preference  |                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+--  ... Metric ...  -+-+-+-+-+-+-+-+-+
|                                                               |
+- .. Metric .. +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               |
+-+-+-+-+-+-+-+-+
```

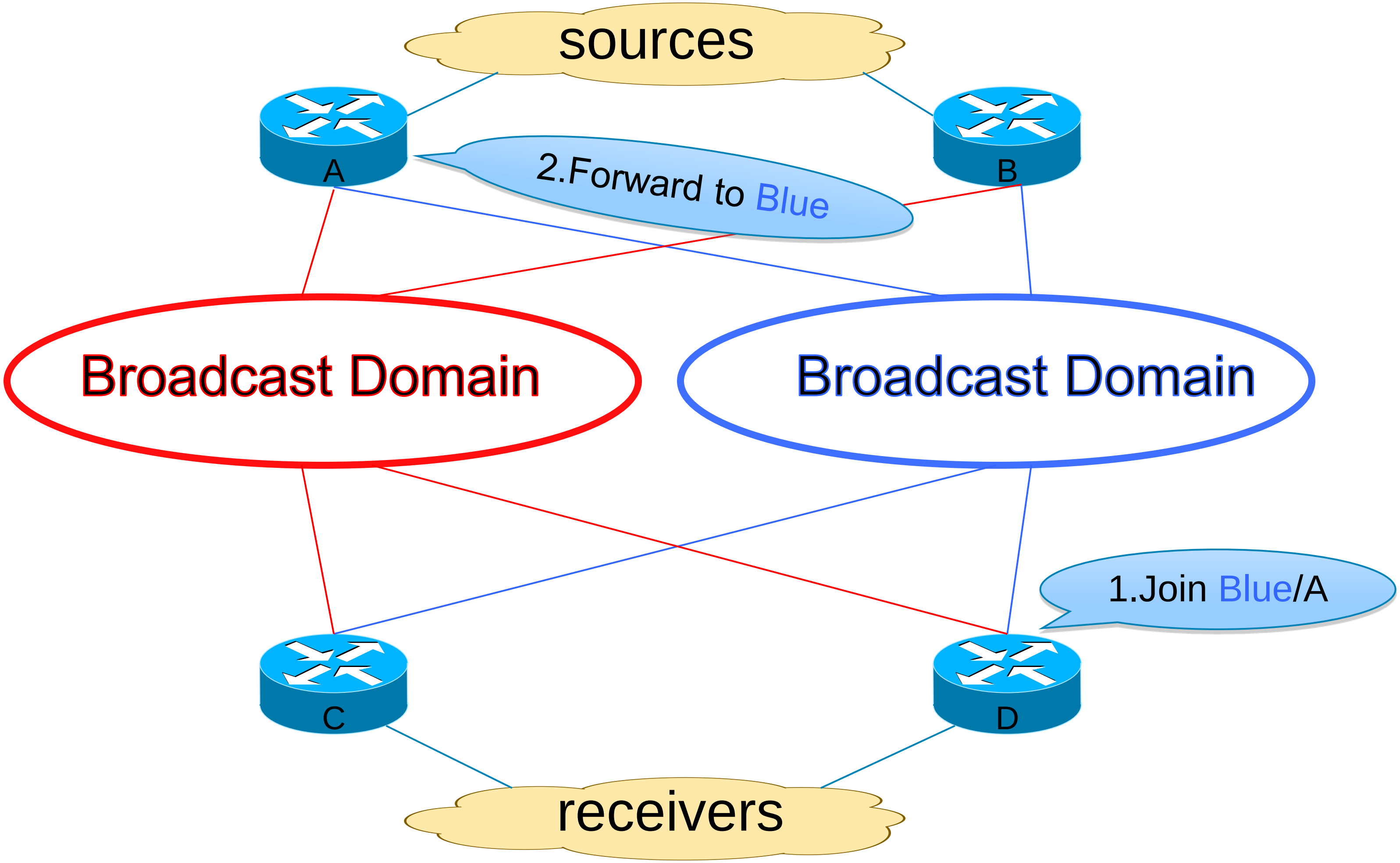
Protecting Against Duplicate Flows



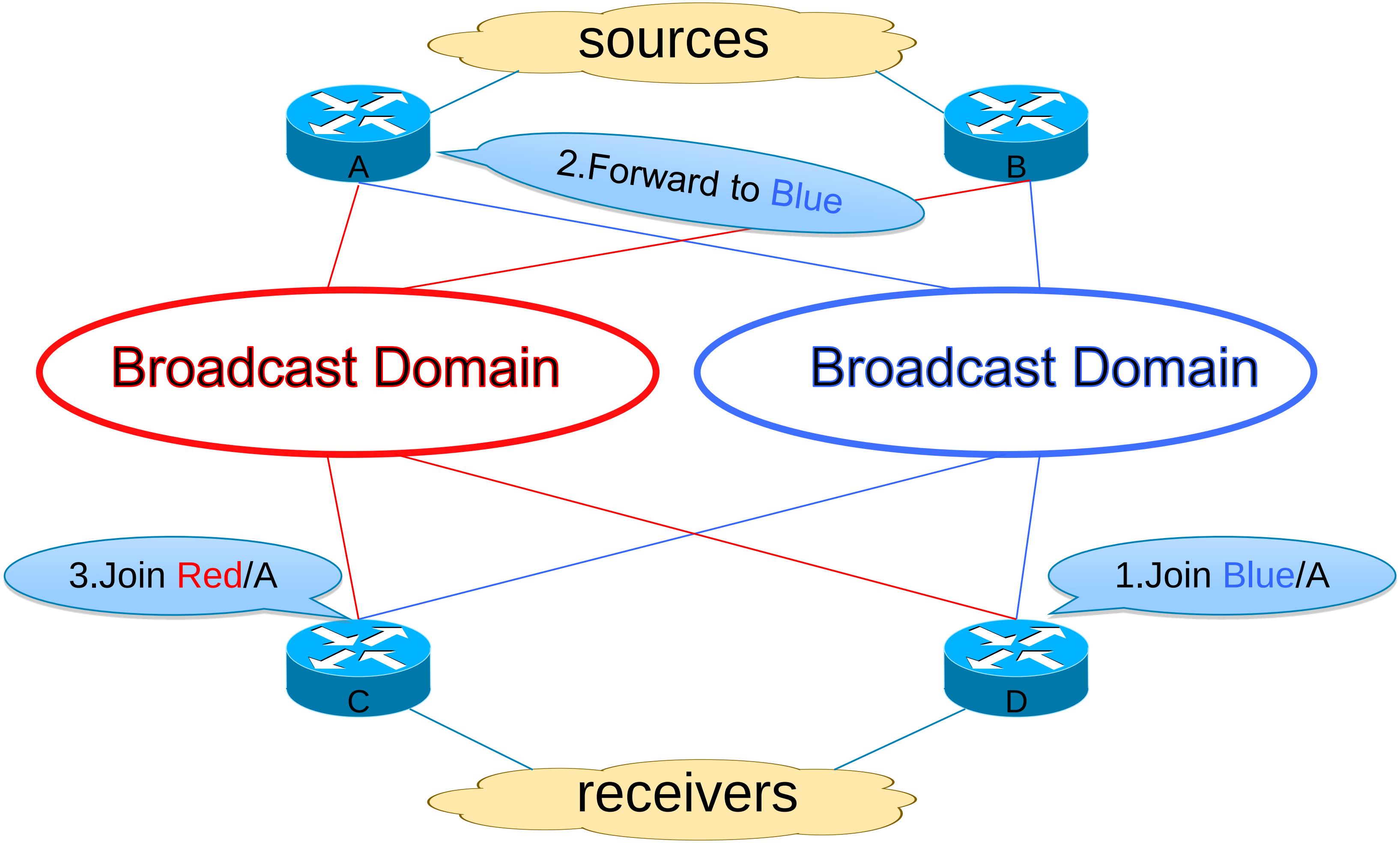
Protecting Against Duplicate Flows



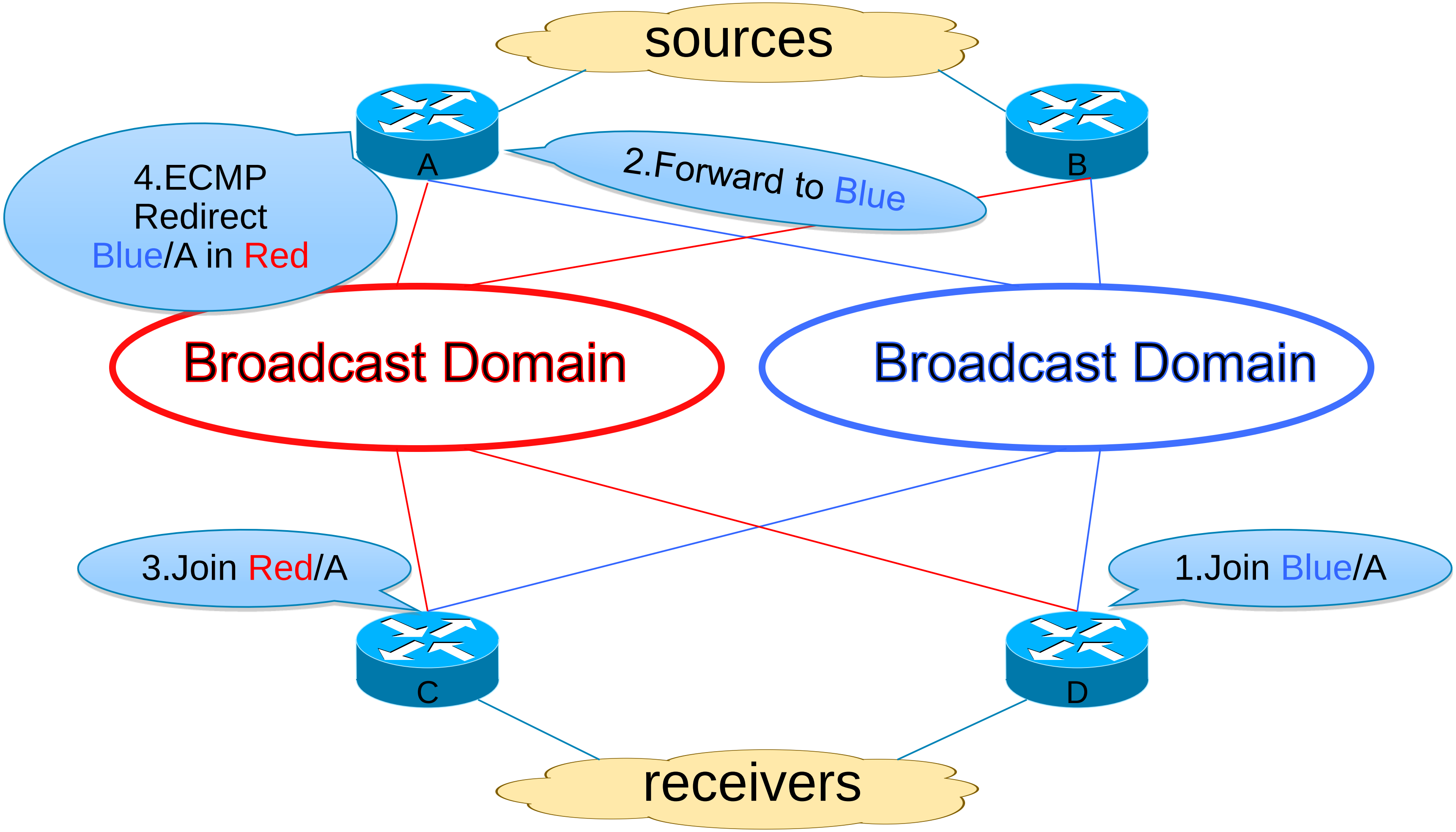
Protecting Against Duplicate Flows



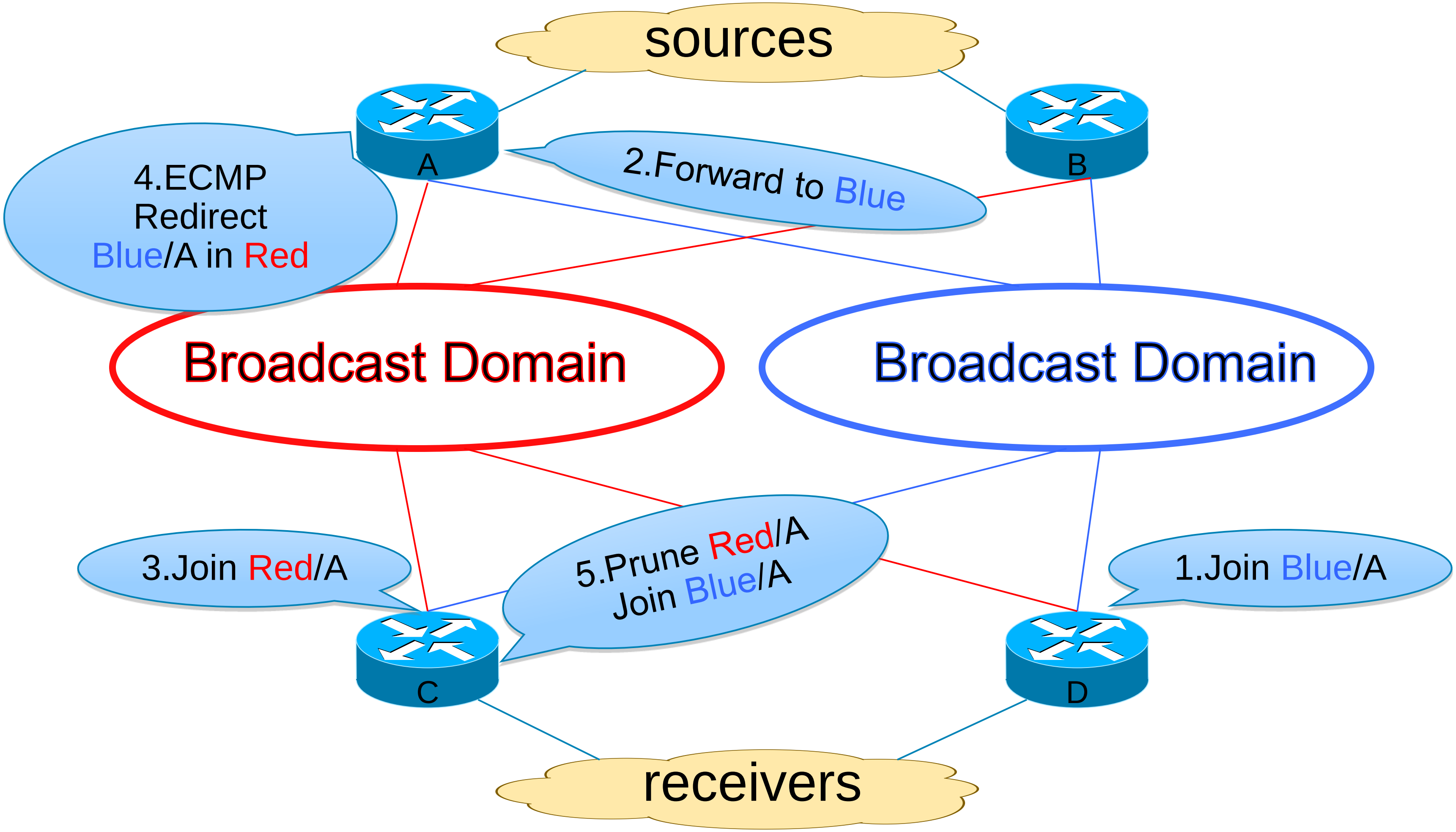
Protecting Against Duplicate Flows



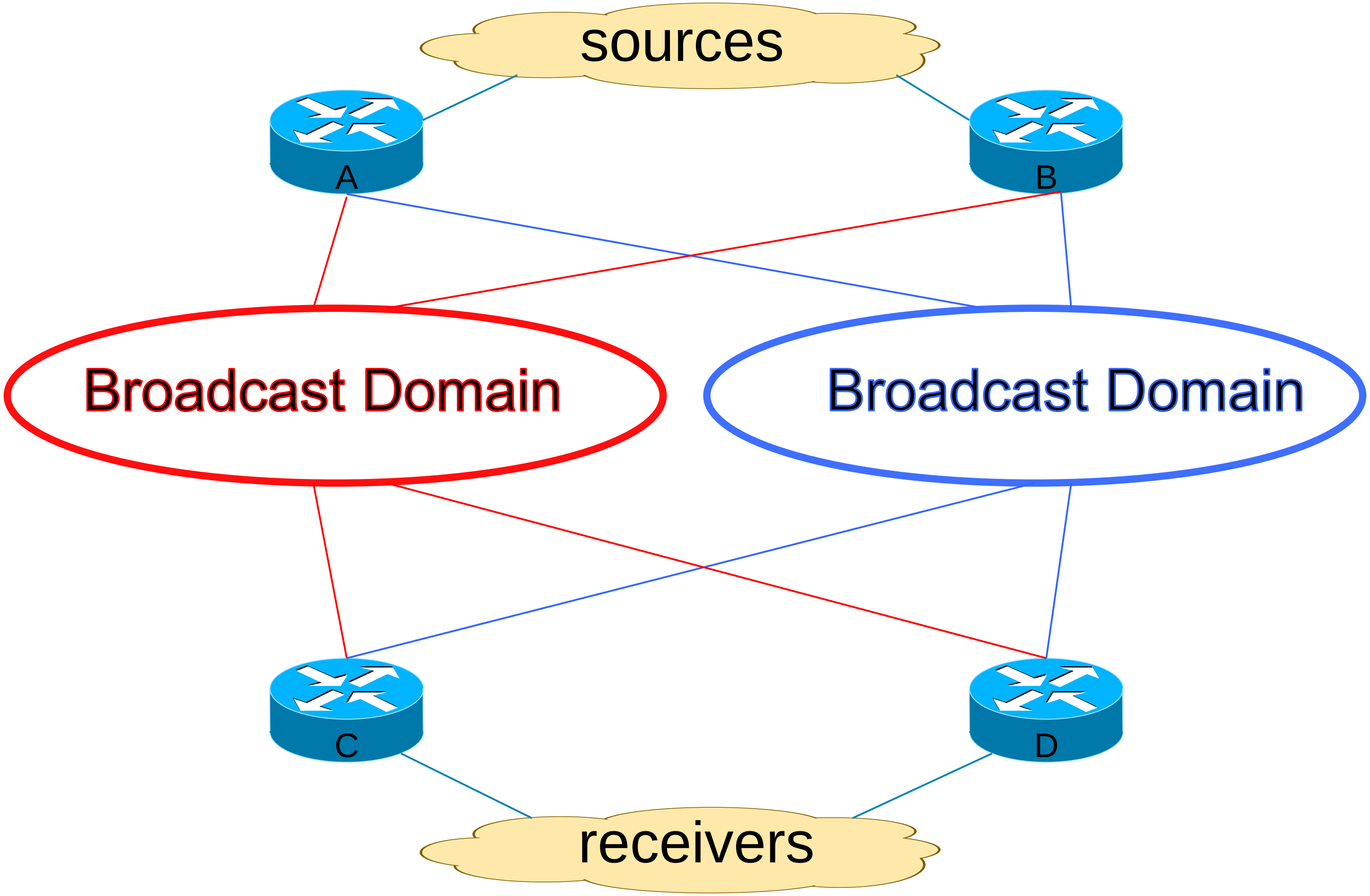
Protecting Against Duplicate Flows



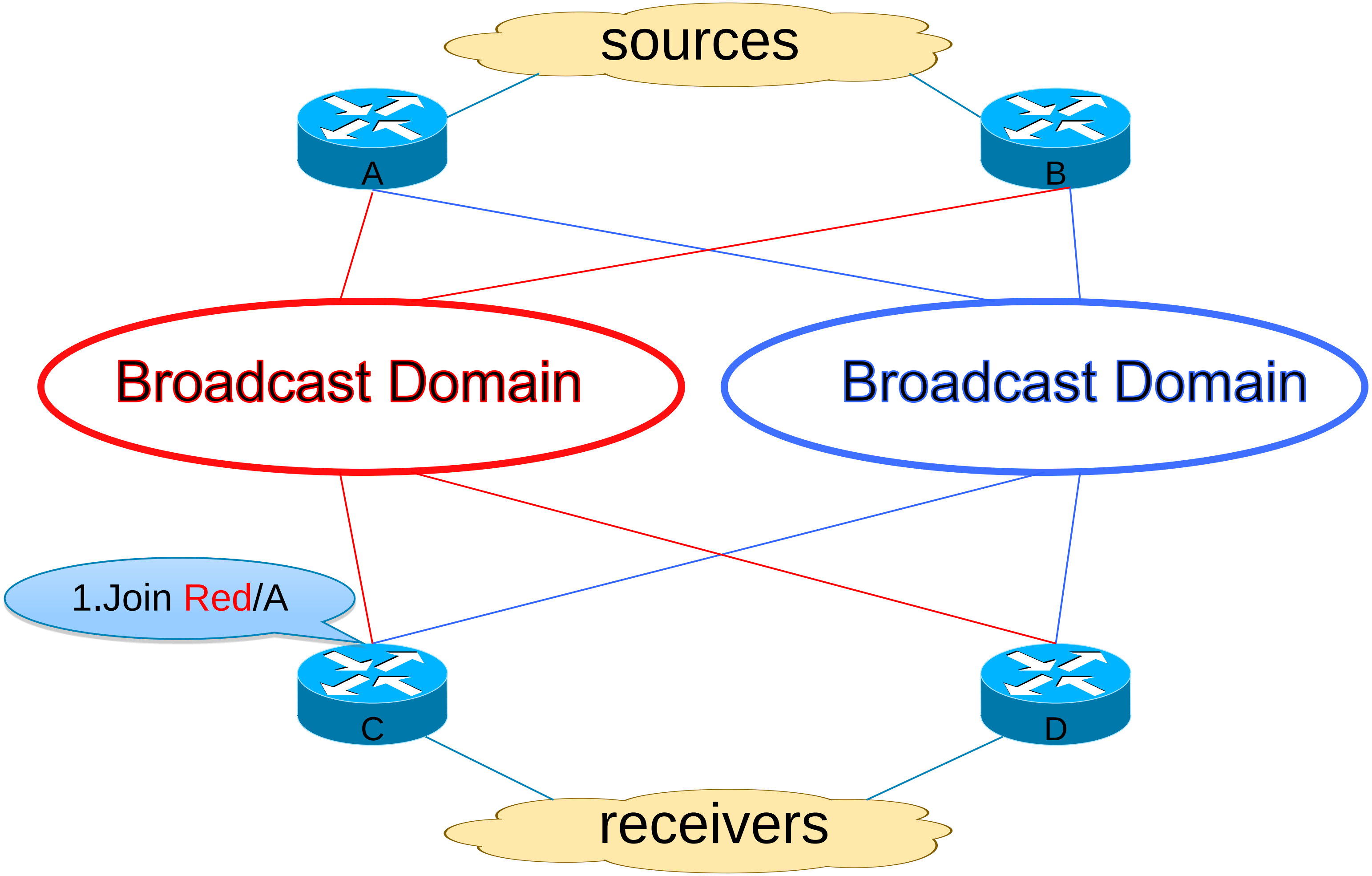
Protecting Against Duplicate Flows



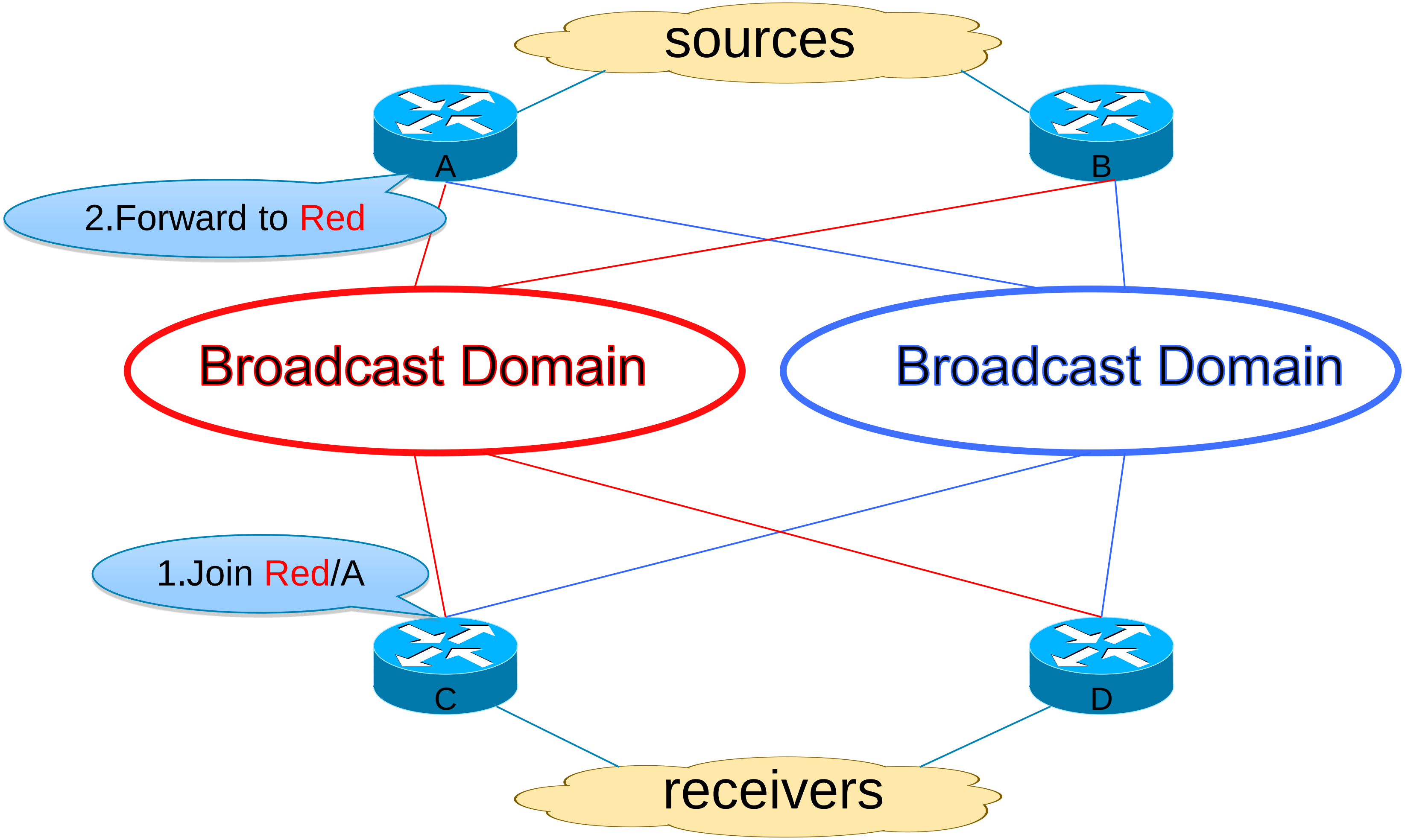
Protecting Against Over-subscription



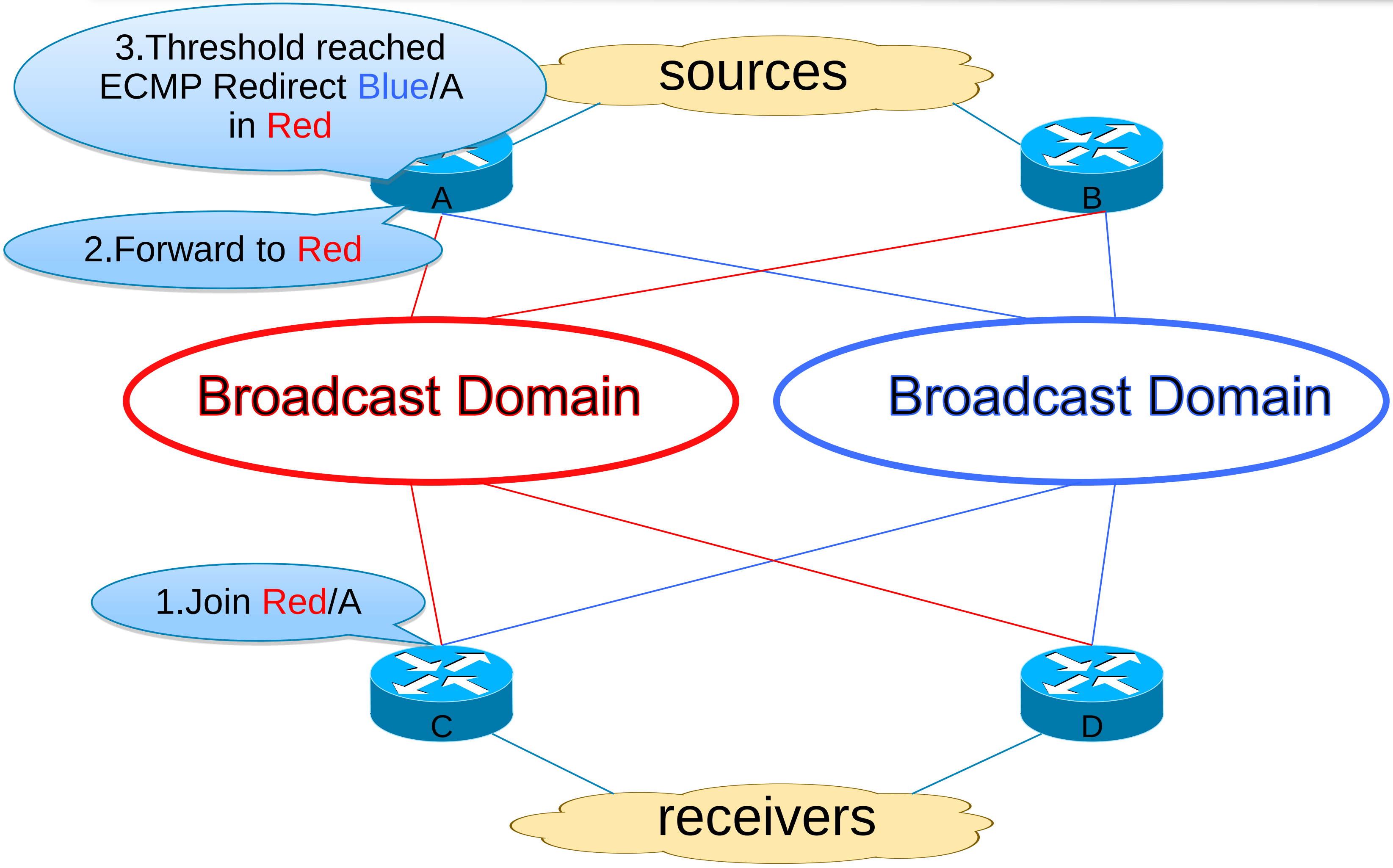
Protecting Against Over-subscription



Protecting Against Over-subscription



Protecting Against Over-subscription

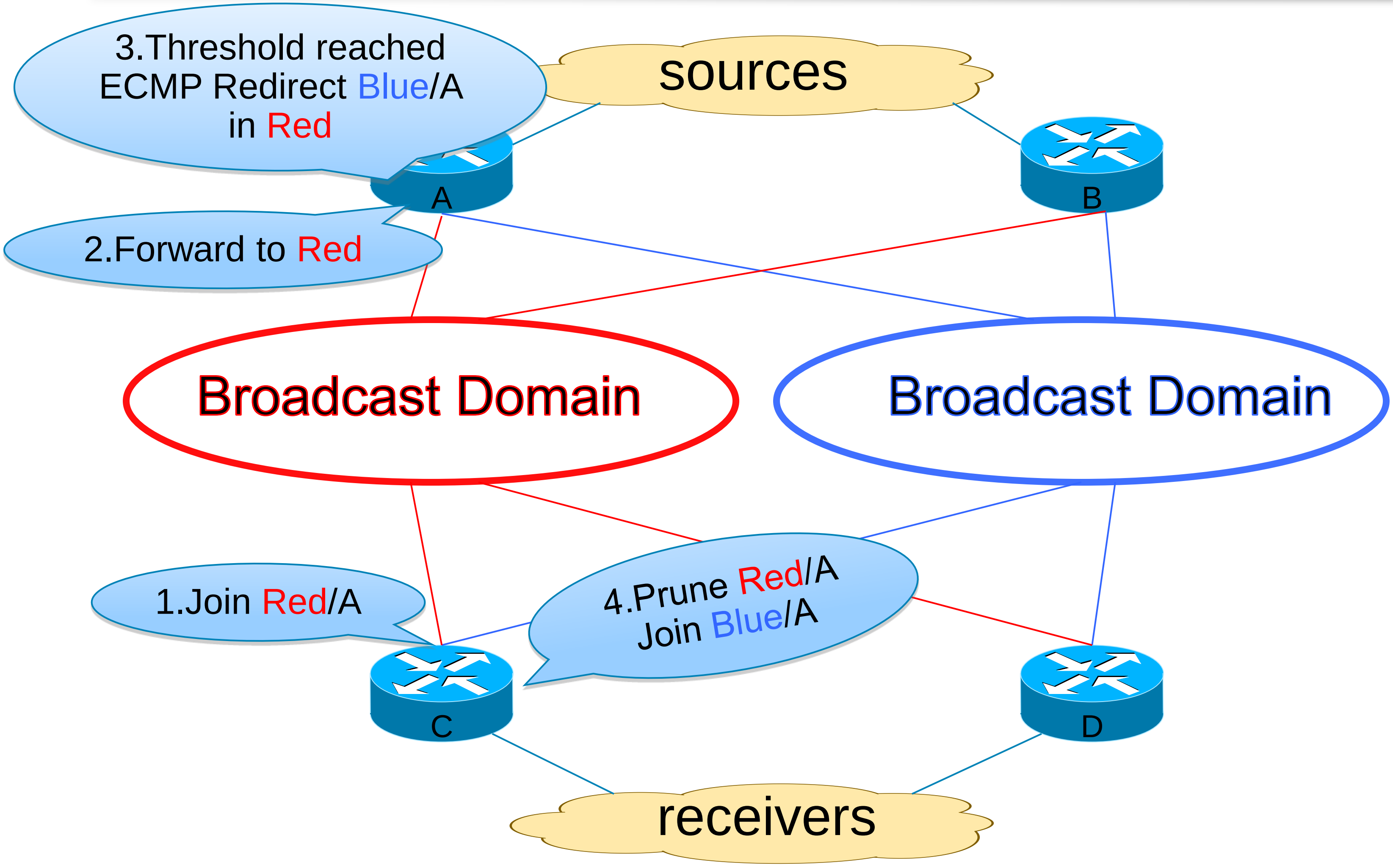


3.Threshold reached ECMP Redirect Blue/A in Red

2.Forward to Red

1.Join Red/A

Protecting Against Over-subscription



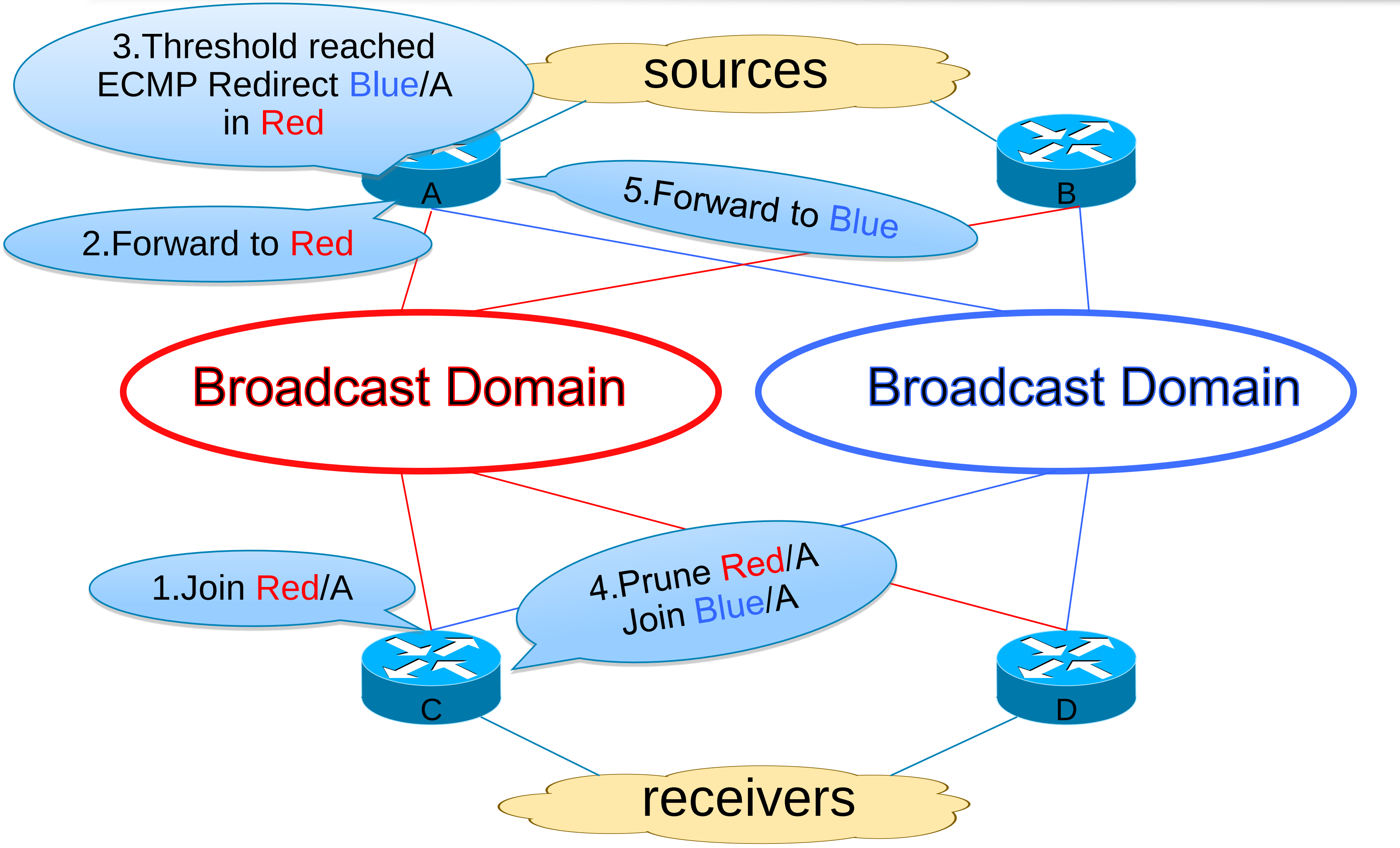
3.Threshold reached
ECMP Redirect Blue/A
in Red

2.Forward to Red

1.Join Red/A

4.Prune Red/A
Join Blue/A

Protecting Against Over-subscription



Bandwidth Aware Multicast Load Balancing Summary

- The core solution
  - Choose RPF interface/neighbor based on available bandwidth instead of address hashing
  - Use PIM ECMP Redirect to preserve bandwidth and protect against oversubscription
- Advantages of using PIM ECMP Redirect
  - Only needed when non-optimal cases happen
  - One new PDU and same PIM machinery
- Automated and more optimal load balancing for paths with same or different available bandwidth and physical bandwidths

References

- draft-hou-pim-ecmp
- draft-ietf-pim-ecmp
- RFC6754 PIM ECMP Redirect