# Border Gateway Protocol v4

Rob Sherwood

Stanford CS144

October 14, 2009

# What

- Intra-domain routing protocols (IGP)
  - Last time
  - OSPF – link state
    - IS-IS: like OSPF but not on IP
  - RIP – distance vector
- Inter-domain (EGP)
  - Today
  - Border Gateway Protocol v4
  - Path vector routing protocol: list possible paths
  - No other EGP's today... why?

# Why Inter vs. Intra?

- Why not just use OSPF everywhere?
  - e.g., hierarchies of OSPF areas

# Why Inter vs. Intra?

- Why not just use OSPF everywhere?
  - e.g., hierarchies of OSPF areas
  - Hint: scaling is not the only limitation

# Why Inter vs. Intra?

- Why not just use OSPF everywhere?
  - e.g., hierarchies of OSPF areas
  - Hint: scaling is not the only limitation
- BGP is a policy control and information hiding protocol
- intra == trusted, inter == untrusted

# Why Study BGP?

- Critical protocol: makes the Internet run
  - Only widely deployed EGP
- Active area of problems!
  - Efficiency
  - Cogent vs. Level3: Internet partition
  - Pakistan accidentally took down YouTube
  - Spammers use prefix hijacking

# Outline

- History (very briefly!)
- Function
- Properties
- Policies
- Example
- Problems and proposed solutions

# History

- Why border gateway protocol?

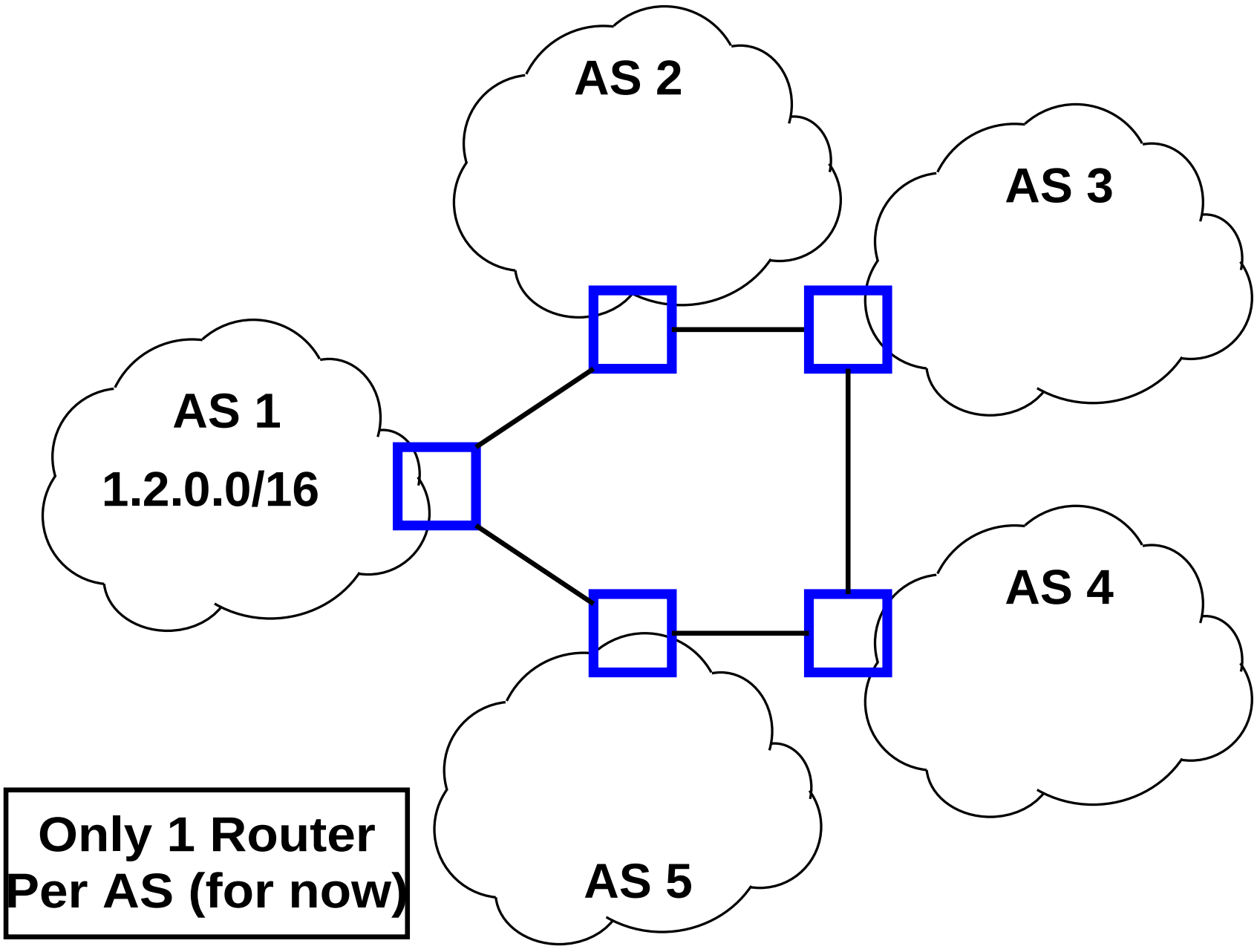
Historical distinction:

1. rfc1105 : BGPv1 1989 : "directional" routing
2. rfc1163 : BGPv2 1990 :
3. rfc1267 : BGPv3 1991
4. rfc1654 : BGPv4 (proposed) 1994
5. rfc1771 : BGPv4 (actual) 1995: CIDR support
    - rfc1772-1774 additional info

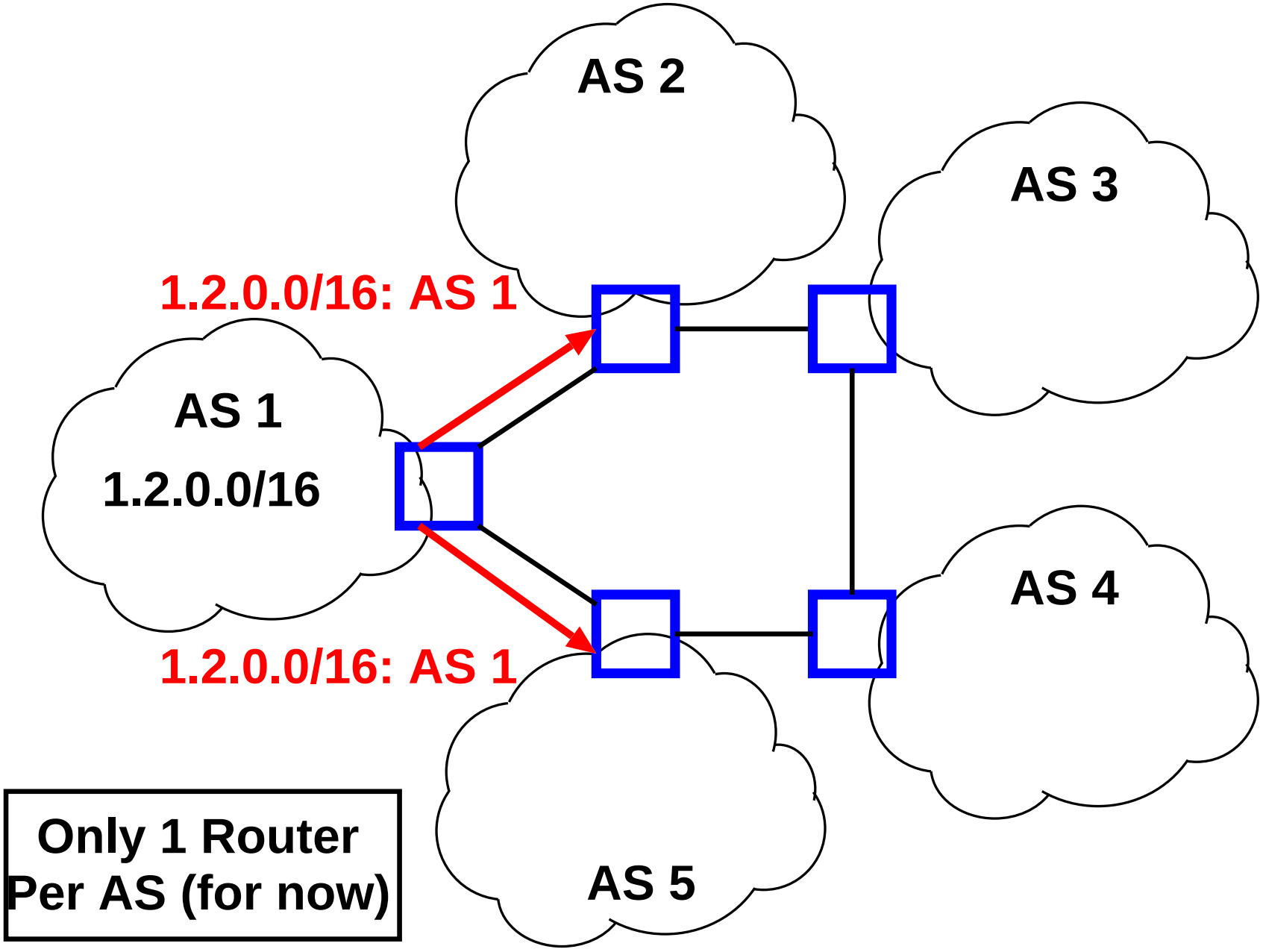
# High Level

- Abstract each AS down to a single node
- Exchange prefix-reachability with all neighbors
- "I can reach prefix 171.67.0.0/14 through AS'es 15444 3549 174 46749 32"
- Select a single path by routing policy
- Critical: learn many paths, propagate only one!
    - Add your ASN to advertised paths

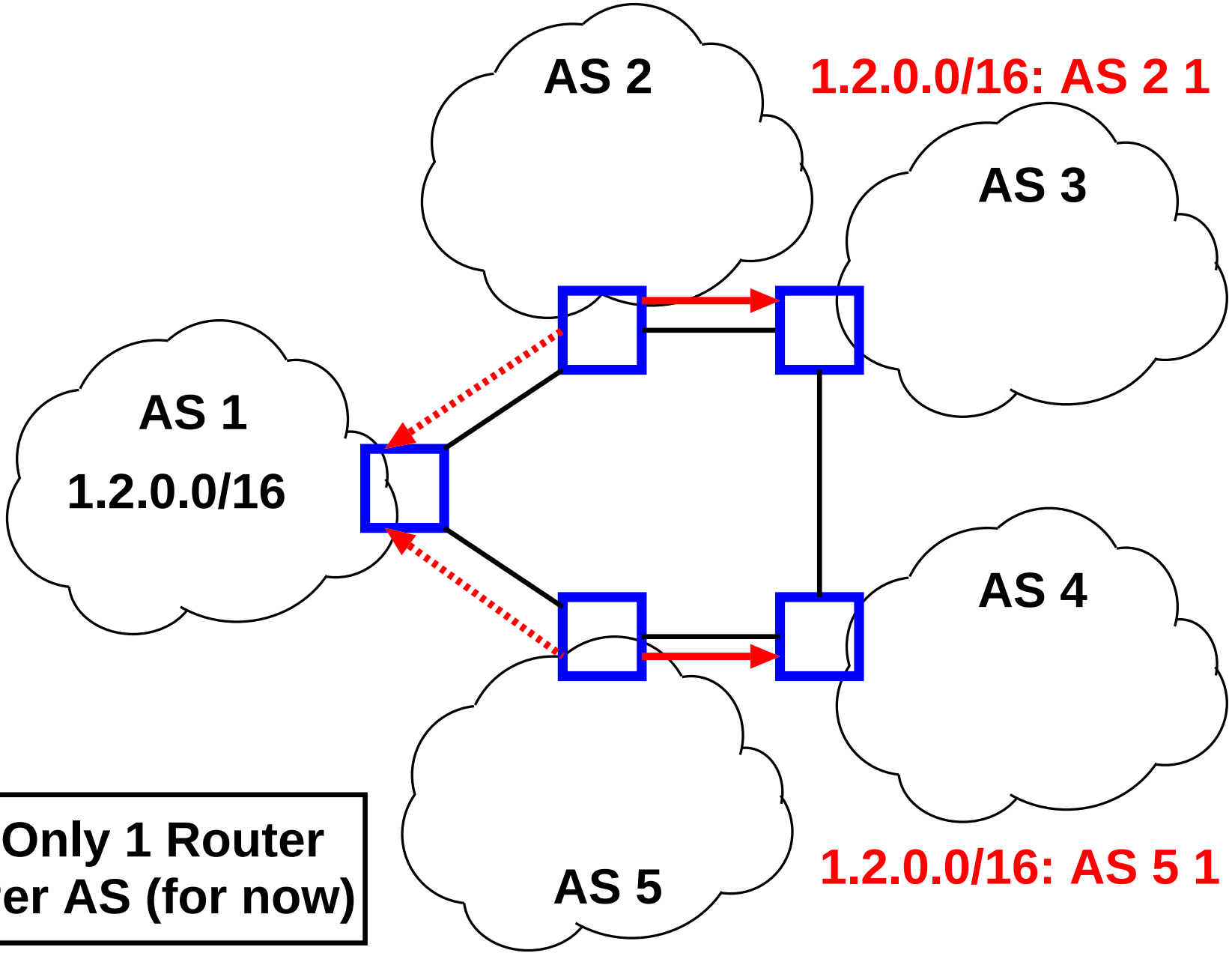
# BGP Example



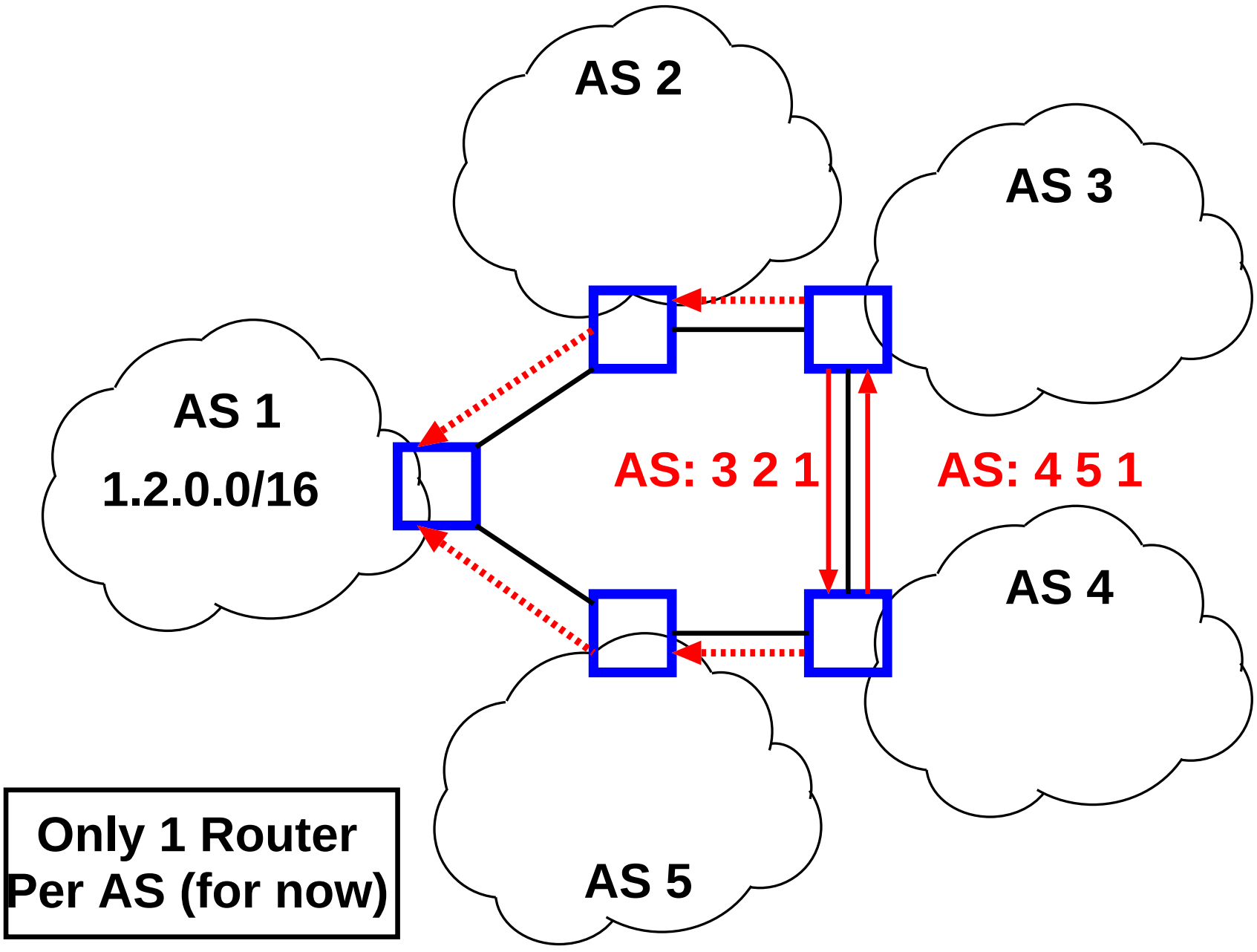
# BGP Example


AS 2
AS 3
1.2.0.0/16: AS 1
AS 1
1.2.0.0/16
AS 4
1.2.0.0/16: AS 1
AS 5
Only 1 Router
Per AS (for now)

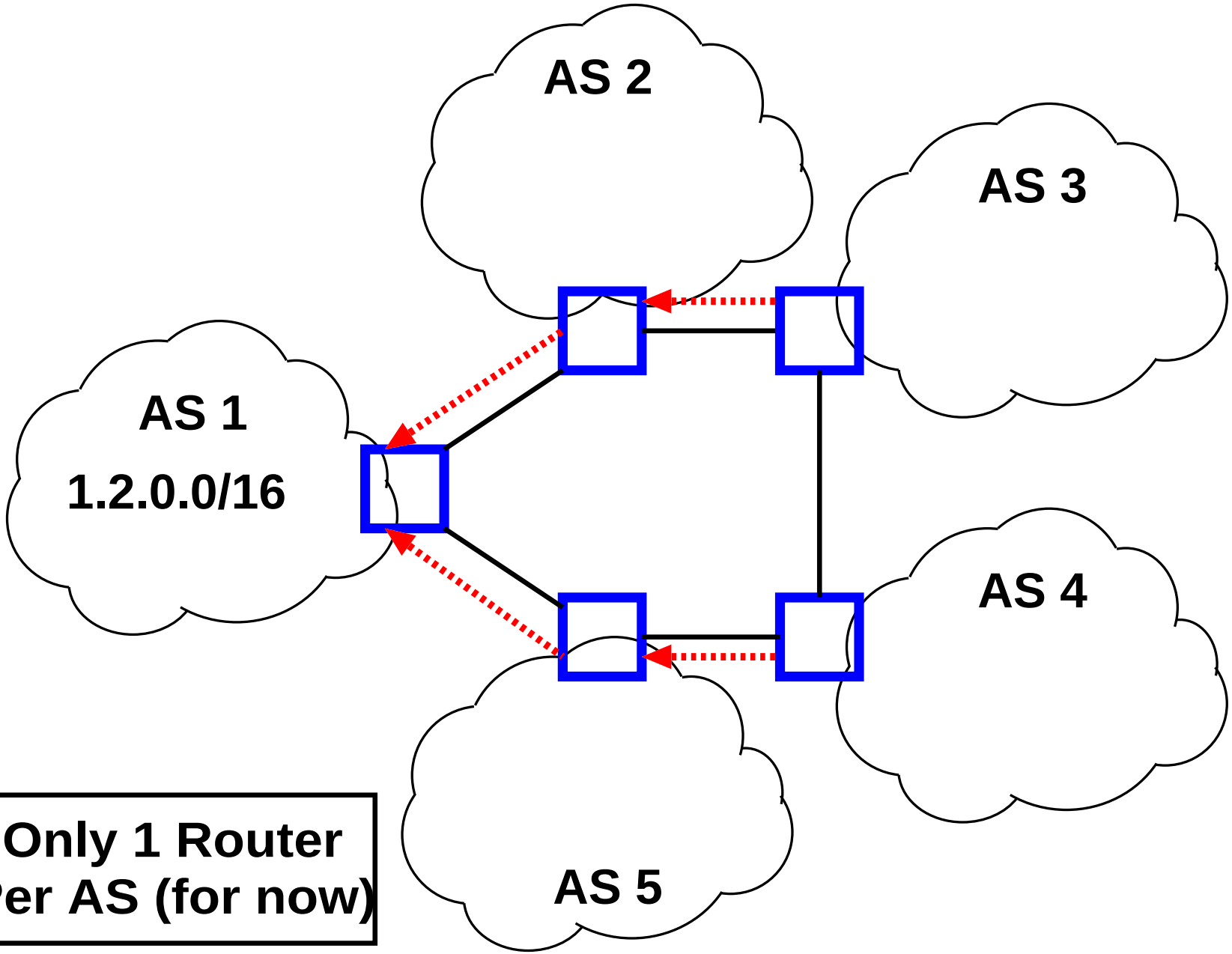
# BGP Example


AS 2
1.2.0.0/16: AS 2 1
AS 3
AS 1
1.2.0.0/16
AS 4
AS 5
1.2.0.0/16: AS 5 1
Only 1 Router
Per AS (for now)

# BGP Example


AS 2
AS 3
AS 1
1.2.0.0/16
AS: 3 2 1
AS: 4 5 1
AS 4
AS 5
Only 1 Router
Per AS (for now)

# BGP Example



# BGP Implications

- Explicit AS path == loop free!
  - Except under churn, IGP/EGP mismatch, etc.
- Not all ASes know all paths
- AS abstraction – loss of efficiency
- Shortest AS path not guaranteed
- Scaling
  - 32K ASes
  - 300K+ prefixes

# Transport Details

1. Border routers must directly connect
2. Connect tcp port 179
3. Negotiate features
4. Full information exchange – expensive!
5. Exchange periodic updates indefinitely

Session resets are expensive (both in CPU and to the entire network!) and should be avoided.

# Advertisements

- Destination prefix: 171.67.0.0/14
- AS Path: ASN 15444 3549 174 46749 32
- Next Hop IP: just like in RIPv2
- Knobs for traffic engineering
  - Metric, Weight, LocalPath, MED, Communities
  - Lots of voodoo

# Getting Your Hands Dirty

RouteViews Project:
http://www.routeviews.org/

1. telnet route-views.linx.routeviews.org
2. show ip bgp 171.67.0.0/14 longer-prefixes
   - note that all paths are learned internally
   - not a production device

# Route Selection 1/2

1. Next-Hop reachable?
2. Prefer highest weight
3. Prefer highest local-pref
4. Prefer locally originated routes
5. Prefer routes with shortest AS path length

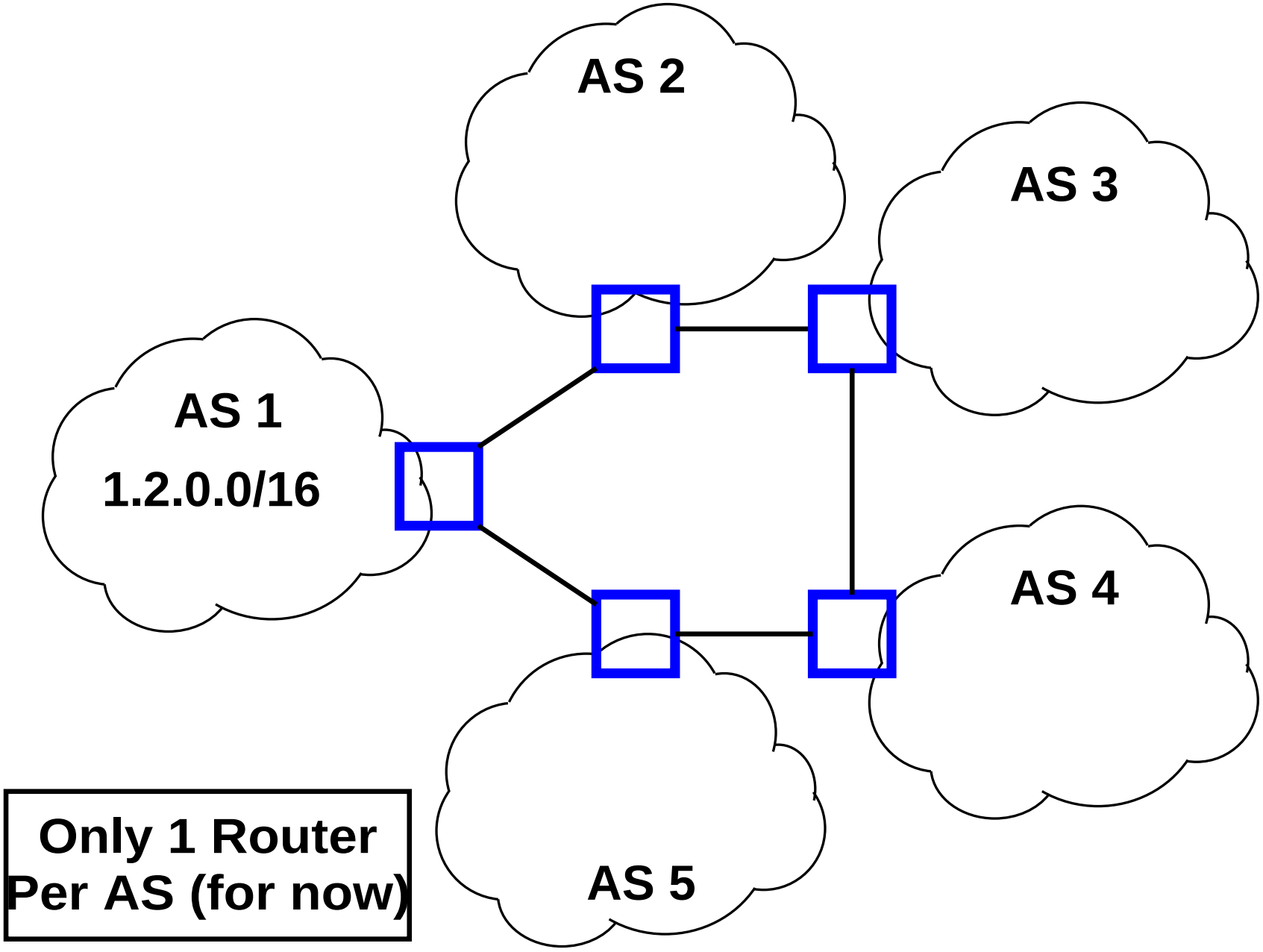
# Route Selection 2/2

6. Prefer path with lowest origin type
7. Prefer route with lowest MED value
8. Prefer eBGP over iBGP
9. Prefer routes with lowest cost to egress point
   - hot-potato routing
10. Tie-braking rules
    - e.g., lowest router-id, oldest route

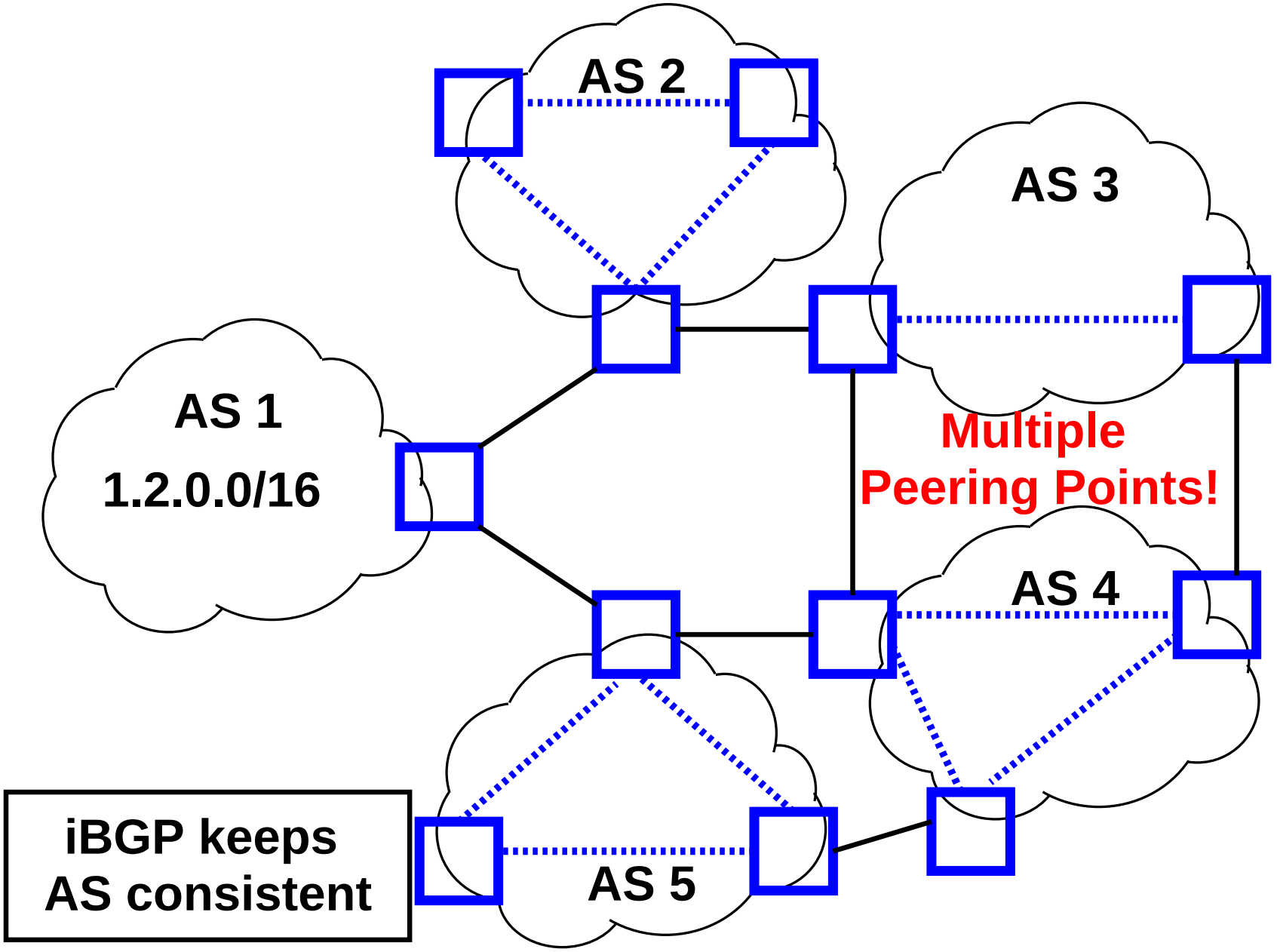
# Revisit RouteViews Data

- Why was that route selected?
- Why are there two routes to Stanford?

# External vs. Internal BGP



# External vs. Internal BGP



# BGP Relationships 1/2

Customer/Provider:

- Customers pay for connectivity
- e.g., Stanford pays Cogent
- Customer is a stub, provider is a transit
  - Amount and cost structure can vary wildly
- Many customers are multi-homed
  - Stanford also connects to Calren/Internet2
- Typical policy: prefer routes from customers

# BGP Relationships 2/2

Peers:

- ASes agree to exchange traffic for free
  - Penalties/renegotiate if imbalance
- Tier 1 ISPs have no default route: all peer with each other
- You are Tier $i + 1$ if you have a default route to a Tier $i$

# BGP Relationship Drama

Cogent vs. Level3

- http://www.isp-planet.com/business/2005/cogent_level_3.html
- Level3 and Cogent were peers
- In 2005, Level3 decided to start charging Cogent
- Cogent said No
- Internet partition: Cogent's customers couldn't get to Level3's customers
  - other ISPs were affected as well
- They came to a new, undisclosed agreement 3 weeks later

# BGP Problems and Solutions

1. Security
2. Convergence
3. Scaling (route reflectors)
4. Traffic engineering - AS preprending
5. Multiple stable solutions - BGP "Wedgies"

# BGP Security

- Anyone can source a prefix announcement
  - BGP is not very secure :-(
- YouTube's prefix is 208.65.152.0/22
- To block YouTube (by government directive), a PieNET advertised 208.65.152.0/23 and 208.65.152.128/23 (longest prefix match)
- Spammers steal unused IP space to hide

Secure BGP is currently being deployed

# BGP Convergence

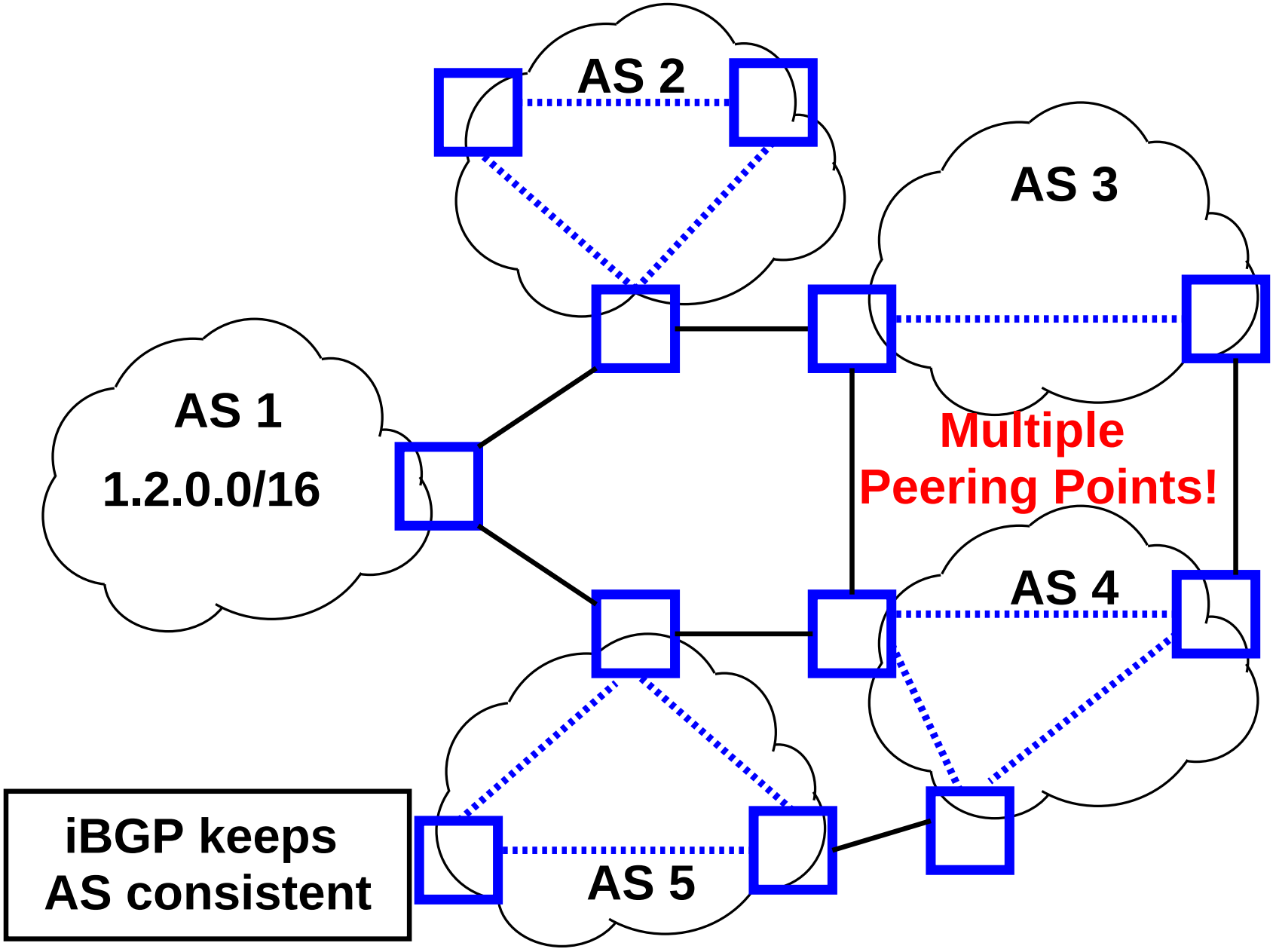
Given a change, how long until the network re-stabilizes?

- ... depends on the change: sometimes never.
- Open research problem: "tweak and pray"
- Distributed setting is challenging

Easier: does there exist a stable configuration?

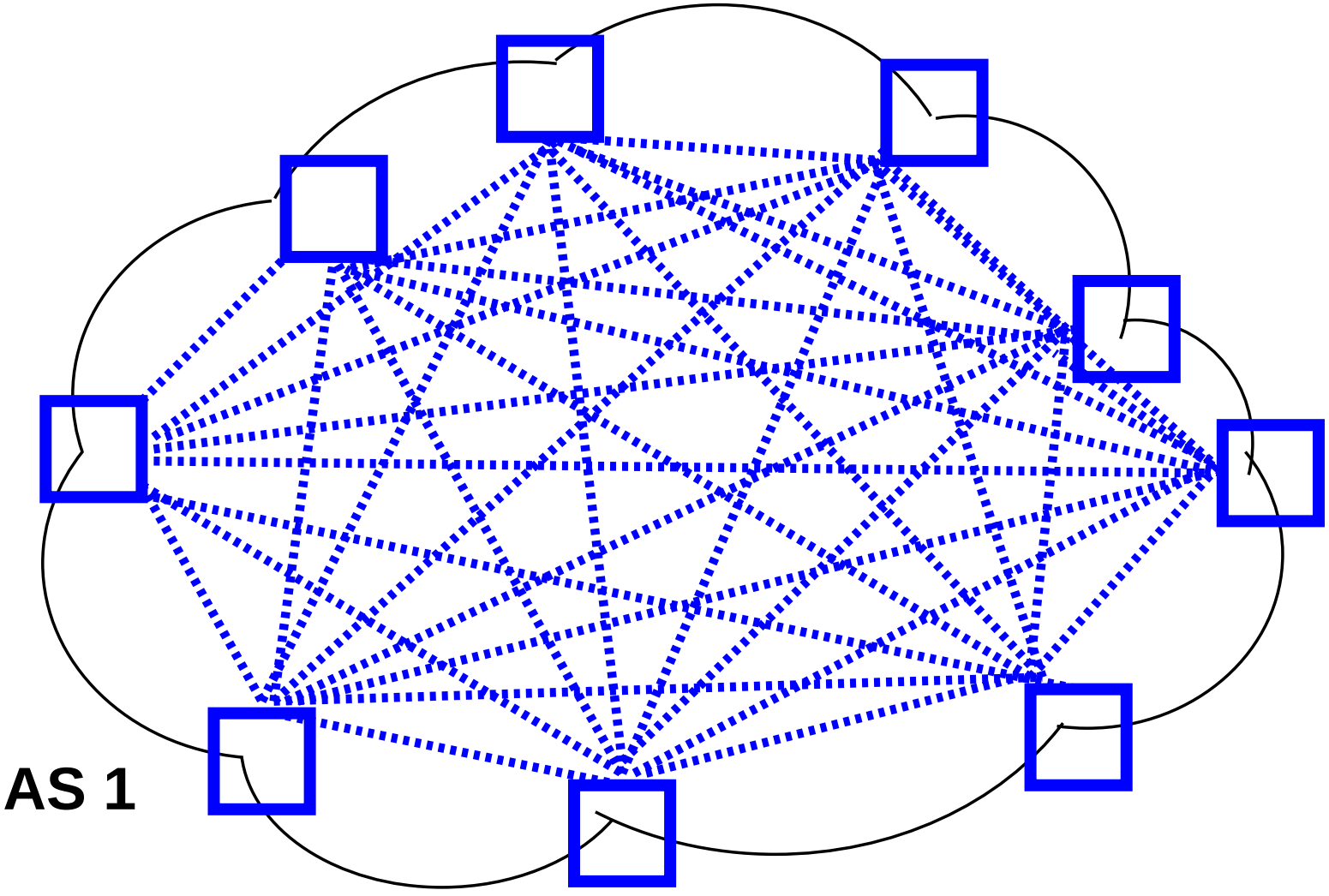
- Distributed: open research problem
- Centralized: NP-Complete problem! [Griffin-Sigcomm99]

# Scaling iBGP: Route Reflectors



# Scaling iBGP: Route Reflectors

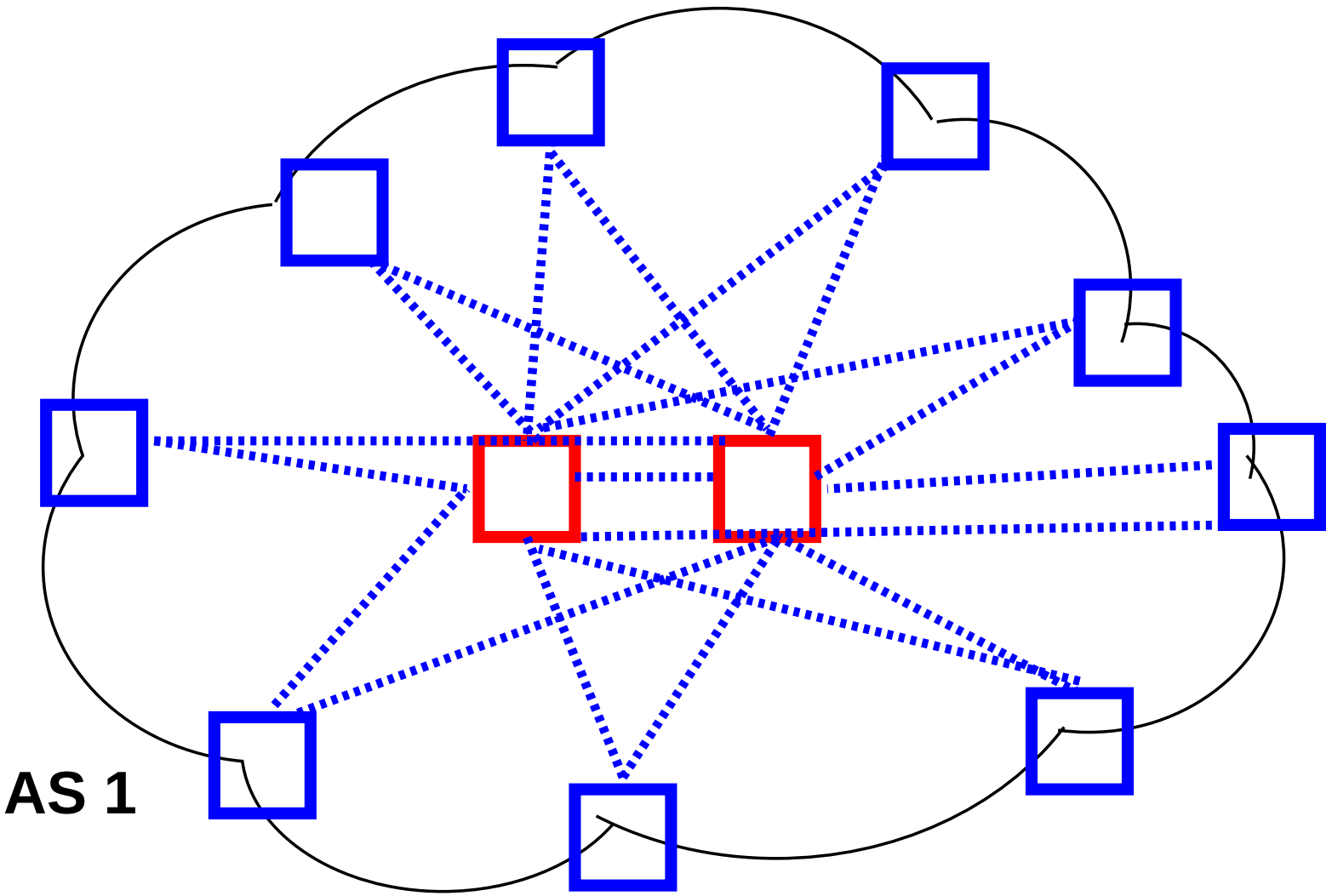
**iBGP Mesh == O(n^2) mess**



# Scaling iBGP: Route Reflectors

**Solution: Route Reflectors**

**O(n*k)**



# Traffic Engineering

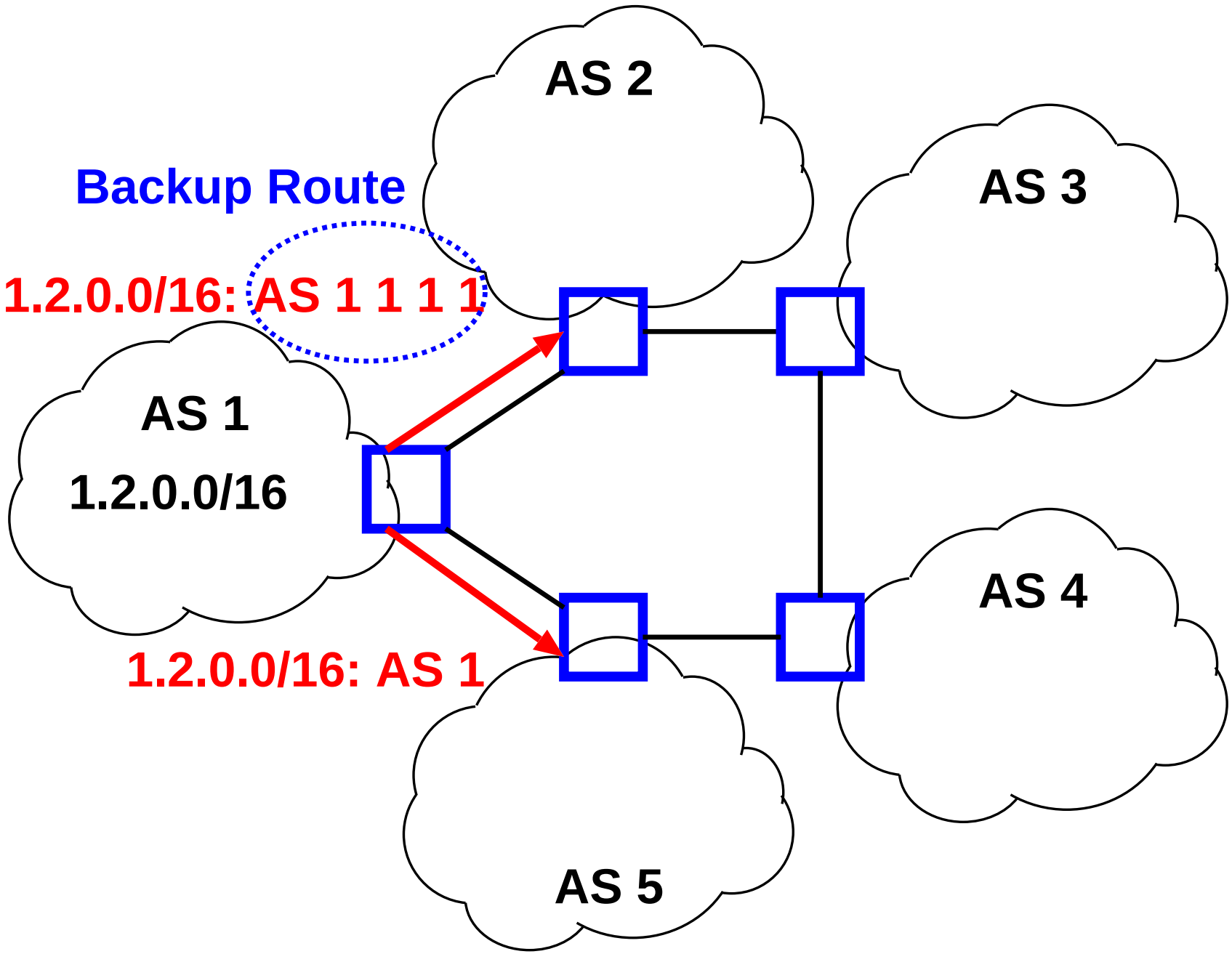
- "Route-map" programs to set weights
- Route filtering: input and output
- More specific routes: longest prefix
- AS prepending: "32 32 32 32"
- Imprecise science

# rfc4264: BGP Wedgies

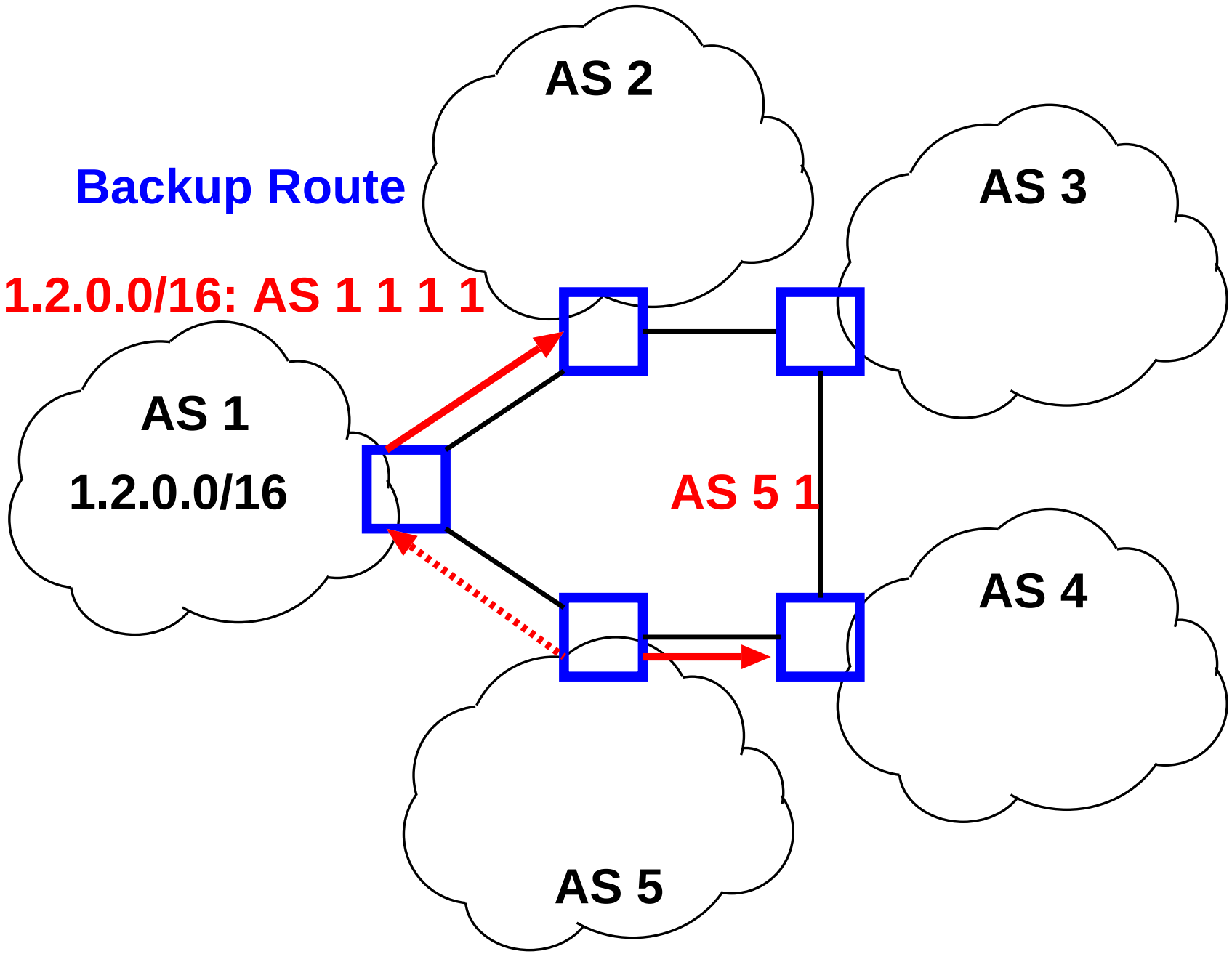
A Common config:

- Prefer customer routes over non-customer
- Then prefer shortest AS path

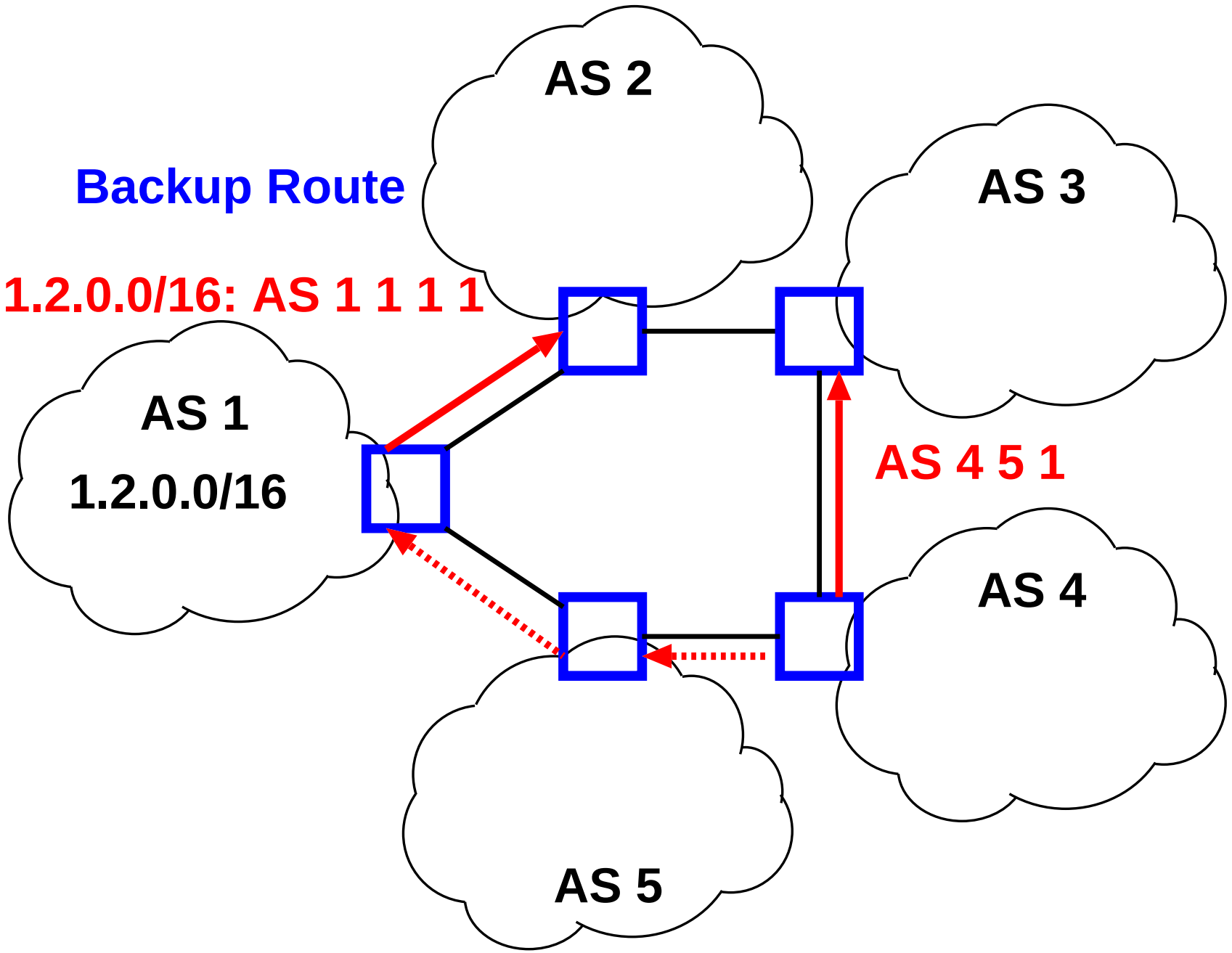
# rfc4264: BGP Wedgies



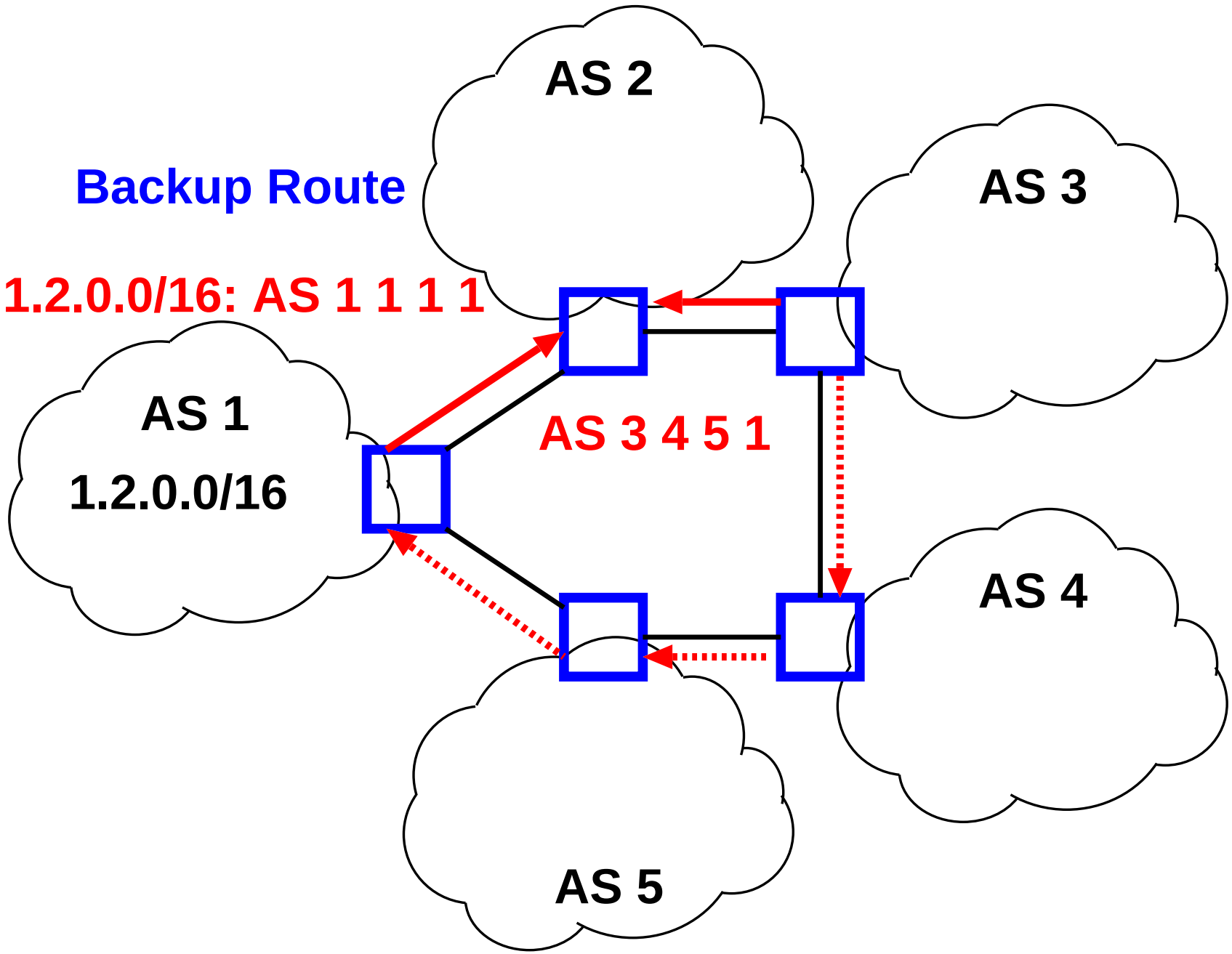
# rfc4264: BGP Wedgies



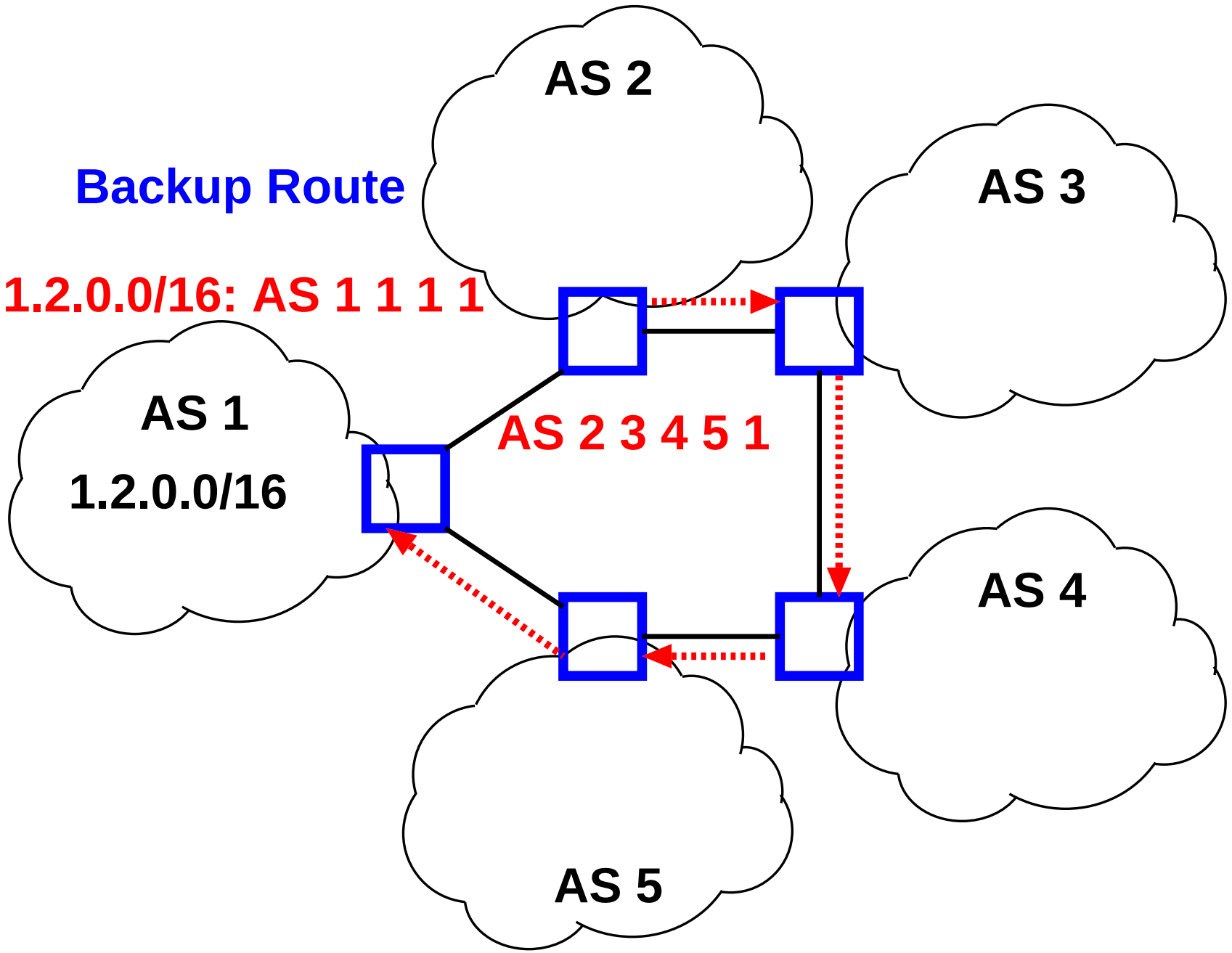
# rfc4264: BGP Wedgies



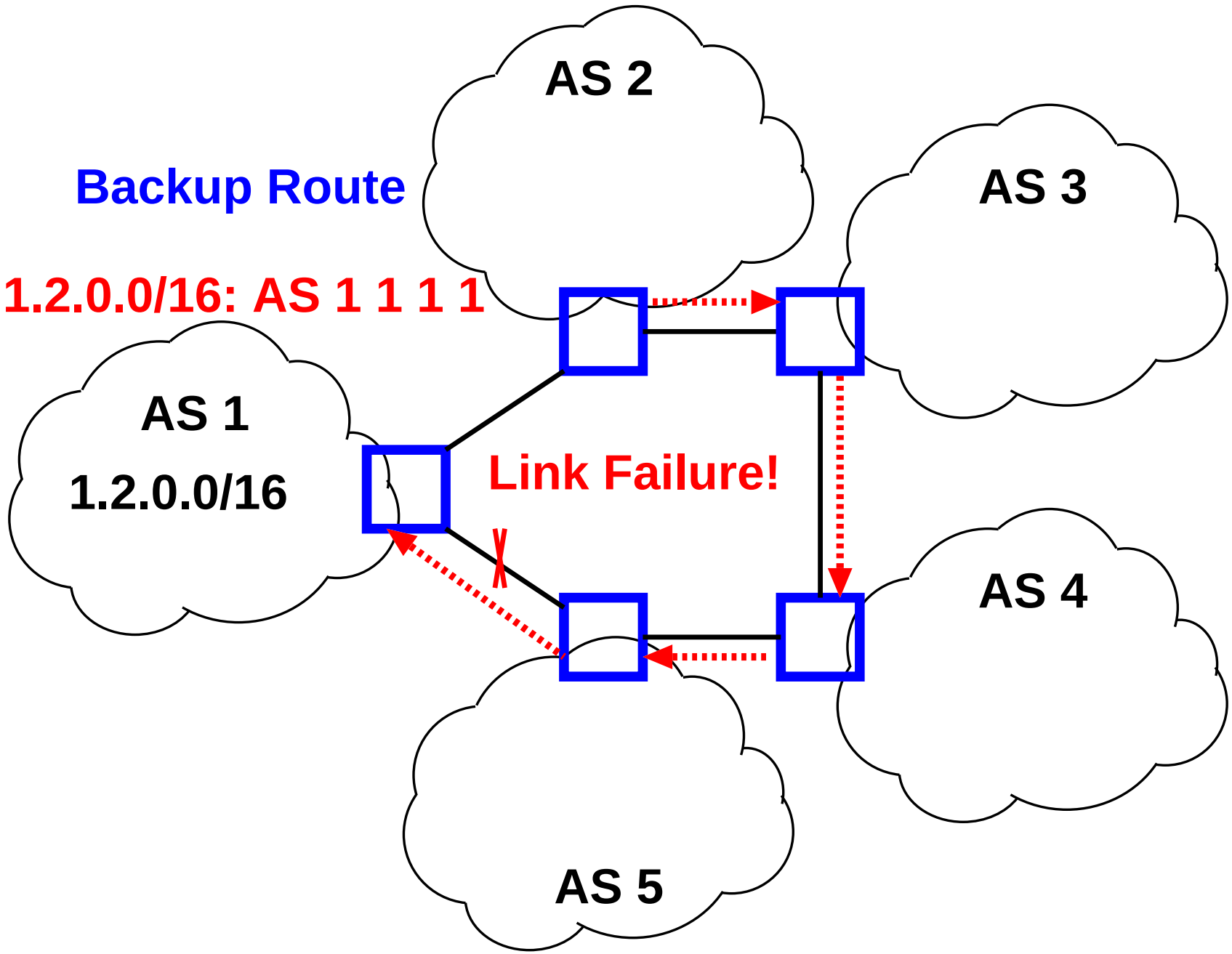
# rfc4264: BGP Wedgies



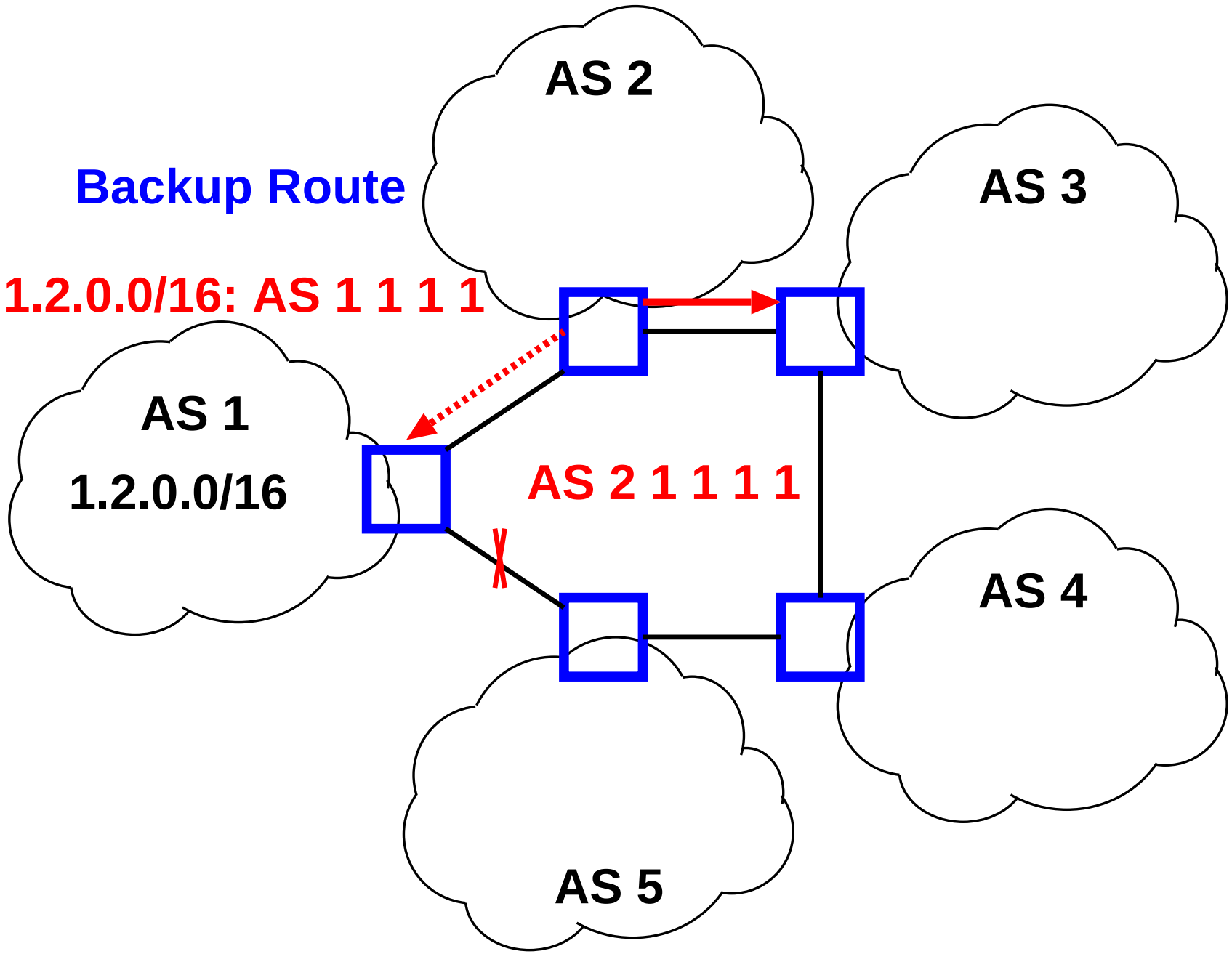
# rfc4264: BGP Wedgies



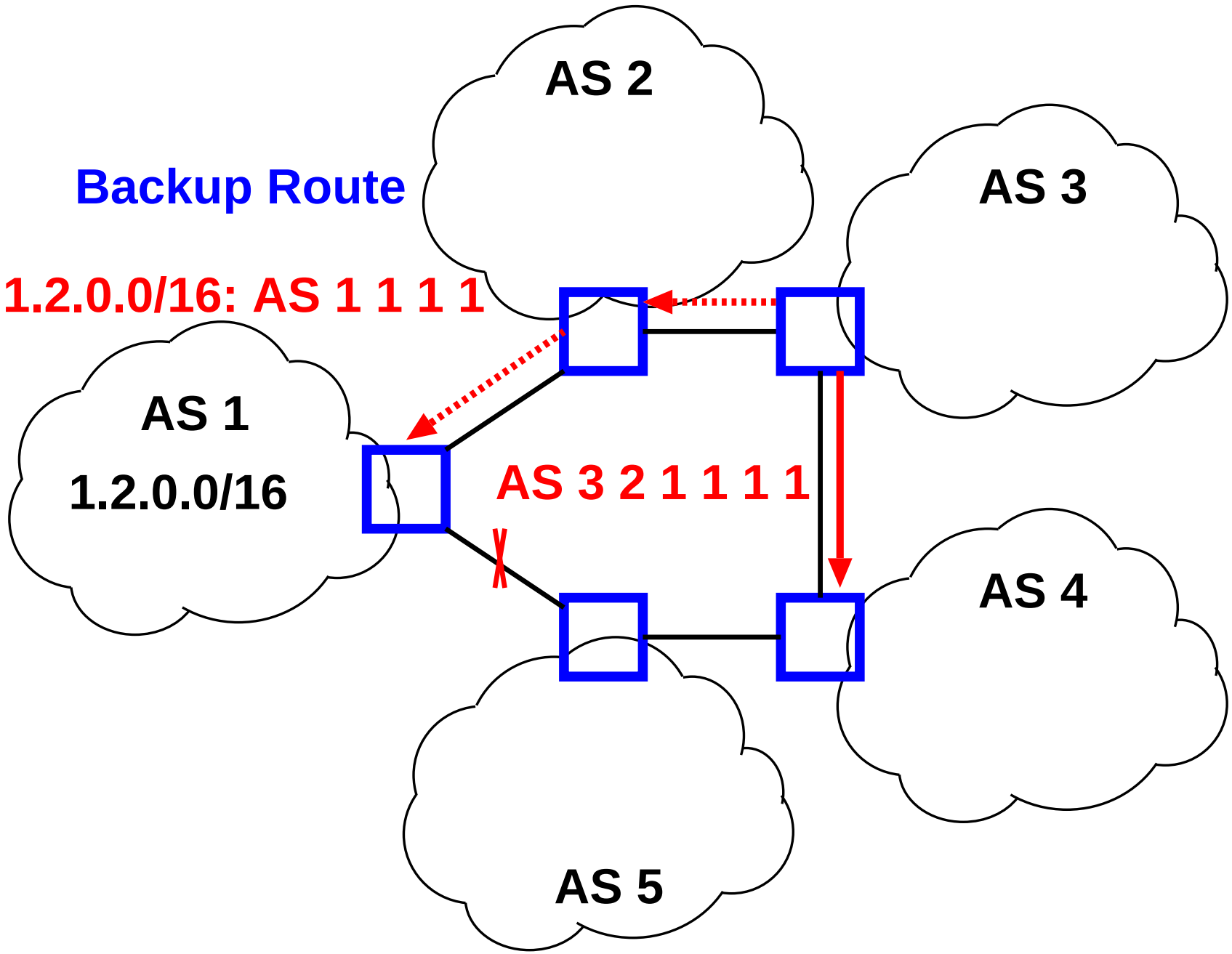
# rfc4264: BGP Wedgies



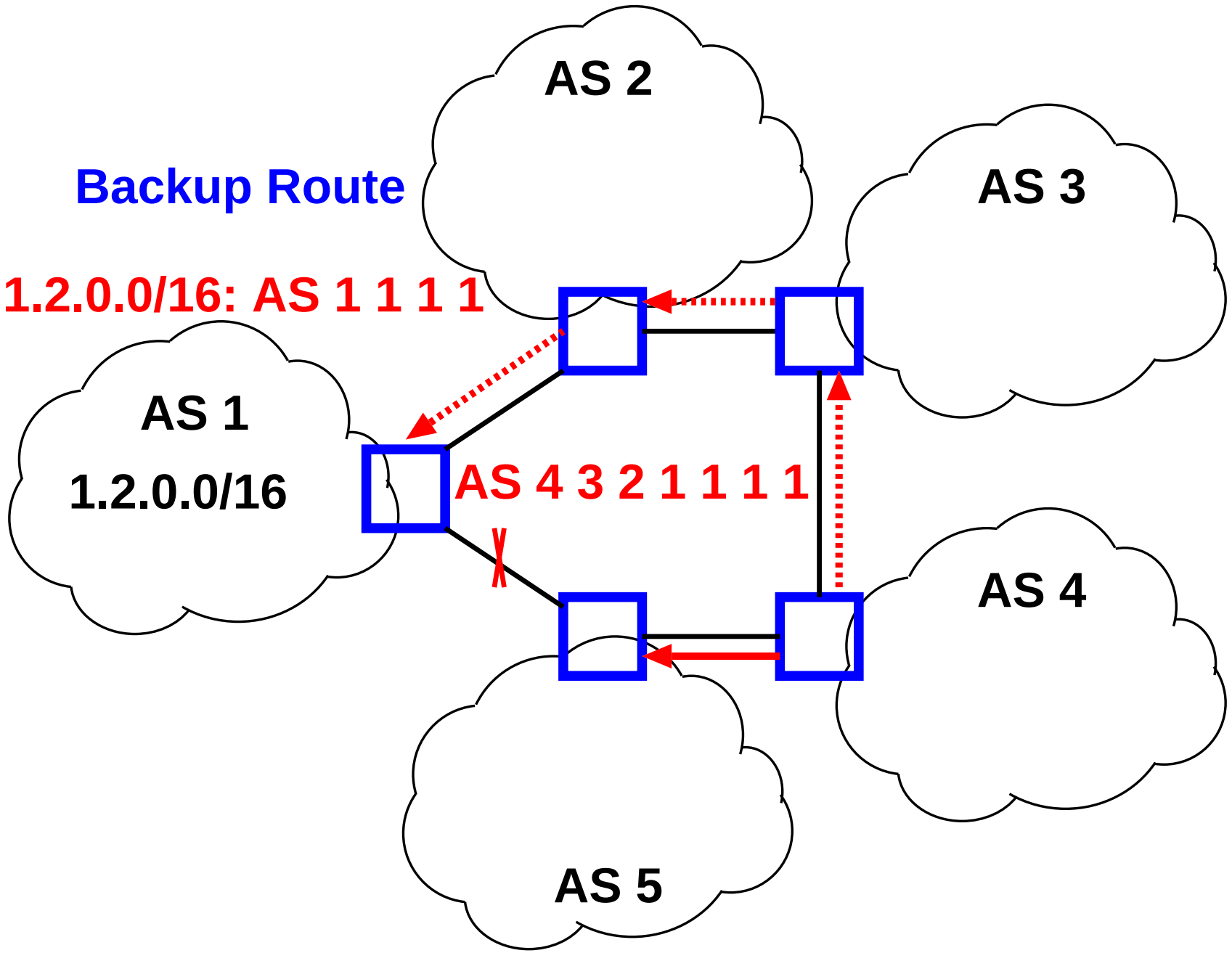
# rfc4264: BGP Wedgies



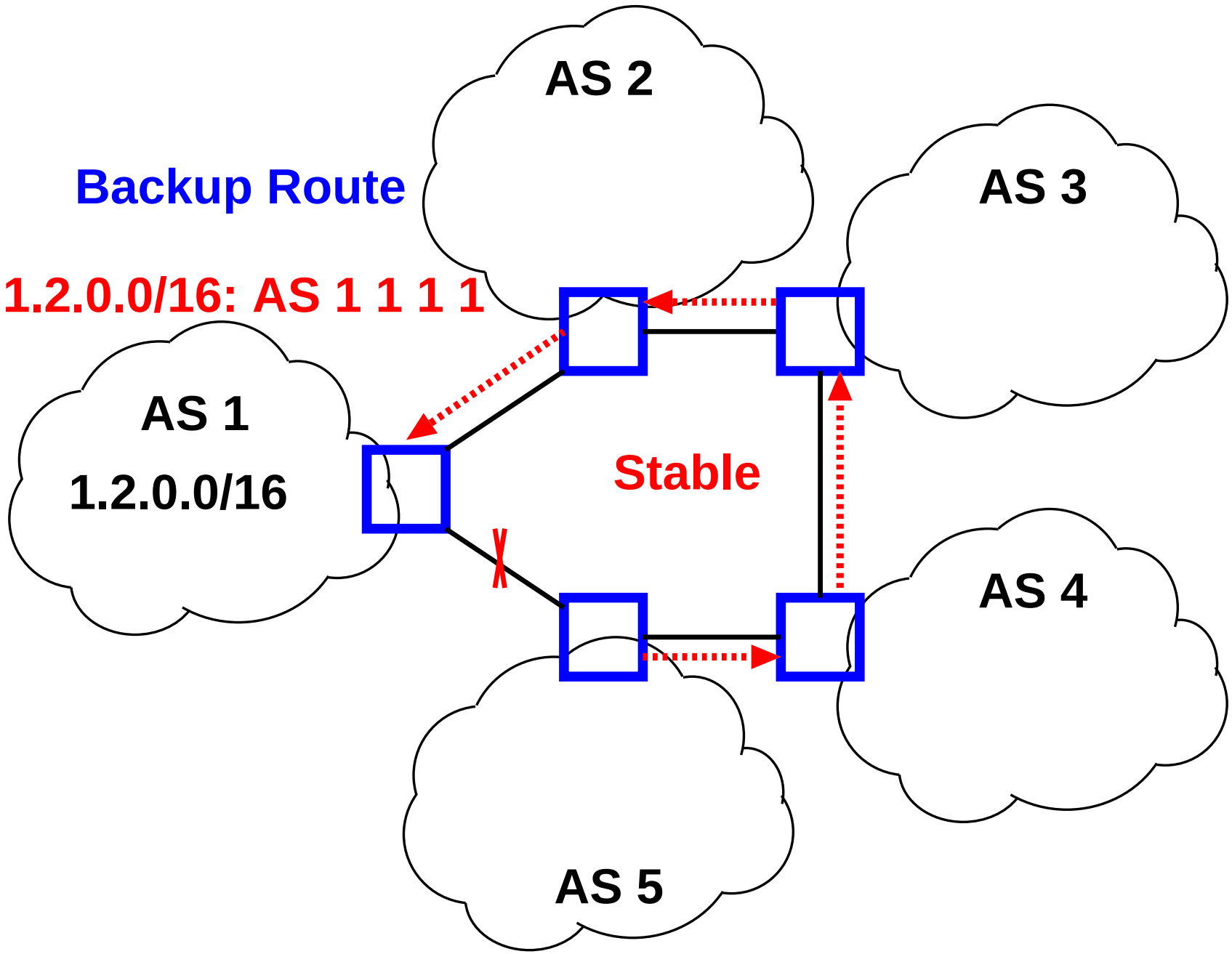
# rfc4264: BGP Wedgies



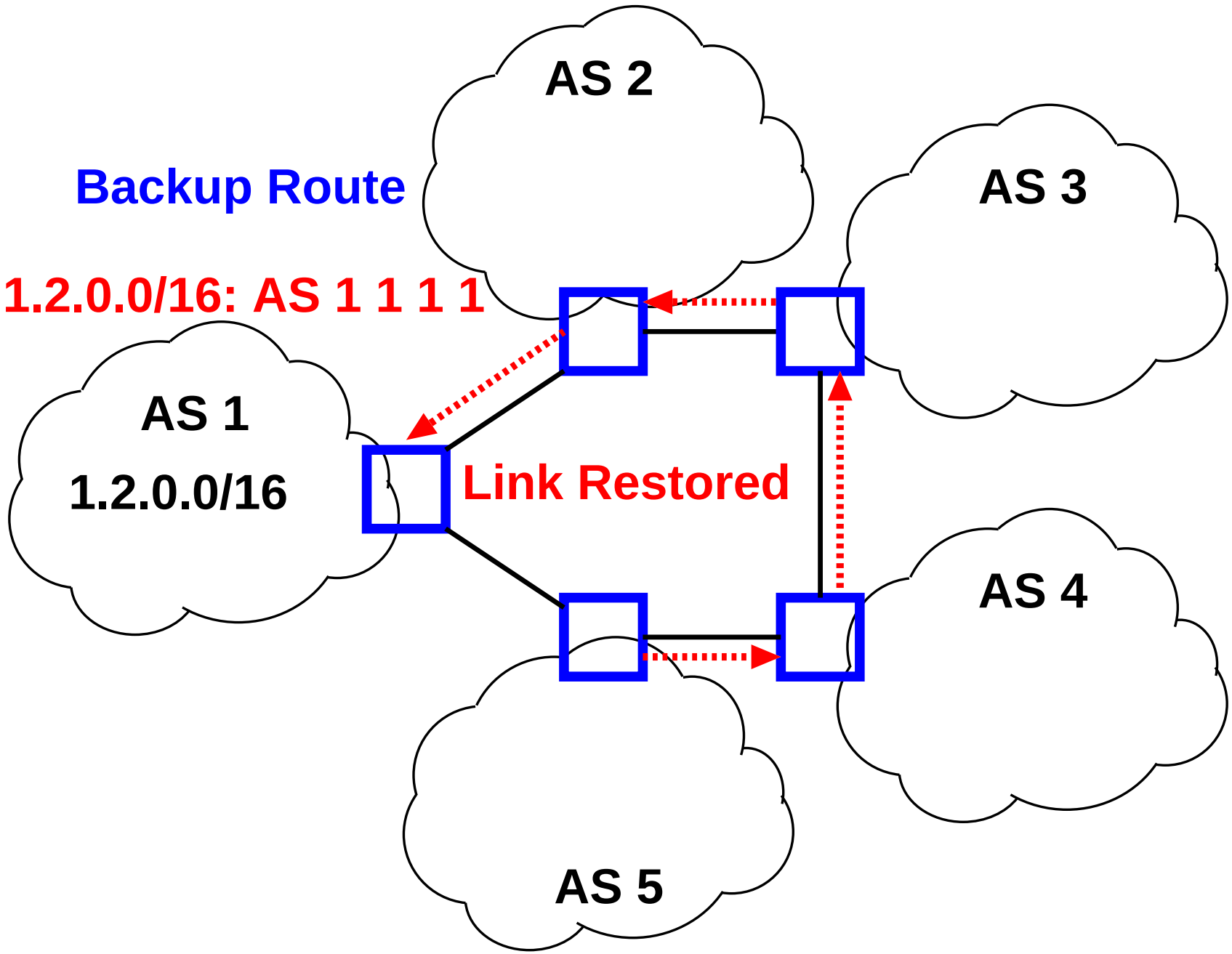
# rfc4264: BGP Wedgies



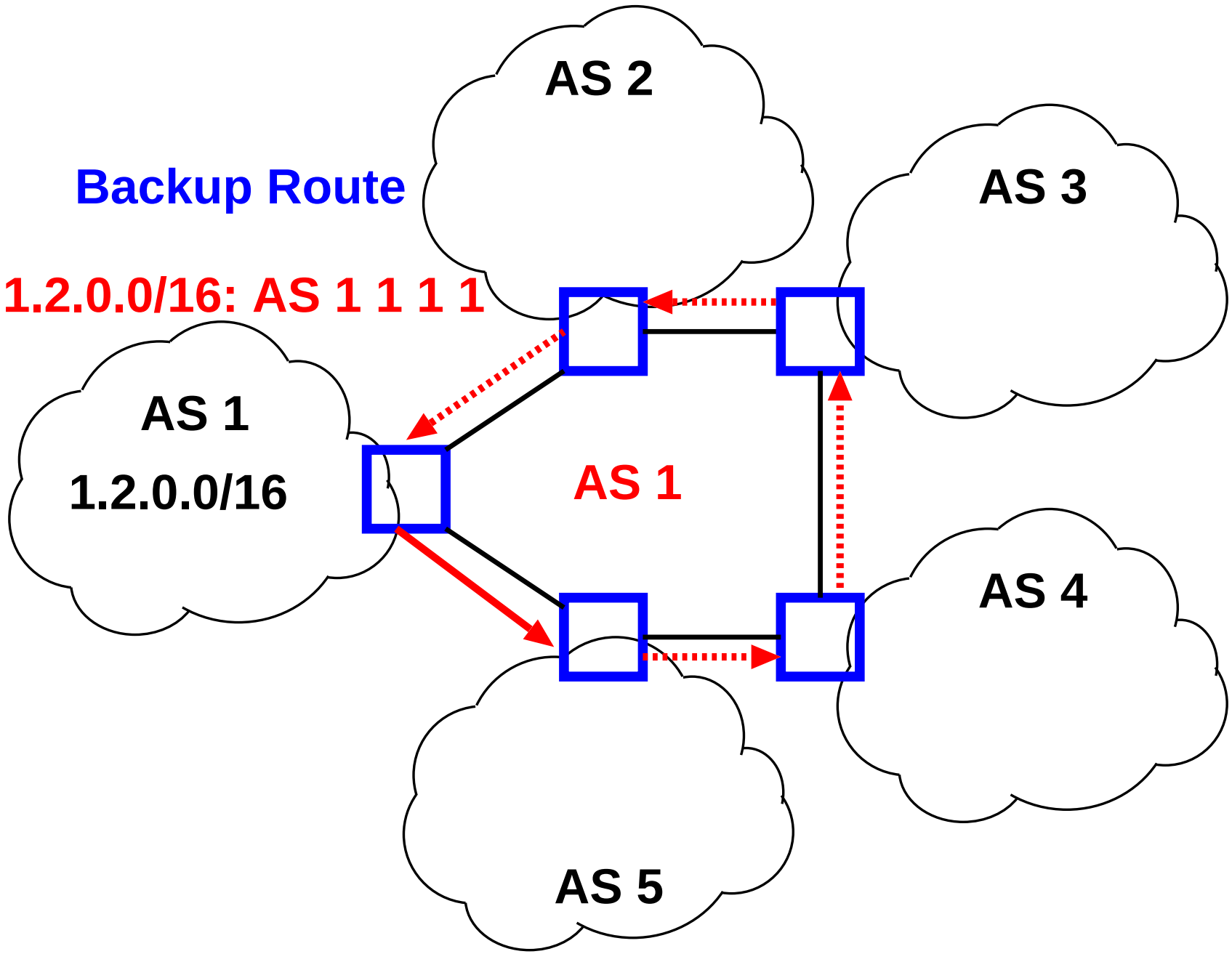
# rfc4264: BGP Wedgies



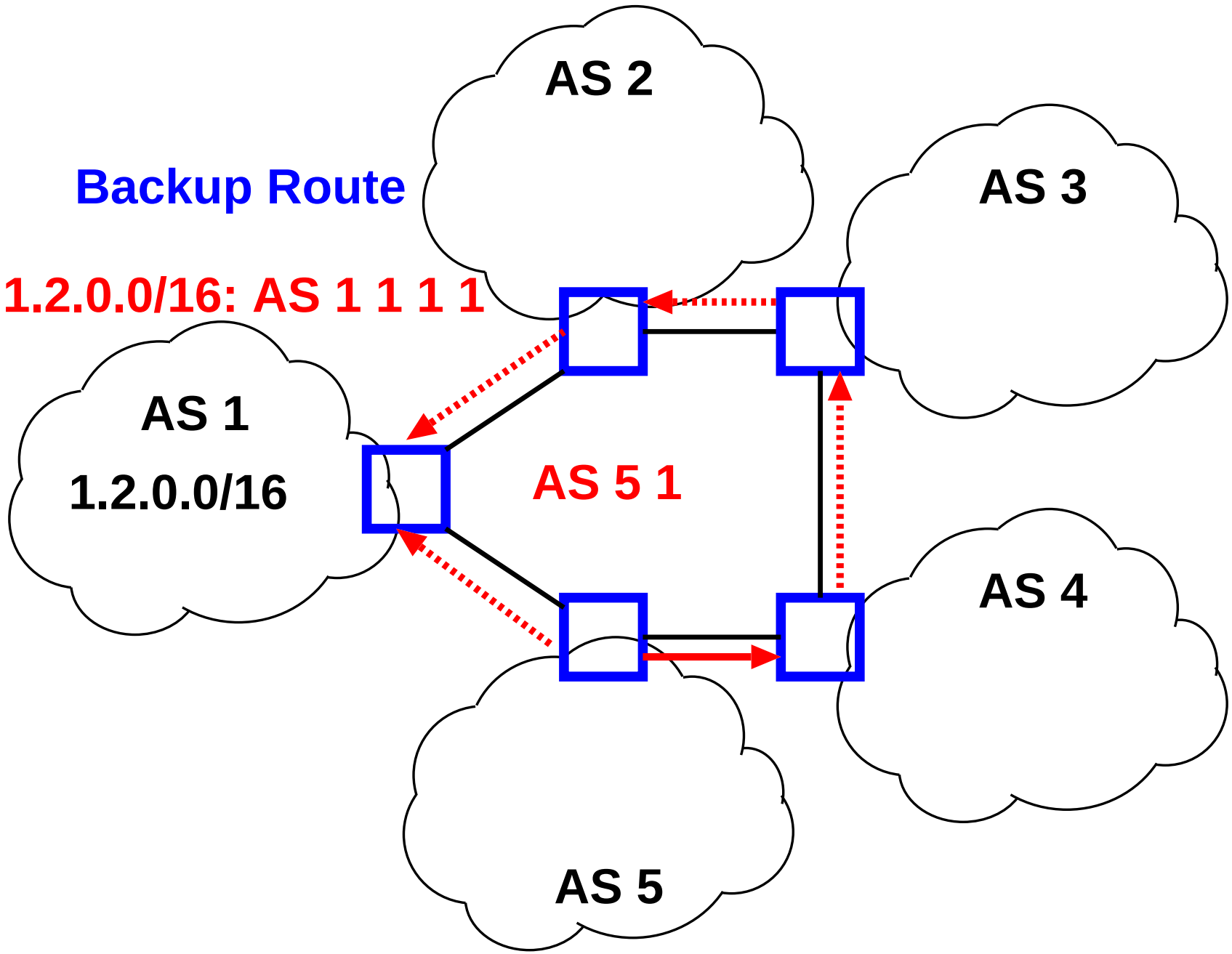
# rfc4264: BGP Wedgies



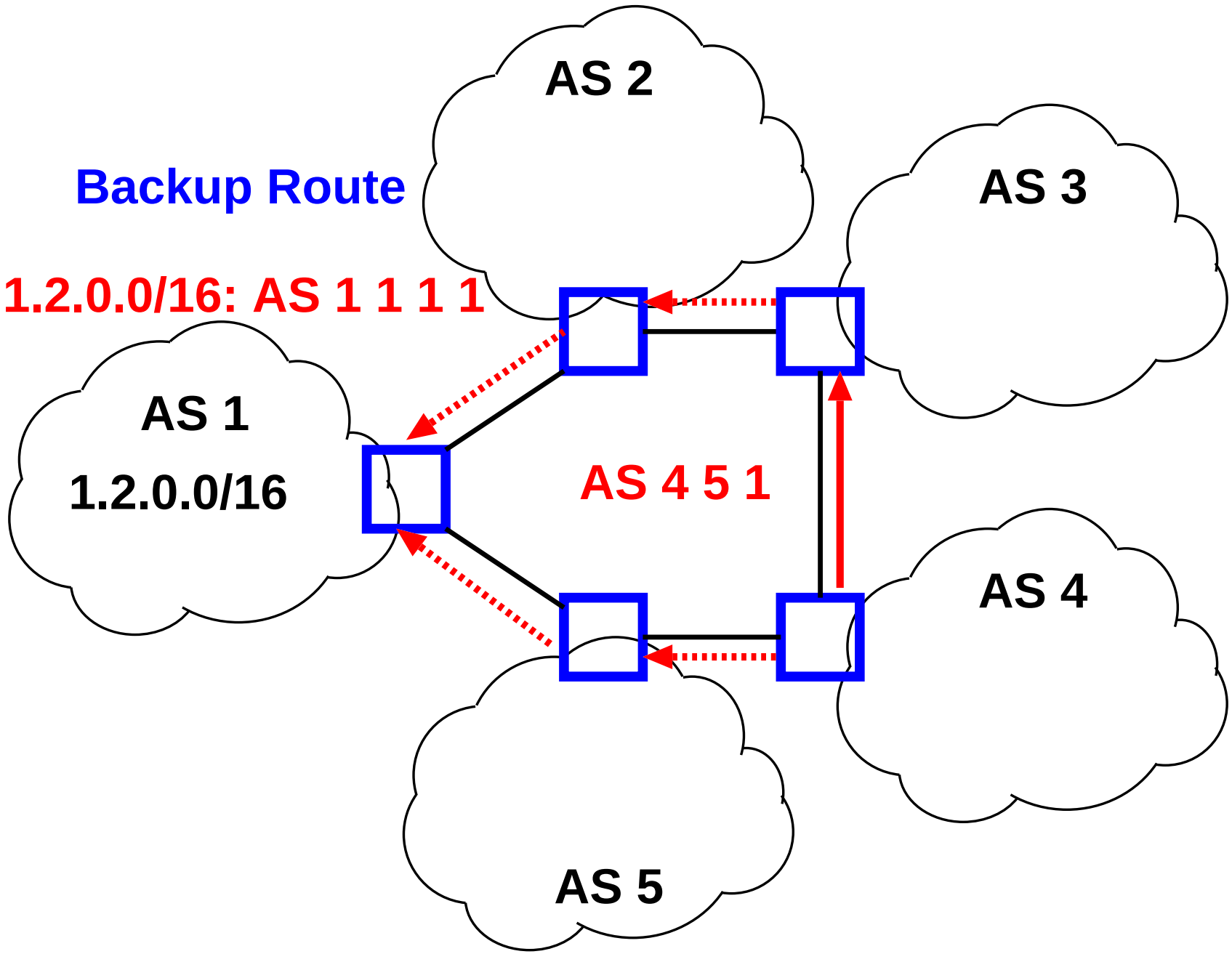
# rfc4264: BGP Wedgies



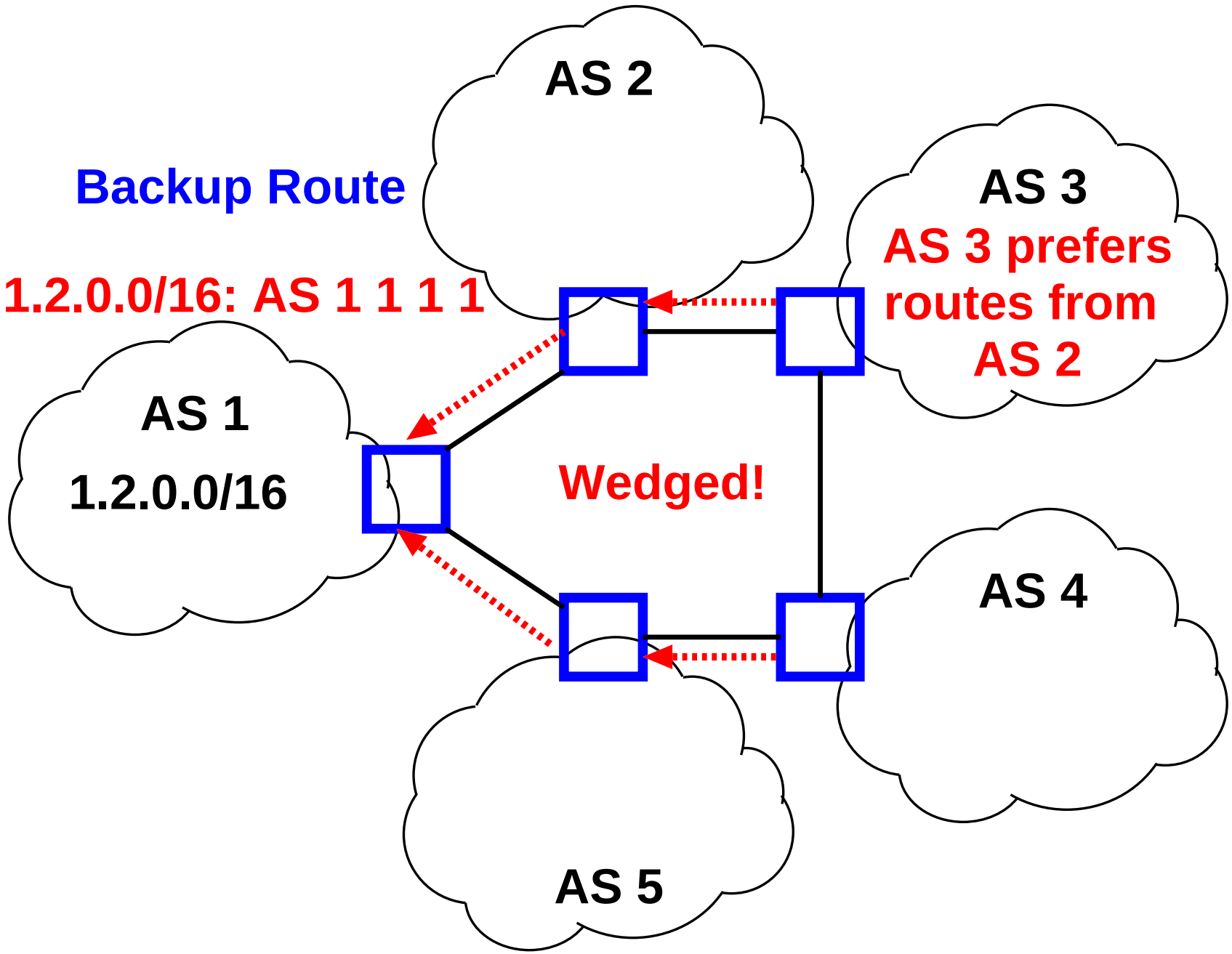
# rfc4264: BGP Wedgies



# rfc4264: BGP Wedgies



# rfc4264: BGP Wedgies



# Conclusion

- BGP is critical
- BGP polices make it complex
- Slides (will be) available online
- Questions: rob.sherwood@stanford.edu