Alcatel·Lucent

Border Gateway Protocol (BGP)

Gordon Wilfong

Algorithms Research Department

Mathematical and Algorithmic Sciences Research Center

Bell Laboratories

August 2007

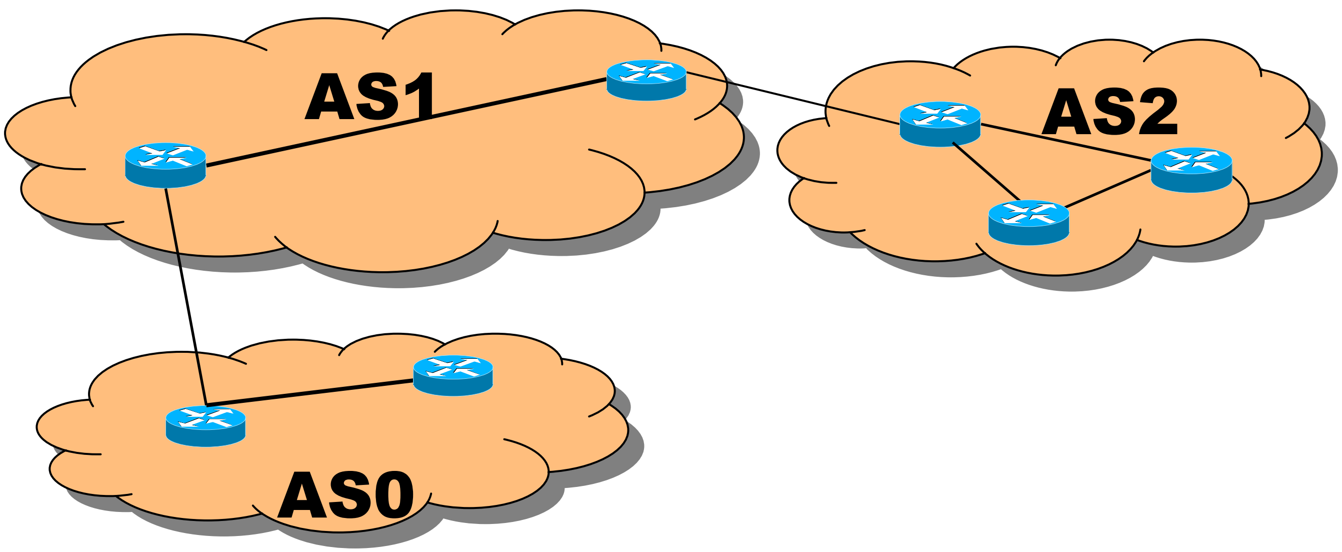
Internet Architecture

Partitioned into Autonomous Systems (ASes)

- Collection of routers under independent administrative control
- Service provider, university, corporate campus…

Hierarchy of Autonomous Systems

- Large, tier-1 provider with a national backbone
- Medium-sized regional provider with smaller backbone
- Small network run by a single company or university



Alcatel·Lucent

Internet Routing

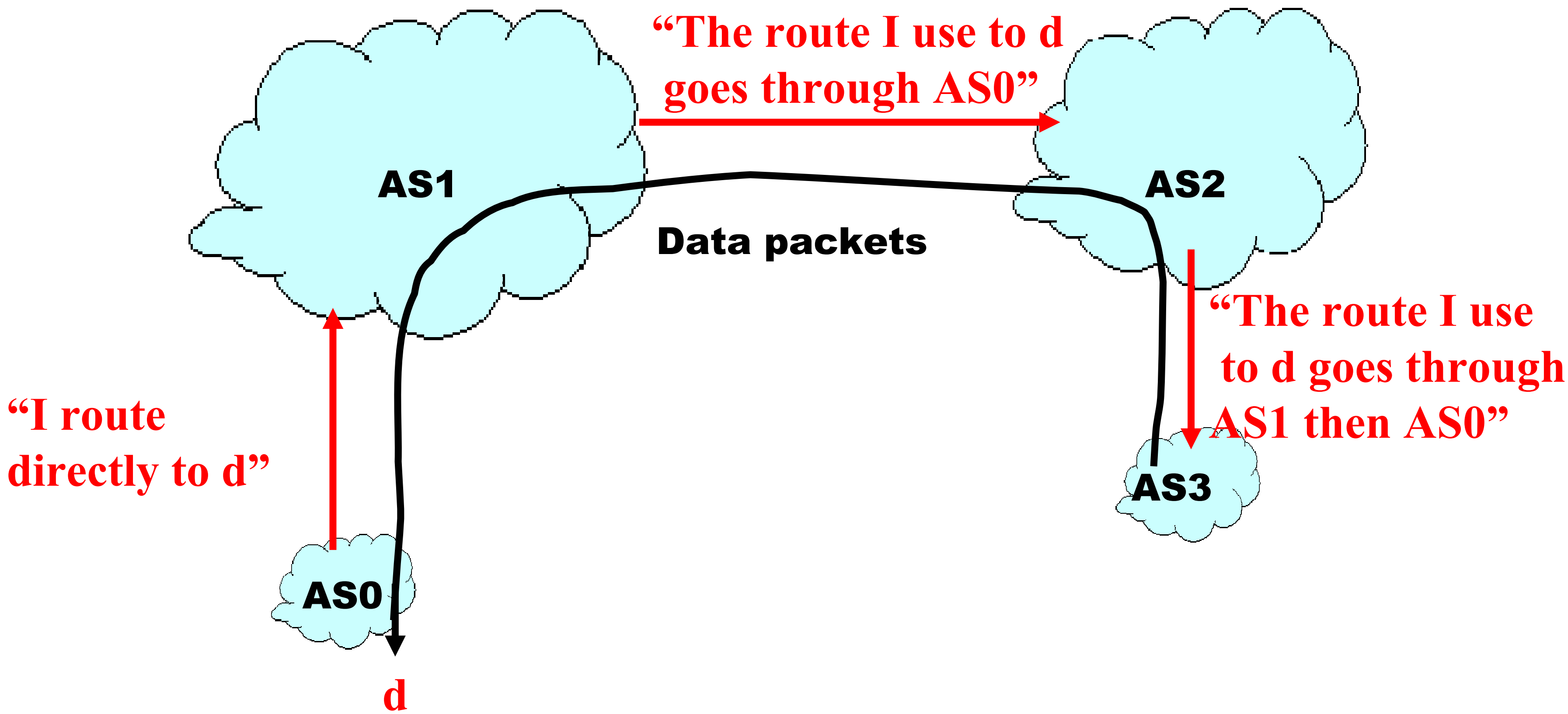
AS has its own economic incentives to:

- cooperate so as to achieve connectivity
- to minimize other's traffic across its network

Routing between ASes achieved by the Border Gateway Protocol (BGP):

- AS implicitly ranks paths to the destination
- AS selects highest ranked route it knows about from neighbors
- AS selectively announces to neighbors its chosen route

Border Gateway Protocol (BGP)



BGP Attributes

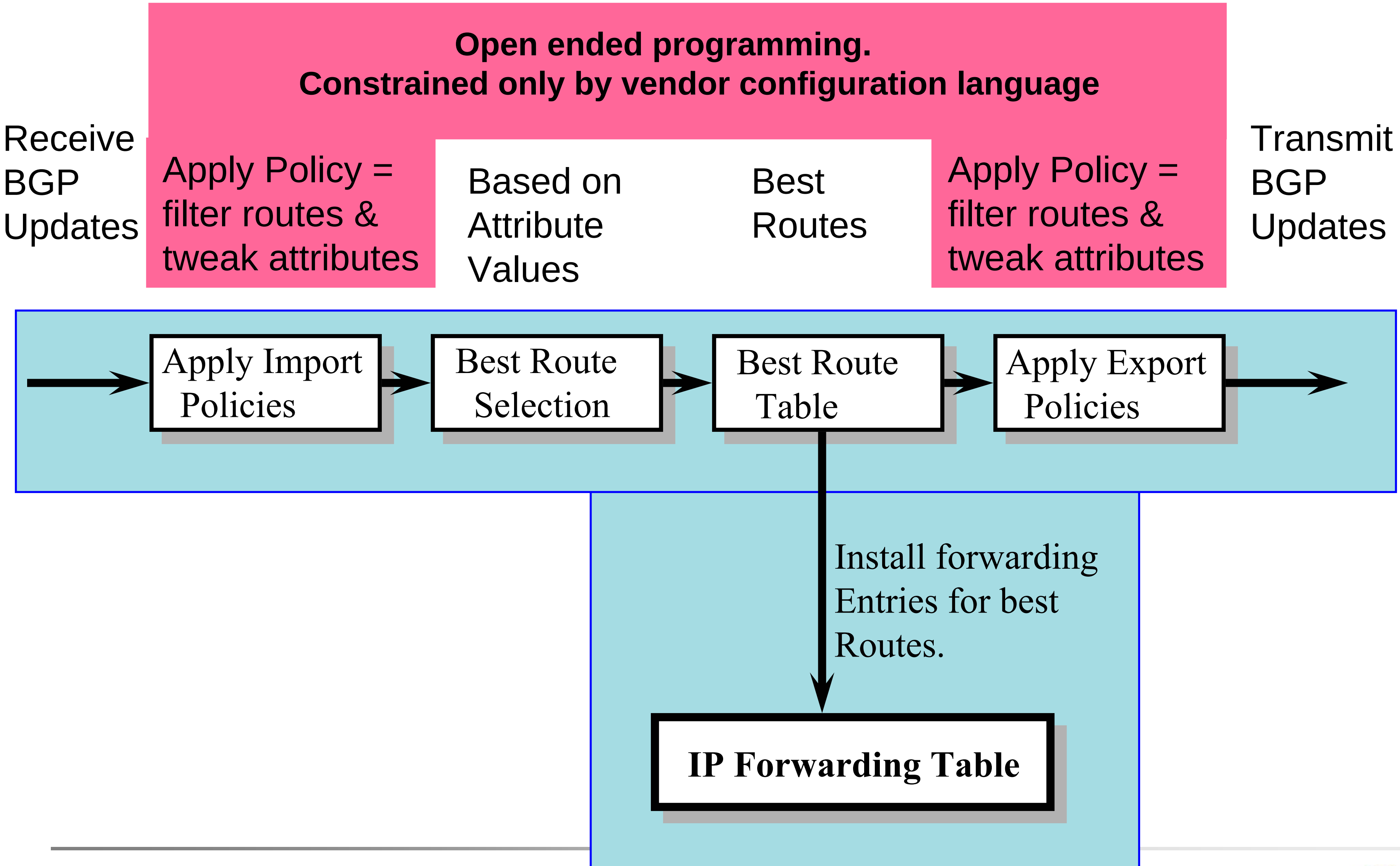
Code	Reference
ORIGIN	[RFC1771]
AS_PATH	[RFC1771]
NEXT_HOP	[RFC1771]
MULTI_EXIT_DISC	[RFC1771]
LOCAL_PREF	[RFC1771]
ATOMIC_AGGREGATE	[RFC1771]
AGGREGATOR	[RFC1771]
COMMUNITY	[RFC1997]
ORIGINATOR_ID	[RFC2796]
CLUSTER_LIST	[RFC2796]
DPA	[Chen]
ADVERTISER	[RFC1863]
RCID_PATH / CLUSTER_ID	[RFC1863]
MP_REACH_NLRI	[RFC2283]
MP_UNREACH_NLRI	[RFC2283]
EXTENDED COMMUNITIES	[Rosen]

...

From IANA: http://www.iana.org/assignments/bgp-parameters

Not all attributes need to be present in every announcement

BGP Route Processing



Interdomain Routing

Must scale

- Destination address blocks: 150,000 and growing
- Autonomous Systems: 20,000 and growing
- AS paths and routers: at least in the millions…

Must support flexible policy

- Route selection: selecting which route to a particular destination the AS will use is based on local policy
- Route export: selecting routes to advertise allows control over who can send packets through the AS

Results in convergence problems

- BGP can take several (tens of) minutes to converge
- There are cases where BGP actually fails to converge at all!

E-BGP

External BGP (E-BGP) is the mode of BGP that propagates routes between autonomous systems.

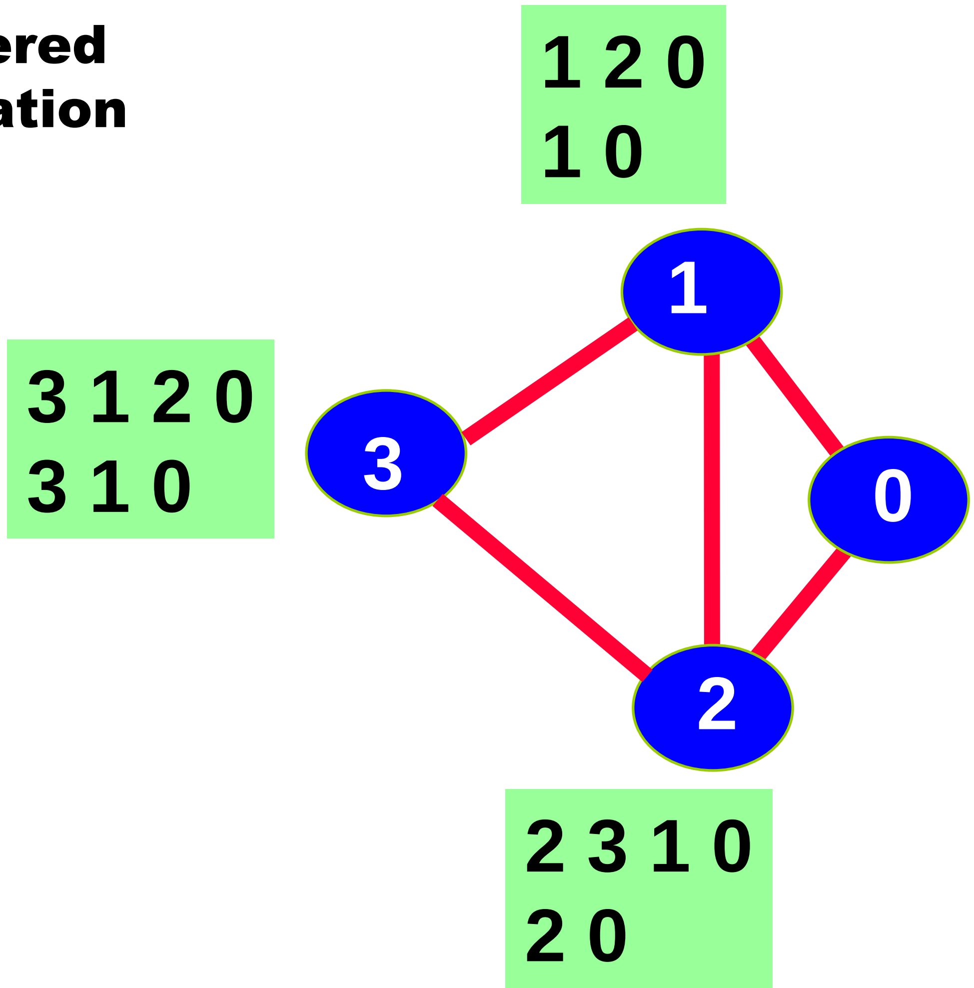
OSPF, RIP, ISIS distributed algorithms for solving shortest paths.

BGP distributed algorithm for solving the Stable Paths Problem (SPP)

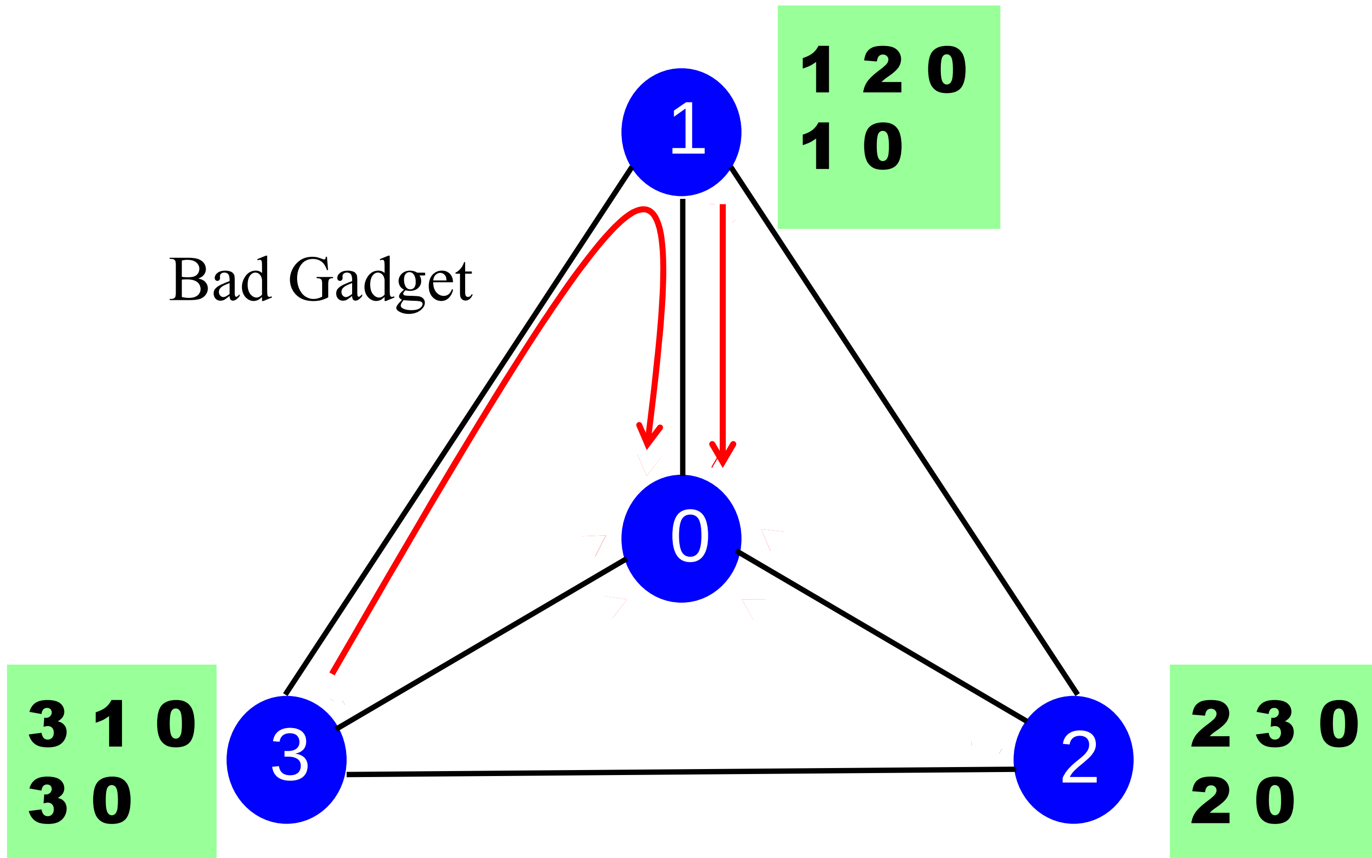
SPP

- graph G=(V,E)
- for each vertex, ordered list of paths to destination

Solution: each vertex chooses a path (consistent with the paths chosen by vertices on the path) and no vertex can choose a more preferred path

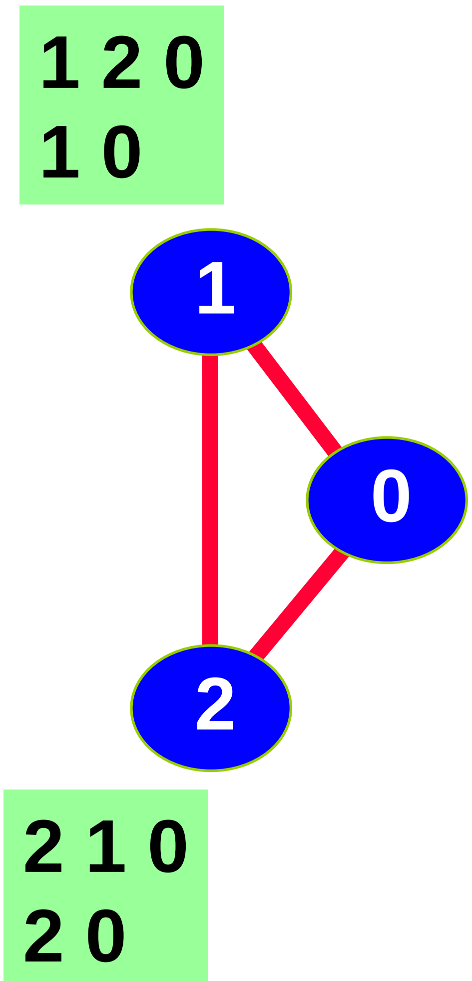


Conflicting Policies Cause Convergence Problems

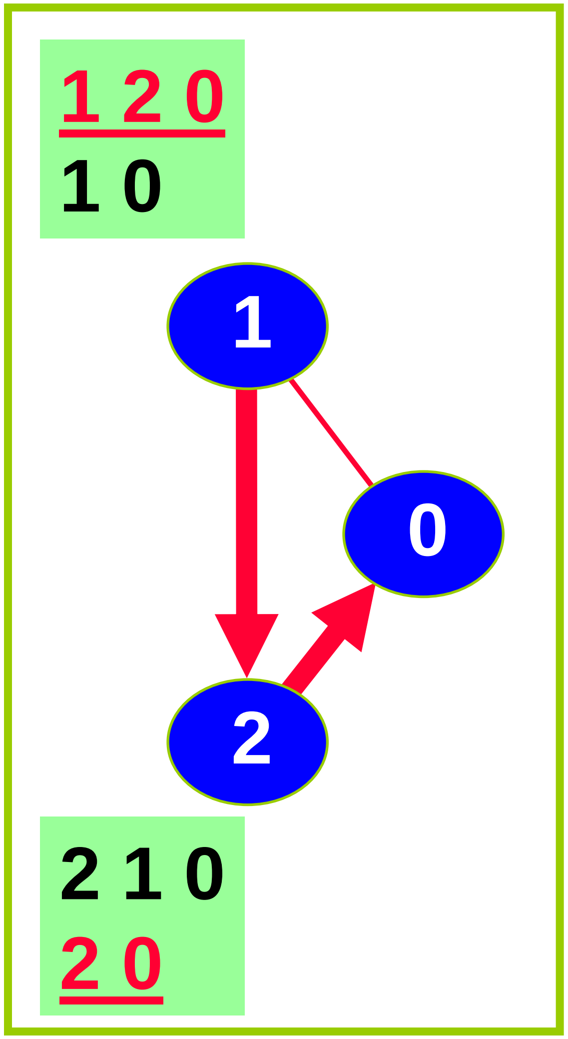


Pick the highest-ranked path consistent with your neighbors' choices.

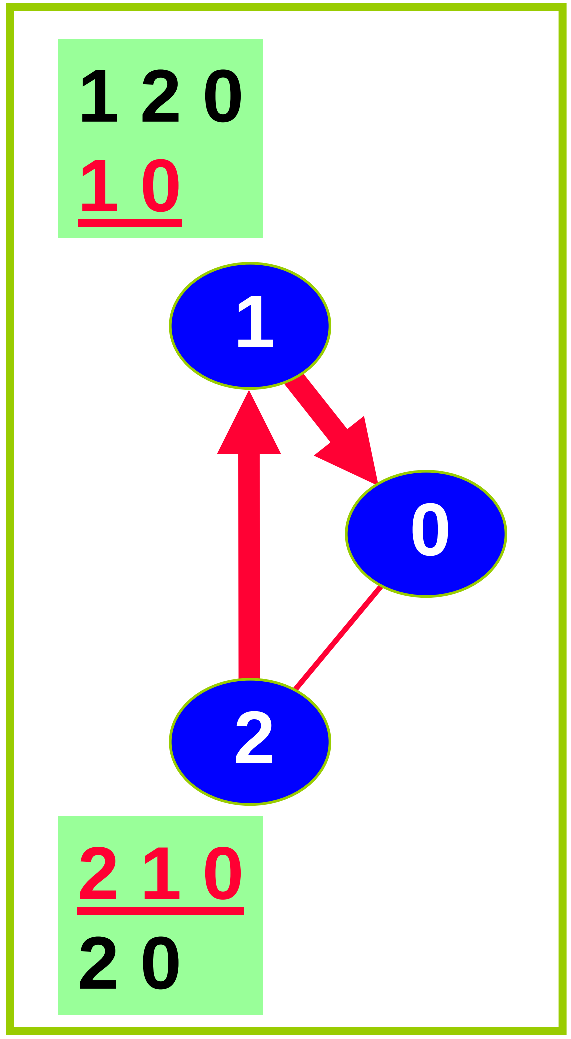
May Be Multiple Solutions



DISAGREE



First solution



Second solution

Complexity of determining existence of a solution

Variables V = {X1, X2, …, Xn}

Clauses C1 = X17 or ~X23 or ~X3,
C2 = ~X2 or X3 or ~X12
….
Cm = X6 or ~X7 or X18

Question Is there an variable assignment
A : V → {true, false} such that
each clause C1, … ,Cm is true?

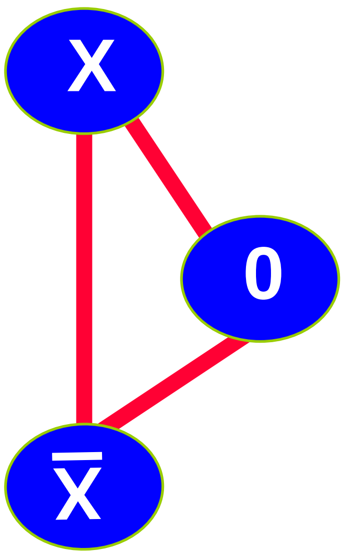
3-SAT is NP-complete

Alcatel·Lucent

Modeling Assignment to Variable X

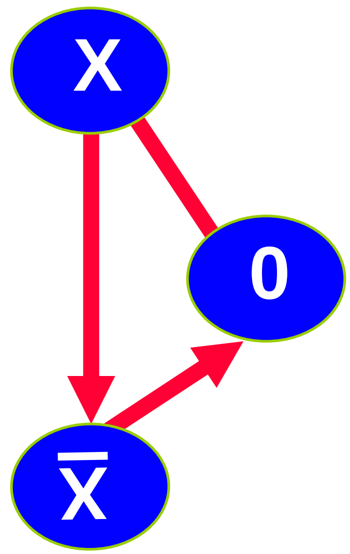
X $\bar{X}$ 0

X 0

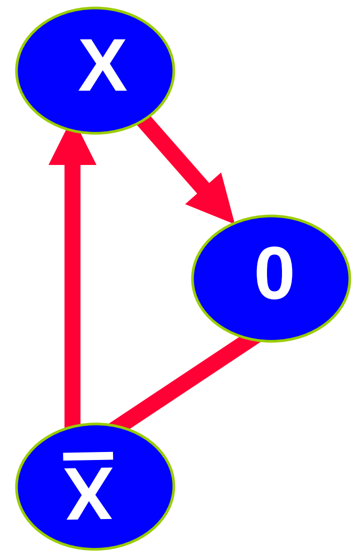


$\bar{X}$ X 0

$\bar{X}$ 0



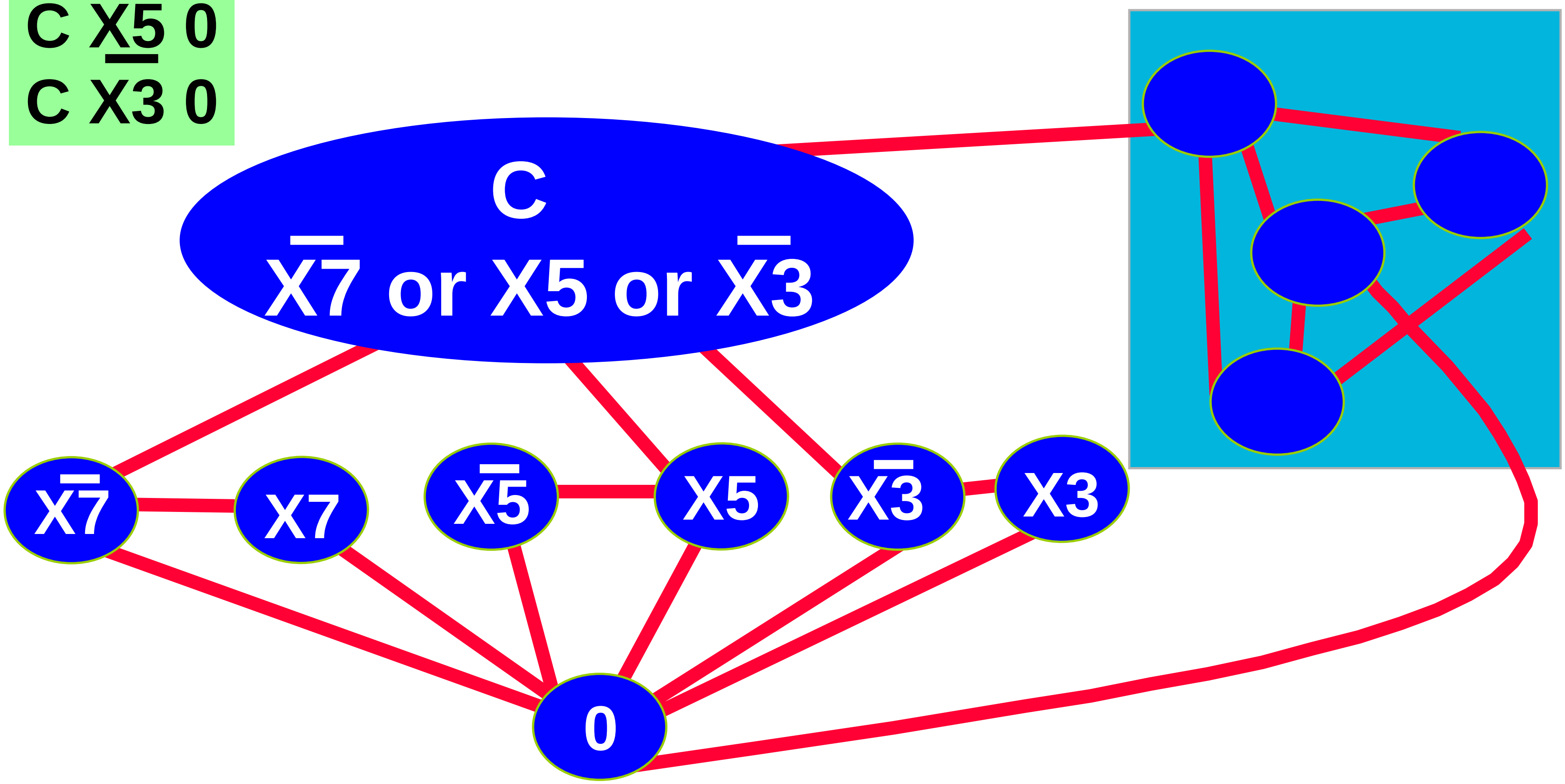
X = false



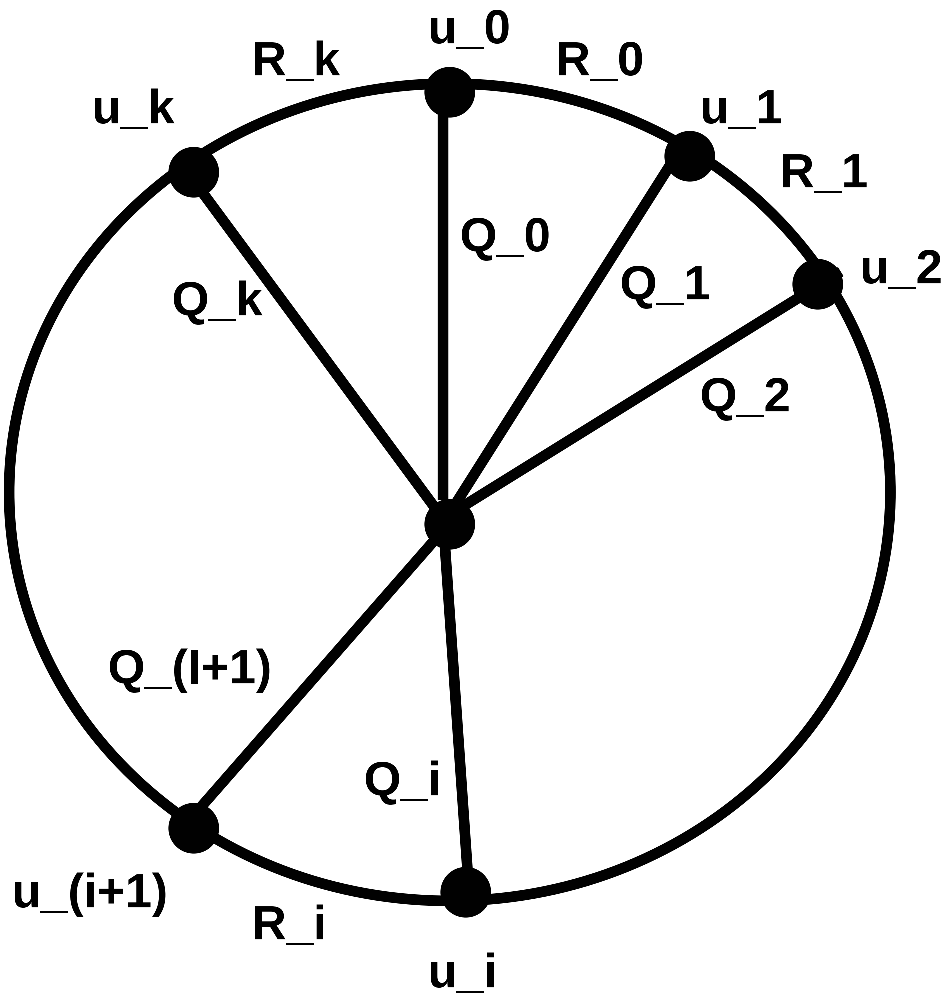
X = true

Convergence Is NP-complete

C $\overline{X7}$ 0
C X5 0
C $\overline{X3}$ 0



Dispute Wheel



At u_i, rank of Q_i is less than or equal to rank of R_iQ_(i+1)

Alcatel·Lucent

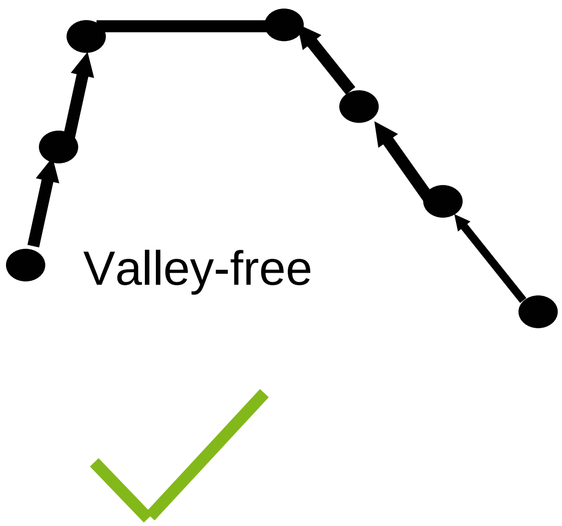
Open Problem

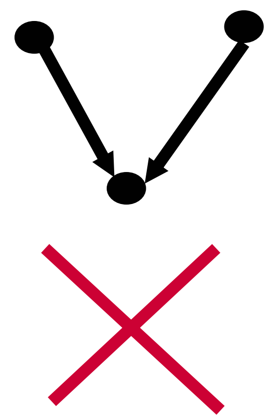
What is the complexity of deciding if an instance of SPP is guaranteed to converge?

What Can Be Done?

Possible approaches:

- Use only configurations that guarantee no problems [GR00]
  - ➢No modifications to BGP required
- Prevent problems for any configuration
  - ➢Modification to BGP required [GW00]



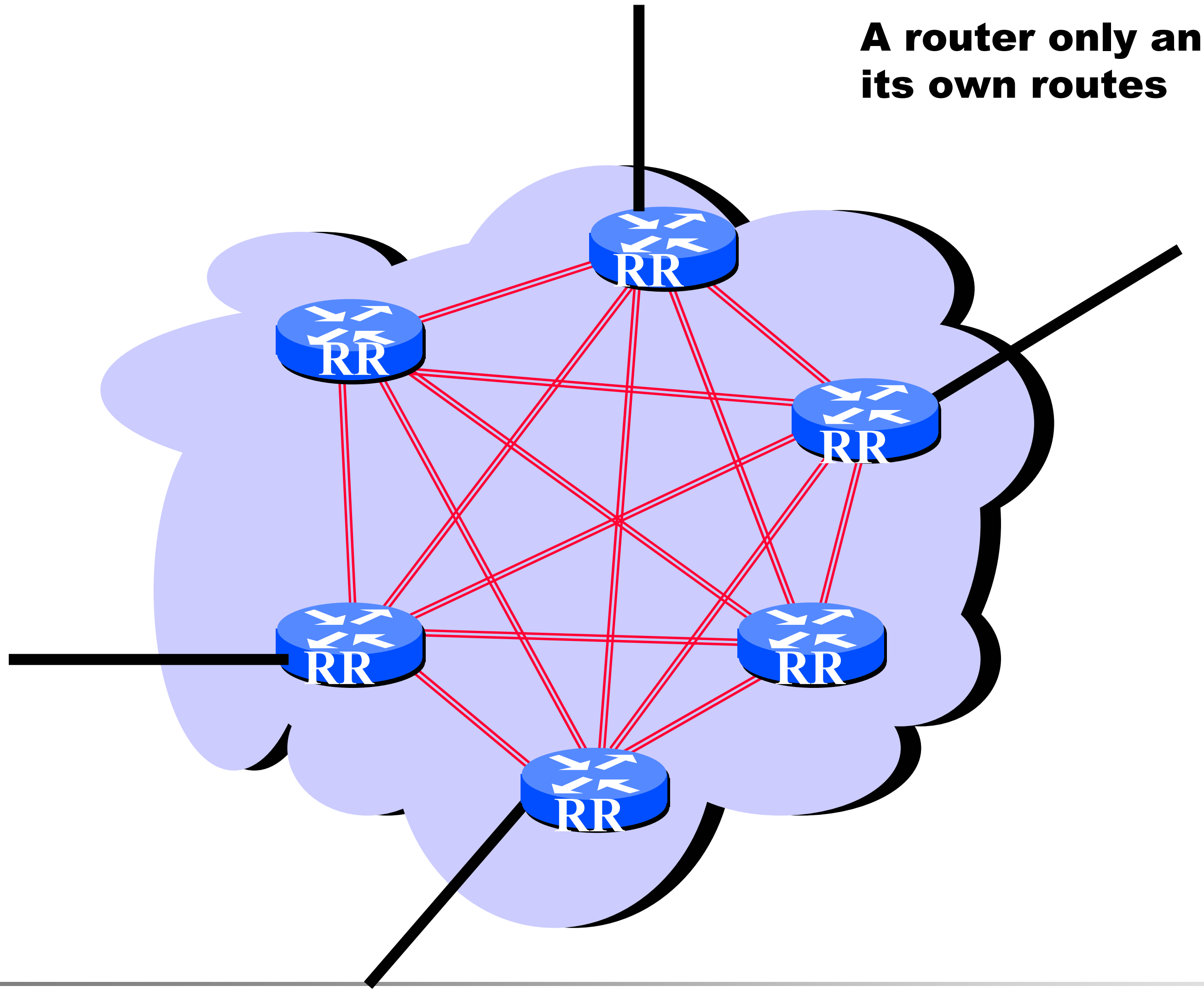


[EHPV]

I-BGP

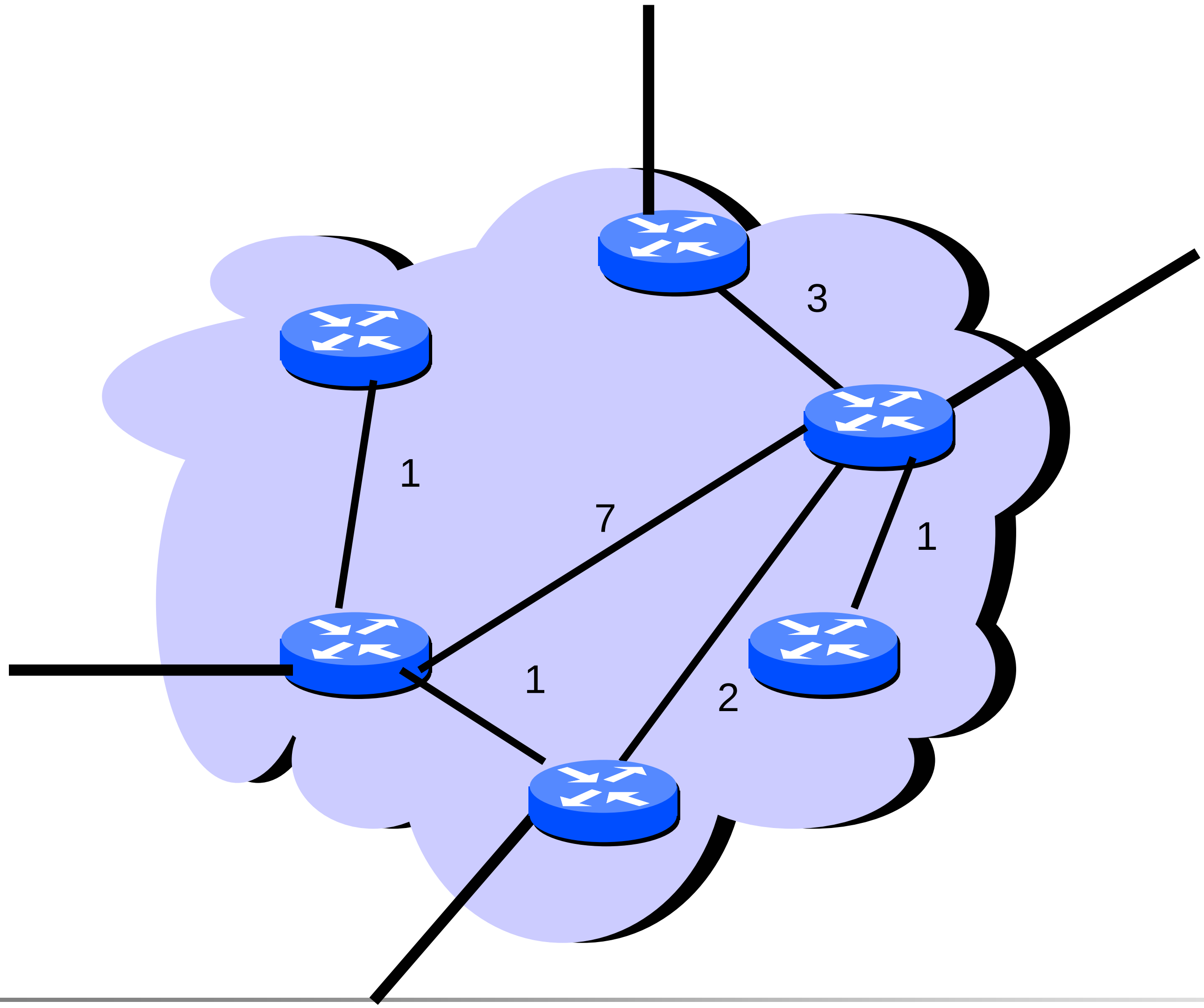
Internal BGP (I-BGP) is the protocol used to propagate external routes within an autonomous system.

Fully Meshed



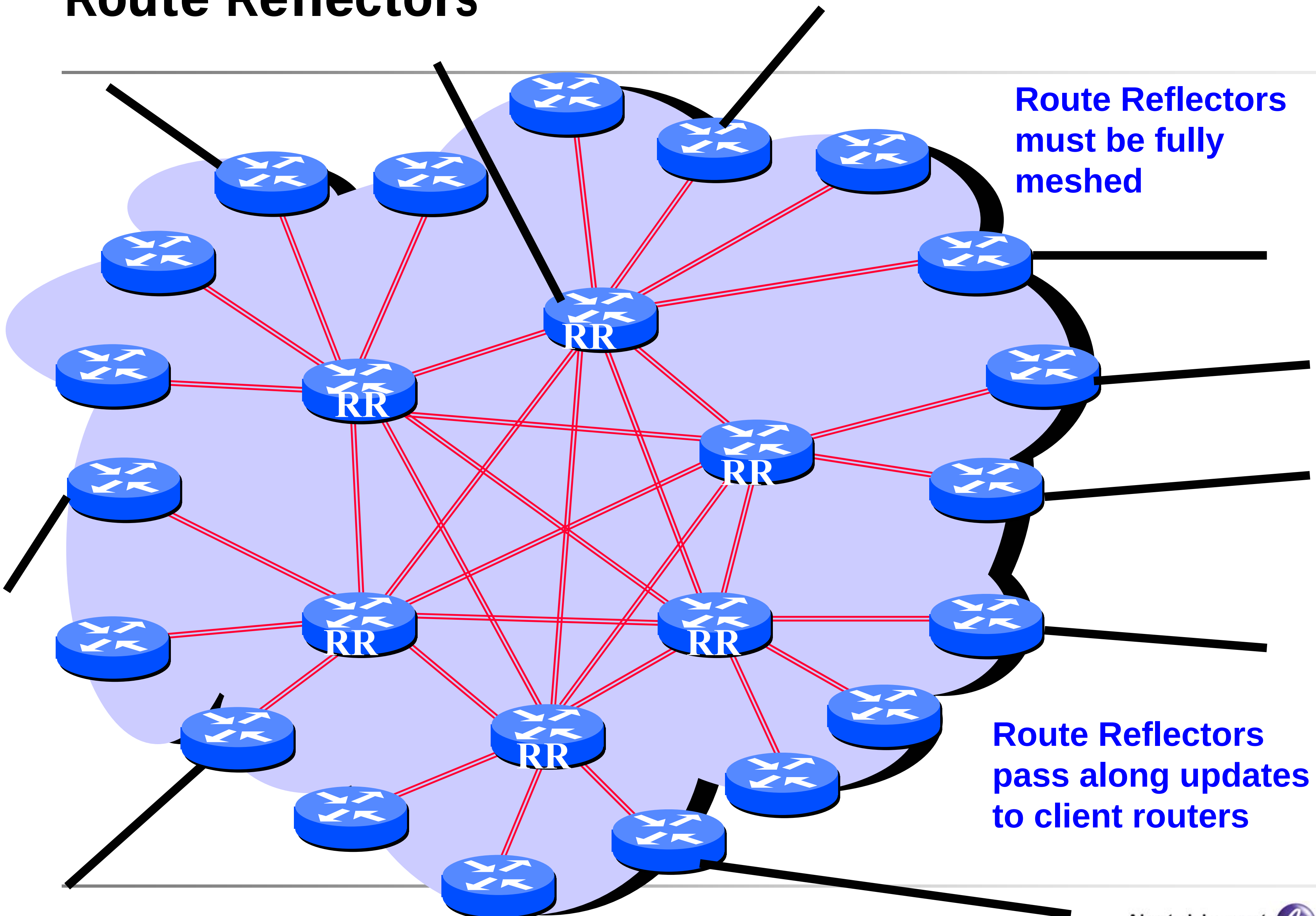
A router only announces its own routes

Shortest Paths to Border



Alcatel·Lucent

Route Reflectors



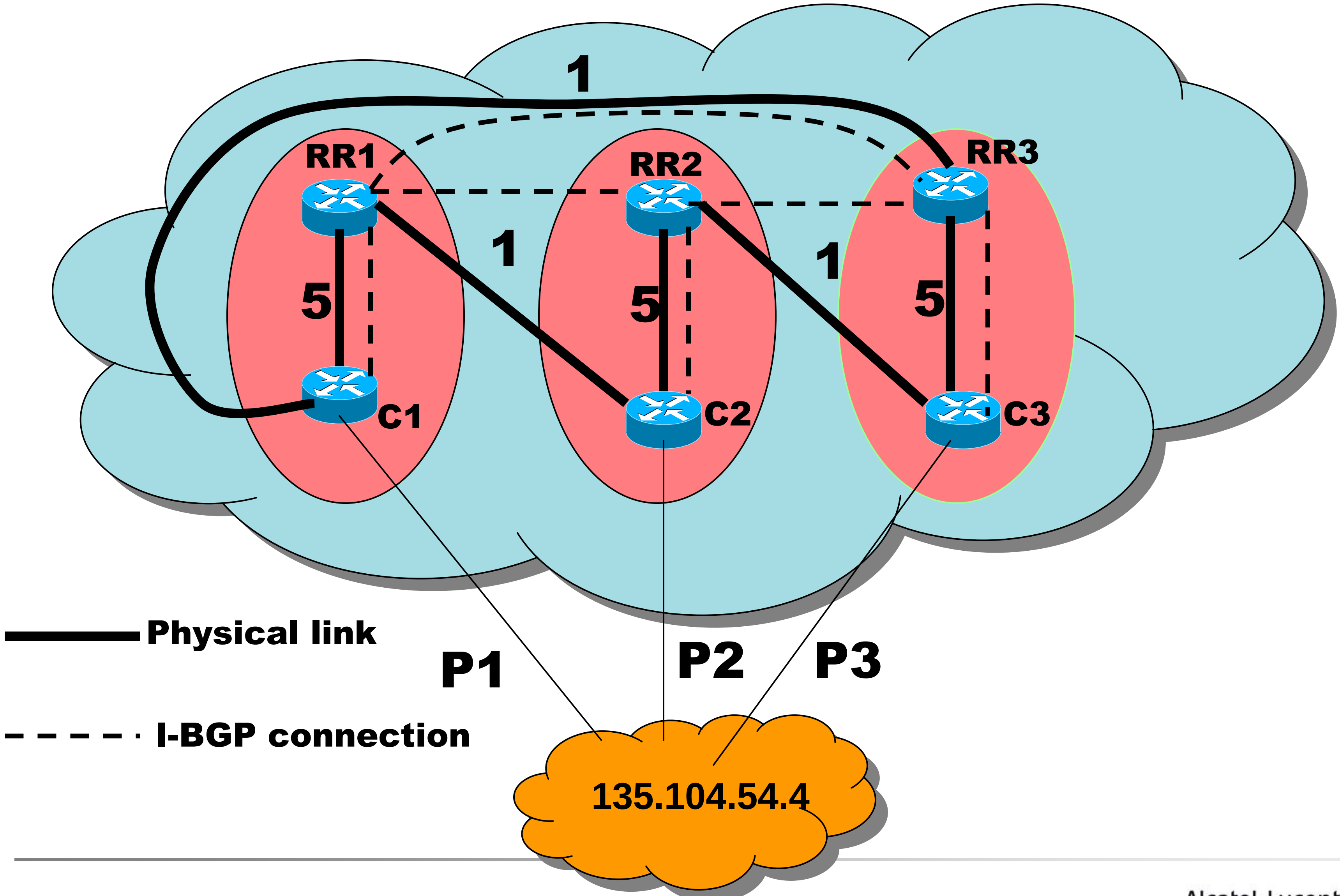
Route Reflectors must be fully meshed

Route Reflectors pass along updates to client routers

Route Selection Summary

Highest Local Preference	**Enforce relationships**
Shortest ASPATH **(Lowest MED (if same next AS))** **i-BGP < e-BGP** **Lowest IGP cost to BGP egress**	**Traffic engineering**
Lowest router ID	**Throw up hands and break ties**

Oscillations in I-BGP

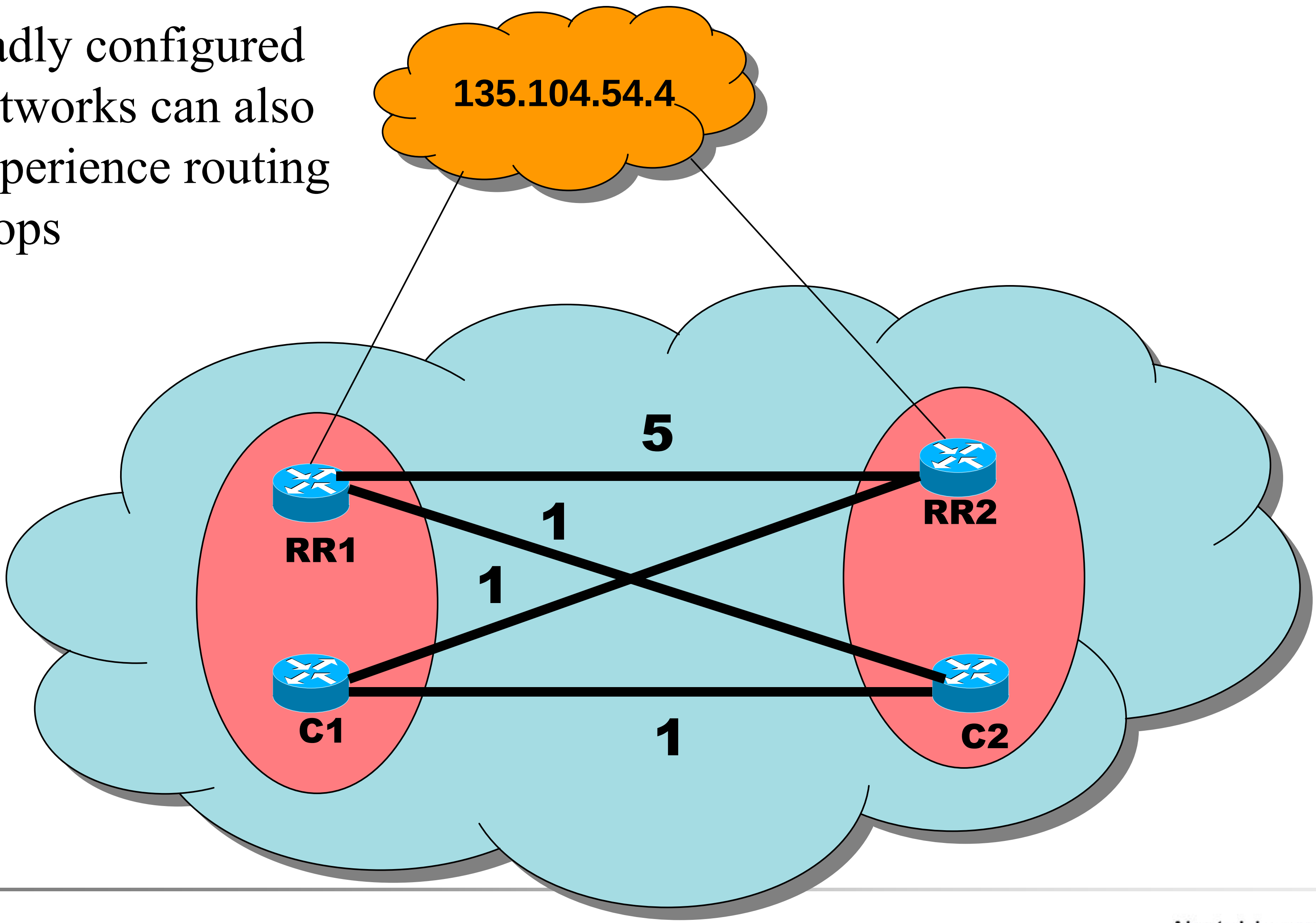


Signaling Safety

- **A configuration is said to be *signaling safe* if I-BGP converges for all possible learned external routes**
- **Determining signaling safety is NP-hard**
- **Sufficient conditions to guarantee signaling safety:**
  - **the directed graph consisting of arcs from clients to route reflectors contains no directed cycles**
  - **route reflectors prefer routes heard about from clients over routes heard from other route reflectors**

Routing Loops

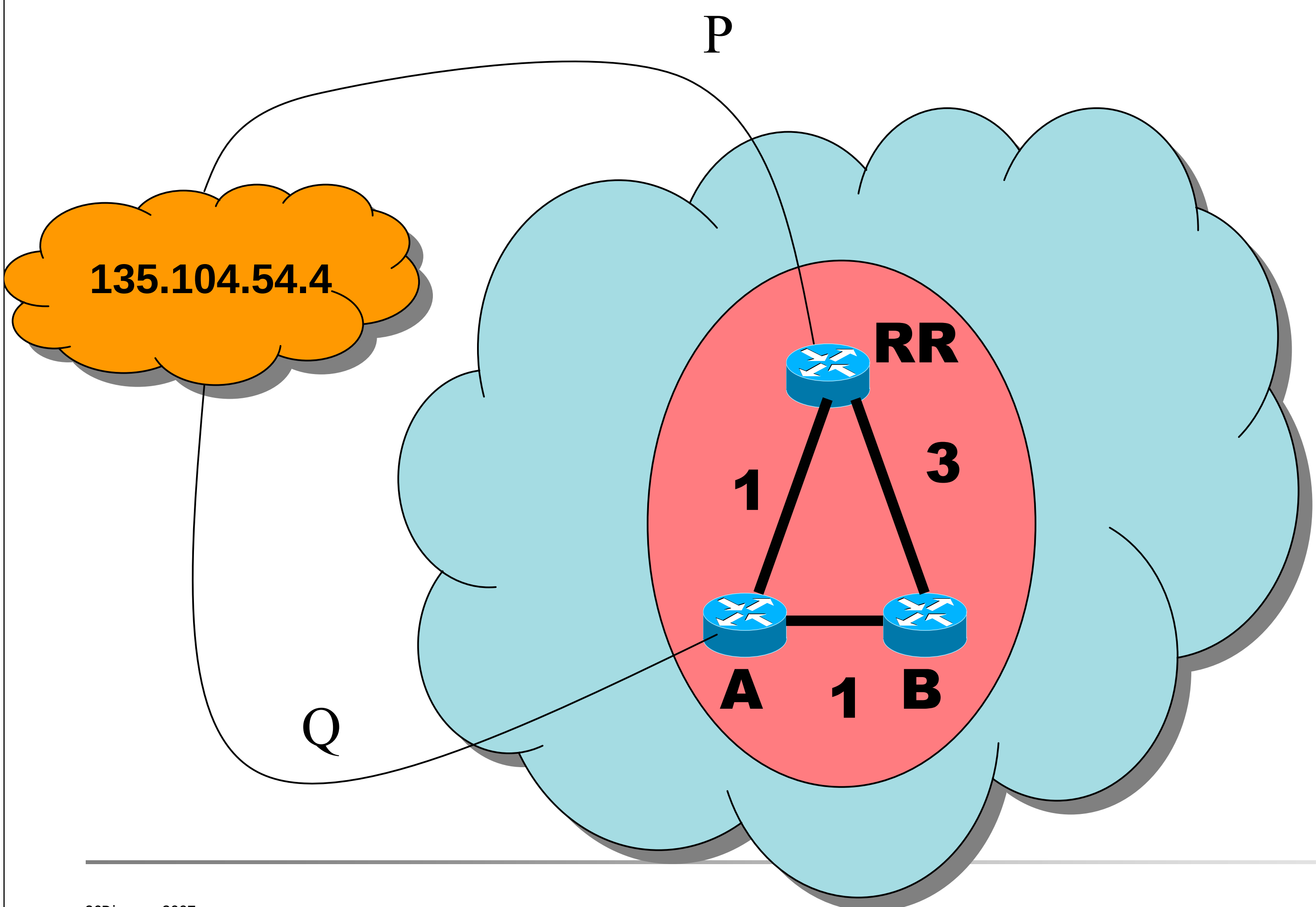
Badly configured networks can also experience routing loops



Alcatel·Lucent

Deflections

Packets can be diverted out of a network unexpectedly.



Alcatel·Lucent

Forwarding Correctness

- **A signaling safe configuration is** *forwarding correct* **if there are no deflections (and hence no loops) for any set of learned external routes**
- **Determining forwarding correctness is NP-hard**
- **Sufficient conditions to guarantee forwarding correctness:**
  - **shortest path between two nodes is a signaling path**
  - **route reflectors prefer client routes to others**

Route Selection Summary

Highest Local Preference — **Enforce relationships**

Shortest ASPATH

(Lowest MED (if same next AS))

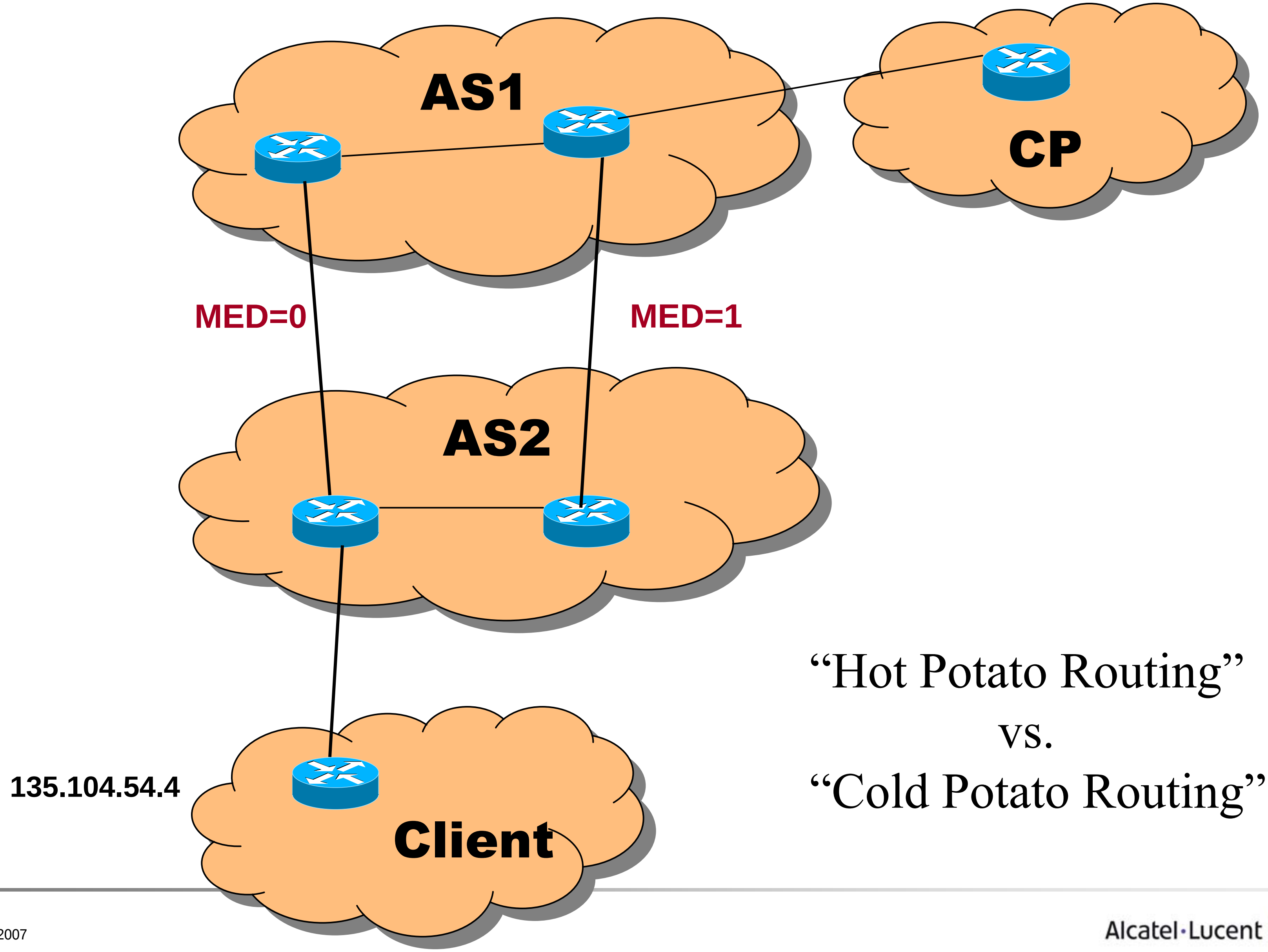
i-BGP < e-BGP

Lowest IGP cost to BGP egress

— **Traffic engineering**

Lowest router ID — **Throw up hands and break ties**

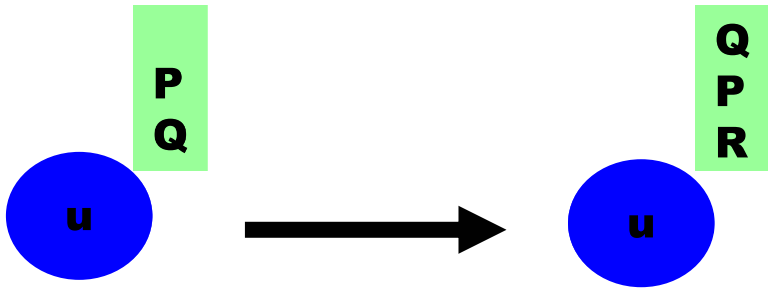
What Are MEDs?



“Hot Potato Routing”
vs.
“Cold Potato Routing”

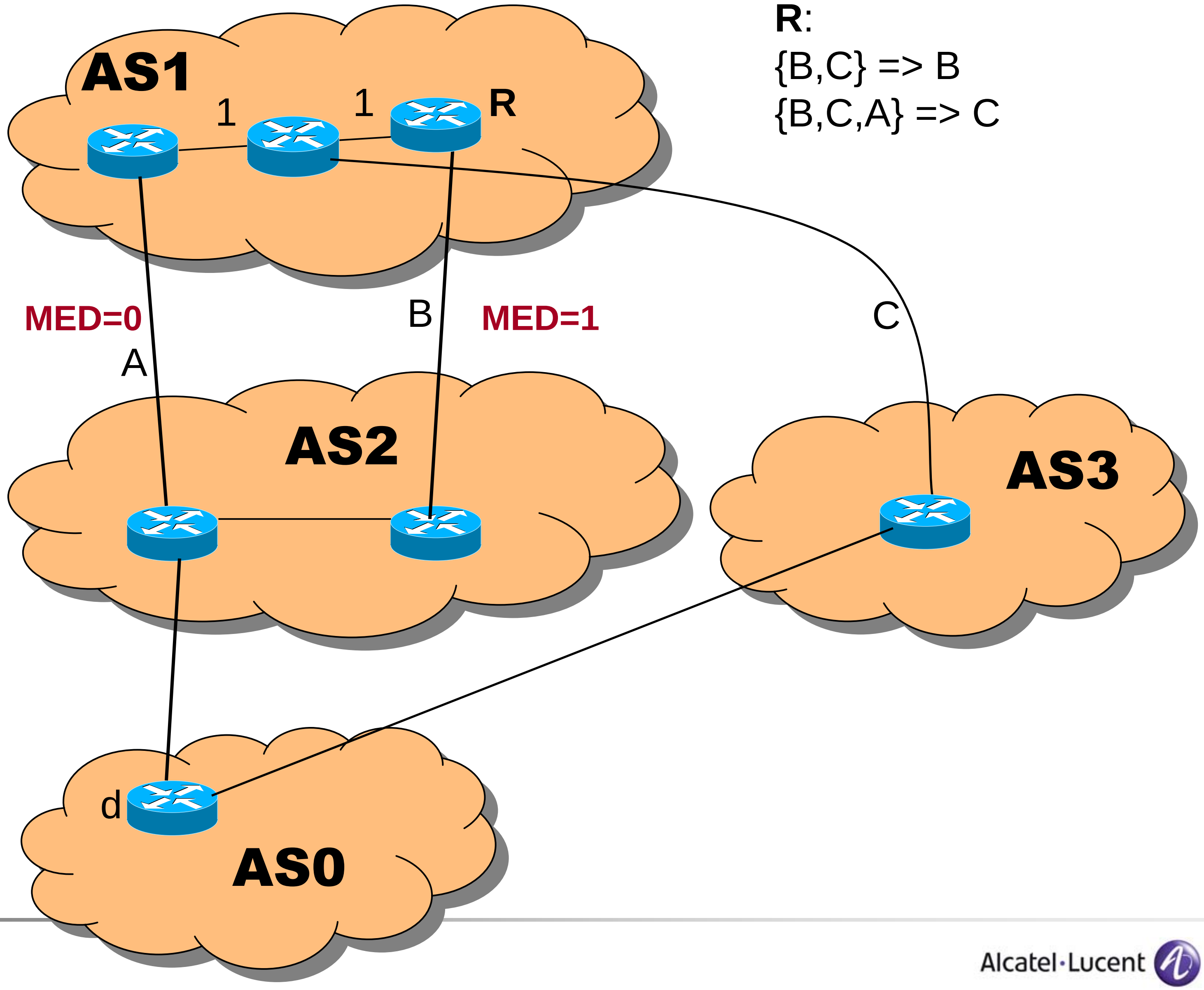
MED Attribute

- MED disobeys independent ordering
  - the presence of a route may change the rank ordering of other routes



Alcatel·Lucent

Not order preserving



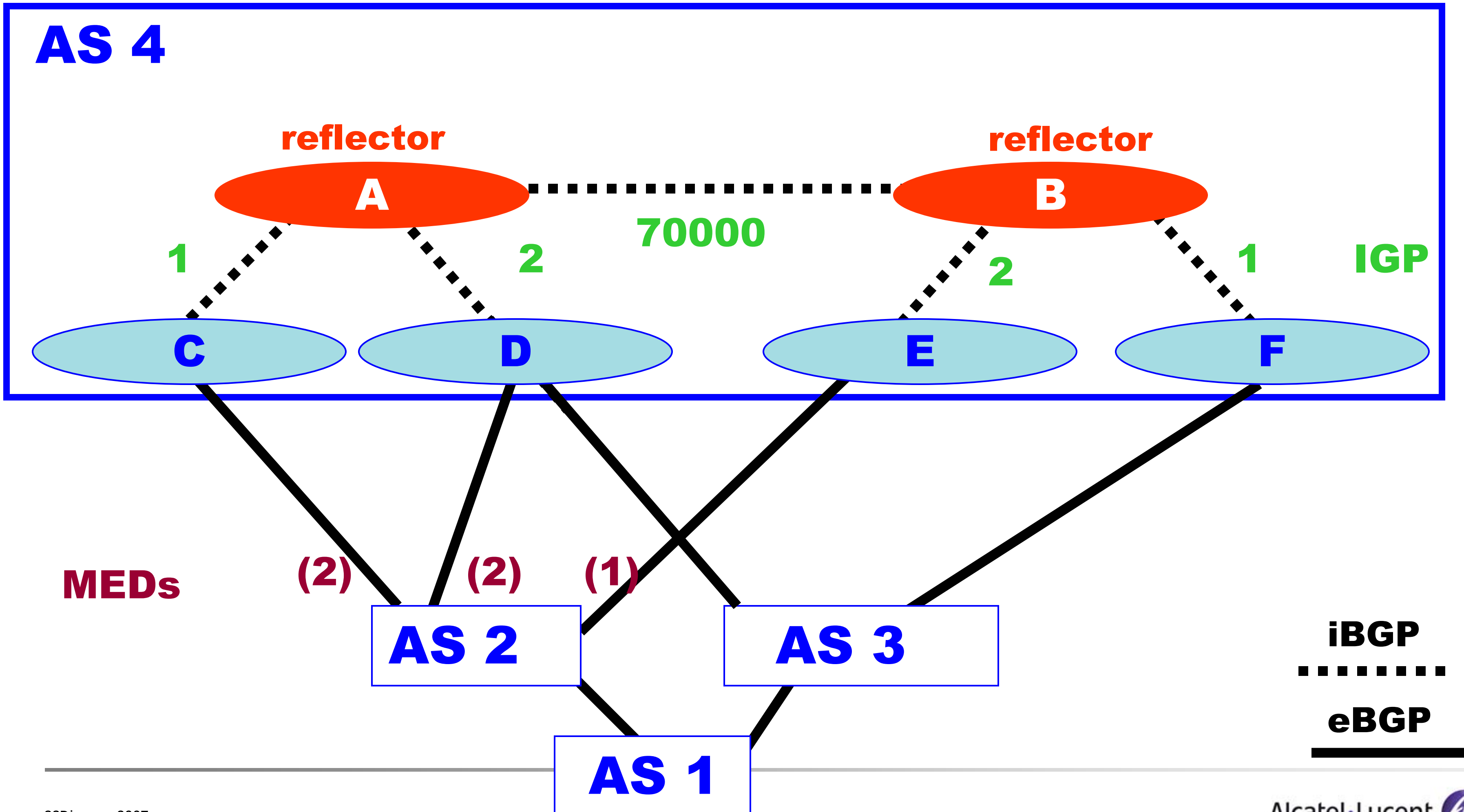
R:
{B,C} => B
{B,C,A} => C

A Real-world MED Oscillation Example

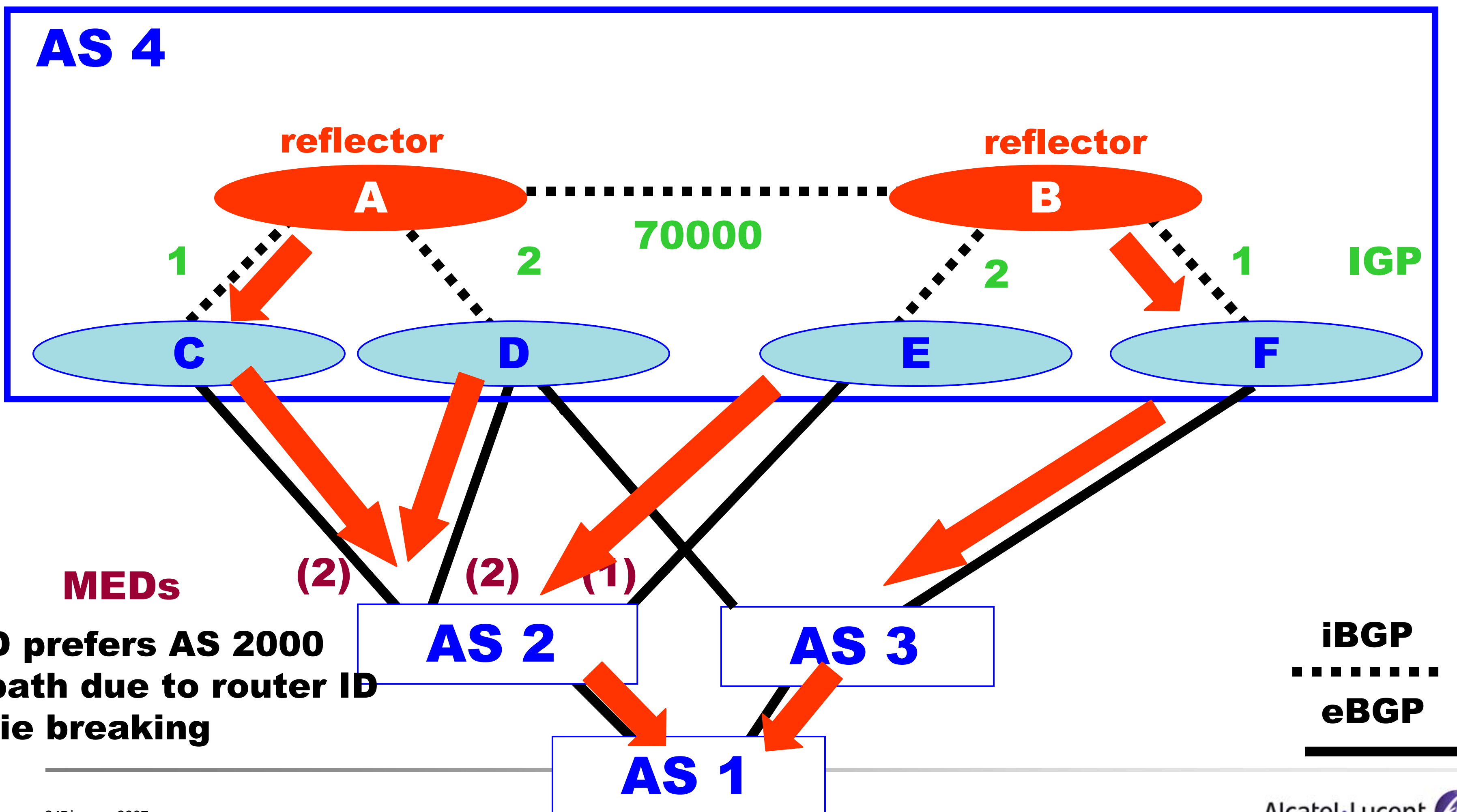
- A functioning network breaks into a state of persistent route oscillations when a BGP session goes down
- First thought to be a hardware problem
- Analysis shows that route oscillations caused by the use of the MED attribute

Initial State

Only AS 2 sends MEDs to AS 4



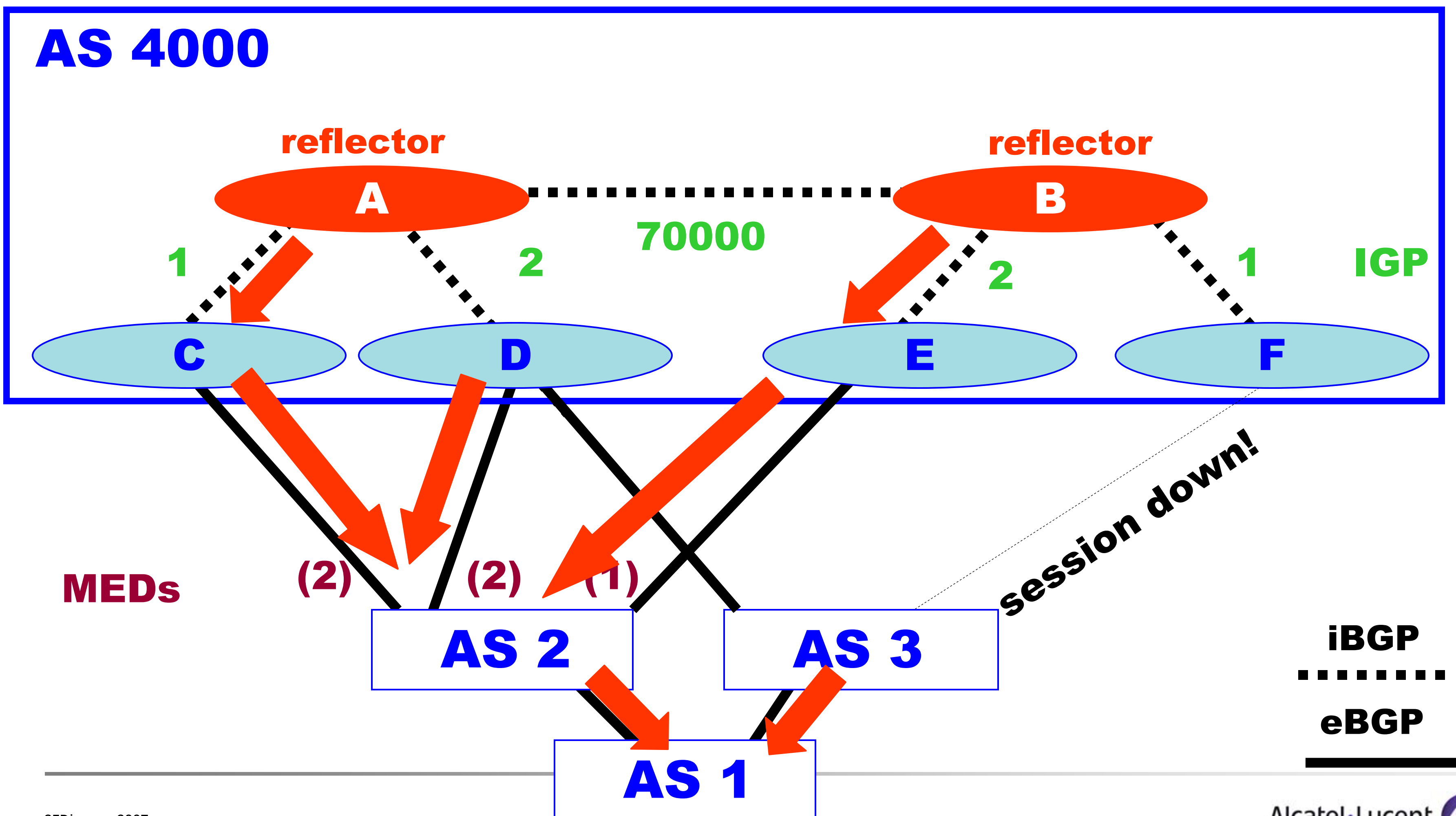
Initial Routing



D prefers AS 2000 path due to router ID tie breaking

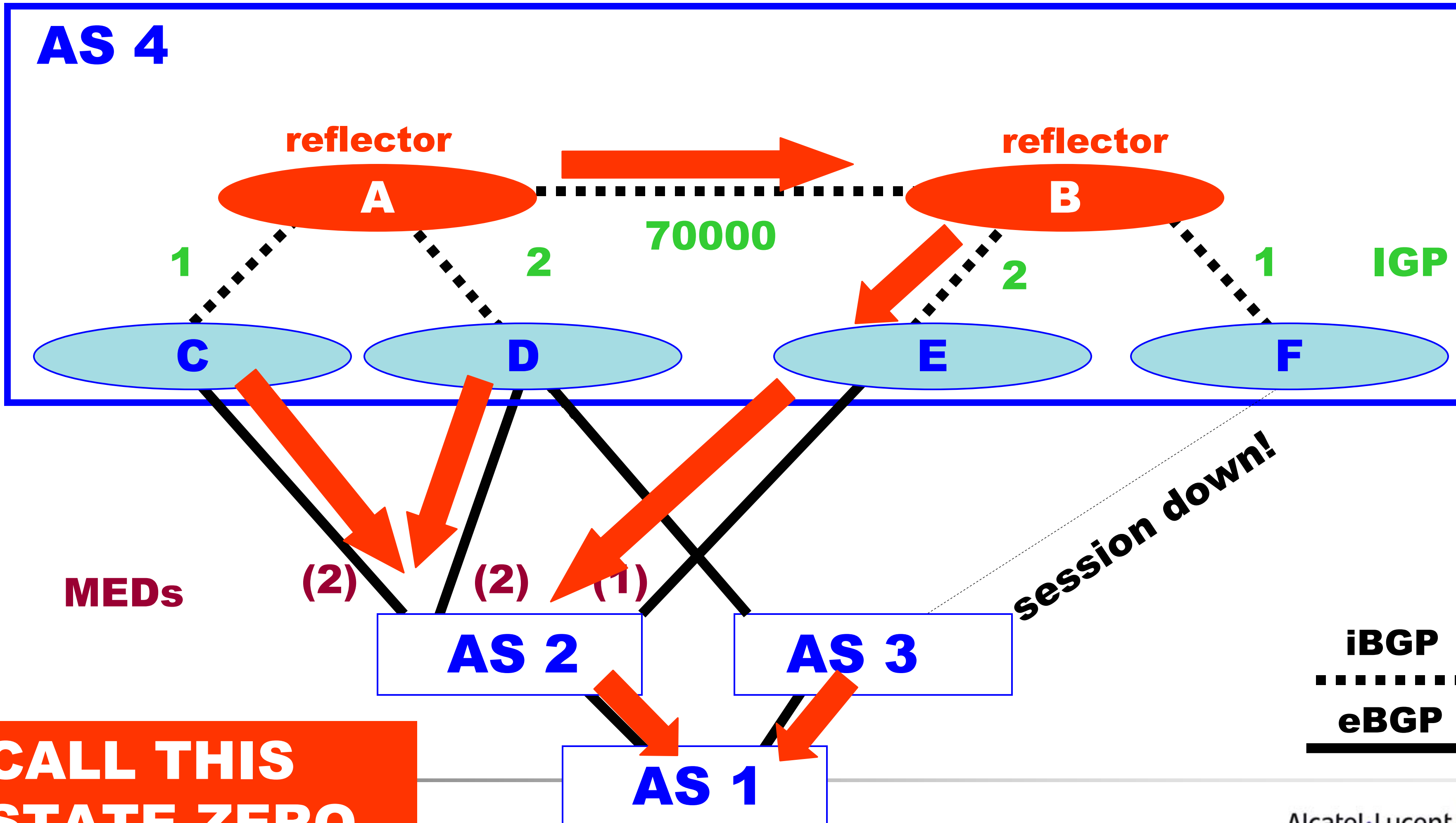
B Changes Its Route

The AS 4 ←→ AS 3 BGP Session is dropped



A Changes Its Route

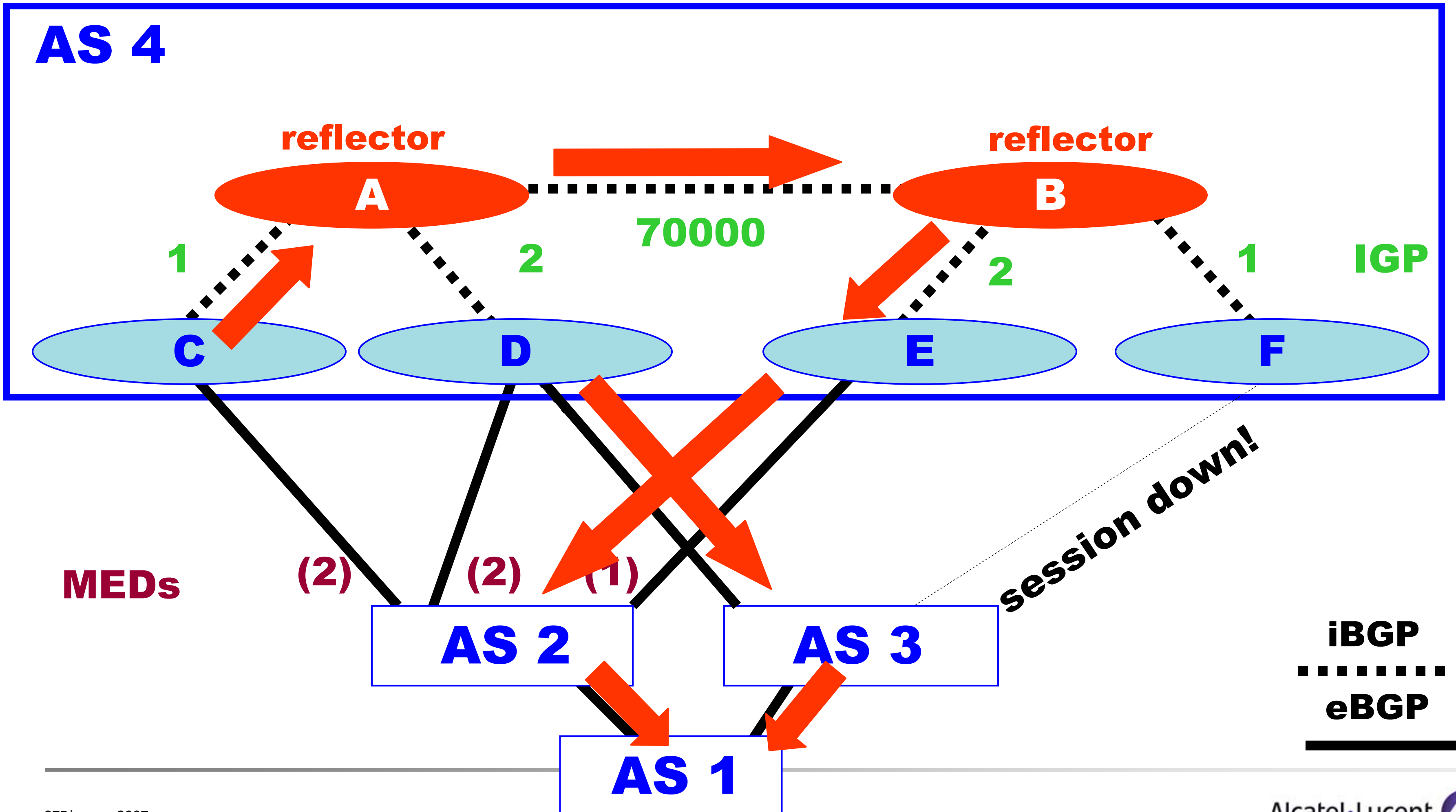
The MED 1 route from B beats the MED 2 routes that A sees from its clients....



CALL THIS STATE ZERO

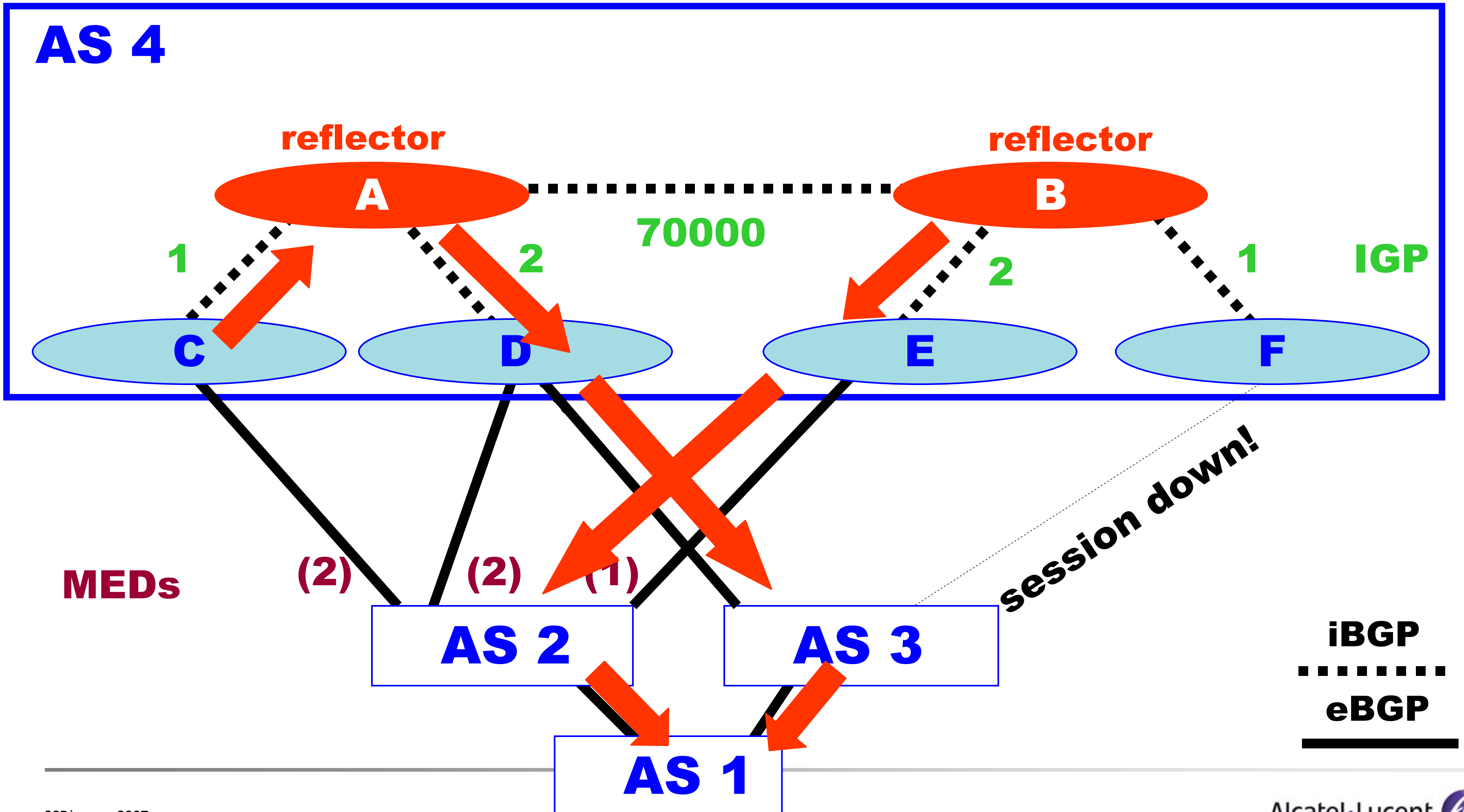
C & D Change Routes

The MED 1 route from A knocks both MED 2 routes out of the picture for C & D ...



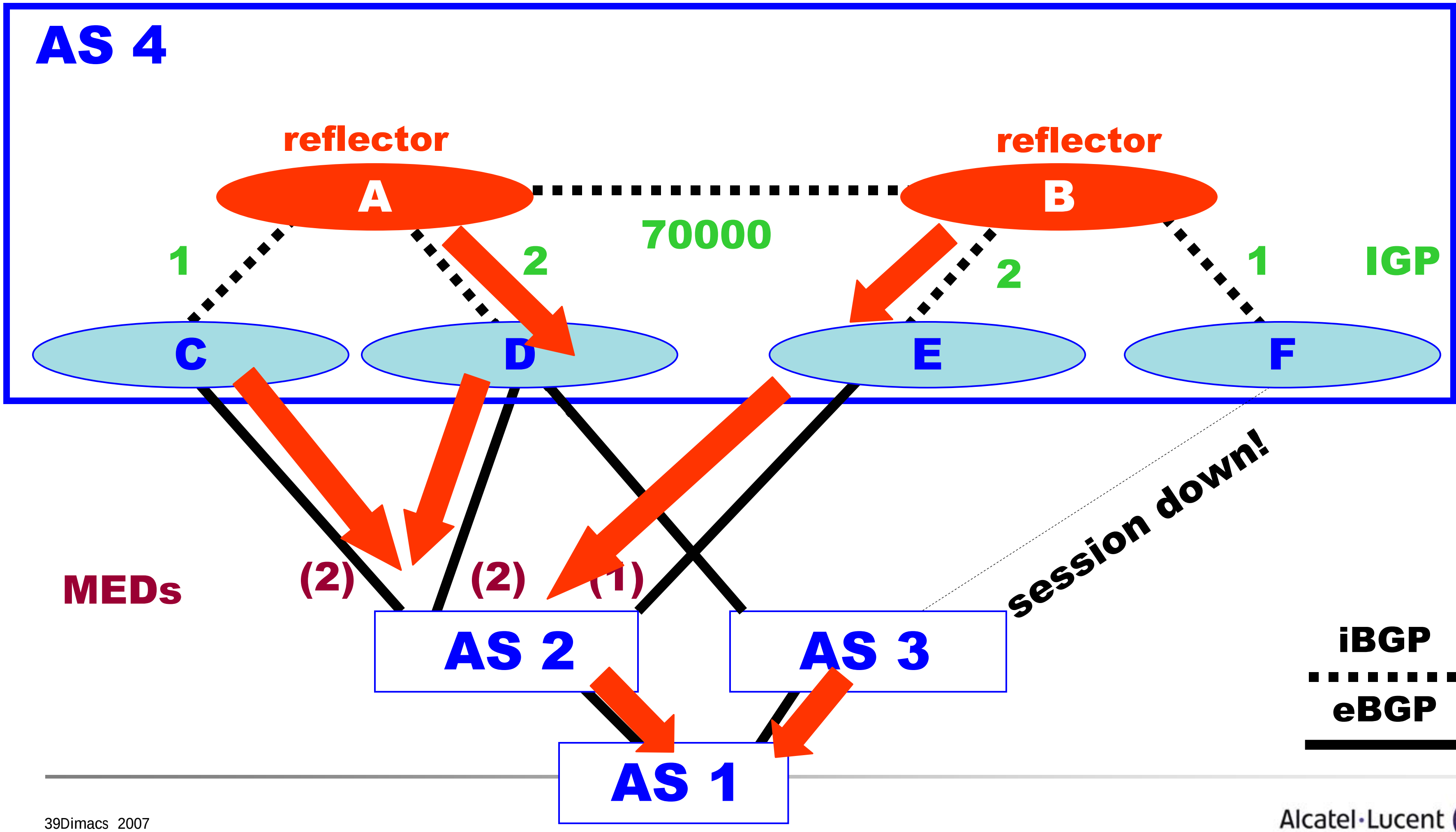
A Changes Route Again

A now sees the route from D through AS 3, and it is closer IGP-wise than the route from B...



C&D Return to Initial Routes

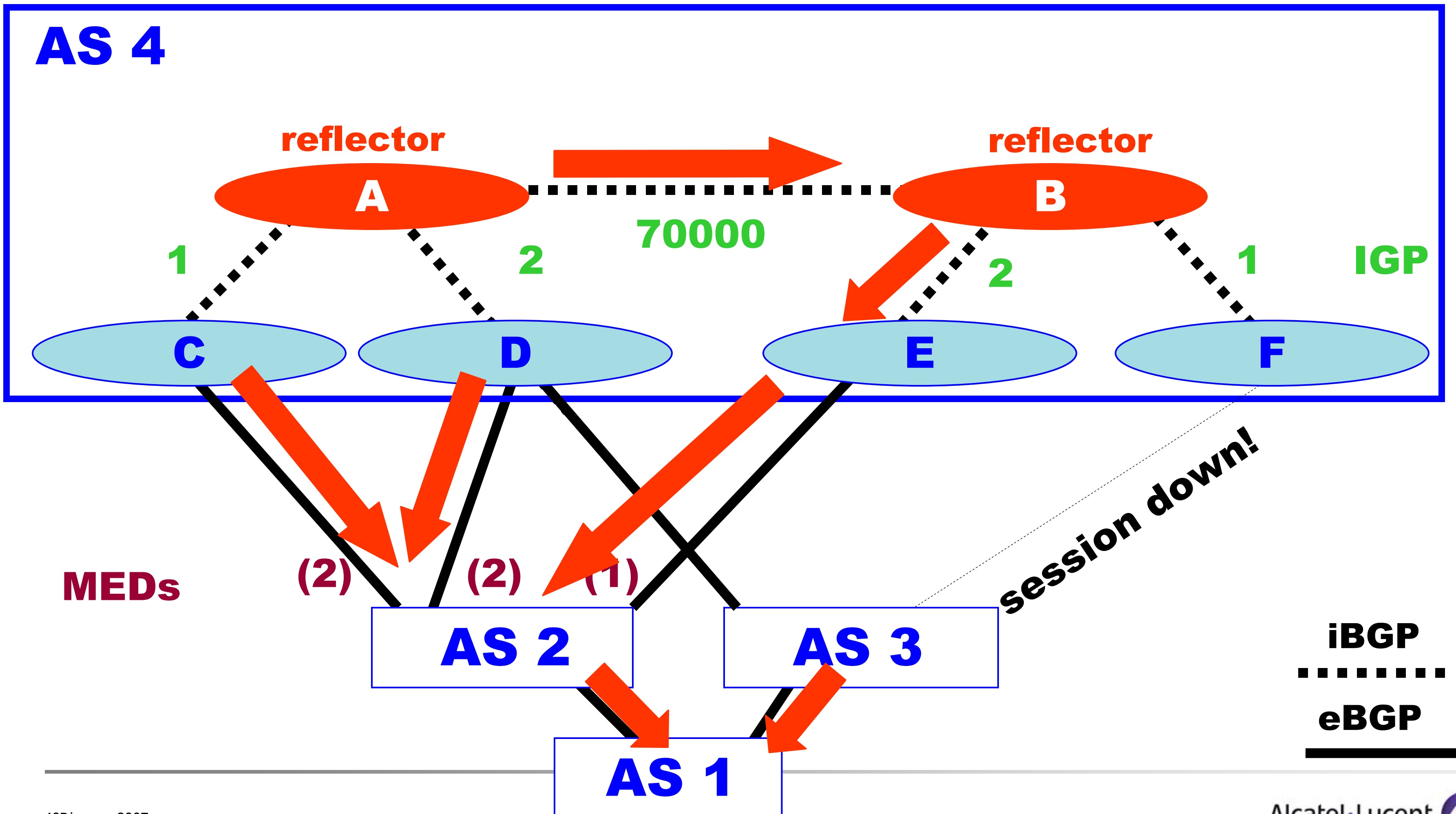
C & D no longer see MED 1 route from A, so they return to the eBGP routes with MED 2…



Alcatel·Lucent

Back to State Zero!

A switches back to MED 1 route through B.



What Can Be Done?

Possible Approaches:

- Use only configurations that guarantee no problems
  - No modification to BGP required
  - Previous example shows this might be difficult
- Prevent problems for any configuration
  - Modification to BGP required

I-BGP Modification

(1) Run selection process up through MED-comparison stage resulting in set of routes S

(2) Run remainder of selection process to determine best route R

(3) Advertise all routes in S (as opposed to announcing only best route R)

Highest Local Preference
Shortest ASPATH
Lowest MED (if same next AS)

S

i-BGP < e-BGP
Lowest IGP cost to BGP egress
Lowest router ID

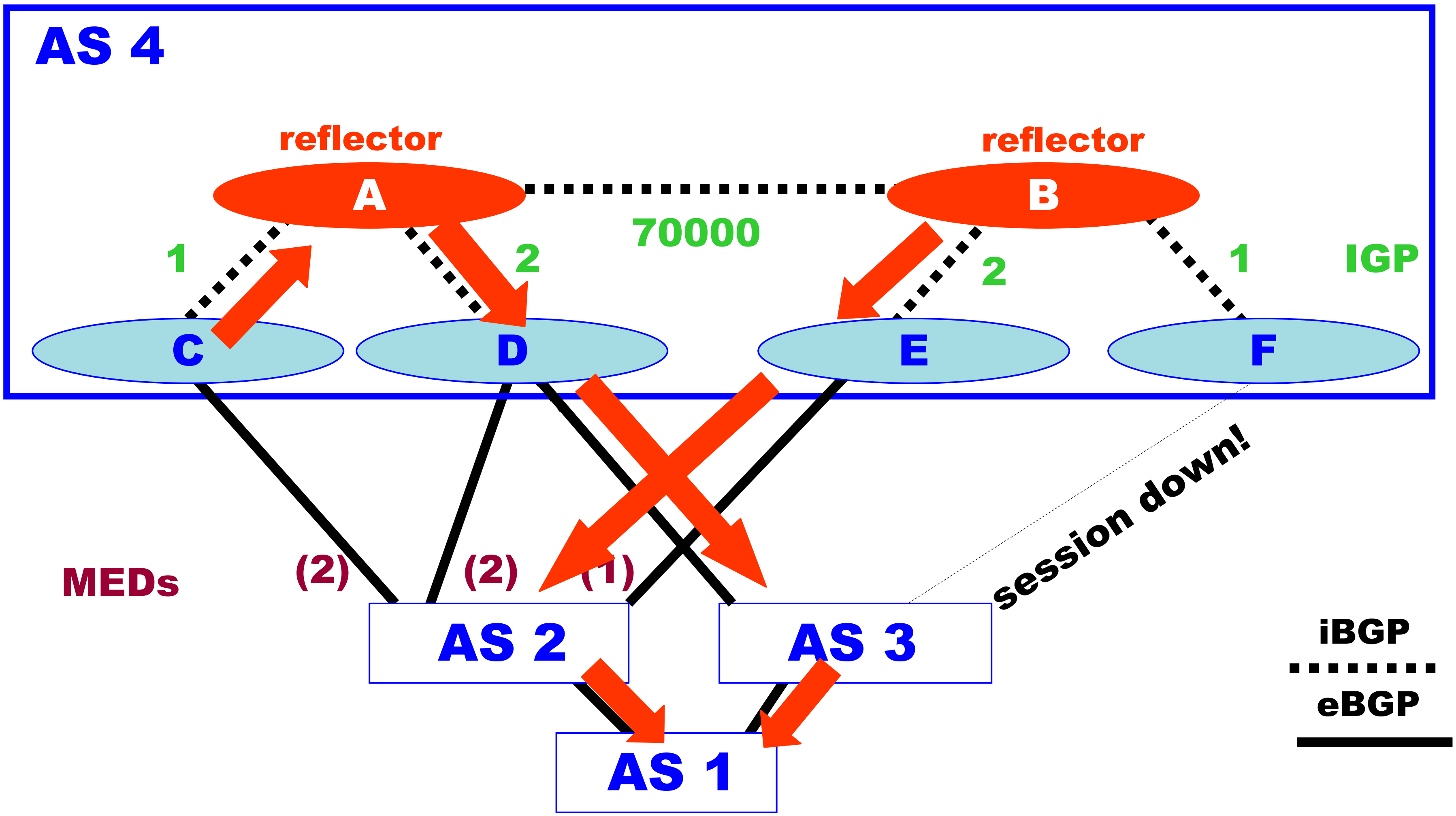
R

Modified I-BGP

(1) Modified I-BGP provably always converges (i.e., it's signaling safe)

(2) Modified I-BGP guarantees no forwarding loops (I.e., it's (almost) forwarding safe although there might be simple deflections)

Solution with modified BGP



A Fractional Model of BGP

Joint work with P. Haxell (U. Waterloo)

August 2007

Fractional SPP (fSPP)

- **Instance of fSPP: same as instance of SPP**
- **Solution to fSPP: assignment of non-negative weights w(P) to each path P that satisfy:**

(1) Unity: total weight of paths starting at each v, W(v), is at most 1

(2) Tree: For each vertex v and path S, total weight on paths from v that end with S is at most w(S)

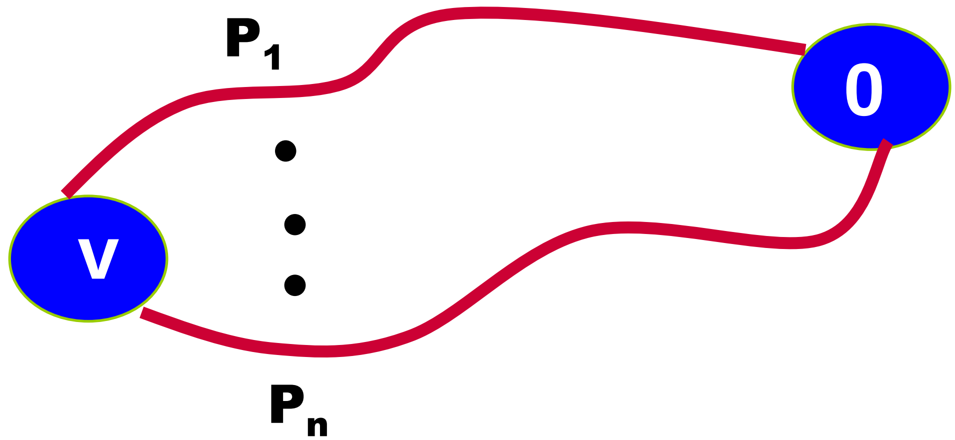
(3) Stablity: If Q starts at v then either:

(i) W(v)=1, if P starts at v and w(P)>0 then v prefers P to Q

(ii) there is a proper final segment S of Q where total weight on paths from v ending in S, ie $W_s(v)$, is s.t. $W_S(v)$=w(S) and if P starts at v with final segment S and w(P)>0 then v prefers P to Q

Unity Condition

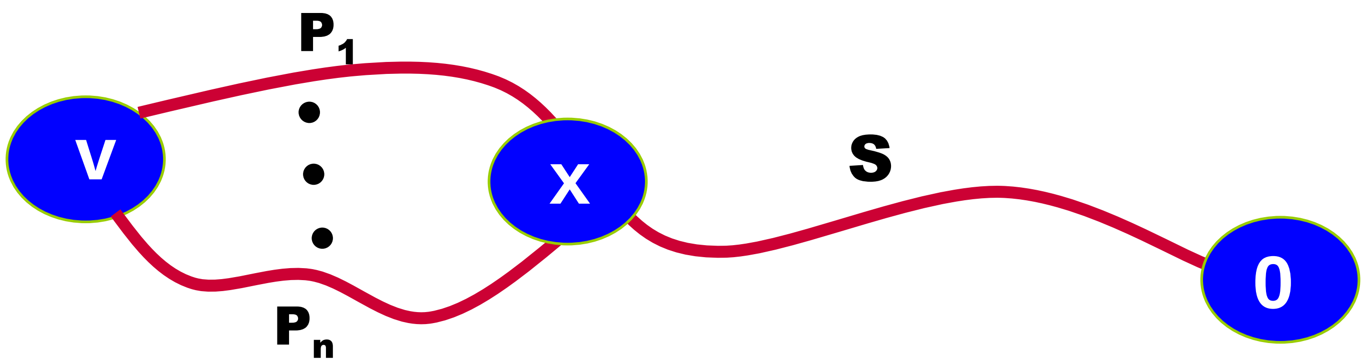
(1) Unity: total weight of paths starting at each v, W(v), is at most 1



$$w(P_1) + \dots + w(P_n) \leq 1$$

Tree Condition

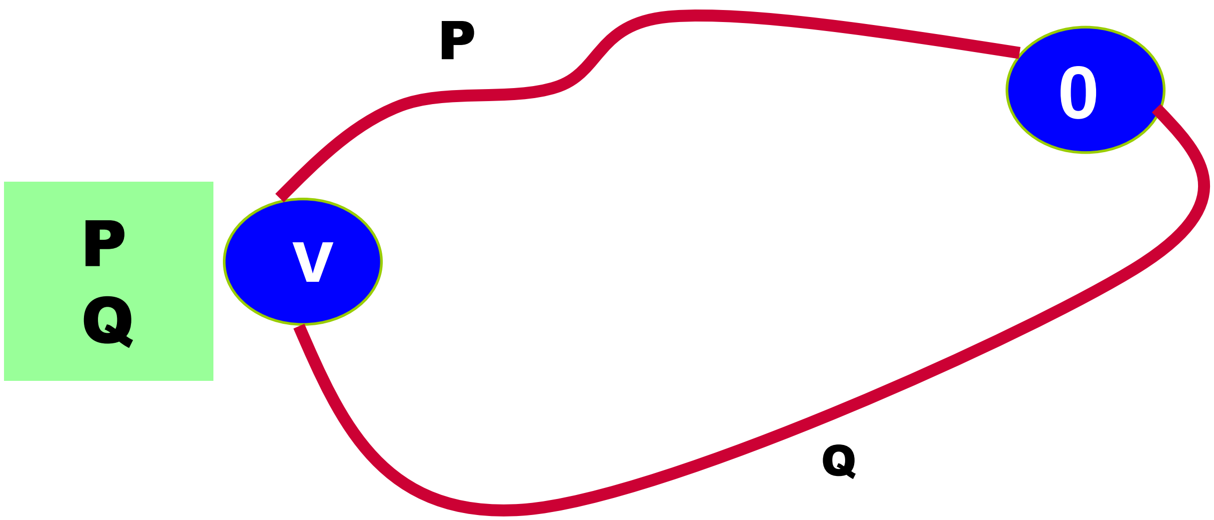
(2) Tree: For each vertex v and path S, total weight on paths from v that end with S is at most w(S)



$$w(P_1) + \ldots + w(P_n) \leq w(S)$$

Stability Condition (i)

(3) Stablity: if Q starts at v then either:
(i) W(v)=1, if P starts at v and w(P)>0 then v prefers P to Q



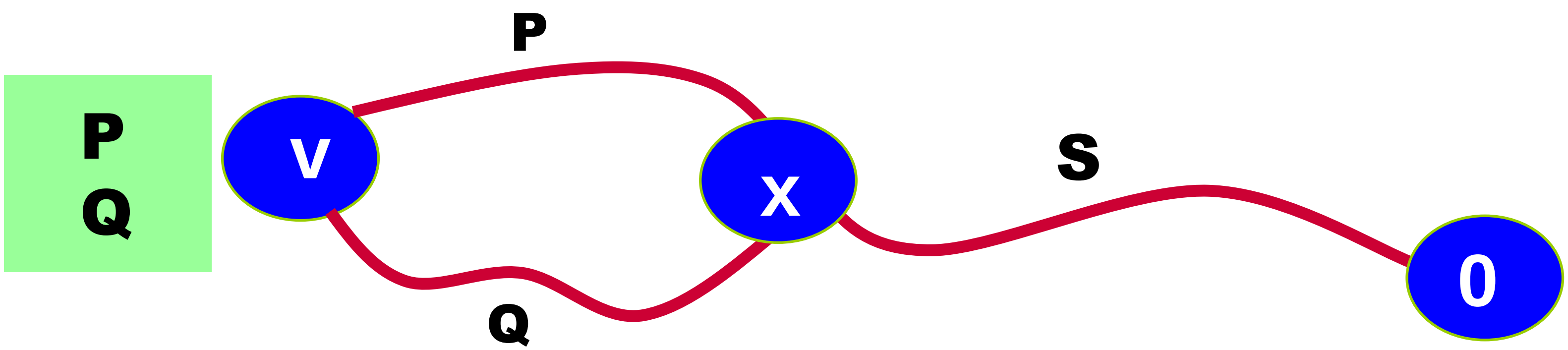
Σ w(R) = 1

w(P) > 0

Stability Condition (ii)

(3) Stablity: if Q starts at v then either:

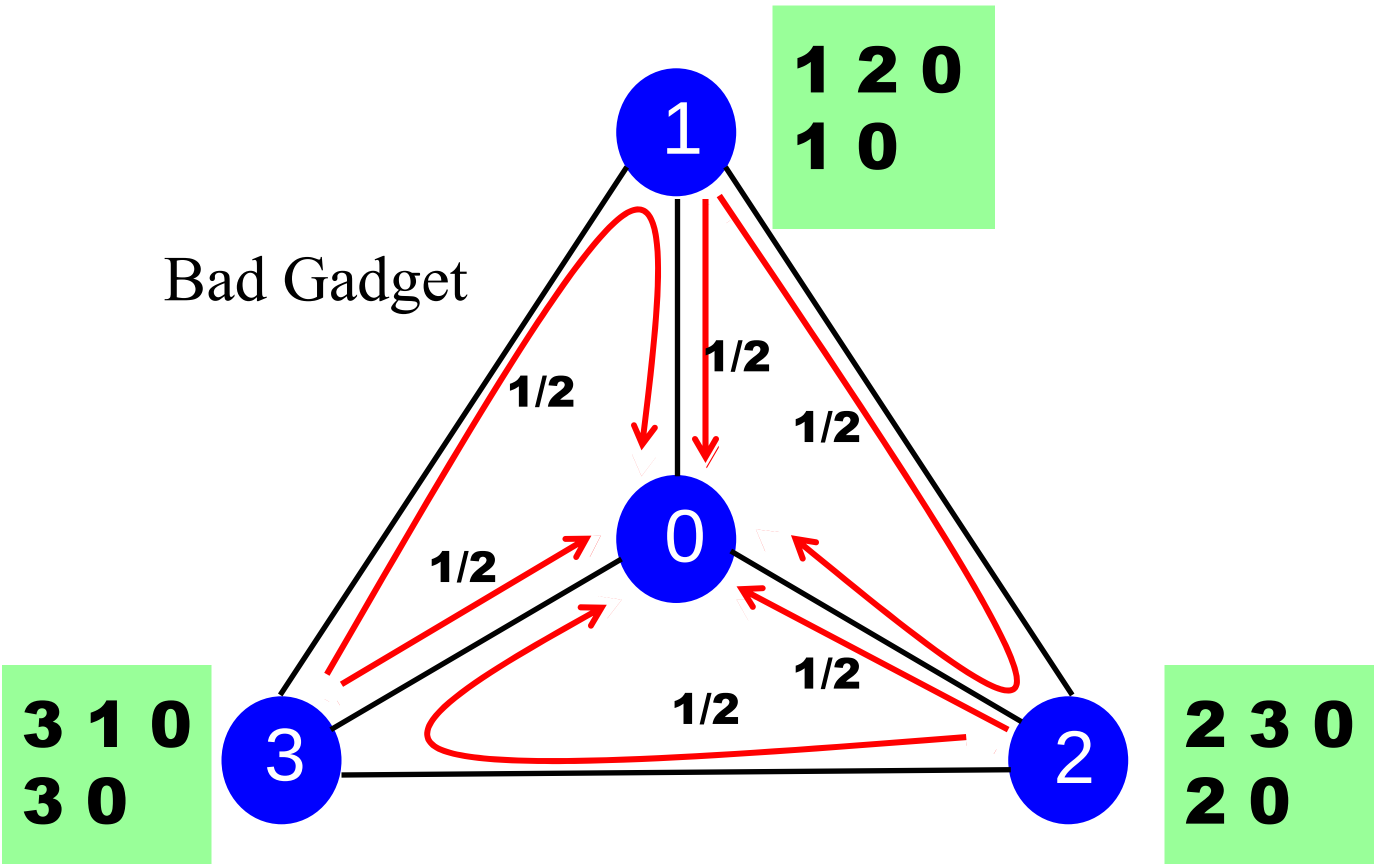
(ii) there is a proper final segment S of Q with $W_S(v)=w(S)$ and if P starts at v, w(P)>0 then v prefers P to Q



$$\Sigma_R\, w(R) = w(S)$$

$$w(P) > 0$$

Solution for Bad Gadget!



Solutions for fSPP

Theorem: A solution to fSPP always exists.

Scarf's Lemma

Let $n < m$ be positive integers, $b \in R^m_+$, B and C nxm matrices such that:

- first n columns of B are the identity matrix
- the set $\{x \in R^m_+ : Bx=b\}$ is bounded
- for c_{ik} $k>n$, $c_{ii} < c_{ik} < c_{ij}$ for each $j \neq i$, $j<n$.

Then there is $x \in R^m_+$ where $Bx=b$ and the set of columns S of C that correspond to supp(x)= $\{k : x_k \neq 0\}$ are such that for all columns j there is a row i such that $c_{ik} < c_{ij}$ for all $k \in \text{supp}(x)$.

A Little Game Theory

- **V a set of players**
- **S(v) a set of "strategies", for each player v**
- **strategy vector $(s_1, \ldots, s_n)$, s_i a strategy for v_i**
- **$P_i((s_1, \ldots, s_n))$, payoff for v_i given choice of strategy s_j for v_j**

A Nash equilibrium is a strategy vector such that no player can change its strategy and improve its payoff.

A Non-cooperative Game

- **The BGP Game**
  - **nodes of SPP instance are players**
  - **a player's strategies are choices of paths to the destination**
  - **payoff to a player directly related to preference of path chosen**
  - **payoff is -1 if choice not "consistent" with strategies of other nodes on the path**

Nash Equilibria of BGP Game

- **The solutions of SPP are the Nash equilibria of the BGP game and vice versa**
- **Some instances of BGP game have no Nash equilibria (Bad Gadget)**

Payoffs for fractional game

- **Utility of path P is some number directly related to preference or -1 if P is not consistent with the strategies of other players along P**
- **payoff to v is the weighted sum of the utilities of the paths originating at v**

Defines fractional BGP Game that is guaranteed to always have a Nash equilibrium.

Alcatel·Lucent

Degree Constrained Network Flows (STOC 07)

Joint work with P. Donovan, B. Shepherd, A. Vetta (McGill U)

August 2007

(Single Sink) Unconstrained Network Flows

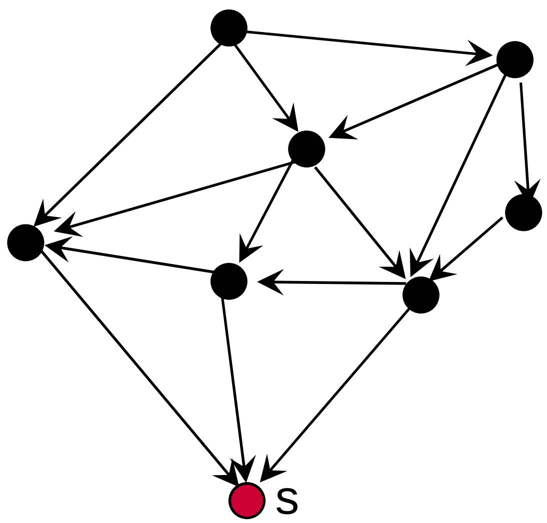
Given:

(1) directed network G=(V,A)

(2) sink node s

(3) demands d(v) from nodes v in V to s

find a flow that minimizes the max load at any non-sink node.



Alcatel·Lucent

Degree Constrained Network Flows

Given:

(1) directed network G=(V,A)

(2) sink node s

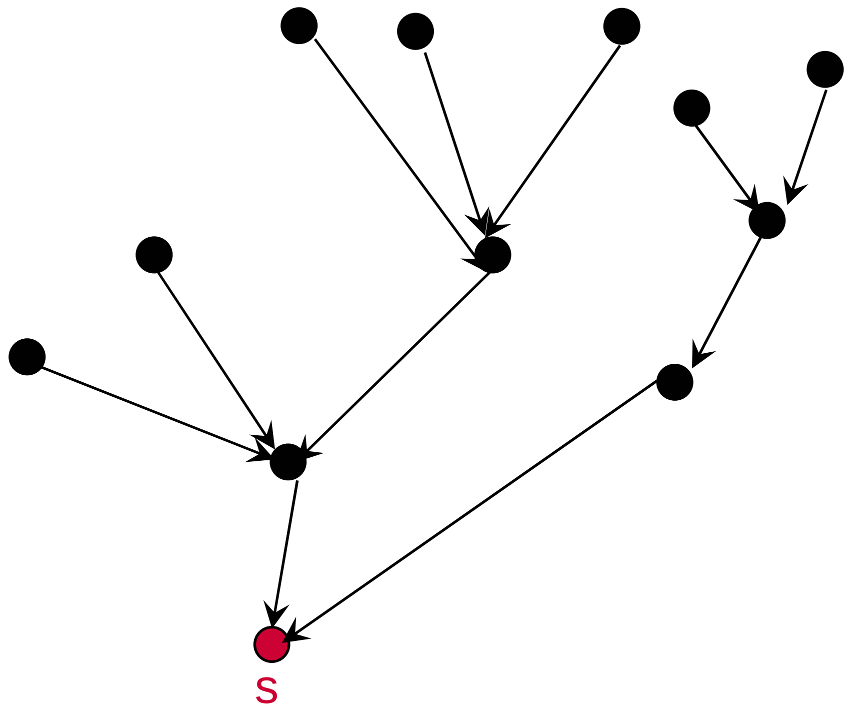
(3) demands from nodes in V to s

(4) outdegree bound d

find a flow that minimizes the max load at any non-sink node where for each node v in V the flow out of v is on at most d out-arcs.

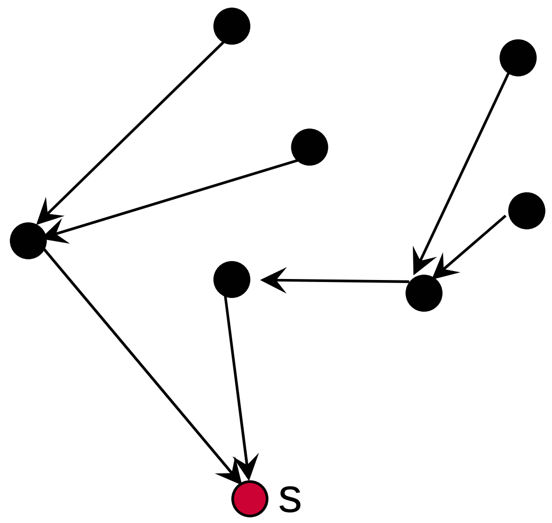
Confluent Flows

BGP results in the flow to a given destination s
to be a *confluent flow*.



Confluent Flows (d=1)

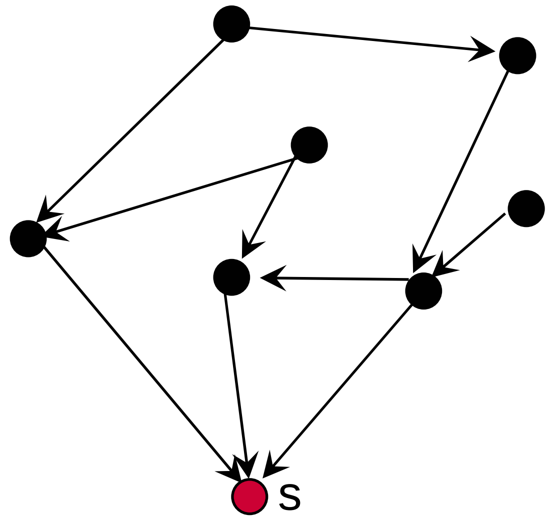
A *confluent flow* allows flow out of each node on 1 out-arc.



If unconstrained max load = 1, then can always find a confluent flow with max load O(log n) (and this is tight). [Chen et al 04]

Bifurcated Flows (d=2)

A *bifurcated flow* allows flow out of each node on at most 2 out-arcs.

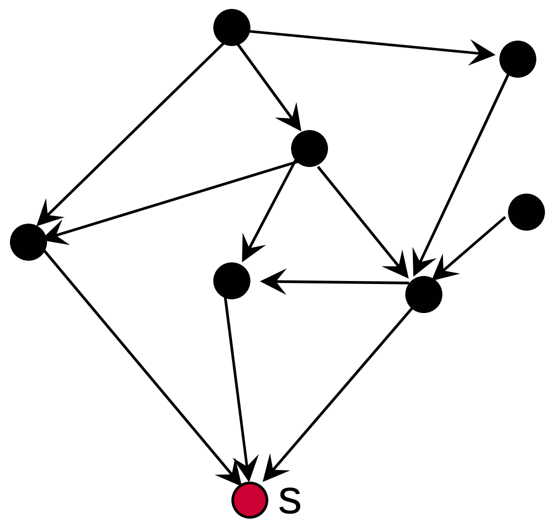


If unconstrained max load = 1, then can always find a bifurcated flow with max load $\leq$ 2 (and this is tight). [DSVW 07]

Min_load_bifurcated_flow is maxSNP-hard. [DSVW 07]

d-furcated flows (d $\geq$ 2)

A *d-furcated flow* allows flow out of each node on at most d out-arcs.

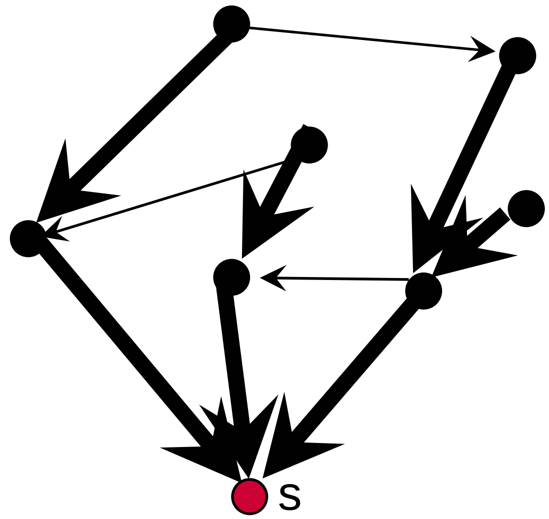


If unconstrained max load = 1, then can always find a d-furcated flow with max load $\leq$ d/(d-1) (and this is tight). [DSVW 07]

Alcatel·Lucent

B-confluent flows (what happens between d=1 and d=2 ??)

A *B-confluent flow* is a bifurcated flow where at each node v at least a B-fraction of the flow goes out on one arc.



For any B in [1/2,1) there is a B-confluent flow with max node load $\leq 1/(B-1)$.

Algorithm outline for bifurcated flow

(1) Find fractional single-sink flow F with minimum max load.

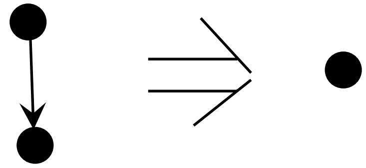
(2) Manipulate F into a "simpler" flow F' without changing max load.

(3) Transform F' into a bifurcated flow F'' with max load ≤ 2.

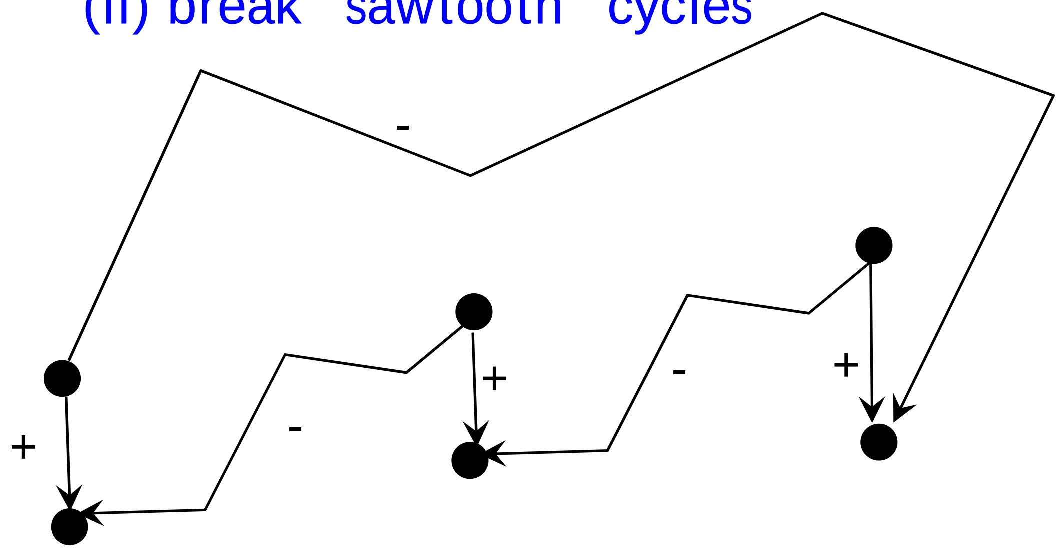
Algorithm outline (2)

(2) Manipulate F into a “simpler” flow F’ without changing max load.

(i) contract node with outdegree 1

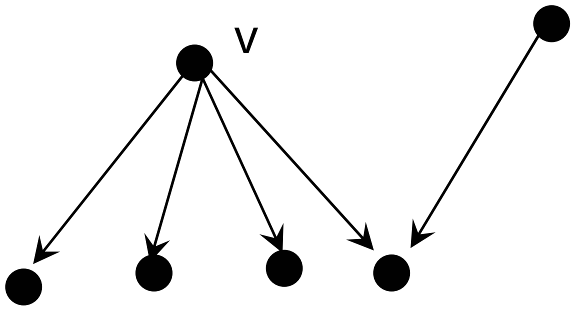


(ii) break “sawtooth” cycles



Algorithm outline (3)

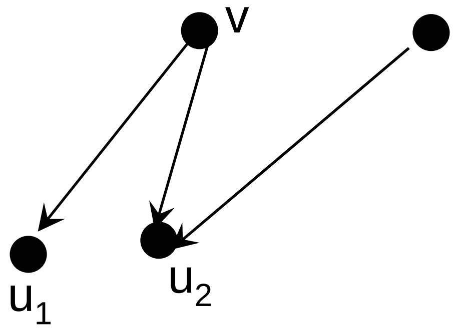
Theorem: There exists a source node v where all but at most 1 of v's neighbors has in-degree 1.



(3) Transform F' into a d-furcated flow F'' with max load $\leq$ 2.

Algorithm outline (3)

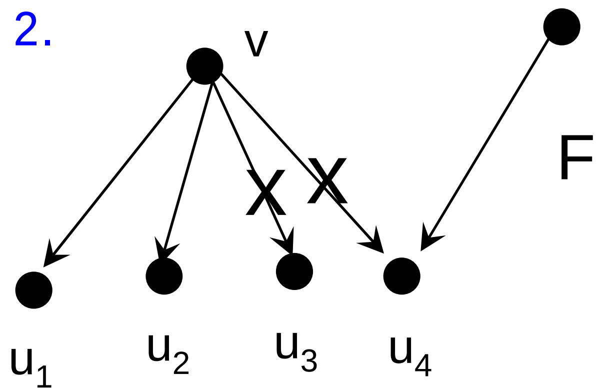
Case 1.



$f(u_1) = f(u_1) + extra(v) \leq f(u_1) + 1$

$f(u_2) = f(u_2)$

Case 2.



For i=1,2,

$$f(u_i) = f(u_i) + [extra(v) + f(u_3) + f(u_4)] / 2$$
$$\leq f(u_i) + (1 + 1) / 2$$
$$= f(u_i) + 1$$

Future

(1) Halfluent flows: a bifurcated flow where load is split evenly on outgoing arcs. Cases where halfluent load is twice as bad as bifurcated flow. Is this worst possible?

(2) What about capacitated versions?

(3) costs?

(4) multiple sinks?

Conclusions

- **BGP is extremely flexible, allowing operators to easily make obscure errors that are difficult to find and correct.**

- **Policy based routing is difficult to get right.**

- **Lots of open algorithmic problems in interdomain routing remain.**