Universität Karlsruhe (TH)
Institut für Telematik

TELEMATICS TECHNICAL REPORTS

Proceedings of the 1st Inter-Domain Routing Workshop (IDRWS 2003) 18th of September Karlsruhe, Germany

Lichtwald, Götz

May, 26th 2004

TM-2004-2

ISSN 1613-849X

http://doc.tm.uka.de/tr/

Institute of Telematics, University of Karlsruhe
Zirkel 2, D-76128 Karlsruhe, Germany

IDRWS 2003

Inter-Domain Routing Workshop 2003

Sponsor

Warum sind wir hier?

- Plattform für Diskussionen von Forscher, Betreiber und Hersteller
- Austausch von Ideen und Erfahrungen
- Praktische Aspekte
  - betriebliche Erfahrungen
  - Beobachtungen aus ISP - Sicht
- Neuartige Konzepte und Untersuchungen
- Lösungsansätze oder Entwicklungen in Richtung
  - zukünftiger IDR-Verbesserungen
  - neuer Routing-Architekturen

Agenda

11:00 – 11:30	Welcome
11:30 – 12:00	Götz Lichtwald; *Towards an Improvement of BGP Failure Handling*
12:00 – 12:30	Uwe Walter; *Explicit Routing Concepts*
12:30 – 13:30	Mittagessen
13:30 – 14:00	Olaf Maennel; *Observed properties of BGP convergence*
14:00 – 14:30	Lx Manhenke; *Routing-Konvergenz von RFC2547bis-VPNs*
14:30 – 15:00	Kaffeepause
15:00 – 15:30	Thomas Schwabe; *Policy based Calculation of the Internet Topology*
15:30 – 16:00	Stefan Mink; *IGB - full mess^Hh*
16:00 – ca. 17:00	Besichtigung des Rechenzentrums von Schlund+Partner AG

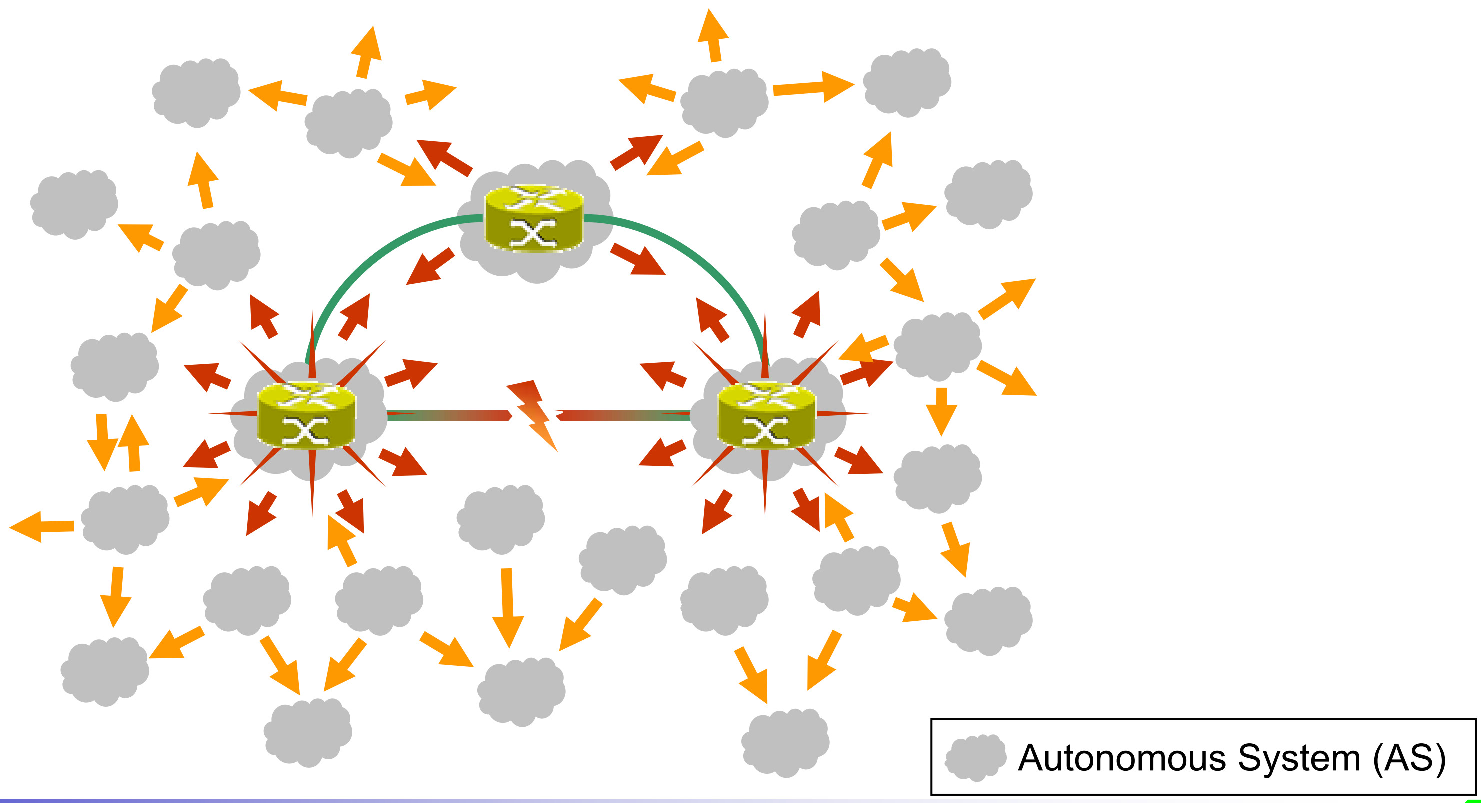
Towards an Improvement of BGP Failure Handling

IDRWS 2003

Roland Bless, **<u>Götz Lichtwald</u>**, Markus Schmidt, Martina Zitterbart
Institute of Telematics
University of Karlsruhe
Germany

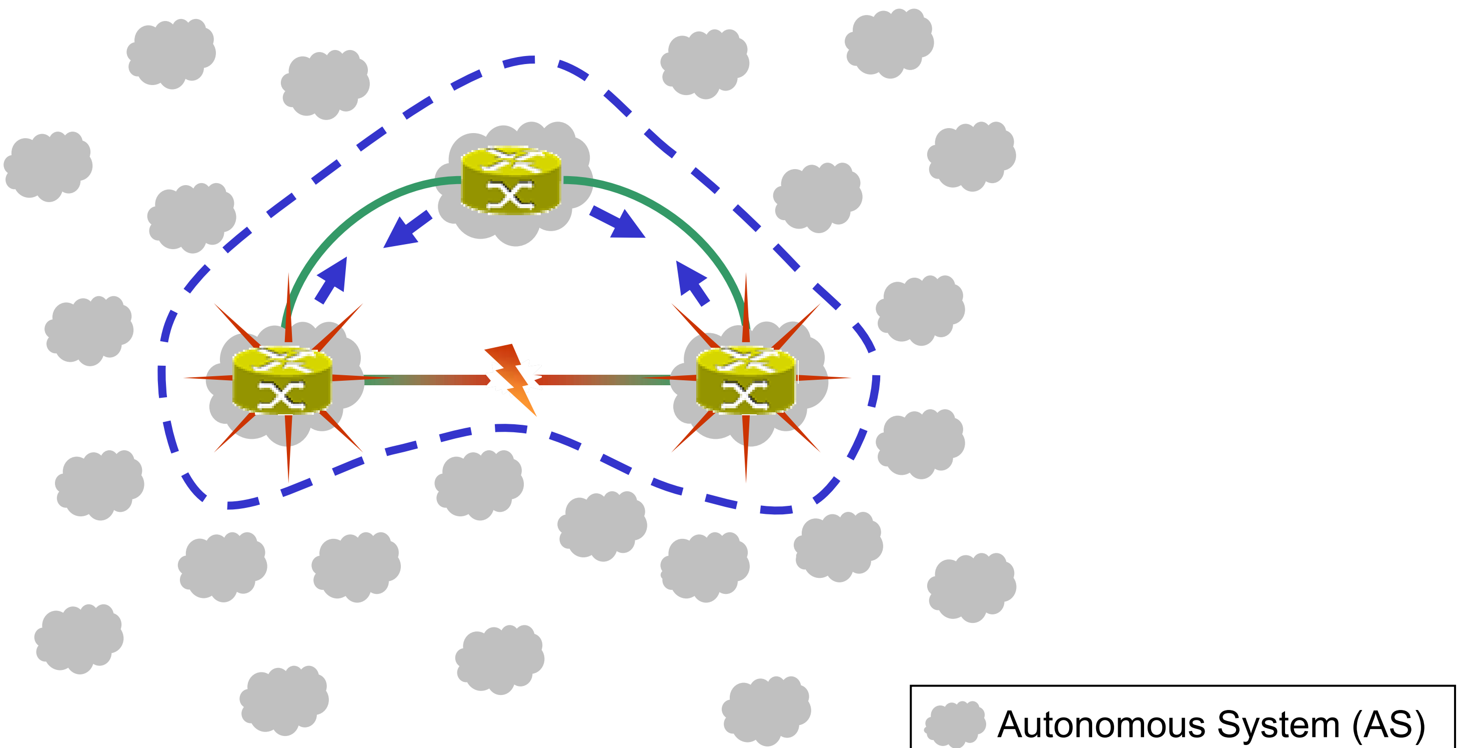
BGP suffers from:

- Slow convergence (2min – 10 min) ➔ bad for VoIP
- Too many update messages ➔ OS-Bugs, …
- Scope of update messages is not restricted
- Updates stress routers unnecessarily



Objectives

- Providing fast inter-domain failure reaction (faster than BGP)
- Improving convergence time
- Limiting propagation scope of update messages
- Reduction of resource consumption



Other approaches

Approaches like

- Route Flap Damping [RFC 2434]
  - Affects only flapping routes
  - Routes are suppressed although they are up again
    ➔ considered to be bad
- Graceful Restart [draft-ietf-idr-restart-06.txt]
  - Limits update storms on router restart

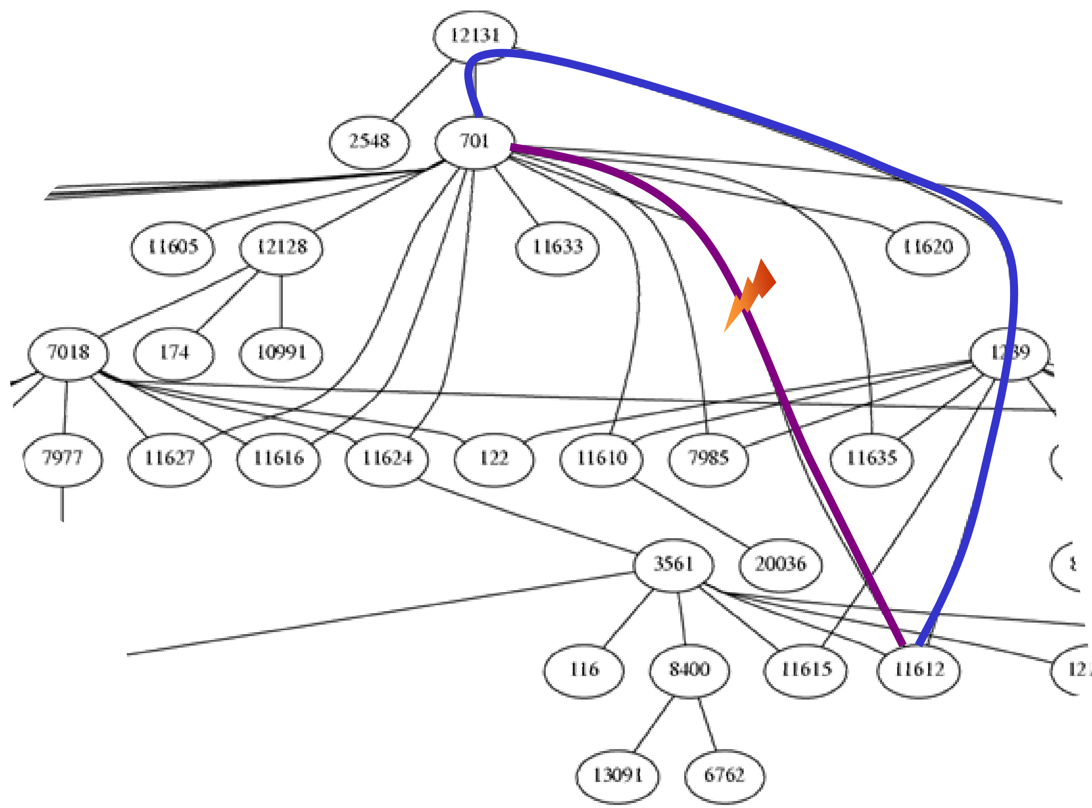
alleviate only a special symptom

Fast Scoped Rerouting

- Provides fast failure reaction
- Limits update storms
- Does not suppress good routes

Real AS Topology

Topology proves that often multiple bypasses are available



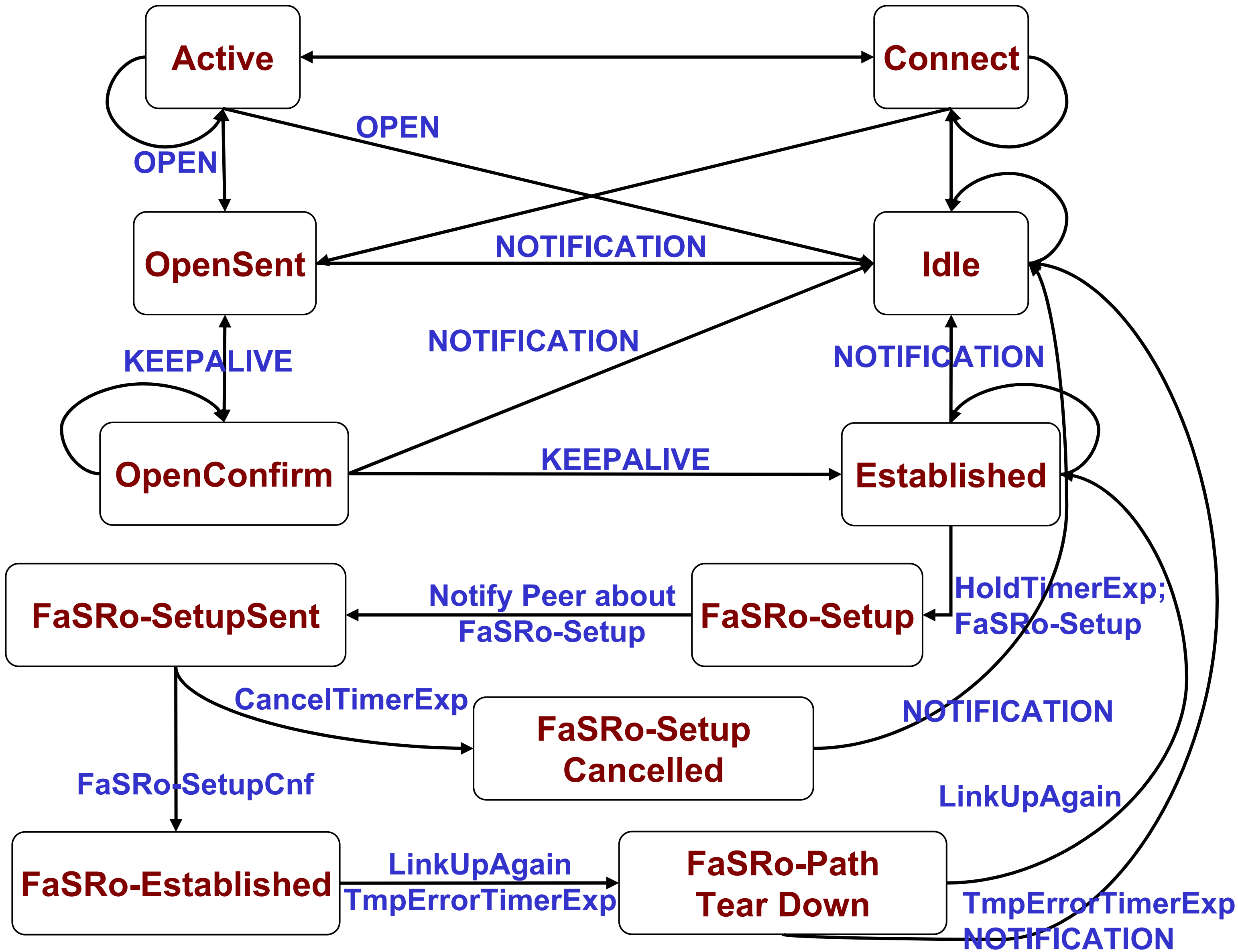
Source: http://www.routeviews.org/

Basic Concept

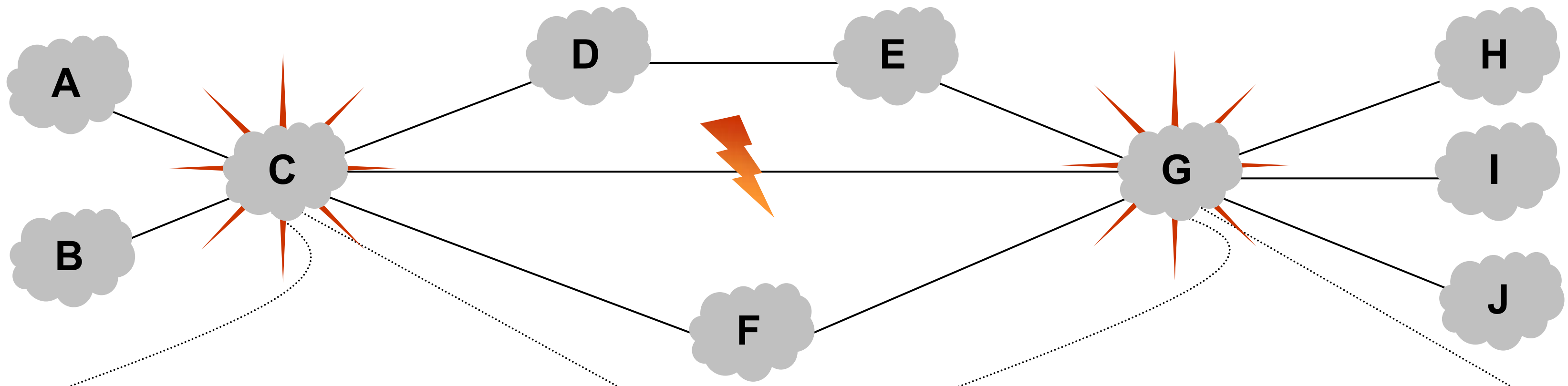
- No link failure ➔ $\text{BGP}_{\text{Fast Scoped Rerouting}} \approx$ BGP version 4
- Link failure ➔ Fast Scoped Rerouting (FaSRo) takes over
- Failure handling on **two time** scales:
  - Fine granular time scale (≤ 10 min) ➔ FaSRo
    - Setting up the FaSRo-Path
    - Traffic redirected to FaSRo-Path
    - link recovers
      - Switch back to BGP
    - link failures seems persistent
      - Switch back to BGP and start BGP update process
    - Short time link failure ➔ BGP is not affected
    - Failure duration exceeds a certain threshold ➔ BGP takes control for the failure reaction
  - Coarse granular time scale (> 10 min) ➔ BGP

Extending the FSM of BGP

- Motivation
- Objectives
- Other approaches
- Example
- Basic Concept
- **FSM**
- How it works
- Pros and Cons
- Evaluation
- Conclusion
- Outlook



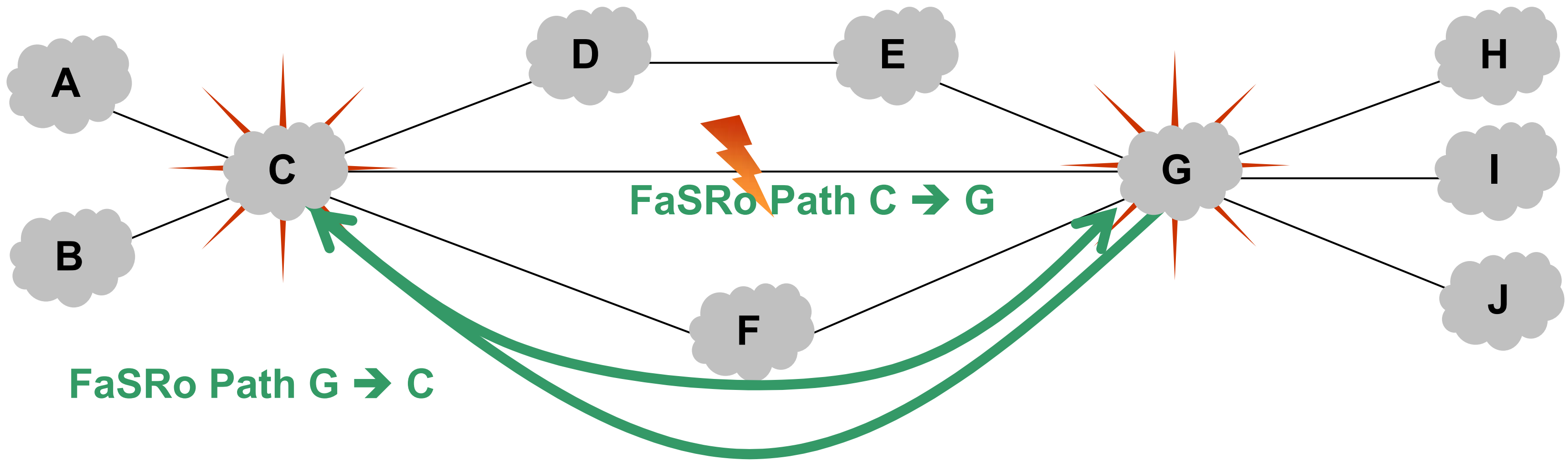
How it works (1)



	Network	Next Hop	Path
	…		
*>	network (H)	G	G H
*		D	D E G H
*		F	F G H
*	network (I)	D	D E G I
*		F	F G I
*>		G	G I
*>	network (J)	G	G J
*		D	D E G J
*		F	F G J
	…		

	Network	Next Hop	Path
	…		
*	network (C)	F	F C
*>		C	C
*		E	E D C
	…		

How it works (2)

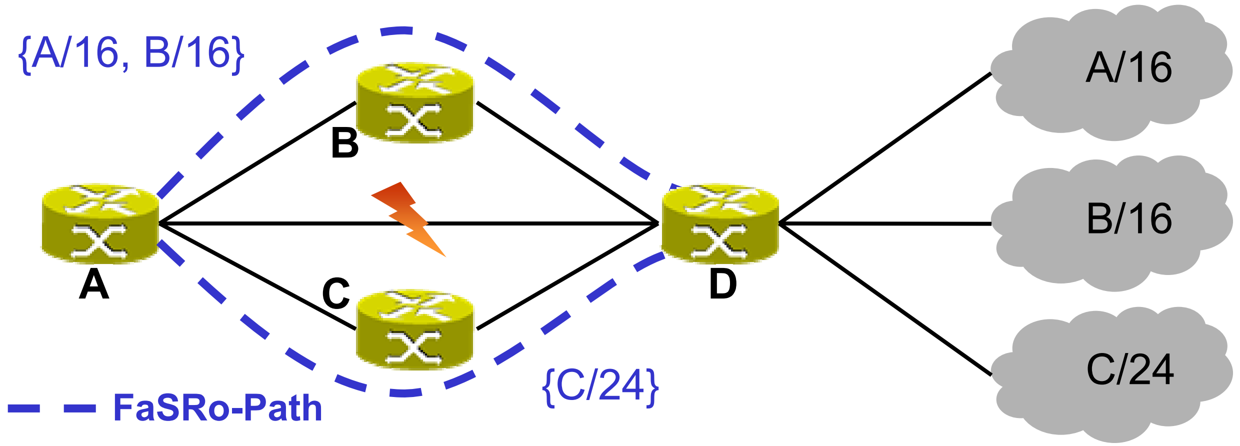


	Network	Next Hop	Path
	…		
~~*>~~	~~network (H)~~	~~G~~	~~G H~~
*		D	D E G H
*>		F	F G H
*	network (I)	D	D E G I
*>		F	F G I
~~*>~~		~~G~~	~~G I~~
~~*>~~	~~network (J)~~	~~G~~	~~G J~~
*		D	D E G J
*		F	F G J
	…		

	Network	Next Hop	Path
	…		
*>	network (C)	F	F C
~~*>~~		~~C~~	~~C~~
*		E	E D C
	…		

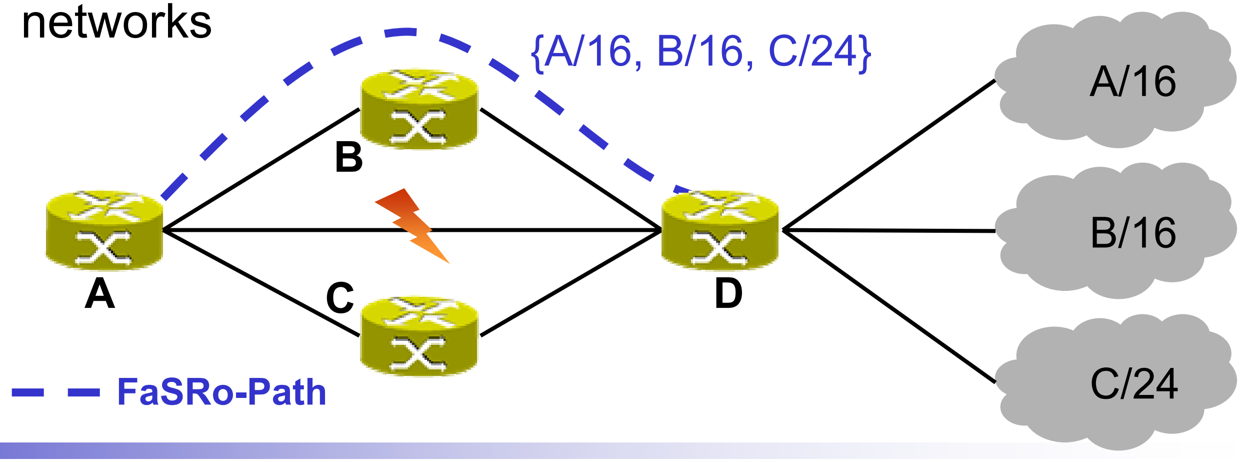
FaSRo-Fan

- Setting up a FaSRo-Path per destination network



Only ONE FaSRo-Path

- Providing only one FaSRo-Path for all destination networks



Pros and Cons of the FaSRo-Variations

Fan-Variant

- ☺ No policy violation
- ☺ Switching traffic to alternative path
- ☺ Simpler than BGP
- ☹ Signaling overhead (FaSRo-Path per destination network)
- ☹ Per destination network a FaSRo-Path has to be maintained
- ☹ Not optimal routes for a short period of time

One-Path-Variant

- ☺ Only one substitution for the broken link
- ☺ Less signaling effort to set up and maintain the FaSRo-Path
- ☺ Simpler than BGP
- ☹ Short time policy violations
- ☹ Risk of bandwidth scarcity
- ☹ Not optimal routes for a short period of time

One-Path-Variant makes sense, as only short time failures are handled!

Simulation results

	16 ASe (hand made)		10 ASe (BRITE)		20 ASe (BRITE)	
	FaSRo	BGP[1]	FaSRo	BGP[1]	FaSRo	BGP[1]
Message ratio	33%	100%	8%	100%	5%	100%
Convergence time ratio	88%	100%	58%	100%	49%	100%

[1] – Light weight BGP implementation

- Definition of **Message ratio**
  - BGP, FaSRo : $\frac{\#updates}{Total\ number\ of\ BGP\ updates}$
- Definition of **Convergence time ratio**
  - BGP, FaSRo : $\frac{\Delta(start\_failure, failure\_handled)}{\Delta(start\_failure_{BGP}, failure\_handled_{BGP})}$

Conclusion

- Failure handling based on **two time** scales
- Simple and fast mechanism
- Fast inter-domain failure reaction
- Improving convergence time
- Limiting propagation scope of update messages
- Reduction of resource consumption
- Easy to install

Outlook - Next Steps

- Looking at Policy violations
- Whispering Withdraw
- Extending simulations

Questions ?

Comments !

Götz Lichtwald

lichtwald@tm.uka.de

Institut für Telematik
Universität Karlsruhe

Explicit Routing Concepts

Uwe Walter

Classification of Routing Architectures

exchanged Data / Path selection	Distance Vector (DV)	Map Distribution (MD)
Hop-by-Hop	Internet Today ❑ BGP ❑ RIP ❑ IGRP	❑ OSPF ❑ IS-IS
Local / Explicit (Source Routing)		❑ IDPR, PNNI, Nimrod ❑ Bananas ❑ NIRA

Intermediates...

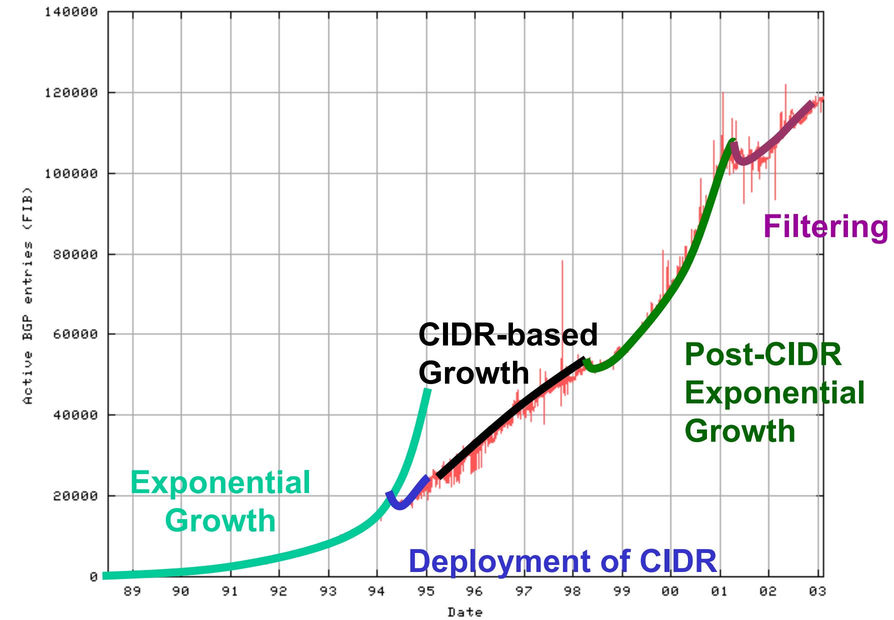
Future Routing Alternatives?

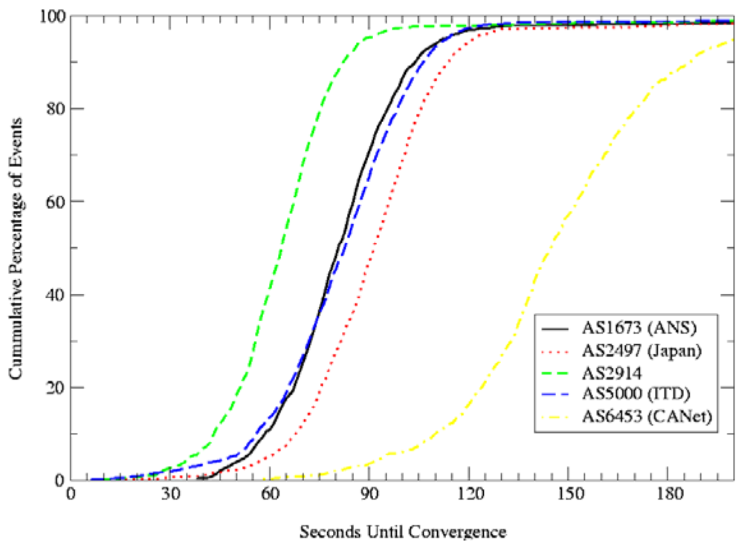
Some Advantages and Disadvantages

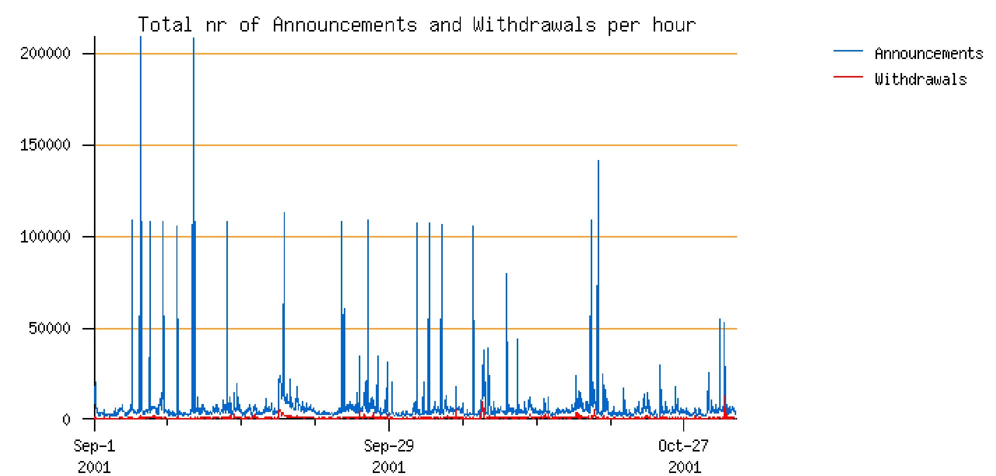
- Hop-by-Hop Path selection
  - ☹ Need global consistency to avoid loops
- Explicit Path selection
  - ☺ Allow trivial testing and deployment of new algorithms
  - ☺ More user control about path selection
  - ☺ More immune against loops
  - ☺ Higher robustness against attacks (at least: scope is reduced)
  - ☺ Some optimization problems cannot be handled by Hop-by-Hop architectures
- Distance Vector
  - ☺ Need fewer ressources (CPU & Memory) than Maps
- Map Distribution
  - ☺ Easier to harden against attacks
  - ☺ Can react faster to (e.g. topology) changes and stabilize faster
  - ☺ Need less bandwidth during significant changes
  - ☺ Policy Routing (e.g. QoS, security) possible

Motivation

- Some BGP problems / problems-to-come:
  - Scalability (Prefix growth)
  - Update message load & Convergence time
  - Growing Policy Configuration „nightmares“
  - Only metric: Shortest-AS-path

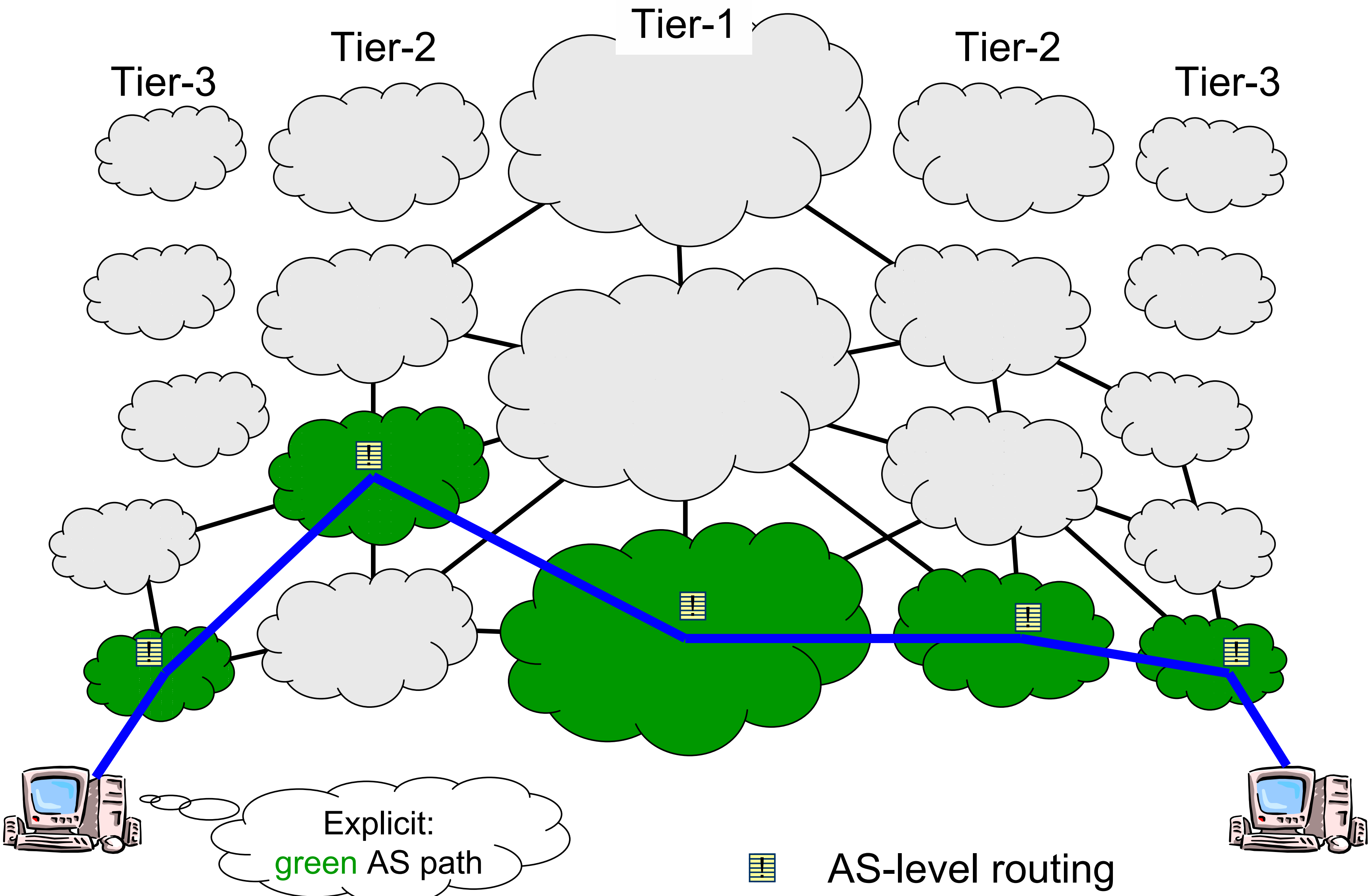






What?

Concept: Explicit Routing on AS-level



Advantages

- Smaller routing tables (currently factor 10 + ASes grow slower than hosts)
- Transparent migration possible (no flag day necessary)
- Enables explicit path selection
  (Users may enforce their own policies)
  - Flexible routing options possible
    (QoS, Inclusion, Exclusion, etc.)
  - Trivial testing and deployment of new routing algorithms
  - Fast failure reaction (user just selects new path)
  - Encourages network modernization and advances new technologies by stimulating competition
  - More immune to router errors
  - BUT: provider compensation must be ensured

Disadvantages & Problems

- Probably: No fixed header length (breaks fast-path?)
- Asymmetric routing: Path back?
- Fine granularity of prefix-based routing and its advantages could easily be lost
  - Load-balancing
  - Multi-Homing
  - Fine grained Policies

How?

Header options

Version (4)	Header Length (4)
Type of Service (8)	
Total Length (16)	
Identifier (16)	
Flags (3)	Fragment Offset (13)
Time to Live (8)	
Protocol (8)	
Header Checksum (16)	
Source Address (32)	
Destination Address (32)	
Options: **Destination AS: 553**	
Data (variabel)	

Version (4)	Header Length (4)
Type of Service (8)	
Total Length (16)	
Identifier (16)	
Flags (3)	Fragment Offset (13)
Time to Live (8)	
Protocol (8)	
Header Checksum (16)	
Source Address (32)	
Destination Address (32)	
Options: **AS Path: 340 720 1060 332 553**	
Data (variabel)	

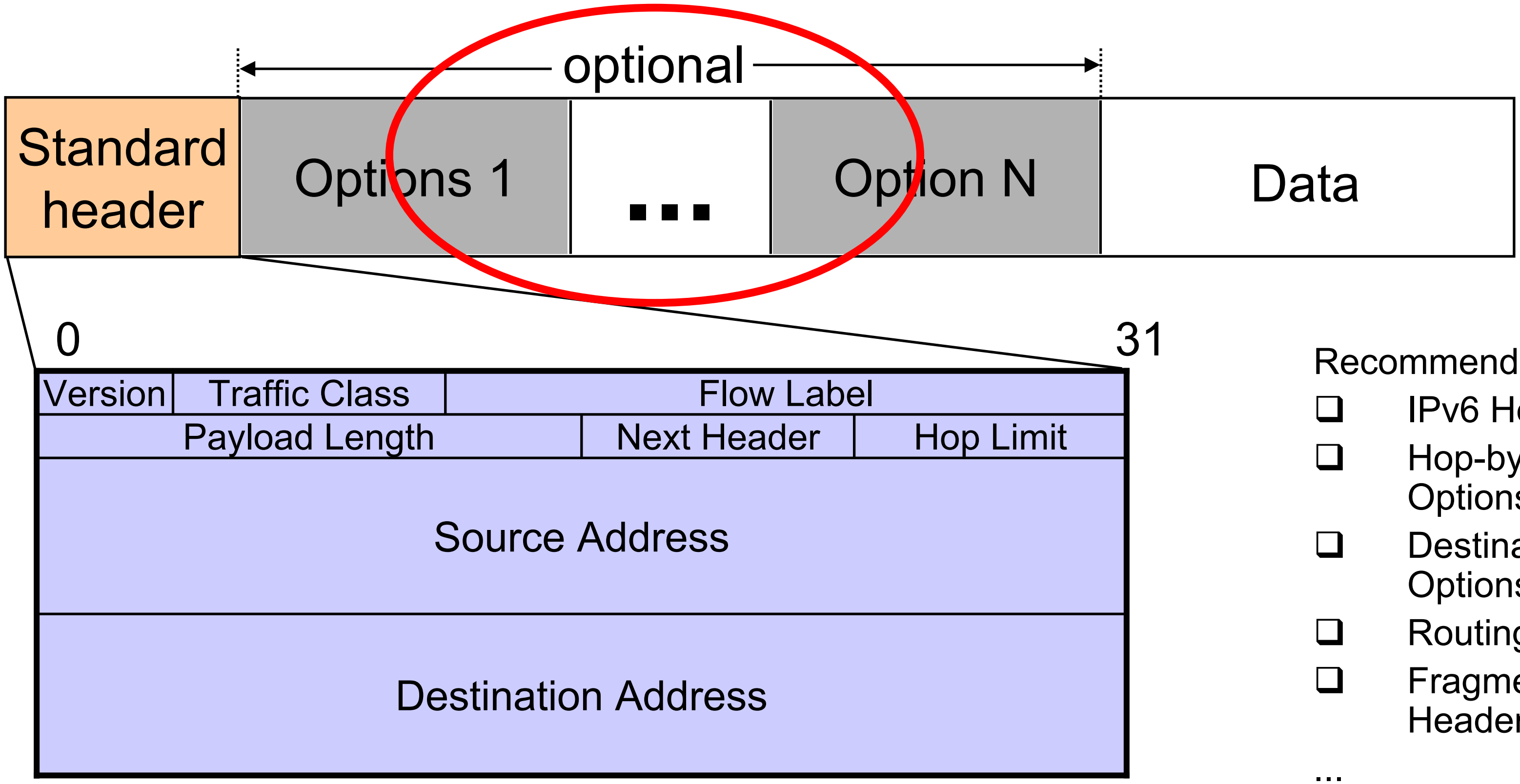
or: Prepend with MPLS-like header?

Problems:

- Overhead
- Routers‘ Fast-Path

Header options (IPv6)

Header options (similar to IPv4):



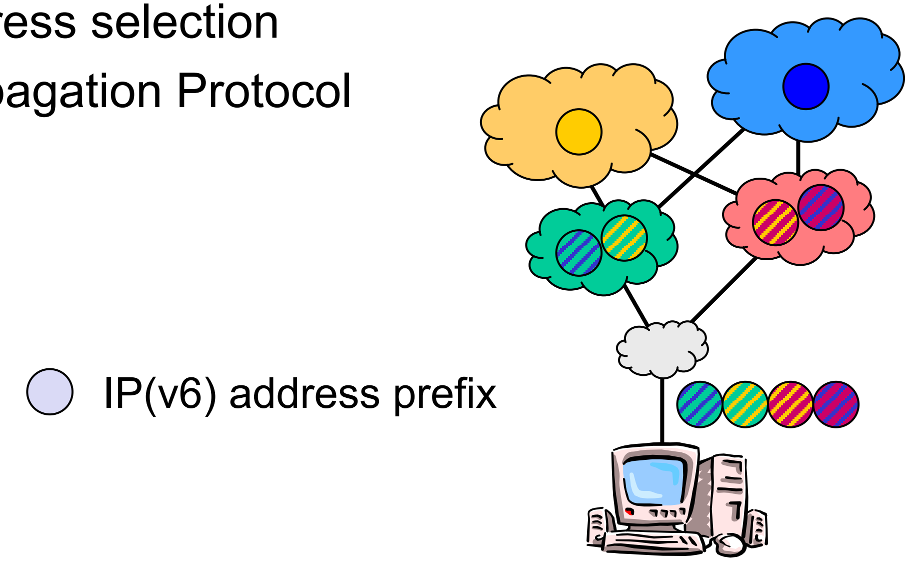
Recommended order:

- IPv6 Header
- Hop-by-Hop Options
- Destination Options
- Routing Header
- Fragment Header

...

Related work

- NIRA – New Internet Routing Architecture
  - Strictly hierarchical provider-rooted IP(v6) address scheme
  - AS Path selection via address selection
  - Topology Information Propagation Protocol



- BANANAS – Evolutionary Framework for Explicit and Multipath Routing
  - see SIGCOMM 2003
  - Encoding of Paths through global hash IDs
- NIMROD (1996)
- ...

Path Selection

1. Client does not set path at all ⇨ ISP choses default one

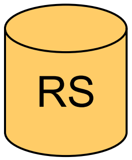
   (e.g. when receiving the packet at the first-hop-router)

2. Client downloads policies and selects path himself

3. Client requests suitable paths (according to its criterias)

   from ISP‘s *Routing Server*

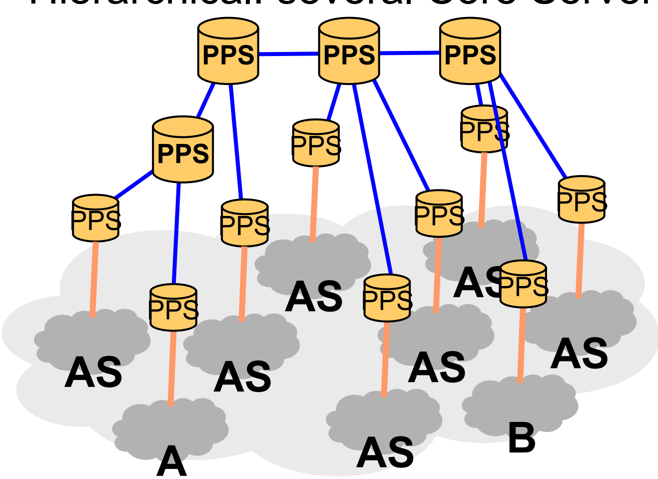
RS - Routing Server



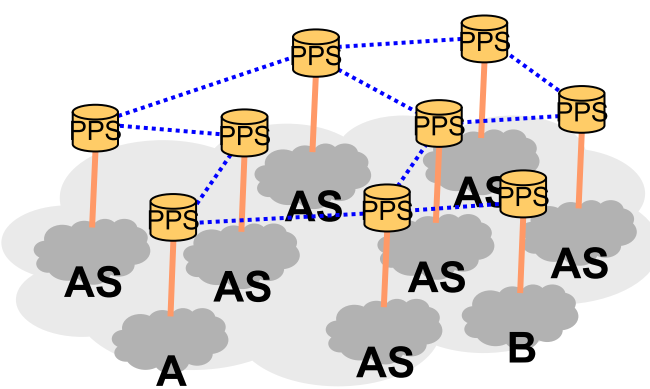
- Stores information about the complete Internet topology
- Offers (public) policies to be queried
- Information gathering:
  - Looking-glass concept:
    Listening to BGP
    - Cannot gather unfiltered information!
  - Active notifications of all ASes:
    ISPs upload their policies
    - information out-of-date problems

Distributed System:

- Hierarchical: several Core Server (DNS-like)



- Overlay-Net (Peer2Peer-like)



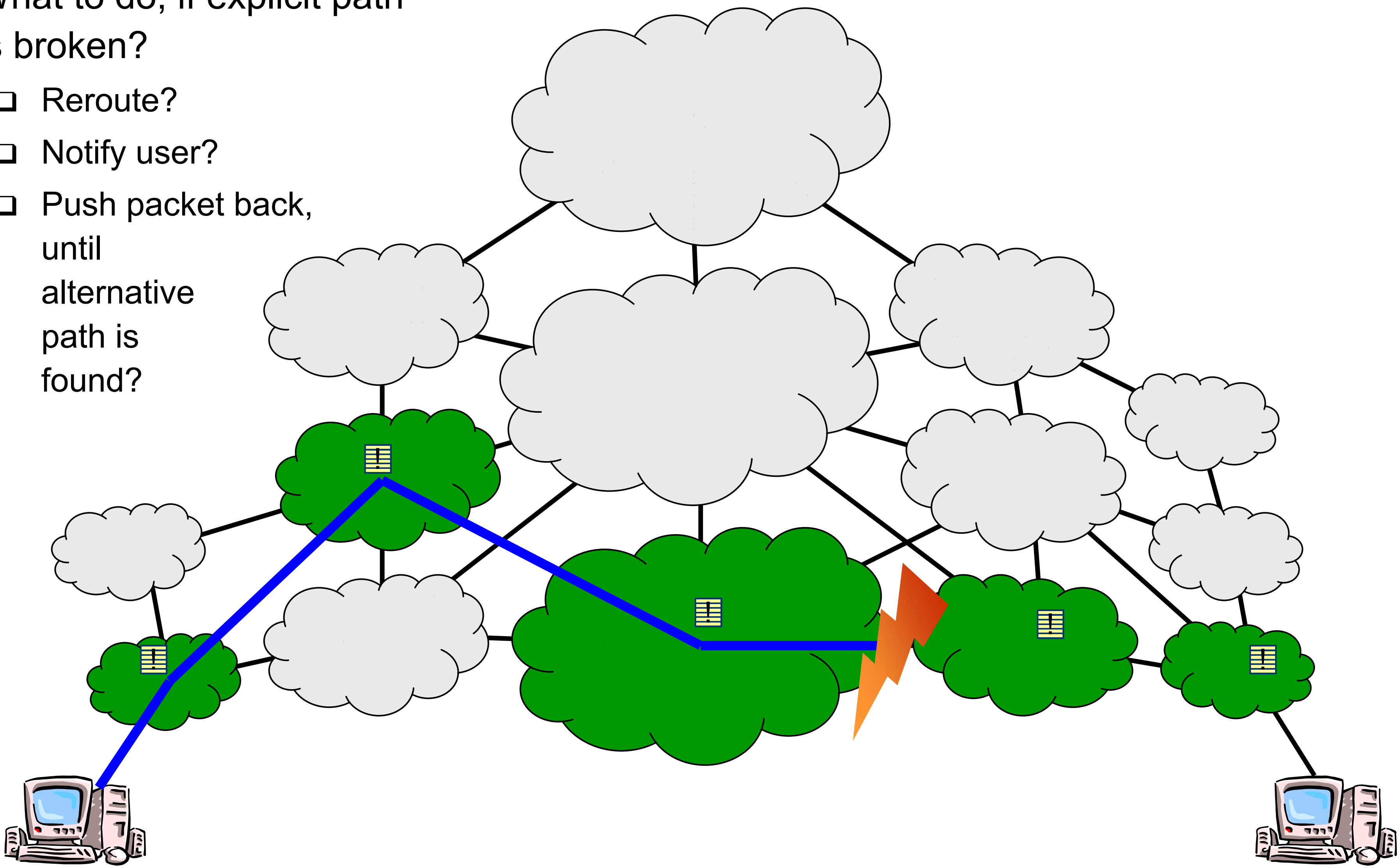
Routing Server: Pros and Cons

- ☺ Pro
  - New policies can be verified (interesting!?)
  - Propagation of BGP updates could be limited
  - All (public) policies are retrievable
  - Coherent view of topology possible
  - More additional information storable
    (AS supports QoS, NGN services, etc.)...

- ☹ Cons
  - Failure Case unresolved (see next slide)
  - Some (important?) Policies are confidential
  - New Point-of-Failures
  - Routing Server-Load (Update storms, DDOS attacks)
  - Memory & Processing power?

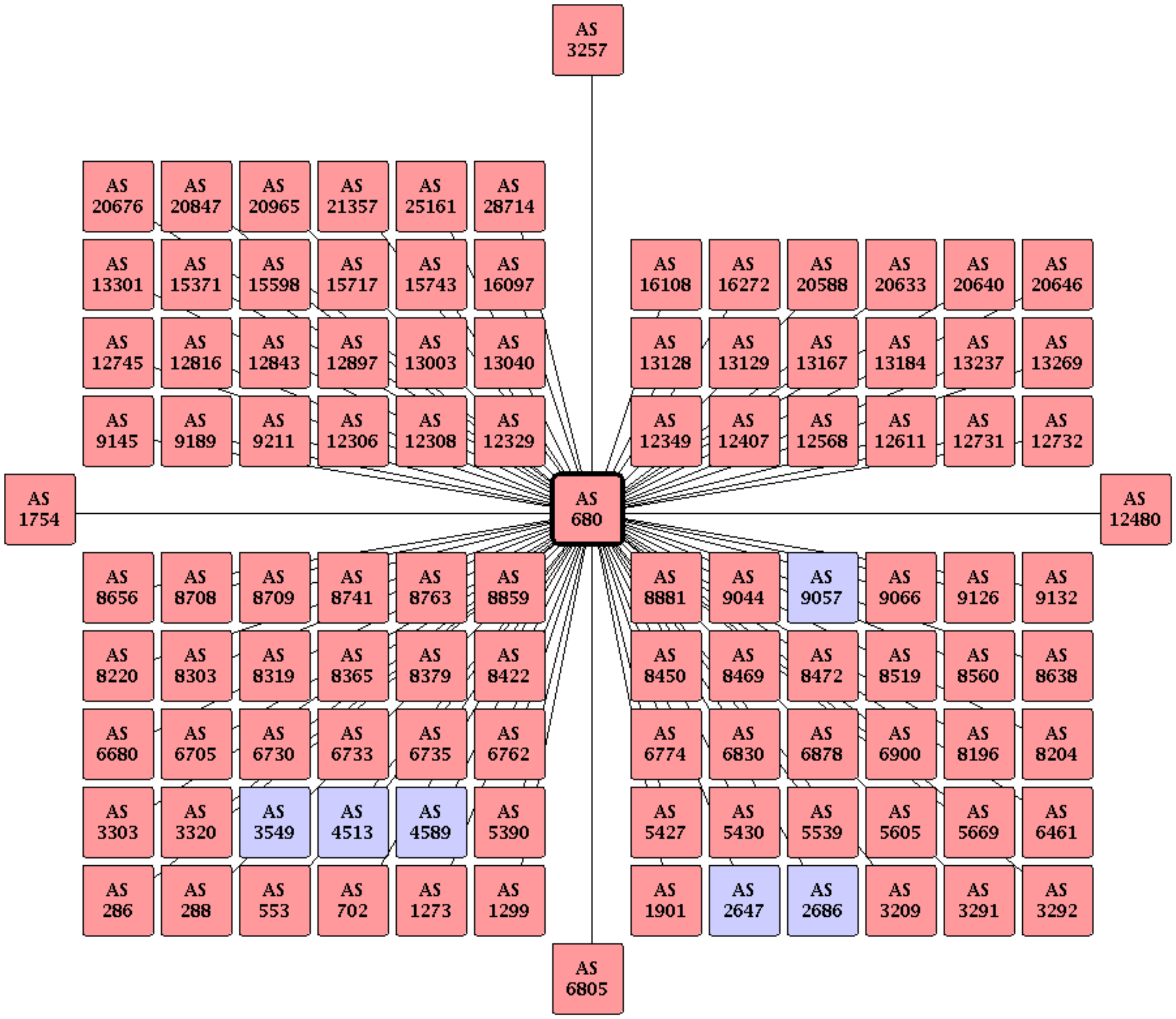
Failure Case

- What to do, if explicit path is broken?
  - Reroute?
  - Notify user?
  - Push packet back, until alternative path is found?



Some Estimations

Neighbors (imports/exports): AS 680, DFN



Policy size:

16KB

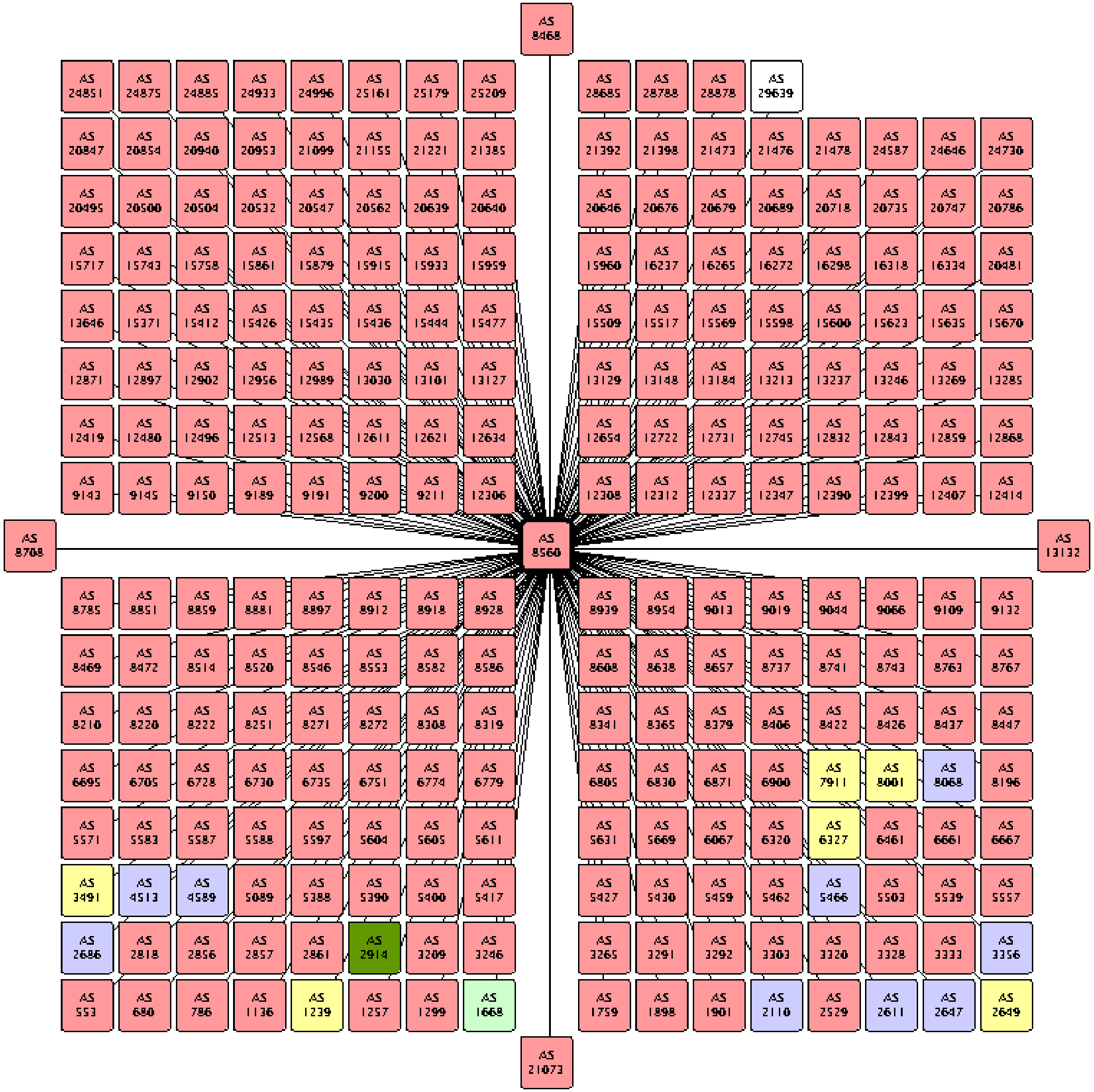
Policy lines:

500

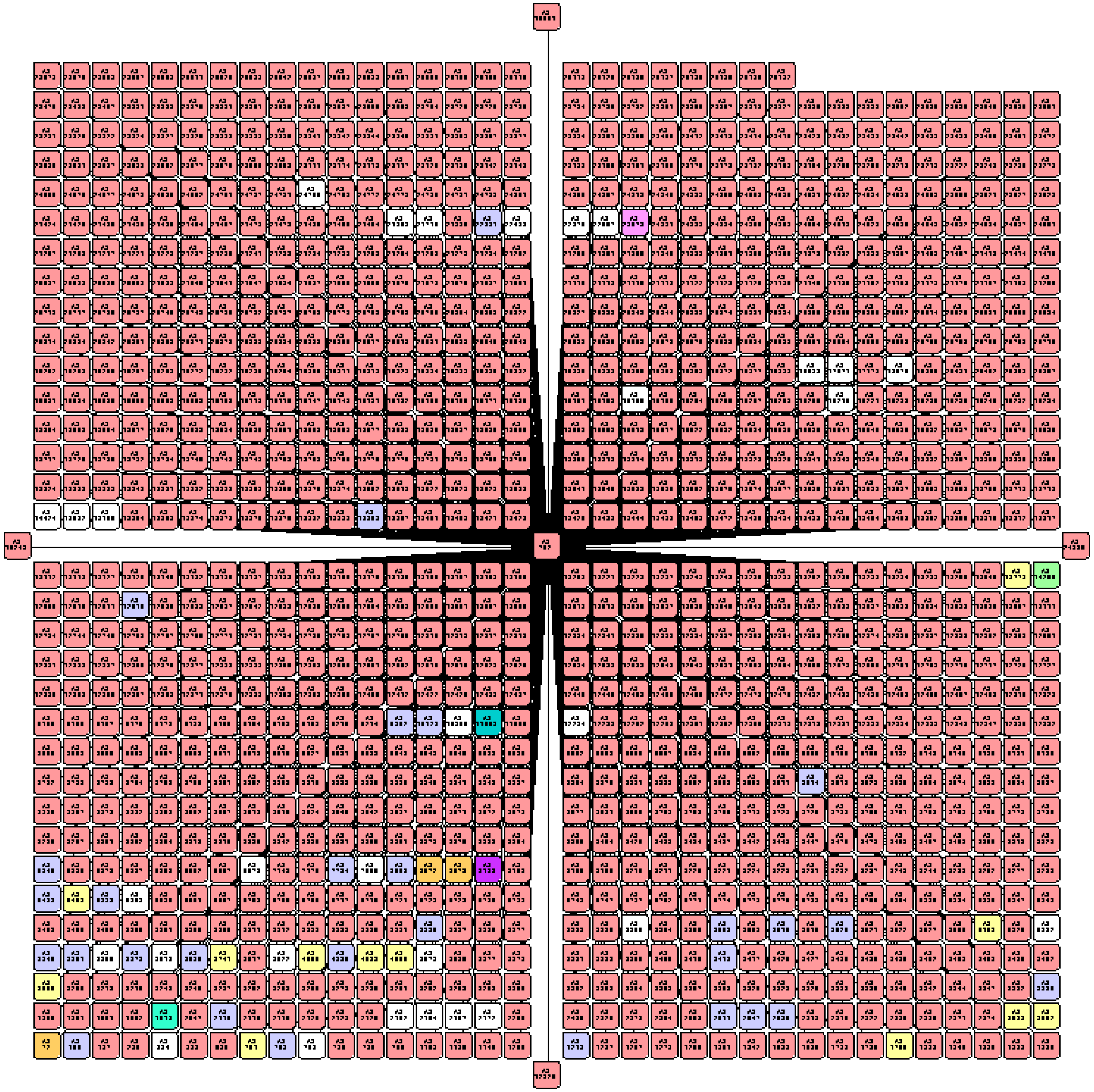
Neighbors (imports/exports):

100

Neighbors (imports/exports): AS 8560, Schlund+Partner AG



Neighbors: AS 702, UUNet Europe



Policy size:

180KB

Policy lines:

2500

Neighbors (imports/exports):

1200

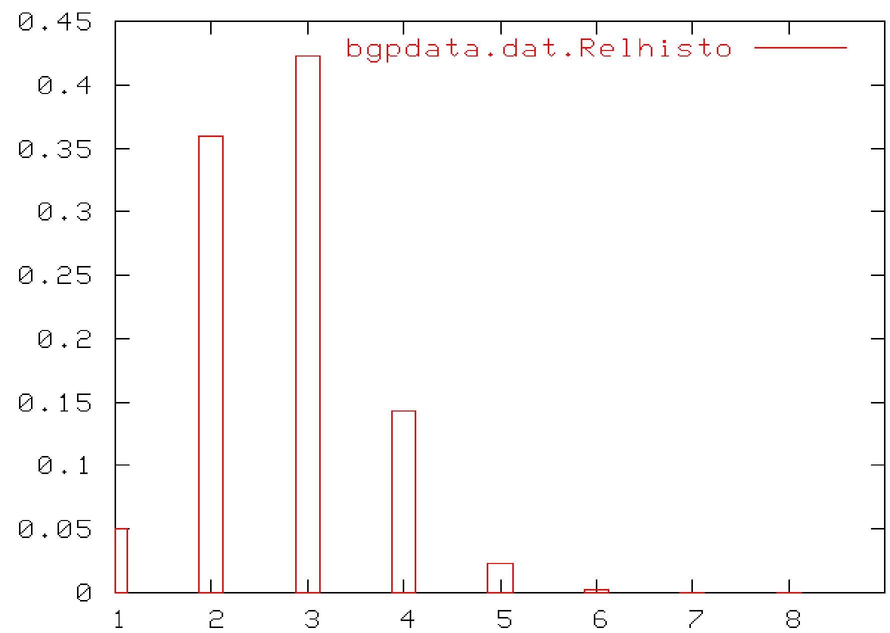
Estimation about Route server table size

15000 ASes, each 250 Neighbors, with 50KB policies:

< 1 GB memory for complete database

3,75 million „links“

Average e2e AS path length histogram:



⇨ The database of the Routing Server should be feasible (and does almost already exist in form of the RIPE and RADB, but heavily out-of-date)

Observed properties of BGP convergence

Olaf Maennel
Technical University Munich

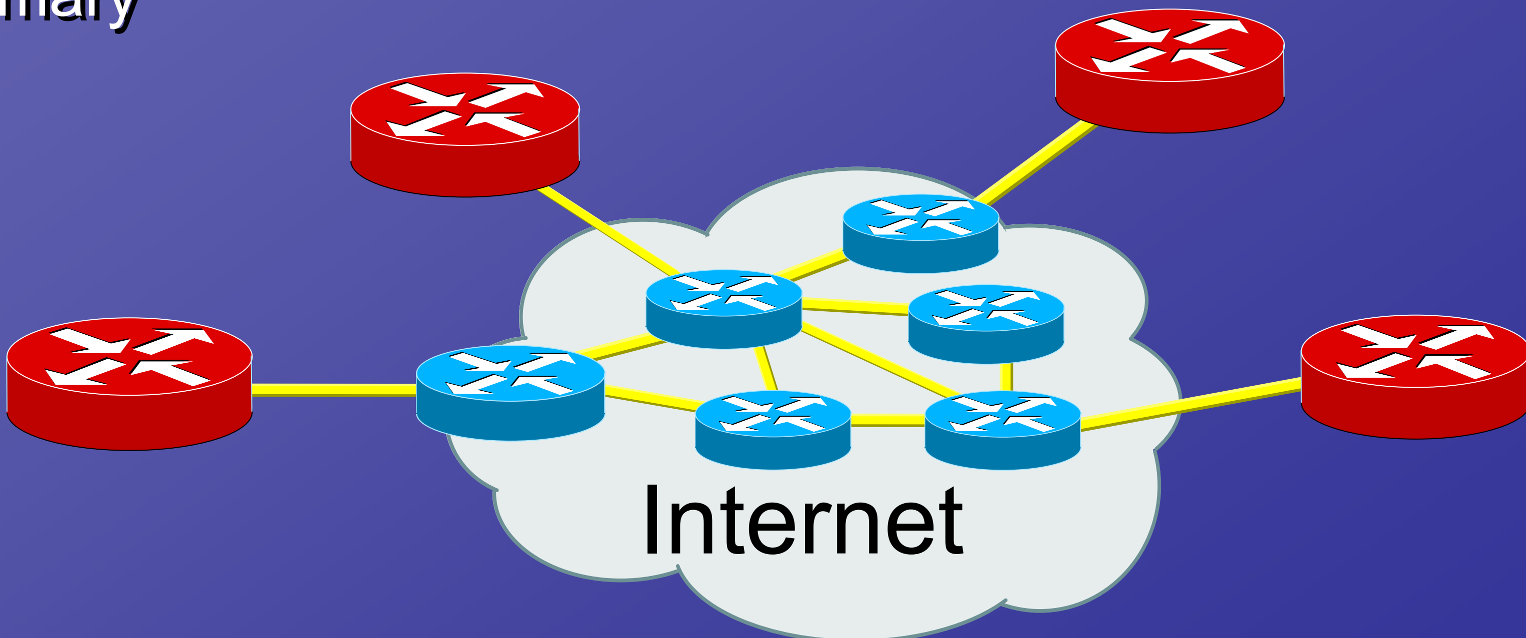
Alexander Tudor
Agilent Laboratories

Sara Bürkle
Saarland University

Anja Feldmann
Technical University Munich

Outline

- One trigger - multiple updates?!!
- Observed BGP convergence properties
  - small timescale behavior
  - larger timescale analysis
  - relationship between multiple viewpoints
- Summary



Data Sets

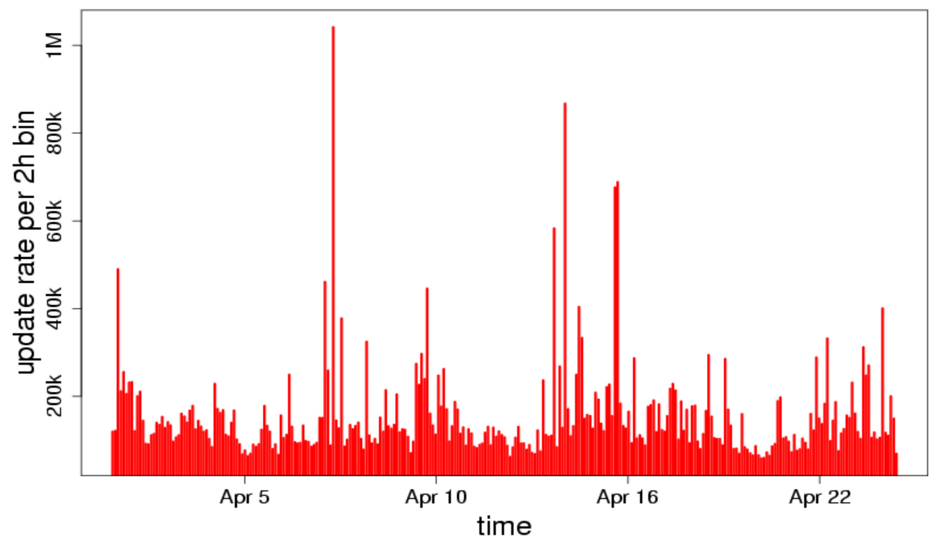
RIPE RIS project : *(http://www.ripe.net/ris/)*

Collector: RRC00

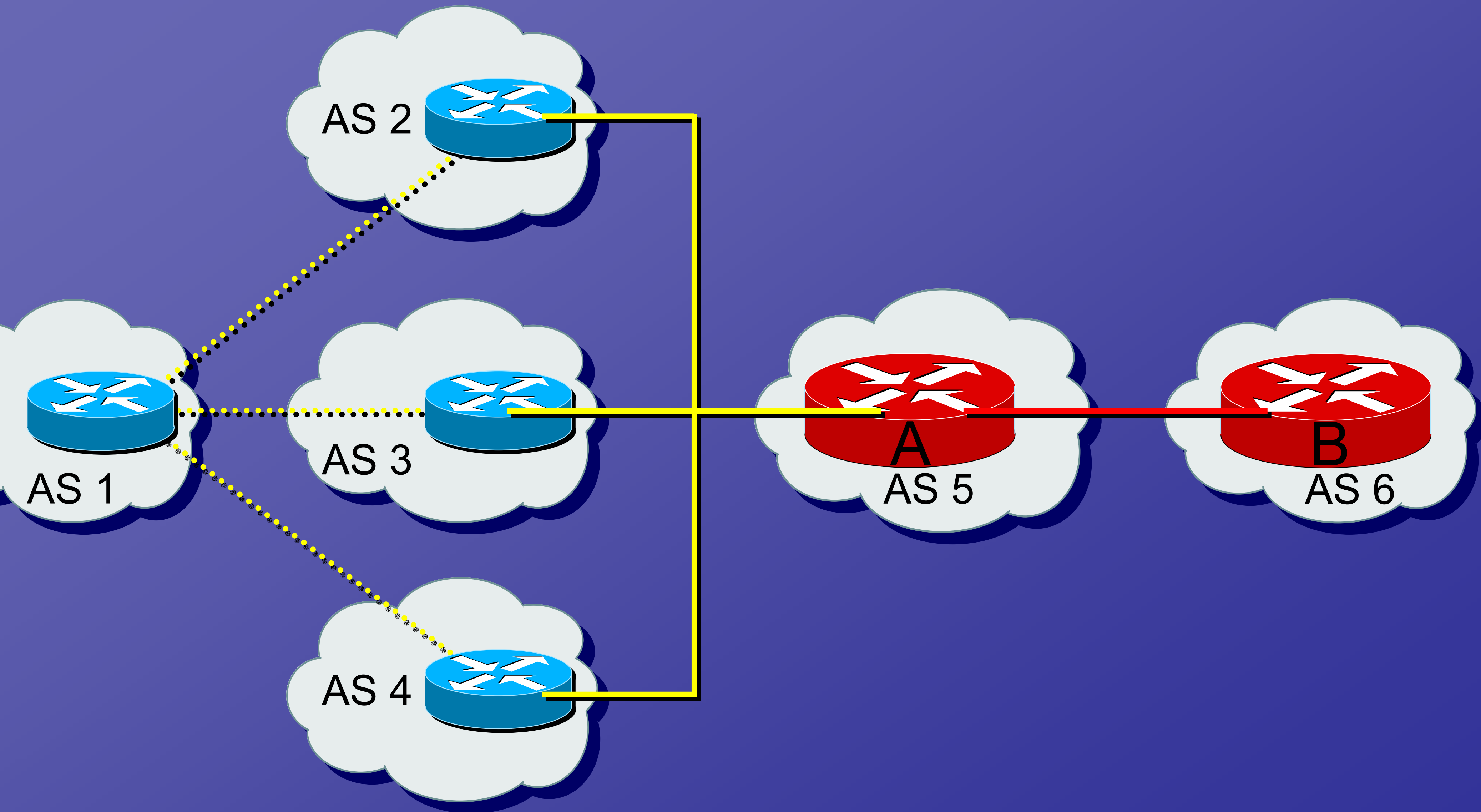
- 100k-123k prefixes
- on 11-13 "default free" peers

- ***January, 2003***
  1/1-1/31 (≈69 Million updates)
- ***February, 2003***
  2/1-2/28 (≈61 Million updates)
- ***March, 2003***
  3/1-3/31 (≈47 Million updates)
  Missing data for 4 hours (3/6 6:08-10:13)
- ***April, 2003***
  4/1-4/24 [17h] (≈44 Million updates)

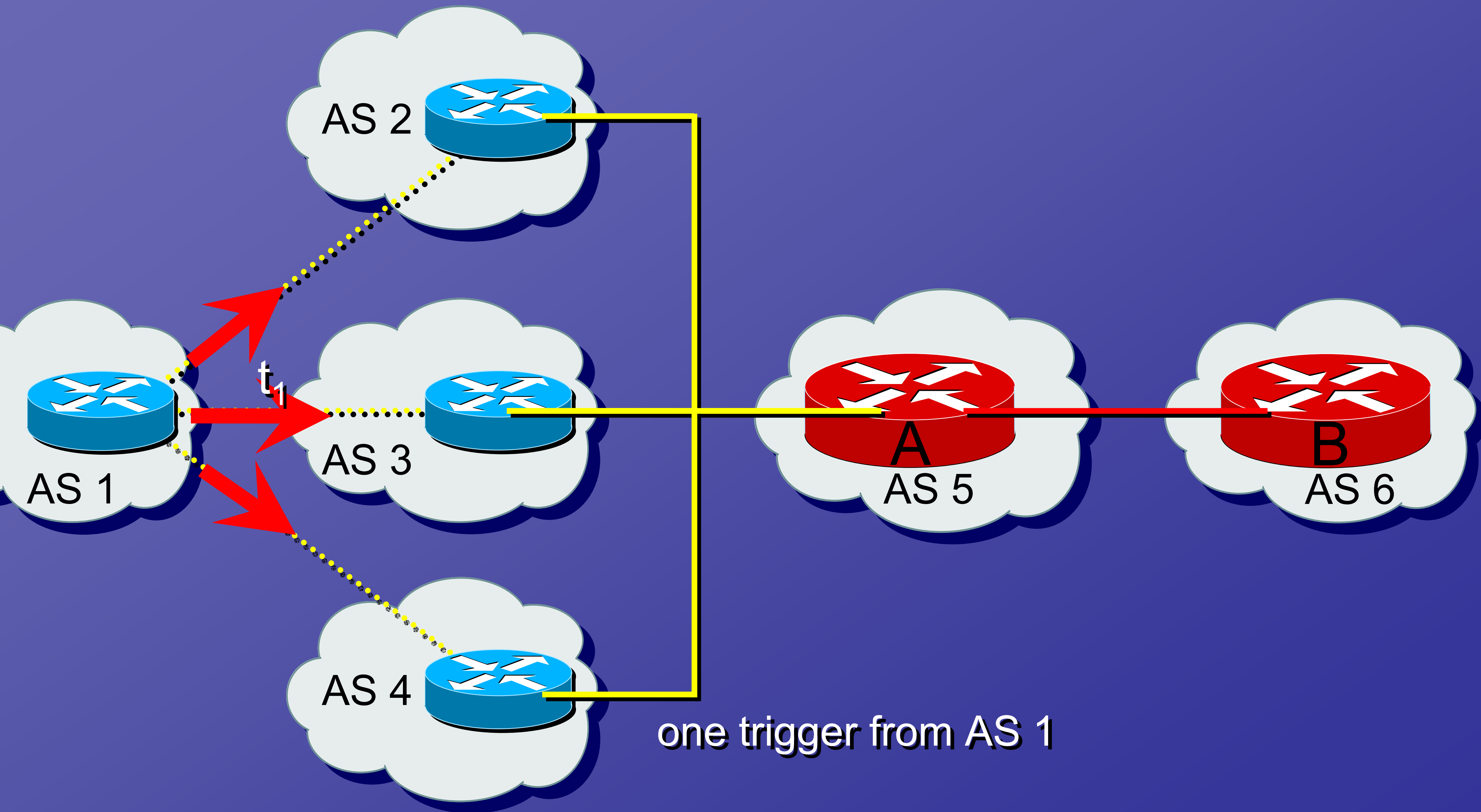
e.g., update rate for April:



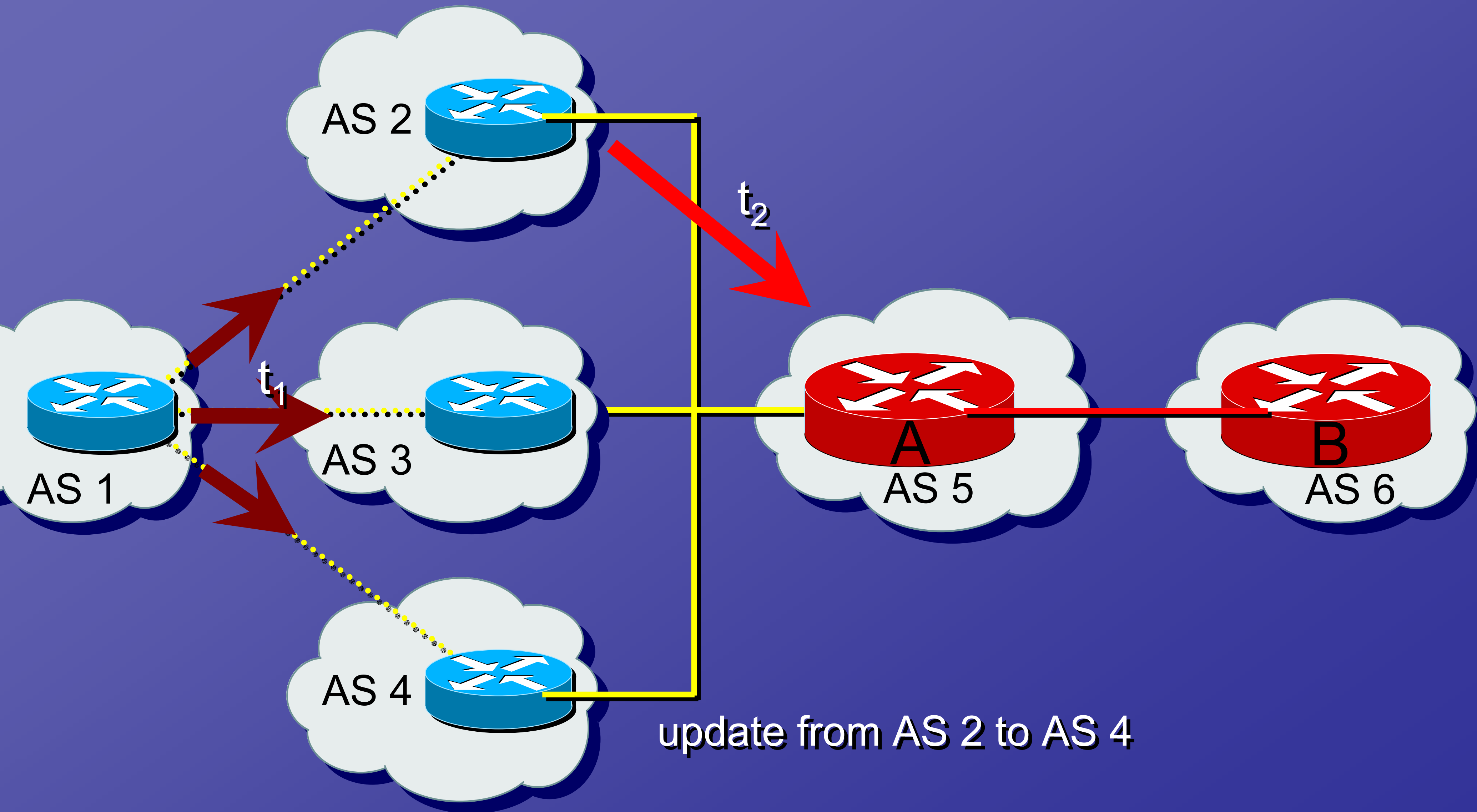
One trigger - multiple updates?!!



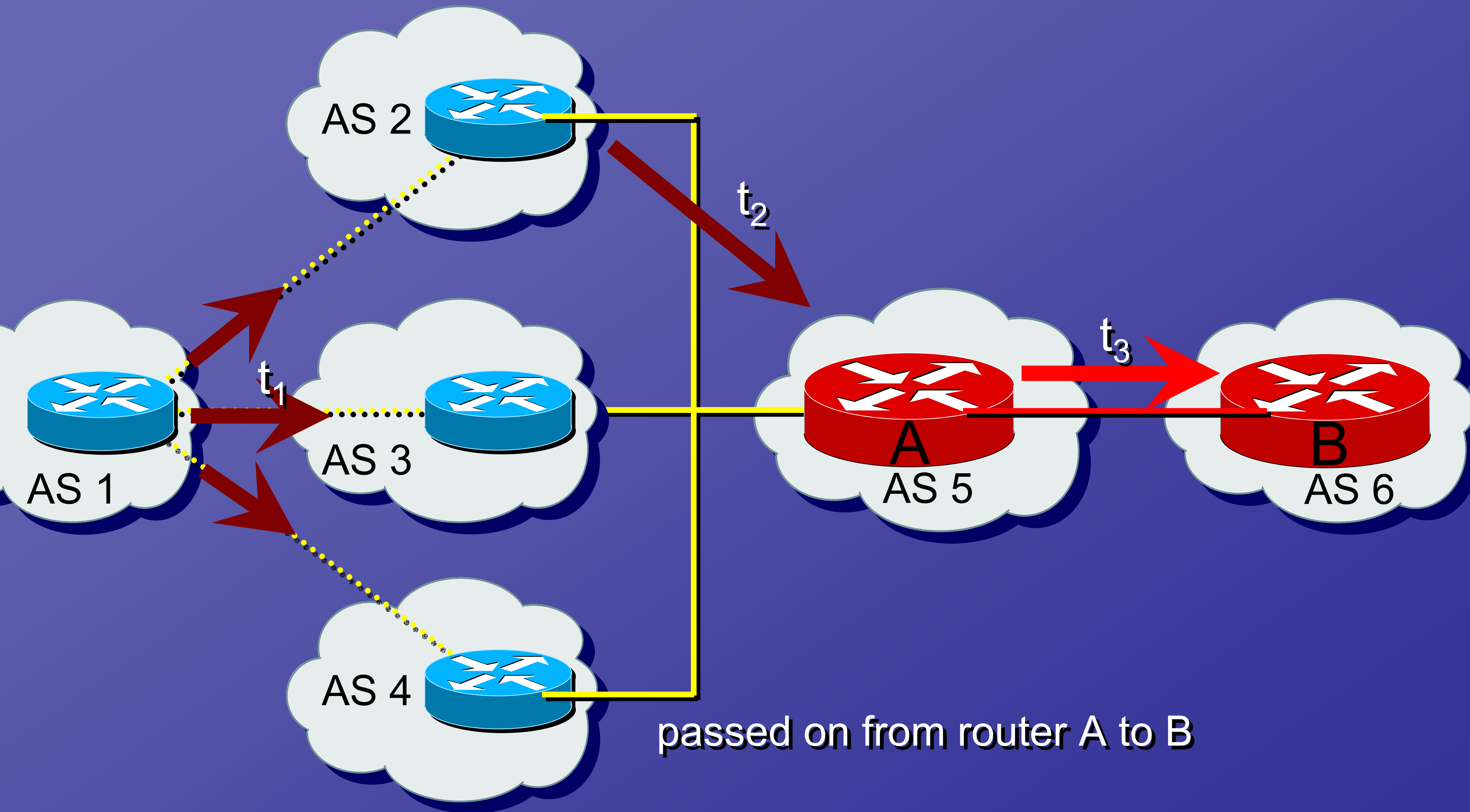
One trigger - multiple updates?!!



One trigger - multiple updates?!!

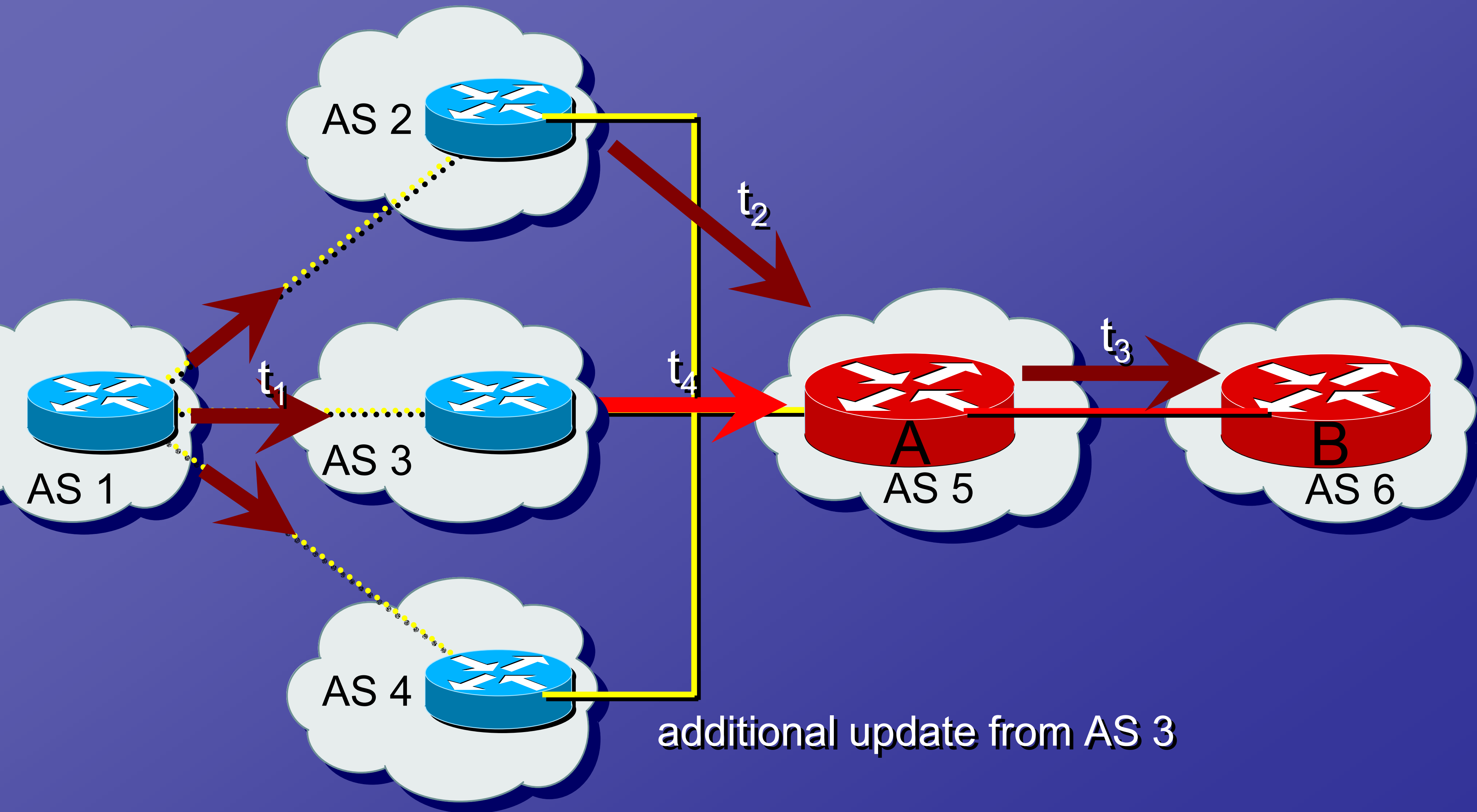


One trigger - multiple updates?!!

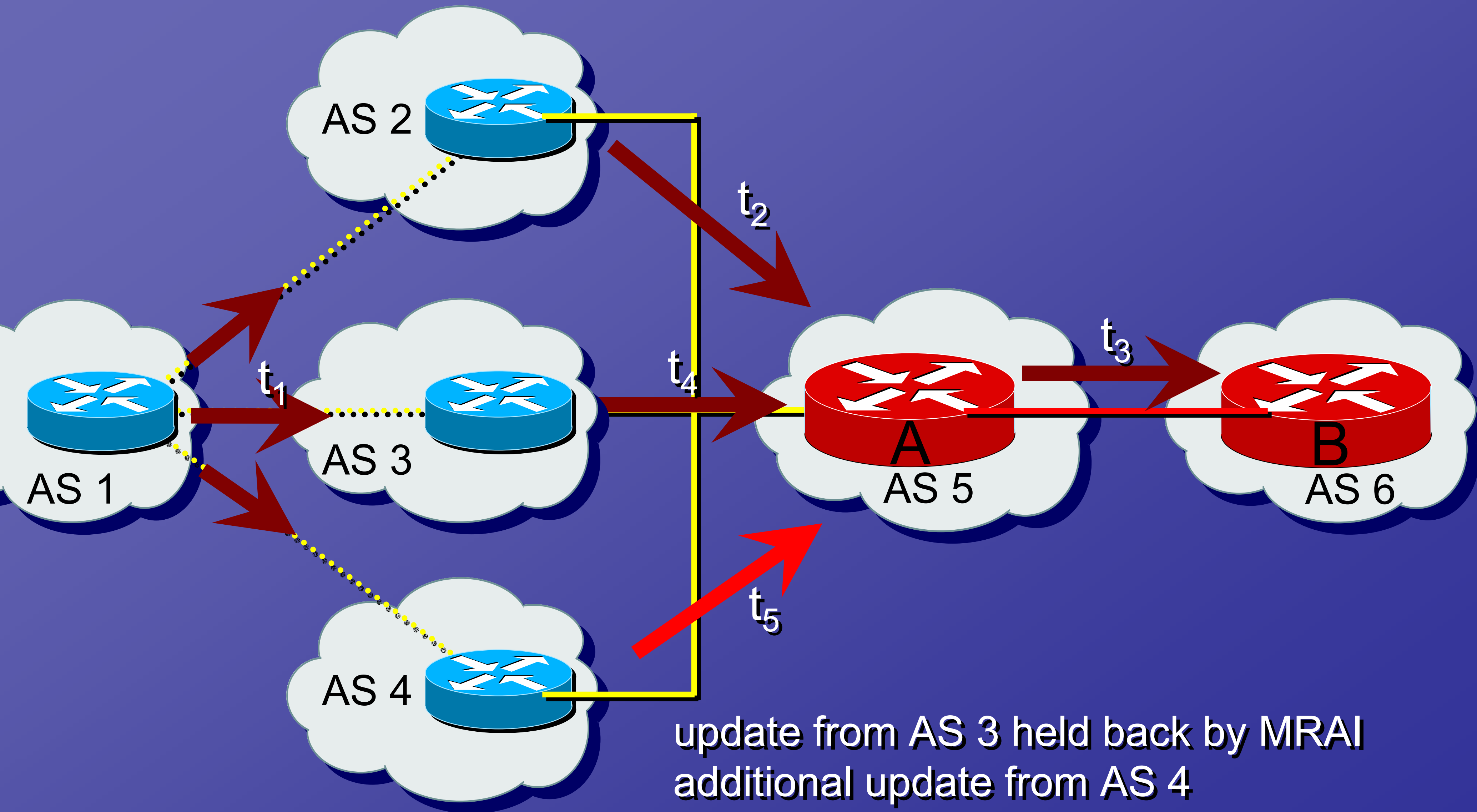


passed on from router A to B

One trigger - multiple updates?!!

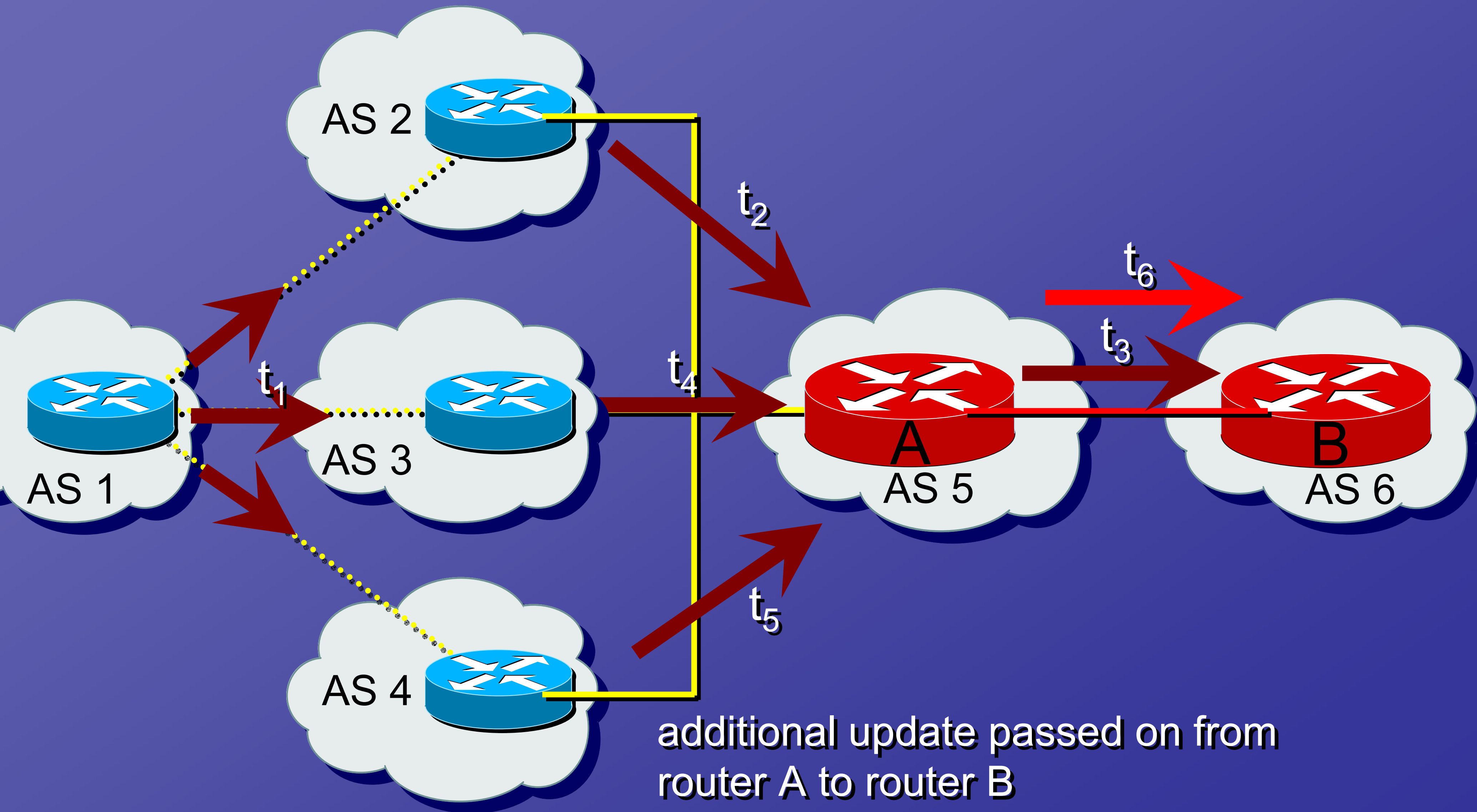


One trigger - multiple updates?!!



update from AS 3 held back by MRAI
additional update from AS 4

One trigger - multiple updates?!!

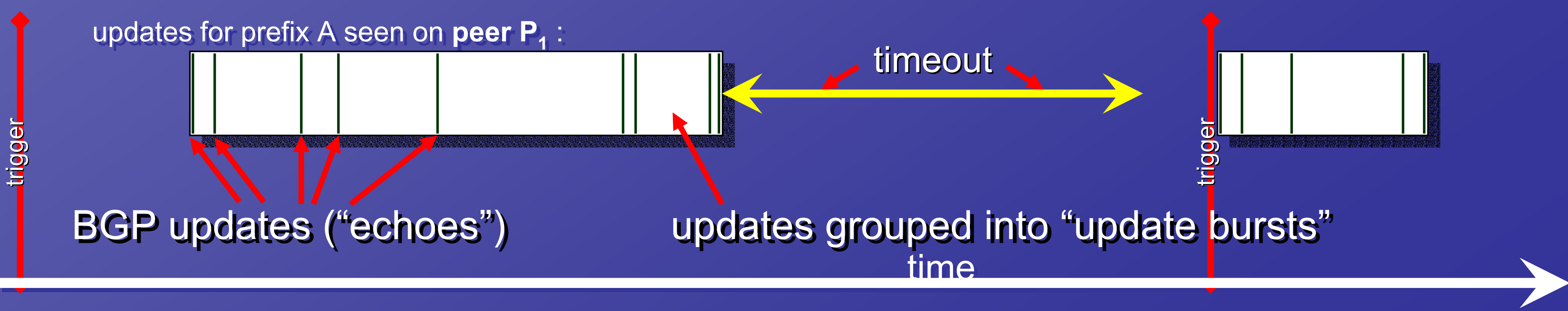


additional update passed on from router A to router B

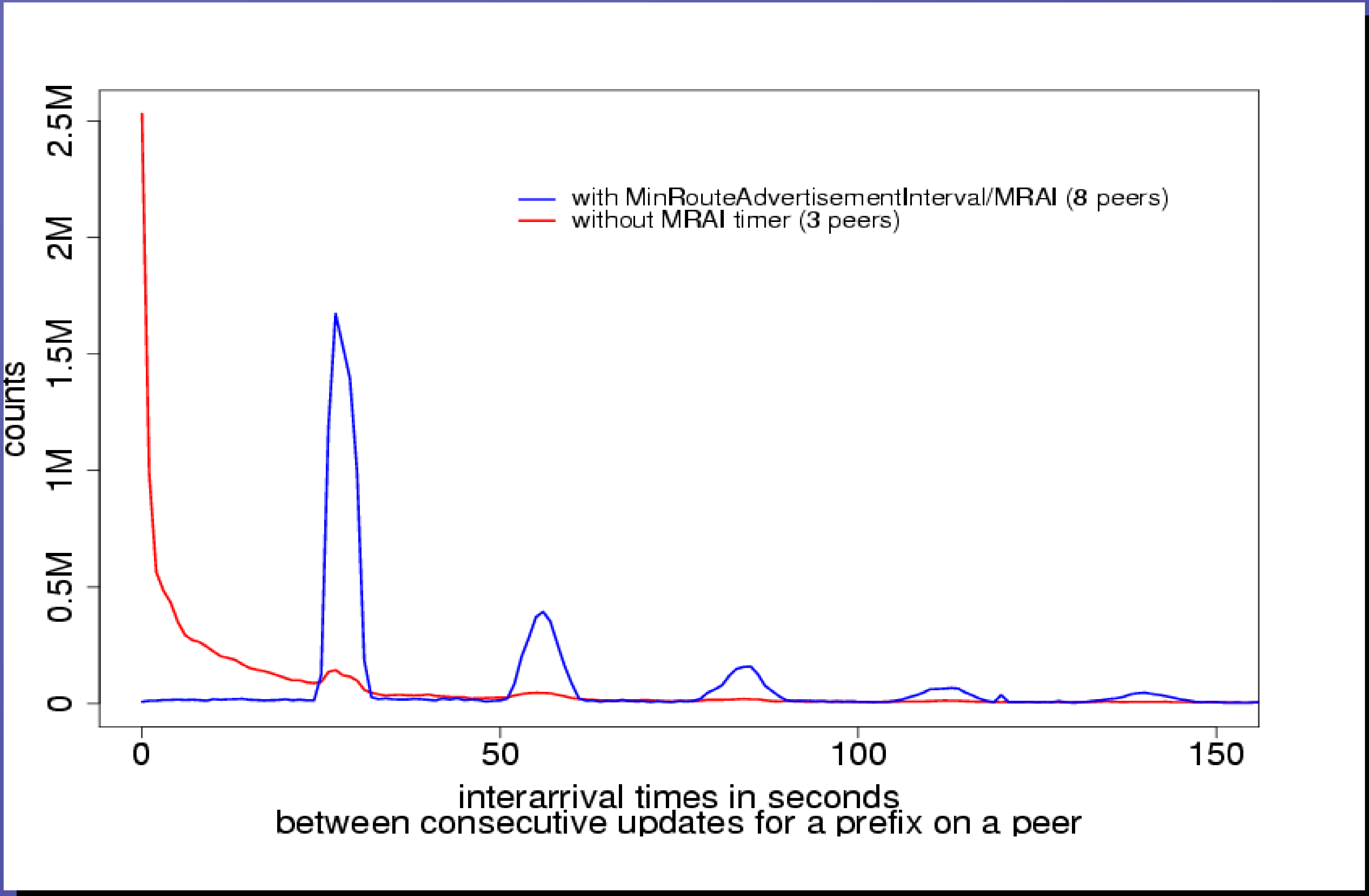
Definition of terms

“echoes” : multiple BGP updates for
- same triggering event
- on one peering session
- for one prefix

Group updates into “update bursts” :
- same prefix / peer
- short time window

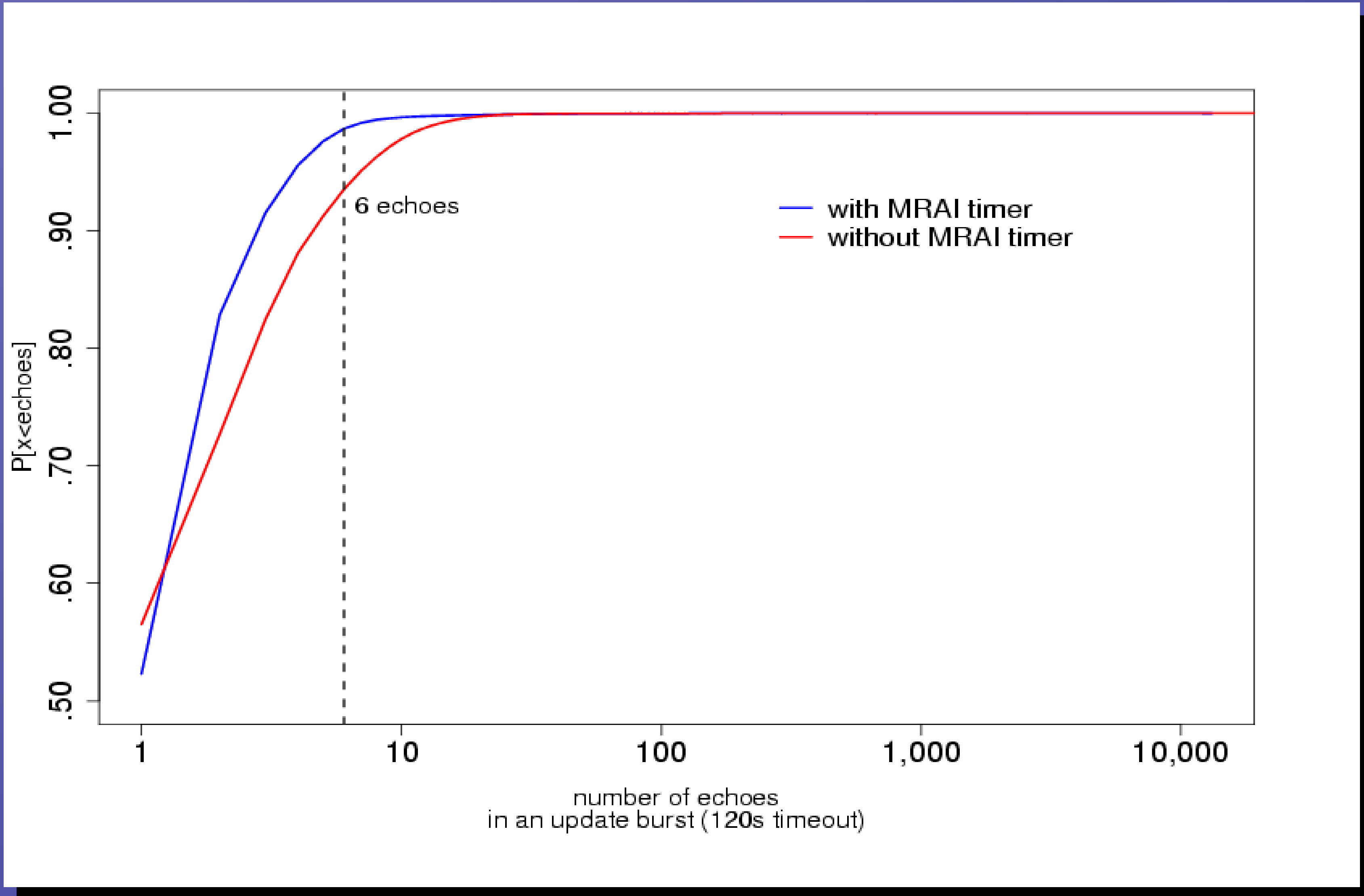


Interarrival time between echoes



peers *without* MRAI: lots of echoes – *with* MRAI: doesn't prevent echoes

Number of echoes in update bursts



damping on peers? *without* MRAI: **8.3%** – *with* MRAI: **2.4%**

Update bursts vs. convergence

Echoes (≥ 6 updates) can cause damping.

e.g.:

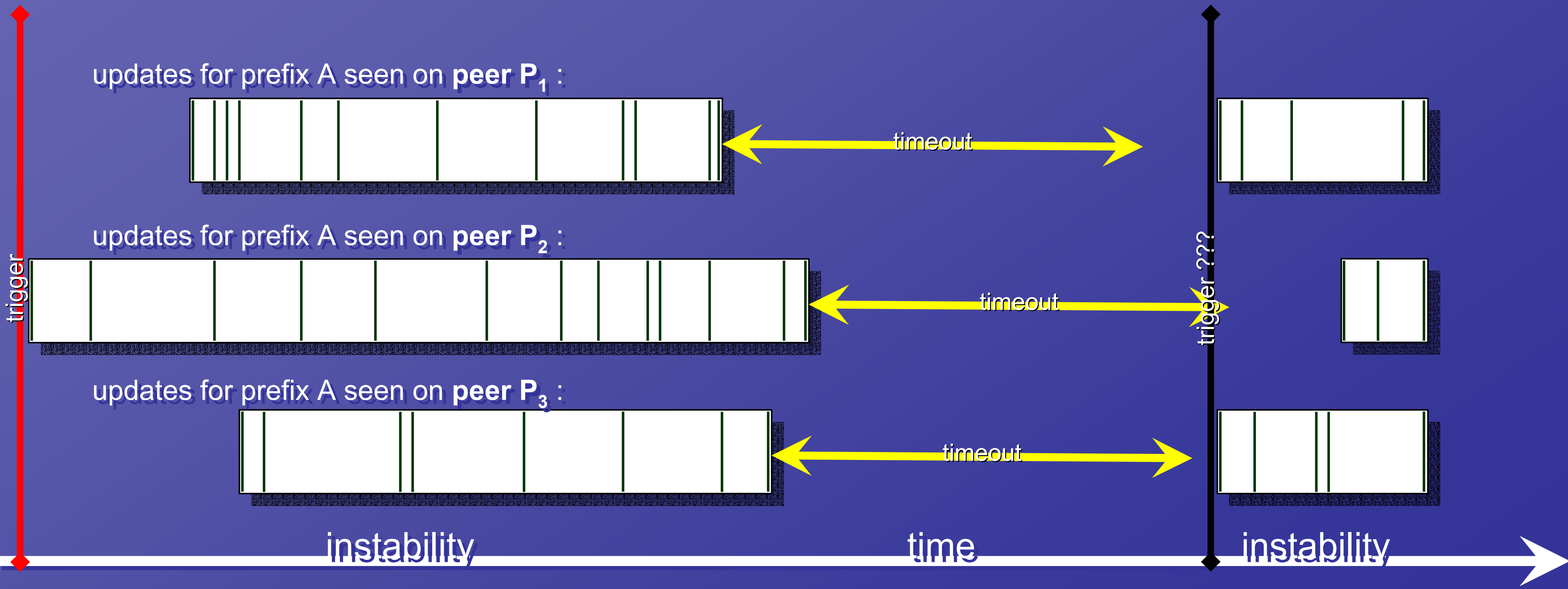
To capture BGP convergence:

- identify "stable" route
- account for damping, overloaded routers, etc.

$\Rightarrow$ timeout > 1 h

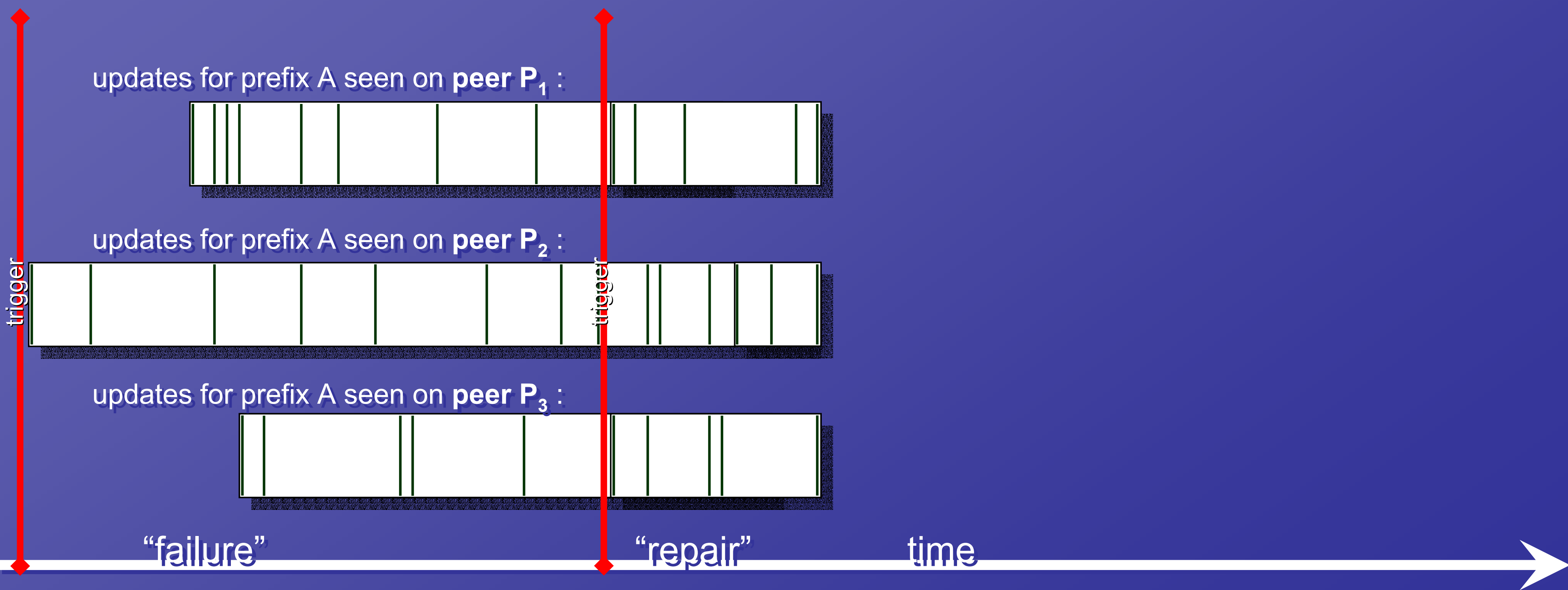
Regarding BGP convergence

- timeout too small: can't capture all effects
- timeout too large: combine several instabilities in one burst

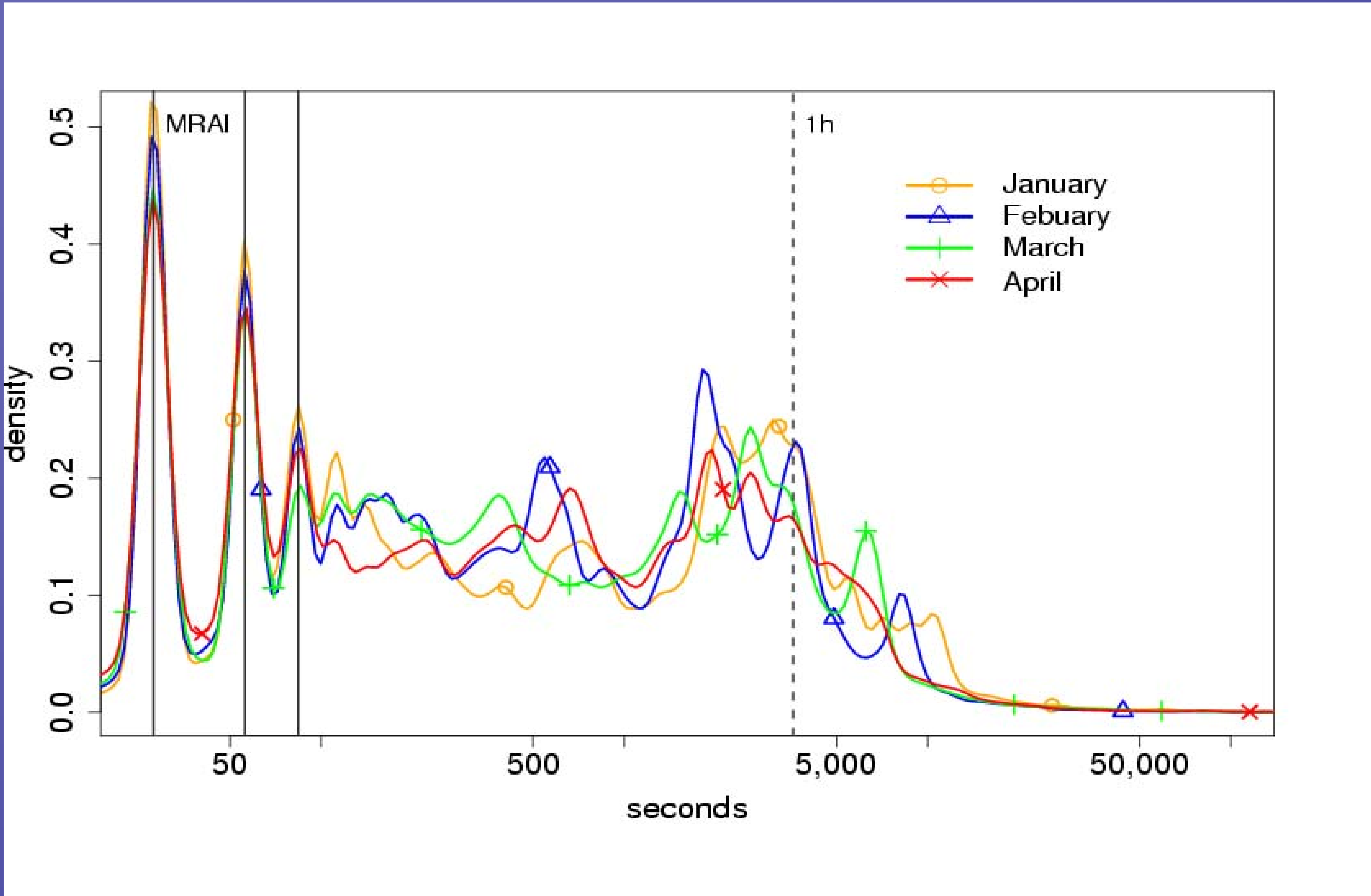


Regarding BGP convergence

- timeout too small: can't capture all effects
- timeout too large: combine several instabilities in one burst



Update burst duration



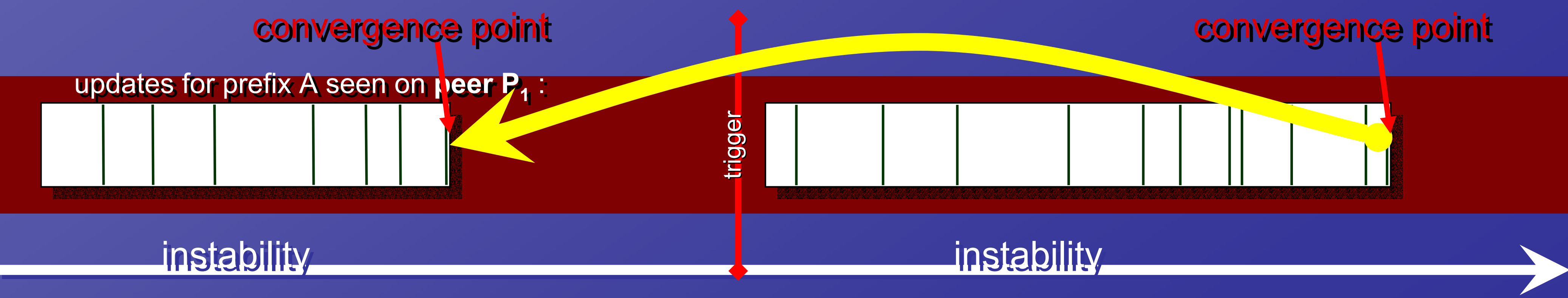
convergence can take rather long…

Important updates in update bursts

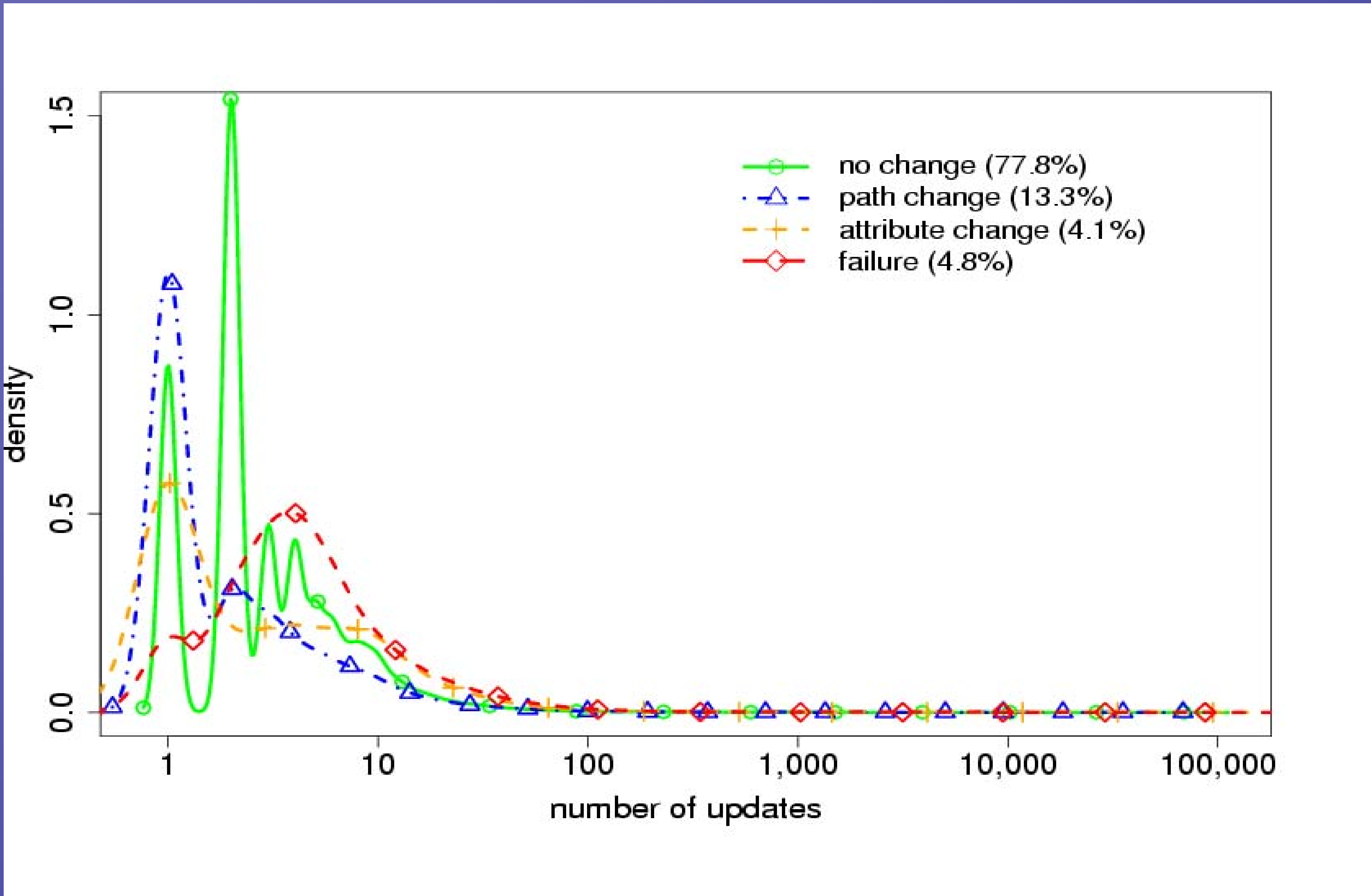
Last update in update burst = "convergence point"
- is result of convergence process
- is a "stable" prefix (at least for some time > timeout).

One prefix – several update bursts:
- how do the convergence points differ?

⇒ compare last updates in update bursts.

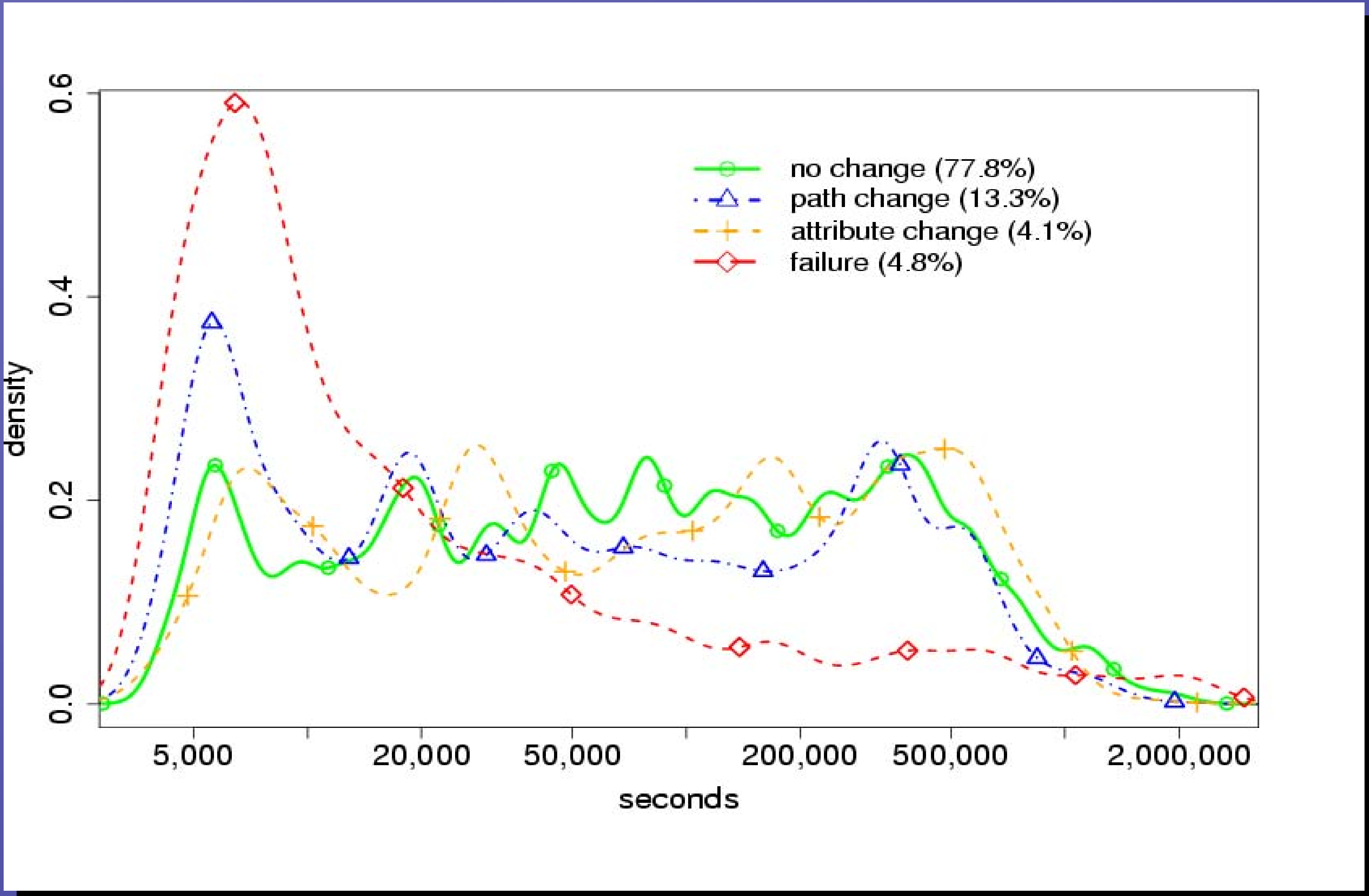


Number of updates in update bursts



most bursts: only a few updates - some bursts: huge # of updates!

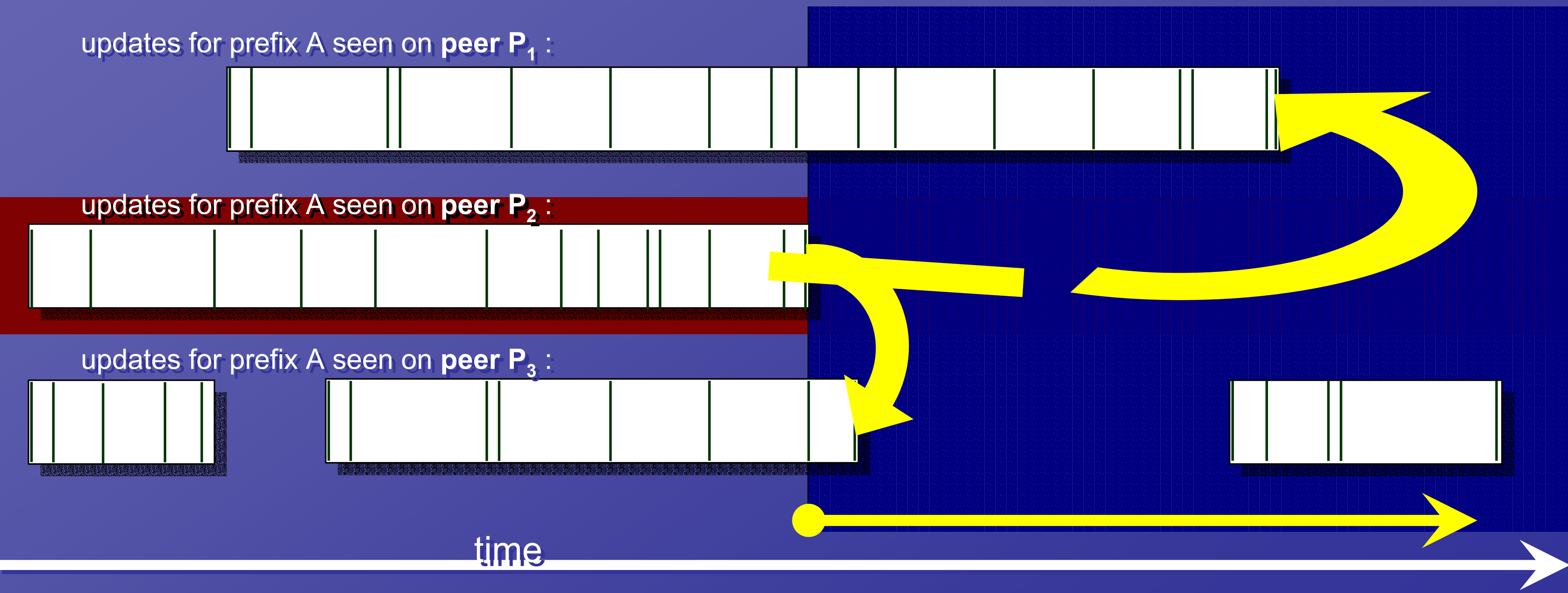
Interarrival time of update bursts



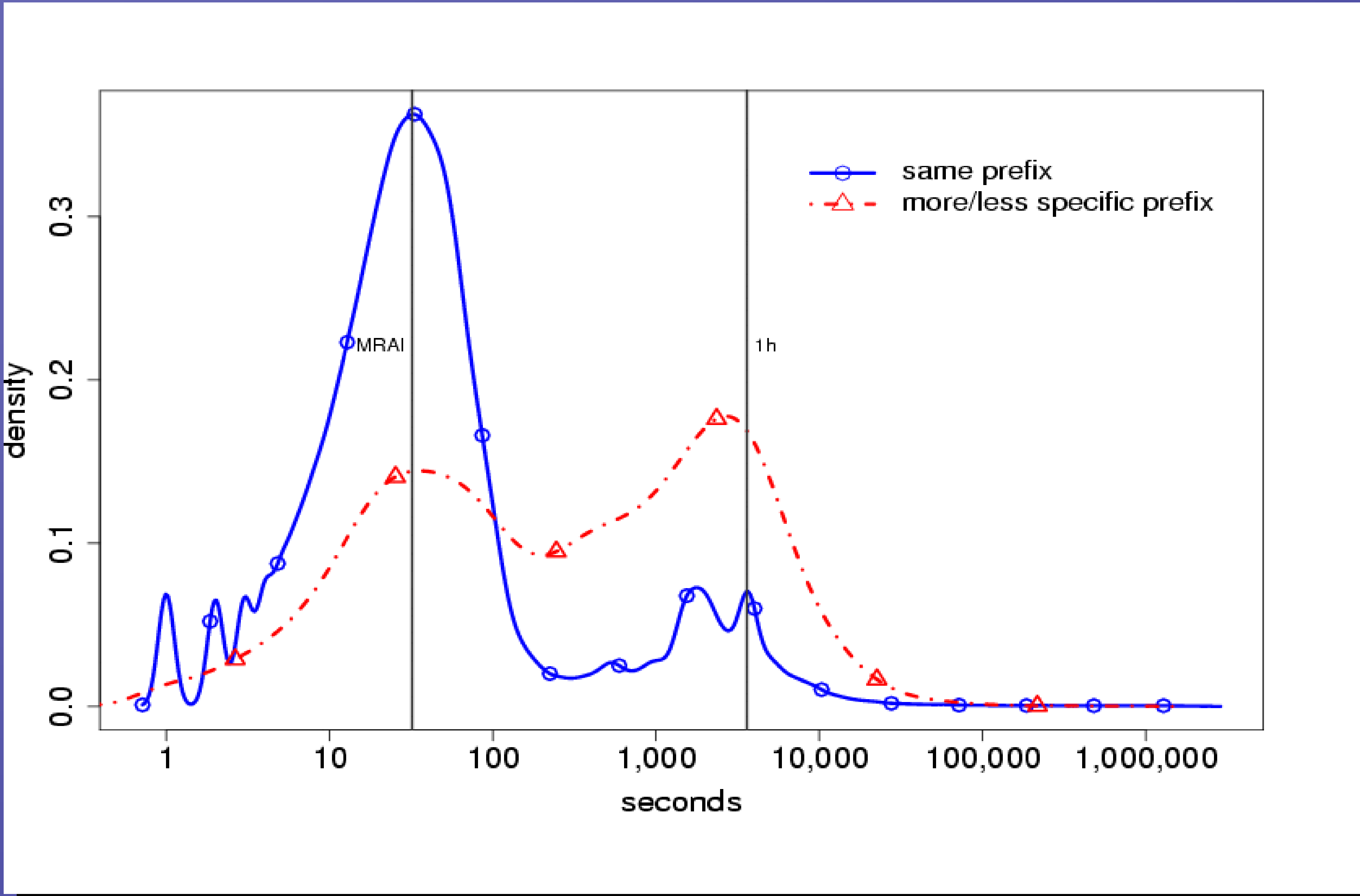
time to next update burst: unpredictable

Convergence points on different peers

Do all peers converge at the same time?
- pick one prefix on one peer
- find other peers with active update bursts
- compute time difference between convergence points



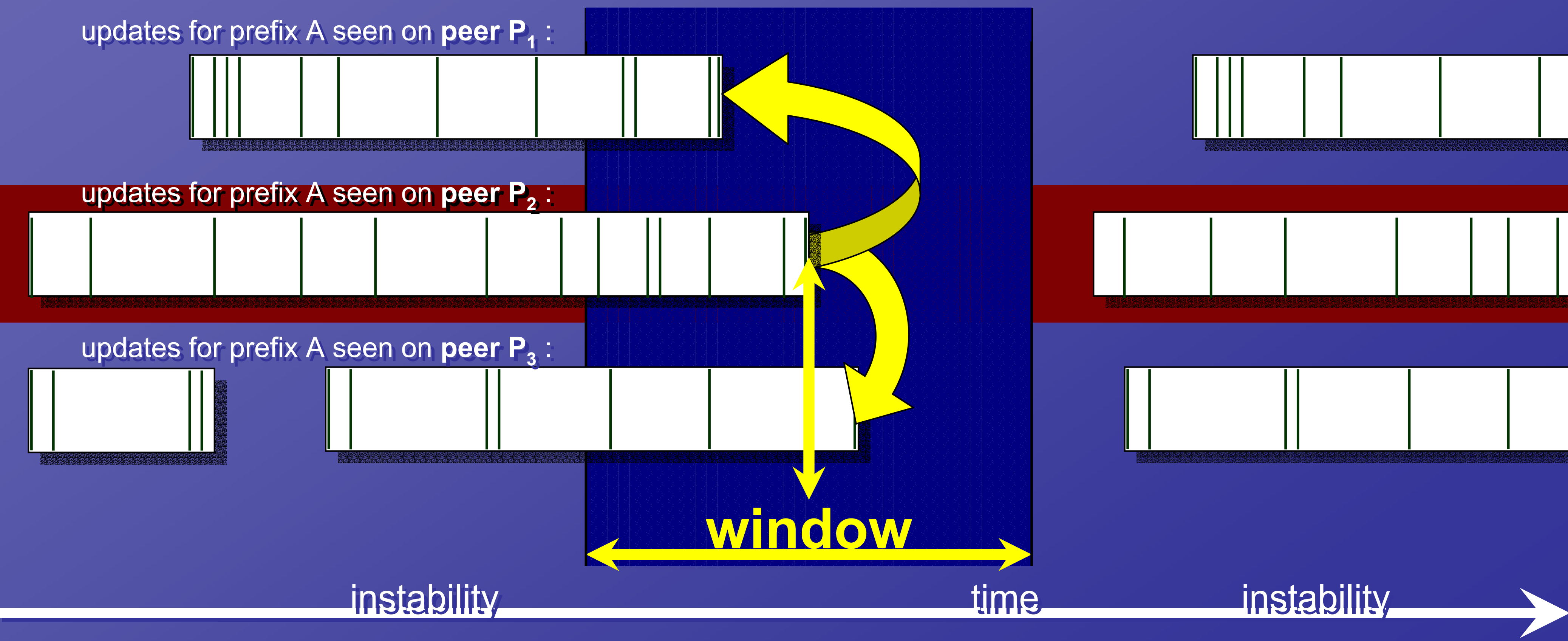
Time difference between convergence points



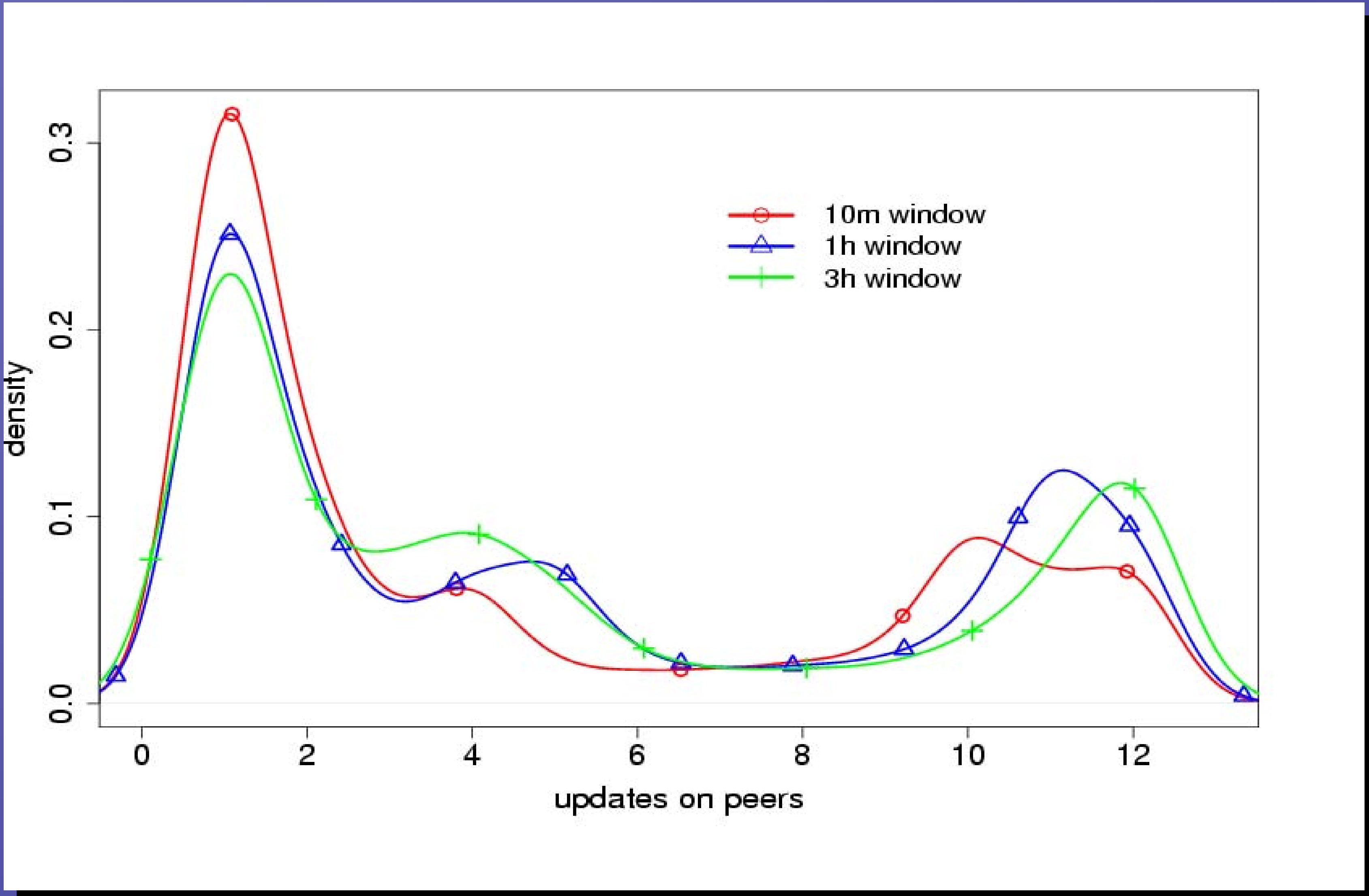
5% of prefixes: with more/less specific update burst

Convergence points on different peers

Problem: depends on which peer is picked
Approach: use sliding window to locate convergence points



Bursts observed on different peers



update distribution: locally or globally visible

Summary

Today's BGP convergence depends on

- MRAI

  shorter MRAI leads to :
  – more echoes and to more damping *and*
  – to faster convergence if damping is not aggressive…

- Damping settings

  – damping occurs for normal prefixes!
  (BGP path exploration may need ≥ 6 echoes,
  and depends on interconnectivity)
  – damping helps for unstable prefixes

Further information

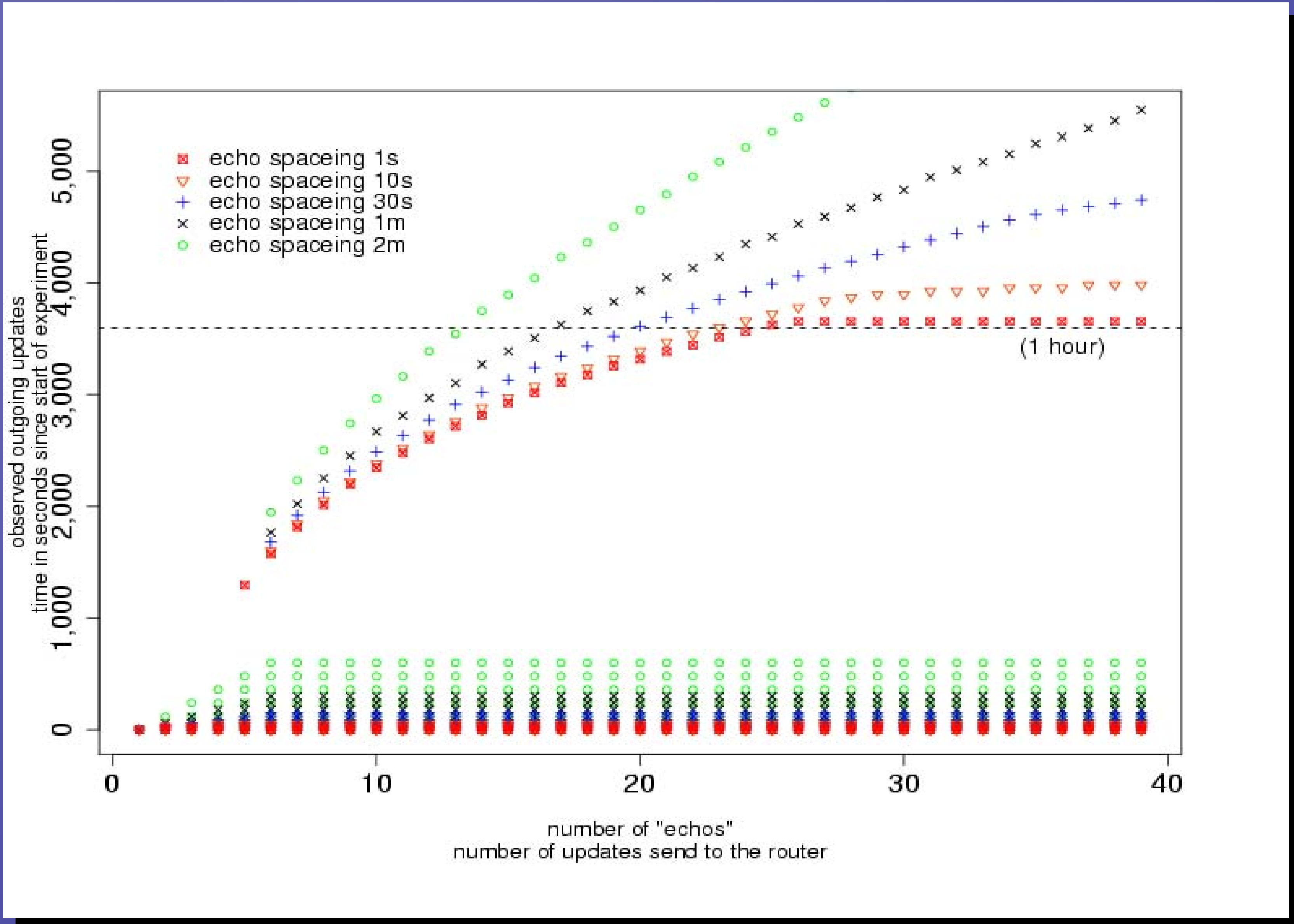
If you are interested, please
visit our website:

http://www.olafm.de/

Questions? Comments?!

Thanks !

Additional slide



Routing-Convergence of RFC2547bis-VPNs

Inter-Domain Routing Workshop 2003

Lx Manhenke, Consulting Engineer, lx@cisco.com

Agenda

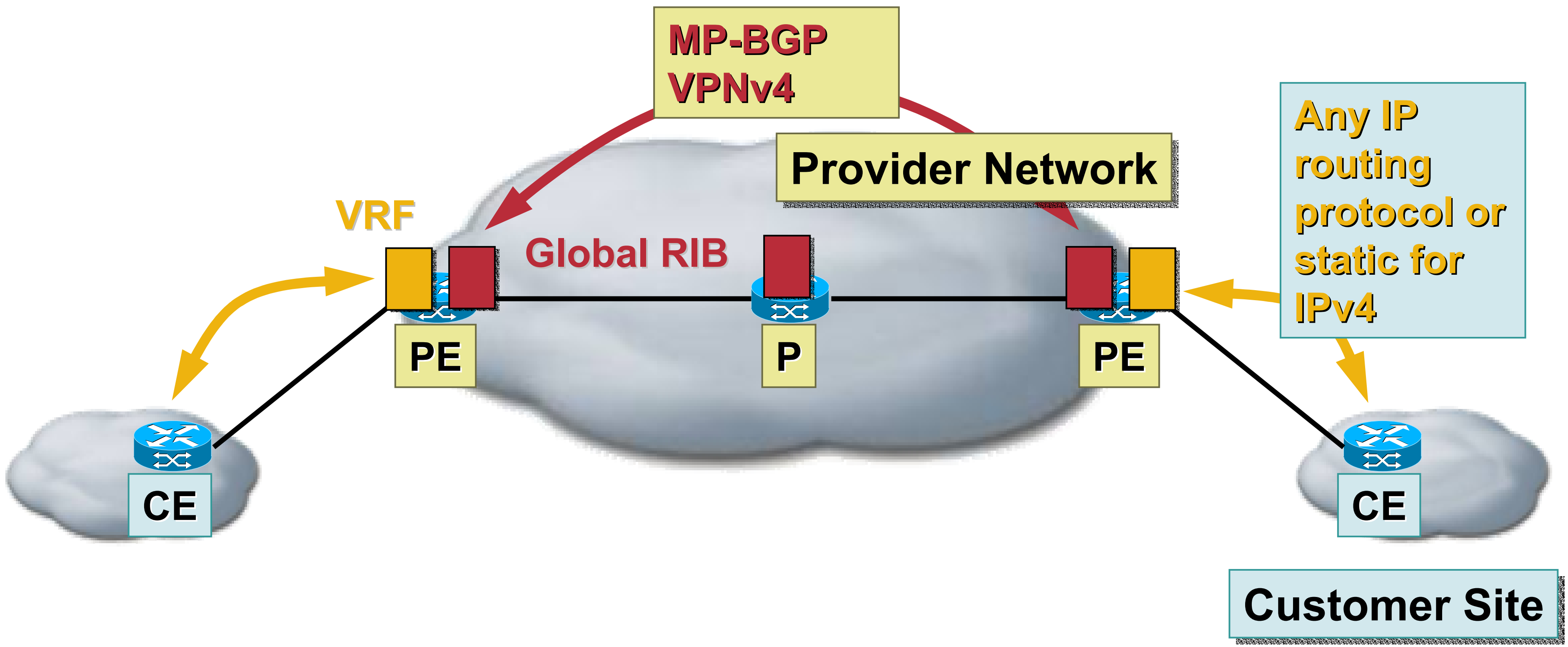
- RFC2547bis Architecture
- Analysis of Convergence Components
- Options for Improvement

RFC2547bis Architecture

Draft RFC2547bis

- Draft based on RFC2547
- Document of the L3VPN working Group of the IETF
- Reference: http://www.ietf.org/internet-drafts/draft-ietf-l3vpn-rfc2547bis-01.txt
- Based on Peer Model – VPN-Sites connect to a cloud (Provider Network) rather than being directly connected
- MP-BGP used as Control Plane Protocol to distribute customer L3 information in form of VPNv4 routes
- VRFs on Edge Routers to separate VPN routing tables
- MPLS or any other tunneling encapsulation to tunnel the VPN data packets through the Provider Network

RFC2547bis Architecture



P – Provider Router
PE – Provider Edge Router
CE – Customer Edge Router
VRF – VPN Routing and Forwarding Instance
RIB – Routing Information Base

Analysis of Convergence Components

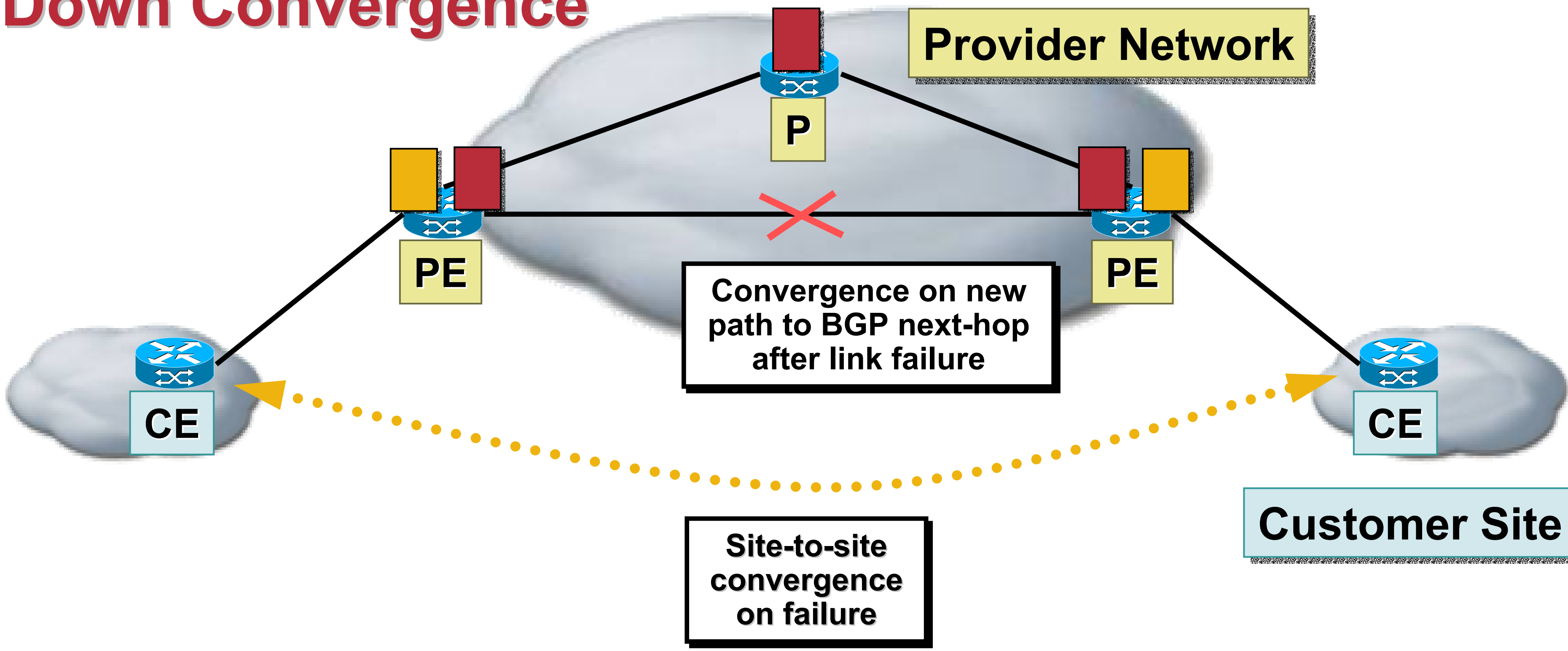
Routing Convergence

Cisco.com

- **Convergence needs to be assessed in two main areas**
  - Convergence within the MPLS/VPN backbone
  - Convergence between VPN client sites
- **And with two separate scenarios…**
  - Convergence on platform/network failure
    - ➡ **DOWN CONVERGENCE**
  - Convergence on new routing information
    - ➡ **UP CONVERGENCE**

DOWN Convergence Analysis

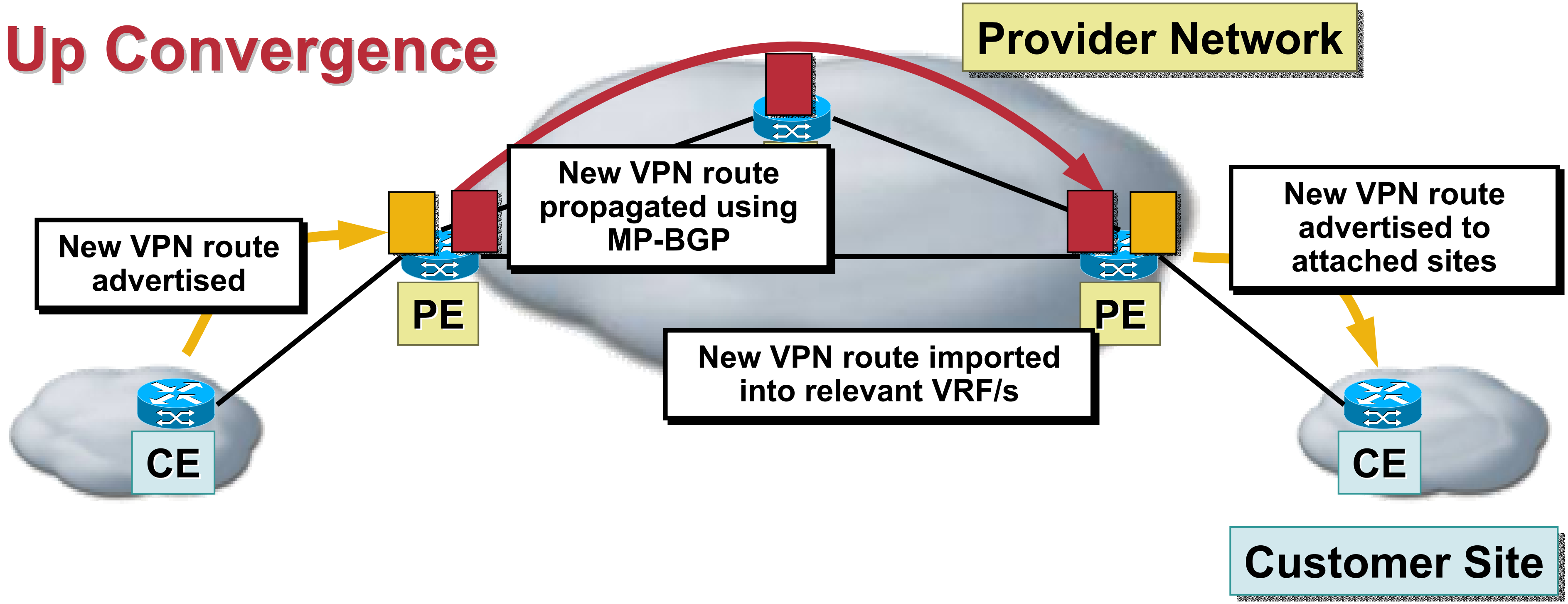
Down Convergence



<u>Site-to-site</u> and <u>MPLS VPN Backbone IGP convergence</u> are independent

UP Convergence Analysis

Cisco.com



<u>Site-to-site convergence</u> on new VPN client routing information

Convergence across Backbone

- **Convergence of MPLS/VPN backbone IGP will not affect client-to-client route convergence**
    - Unless BGP next-hop becomes unavailable;
    - but will affect client-to-client traffic while backbone converges
- **Focus on Site-to-Site Up/Down Convergence**

Site-to-site Convergence Points

- Several main convergence areas

  T1 – Advertisement of routes from CE to PE

  T2 – Placement into VRF

  T3 – Propagation of VPNv4 routes to BGP neighbors

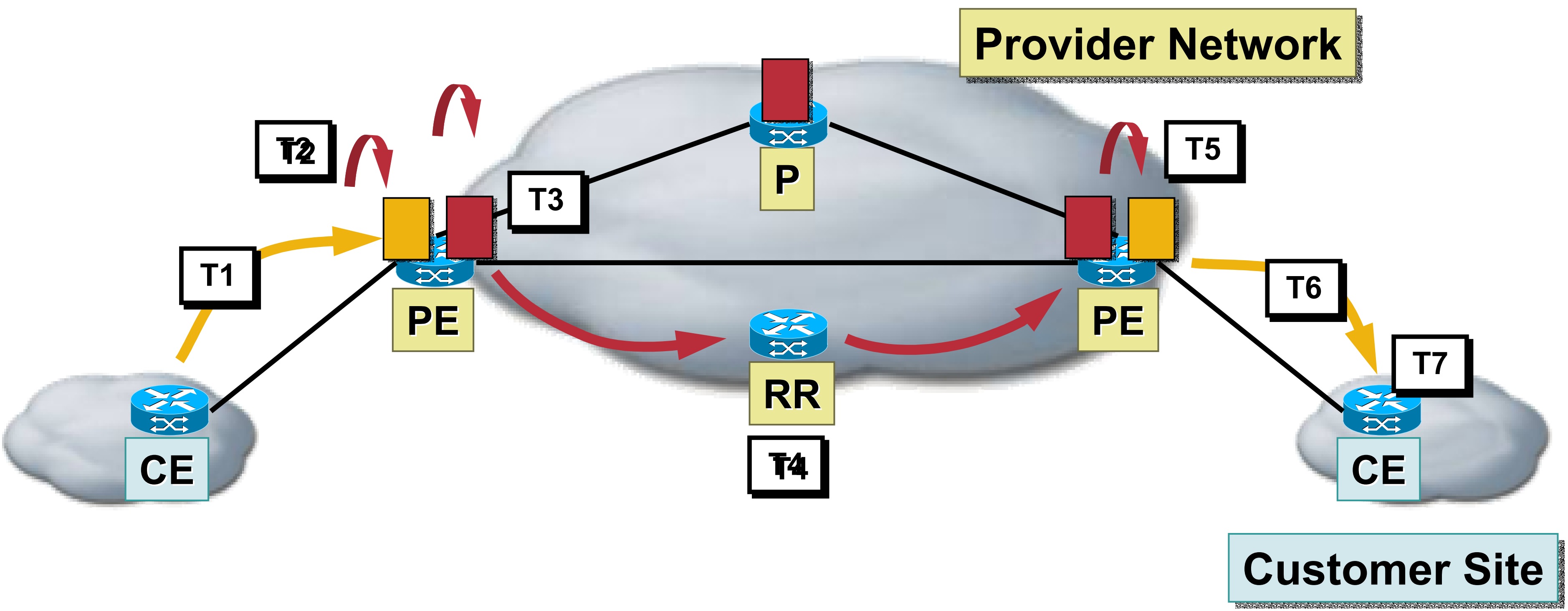
  T4 – Reflection of VPNv4 routes to PEs

  T5 – Import process of these routes into relevant VRFs

  T6 – Advertisement of VRF routes to attached VPN sites

  T7 – Processing of VPN routes on the CE routers

Site-to-site Convergence Points



Site-to-Site Convergence Points

Backbone Route Propagation

- **Change are not propagated to other BGP speakers immediately**
    - Batched together and sent at advertisement-interval
        - Default is every 5 seconds for iBGP, 30 for eBGP
- **Can be tweaked using the neighbor advertisement-interval command**
    - Needs to be changed for both backbone and CE routers if BGP between PE and CE

Scanner Process

- **Scanner process will also have an affect on convergence**
  - Used to check next-hop reachability and to process any network commands within the BGP process
  - Invoked every 60 seconds by default for IPv4 and VPNv4 prefixes
  - Can be tuned with BGP scan-time command
  - Large BGP table and small scan-time can be very CPU intensive—beware!

Import Process

- **Import process uses a separate invocation of the scanner process**

  Every 15 seconds by default

  Can be tuned using the BGP scan-time import command

- **Could take up to 15 seconds for a route to be placed into a receiving VRF**

  And then potentially a further 30 seconds to be advertised to CE if eBGP is in operation!

Maximum UP Convergence Summary

	BGP-4	RIP V2	OSPF	Static
T1_max	30 (Default Adv Int)	30 (0 with Triggered)	Variable (Time for LSA Propagation)	0
T2_max	0	0	0	0
T3_max	0	0	5	0
T4_max	10 ((n+1) * Adv Int)	10 ((n+1) * Adv Int)	10 ((n+1) * Adv Int)	10 ((n+1) * Adv Int)
T5_max	15	15	15	15
T6_max	30 (Default Adv Int)	30 (0 with Triggered)	Variable (Time for LSA Propagation)	0
T7_max	0	0	5	0
Total	~85 Seconds	~85 Seconds or ~52 Seconds (with Triggered)	~35 Seconds	~25 Seconds

Options for Improvement

New Requirements for BGP

- Focus for BGP Implementation in the past was scalability and most important stability for Internet Routing
- Use of scan timers in the implementation of BGP suits very well for that purpose
- Today MP-BGP is used for more than just Internet Routing Control Plane
- Applications like RFC2547bis require additionally to scalability and stability also faster convergence

Options for Improvement

- **MP-BGP as protocol suites very well for the new applications**
- **Current MP-BGP implementations can be improved to achieve better convergence e.g.:**
  - **Use of event triggered implementation and back-off algorithms at the same time**
    - Eliminates delays of scan timers and protects the system in case of event storms
  - **Implementation of event queues and schedulers**
    - Allows to give preference to selected routing information over the rest

Thank You!

Policy based Calculation of the Internet Topology

Thomas Schwabe, Abhijit Chowdhury
TU München

Technische Universität München
Lehrstuhl für Kommunikationsnetze
Prof. Dr.-Ing. J. Eberspächer

Outlook

1. Motivation
2. Border Gateway Protocol - BGP
3. Network Management Databases
   - e.g. RIPE database
4. Idea
5. Example
6. Conclusions and Outlook

Motivation

- Resilience important topic for future networks
  - Need for disjoint backup routes
- How to find disjoint Inter-Domain routes?
  - No trust of BGP path information
- Topology information of an AS depends on BGP policies of its neighbors
- Idea:
  - Combine BGP policies of neighboring ASs
  - Better topology information on IP Layer

Border Gateway Protocol

- Inter-Domain Routing protocol in today‘s Internet – BGPv4
- Selection of the best route
  - Import into the Intra-Domain routing protocol
  - Announce to the neighboring networks
  - Based on
    - BGP Policies
    - Path attributes
    - AS hop count (part of the BGP announcement)

Border Gateway Protocol 2

- It works, but it may be not fulfill the requirements of future Inter-Domain routing:
  - Convergence time
  - Scalability
  - Number of BGP messages

BGP Policies

- Rules for filtering of BGP messages
- Reason for the success of BGP
  - Realization of business relationship of AS provider
- Provides a lot of problems
- Influence of routing decisions of neighboring networks
  - Setting of BGP Attributes
  - Announcement of selected routes

Network Management Databases

- Examples:
  - RIPE
  - ARPIC
  - etc.
- RIPE (Réseaux IP Européens)
  - open collaborative community of organisations and individuals,operating wide area IP networks
  - Founded in 1989
  - Region: Europe, Middle East

AS8208

```
% This is the RIPE Whois server.
% The objects are in RPSL format.
%
% Rights restricted by copyright.
% See http://www.ripe.net/ripencc/pub-services/db/copyright.html

as-block:      AS8192 - AS9215
descr:         RIPE NCC ASN block
remarks:       These AS numbers are further assigned by RIPE NCC
remarks:       to LIRs and end-users in the RIPE NCC region
remarks:       Please refer to these documents
remarks:       <http://www.ripe.net/ripe/docs/ir-policies-procedures.html>
               <http://www.ripe.net/ripe/docs/asnrequestform.html>
               <http://www.ripe.net/ripe/docs/asnsupport.html>
admin-c:       CREW-RIPE
tech-c:        OPS4-RIPE
mnt-by:        RIPE-NCC-HM-MNT
mnt-lower:     RIPE-NCC-HM-MNT
changed:       hostmaster@ripe.net 20010423
changed:       hostmaster@ripe.net 20011024
changed:       hostmaster@ripe.net 20011120
changed:       hostmaster@ripe.net 20020408
source:        RIPE

aut-num:       AS8208
as-name:       CAMELOT-AS
descr:         Teamware GmbH
descr:         Stahlgruberring 11
descr:         81829 Muenchen
descr:         Germany
import:        from AS1273 action pref=50; accept ANY
import:        from AS8767 action pref=100; accept ANY
import:        from AS25063 action pref=100; accept AS25063
export:        to AS1273 announce AS-CAMELOT
export:        to AS8767 announce AS-CAMELOT
export:        to AS25063 announce ANY
admin-c:       AI179-RIPE
tech-c:        AI179-RIPE
tech-c:        FB23-RIPE
remarks:       multi-homed AS
mnt-by:        CAMELOT-MNT
```

Idea

- use BGP policies from a WhoIs database
- combine BGP policies from neighboring ASs
- get connectivity information
- additional relationship between Neighbors
  - Peering – exchange traffic without payment
  - Customer – Provider – customer pays for forwarded traffic of the provider

import: from AS1273 action pref=50; accept ANY

- AS 1273 -> connection between AS 8208 and AS 1273
- Pref=50 – value of the BGP attribute “Local Preference”
- Accept Any – base for relationship decision:
  - peering or Customer - Provider

Example entries

AS25063 (Inotronic)
import: from AS8208 action pref=50; accept ANY
export: to AS8208 announce AS25063

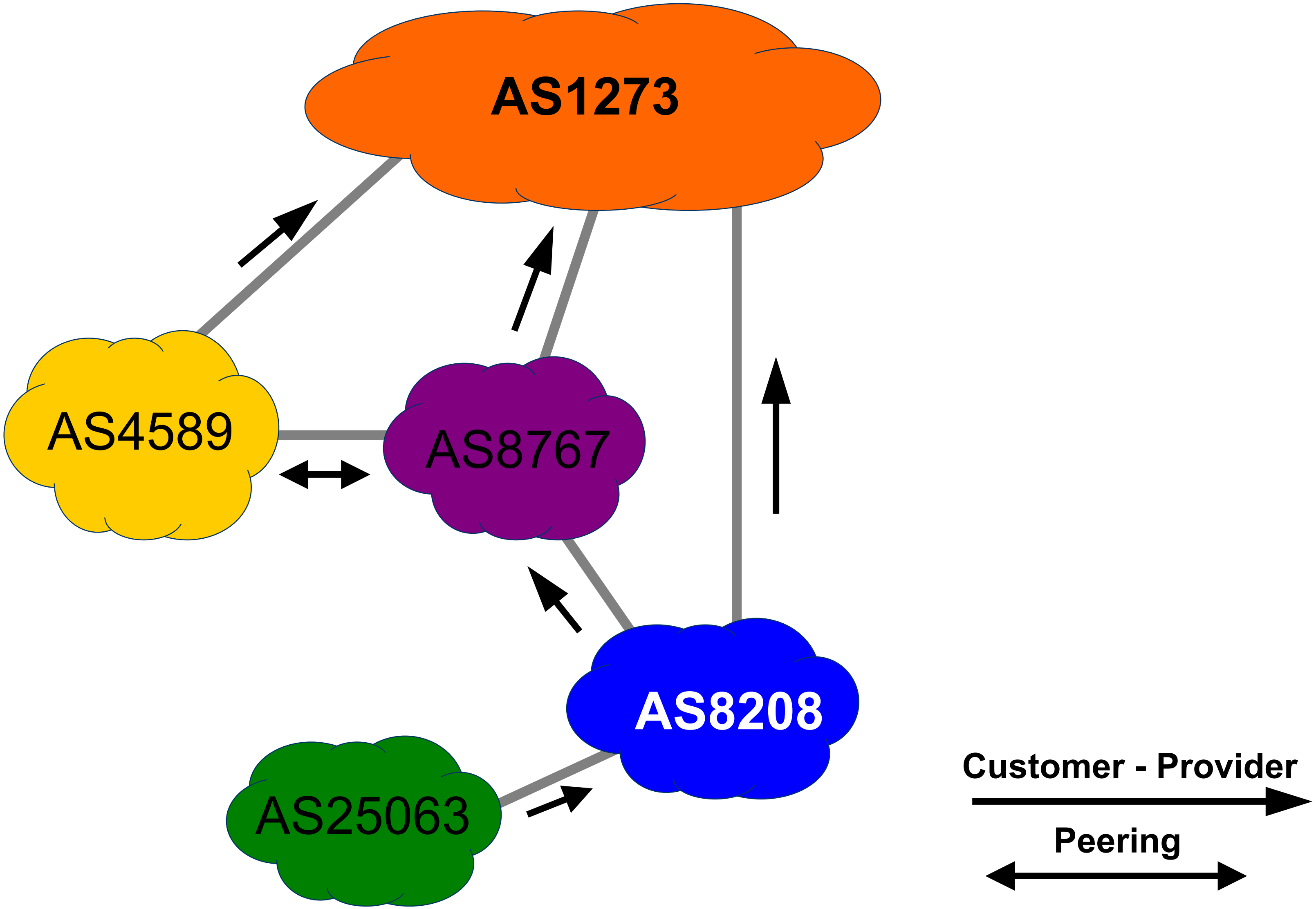
AS8208 (Camelot)
import: from AS1273 action pref=50; accept ANY
import: from AS8767 action pref=100; accept ANY
import: from AS25063 action pref=100; accept AS25063
export: to AS1273 announce AS-CAMELOT
export: to AS8767 announce AS-CAMELOT
export: to AS25063 announce ANY

AS8767 (M'Net)
import: from AS1273 action pref=100; accept ANY
import: from AS4589 action pref=80; accept AS-EASYNET
import: from AS8208 action pref=80; accept AS-CAMELOT
export: to AS1273 announce AS-MNETDE
export: to AS4589 announce AS-MNETDE
export: to AS8208 announce ANY

AS 1273 (Cable&Wireless)
import: from AS8767 accept AS-MNETDE
import: from AS4589 accept AS-EASYNET
export: to AS8767 announce ANY
export: to AS4589 announce ANY

AS4589 (Easynet)
import: from AS1273 accept ANY
import: from AS8767 194.59.190.42 at 194.59.190.8 accept AS-MNETD
export: to AS1273 announce AS-EASYNET
export: to AS8767 194.59.190.42 at 194.59.190.8 announce AS-EASYNET

Example topology



Conclusions 1

- BGP policies can be used for finding Internet topology information
  - Not only connectivity
  - Forecast of Inter-Domain routing
- No bases for Inter-Domain routing

Conclusions 2

- Base for finding of disjoint routes
- Looking for topology between source an destination
- Topology from a specified starting point (AS)
  - sophisticated limits
  - calculation of the whole topology – too complex

Outlook

- Further Work:
  - Comparison with real BGP routing
  - Find topology between specified source destination
  - More efficient tool for evaluation of the BGP policies

IBGP – full mess^Hh?

Probleme – Lösungsansätze – Bewertungen

IDRW-Workshop, Karlsruhe, 18.09.2003

Stefan Mink, Schlund+Partner AG

Recap: EBGP vs. IGBP

- E(xternal)-BGP
  - Zwischen unterschiedlichen ASen
  - Default: Alle aktiven BGP-Routen werden weitergegeben
  - Loop-Erkennung durch AS-PATH-Attribut
- I(nternal)-BGP
  - Session innerhalb eines ASes
  - Multihop-Sessions möglich und üblich
  - Nur aktive, eigene oder direkt empfangene externe Routen werden weitergegeben
  - Loop-Vermeidung durch **Vollvermaschung**

Problem von IBGP: Full Mesh

- Skalierbarkeit
  - Schlund (kleines Netz): 25 BGP-Speaker, hätte somit 25*24/2=300 BGP-Sessions
  - Inbetriebnahme des nächsten Routers
    - 25 neue IBGP-Sessions
    - Änderungen an der Konfiguration jedes Routers
  - Aufbau einer neuer Transit-Verbindung:
    - 25 Session *130K Routen * 50 Bytes/Update: >160 MByte Daten (1 NRLI pro Update, AS-Path-Länge von 3)
    - ½ bei Versendung von 2 NRLIs pro Update

Problem von IBGP – II

- Risiko
  - Konfiguration ist automatisierbar, jedoch Fehler mit globaler Auswirkungen möglich
  - Fehlen/Beendigung einer BGP-Session kann zu Loops und Blackholes führen
- Stabilität
  - Auswirkung von Massen-Versand auf andere Dienstes (CPU, Speicher, Bandbreite)

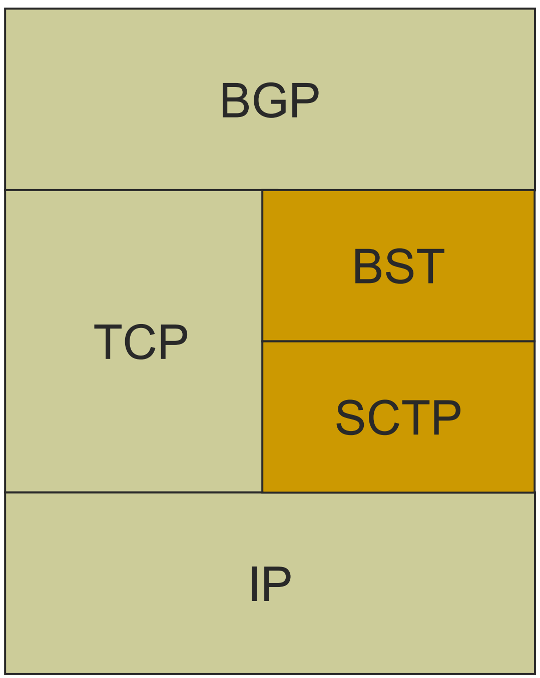
Lösung No. 1: BGP Scalable Transport

- Ansatz von K.Poduri, C. Alaettinoglu, V. Jacobson [BST]
- Grundgedanke von BST
  - BGP an sich ist stabil und gut
  - BGP an sich ist schwer austauschbar
  - Full-Mesh ist reines Transportproblem:
    - Transport gleicher Daten an Empfänger via Pt-to-Pt-Verbindungen
    - Phys. Topologie entspricht nicht den PtP-Verbindungen

BST: Point-to-Multipoint-Protokoll

- Protokollalternativen
  - Multicasting auf L3/4, z.B. *Pragmatic General Multicast* (PGM [RFC3208])
  - Multicasting auf L3/8, z.B. IS-IS [RFC 1142], OSPF [RFC2328])
- BST
  - folgt dem 2. Ansatz
  - Grund: leichtere Umsetzbarkeit

BST: Stack-Integration



BGP Scalable Transport:

(Variante von SRM)

- Reihenfolgetreue
- Zuverlässigkeit
- Robustheit

Stream Transmission Control Protocol [RFC2960]***:***

- Flusskontrolle
- Authentifikation

BST: Datenverteilung

- Fluten auf Anwendungsebene
  - Keine Topologie-Information nötig
  - Keine Konzentration von Updates auf wenigen Verbindungen
  - Redundante Anbindung des Routers resultiert in redundanter Versendung von Routingdaten, somit erhöhter Zuverlässigkeit

BST: Konfiguration

- Vereinfachungen durch
  - Virtualisierung der Empfänger
    - Eine logische Adresse als BGP-Zieladresse
    - Versand von *Hellos* (evtl. authentifiziert mit MD5 + shared secret [RFC2385])
  - Autodiscovery ermöglicht Verzicht auf explizite Konfiguration
  - Kopplung von Nicht-BST-Routern an das BST-Mesh angeblich möglich

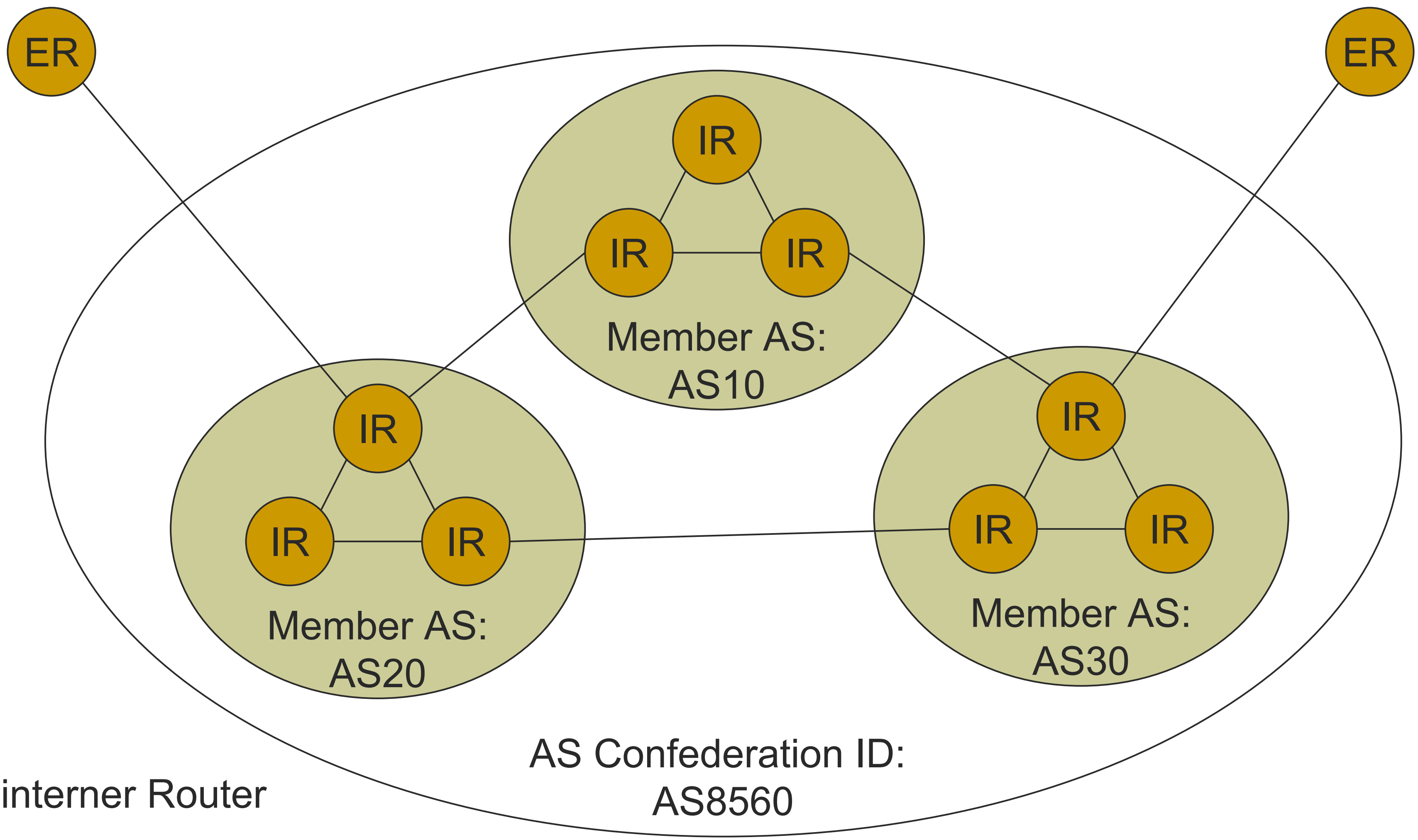
BST: Bewertung

- Vorteile
  - Transparenz für BGP
  - Zuverlässigkeit, Skalierbarkeit
  - Geringere Bandbreite und CPU (?)
- Nachteile
  - Announcements bei neuem Router?
  - Granularität sinkt, keine individuellen Policies mehr möglich (nötig bei IBGP?)
  - Proprietärer Ansatz findet selten Akzeptanz

Lösung No. 2: Confederations

- Erweiterung des BGP-Standards [RFC 3065]
- Klassisches Divide & Conquer
- Lösungsansatz:
  - AS wird zur Konföderation von Unter-ASen
  - Full-mesh innerhalb der Unter-ASe
  - Verbindung zwischen Unter-ASen ähnlich wie zwischen regulären ASen, vermitteln **nur aktive Routen** zwischen Sub-ASen

BGP-Confeds: Visualisierung



ER: interner Router

IR: interner Router

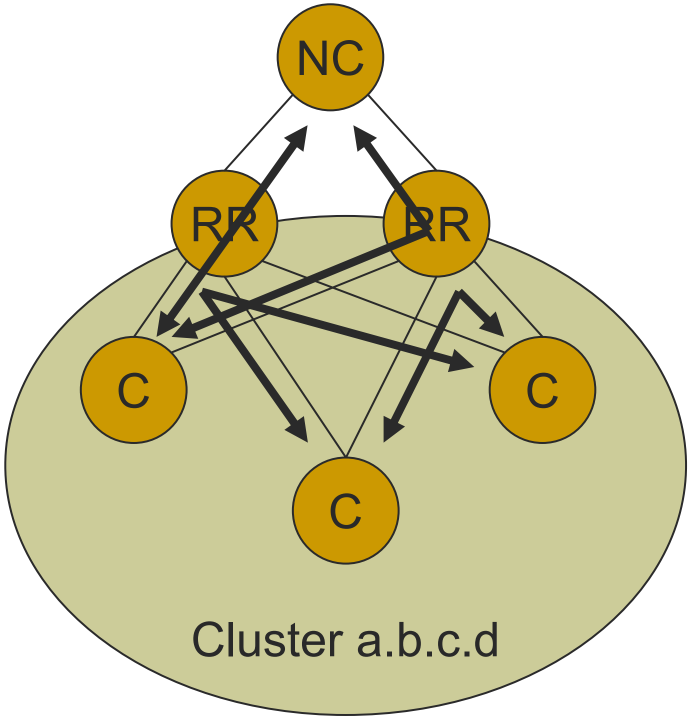
BGP-Confeds: Bewertung

- Vorteile
  - Offener Standard
  - Vollvermaschung nur innerhalb eines Sub-ASes notwendig
- Nachteile
  - Nicht transparent, alle Router müssen es unterstützen
  - Nicht schachtelbar
  - Persistent Route Oscillation (später mehr)

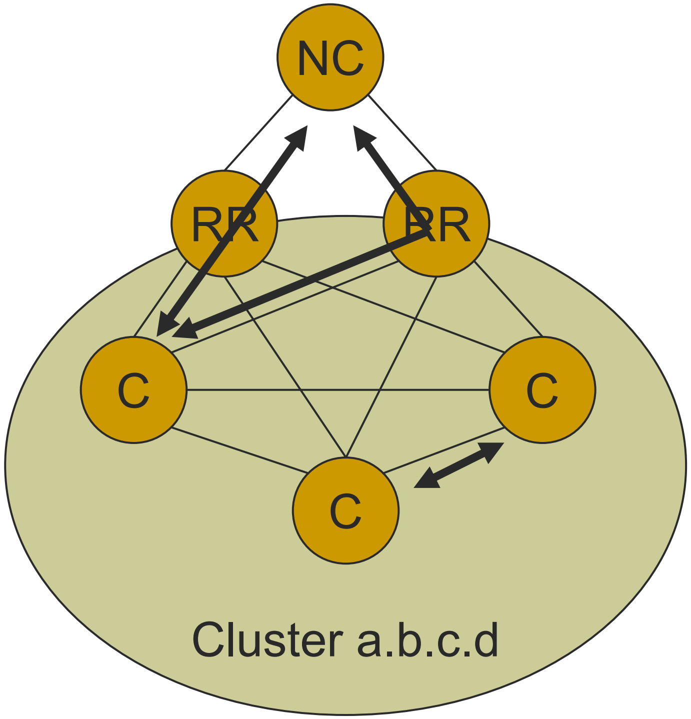
Lösung No. 3: Route Reflector

- Erweiterung des BGP-Standards [RFC 2796]
- Lösungsansatz:
  - Manche Router fungieren als Route-Reflektoren, die zwischen Gruppen von Routern Routen vermitteln
  - Clients: vollvermascht oder isoliert
  - Non-Clients: vollvermascht
  - Reflektoren vermitteln **nur aktive Routen** zwischen beiden Gruppen

RR: Visualisierung



Isolierte Clients (Client-to-Client-Reflection)



Client-Full-Mesh

NC: non-client, C: client, RR: route reflector

RR: Bewertung

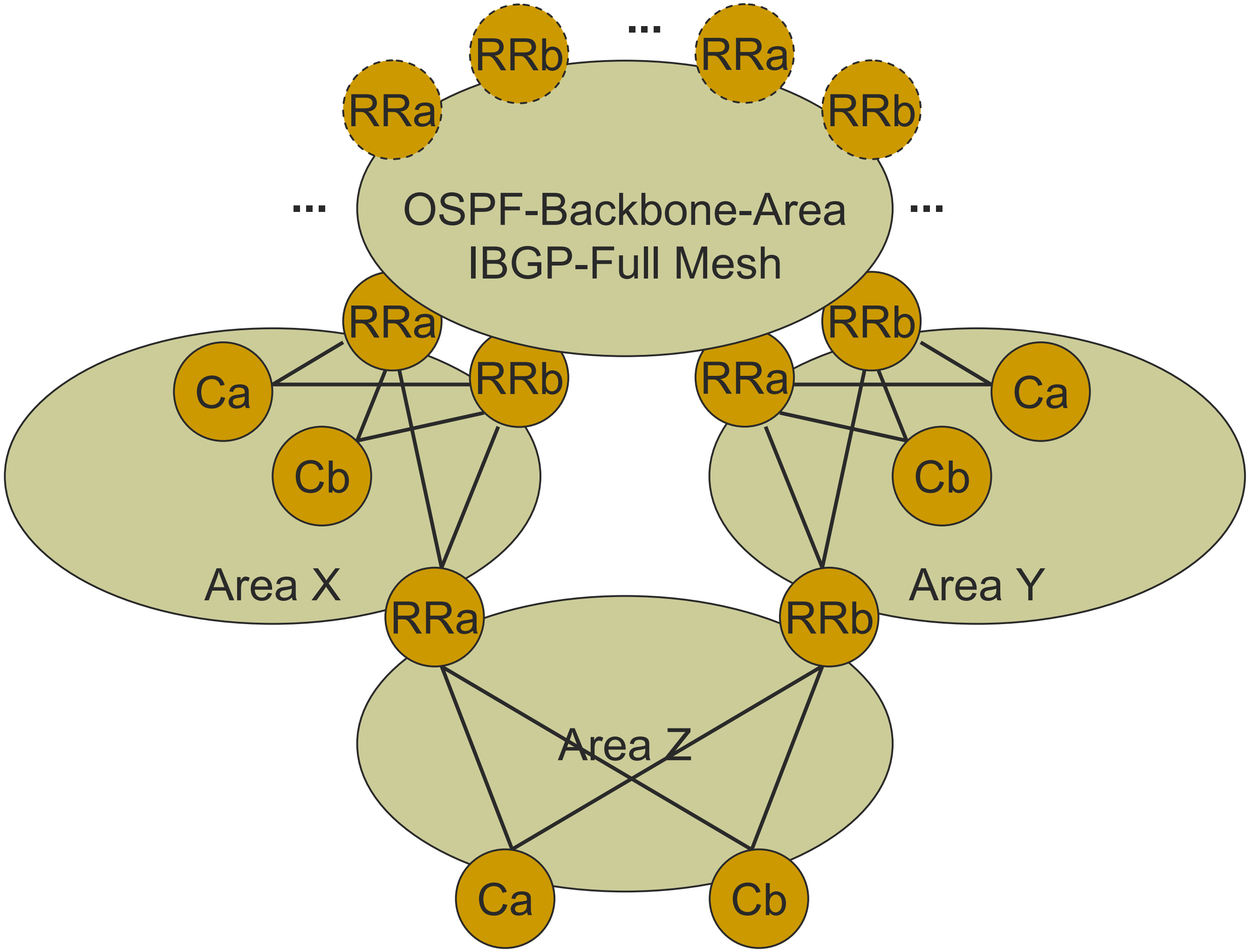
- Vorteile
  - Offene Erweiterung
  - Transparent: nur RRs müssen die Erweiterung unterstüzen
  - Schachtelung von Clustern möglich, Loop-Detection via CLUSTER_LIST

- Nachteile
  - Persistent Route Oscillation möglich
  - Transient Route Oscillation möglich

Status Quo Schlund: RRs

- Schlund
  - reduziert das Full-Mesh durch RRs von 25 auf 10 vollvermaschte Router (45 Sessions statt 300)
  - nutzt homogene Strukturen in IGP und EGP
    - ABRs sind gleichzeitig RRs, fast überall redundant ausgelegt, sogar Hersteller-redundant
    - Router, die via virtuelle Links an die Backbone-Area angebunden sind, stellen zweite RR-Stufe dar

Status Quo Schlund: Visualisierung



Neues Problem: *Persistent Route Oscillation*

- RRs und Confeds resultieren
  - in einer Teilvermaschung
  - in einer Einschränkung der Sicht von Routern außerhalb und sogar innerhalb eines Clusters
- Grund
  - RRs können nur **eine** aktive Route zu einem Ziel announcen (die zweite widerruft implizit die erste), dadurch Filterung von Routen durch RRs
  - Keine Vollordnung bzgl. MED

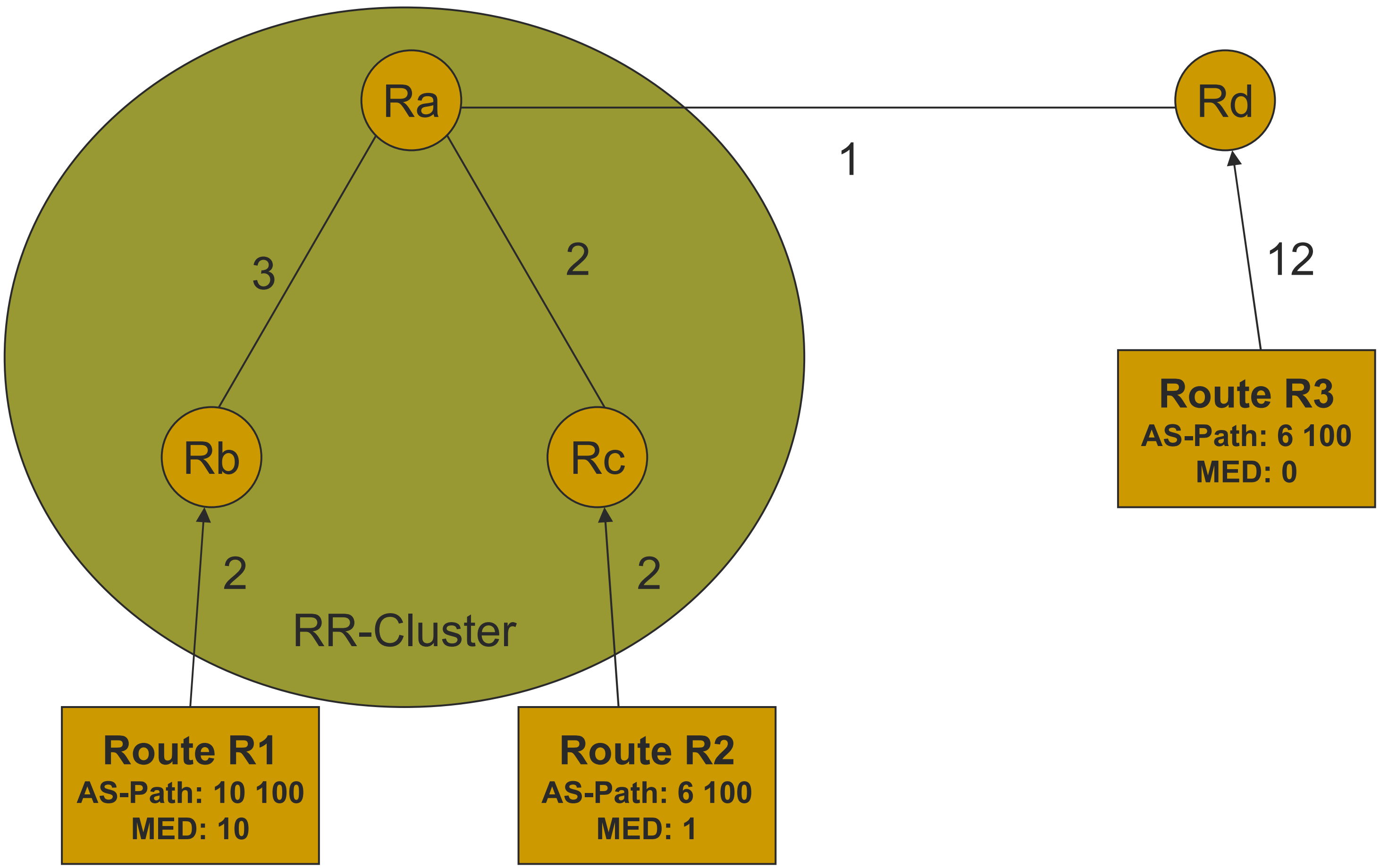
PRO: Ergebnis

- Inkonsistente Zustände im Netz, die eine Konvergenz verhindern
  - Zyklische Abhängigkeiten zwischen Routen-Announcements verhindern Konvergenz
  - Bei Schlund: 5 Backbone-Router „bombadieren“ sich gegenseitig mit >50 Updates pro Sekunde beim „Kampf“ um zwei Präfixe
- **Frage:** Wie sieht so ein Zyklus aus und wie entsteht er ?

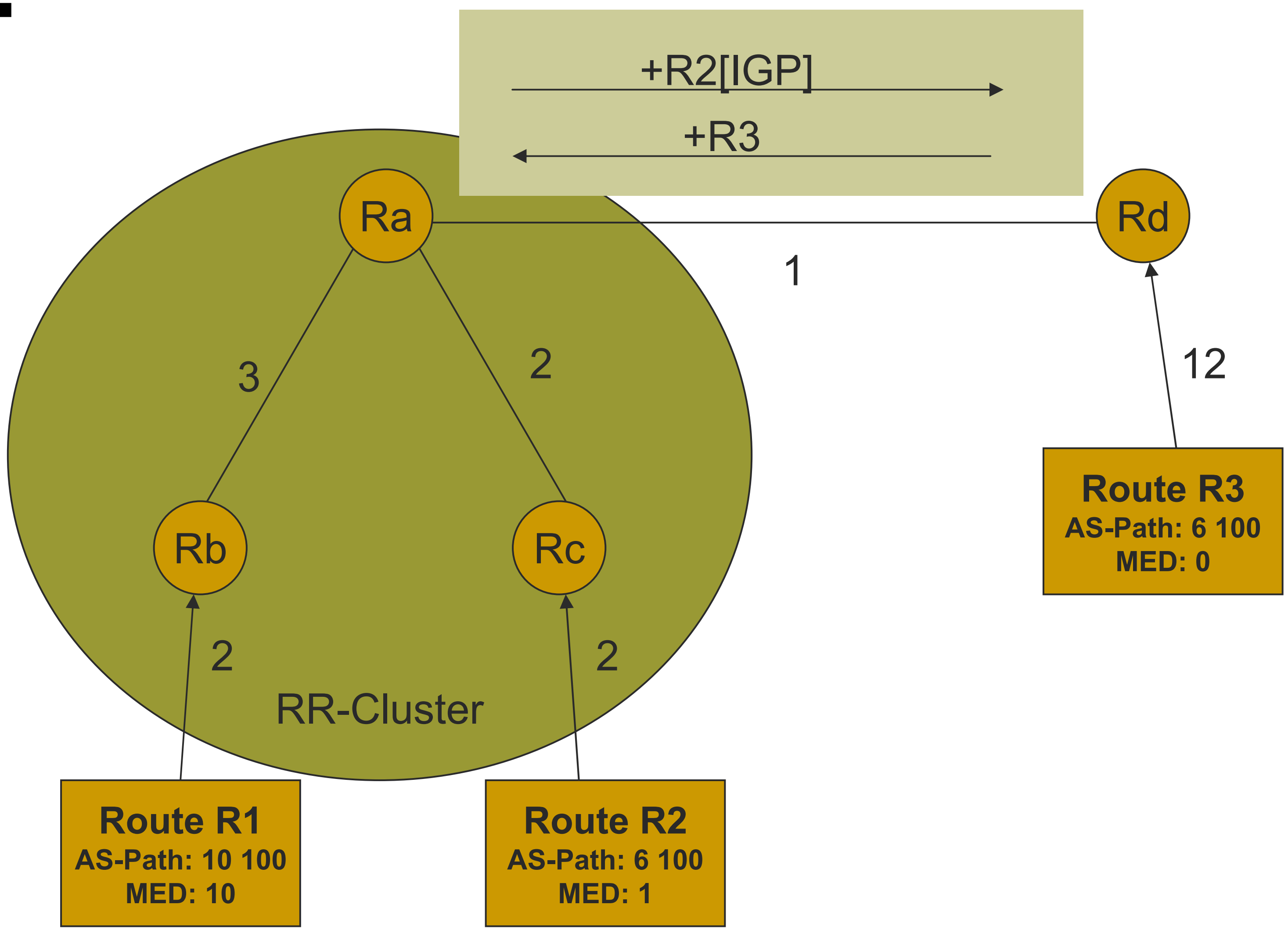
Recap: Routen-Selektion in BGP

- Wesentliche Kriterien:
  - Höchste LocalPreference
  - Kürzester AS-Pfad
  - Geringster MED
  - EBGP vor IBGP
  - Geringste Link-/IGP-Metric
  - Niedrigste Router-ID

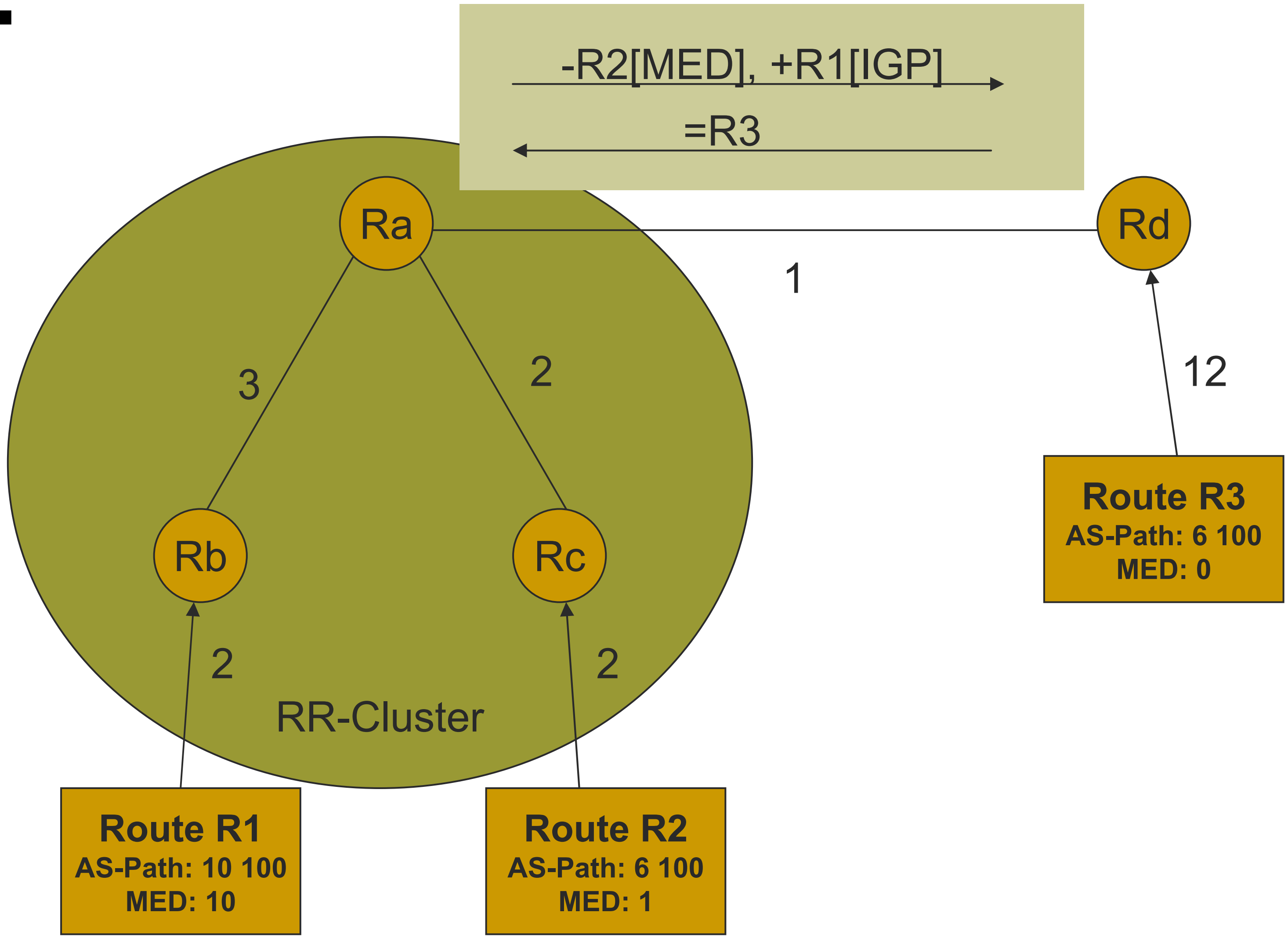
PRO: Beispiel-Szenario



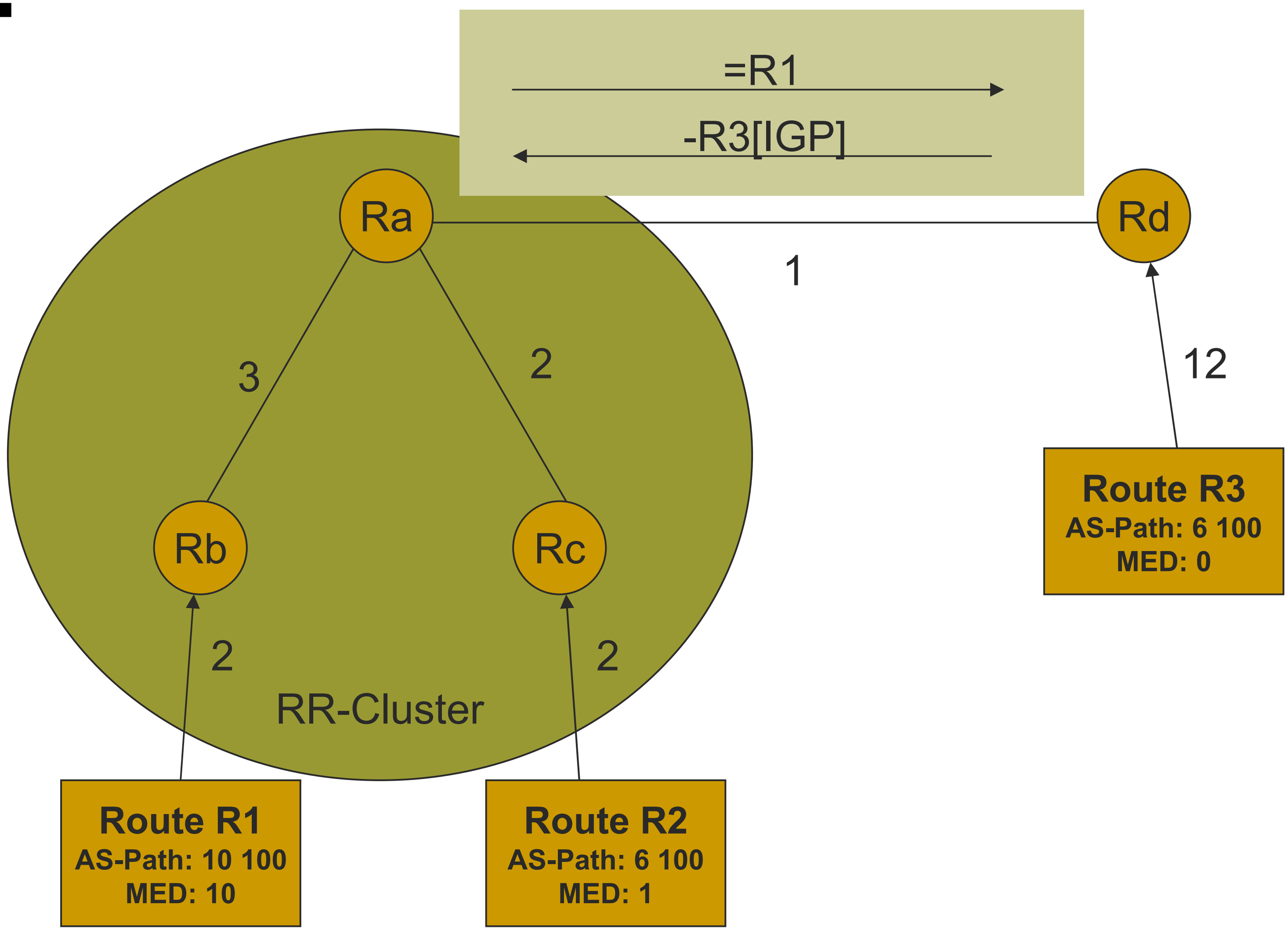
PRO: Beispiel Schritt 1



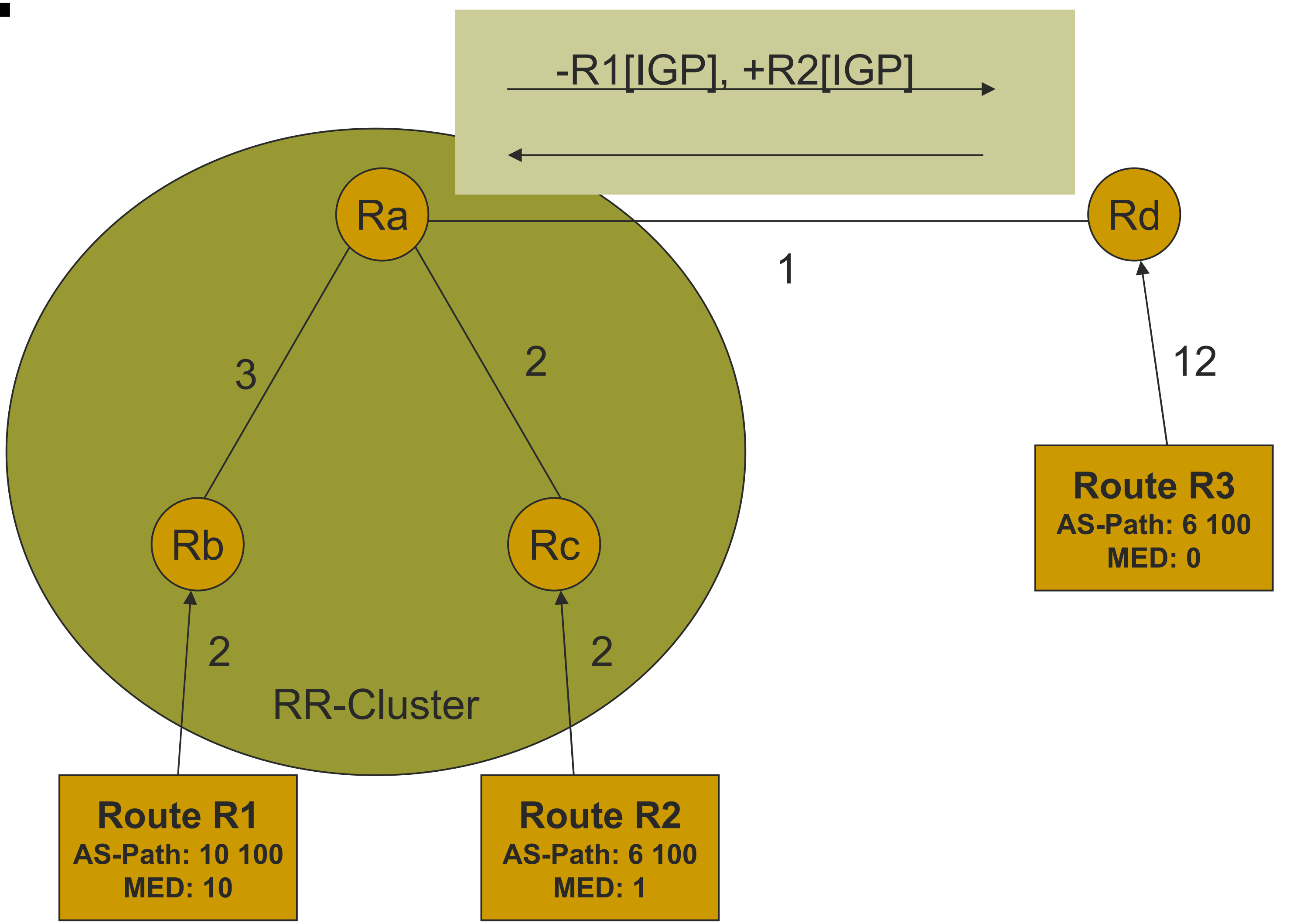
PRO: Beispiel Schritt 2



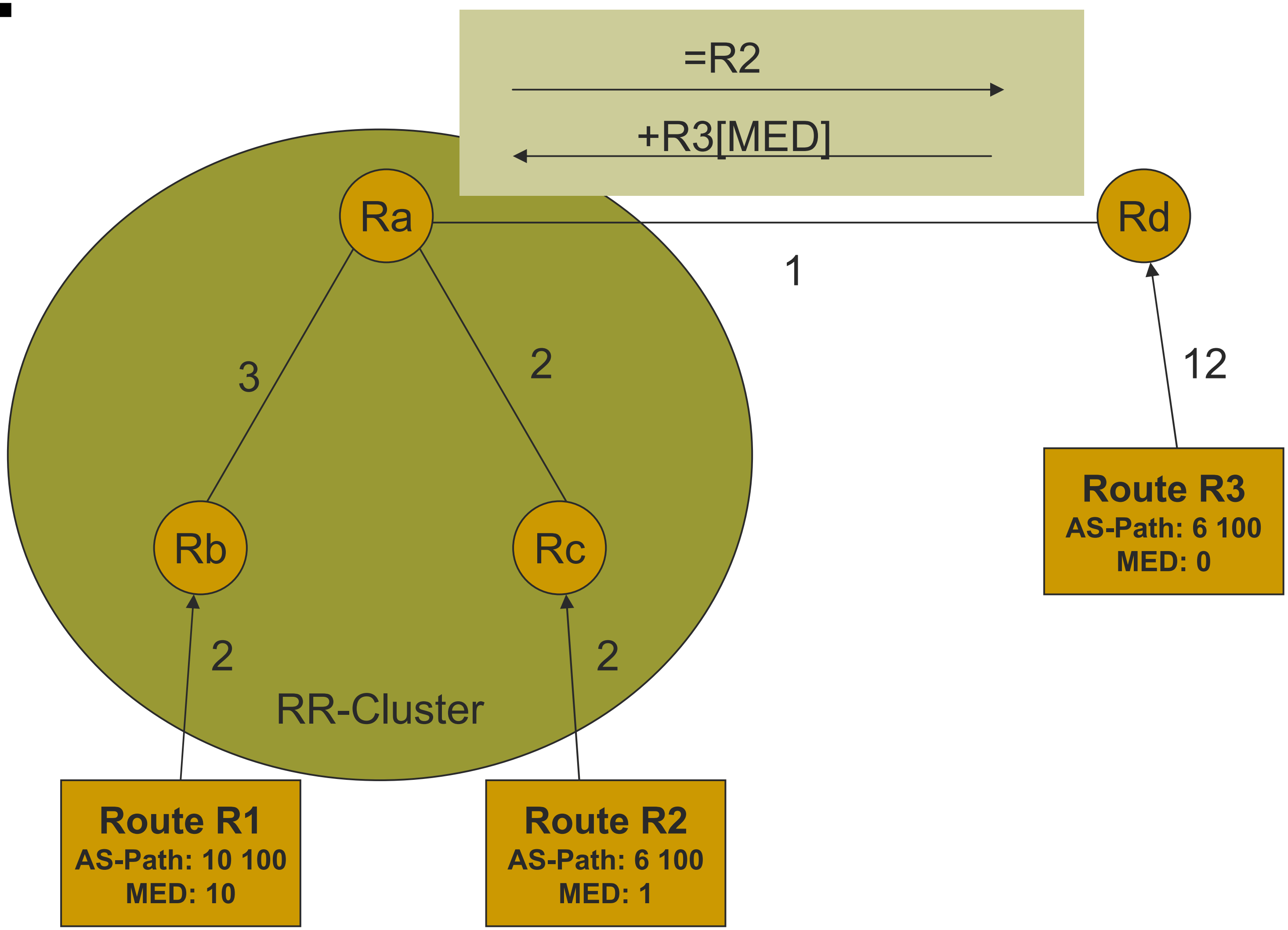
PRO: Beispiel Schritt 3



PRO: Beispiel Schritt 4



PRO: Beispiel
Schritt 5



Weitere Probleme

- [RFC 3345] beschreibt neben dem eben gezeigten PRO-Szenario noch ein weiteres.
- [RO] beschreibt zusätzlich zum gezeigten Szenario noch Szenarien, die nur temporäre Loops erzeugen, sog. *transient route oscillations.*

Lösungsansätze

- [RFC 3345] beschreibt PRO-Szenarien, als Lösungsansätze sind enthalten:
  - Intra-Cluster-IGP-Metriken >> Inter-Cluster-IGP-Metriken wählen, Analog bei Confeds
  - Keine MEDs verwenden
  - MEDs immer vergleichen (auch bei unterschiedlichen next-hop-ASen)
  - [Vollvermaschung einsetzen]

➔ alle Ansätze lösen zwar das Problem, „vergewalti-gen“ aber die Netzadministration und/oder den BGP-Standard

Lösungsansätze – II

- In diversen Drafts/Papers (z.B. [PRO1, PRO2, RO] favorisierte Lösung:
  - Ändern des BGP-Standards, dass ein Router zu einem Ziel **mehrere** verschiedene Pfade announcen kann
  - RRs announcen alle Routen, die nach MED-Auswertung noch übrig sind [RO]
- ➔ Korrekte Lösung des Problems
  - ➔ [RO] enthält Beweis (to be verified)
  - ➔ Bleibt die Frage nach Zeitpunkt der Umsetzung (vor allem in multi-Vendor-Netzwerken).

Status Quo Schlund

- Alle Router mit externen Anbindungen wurden wieder in das Full-Mesh aufgenommen
- Design-Prinzip-Bruch um Divergenz in Zukunft ausschließen zu können
- Warten auf Umsetzung von akzeptable Lösungen, z.B. [RO]

Wake-Up ;)

Fragen, Vorschläge, Anregungen?

Literatur

[BST] K. Poduri et al; NANOG 27: „BST – BGP Scalable Transport"; Februar 2003

[RO] A.Basu et al; SIGCOMM 2002: „Route Oscillations with I-BGP with Route Reflection"; 2002

[PRO1] John Scudder; NANOG 26: „BGP Route Oscillation Elimination"; October 2002

[PRO2] Daniel Walton et al; Internet Draft „BGP Persistent Route Oscillation Solution"; May 2002

[RFC 1142] D. Oran; „OSI IS-IS Intra-domain Routing Protocol"; Februar 1990

[RFC 2328] J. Moy; „OSPF Version 2"; April 1998

[RFC 2385] A. Heffernan; „Protection of BGP Sessions via the TCP MD5 Signature Option"; August 1998

Literatur – II

[RFC 2796] T. Bates;“BGP Route Reflection - An Alternative to Full Mesh IBGP“; April 2000

[RFC 2960] R. Steward; „Stream Control Transmission Protocol“; October 2000

[RFC 3065] P. Traina; „Autonomous System Confederations for BGP“; February 2001

[RFC 3208] T. Speakman et al; „PGM Reliable Transport Protocol Specification“; December 2001

[RFC 3345] D. McPherson et al; „Border Gateway Protocol (BGP) Persistent Route Oscillation Condition“; August 2002