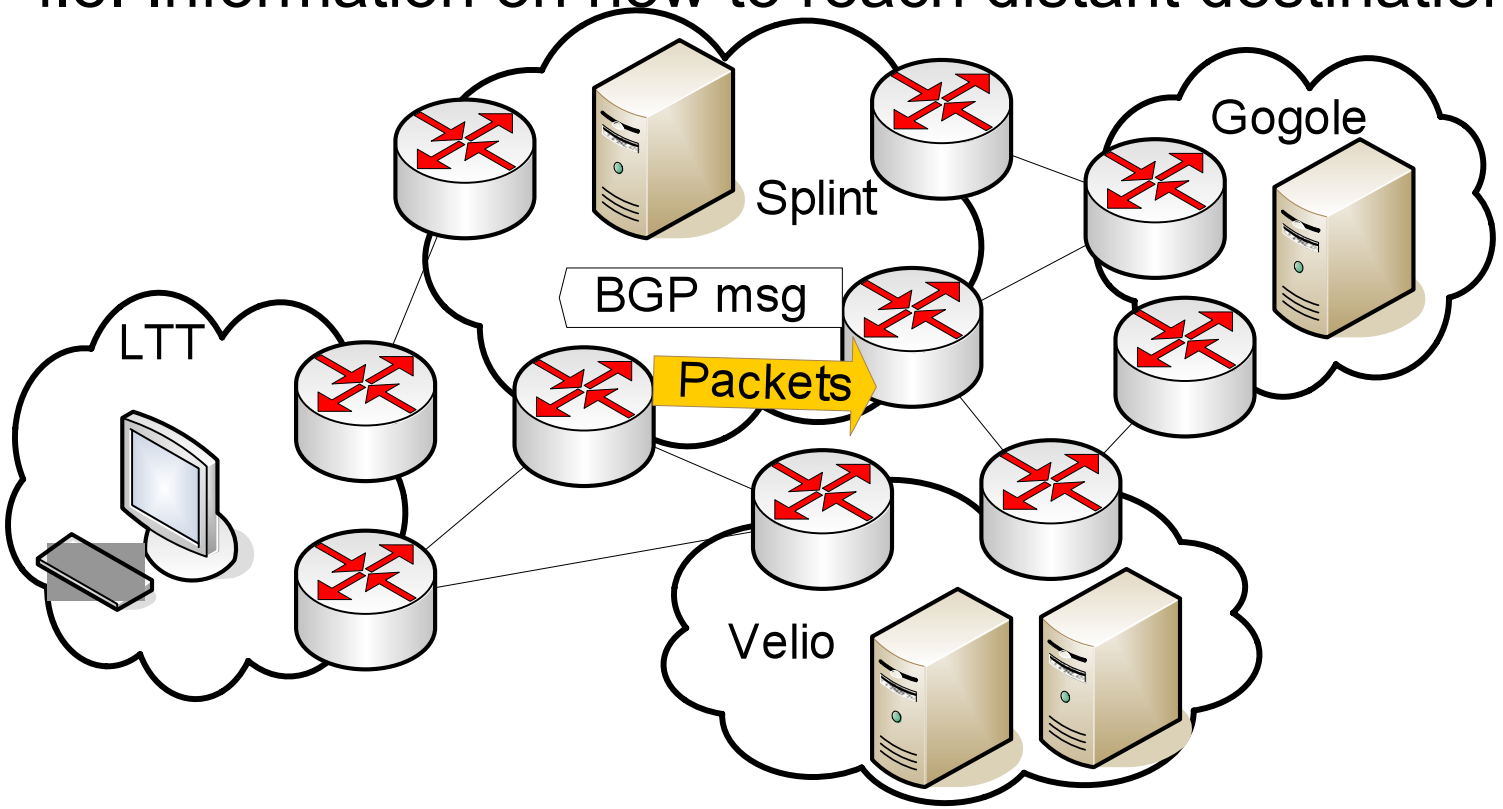
A novel internal BGP route distribution architecture

C. Pelsser (IIJ)
Joint work with A. Masuda, K. Shiomoto (NTT)

Loughborough University
March 2010

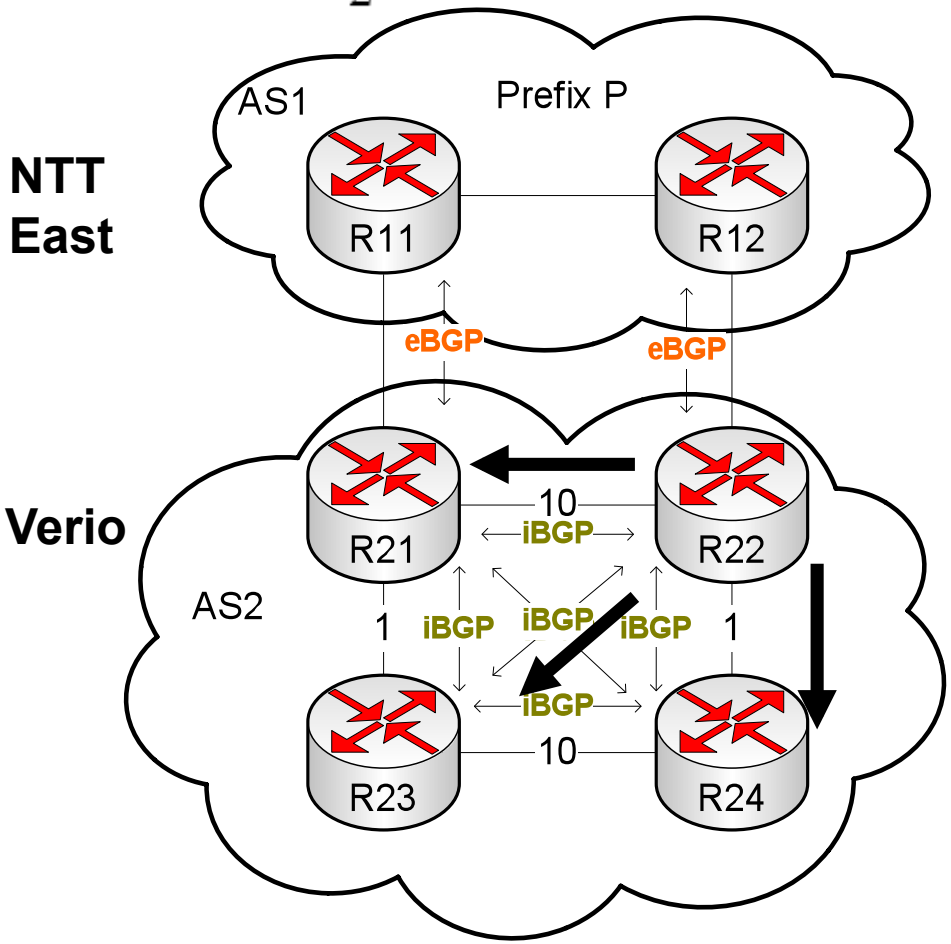
Background

- The Internet is composed of **domains**
  - A domain is also called an Autonomous System (**AS**)
- **BGP** distributes **routes** for destinations outside a domain
  - i.e. Information on how to reach distant destinations

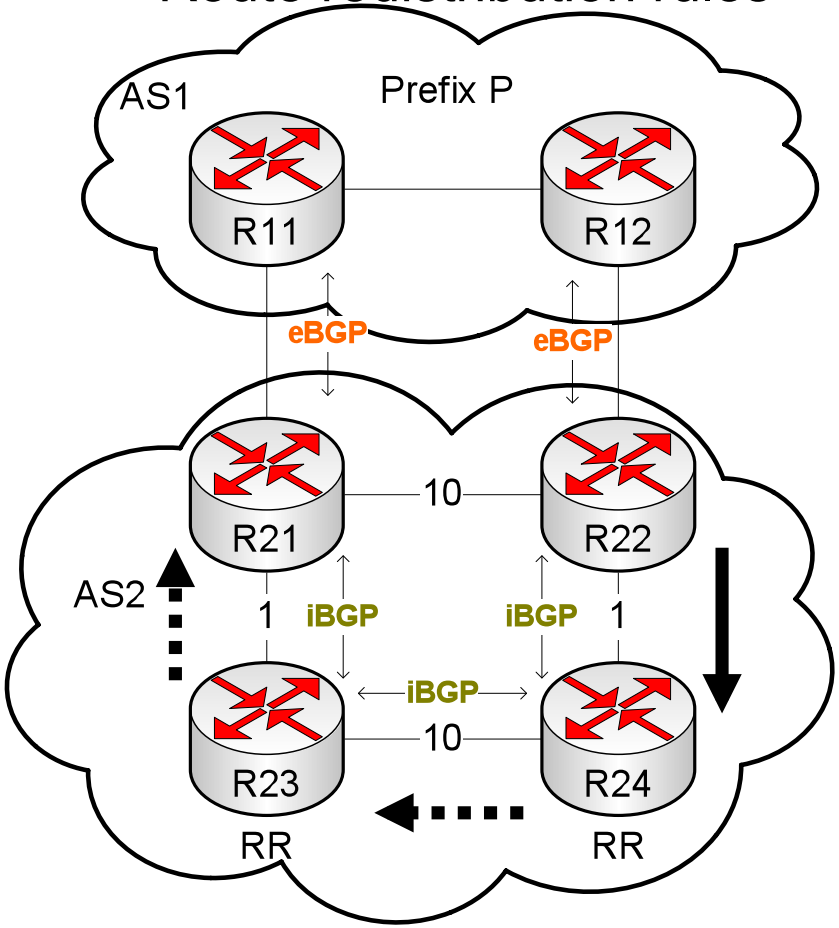


BGP route distribution

- iBGP full-mesh
  - $\frac{n * (n-1)}{2}$ sessions



- Route-reflectors
  - less sessions → scalable
  - Route redistribution rules

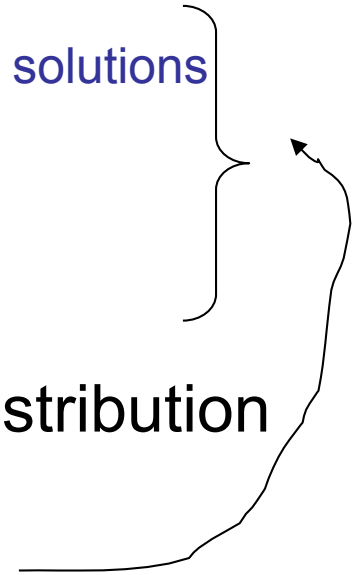


Problem statement

- Issue1: Full-mesh of iBGP sessions have a **scalability** problem

  → Route Reflectors (RR) and confederations have been introduced to solve this problem

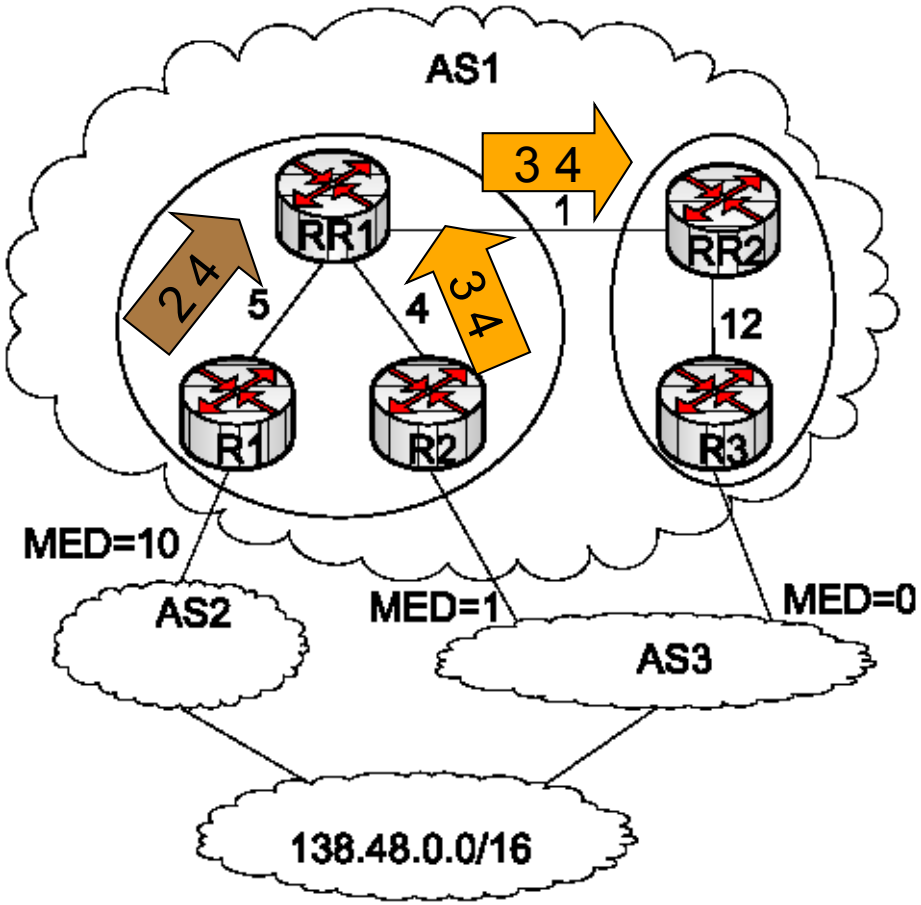
- Issue 2: However, RRs and confederations lead to the following issues
  - Routing convergence issues
    - **routing oscillations** and multiple routing solutions
  - Packet forwarding issues
    - deflection, **loops** and sub-optimal paths
  - Connectivity losses upon a failure



- We propose an internal BGP route distribution
  - Scalable
  - Immune to the issues mentioned above

Sparse iBGP topologies

- Each router **only learns the best route** selected by its peer
  - Not all the external routes
- Each router locally performs its **own selection**

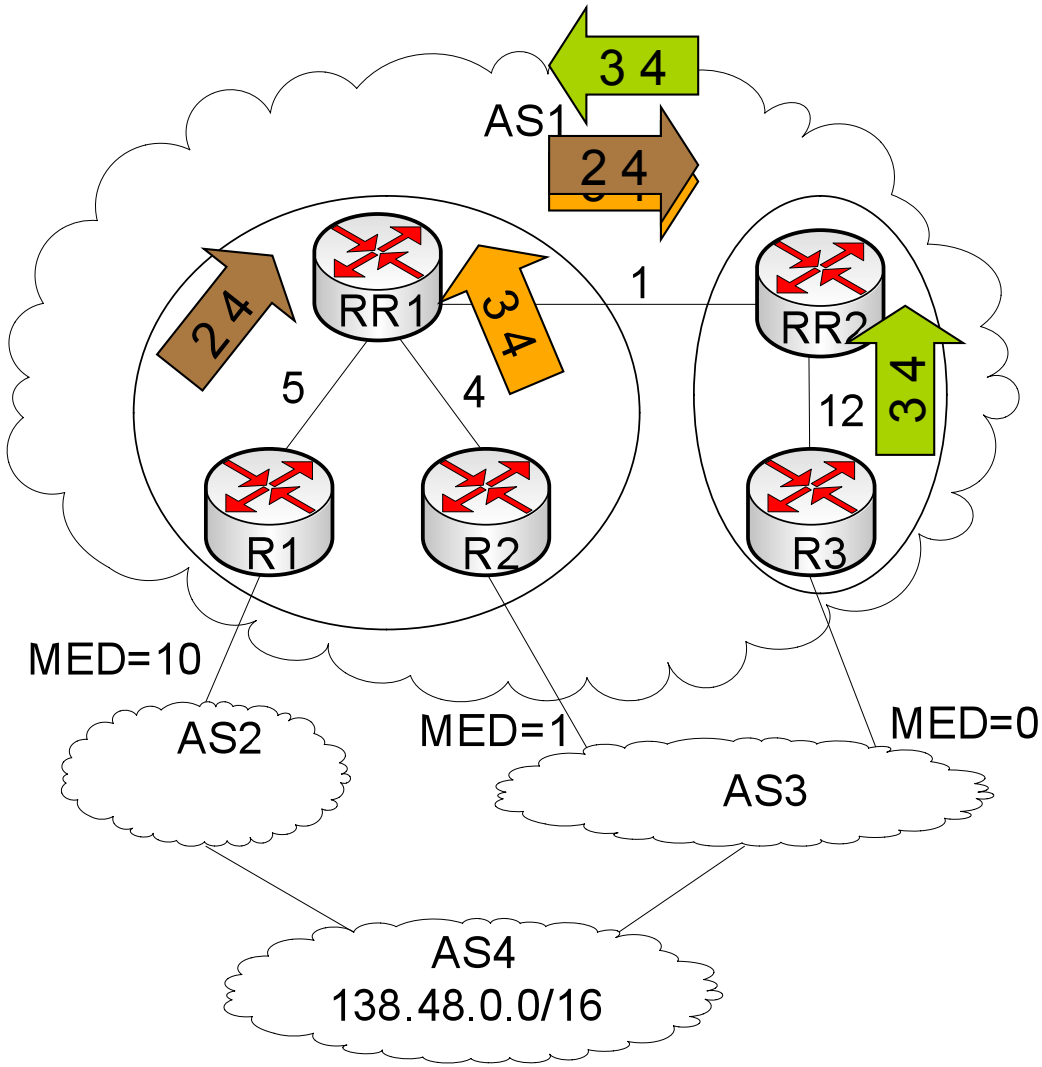


	BGP decision process
1.	Highest Local_Pref
2.	Shortest AS-Path
3.	**Lowest MED**
4.	eBGP over iBGP
5.	**Lowest IGP cost to NH**
6.	Tie-break

Sparse iBGP topology issues are due to a lack of route visibility at the routers

Routing convergence issues:

Example of a permanent route oscillation

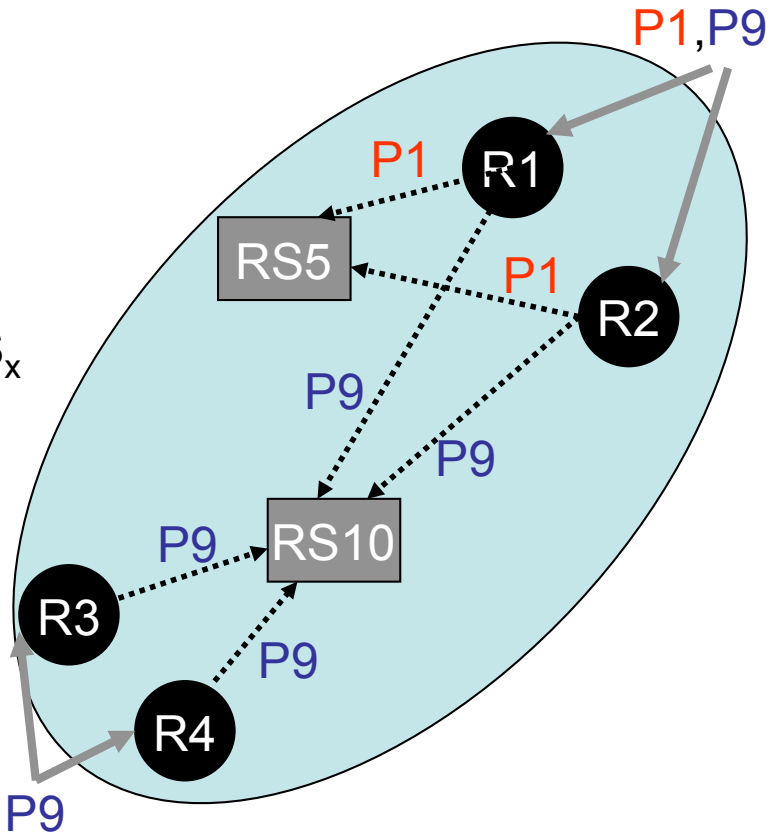


Router	AS-path	MED	IGP
RR1	>3 4	1	4
	2 4	10	5
RR2	>3 4	0	12
	3 4	1	5
RR1	3 4	1	4
	>2 4	10	5
	3 4	0	13
RR2	3 4	0	12
	>2 4	10	6
RR1	>3 4	1	4
	2 4	10	5

Our proposal (Issue 1):
Distributed Route Servers (RS)

Distributed → Scalabiltiy

- A Route Server computes routes on behalf of the routers
- Each RS, RS_x is responsible for a subset of prefixes
  - Prefixes with
    - key ≤ ID of RS_x
    - key > ID of RS_y with ID smaller than RS_x
- Each RS receives all the BGP messages for these prefixes



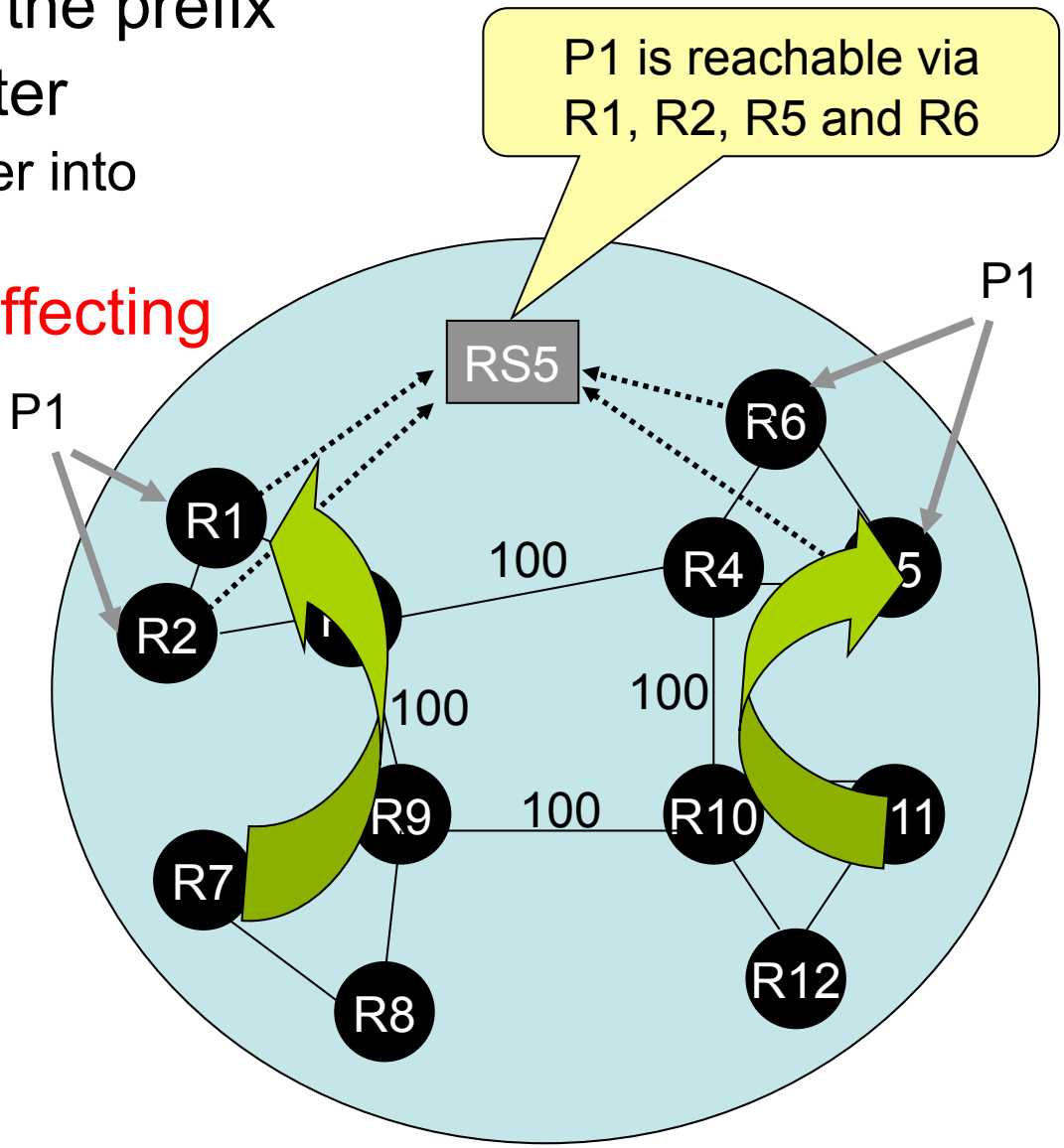
Our proposal (Issue 2):
RS decision process

- RS knows all routes for the prefix
- Route selection per router
  - Take location of the router into account

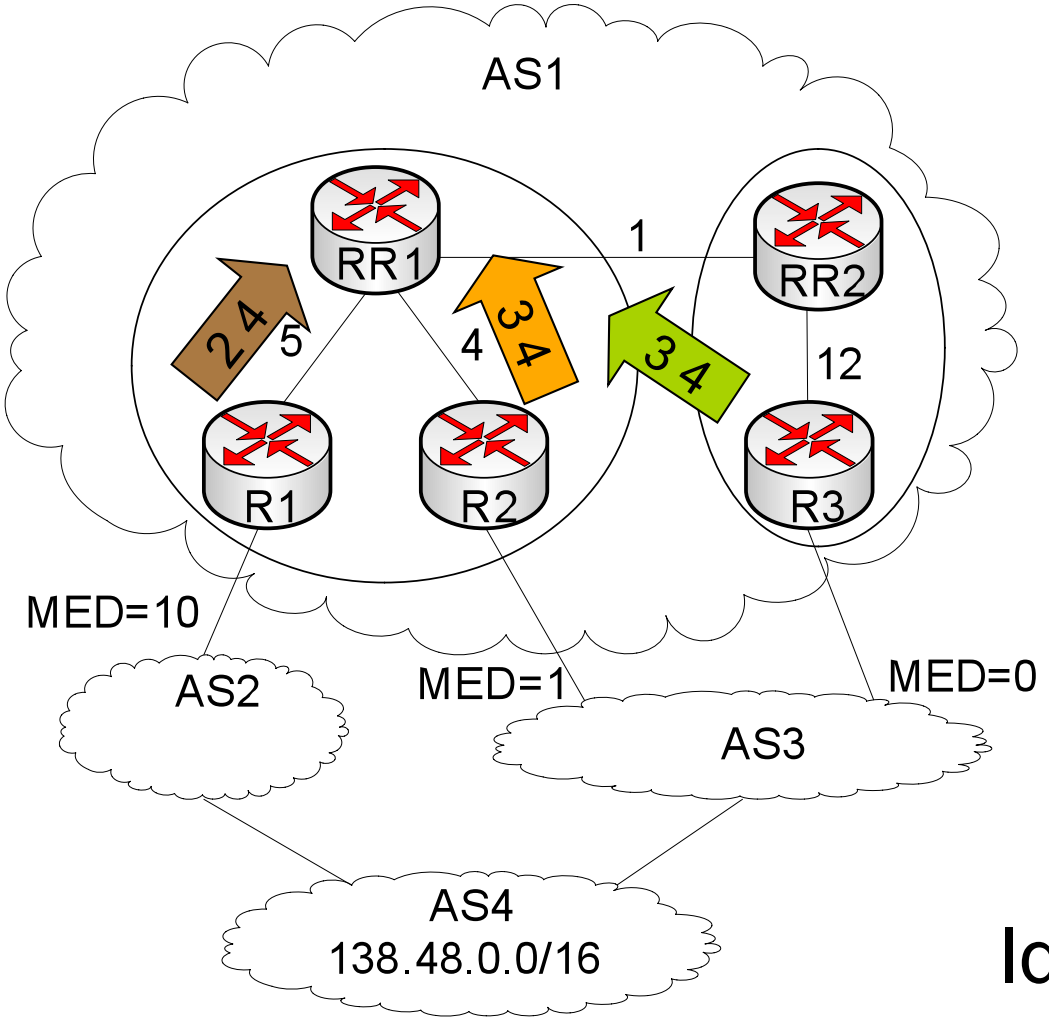
→ Immune to the issues affecting sparse iBGP topologies

	BGP decision process
1.	Highest Local_Pref
2.	Shortest AS-Path
3.	Lowest MED
4.	**eBGP over iBGP**
5.	**Lowest IGP cost to NH**
6.	Tie-break

All links have a cost of 1 except otherwise specified



Resolution for the permanent routing oscillation example



RR1 is RS for 138.48/16

Router	AS-path	MED	IGP
R1	>2 4	10	**0**
	3 4	***1***	9
	3 4	***0***	**18**
R2	>2 4	10	**9**
	3 4	***1***	0
	3 4	***0***	**17**

Idem for the other routers

iBGP issues resolved

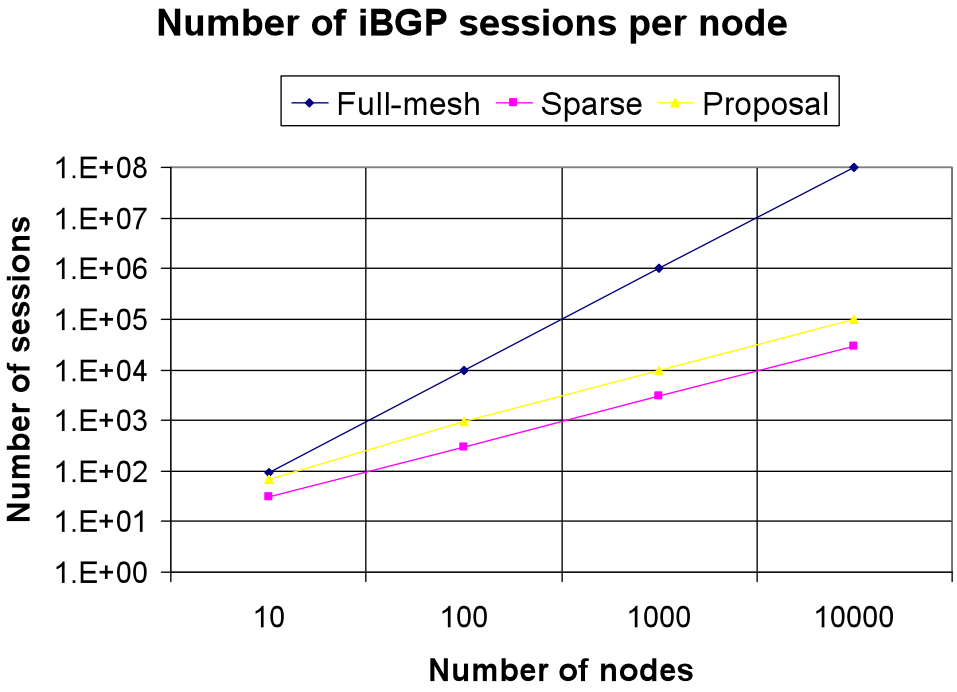
Advantages compared to sparse iBGP topologies	
1	No routing oscillations
2	Single routing solution
3	No forwarding loops
4	No deflection
5	Route diversity
6	Optimal paths according to the BGP DP

RRs and confederations are **sparse** iBGP topologies

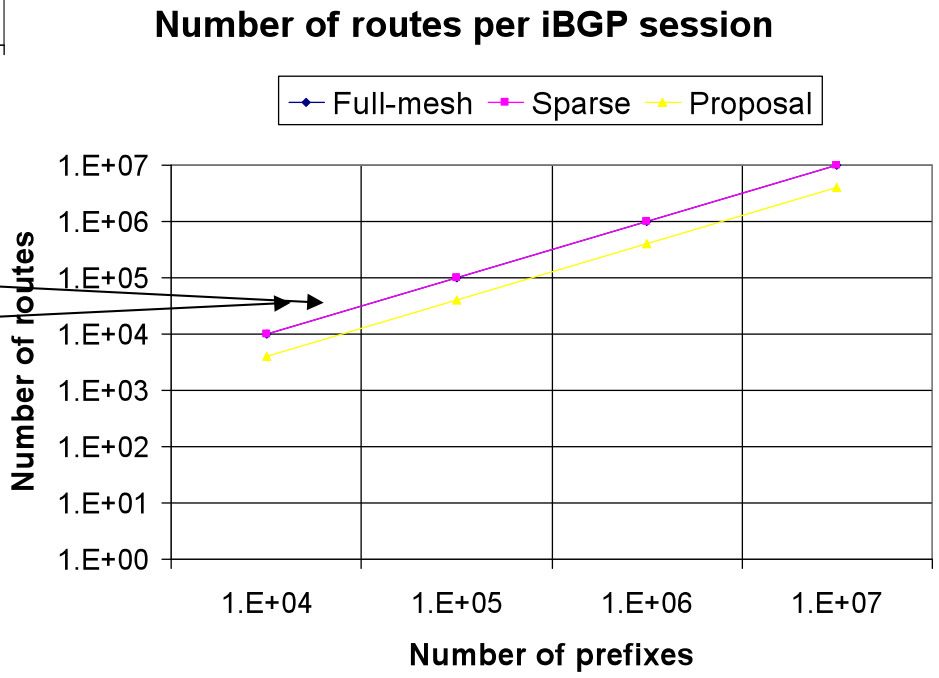
Scalability: metrics

- Number of iBGP sessions
  - Configuration burden
- Number of routes on an iBGP session
  - Bandwidth usage
- Table sizes
  - Memory consumption
- Number of BGP messages
  - Bandwidth usage, CPU load

Scalability



iBGP sessions
- $O(n^2)$ in a full-mesh
- O(n) for sparse
  - fixed average iBGP sessions degree
- O(n) for our proposal
  - fixed number of RS



Routes per iBGP sessions
- O(p) in a full-mesh
- O(p) for sparse
- O(p) for our proposal

Our proposal scales as sparse iBGP topologies

Scalability

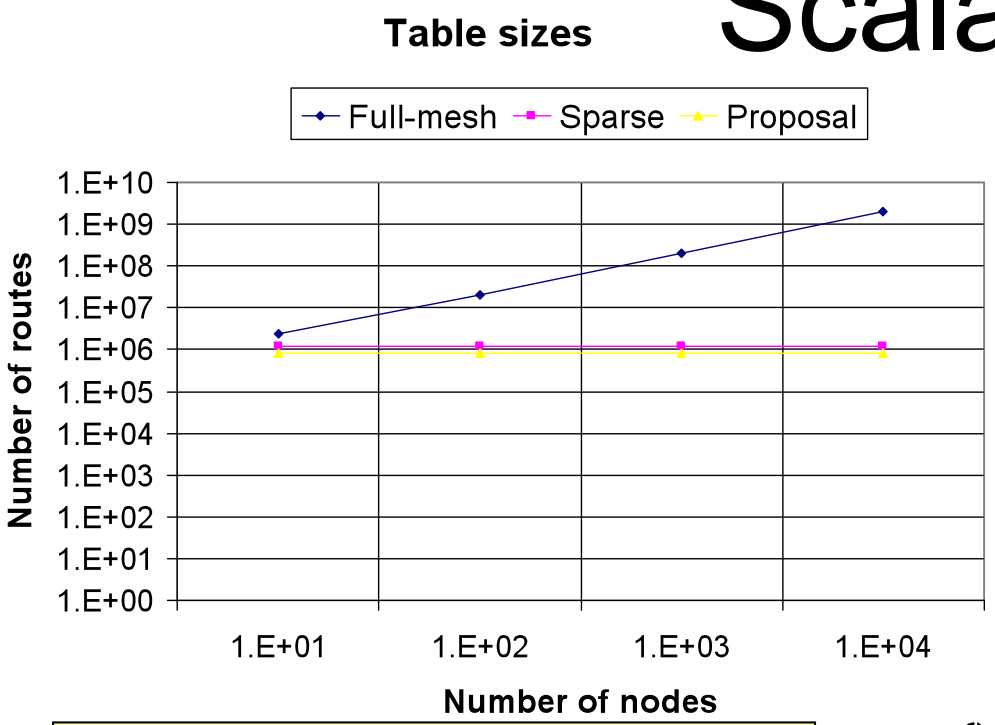
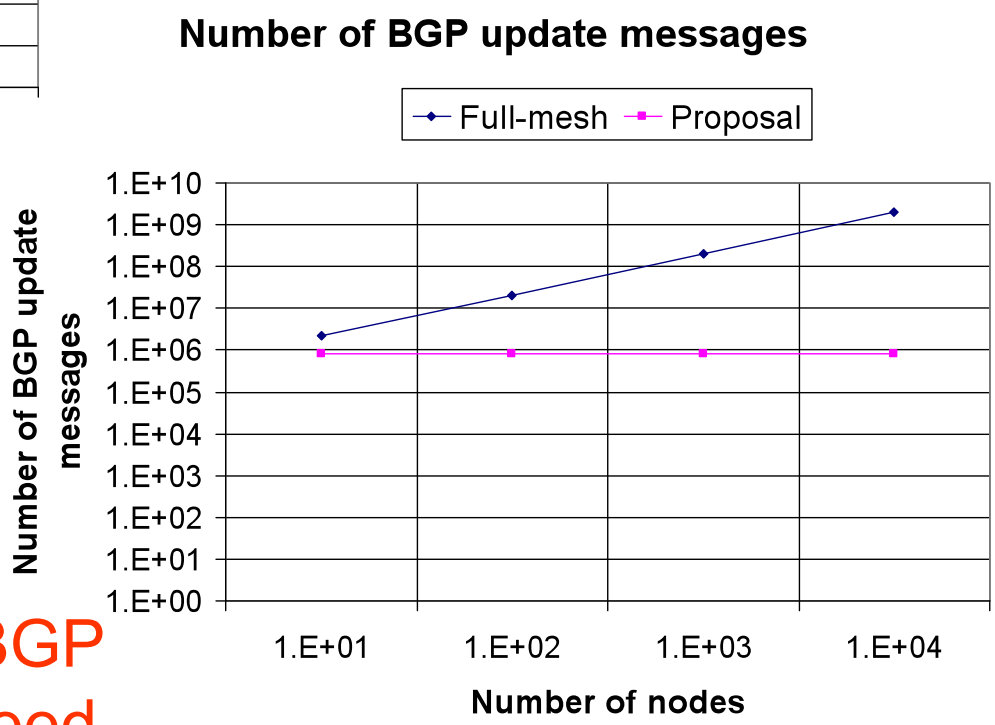


Table sizes
(fixed number of prefixes)
- O(n) in a full-mesh
- O(1) for sparse
- O(1) for our proposal

BGP update messages
(fixed number of prefixes)
- O(n) in a full-mesh
- **Undef** for sparse
- O(1) for our proposal



Our proposal scales as sparse iBGP Topologies + provides a guaranteed convergence

Conclusion

- Contributions
  - Distributed Route Servers
    - New way to distribute the route selection
  - BGP decision process applied on a per router basis

- Scalable solution
- Solve issues of current sparse iBGP topologies

Thanks

- NTT Network Service Systems Laboratories
  - Cristel Pelsser worked on this during her Post Doctorate at NTT in Japan