

# design optimal iBGP route-reflection topologies

## IFIP Networking 2008

Marc-Olivier Buob – Orange Labs, LERIA

Steve Uhlig – Delft University of Technology

Mickaël Meulle – Orange Labs

*thanks to: Olivier Klopfenstein & Jean-Luc Lutton (Orange Labs)*

May 5-9th, 2008, Singapour



# agenda

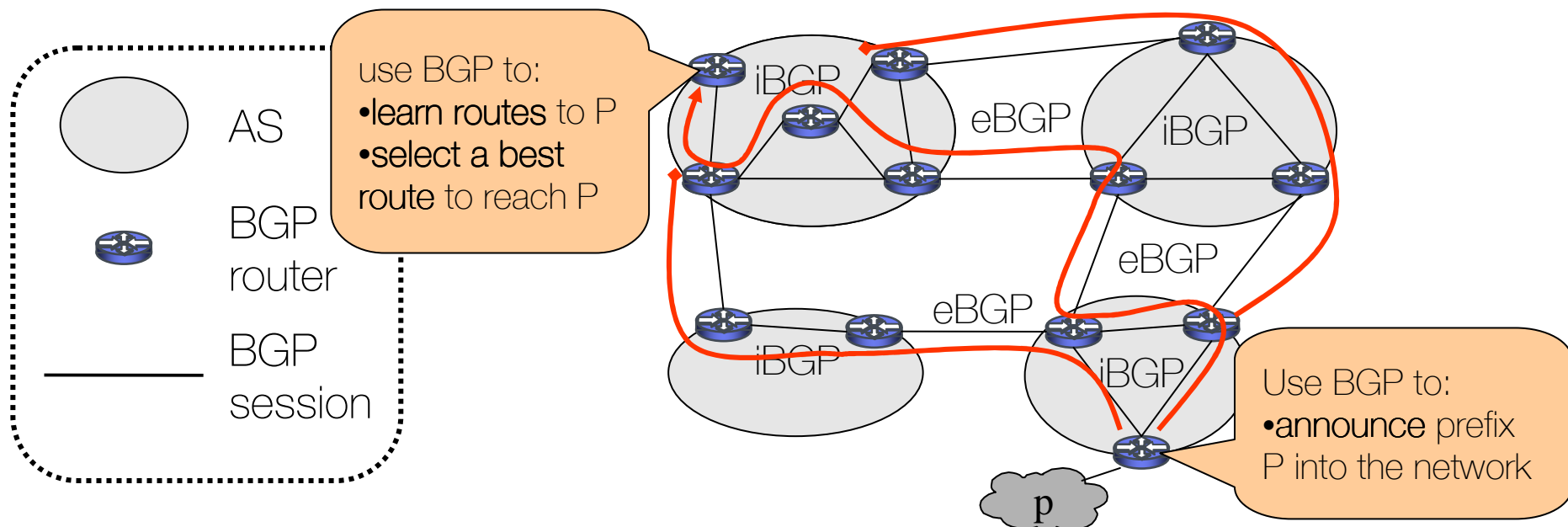
- part 1 background on iBGP routing
- part 2 the iBGP network design problem
- part 3 results : evaluation of our algorithm
- conclusion

## background on iBGP routing

Internet routing with the Border Gateway Protocol (BGP)  
iBGP route-reflection and its drawbacks

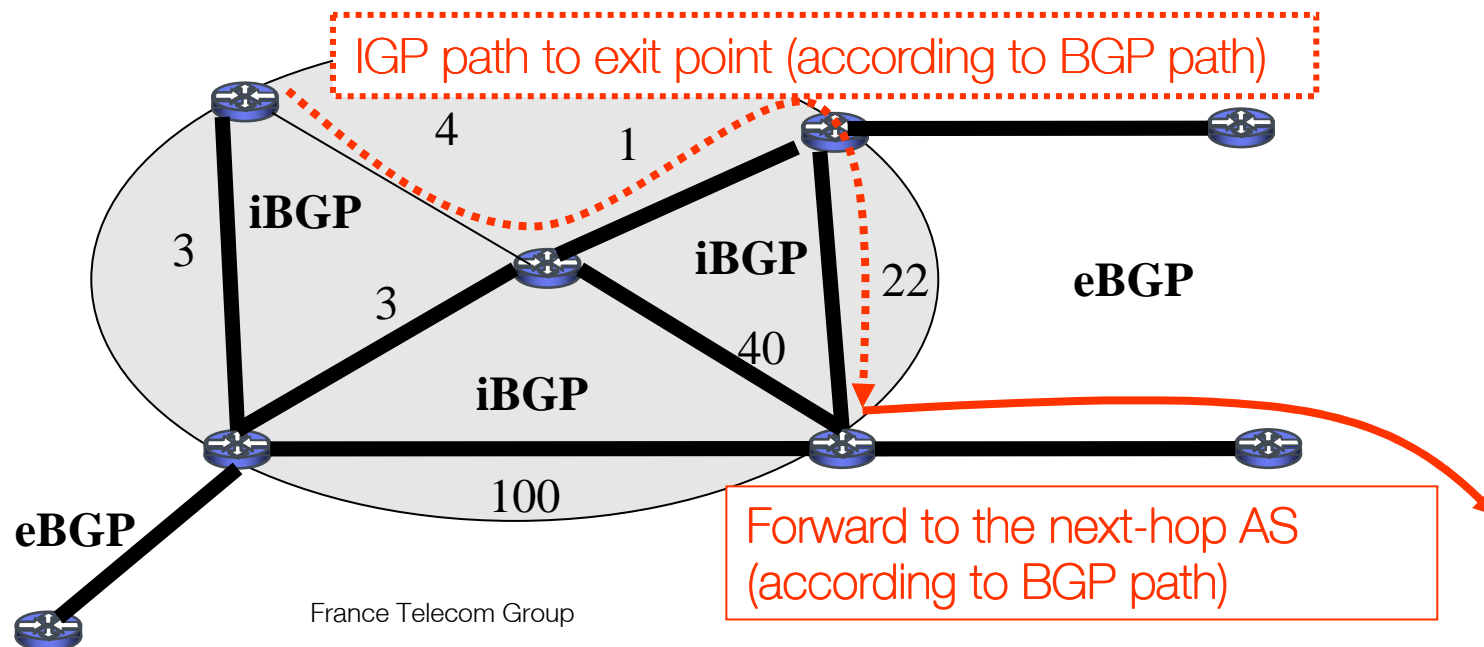
## Internet routing with BGP: basics

- the public IP address space is split into blocks: **prefixes**
- Internet connects together about 30k Autonomous Systems (AS)
- *"An AS is a connected group of one or more IP prefixes run by one or more network operators under a single and clearly defined routing policy."* (source: apnic.net)
- Internet routing : the Border Gateway Protocol (BGP)
  - exchange of reachability information about Internet prefixes between AS



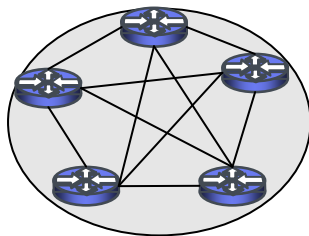
## Routing inside an AS

- Intra-AS routing is handled by an Interior Gateway Protocol (IGP)
  - usually shortest-path routing protocol: OSPF, IS-IS... (paths have IGP cost)
- Routing to internal prefixes and routers inside the AS (including exit points)
  - IGP routing
- Routing to external prefixes
  - BGP routing: the best route to a prefix indicates exit point
  - IGP routing: to the exit point

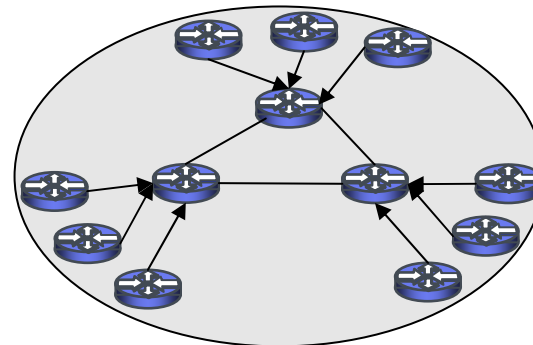


## iBGP routing: scaling with route reflection

- iBGP routers follow the rule:
  - “do not forward a route learned from iBGP to iBGP”
- iBGP routing in practise
  - in small AS: full mesh of iBGP sessions between routers
  - in large AS: needs BGP confederations or route-reflectors to **scale**
- Route reflectors (RR) in iBGP
  - can establish route-reflector-to-client sessions or classic iBGP sessions
    - routes learned from iBGP are propagated to client routers
    - routes learned from clients are propagated to all iBGP neighbours



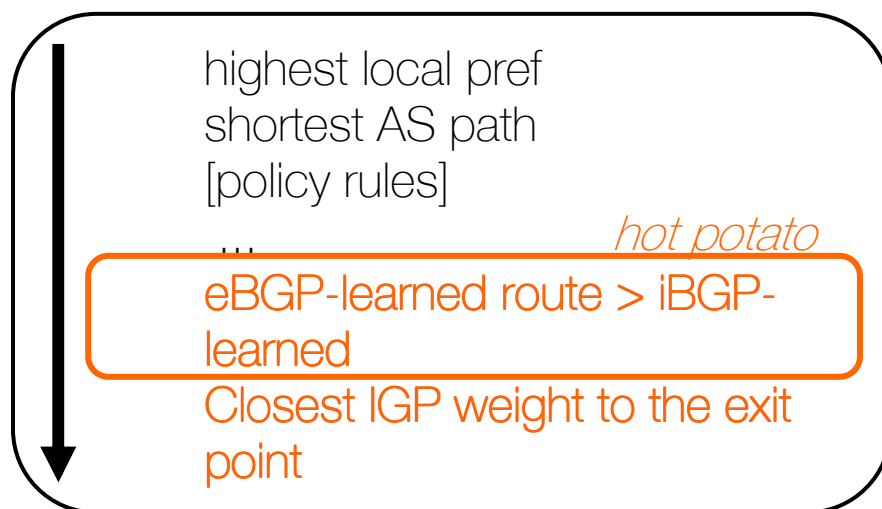
iBGP full mesh



hierarchy example with RR

# BGP routing inside an AS: route propagation and selection

- BGP best route selection: *BGP decision process*
  - for each prefix, execute a step-by-step tie break between concurrent routes



- In large AS, Hot potato occurs! (>70% of prefixes in our network of study)  
[tie breaking rules]  
→ routers make **local decisions** (IGP and iBGP position): deflection...

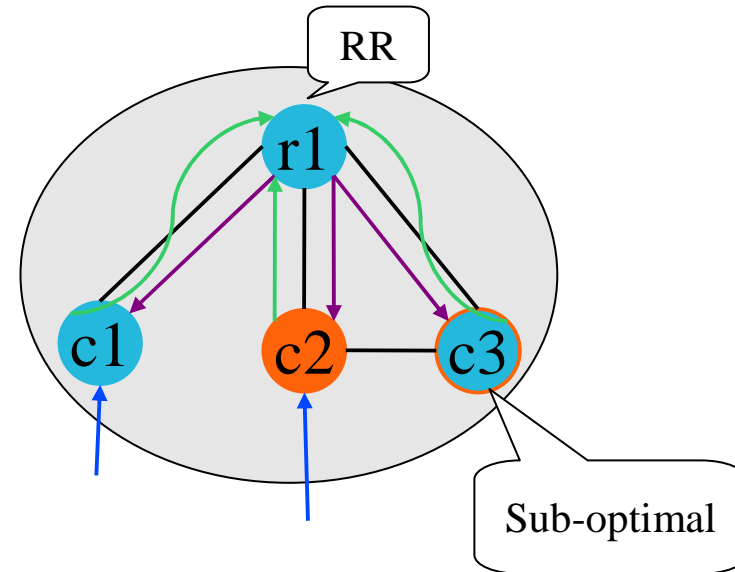
Loose of:  
• algebraic properties  
+  
• Information (diversity)  
=  
**Problems!**

- BGP route propagation inside a large AS: limited resulting diversity
  - for each prefix, a router **only propagates its best route** to neighbours
  - iBGP propagation inside the topology constrained by **iBGP forwarding rules**

# iBGP route-reflection and IGP topology: dangerous cocktail

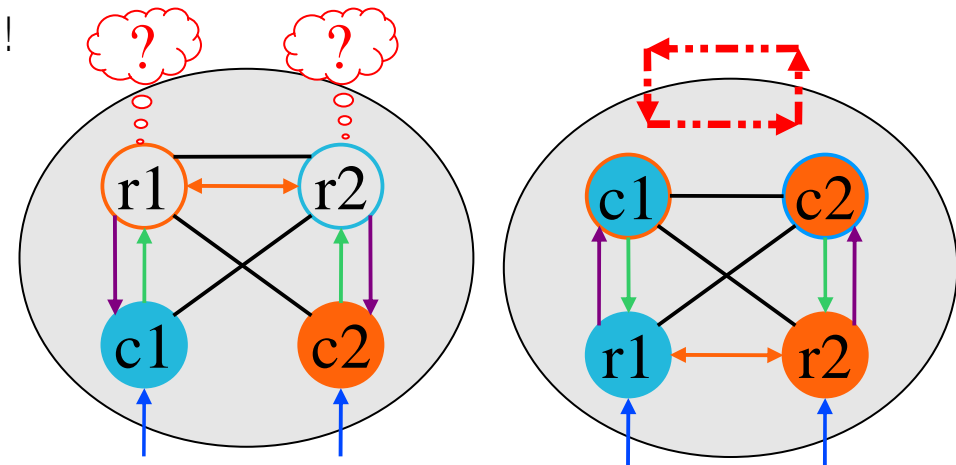
- Issues

- Routing Sub-optimality
- BGP routing oscillations
- Not-deterministic
- Deflection, forwarding loops



- Our solution

- The same routing as in a full-mesh !
  - modify iBGP
  - Correct iBGP topology 😊



# iBGP network design

Related work

Revisited formulation of the problem

Our solving approach: benders decomposition

## Related work

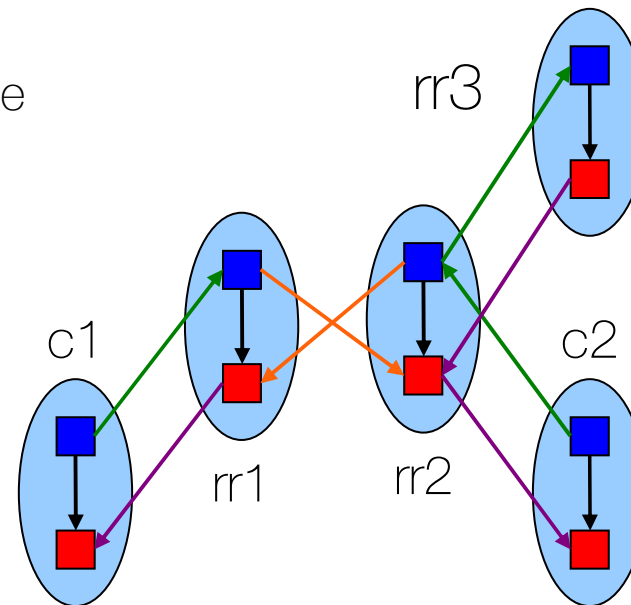
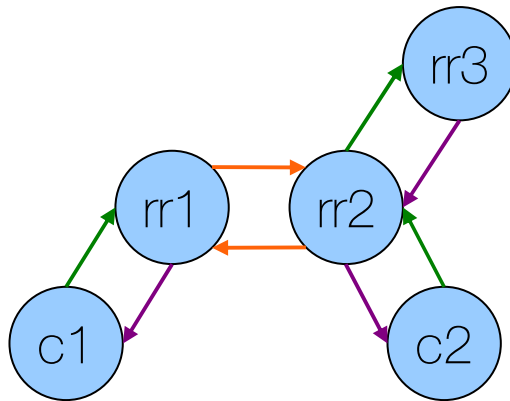
- Make RR more verbose (send all equally good BGP routes, up to IGP cost)
  - Route Oscillation with I BGP route reflection [Basu02]
- Make RR more intelligent (send customized routes for each client)
  - *draft-bonaventure-bgp-route-reflectors-00* [Bonaventure00]
  - Design and implementation of a Routing Control Platform [Caesar03]
- Apply a multicast protocol to distribute messages
  - BST—BGP Scalable Transport [Poduri03]
- Design a good hierarchy of Route reflectors
  - Reliability-aware IBGP Route Reflection Topology Design [Xiao03]
  - Optimizing IBGP Route Reflection Network [Xiao03]
  - Verifying the Correctness of WideArea Internet Routing [Feamster04]
  - How to Construct a Correct and Scalable iBGP Configuration [Balakrishnan06]

## iBGP network design problem

- Inputs
  - IGP topology  $V_{igp}$
  - set of BGP routers "targets" ( $R \subseteq V_{igp}$ )
  - set of border routers "sources" ( $N \subseteq R$ )
- Variables/Output
  - iBGP topology
- Constraints
  - Fm-optimal routing i.e. as in a full mesh topology **with any set of concurrent border routers**  
[ cf. Buob et al., "Checking for optimal egress points in a large AS, DRCN'07 ]
    - implies a loop free and deterministic routing
  - Fm-optimal routing even in case of a link failure
    - implies a loop free and deterministic routing even if a link fails
- Objective
  - iBGP topology should match as much as possible IGP topology
  - Less possible sessions

## How to handle fm-optimality constraints?

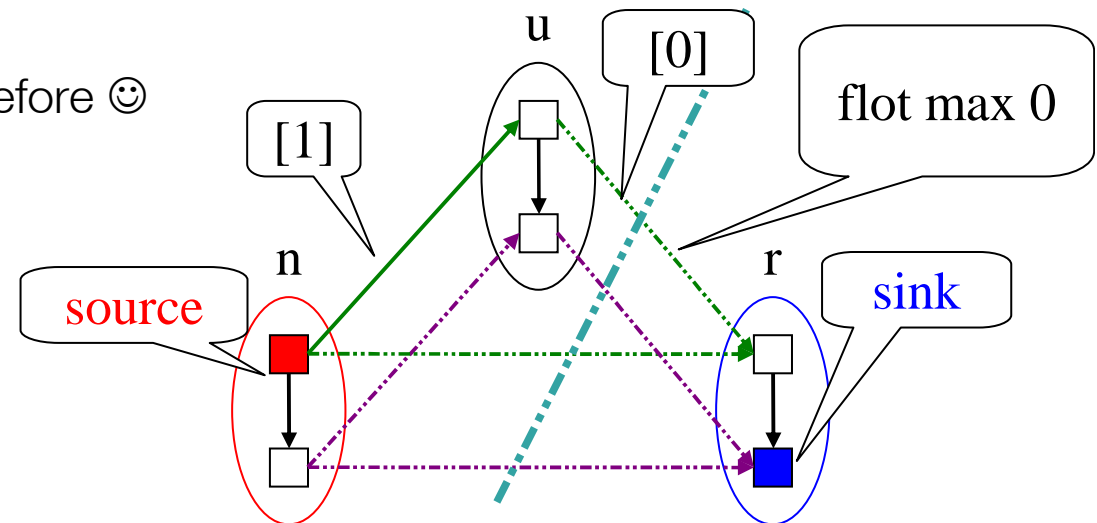
- Give a sufficient condition to guaranty that a router  $r$  (in  $R$ ) may be able to receive the route exiting at border router  $n$  (in  $N$ ) when route entering at  $n$  is the best among all possible
  - [CFIP'07] a router  $w$  is **white** for  $(n,r)$  if it never blocks route propagation from  $n$  to  $r$
  - *If* a valid iBGP path composed of white routers exists between  $n$  and  $r$   
*THEN*  $r$  will learn its fm-optimal route from  $n$
  - condition works for any set of concurrent border router ☺ (only IGP weights needed)
- Use a graph transformation to easily look for valid and white iBGP paths
  - computing a valid iBGP path in a topology  
*is equivalent to* computing a normal path in the corresponding extended graph[CFIP'06]



## Solving the problem with Benders decomposition: divide to conquer

- Satellite problem for each  $(n,r)$  in  $(N,R)$ 
  - It looks for a white iBGP path in the extended graph for a  $(n,r)$  pair
  - A Flow problem is solved and outputs a new constraint if no such path exists
    - Max-flow Min cut, source  $n$ , sink  $r$
- Satellite problem for each  $(n,r)$  in  $(N,R)$  and each IGP link failure
  - Restrict iBGP sessions between routers in the same IGP connected component
  - Solve the same satellite as before 😊

- Robust fm-optimality  
= all satellite problems  
simultaneously satisfied



## Solving the problem with Benders decomposition : Algorithm

Do :

Solve master problem (Integer Linear Program)

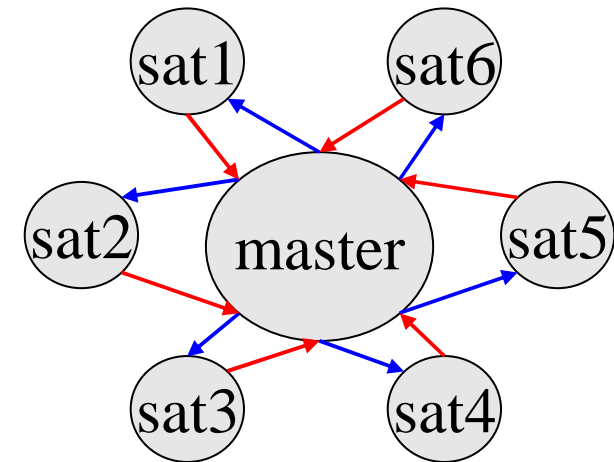
inject solution found into satellites

Interrogate satellites until 1 or more unsatisfied

- Each unsatisfied satellite add a new constraint to the master problem

While at least one satellite unsatisfied

Return optimal solution



- Objective function of the master problem gives incentive for:
  - the iBGP topology to follow IGP topology
  - Minimizing of number of sessions

## Some issues

- Large Mixed-Integer Linear Program
  - Example with  $|N| = |R| = 100$ ,  $|E| = 500$
  - 10 k boolean variables
  - 5 M satellite problems
- Tweaks needed to better scale the approach:
  - Restrict types of BGP sessions to RR-to-client only because of degenerate solutions
  - Reduce number of installable sessions (without losing global satisfaction)
  - Aggregate redundant satellite problems (presolve)
  - Reduce the number of constraints thanks to Benders approach

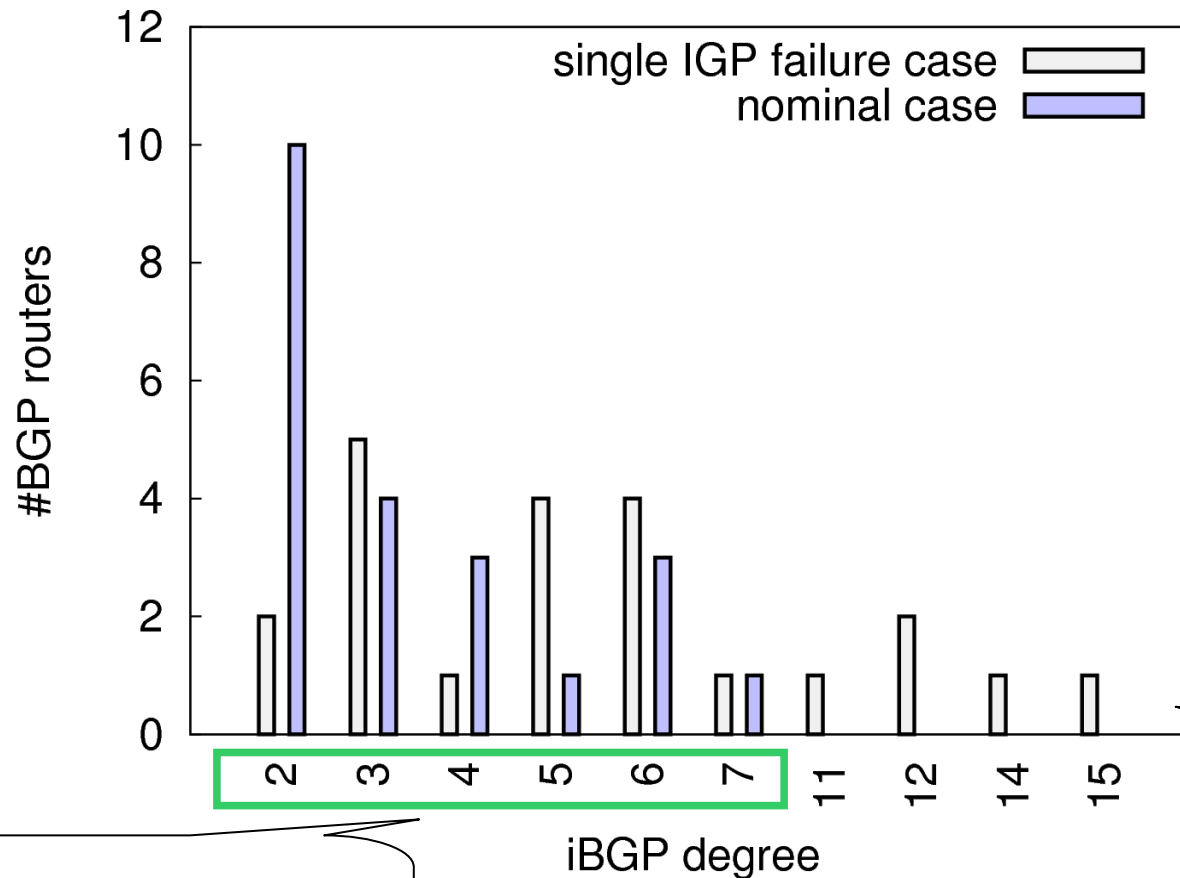
## results

Random topologies & GEANT  
OpenTransit Network

## Overview of topologies

- Realistic topologies generated using iGEN
  - <http://www.info.ucl.ac.be/~bqu/igen/>
  - 25 NODES, 2 distributions (W,NA)
  - 2 network design heuristics (Delaunay triangulation and Two Trees)
- GEANT TOPOLOGY
  - 22 nodes
- Similar results for each instance
  - Solving time:
    - with failures:  $25s < < 120s$
    - Without failures:  $< 10s$
  - We focus on GEANT

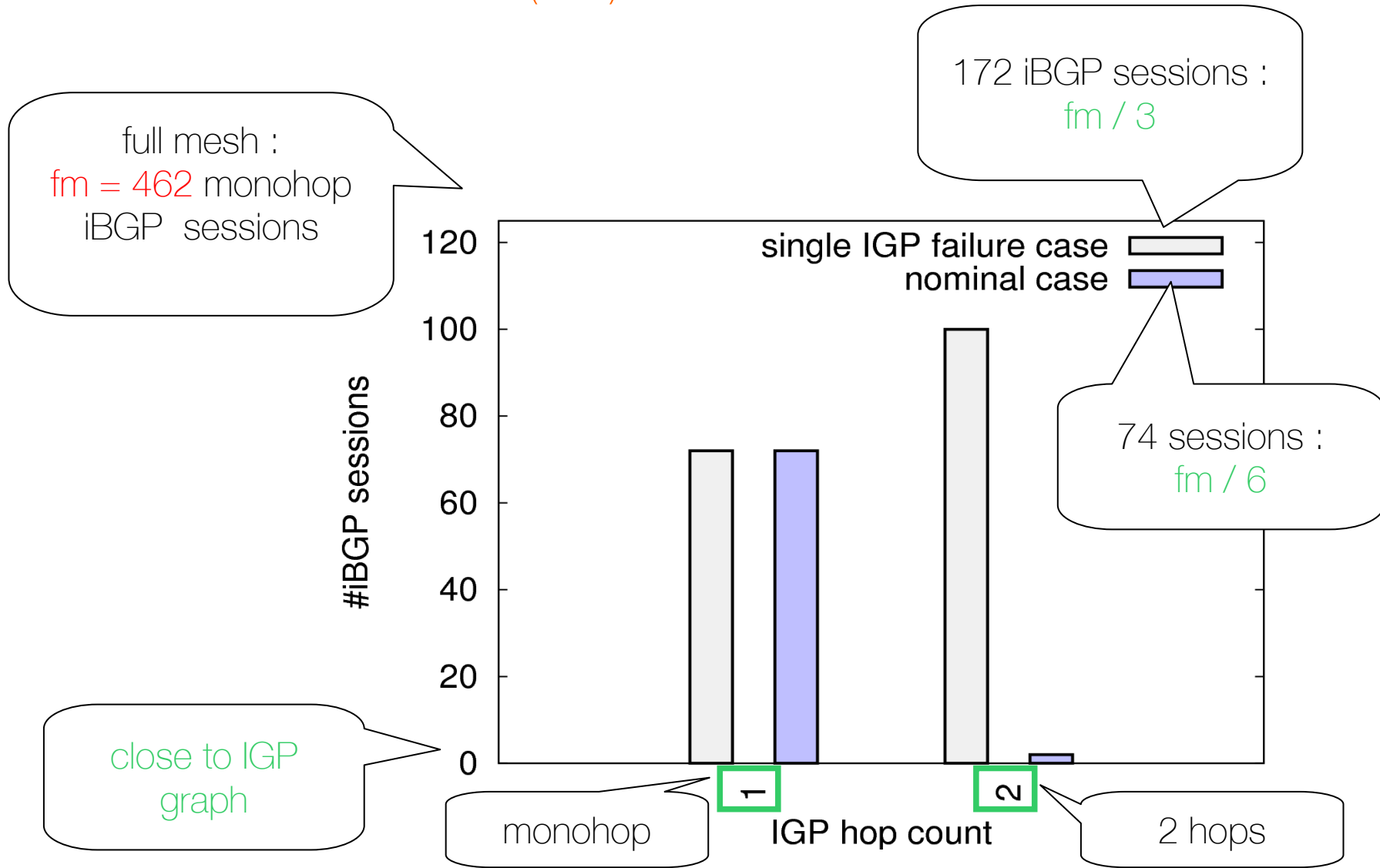
## illustration : GEANT (1/3)



few iBGP sessions  
per router

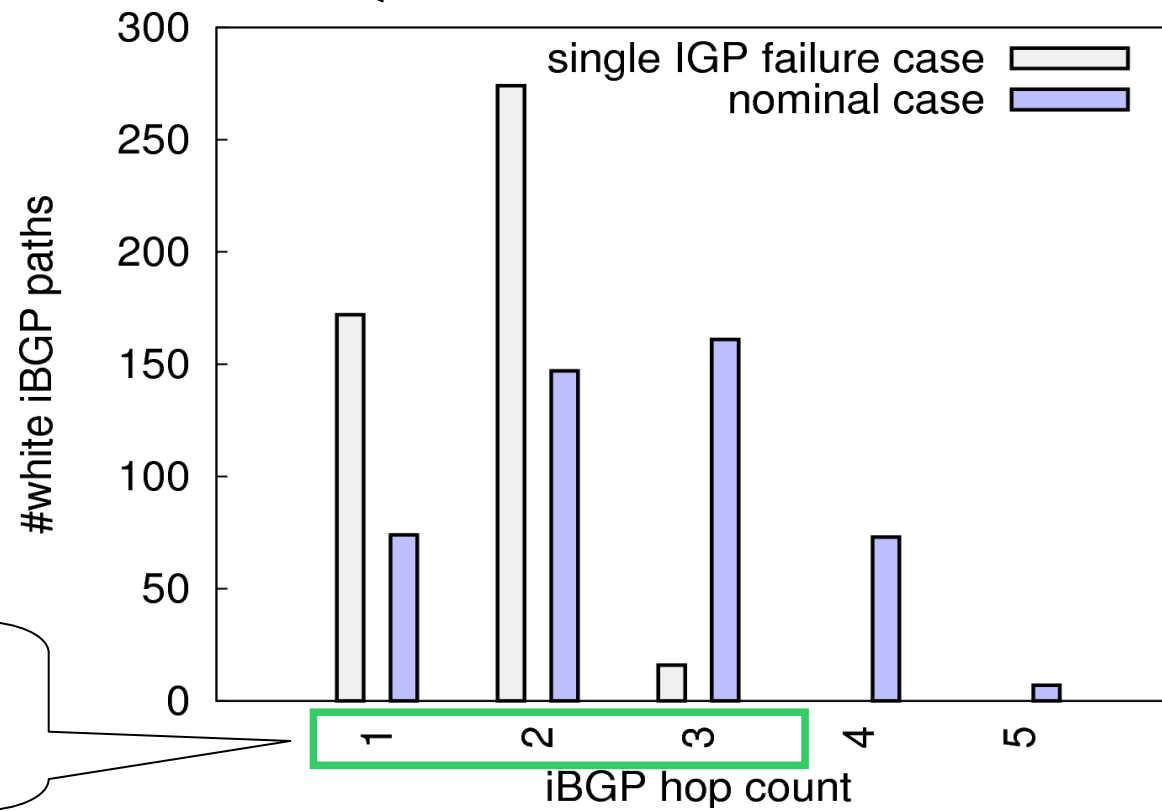
full mesh :  
21 iBGP  
peers

# illustration : GEANT (2/3)



## illustration : GEANT (3/3)

Out of 462 o-d paris



Small iBGP paths

## Open Transit (AS5511) 2005 and 2008

- Computational time
  - 2005 : without failure : 3d ; with failure : 6d
  - 2008 : without failure : 15min ; with failure : 8h
- characteristics of the solutions
  - Not hierarchical, significantly different from real topology
  - BUT realistic compared to the real topology:
    - iBGP paths with similar length (convergence)
    - Approx. about the same number of iBGP sessions in the robust case, 2 times less otherwise
  - Very few multi-hops sessions
  - robust

## Conclusion

- iBGP problems come from a tricky combination
  - iBGP pattern + best route DP + local choice of routers
- iBGP topology design
  - Fm-optimality guaranty : same routing as in a full mesh
    - optimality, no deflection, deterministic...
  - robust approach (the only one ?)
  - Scalable to real large network topologies, realistic solutions
  - ... What happened if the IGP graph changed ?
- Future work : a new iBGP...

Thanks!

