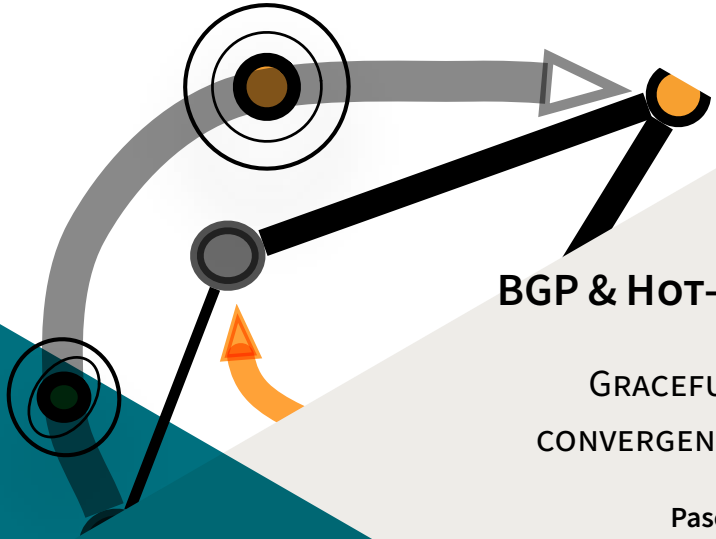
Université de Strasbourg



BGP & Hot-Potato Routing

Graceful and optimal convergence for IGP events

Pascal Mérindol

30/6/2023

A brief presentation: who is this guy?

My main Research Contributions

Université de Strasbourg

Two Main Research Topics to Improve the Internet

Both about Routing but from two distinct perspectives

1. **Routing Algorithms & Protocols**: **Compute and Deploy Valuable IP Routes**
   - Fast and correct convergence for several change models: link and node failures, maintenances or BGP hot potato
   - Multi-metric and multi-paths for traffic engineering
2. **IP Measurements** in Transit Networks: **Topology Discovery & Large Scale Monitoring**
   - Reveal Hidden MPLS Tunnels and Forwarding Detours
   - Troubleshoot ISP (Internet Service Provider) Networks

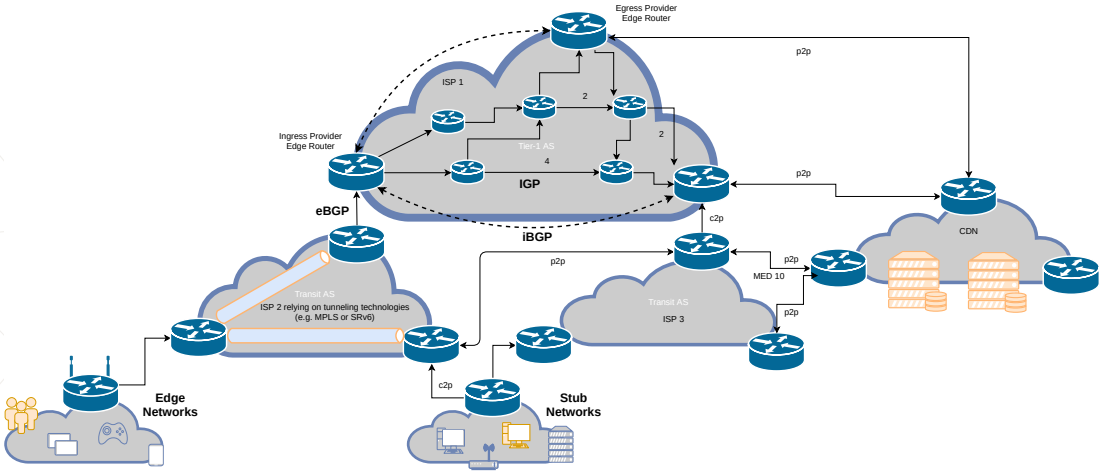
Context

Internet: the Worldwide Distributed System

- **Large scale**: soon $100,000$ Domains (including $\approx [100 \ldots 10,000]$ routers), $1M$ IP prefixes and much more end-devices than human beings!
- **Complex**: numerous heterogeneous hardware and software components
- **Innovation is difficult and challenging**:
  - The universal common technology for connecting all devices lies in the **IP data-plane**
  - The **IP control-plane (i.e. the routing)** provides means to install and manage forwarding routes, e.g. with MPLS or Segment Routing (SR) w.r.t. the type of service

The Big Picture

Overall, a complex ecosystem!



Two Scales and Several Technologies

Two scales of routing operations:

- inter-domain or AS-wide level: BGP
- intra-domain or router level within an AS: IGP

Two Scales and Several Technologies

Two scales of routing operations:

- inter-domain or AS-wide level: **BGP**
- intra-domain or router level within an AS: **IGP**

Several technologies and paradigms:

- **Hop-by-hop** packet forwarding (for best-effort traffic), or **loose source routing**, e.g. with **SR or MPLS tunnels**
- With packet encapsulation, one can enable **Traffic Engineering (TE)** for premium flows, or deploy **Load Balancing (LB) and Fast-ReRoute** (FRR) for all services.
  A single deviation from the best-effort usual IP forwarding scheme is often enough

Background: Internet Routing Protocols

Two respective paradigms for inter- and intra-domain routing:

- with BGP, preferences are local and selfish: the path vector protocol may diverge
  - sufficient conditions for convergence a priori met in practice (*valley free* routes and usual economical incentives)
- within IGP link-state routing protocols (e.g. OSPF, IS-IS), the objective is globally consistent, typically a *minimum function* applied on an *additive metric*: $(min, +)$
  - convergence is granted (thanks to the *isotonicity* & *monotony* of routing operations), but **anomalies, e.g. forwarding loops and path sub-optimality,** still occur during transient periods of change

The IGP and BGP are entangled

Hot and Cold Potato Routing: MED & IGP Distance

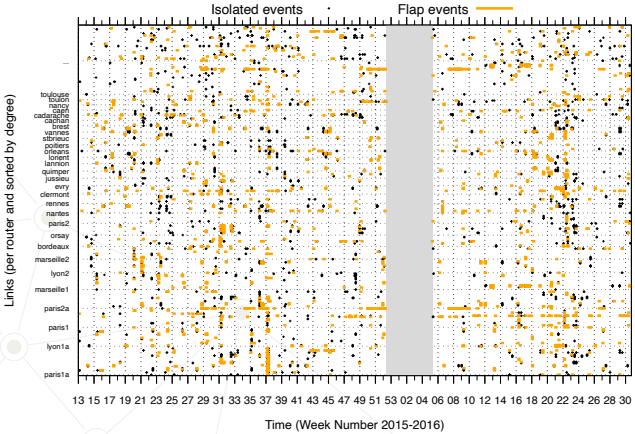
Step	Criterion	
1	Prefer highest `local-pref`	(economical relationships)
2	Prefer shortest `as-path`	(global optimality)
3	Prefer lowest `origin`	
4	Prefer lowest `med`	cold potato routing
5	Prefer routes learned over `ebgp`	
6	Prefer lowest `igp` cost	hot potato routing!
7	Prefer lowest `router-id`	(arbitrary tie-break)

Let's recall some IGP problems first...

Routing Changes: Illustration in RENATER

Université de Strasbourg

The **DCART** project deployed in the French Educational & Research Network

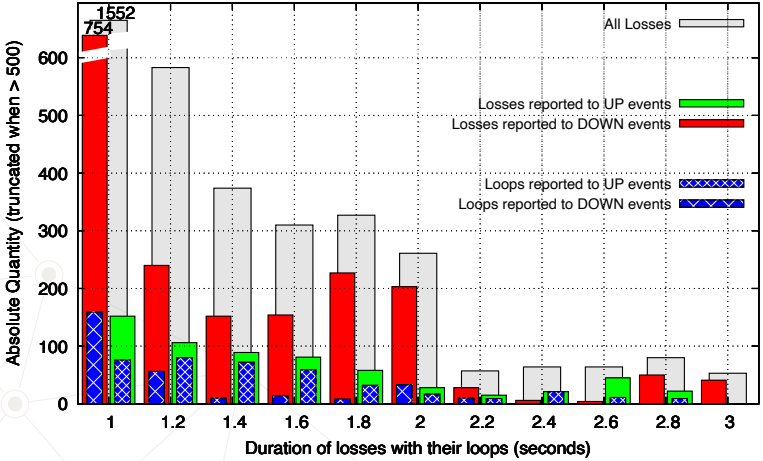


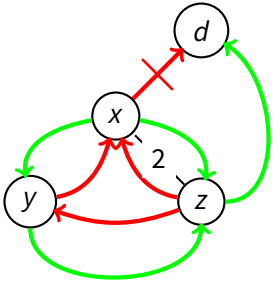
RENATER

GÉANT
Networks • Services • People

Topological changes are common and links are often **flapping**

Routing Changes: Illustration in RENATER



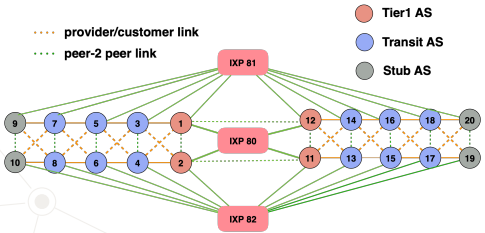


A not so simple ECMP forwarding loop

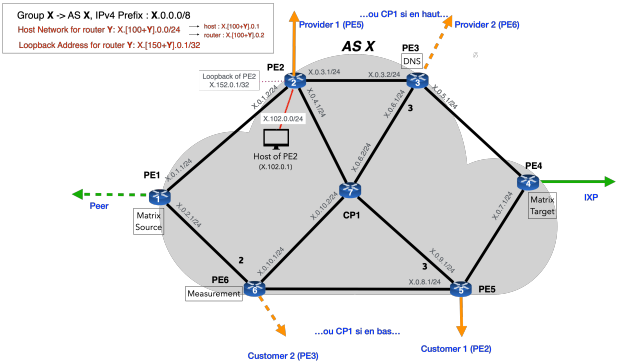
Forwarding loops do occur, in particular when routers reboot!

BGP is not just external connections

Peering with internal neighbors is also necessary!



eBGP connections



iBGP + IGP redistribution

Scalability issues: replace full mesh with route reflections

Convergence of Routing Systems

Fast-ReRouting

IGP wide: From LFA to TI-LFA

To **avoid inconsistent transient forwarding states to occur**, backup paths should be **loopfree**: during the convergence, *packets must be safely released at a node where the failed link is not used even before its failure*

Theorem (one segment is enough with symmetric weights)

In symmetric networks (w is a symmetric valuation function), post-convergence paths require only one intermediary detour to be safely deployed.

$\Rightarrow$ *TI-LFA benefits from this property to deploy post-convergence paths* ***using one node or one adjacency segment at worst****. At best, one local LFA is enough.*

TI-LFA

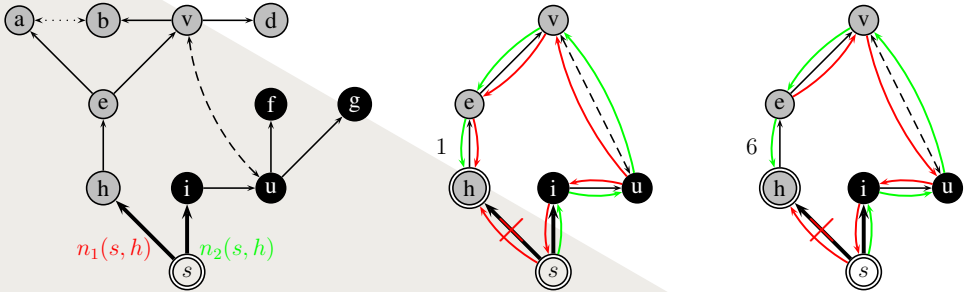
And its most efficient underlying algorithm (TBFH)

Definition (post-convergence paths)

A post-convergence path $p_2(s, d)$ is an optimal path w.r.t the failure of link $(s, n_1(s, d))$.

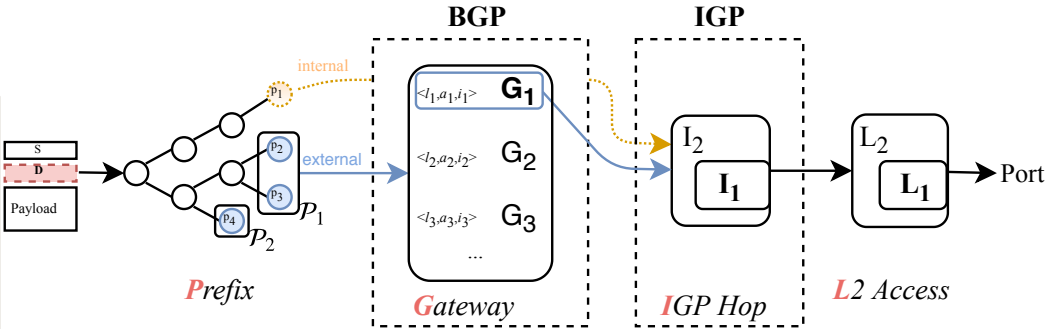
Theorem (post-convergence paths have the same cost as 1-alternate paths)

TBFH returns all $p_1(s, d)$ and $p_2(s, d)$ for a given $s \in V$ and $\forall d \in V$ in only 2 SPC runs.



Hierarchical Forwarding

A flexible data-plane



From the LPM lookup up to L2 interface!

Flat forwarding engines are much less efficient to manage...

Prefix Independent Convergence (PIC)

Provide and exploit backup options

PIC handles two main cases:

1. **PIC-core**: an internal event occurs, so just relying on the hierarchical FIB is enough...
2. **PIC-edge**: the failure of a border router occurs, so what? provision a secondary backup gateway!

Limitations of PIC (simple but naive):

- Not optimal for core: PIC assumes the gateway remains the same after the IGP event
- PIC-edge does not provide enough gateways: **ensuring optimality for any IGP event require more gateway diversity**
- And overall, not enough resiliency: what if bi-connectivity towards a given gateway is not ensured?

BGP Add-Path

Increase the internal visibility (iBGP diversity)

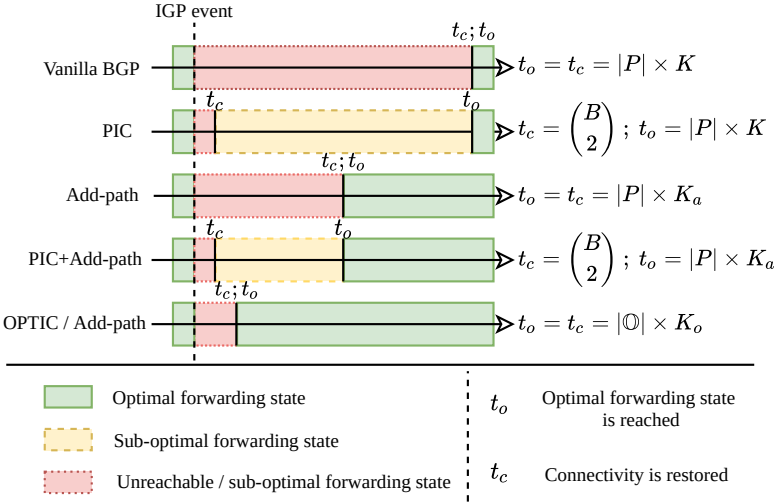
The goal of Add-Path is not really about enabling BGP multi-path, but it is rather about:

- **mitigating the churn**: avoid superfluous iBGP signalisation (most becomes unnecessary with more diversity)
- **mitigating internal MED oscillations**: more aware route reflectors!
- and even more...

Indeed, for the second item, some options of Add-Path, like the most extreme one (Add-Path `All`), enables to fix the absence of total order due to the MED.
However, it **does not scales well** on the contrary of other options like `2 AS wide` (but that comes with no guarantee).

Towards a New Model?

Nor PIC neither Add-Path are sufficient for ideal convergence!



OPTIC, THE IGP SHIELD FOR BGP

Optimal Protection Technique for Inter-intra domain Convergence (OPTIC)

Université de Strasbourg

Goal

Fast optimal **BGP convergence for IGP events**

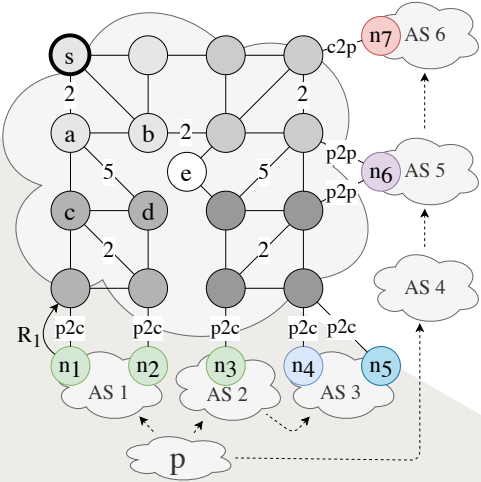
Challenge and Originality

OPTIC solves the **hot-cold potato problem** while existing schemes only mitigate it; with enough *iBGP visibility* for *stable deterministic MED* & *IGP diversity* for *hot-potato routing*, define **sufficient conditions to optimally protect** each IP prefix for any IGP event

The Algorithm in a Nutshell

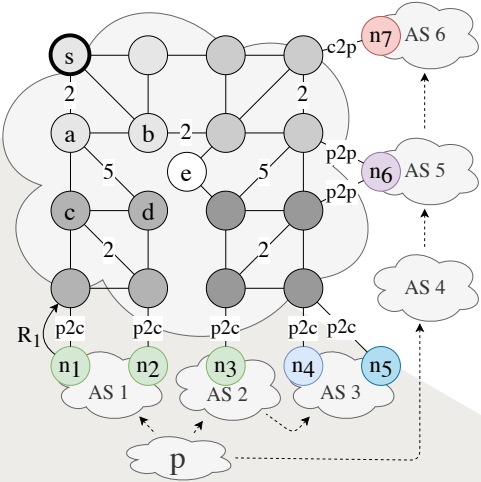
- Stack enough MED-aware Rounded (MR) sets to ensure k-node connectivity ($k > 1$) $\longrightarrow$ Optimal Protecting Rounded (OPR) set

Problem Statement: Protect the Potato



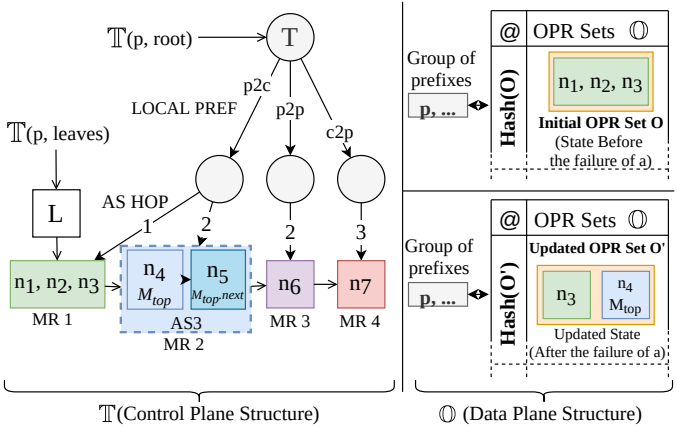
- **Optimally protect the traffic from s, the ingress PER, towards external prefix p:**
  - considering any single internal failure
  - as well as the failure of the egress PER
- **Which routes s needs to protect p (OPR)?**
  - obviously, the best route R_1: via n_1 at AS 1
  - but also *all* the routes that *may turn to be* the best after *any* single event: via n_2 and n_3

Problem Statement: Protect the Potato

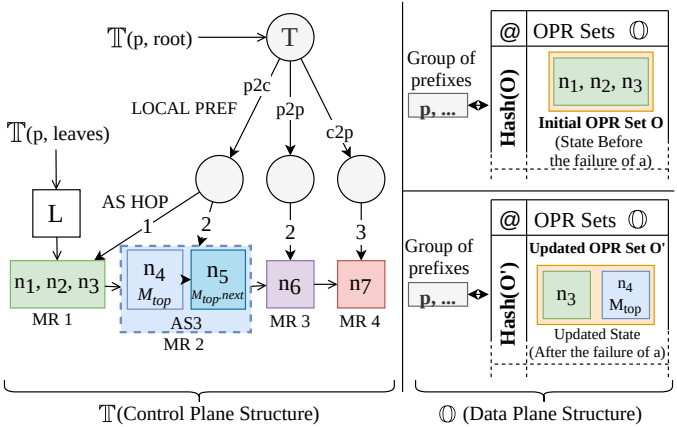


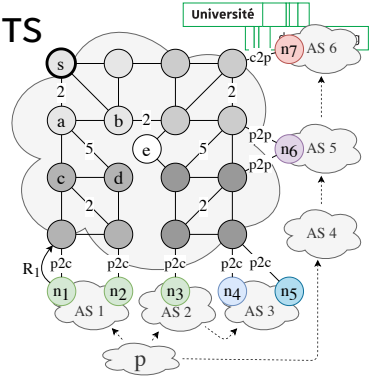
- **Optimally protect the traffic from s, the ingress PER, towards external prefix p:**
  - considering any single internal failure
  - as well as the failure of the egress PER
- **Which routes s needs to protect p (OPR)?**
  - obviously, the best route R_1: via n_1 at AS 1
  - but also *all* the routes that *may turn to be* the best after *any* single event: via n_2 and n_3
- **Current deployed approaches (e.g., PIC+Add-Path) neglect several aspects**
  - no explicit check for k-node disjoint paths
  - intermediary transient sub-optimal routes
  - insufficient diversity for MED re-convergence

The Data-Structures: MR and OPR sets



The Data-Structures: MR and OPR sets





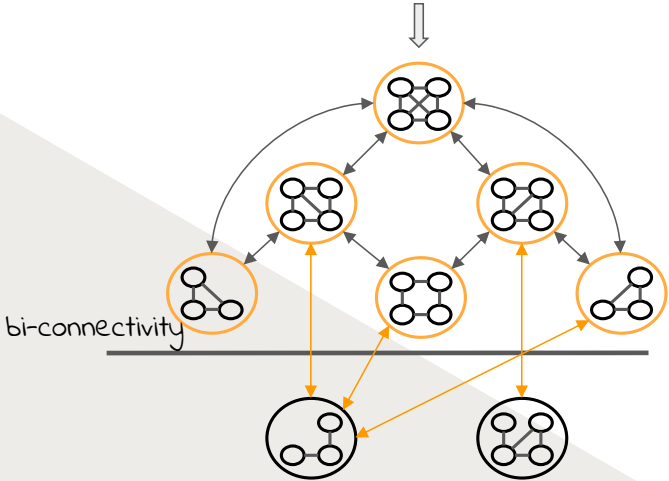
Routes are sorted in a tree

Typically 1 or 2 MR for 1 OPR set

The OPR is updated only if necessary

Stability of the OPR sets

OPTIC has at least one step ahead



Network states

Events

Events requiring set updates

Analytical Results

OPTIC in Math

$$|\mathbb{O}_{B,P,ps}| = \sum_{n=2}^{b} \binom{B}{n} P_{B,P,ps,n} + B\binom{B-1}{n-1} P'_{B,P,ps,n} \tag{1}$$

$$P_{B,ps,P,n} = 1 - \left(1 - \binom{B}{n}^{-1}\right)^{p_n P} \quad , \quad P'_{B,ps,P,n} = 1 - \left(1 - \left(B\binom{B-1}{n-1}\right)^{-1}\right)^{p'_n P} \tag{2}$$

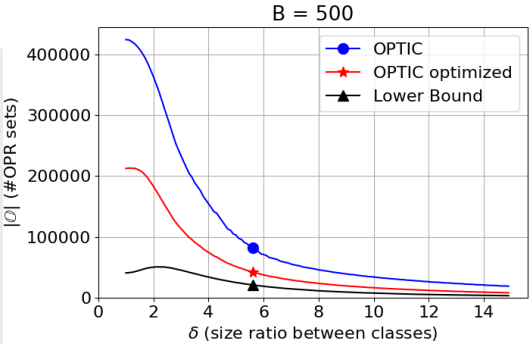
$$p_n = \sum_{i=1}^{ps} \binom{b}{n} \frac{1}{ps^n} \left(1 - \frac{i}{ps}\right)^{b-n} \quad , \quad p'_n = \sum_{i=1}^{ps} b\binom{b-1}{n-1} \frac{1}{ps^{n-1}} \frac{i-1}{ps} \left(1 - \frac{i}{ps}\right)^{b-n} \tag{3}$$

Type of AS	# gateways per class	# prefix per class	# distinct OPR sets	OPR sets median size	Lower bound
Stub	(10; 20; 0)	(700K; 100K; 0K)	3945	4	235
Tier 3	(10; 50; 100)	(500K; 200K; 100K)	46 010	3	6219
Tier 2	(5; 500; 2000)	(500K; 200K; 100K)	263 219	2	197 194
Tier 1	(0; 50; 5000)	(0K; 600K; 200K)	232 180	2	199 633

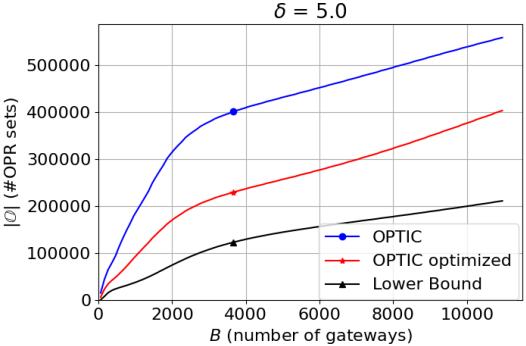
Performance Evaluation

The number of groups (# groups) depends on B (# gateways) and...r?

The class breakdown (r): this ratio determines the # prefixes reported to B in each class!

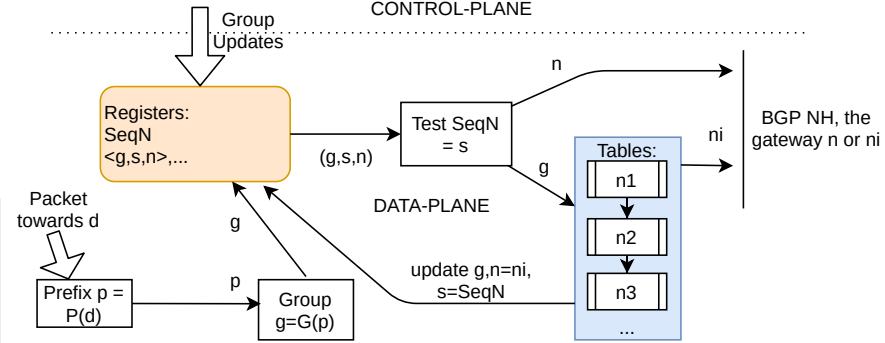


groups / r ($B = 500$)



groups / B ($r = 5$)

Data-plane updates: Illustration with OPTIC



With a programmable switch, updates can be conccurently performed both at the control and data-planes.The number of groups or prefixes is not anymore a processing limitation (but only a spatial one)!

How to efficiently retrieve the new optimal BGP NH after an IGP change?

Ongoing Steps

Université de Strasbourg

- Develop a prototype in P4 for our Tofino hardwares and combine it with FRR at the control-plane...
- Evaluate our proposals with respect to several competitors within open simulation/emulation frameworks (e.g., with NS3/BIRD or FRR)
- Finally, design a **comprehensive fast-reroute framework** enabling a combination with OPTIC (e.g., based on SR for IGP) and BGP wide techniques

Conclusion & Perspectives

Summary

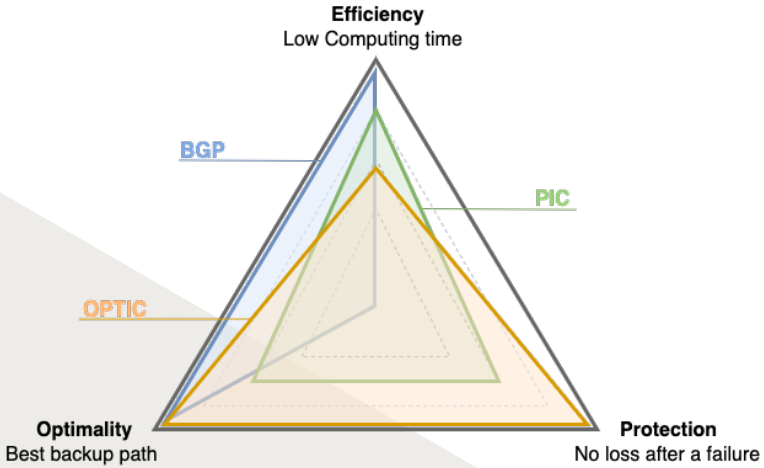
Université de Strasbourg

OPTIC enables a fast convergence of BGP in case of IGP events. In addition, it is:

- efficient, i.e. almost no overhead (w.r.t. vanilla BGP)
- truly optimal, w.r.t. PIC, it avoids transient sub-optimal states
- more robust than usual approaches: prevent most MED oscillation problems and corner cases with no enough network resiliency!

A Radar Chart...

...to rule them all!



Future works

Complete ongoing steps and go even further...

- Move (most of) OPTIC from the control-plane to the data-plane;
- Evaluate it with more realistic (and likely more favorable) scenarios;
- Overall, build a FRR+P4 setup for a comprehensive fast-reroute prototype!
- Go further looking at **protection for Traffic Engineering routes** within the same context (IGP events for transit traffic)

Acknowledgements

Here is the list of people involved in this work:

- Jean-Romain Luttringer (the still ex-best PhD student)
- Quentin Bramas (a great folk)
- Cristel Pelsser (a PhD advisor)
- Thomas Alfroy (the great internship)

And the two papers mentioned in this talk:

- **INFOCOM 2021**: *A Fast-Convergence Routing of the Hot Potato*
- **CORES 2022** (in french): *L'Art d'Anticiper les Changements IGP pour Acheminer Optimalement la Patate en Transit*