Action Concertée Incitative

SÉCURITÉ & INFORMATIQUE

Scientific Description of the project

Internet Inter-domain Routing Security

1 Goal and context

The main issue of the Internet routing protocols is to provide worldwide connectivity. The Internet connects thousands of Autonomous Systems (AS) operated by various institutions such as Internet service providers, universities, companies or organizations. Within an AS, routing is controlled by intra-domain protocols such as OSPF or RIP. The Autonomous Systems are connected by dedicated links or public network access points. They exchange information about routes using the Border Gateway Protocol (BGP). BGP is an inter-domain routing protocol that allows AS to apply local policies for the route selection and the propagation of control information. Such a hierarchical routing follows the hierarchical and distributed nature of the Internet. But, the routing protocols have not been designed to manage the large number of hosts we now have in Internet. The network is fault tolerant as most of its functions are distributed. However, new types of faults like worm propagation, distributed denial of service and attacks of the DNS servers have shown that the network is much less reliable and secure that we may have thought. Furthermore, the lack of guarantee prohibits most of the value added services.

Clearly, a hierarchy of routing protocols is necessary because Internet is a collection of smaller networks with their own rules. And the inter AS routing protocol has a large impact on delays experienced by packets, due to the routes selected. The main property for a routing algorithm on a large network is to deal with the dynamics of the topology, the available links and the connected routers. As the whole Internet is dynamic, the routing tables must change frequently. BGP routers must inform their neighbors when paths change due to a failure or a correction. And they must propagate this information when they receive it. Thus the routing protocols have to manage these updates efficiently. Many authors (for instance, Govindan [1], Labovitz [2], [3]) have reported BGP oscillations for routes. Labovitz et al. report in [2] that the number of exchange messages is several orders of magnitude larger than the number of real faults. These oscillations imply a considerable loss of bandwidth but also larger delays. Furthermore, it is still possible to have an infinite loop, and BGP does not converge if the AS use general rules of routing (see Govindan et al. [4]). Of course the TTL, timers and retransmit techniques will insure that packets never stay too long in the network and that they will be sent again. We advocate that this bandwidth lost will have been better used by a more efficient routing protocol.

Thus, BGP suffers typically from a *reliability* problem since an elementary breakdown of router causes serious transient problems. But BGP and the routers also exhibit *safety* problems which threaten Internet connectivity. The propagation of worms using random IP addresses recently caused important breakdowns of BGP routers [16], [17]. Indeed the random IP address will more likely create a cache miss (due to the large size of the BGP table) and the router is stressed by a large number of cache misses and crashes.

BGP thus suffers from multiple problems of safety: first, reliability by its reactions to links or routers breakdowns, reliability also when inconsistent local policies obstruct the convergence of the routing algorithm [4] or causes an overload of the network because of the flooding by obsolete update messages, and also safety when attacks on other components of the network stress the router and causes a breakdown. Note that the IP spoofing technique may also be used to provide wrong information to a BGP router as the routes are advertised by TCP packets. We will not address this spoofing problem. Classical cryptographic techniques have already been proposed to make secure the exchange of informations (for instance, some Cisco routers use a signature based on MD5).

The BGP instability is a well-known subject among the networking and distributed algorithms teams. Several papers on this topic have been published in the most important networking conferences (Sigcomm and Infocom) and journals. Most of the researches on this topic are carried in the US universities. Some routers manufacturers have also tried to correct this instability problem. For instance Cisco has developed the "route dampening" to minimize the instability caused by route flapping and oscillation over the network. Unfortunately, it has been proved that route dampening creates more complex oscillations.

The purpose of the project is to improve the reliability of the routing infrastructure. To achieve this goal, we will study how to improve the routing and flooding algorithms, design correct local policies and perform statistical controls to detect traffic variations. The tools used in this analysis will include: algorithmic (worst case) analysis, and probabilistic (average case) studies through simulations and analytical approaches (stochastic processes, queuing theory). The contribution of all the teams would span several subtopics of the project as the difficulty of the problem requires a collaboration to tackle the various aspects of BGP security. The focus of the analysis will be

- the theoretical convergence properties: does the network stabilize in response to a change in link states (failure, administrative updates, flow changes);
- the study of the time needed to obtain this stabilization;
- the amplitude of traffic oscillations observed before convergence occurs, if any.
- the monitoring of the network which may help to understand the real topology and load. Active and passive tomography will recover more information about the network. These techniques can also help to limit the propagation of worms and maybe the Distributed Denial of Service.

Starting with the standard BGP, we want to ask what combinations of network information and router behavior can improve the stability and robustness of BGP. The underlying idea is the construction of a "more realistic" image of the networks and the routes to limit the exchanges of messages, accelerate convergence and improve BGP and routers robustness.

2 Project description

First we present the fundamental questions about inter-domain security we want to address, then we develop the project, and present briefly the teams involved in the project.

2.1 The fundamental questions we have to address

Most of the work is related to the algorithmic aspects of BGP. But the whole project is much larger. We also need some techniques about traffic and topology monitoring to help the routing algorithm and the routers. The main idea is first to give enough information to the routers to know:

- Where is the failure ?

  Typically BGP messages withdraw a route but do not explain why. Giving more information about the link or router failure help the router to chose a good path. Indeed, as the complete path is in the table, it is possible to check that a path still exists after the failure.

- When ?

  First, as usual with distributed control, we need a time step or a number to identify correctly the events and to avoid the duplication of update messages.

  But it is also possible to compute a statistical profile of messages per origin to identify Ases and routers which have the largest influence on the convergence time and whose informations are not reliable.

- What about the traffic ?

  Can we build a profile for the traffic and is the traffic consistent with the profile ? If not, may I infer mixture with uniform traffic or with hot-spot traffic. ?

- Is the router OK and if not, try to answer why ? And of course, in this case, try to correct or control ?

Thus we have to study the algorithmic aspects of BGP, the topology monitoring, the traffic monitoring, the control and the interaction between these techniques. Furthermore, even if some theoretical aspects can be proved in isolation, most of the interaction must be validated using simulation. In the first important step, we focus on distributed routing algorithms in graphs modeling Internet networks.

2.2 Routing algorithms to solve BGP instability

Considering first routing tables only, from a theoretical point of view, a routing in an (Internet) graph is a set of paths, one path for each pair of nodes. Since the routing tables do not take care of the origin of a packet but only its destination, the routing function in this context are said to be *consistent* : if the path $R(u, v)$ in the routing from node u to node v crosses node w, then it contains the path from w to v (i.e., $R(w, v) \subset R(u, v)$. BGP protocol is a way of determining such routings. It can be seen as a greedy algorithm to obtain a consistent routing : each node knows the routes used by its neighbors and it chooses one such route for each destination.

One of the main problem to be focused on about BGP concerns the instability of the network in case of link failures. When a link fails, all the paths using it in the routing have to be changed. One can define the instability as the time between a link failure and the updated of routing tables making together a new correct consistent routing (we talk about convergence time). The instability could also be evaluated by the perturbation of the traffic during the convergence time. Moreover, these problems of instability make that some routes chosen by some nodes in BGP are not the real ones induced by the routing tables, in particular if the average time between two failures is less than the average convergence time. This inconsistency between tables and BGP protocol wanes the interest of using BGP. Such an inconsistency could also occur in case of Byzantine behavior of some BGP nodes, i.e., if they communicate false and/or unchosen routes to their neighbors. At end, some problems can occur in BGP if different linked BGP nodes have incompatible routes choice policies.

Thus, to deal with these instability and inconsistencies problems, we will focus on different aspects.

- *Decreasing the convergence time.* The first natural theoretical question is to know if there are good and bad consistent routing in terms of convergence time. If the answer is yes, the second question is how to modify BGP policies to obtain good consistent routings.

  Another point is to focus on some alternative algorithmic methods to deal with link failures. First, in some particular consistent routings, if an edge fails then the routing could be repaired without changing the routing tables of each node being the origin of a path using this failed edge. Once could use some properties of alternative paths for each pair of nodes with some common edges properties. Secondly, a k-spanner of an edge in a graph is a path of length at most k. In case of link failures, its spanner can be used as a shunt path being transparent for the routing (this could be used only for short time link failures).

- Inconsistency. We will focus on two aspects: detecting inconsistency and avoiding as far as possible inconsistency.

  To detect inconsistency, three approaches could be used. First, all the routes communicated to a node, even the ones it has not chosen, give to it a (partial and more or less reliable) vision of the network moving time after time. This vision could be used by each node to evaluate the reliability of each information given by its neighbors. Secondly, the logical informations obtained from the comparison of the routes proposed by different neighbors could be used. For example, consider two different neighbors u and v of a node, each one proposing to it a route to a same destination with respective lengths d_u

and d_v, such that $d_u >> d_v$. Then, the target node can conclude that either u does not consider a short length policy to choose routes or u and/or v has a Byzantine behavior. At end, each BGP node can use some passive (traffic inferring) and/or active tomography techniques to check the consistency of the routes it knows.

To avoid (as far as possible) inconsistency and divergence aspects in BGP, since each operator should have its own private policy for choosing routes, it seems difficult to verify the compatibility between all these policies (except if we assume a global arbiter knowing all these informations). Thus, we could only think of giving some basic rules all the operators have to respect or to propose some hierarchical policies where the second one operates if no routes corresponding to the first policies are known since a long time, and so on.

To deal with the first point concerning convergence time, we have to focus on the following consecutive points.

1. For any graph modeling a network, definition of parameters to evaluate the convergence time and the traffic perturbation behavior of a consistent routing.
2. Given a network topology, design of a polynomial algorithm to (randomly) obtain consistent routing being good or bad in terms of the parameters defined.
3. Simulation the behavior of a network where links failure occur to compare the different routings obtained in the previous point.
4. If the simulations show the interest of good routings, how to modify BGP to obtain as far as possible such routings.

One example of convenient routing parameter to be studied concerns load balancing. The load of a routing function have been intensively studied in the context of topologies of parallel architectures. Many works focused on the *forwarding index* of a graph being the the minimum value of the maximum load of an edge considering all the possible routings. This problem have been shown to be NP-hard. Even if these works didn't focus on consistent routing, it seems that determining a consistent routing with minimum maximum load is a hard problem.

2.3 Convergence and self-stabilization

The concept of self-stabilization was first introduced by Edsger W. Dijkstra in 1974 [5]. It is now considered to be the most general technique to design a system to tolerate arbitrary transient faults. A self-stabilizing system guarantees that starting from an arbitrary state, the system converges to a legal configuration in a finite number of steps, and remains in a legal state until another fault occurs (see also [6]). It is desirable that even if the error occurs rarely in the system, the networks should recover from those faults automatically [7].

In the context of computer networks, resuming correct behavior after a fault occurs can be very costly [8]: the whole network may have to be shut down and globally reset in a good initial state. While this approach is feasible for small networks, it is far from practical in large networks such as the Internet. Self-stabilization provides a way to recover from faults without the cost and inconvenience of a generalized human intervention: after a fault is diagnosed, one simply has to remove, repair, or reinitialize the faulty components, and the system, by itself, will return to a good global state within a relatively short amount of time.

Routing instability in the Internet has a number of origins including route configuration errors, transient physical and data link problems, software bugs, and sometimes memory corruption. In a recent work of Varghese and Jayaram (see [9]), it is proved that the crash failures of routers can lead every other router in the network to an arbitrary state, and that if links can reorder and lose messages, then any incorrect global state is reachable. Therefore, it is very important to have a self-stabilizing [5] routing protocol which can recover from any arbitrary faults without any external intervention.

So far, the research about BGP focused on the *stability* of the protocol. Assume that a computer network is started from an initial coherent state (where known paths to the destination do exist, where routers are properly configured, etc.), then a particular BGP instance (*i.e.* a particular route policy configuration of BGP in the network) is stable if it reaches a final global state where all nodes have a stable path toward the destination. Unfortunately, knowing if a particular instance of BGP is stable is an NP-complete problem (see [10]). Thus, proposals to guarantee stability include a global sufficient condition on the system (see [11]), a local condition on the BGP routers' policies (see [12]), and a dynamic additional algorithm that is run on each BGP router (The Safe Path Vector Protocol as defined in [13]).

One aspect of this project is to focus on the *self-stabilizing* properties of the protocol. Assume that a computer network is started from an arbitrary initial configuration, (where known paths can be arbitrary and where routers can be completely mis-configured), then a self-stabilizing BGP solution guarantees, as soon as faults cease, that the network eventually reaches an initial coherent state, and then a stable final configuration (assuming a stable BGP is run).

None of the approaches we mentioned (that guarantee BGP stability) is self-stabilizing. Our approach in this project is twofold:

1. *Design self-stabilizing versions of the BGP protocol.* For far, and to the best of our knowledge, only two self-stabilizing versions of the BGP protocol exist. The first one, [14], is a self-stabilizing version of the Safe Path Vector Protocol of [13]; it assumes realistic communications between nodes, and provides all available routes, but and has high memory requirements. In contrast, [15] has smaller memory requirements than [14], but uses a theoretical reliable communication model, and may removes routes that fit all policies of the system. One important issue would be to design a self-stabilizing version of BGP that is realistic (*i.e.* implementable) and that does not have high resources consumption.

2. *Study the self-stabilizing properties of the BGP protocol itself.* Previous approaches to guarantee BGP stability provide sufficient conditions on the acceptable routing policies. If such conditions are satisfied, the resulting protocol is stable (*i.e.* starting from an initial well known configuration, it provides a loop free route to every autonomous system). Since BGP is designed to accept the arrival and departure of autonomous systems dynamically (along with their corresponding routing policies), it should be able to start from an arbitrary configuration (resulting from a brutal modification of the autonomous systems graph topology). It is still unclear whether the sufficient conditions that have been provided so far for stability are also sufficient conditions for self-stabilization. In a high level communication model, at least two approaches (see [11, 12]) proved to be self-stabilizing. Further investigation is needed to characterize the minimal possible sufficient condition for stability and self-stabilization.

In a transversal way, none of the currently proposed self-stabilizing BGP versions has been tested in a realistic environment (*e.g.* in a distributed simulator). It is needed to run benchmarks to analyze and compare non-stabilizing and self-stabilizing versions of BGP according to several criteria: fault-tolerance, speed of convergence toward a loop-free routing, memory consumption.

2.4 Topology Monitoring

Suppose now that we have the following assumptions realized by a path-vector protocol :

- We can model an AS by a node in a graph.
- Each node (AS) selects only one path among its available paths to a destination and uses it to forward packets. This route will be advertised to its neighbors.
- When a neighbor receives the advertised route, it keeps it as a candidate for path selection instead of the previous one sent by the same neighbor.

This what authors of [20] called a simple path vector protocol. As far as BGP is concerned, it has observed that it cannot obey to this model. Indeed, on 06/08/2001, the Oregon Route Views Server [22] showed that

AS701 announced two different routes to prefix 169.131.0.0/16 ; it announced the route (701 6079 4527) to AS1, whereas AS1740 has learned the route (701 6347 4527).

This problem is principally due to the fact that one AS may contain many BGP routers through which it may advertise multiple routes to a single destination. Furthermore, due to some internal link failures, an AS could be partitioned into multiple parts and routers belonging to different partitions might select and use different routes to one destination. As a result, in BGP model, an AS cannot simply be modeled as a node!

However, in [20], authors introduce consistency assertions for BGP [1]. For example, they suggest to virtually divide an AS X into multiple *logical ASes* with some assumptions where each logical AS is uniquely identified in the Internet by the couple (X,RID) and RID is the entry router identifier. All the BGP routers in AS X selecting the best route from the entry routerID belong to the logical AS (X,RID). Note that the logical AS is defined in the context of one particular destination. Therefore one real AS may have different divisions of the logical ASes for different destinations. (See [21] for more details).

Consequently, The BGP model that seems to be more realistic is to divide each AS into logical ASes and to model each logical AS by a node.

Starting with the standard BGP, we want to ask what combinations of network information and router behavior can improve the stability and robustness of BGP. By network information, we mean any degree of knowledge on network parameters:

- at the topological level: local neighborhood, set of routes
- at the metric level: simple hop counts, link bandwidths, detailed traffic metrics (throughputs, flow characteristics...);
- at the temporal level. Other types of informations such as route change advertisements or other alarms (failures, overloads,...) produced at the network control level may be used as well.

The concern is to achieve the best possible compromise between the speed of convergence and the amount of information needed by the control mechanism. An important factor to take into account is also the "cost" of obtaining the information involved within a reasonable delay, and the overhead for the network, as compared to the standard BGP protocol.

The first idea consists to add failure information to BGP update messages. When a link failure occurs, the withdrawal message can be augmented to explain which link is down and a time-stamp. All the BGP routers which receives this message, can now choose a new route which does not use this particular link. In the standard BGP approach, the reason of the withdrawal is not given and another bad path may be selected to replace the one withdrawn. Remember that BGP uses TCP to send these messages and the messages are not synchronous. Similarly, routers going down can be detected by an "hello" protocol and in case of failure, this information has to be sent with the withdrawn messages. Thus extending BGP messages, we can obtain a more realistic image of the real network. Of course, we must add timestamps to be sure that the information we get are still valid. Once the messages contain such a time-stamp, it is possible to filter much more efficiently the withdrawal of messages and to avoid sending several update message for the same event.

Another idea consists in the analysis of the paths advertised by BGP. Clearly, a path is a set of links but this information is not used in BGP. Finally, we may also study the temporal profile of the udpate traffic to identify routes which behaves poorly. A route which is flapping gets a penalty for each flap. As soon as the cumulative penalty reaches a predefined value, a control can be realized. In route dampening technique, this control consist in the transient suppression of the route publication. More complex control will be studied.

This activity is clearly dependent on the results of the part of the project concerned with tomography and statistics of traffic: the information on the network state (routes, traffic,...) or the router states (normal, loaded, overloaded,...) have to be collected somehow. As a result of the analysis, extensions to the BGP protocol giving it superior security properties will be be suggested. The changes will probably concern:

[1] The consistency definition given in [20] is different from our definition because it relates to protocols not to routing

- the information contained in BGP messages;
- the frequencies of message exchanges;
- rules for router algorithms.

2.5 Traffic Monitoring and Control

The recent worm propagations [17], and Distributed Denial of Services have shown Internet vulnerability. Some worms propagate using random IP addresses [18] but these addresses will almost surely provoke a cache miss. Indeed despite aggregation based on CIDR, the size of routing table is still growing. As during infection, the worms represent a significant part of the traffic, the routers experiences a large number of cache misses. This may provoke crashes (at least there exist a statistical evidence between worms propagation and routers crashes).

Clearly, using a statistical profile for traffic, it is possible to infer a modification of the traffic and to try to adapt to this new traffic. For instance, it may be possible in this case to route the packets in a different fashion to avoid stressing the router because of cache misses (for instance Deflection (or Hot Potato routing)). The question is twofold:

- How to build a statistical profile for traffic and how to infer from new data ?
- How to control this new traffic ?

We have already studied the first question in another context: the aggregated flows coming from the access network into the UMTS core network. The principle of operation at the UMTS MAC layer on the radio interface will smooth the flows generated by the mobile users. So we decided to use an MMPP model with few number of phases (2, 4) to represent the aggregated traffic flows at the border gateway of the core network. In order to determine the MMPP statistical parameters, we used a HMM (Hidden Markov Model) approach. It is possible to show that the best estimation methods for MMPP are those based on Maximum Likelihood Estimates (MLE). We used the iterative algorithm to reach the convergence of the estimate parameters with a predetermined error precision. The speed to reach the convergence for the algorithm becomes important when the estimation will be made in real time Within the Internet, several studies have been made in order to characterize the aggregated flows statistics. In [23], it is shown that TCP flows are of two types (long, short). In [25], the authors proposes a more general approach in order to partition the Internet flows into an arbitrary number of classes. In the perspective of the introduction of the QoS mechanisms within the Internet, the network management within an operated and controlled IP network should deal with a limited number of traffic aggregated classes, each one is supposed to be sufficiently well known. A first issue will consist in the determination of a reasonable number of classes allowing a sufficiently precise characterization of the traffic exchanged between the border gateways. Within border routers, another important issue will be the detection of significant variation of the flows behavior in terms of statistical parameters variations. The significant variations occur most likely following an abnormal network event like an important nodes / links failures, Distributed Denial of Service or viruses attacks. The variation detection should be done in real time within the border gateway. The traffic parameters estimation will use a network tomography techniques. The decision process based on the parameters estimation will permit the update of the routing tables in terms of rerouting decisions possibly in relation with the QoS class.

The second question (control) is more complex. We want to study two techniques based on routing or deletion and rate reduction.

- The first idea consists in a modification of the routing algorithm to avoid cache misses. We will study a convergence routing on the whole net or a sub-net. The convergence routing is based on a fixed route such that at each step along the path your distance to destination will decrease surely. But the path is not a shortest path. Another approach is based on the deflection or hot potato routing.

- A more aggressive control consists in packet deletion (short time answer) and rate reduction (long time). The deletion can be done immediately at the router which detects an abnormal traffic. Such a deletion is of course probabilistic in nature. Furthermore, if the BGP routers are able to identified themselves and want to collaborate, the router which detects the traffic problem can build the vector of flow inputs for the whole network and find the routers which send him more packets than expected. In a collaboration, it may ask to these routers to drop the packets and to limit the bandwidth available. This traceback at the BGP level looks more efficient that the other proposals based on IP-traceback [19] or ICMP traceback. Of course such a control has to be carefully designed and tested as we propose to drop packets to save Internet connectivity.

For the test of routers stress and crash, it may be possible that we can use an Alcatel simulator for their own routers.

2.6 Simulation and Distributed Simulation

Another aspect on which there is a clear interdependence of activities is the development of the BGP simulator. On the one hand, the ideas of new protocols will have to be tested on the simulator, and on the other hand, the development of the simulator will developed so as to accommodate seamlessly new protocol features and router behaviors.

As usual, with simulation of routing algorithms, we must check the techniques on a large scale simulation. Drawbacks or improvements of the algorithms may only appear for a large network (more than 200 nodes). This is the first difficult point here. Let us now review the problems:

- BGP messages are exchanges using TCP packets and these packets are subject to the same congestion problems than ordinary packets which must be explicitly represented in the simulator. Thus we have to simulate BGP flows but also TCP flows (or at least aggregation of TCP flows), TCP control by window size modification and time out. As the two flows are not based on the same time-scale, the simulation time must be quite long.

- To evaluate routing algorithms we must represent explicitly a large number of hosts and correct topology. The first point will need some processing power. The second point is related to the topology of Internet, an open problem at this time, even if the number of publications on this subject is quite large. Hopefully, we make the simulation at the AS level and the topology of the ASes is much more understood because of the routing tables periodically dumped and available on the Internet.

- The simulator must represent some local politics to select the best route. Local policies may be quite complex. The Cisco description use 10 nested tests which begin like this:

    1. If NextHop is inaccessible do not consider it.
    2. Prefer the largest Weight.
    3. If same weight prefer largest Local Preference.
    4. If same Local Preference prefer the route that the specified router has originated.
    5. If no route was originated prefer the shorter AS path.

- The simulator must carry the same type of information we add in the extension of the protocol. Therefore, me must have some TCP and UDP traffic flows.

- Furthermore, the simulator has to be efficient. This prevents to use the BGP simulator based on SSFNET [26].

One possible direction to obtain enough processing power is the distribution of the simulator. At PRiSM, we have developed several simulators (sequential but also on a parallel machine) for routing algorithms and complex networks and we have an expertise in distributed simulation. The LIRMM has also been associated

to parallel simulators developed at INRIA Sophia. The first prototype will be sequential to help for the test but the final version will certainly be based on a conservative approach. These two simulators will be freely available for universities and researchers.

2.7 About the Teams

We now briefly present the three teams or groups which are gathered into this project. It is also worthy to remark here that PRiSM laboratory has a long time collaboration with Alcatel research center (Marcoussis) on routers (optical or IP) and routing algorithms. Olivier Marcé (from Alcatel, email : Olivier.Marce@alcatel.fr) will be associated to this project, even if Alcatel is not an official member of the project.

- APR (Algorithmic and Performances of Networks) Team (LIRMM) is devoted to the analysis of network protocols with a focus on service guarantees (Quality of Service). The members of the project have an established expertise in the development of algorithmic mechanisms for *routing with constraints*: constraints on the throughput and delay in multicast applications (variants of the Steiner tree problem), constraints on the wavelength choice in all-optical networks (variants of the graph coloring problem) of geographical constraints (routes with obliged passing points).

  Moreover, the project has contributed to the past in the area of the conception of distributed algorithms and their analysis (both qualitative and quantitative), as well as on the probabilistic analysis of algorithms. The APR project is connected to the MAESTRO project of INRIA Sophia-Antipolis (A. Jean-Marie, DR INRIA, is a member of MAESTRO). This offers the opportunity of fruitful exchanges of information and expertise on traffic modeling and control, which is a topic of interest to MAESTRO.

- Inside PRiSM, the group "Réseaux, Routage, Performance" (Network, Routing and Performance) gathers researchers from three teams (Networking, Algorithmic, and Network Performance Evaluation) to study new routing techniques. This collaboration has a successful collaboration with Alcatel and France Telecom for various aspect of pricing and routing (optical packet networks and more recently BGP). The group has an established expertise in routing algorithms, Markovian techniques for performance evaluation, simulation and distributed simulation, telecommunication and networking. More recently, new topics such as network performance monitoring, tomography and statistical analysis of performance data have been explored.

  The group has already studied several routing techniques which are not based on the shortest paths: for instance Deflection routing and Convergence routing. We have also shown that some of these techniques are almost as efficient as shortest path but are much more reliable when faults appear. The other approach we advocate is based on active and passive tomography. The main idea is to build for each router an image of the real topology that the routers know because of the data and control packets he received and also because he can probe some hosts or routes using the same tools like traceroute.

- LRI

  The LRI team is the pioneer on self-stabilization in France. The first Ph.D. thesis defended in France on Self-stabilization was by S. Delaët in 1995. Since then, self-stabilization spread in many regions such as Amiens, Bordeaux, Compiègne and Reims. Ongoing research on emerging trends in self-stabilization has been carried out at LRI for a decade. As routing is the main current application of self-stabilization (because it guarantees adaptivity, dynamicity, and fault tolerance in a single package), most of the current work of the LRI team is the study and design of self-stabilizing routing and communication protocols, with emphasis on contexts where communications are unreliable (messages can be lost, duplicated, or reordered).

The contribution of all teams would span several subtopics of the project and will be fostered by collaboration between the teams. Most of the researchers involved in the project came from various LRI teams and have already collaborated in research projects. The following points will certainly be developed by an effective collaboration between the teams: the analysis of routing rules and constraints imposed to BGP

routers, the theoretical study of the speed of convergence of distributed routing protocols, the improvement of the BGP protocol, the development and test of the simulation testbed. Note that the GRiD project managed by INRIA action Grand Large may be used to run a distributed simulator and all the teams will use this simulator.

3 Intended results

In our opinion, the border gateway protocol misses the following properties:

- reliability
- support for pricing
- security
- support for Quality of Service
- support for load balancing
- simplicity of configuration

As most of the providers use shortest path selection as local preference algorithm, real examples of BGP oscillation problems are not that frequent. However the lack of reliability and security is a clear limitation to an Internet with QoS and large scale value added services. So the project may have a large impact on the design of routers and the network.

Clearly, the solutions we will study imply more information exchanges and a more complex control with more powerful routers and more bandwidth. Such a trade-off is possible because there exist a lot of unused fibers (and bandwidth) and the routers technology is quite simple and based on ordinary PC architectures. Internet has a lot of security problems (BGP is only one part of the problem) and even a partial answer to the questions we want to address may have a considerable impact. The results expected are:

- the theoretical and experimental evaluations of distributed routing protocols based on different levels of information and different routing decision rules;
- proposals for improvements of the BGP protocol giving it higher *robustness*, *responsiveness* and *security* levels, with their validation.
- A distributed simulation testbed (on a GRID) for the study of large scale routing techniques.

Finally, we show a possible timetable for this project:

Timetable

T0+3	Analysis of BGP security problems.
T0+6	Active and passive tomography for BGP.
T0+6	Functional description of the simulator.
T0+12	First version of the simulator
T0+12	Worst-case evaluation of BGP extensions.
T0+12	How Traffic monitoring on a router may help to detect an attack ?
T0+18	Test and diffusion of the simulator.
T0+24	Stochastic Analysis of the traffic, proposition for control

T0+30 Experimental average-case evaluation of BGP extensions.

T0+30 Stochastic analysis of the convergence.

T0+36 Experimental analysis of traffic monitoring and control

4 References

References

[1] R. Govindan and A. Reddy, "An analysis of Internet inter-domain topology and route stability", IEEE INFOCOM, 1997, pp 850-857.

[2] C. Labovitz, G. R. Malan, and F. Jahanian, "Internet Routing Instability", ACM SIGCOM 97, France, pp 115-126

[3] C. Labovitz, G. R. Malan, and F. Jahanian, "Origins of Internet Routing Instability", IEEE INFOCOM 99, USA

[4] Kannan Varadhan and Ramesh Govindan and Deborah Estrin, "Persistent route oscillations in inter-domain routing", Computer Networks (Amsterdam, Netherlands, 1999, volume32, number 1, pp 1–16,

[5] E.W. Dijkstra", "Self stabilizing systems in spite of distributed control", Communications of the ACM, V 17, 1974, pp 643-644

[6] S Dolev, "Self-stabilization", 2000, The MIT Press

[7] G. Varghese, "Self-stabilization by counter flushing", International ACM Conference on Principles of Distributed Computing, 1994, pp 244-253

[8] Raida Perlman, "Interconnections: Bridges, Routers, Switches, and Internetworking Protocols", 2000,Addison-Wesley Longman

[9] G Varghese and M Jayaram, "The Fault Span of Crash Failures", 2000, Journal of the ACM, V 47, N 2, pp 244-293

[10] T Griffin and G Wilfong, "An analysis of BGP convergent properties", ACM Sigcomm 99, 1999.

[11] T Griffin and FB Shepherd and G Wilfong, "Policy disputes in path-vector protocols", International Conference on Network Protocols (ICNP'99), 1999.

[12] L Gao and T Griffin and J Rexford, "Inherently Safe Backup Routing with BGP", IEEE INFOCOM'01, V 1, pp 547-556, 2001

[13] T Griffin and G Wilfong, "A Safe Path Vector Protocol", IEEE INFOCOM'00, pp 490-499, 2000

[14] Yu Chen and Ajoy K. Datta and Sébastien Tixeuil, "Stabilizing Inter-domain Routing in the Internet", Europar 2002, LNCS 2400, Paderborn, Germany, pp 749-752

[15] Jorge Arturo Cobb and Mohamed G. Gouda and Ravi Musunuri, "A Stabilizing Solution to the Stable Path Problem", Self-Stabilizing Systems 2003, 6th International Symposium, SSS 2003, Lecture Notes in Computer Science 2704, pp 169-183, San Francisco, CA, USA

[16] L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S.F. Wu, and L. Zhang, "Observation and Analysis of BGP behavior under stress", ACM SIGCOMM USENIX IMW, pp 183-195, 2002

[17] M. Liljenstam, Y. Yuan, B.J. Premore, D. Nicol, "A mixed abstraction level simulation model of large-scale Internet worm infestations", IEEE MASCOTS 2002, USA, pp 109–116.

[18] S. Staniford, "Code Red analysis pages", http://www.silicondefense.com/cr

[19] D. X. Song and A. Perring, "Advanced and authenticated marking schemes for IP traceback", Infocom 2001.

[20] D. Pei, X. Zhao, "Improving BGP Convergence through consistency assertions" In Proc.IEEE INFOCOM, 2002

[21] Y. Rekhter and T. Li, "A border Gateway Protocol4 (BGP-4)", RFC 1771. 1995.

[22] D. Meyer, "The Route Views Project", http//:www.antc.uoregon.edu/route-views/

[23] Yin Zhang, Lee Breslau, Vern Paxson, Scott Shenker. "On the Characteristics and Origins of Internet Flow Rates", SIGCOMM.02, August 2002, Pittsburgh, Pennsylvania, USA.

[24] Tobias Ryden, "An EM algorithm for estimation in Markov Modulated Poisson Processes", Computational Statistics and Data Analysis, Volume 21, Issue 4, 1996

[25] Augustin Soule, Kavé Salamatian, Nina Taft, Richard Emilion and Konstantina Papagiannaki, "Flow Classification by Histograms or How to Go on Safari in the Internet", ACM Sigmetrics 2004 / Performance 2004, June 2004, New York, USA

[26] Brian J. Premore, "An Analysis of Convergence Properties of the Border Gateway Protocol Using Discrete Event Simulation" PhD thesis, Dartmouth College Department of Computer Science, May 2003.

5 Bibliographical references of the researchers involved in the project

- O. Cogis, J.-C. König and J. Palaysi: "On the List Colouring Problem". *Proc. ASIAN'02*, LNCS 2050, pp. 47-56, dec. 2002.
- A. Irlande, J.-C. König and C. Laforest, "Construction of low cost and low diameter Steiner trees", *SIROCCO2000*, Carleton Scientific, 197-210, 2000.
- A. Irlande, "Structures de Communication pour les Groupes Multipoints", PhD thesis, University Montpellier II , 2002.
- A. Bouabdallah, J.-C. König, "A distributed algorithm for the mutual exclusion problem", Parallel and Distributed Computing in engineering systems, Tzafe staset al. (Eds), Elsevier (1992) 285–290.
- A. Bouabdallah., J.-C. König, M. B. Yagoubi, "A fault-tolerant algorithm for the mutual exclusion in real-time distributed systems", Journal of computing and Information, Vol. 1 No. 1 (1994), pp. 438–454.
- J.-C. Bermond, J.-C. König, M. Raynal: "General and Efficient Decentralized Consensus Protocols", WDAG 1987, 41–56, 1987.
- A. Jean-Marie and F. Baccelli, "Communication and time complexity of a distributed election protocol", QUESTA, 9, pp. 83–112, 1991.
- D. Barth and P. Berthomé. "The Eulerian stretch of a network topology and the end guarantee of a convergence routing", A paraitre dans Journal of Interconnection Networks (JOIN), **2004**.

- T. Atmaca, D. Barth, P. Berthomé, D. Chiaroni, and al. "Multiservice optical network: main concepts and first achievements of the ROM program", Journal of Lightwave Technology, 19:23–31, January 2001.
- M. Sbai, "BGP as a consistent routing algorithm : improvements and tomography requirements", Preprint of the (Alcatel-PRiSM) TOMO-BGP contract, 2004.
- D. Barth and C. Laforest, "Scattering and multi-scattering with a local routing without buffering", Parallel Computing 25(8), pp. 1035-1057 **1999**
- D. Barth and P. Fragopoulou, "Compact multicast algorithms on grids and tori without intermediate buffering", Parallel Processing Letters 12(1), pp. 31-39, **2002**.
- C. Durbach and J.M. Fourneau, "Performance Evaluation of a Dead Reckoning Mechanism",, Workshop IEEE on Distributed Interactive Simulation and Real Time Applications, Montreal, juillet 98.
- C. Durbach and J.M. Fourneau, "Distributed Interactive Simulation : Bandwidth optimization and group multicast", 14th Int. Symp. Comp. and Inform. Sciences (ISCIS) , Izmir, Turquie, 1999
- D. Barth, P. Berthome, A. Borrero, J.M. Fourneau, C. Laforest, F. Quessette, S. Vial, "Performance comparisons of Eulerian routing and deflection routing in a 2D-mesh all optical network", European Simulation Multiconference (ESM2001), Prague, Rep. Tcheque, 2001
- D. Barth, P. Berthome, T. Czachorski, J.M. Fourneau, C. Laforest, S. Vial, "A mixed defflection and Eulerien routing : design and performance", Europar 2002, Paderborn, Allemagne, LNCS 2400.
- T. Czachorski, J.M. Fourneau, F. Quessette, S. Nowak, "Performance of a Mixed Deflection and Convergence Routing Algorithm", Proceedings of the second Polish-German Teletraffic Symposium, Gdansk, 2002, pp 47–54.
- Paraskevi Fragopoulou, Selim G. Akl, "Edge-Disjoint Spanning Trees on the Star Network with Applications to Fault Tolerance", IEEE Trans. Computers 45, 174-185 (1996)
- David W. Krumme, Paraskevi Fragopoulou, "Minimum Eccentricity Multicast Trees", Discrete Mathematics and Theoretical Computer Science 4, 157-172 (2001)
- H.Aouad, A.Ibrahim, S.Tohme, "UTRAN Traffic Parameters Estimation", IEEE CCNC2004 Conference, Las Vegas, Nevada, USA. January 2004
- R.Naja, S.Tohme, "Buffer Management for smooth and anticipated handovers considering the optimized uplink routing in cellular IP networks", IFIP Conference on Personnal Wireless Communication PWC 2003. September 2003. Venice, Italy
- R.Abi Fadel, S.Tohme, "Connection Admission Control CAC and Differentiated Resources Allocation RA in a Low Earth Orbit LEO Satellite Constellation", IFIP Networking'2002 Conference, Published by Springer-Verlag LNCS Series, May 2002, Pisa, Italy.
- S. Delaët and S. Tixeuil, "Tolerating Transient and Intermittent Failures", Journal of Parallel and Distributed Computing, 2002, 62, n5, pp 961-981,
- S. Delaët and S. Tixeuil, "Un algorithme auto-stabilisant en dépit de communications non fiables", Technique et Science Informatiques, 1998, V 17, N 5, Hermès
- S. Delaët and Duy-So Nguyen and S. Tixeuil, "Stabilité et Auto-stabilisation du Routage Inter-Domaine dans Internet", RIVF 2003, Studia Informatica Universalis, pp 139-144, 2003, Hanoi, Vietnam Ed. Suger

- J. Beauquier, S. Delaët, S. Dolev and S. Tixeuil, "Transient Fault Detectors", DISC'98, pp 62-74, 1998, LNCS 1499, Andros, Greece
- B. Ducourthial and S. Tixeuil, "Self-stabilization with Path Algebra", Theoretical Computer Science, 2003, V 293, N1 , pp 219-236
- B. Ducourthial and S. Tixeuil, "Self-stabilization with r-operators", Distributed Computing, 2001, V 14, N 3, pp 147-162
- Ajoy Kumar Datta and Jerry L. Derby and James E. Lawrence and Sébastien Tixeuil, "Stabilizing Hierarchical Routing", Journal of Interconnexion Networks, 2000 V 1, N 4, pp 283-302.
- C. Johnen, F. Petit and S. Tixeuil, "Auto-stabilisation et Protocoles Réseau", Technique et Science Informatique, 2004, to appear.
- Yu Chen, A. K. Datta and S. Tixeuil, "Stabilizing Inter-domain Routing in the Internet", Europar 2002, pp 749-752, 2002, LNCS 2400, Paderborn, Germany