Essay on Achieving Sub-Second IGP Convergence in Large IP Networks

André Rosa
DI/FCT/NOVA University of Lisbon, Lisboa, Portugal
Student № 48043
af.rosa@campus.fct.unl

I. Introduction

Large-scale deployments of real-time applications require that messages are delivered to their destinations as fast as possible. However, when there are modifications in the network's topology, either by failures or joining of new nodes and links, the information used to forward packets, that is maintained by each router, may become stale. Consequently, this information must be updated, which temporarily leads to some routers having their state updated while others remain with stale information, i.e. routers having different global information. This phenomenon is called inconsistent states and is caused due to the difficulty of all routers become aware of a topology update at the same time, since the propagation of topology changes across the networks is not instantaneous.

Unfortunately, these inconsistent states may lead to routing loops when forwarding packets across the network, which leads to packets not arriving at their intended destination. Thus, minimizing the convergence time of routing information of an update (i.e. time required to all the routers of the network to reach a new consistent state) is of the utmost importance to avoid compromising the delivery of packets on the network.

Therefore, in this essay, we inside on the article [1] which reviews how sub-second convergence times are achieved in Interior Gateway Protocols (IGPs), i.e. link-state routing protocols executed by routers within autonomous systems[1].

II. Overview of Link-State Routing Protocols

In general, in link-state routing protocols each router maintains a global view of the entire network, by receiving control messages from every router in the network, containing the routers to which they are directly connected, i.e. their neighbors or local topology. Then, whenever a router detects a link failure or a new link, it broadcasts a new control message containing their neighbors. Upon the reception of this control message, each router updates its global view of the network, and then locally recomputes the best path, according to some metric, to every other router in the network. Next, from these paths, each router determines through which one of its interfaces each packet received should be forwarded, to reach its intended destination trough the most suitable path.

For this, in link-state routing protocols, each router generally leverages the following four structures:

- *Link State Database (LSDB)*: contains the local topology information of every router of the network, i.e. the global network topology.
- *Shortest Path Tree (SPT)*: is a tree formed by the shortest paths from the current router to all the other routers of the network. The SPT is computed from the LSDB through an algorithm called Shortest Path First (SPF).
- *Routing Information Base (RIB)*[2]: assigns to each network prefix the address of the router to which the packets, whose destination matches the prefix, should be sent to.
- *Forwarding Information Base (FIB)*[3]: assigns to each network prefix the current router's network interface from which the packets, whose destination matches the prefix, should be sent.

Then, each router periodically transmits a type of control packet, called HELLO packet, through each one of its interfaces, to discover links with their neighbors, i.e. determining their local topology.

Next, upon detecting a topology change (i.e. a new link or a link failure), routers broadcast their local topology information, through reliable flooding, inside a special packet called Link State Advertisement (LSA) or Link State Protocol data unit (LSP). Reliable flooding consists in transmitting a packet through every interface, excluding the one from which the packet was received[4], and ensuring the routers at the other end of the link receive it, usually through the usage of acknowledgment (ACK) packets. Routers can also periodically broadcast LSPs (even when no topology changes occurred) to avoid existing incorrect information due to undetected memory or transmission errors.

Whenever a router receives an LSP, it updates its LSDB and floods the LSP. From the updated LSDB, the current router performs a two-way connectivity check on every pair of routers, verifying if both advertise each other. In the case this verification fails, i.e. only one of the routers advertise the other, the link between those routers is removed from the LSDB. Afterward, the router computes the new SPT from the LSDB, determining the shortest paths to each destination, which is then leveraged to update the RIB, assigning to each prefix the neighbor router which belongs to the shortest path between the current router and that prefix.

This essay was written in the context of the APRC course.

[1] e.g. an Internet Service Providers (ISPs)

[2] also called routing table.

[3] also known as a forwarding table.

[4] In the case of the router that initiates the broadcast, the packet is transmitted through all interfaces

Then, the RIB is utilized to update the FIB, assigning to each prefix the network interface that connects the current router with the neighbor router assigned to that prefix in the RIB.

Posteriorly, when a packet is received by a router, the router consults its FIB to determine through which network interface that packet should be forwarded to.

III. Convergence of Routing Protocols

Taking into account how a link state routing protocol updates its information, the convergence time between consistent states can be decomposed into: i) link failure detection time; ii) delays before originating new LSPs; iii) broadcast latency; iv) SPF computation time; and v) RIB and FIB computation time. Next, we will discuss each one individually.

A. Link Failure Detection Time

The communication technologies used to inter-connect routers enables them to detect link failures in a few tens of milliseconds, at the hardware level (physical layer). This characteristic is very important to achieve sub-second convergence times.

B. Delays Before Originating new LSPs

Upon the detection of a link failure, it can be employed a configurable delay to allow link protection mechanisms[5] to activate. If no such mechanisms are available, this delay can be configured to be zero, and thus the link failure detection is immediately handled, i.e. a new LSP is generated and broadcast.

When a new link emerges, it can also be applied a delay before informing routing protocols of the availability of the correspondent interface. This enables routers to deal with network instability due to flapping links. However, these delays negatively impact the convergence time of routing protocols. Thus, to obtain fast and stable convergence, *dynamic timers* can be employed which have small duration when the network is stable, and an exponentially increasing duration on times of network instability.

C. Broadcast Latency

The time required to propagate each LSP throughout the whole network, i.e. the broadcast latency, is another important component of the convergence time. This latency depends on the topology of the network itself, instead of the computational power of each device. Thus, the higher the overall link delays, the higher the broadcast latency.

To ensure the quickest propagation possible, each router should perform *fast flooding* which can be achieved by retransmitting the received LSPs before computing the new RIB (since this is very slow) and not employing pacing timers[6] on LSPs resultant of topology modifications.

[5]mechanisms take "recover" a link between two routers by, for instance, leveraging redundant cables connecting the routers

[6]delays applied to ensure intervals between consecutive LSPs transmissions/retransmissions

D. SPF Computation Time

An additional component of the convergence time is the time required to execute the SPF algorithm, which computes the shortest paths to all destinations. In this case, the higher the number of routers on the network, the higher the SPF computational time. There can be employed two strategies to minimize this time: i) minimize the number of routers of the topology (grouping some routers under one representative in broadcast networks such as Ethernet) and ii) leverage *incremental SPF (iSPF)* which recalculates only the routes affected by the topology changes. *iSPF* is the main factor in reducing the SPT computation time.

In case several LSPs are received in a short amount of time, this triggers several executions of the SPF algorithm. Therefore, it might be useful to wait for a short period before executing the SPF, called *initial SPF wait*, to enable routers to receive multiple LSPs before executing the SPF algorithm. Furthermore, *dynamic timers* can also be applied to control the execution of SPF to achieve fast convergence when the network is stable, and reasonable processing overhead otherwise.

E. RIB and FIB Computation Time

Finally, the last component of the convergence time is the time required to compute the RIB and FIB, and which is directly proportional to the number of network prefixes belonging to routers whose the shortest path to was modified by *iSPF*. This is usually the main bottleneck in the convergence time minimization, since each router announces a large amount of network prefixes.

There are two approaches to minimize the RIB and FIB computation time: i) *incremental FIB and RIB updates*, and ii) *prioritize prefixes*. *Incremental FIB and RIB updates* consists in partially updating the FIB and RIB at successive intervals, which allows some entries to be updated faster than others. By *prioritizing the prefixes*, it is possible to select the ones that should be updated first. By combining these techniques, the FIB and RIB update time can be greatly reduced for high priority prefixes.

IV. Experimental Evaluation

The authors of [1] performed an experimental evaluation to measure the impact of the previous mechanisms and techniques for achieving sub-second convergence time. The experimental evaluation was performed on a simulator on two distinct networks: the pan-European Research Network (GÉANT), a small network with 22 routers; and a Tier 1 ISP network, a large network with 200 routers. During this evaluation, two scenarios were studied: link failures and router failures, which are discussed next.

A. Link Failures

Link failures are the most frequent event that modifies networks' topologies.

When a link between two routers fails, both detect that failure, not necessarily at the same time, and each broadcast a

new LSP, reflecting that topology modification. Due to the two-way connectivity check performed, only one of those LSPs is required to be received by any other router, to correctly update the FIB. Therefore, small *SPF initial waits* were shown to be most suitable since it is not required to wait for the other LSP. Additionally, this also causes *Fast flooding* to have little impact since the SPF and FIB updates are the components that dominate the convergence time. Consequently, the *incremental FIB and RIB update* has a greater impact on improving the performance of convergence times, specially on the ISP network, since the number of prefixes of that network is much higher.

B. Router Failures

When a router fails, all its neighbors broadcast a new LSP. In this case, it may be necessary for all the routers to receive all these LSPs. Consequently, some routers may have to update their FIB multiple times, which increase the convergence time. Thus, *fast flooding* with high *SPF initial waits* are necessary to achieve sub-second convergence times. The *SPF initial waits* must be carefully configured to ensure that for the majority of router failures, all LSPs are received by all routers before they start executing SPF. In small networks, *SPF initial wait* can be small. However, in large networks, this value should be large. Furthermore, while routers compute the SPT and update the RIB and FIB, they cannot perform other actions, like immediately flooding other LSPs received meanwhile. As a result, *SPF initial waits* and *incremental RIB and FIB updates* are beneficial to allow processing those LSPs faster and consequently improving the performance of convergence times.

The *dynamic timers* applied before SPF computation were shown to have little to no impact on the overall convergence time.

V. Conclusion

In summary, in this essay we discussed the strategies and techniques employed to achieve sub-second convergence times to update routing information, in link-state routing protocols. The experimental evaluation of these strategies and techniques revealed that *fast flooding*, *iSTF*, and *incremental RIB and FIB updates* are very important to achieve low convergence times. Additionally, in large networks, the value of *SPF initial wait* was shown to depend on the type of expected failures on the network. Thus, if link failures are the most frequent type of failures, low *SPF initial wait* values should be employed. On the other hand, if router failures are the most frequent type of failures, high *SPF initial wait* values are fundamental to ensure sub-second convergence times.

References

[1] P. Francois, C. Filsfils, J. Evans, and O. Bonaventure, “Achieving sub-second igp convergence in large ip networks,” *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 3, p. 35–44, Jul. 2005. [Online]. Available: https://doi.org/10.1145/1070873.1070877