

# Achieving sub-second IGP convergence in large IP networks

## Article Summary

José Duarte

March 2019

## 1 Introduction

The paper describes and analyses the various factors influencing convergence time in IGP networks, presenting a short description of the IS-IS protocol followed by a description of the convergence time. Showing that the main problems achieving sub-second convergence lie on the Routing Information Base (RIB) and Forwarding Information Base (FIB) update, being able to reduce its influence by introducing prefix reduction during the design stages, prefix prioritization during RIB/FIB updates and incremental updates to the FIB.

A simulation model was used to study the convergence time in larger networks concluding that the sub-second convergence goal is achievable on ISP scale networks.

## 2 IS-IS Protocol

In the IS-IS protocol the router exchanges HELLO PDUs with its neighbors to determine its local topology, afterwards it will flood a link-state packet (LSP) describing its local topology, this packet will contain at least the identifier of its neighbors.

For broadcast networks, IS-IS routers will elect a router to "represent" the broadcast network, the router will then generate a LSP describing this network and all attached routers.

Two situations may force the flooding of the LSP, the information contained in the LSP changed, meaning that a new LSP must be generated and flooded, or the LSP lifetime ended and it must be flooded again. In order to ensure reliability when flooding the network, each LSP is acknowledged on each link.

When a router receives a LSP describing a topology change, it updates the Link State Database (LSDB), this event triggers the update of the RIB which in turn triggers the update of the FIB. In order to update the RIB a new Shortest Path Tree (SPT) must be computed based on the information contained in the LSDB.

## 3 Convergence Time Components

The convergence time can be characterized as  $D + O + F + SPT + RIB + DD$  where:

- $D$  - Link failure detection time
- $O$  - Time to originate the LSP describing the new topology
- $F$  - Complete flooding time from the node detecting the failure to the rerouting nodes that require a FIB update to bring the network to a consistent forwarding state
- $SPT$  - Shortest Path Tree computation time

- *RIB* - Time taken to update the RIB and the FIB
- *DD* - Time to distribute the FIB updates to the linecards (in the case of a distributed router architecture)

### 3.1 Router Architecture, Processor Performance, Operating System

One of the main bottlenecks on convergence time is the time taken updating the RIB and FIB components, it is clear that the faster the processor, the faster the convergence.

A distributed router architecture with hardware packet processors is presented as a very well suited solution to the problem given that the CPU (RP) is able to dedicated all power to the control plane operation. Handling all routing operations and delegating to the linecards the write of the FIB updates to the hardware packet processors.

The operating system (OS) running on the RP and the LineCard (LC) CPU's implements a process scheduler with multiple priorities and preemption capabilities, allowing for the IS-IS process to be scheduled immediately upon link failure.

During convergence on a distributed platform, two processes share the CPU: the IS-IS process to update the RIB and the FIB and the process distributing the FIB updates to the LC CPU's.

Given that the main bottleneck is present in the RIB update, the IS-IS process will start by updating the prefixes with the higher priority, in order to ensure the update of the LC's a quantum is used, when the quantum is over the IPC process is scheduled, distributing the updates to the LC's, going back to the prefix update when the quantum is over, repeating the process.

### 3.2 Link Failure Detection

The use of Packet over SDH/SONET (POS) links is pointed as being one of the major enablers of sub-second IGP convergence given their ability to detect failures in tens of milliseconds.

Mechanisms present in SDH and SONET allow the LC hardware to detect failure in less than 10 milliseconds. When a failure is detected an high-priority interrupt is fired causing a POS routine to be executed, which enforces a user-defined hold-time delaying the communication of the failure to the central CPU and allowing for SDH/SONET protection to occur. If protection is not available, the failure is immediately signaled to the common CPU, updating the interface and scheduling IS-IS for reaction.

The measurements confirm that in most cases failure detection takes less than 20ms, under heavy load, although rare, it is possible that another process was owning the CPU when the LC detected a failure, in the worst-case, the process has the same priority as IS-IS, it was scheduled just before the failure and busy enough to consume the its full quantum.

When link or path alarms are cleared, timers are used to hold down the interface for an additional 10s before informing the routing protocols, ensuring robustness against unstable situations.

### 3.3 LSP Origination

In order to achieve a rapid convergence while avoiding the generation of unnecessary LSP's, generation timers with fixed values have been used, these values have been set to limit the overhead during network instability, however they impact the convergence time by introducing unnecessary wait time when the network is stable.

Introducing dynamic timers to control LSP generation solves the problem. Adapting to the current network conditions, with short generation times when the network is stable, allowing fast

information exchange, and increasing exponentially when the network is unstable, introducing overhead to allow the network to settle down.

### 3.4 Flooding

The flooding time from the Failure node to the Rerouting nodes is the sum at each hop of the bufferization, serialization, propagation and the ISIS processing time.

Serialization and bufferization is considered to be negligible and only the propagation delay impact on convergence is measured in the simulation.

There are several optimizations that can be implemented for the fast-flooding process.

For a single-threaded IS-IS implementation the LSP must be flooded before the RIB is updated, given the time the latter can take jeopardizing the network convergence when the local node is not the only rerouting node.

The IS-IS specification suggests a *33ms* pacing timer which is considered outdated, given the evolution in CPU and link speeds over 15 years, and damaging to the convergence time. Fast flooding is introduced as a solution to the pacing problems, bypassing pacing on LSPs describing a new link-state change event and applying pacing on Refresh and TE LSPs. However, in the case of link flapping bursts of LSPs must be avoided, in order to do so, burst size is controlled and pacing is re-applied when a configurable certain amount of LSPs has been flooded over a configurable certain amount of time.

The measurements for percentile-90, 95 and 100 were respectively *2ms*, *28ms* and *52ms*. In practice, the worst-case scenario should be neglected due to the low probability of occurrence, given that for it to really occur, this worst-case must occur in the multiple parallel paths that exist in a meshed network.

It is concluded that the time to flood a single LSP is negligible compared to the sub-second convergence objective.

### 3.5 SPT Computation

The dynamic timers from the section 3.3 were also applied for the SPF calculations allowing for IGP to be tuned such that when the network is stable, the timers will be short, reacting to topology changes within a few milliseconds, when the network is unstable SPF recalculation timers will increase in order to throttle the rate of response to network events.

This ensures fast convergence when the network is stable and only incurs moderate overhead when the network is unstable.

In IGP networks designed for fast convergence it is best practice to minimize the number of nodes in the topology.

An important optimization to SPF is the incremental SPF (iSPF). iSPF analyses the impact of new LSPs on the previously computed SPF and minimizes the number of computations required updating only the necessary parts of the SPT. For failures in smaller subsets of the network, the bigger the iSPF computation gain when compared to a full SPF.

Results show that the SPT computation can be executed in tens of milliseconds, without compromising network stability.

### 3.6 RIB and FIB update

The RIB/FIB update duration is linearly dependent on the number of modified prefixes. As expected, the update time depends primarily on processor performance as this is the main bottleneck in the convergence process. It is also shown that the time taken per RIB/FIB update is  $\sim 146\mu s$ .

In order to minimize the RIB/FIB update component three approaches can be taken: implementing a network design rule to minimize the number of IGP prefixes, protocol and implementation optimization to allow the prioritization of some prefixes during the update and the optimization of the table management code. The paper discusses the first two.

For a large ISP there are usually a few hundred important prefixes, the other only links interconnecting routers and since these prefixes are rarely used as destination addresses they mostly introduce overhead to the RIB/FIB updates.

Introduction prefix prioritization solves the problem, the important prefixes will be updated first making the worst-case update duration scale based on the number of important prefixes instead of all prefixes.

## 4 Simulation

### 4.1 Router model

The link failure component was modeled in a way that both routers will not detect the failure at the same time. For low delay links the lab measurements were used and for high delay links the time it takes to propagate the failure detection signal from location A to another random location.

The LSP generation timers were not modeled and the simulator allowed the router to flood its LSP immediately.

The exponential backoff mechanism from section 3.5 was also modeled and it is configured with three parameters: *spf<sub>i</sub>initial<sub>wait</sub>*, *spf<sub>e</sub>exponential<sub>i</sub>ncement* and *spf<sub>m</sub>aximum<sub>wait</sub>*.

The SPT computation time was modeled as a function of the number of nodes in the network with some jitter in order to take into account other processes that may be running on the router CPU. Only the full SPT computation was considered and not its incremental variants.

To model the time required to update the FIB, the number of prefixes whose FIB entries have changed and then multiplying the number of FIB updates with the time required to update one entry. The simulator models two kinds of FIB updates: *static* and *incremental*.

### 4.2 Convergence time

When there are no failures within the network, the routing is consistent, however on a link failure, the routers using the failed link need to update their FIB, while the FIBs are being updated the routing may not be consistent anymore.

The *instant of convergence*