June 2023

HITLESS GRACEFUL INSERTION AND REMOVAL OF A ROUTER/ SWITCH IN HIGHLY RELIABLE MULTICAST NETWORKS

Francesco Meo

Ramakrishnan Chokkanathapuram Sundaram

Stig I Venaas

Martin Hospodar

Swetha Velamuri

HITLESS GRACEFUL INSERTION AND REMOVAL OF A ROUTER/SWITCH IN HIGHLY RELIABLE MULTICAST NETWORKS

AUTHORS:
Francesco Meo
Ramakrishnan Chokkanathapuram Sundaram
Stig I Venaas
Martin Hospodar
Swetha Velamuri

ABSTRACT

Currently, the support for graceful insertion and removal (GIR) within a Protocol-Independent Multicast (PIM) environment is limited. As a result, there is no mechanism today that allows for a soft migration of flows when, for example, planning for a maintenance window. Techniques are presented herein that support a hitless upgrade capability that avoids impacting any flow during the upgrade and reload of a spine switch. This is the most difficult task today for customer networks, where multicast flows run all of the time and no disruption is acceptable. Aspects of the presented techniques include a new PIM hello message type-length-value (TLV) option. Such an option may be referred to herein as a progressive graceful insertion and removal (PGIR) capability option.

DETAILED DESCRIPTION

As an initial matter, it will be helpful to confirm the meaning of a number of terms that are employed in the narrative that follows. Table 1, below, such terms and their intended meaning within the scope of this submission.

Table 1: Selected Terms and Meaning

Term	Meaning
ASM	Any source multicast
FHR	First hop router (a switch where a source is connected)
Flow policy	A configuration policy that maps a flow with an administratively defined capability such as required bandwidth or priority
IPFM	Internet Protocol (IP) fabric for media
LHR	Last hop router (a switch where receivers are connected)
PGIR	Progressive graceful insertion and removal
RPF	Reverse path forwarding interface (an expected incoming interface for a multicast traffic stream)
SSM	Source-specific multicast

Techniques are presented herein that support a method for achieving hitless graceful isolation, with respect to multicast flows, for a switch or a router in order to enable that device’s upgrade or reload, during a planned maintenance window, with no traffic impact. Currently, graceful insertion and removal (GIR) within a Protocol-Independent Multicast (PIM) environment is only supported in scenarios in which a forwarder role changes. However, such an approach is in no way hitless or even close to that. At present, there is no solution in the market that allows for the upgrade of a router in a multicast network that guarantees a hitless upgrade.

As described above, an administrator currently has no way of performing a progressive or graceful removal of a device. This is a particular requirement that is present mostly in the media industry. Large operators present shows that are broadcast at any time of the day, which cannot experience any interruptions, and that can last for many hours. While a leaf may be easily upgraded or removed whenever there is no sender or receiver on that particular leaf (and in case of virtual workloads, this is even easier), a spine’s upgrade or removal poses significant problems. Traditional GIR approaches cannot be used because they still create a disruption in the traffic streams that are always present in the network.

Consequently, a need exists for a mechanism under which, when downtime is planned for a spine, any current stream may continue to be serviced until completion, but all new streams may not select that spine. Once all of the streams on that spine complete their duration, the spine will remain empty and can then be upgraded or removed with no traffic impact at all.

Under the presented techniques, a multicast router can be provided that supports a graceful migration capability can be used to indicate the same to all of its PIM neighbors using a new PIM hello message type-length-value (TLV) option. Such an option may be referred to herein as a progressive graceful insertion and removal (PGIR) capability option. That capability will ensure that a router that is going into a planned maintenance mode will inform all of its capable neighbors that they need to progressively stop using it as a viable or usable path for stitching new multicast flows and, instead, employ alternative paths that go through other routers that are not being placed into planned maintenance.

The new PGIR capability option that is available under the techniques presented herein provides for an efficient facility from a multicast perspective since PIM hello messages are already exchanged by neighbors. Every router maintains a list of GIR-capable neighbors for multicast flows. A triggered hello message may be sent from a GIR router to all of its neighbors indicating the start of a PGIR procedure. Figure 1, below, presents elements of a data frame that may be employed for a PIM hello message TLV option for GIR according to the techniques presented herein.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver| Type  | Reserved  | Checksum                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Option Type(GIR)  | Option Length (1)                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Option Value  |
+-+-+-+-+-+-+-+-+
```

Figure 1: PIM Hello TLV Message Format for GIR

As indicated in Figure 1, above, a new PIM option type of GIR may be employed. That option type may have a value that is to-be-determined and a length of one byte. Additionally, one of two option values may be employed. A value of '01' may indicate a GIR capability and periodic hello messages may carry this value. Alternatively, a value of

'02' may indicate a GIR notification (e.g., a router that is undergoing isolation in connection with maintenance may send this value). Such a value may be sent during the time that the router stays in maintenance mode until the moment that it becomes empty and is reloaded. Once the router is reloaded, this flag may be reset, thus notifying all of the neighbors that this switch is no longer in maintenance.

The techniques of this proposal may be performed, as follows. For example, during a first step, a router that will be undergoing PGIR may send a PIM hello message with a new TLV option (as described above) to all of its PIM neighbors indicating that the router wishes to go into a planned maintenance mode. This may be done when, for example, an "isolate progressive" command is issued as part of the maintenance process. The router may keep sending PIM hello messages with the flag '02' set until it is upgraded and reloaded.

Under a second step, in contrast to a normal GIR process, PIM will not bring down PIM neighborhsip and unicast communications will not withdraw all of the routes because that would trigger an RPF change in the downstream and upstream routers and create a traffic loss. The PIM neighbors may mark the interfaces toward the instant router (i.e., the device that is scheduled to go under maintenance) with a flag such as, for example, "LIMITED USE." When the neighbors have appropriately marked those interfaces, nothing will change for the existing flows. Importantly, no RPF change has to be completed.

During a third step, existing devices will keep using the same path traverse a given device that is scheduled to for maintenance. However, new flows on the neighboring switches will not choose any interface that is marked as “LIMITED USE” as an RPF interface. This means that all of the new flows will never traverse the given device that is to be upgraded or reloaded. Eventually, all of the existing or old flows will end and will expire when the source stops (at, for example, the end of a show or at the end of a feed). Since no new flow can be stitched through the device, after some period of time, the device will eventually be free or empty of multicast flows due to the above-described starvation mechanism. When this happens, the device may be easily upgraded or reloaded with no traffic impact to any existing flow.

Under a fourth step, when the given device becomes operational again, the old multicast flows will continue to use the paths that they were using before the GIR router came back up (through, for example, a resilient hashing option for switches and routers).

However, new flows, given that all of the interfaces toward this device will no longer be marked as "LIMITED USE," can begin using it as a viable path and, consequently, insertion will be hitless as well.

Use of the techniques presented herein may offer a number of benefits. For example, in the case of media customers, especially under an IPFM solution, the above-described infrastructure may be expanded and made much more intelligent. In a capability-aware multicast solution, such as a bandwidth-aware multicast environment, currently a flow is typically provisioned as requiring a configuration of flow capabilities. In the case of an IPFM solution, flow policies may be employed to configure flow parameters for a given multicast flow. One possibility encompasses the addition of two new parameters – "START_TIME" and "DURATION" – to a stream. Such an addition would allow for the easy calculation of a time window during which a multicast flow will come and be alive on the switches.

In some instances, a command-line interface (CLI) may be extended to capture the date and the time at which an upgrade will occur. That information may then be relayed to all of a device's neighbors through another PIM hello TLV Option Type and Option Value pairing (similar to the approach that was presented above regarding Figure 1).

After the neighboring switches have both the information related to the exact time of a reload and the time window during which a multicast flow will be on the air, they can continue to use the interfaces through the instant GIR switch (that is scheduled to go under maintenance) if a flow ends before the reload time. When the reload time arrives, the GIR router will be empty and a reload may occur without any manual intervention, allowing for the creation of a completely automated upgrade procedure.

In summary, techniques have been presented herein that support a hitless upgrade capability that avoids impacting any flow during the upgrade and reload of a spine switch. This is the most difficult task today for customer networks, where multicast flows run all of the time and no disruption is acceptable. Aspects of the presented techniques include a new PIM hello message TLV option. Such an option may be referred to herein as a PGIR capability option.