Overview

- Equal Cost Multi-path (ECMP) provides the ability to load-share traffic across multiple next-hops.
- When a next-hop fails or is deleted all flows are affected. This is due to the nature of the load-balancing algorithm which re-calculates a new hash for the flows based on the remaining active next-hops.
- Resilient ECMP keeps the total number of next-hops same by replacing the deleted next-hop with one or more of the remaining active next-hops and maintain the order of next-hops in the ECMP groupset.
- This feature is currently supported on the 7050, 7160, 7280, 73xx, 75xx platforms

Benefits

- When a next-hop fails, only the flows forward to the failed next-hop will be affected. The flows that were being forwarded to the other active next-hops will not be affected by the failure.
- Flows which would otherwise be matched to the failed/deleted next hop will be fairly divided across the remaining active next-hops.
- When adding a new next-hop only the flows matching the new next-hop will be affected.

Configuration

Resilient ECMP is configured by specifying a prefix, mask, capacity value and redundancy value like,

```
ip hardware fib ecmp resilience <prefix/mask> capacity
 <value> redundancy <value>
```

The **<prefix/mask>** can match to a specific network, or can be a “parent” range, from which specific routes are gleaned. For example, using 50.10.1.1/32 specifies only that one host route, while 50.10.0.0/16 would pick up that network, and any other more specific routes that fall within its boundaries.

The **capacity** <value> should match the maximum number of next-hops expected to be used by the network.

The **redundancy** <value> is a multiplier. Multiplied by the **capacity**, this produces the number of entries in the ECMP resiliency table for the **prefix** in question.

We generate a log message ROUTING_ECMP_ROUTE_MAX_PATHS_EXCEEDED if a prefix exceeds **capacity** next-hops and ignore the excess next-hops.

Example

The following example is applicable to all supported platforms. The platform show command and output format are a little different across the supported platforms. The example below is illustrated on the 7160 platform.

We have a prefix 50.10.0.0/24 pointing to 4 different equal-cost next-hops:

```
B E    50.10.0.0/24 [200/0] via 192.1.1.1, Vlan2001
                            via 192.1.1.3, Vlan2002
                            via 192.1.1.5, Vlan2003
                            via 192.1.1.7, Vlan2004
```

This gives us a standard 4-way ECMP table for this prefix. This can be seen by running the following command:

```
Switch#show platform xp ip route | grep 50.10.0.0/24
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2003|  44:4c:a8:2f:cc:b
5|   1659|     0|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2001|  44:4c:a8:2f:cc:b
5|   1660|     1|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2004|  44:4c:a8:2f:cc:b
5|   1661|     2|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2002|  44:4c:a8:2f:cc:b
5|   1662|     3|
```

We configure the resilience entry with the following command:

```
Switch(conf)#
ip hardware fib ecmp resilience 50.10.0.0/24 capacity 4 redundancy 2
```

This creates an 8-entry ECMP table (4×2) for the 50.10.0.0/24 prefix:

```
Switch#show platform xp ip route | grep 50.10.0.0/24
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2003|  44:4c:a8:2f:cc:b
5|   1659|     0|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2001|  44:4c:a8:2f:cc:b
```

```
5|   1660|     1|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2004|  44:4c:a8:2f:cc:b
5|   1661|     2|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2002|  44:4c:a8:2f:cc:b
5|   1662|     3|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2003|  44:4c:a8:2f:cc:b
5|   1663|     4|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2001|  44:4c:a8:2f:cc:b
5|   1664|     5|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2004|  44:4c:a8:2f:cc:b
5|   1665|     6|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2002|  44:4c:a8:2f:cc:b
5|   1666|     7|
```

If we then remove one of the next-hops (in this case, the one pointing to 192.1.1.5, VLAN-2003), the table will remain constant with 8 entries, and the flows which would otherwise be matched to the VLAN-2003 entry will be fairly divided across the remaining active next-hops (next-hop VLAN-2003 is replaced with the VLAN-2001 entry in the first slot and with the VLAN-2004 entry in the fifth slot):

```
B E    50.10.0.0/24 [200/0] via 192.1.1.1, Vlan2001
                            via 192.1.1.3, Vlan2002
                            via 192.1.1.7, Vlan2004

Switch#show platform xp ip route | grep 50.10.0.0/24
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2001|  44:4c:a8:2f:cc:b
5|   1659|     0|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2001|  44:4c:a8:2f:cc:b
5|   1660|     1|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2004|  44:4c:a8:2f:cc:b
5|   1661|     2|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2002|  44:4c:a8:2f:cc:b
5|   1662|     3|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2004|  44:4c:a8:2f:cc:b
5|   1663|     4|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2001|  44:4c:a8:2f:cc:b
5|   1664|     5|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2004|  44:4c:a8:2f:cc:b
5|   1665|     6|
|    2|      50.10.0.0/24|  FWD|  Ethernet5/1| 2002|  44:4c:a8:2f:cc:b
5|   1666|     7|
```

Syslog messages

As of EOS-4.24.2F, we syslog the following message when the cumulative total number of next-hops for all the resilient routes under a given resilient parent prefix throughout its lifetime exceeds the configured capacity of the resilient parent prefix:

```
%ROUTING-6-RECMP_CAPACITY_EXCEEDED: Total next hop count (5) under REC
MP parent prefix 50.10.0.0/24 has exceeded the parent prefix's configu
red capacity
```

Note that the resilient ECMP feature will still work when this condition is hit. It only has implications for resilient ECMP FEC sharing, wherein resilient ECMP FECs that share the same set of next-hops in steady state may now not share the same resilient ECMP FEC in hardware. The workaround for this is to configure the capacity for the resilient ECMP parent prefix to the one specified in the syslog message.

Caveats

- While the removal of an ECMP next-hop with resiliency enabled allows traffic to maintain the path to which it is hashed, adding an additional ECMP next-hop will disrupt the flows hashed to a given slot since the hash tables are recomputed. Existing next-hops on other slots are unaffected.
- **Capacity** value range is **<2 – Max Value>** (the Max Value varies per platform), and the **Redundancy** value range is <1-63>.
    - For example, the 7160 platform has ECMP table that maxes out at 127 entries, so no value of **(Capacity*Redundancy)** that is greater than 127 will be accepted.
- As noted, the configuration of **<prefix/mask>** provides the flexibility of assignment which can let one specify a single network or a range of subnets whose boundaries are contained within the **<prefix/mask>** range. Assigning an especially wide range (e.g. 0/0) will prompt all ECMP routes to have resiliency tables created, which can drastically increase the amount of memory used for ECMP routes. Therefore, use this form of configuration with care.