LACP Fallback Feature for SONiC

Introduction 2

Overview 2

Requirements 2

Assumptions 3

Limitations 3

Background 4

LACP Fallback Design 6

LAG with single member port 6

*LAG with multiple member ports* 7

LACP Fallback Config 9

JSON Config 9

Minigraph Config 9

CLI for LACP fallback [TBD] 9

Test Plan [TBD] 10

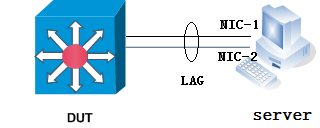
References 10

# Introduction

## Overview

The LACP Fallback Feature allows an active LACP interface to establish a Link Aggregation (LAG) before it receives LACP PDUs from its peer.

This feature is useful in environments where customers have Preboot Execution Environment (PXE) Servers connected with a LACP Port Channel to the switch.  Since PXE images are very small, many operating systems are unable to leverage LACP during the preboot process.  The server’s NICs do not have the capability to run LACP without the assistance of a fully functional OS; during the PXE process, they are unaware of the other NIC and don’t have a method to form a LACP connection. Both the NIC’s on the server will be active and are sourcing frames from their respective MAC addresses during the initial boot process.  Simply keeping both ports in the LAG active will not solve the problem because packets sourced from the MAC address of NIC-1 can be returned to the port on which NIC-2 is attached, which will cause NIC-2 to drop the packets (due to MAC mismatch).



With the LACP fallback feature, the switch allows the server to bring up the LAG (before receiving any LACP PDUs from the server) and keeps a single port active until it receive the LACP PDUs from the server. This allows the PXE boot server to establish a connection over one Ethernet port, download its boot image and then continue the booting process. When the server boot process is complete, the server fully forms an LACP port-channel.

## Requirements

1. LACP fallback feature can be enabled / disabled per LAG.
2. LACP fallback timer is configurable. This timer specifies the wait time of a LAG changed from defaulted state to fallback state, without receiving an LACP PDU from its peer.  The default fallback timeout period is 3 seconds.
3. When there are multiple ports present in the LAG in fallback mode, one and only one port will be selected as active
4. LACP port priority determines the port that is active in a LAG in fallback mode. This command is optional, but gives a deterministic way to allocate the port which will be active.
5. The LAG will be moved out of the fallback state if it receives LACP PDU from its peer.
6. Interoperability with other devices running standard 802.3ad LACP protocol.
7. User can show the current status of the LAG, whether it’s in LACP fallback mode or not.
8. The LACP runner behavior is not changed if fallback feature is disabled

## Assumptions

1. The LACP fallback feature is implemented on top of the open source libteam (<https://github.com/jpirko/libteam>) adopted by SONiC
2. The LAG and member ports need to be shutdown to enable/disable the LACP fallback feature.
3. The changes are restricted to the libteam library only, the APP DB/SAI DB is not aware of the fallback state.

## Limitations

1. Currently can only support LAG with only 1 member port (phase 1).
2. LACP fallback mode may also kick in during the normal LACP formation process due to the timing, which might cause some unexpected traffic loss. For example, if the LACP PDUs sent by peer are dropped completely, local LACP may still enter fallback mode, which might end up with data traffic drop.

# Background

LACP fallback feature is implemented on the receiver side to establish a LAG before it receives LACP PDUs from its peer. So this section presents a formal description of the standard LACP Receive Machine.

**Receive Machine States and Timer**

The receive machine has four states:

• Rxm\_current

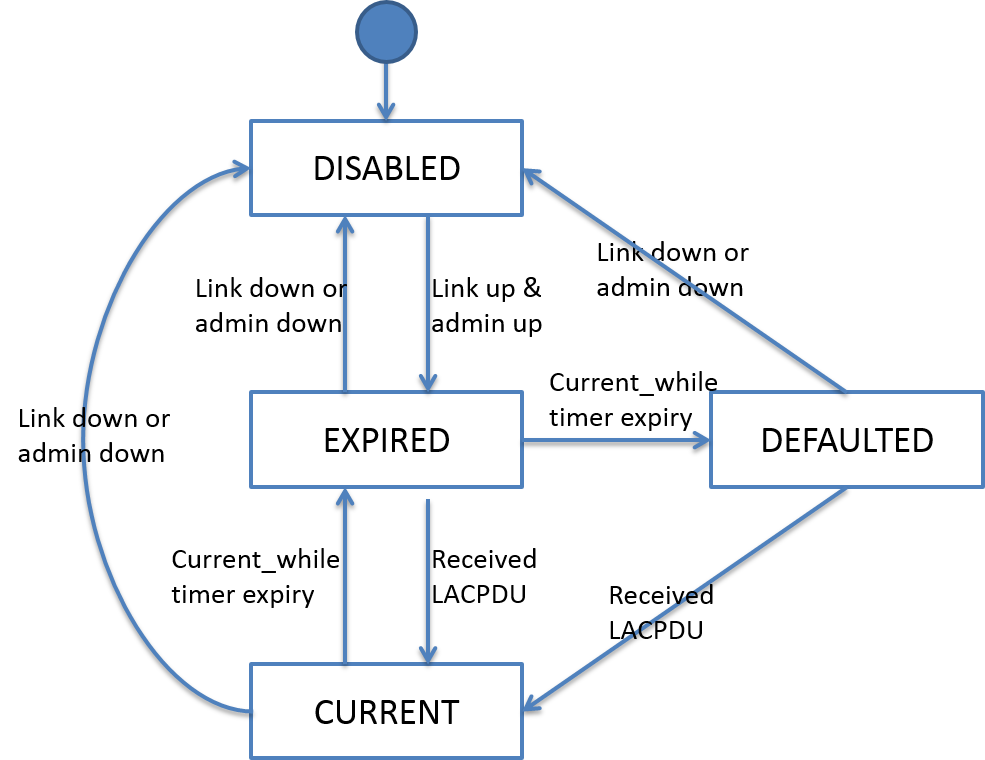
• Rxm\_expired

• Rxm\_defaulted

• Rxm\_disabled

One timer:

Current while timer that is started in the Rxm\_current and Rxm\_expired states with two timeout: Short timeout (3s) and Long timeout (180s) depending on the value of the Actor’s Operational Status LACP\_Timeout, as transmitted in LACPDUs.



**Receive Machine Events**

The following events can occur**:**

• Participant created or reinitialized

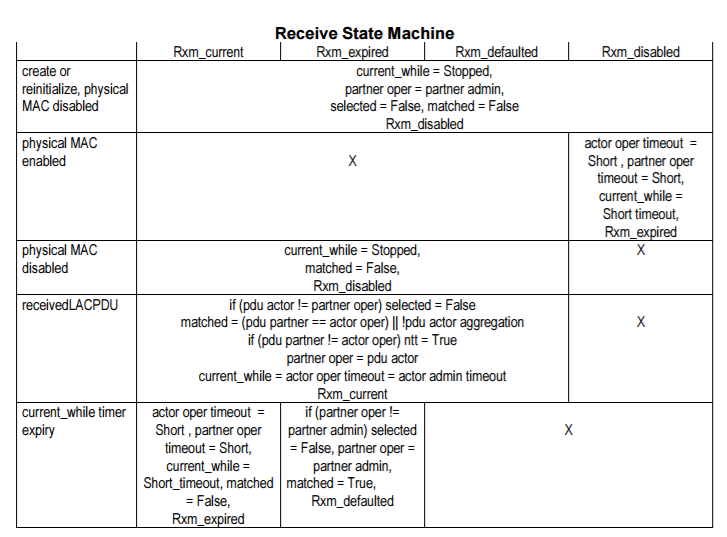
• Received LACP PDU

• Physical MAC enabled

• Physical MAC disabled

• Current while timer expiry

The physical MAC disabled event indicates that either or both of the physical MAC transmission or reception for the physical port associated with the actor have become non-operational. The received LACPDU event only occurs if both physical transmission and reception are operational, so far as the actor is aware.

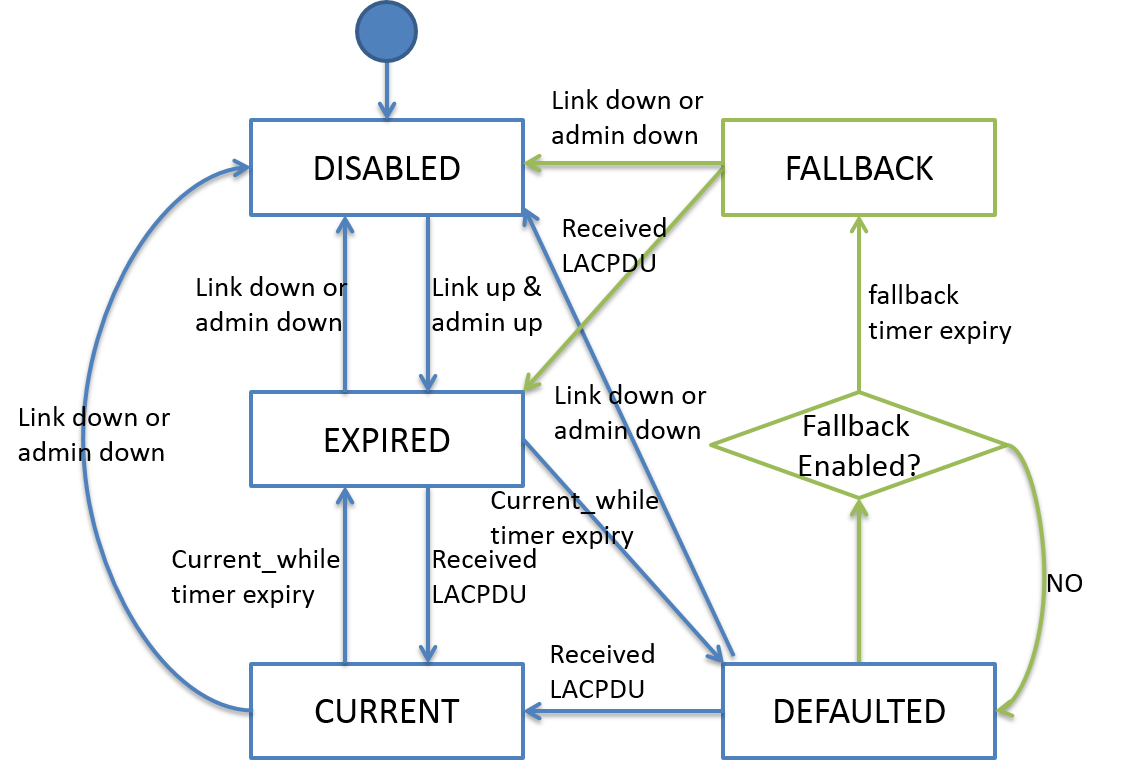


# LACP Fallback Design

The development will be divided in to two phases. In phase 1, we will focus on developing and testing the fallback feature support for LAG with single member port. In phase 2, we shall support fallback feature for LAG with multiple member ports, which also takes care of the port selection and link switchover logic in this phase.

## LAG with single member port

To support LACP fallback mode for LAG with single member port, a new state rxm\_fallback is being introduced into current LACP state machine implementation. Also a new timer fallback\_timer is being added, which specifies the wait time of a LAG changed from defaulted state to fallback state, without receiving any LACP PDU from its peer over the LAG.  The default fallback timeout period is 3 seconds.



Fallback state:

In this state, both the current while timer and fallback timer will be turned off, so there is no timeout event. And the port selected bit is being set, which means the port is selectable and can be aggregated into the LAG. If any LACP PDU is being received over the LAG during this state, the port will be moved to expired state, and restarts the LACP negotiation with peer.

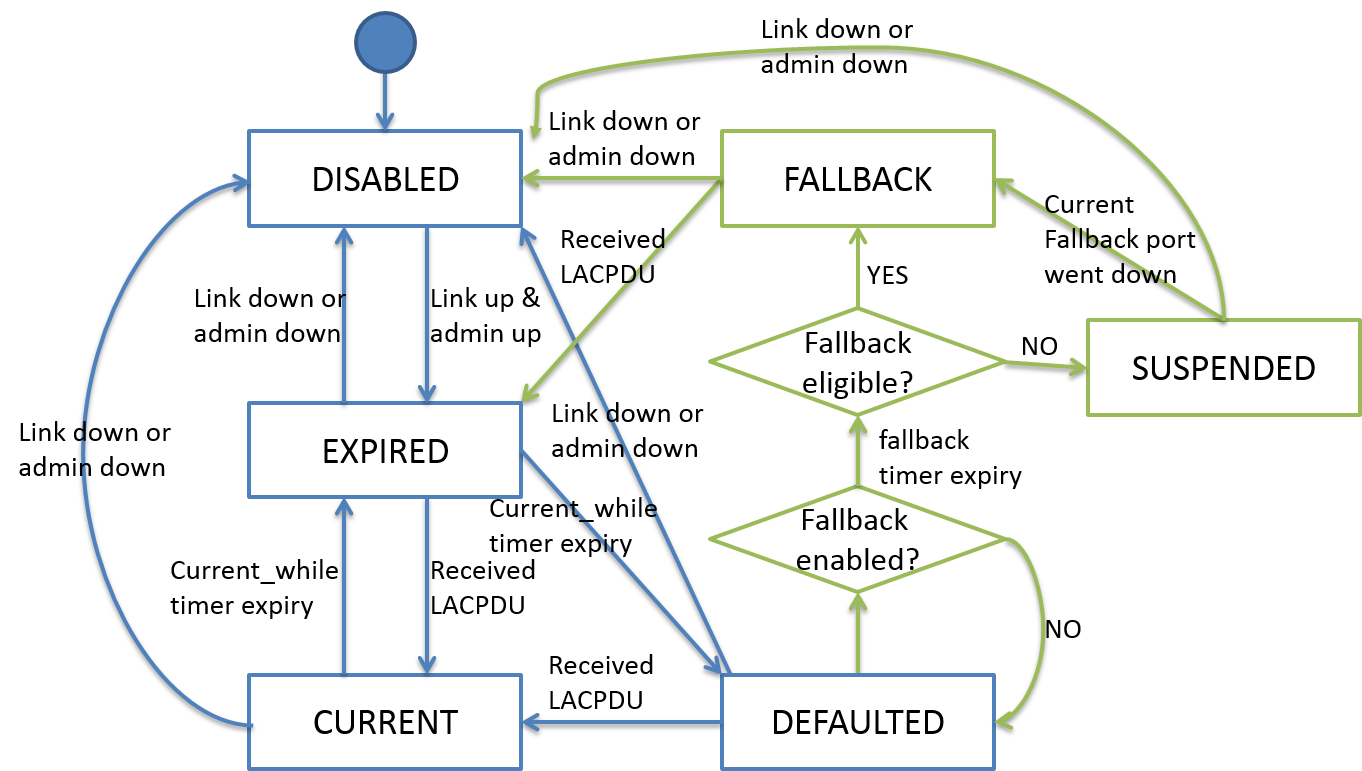
Fallback Enabled:

This checks whether LACP fallback is enabled on this LAG. Only when fallback config is enabled the fallback timer will be started.

Fallback Timer:

This timer will be started if fallback is enabled on this LAG, and the port enters defaulted state due to not receiving LACP PDUs from peer.

## LAG with multiple member ports



To support LACP fallback mode for LAG with multiple member port, state rxm\_suspended, port selection, and port switchover logic need to be implemented in addition to phase 1 development.

Suspended state:

When a port is being selected as fallback port, the rest ports in the fallback LAG will be suspended. In this state, the MAC needs to be disabled (the physical link is being kept down) for the suspended ports. So the remote peer won’t use the suspended ports to send/receive any data traffic over the LAG. Otherwise, it will cause the traffic being blackholed if the peer send the traffic over the port not being selected as fallback.

Port Selection:

The following conditions needs to be met before selecting a port.

1. There is currently no port in fallback/current mode within the same LAG.
2. The port has the won the election :

When there are multiple ports present in the LAG in fallback mode, one and only one port will be selected as fallback port.

* 1. To LACP port priority determines the port that is active in a LAG in fallback mode. The port with highest priority with be selected as fallback port in this LAG.
  2. If all ports in a port-channel have the same port-priority (default), the switch has an internal algorithm that compares interface names (Ethernet 1, Ethernet 2, etc.) in order to break the tie. The algorithm compares two interface names by base name, then by module number (if any) and finally by port number (if any) and selects the port that is lowest in this ordering; i.e. Ethernet1 < Ethernet2, hence we choose Ethernet1 to become active.

Port switchover and recovery:

We need also address the following scenarios may happen during the fallback state.

1. The fallback port is disabled due to link failure/admin shutdown. In this case, the port with highest priority among the suspended port will be prompted as fallback port.
2. A new port comes up with higher priority than the port in fallback mode without receiving any LACP PDU from peer. In this case, we prefer to keep current fallback port active and the new port in suspended state to avoid any traffic interruptions on the current fallback port.
3. A new port is being added to the LAG in fallback mode. And the newly-added port receives LACP PDU immediately before being suspended. As a result, the newly-added port will enter current state with LACP up and running. In this case, the ports in fallback/suspended state will be moving to expired state to restart LACP negotiation with peer.
4. When any LACP PDU is being received on the fallback port, the port will be moved into expired state, and start LACP negotiation with peer. Once the fallback port enters the current state with LACP up and running, all suspended ports will be moving out of suspended state into expired state.

# LACP Fallback Config

## JSON Config

teamd is configured using JSON config string. This can be passed to teamd either on the command line or in a file. JSON format was chosen because it's easy to specify (and parse) hierarchic configurations using it.

Example teamd config (teamd1.conf):

{

"device":"team0",

"runner":

{

"name":"lacp",

"active": true,

"fast\_rate": true,

"fallback": true,

"fallback\_timer": “3”,

"tx\_hash": ["eth", "ipv4"]

},

"link\_watch":{"name":"ethtool"},

"ports":

{

"Ethernet0":{}

}

}

## Minigraph Config

<PortChannelInterfaces>

<PortChannel>

<Name>PortChannel01</Name>

<AttachTo>Ethernet0</AttachTo>

<Fallback>true</Fallback>

<SubInterface/>

</PortChannel>

</PortChannelInterfaces>

## CLI for LACP fallback [TBD]

CLI configuration commands for LACP:

The following set of Show commands relevant for LACP will be supported:

Teamshow

Teamdctl teamdevname state

# Test Plan [TBD]

|  |  |
| --- | --- |
| Test Case | Verify LACP fallback feature with single member port LAG |
| Testbed Setup: |  |
| Steps: | 1. Connect DUT and server with single Ethernet port shown as above without any bundling/LACP configurations, and keep the port shutdown on both sides 2. On DUT side, add the port to port-channel1 with LACP Fallback feature on, and unshut both port and port-channel1 3. On server side, unshut the port. Check expected result 1 4. Configure the port-channel with ip address 10.1.1.2/24 on DUT side, and port with ip address 10.1.1.1/24 on server side. Ping 10.1.1.2 from server side. Expected result 2 5. On server side, add the port to port-channel1 with LACP running, and unshut both the port and port-channel1. Check expected result 3 6. Configure the port-channel with ip address 10.1.1.3/24 on server side. Ping 10.1.1.2 from server side. Check expected result 2 |
| Expected result: | 1. The port is being selected into port-channel1 on DUT in fallback mode. 2. Ping works 3. The port is being aggregated into port-channel1 on both DUT and server side, and the LACP is formed successfully |
| Description: |  |
| Test result: |  |

# References

1. SONiC Configuration Management
2. Open Source libteam <https://github.com/jpirko/libteam>
3. IEEE 802.3ad Standard for LACP <http://www.ieee802.org/3/ad/public/mar99/seaman_1_0399.pdf>