1 Controlling Switch Redundancy Protocol

NOTE 1 – This document is written in terms of native Fibre Channel operations. The same operations apply also to FCoE, with physical FC links replaced by FCoE Virtual Links. Specific references to a protocol are provided when needed.

1.1 Definitions

1.1.1 A_Port (Adjacent Port): The combination of one PA_Port and one VA_Port operating together.

1.1.2 AISL (Augmented ISL): an E_Port to E_Port link used by the redundancy protocol.

1.1.3 AISL Set: The set of AISLs that connect the two Controlling Switches that are part of a Distributed Switch.

1.1.4 ASL (A_Port Switch Link): An A_Port to A_Port link.

1.1.5 Controlling Switch: A Switch able to control a set of FCDFs in order to create a Distributed Switch.

1.1.6 Controlling Switch Set: The Switch_Names of the up to two Controlling Switches that are part of a Distributed Switch.

1.1.7 Distributed Switch: A set of FCDFs associated with at least one Controlling Switch, that controls the operations of the set of FCDFs. A Distributed Switch is defined by the administrative configuration of the Controlling Switch Set and of the FCDF Set.

1.1.8 FCDF (FC Data-Plane Forwarder): A simplified FC switching entity that forwards FC frames among VA_Ports and VF_Ports through a FCDF Switching Element. An FCDF shall support at least one VA_Port operating together with a PA_Port (i.e., an A_Port) and may support one or more F_Ports.

1.1.9 FCDF Set: The Switch_Names of the FCDFs that are part of a Distributed Switch.

1.1.10 PA_Port (Physical A_Port): The LCF within the Fabric that attaches to another PA_Port through a link.

1.1.11 VA_Port (Virtual A_Port): An instance of the FC-2V sublevel of Fibre Channel that connects to another VA_Port. A VA_Port is uniquely identified by an A_Port_Name Name_Identifier and is addressable by the VA_Port connected to it through the A_Port Controller address identifier (i.e., FFFFF9h).

1.2 Overview

The purpose of the Controlling Switch Redundancy protocol is to avoid any single point of failure in a Distributed Switch. Figure 1 shows an example of redundant Distributed Switch.

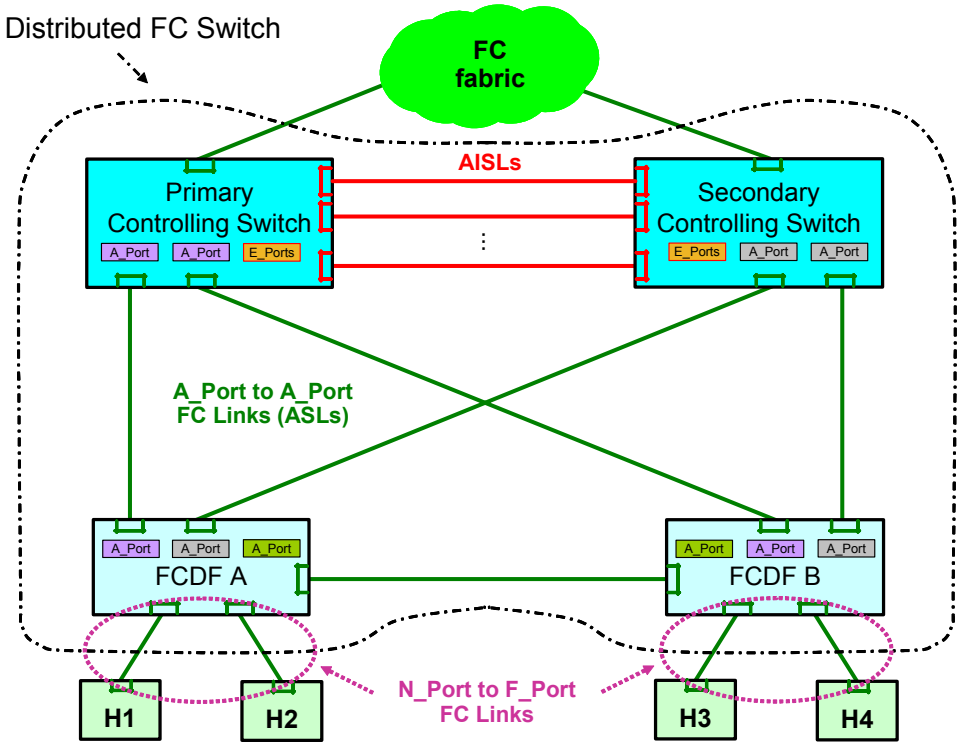

Figure 1 – Example of Redundant Distributed Switch

The Controlling Switch Redundancy protocol uses a set of Augmented E_Port to E_Port links (AISLs) between the Primary and Secondary Controlling Switches. This set is referred to as the AISL Set. There shall be at least two AISLs in the AISL Set, in order to distinguish the case of an AISL failure from the case of a Controlling Switch failure. Additional AISLs provide additional resiliency.

In a Redundant Distributed Switch the Primary Controlling Switch generates the LSR describing the Virtual Domain_ID in the Distributed Switch. In addition, both Primary and Secondary Controlling Switch list the Virtual Domain_ID as a directly attached Domain_ID in their LSR. The resulting FSPF topology is depicted in figure 2, where Z is the Virtual Domain_ID and X and Y are the Domain_IDs of the two Controlling Switches.

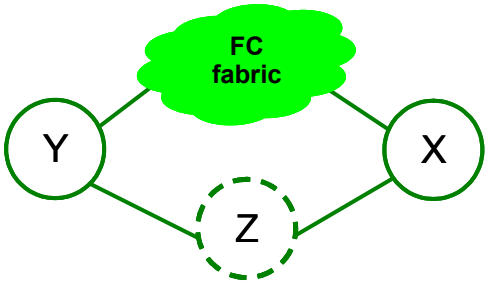

Figure 2 – Distributed Switch FSPF Topology

1.3 Redundancy Protocol State Machine

The redundancy protocol state machine uses the following time intervals and timers:

RHello_Interval: Time interval between RHellos, expressed in milliseconds. The default value is 400 ms.

Down_Interval: Time interval for a Controlling Switch to declare the other one down. Calculated as 2.5 * RHello_Interval.

ASL_Fail_Interval: Time interval for the Primary to relinquish its role after detecting an ASL failure, if appropriate. Calculated as 25 * RHello_Interval.

Down_Timer: timer that expires when RHellos have not been received for Down_Interval over any AISL.

To determine which Controlling Switch behaves as Primary and which one as Secondary, the redundancy protocol uses a Priority value associated to each Controlling Switch. Priority values are shown in table 1.

Table 1 – Controlling Switch Priority Values

Value	Description
00h	Reserved
01h	Highest Priority value. This value is administratively configured to force the election of a Controlling Switch to Primary.
02h[a]	Primary Controlling Switch priority. This value is used by the Redundancy protocol to identify a Controlling Switch as Primary.
03 .. FEh	Higher to lower Priority values. The default value is 128.
FFh[a]	This value indicates that a Controlling Switch is not willing to operate as Primary. This is used by the Primary Controlling Switch to trigger a transition of the Secondary Controlling Switch to Primary without having to wait for the current Primary to timeout, if appropriate.
[a] These values are used by the Redundancy protocol and not available to an administrator.	

Figure 3 shows the redundancy protocol state machine.

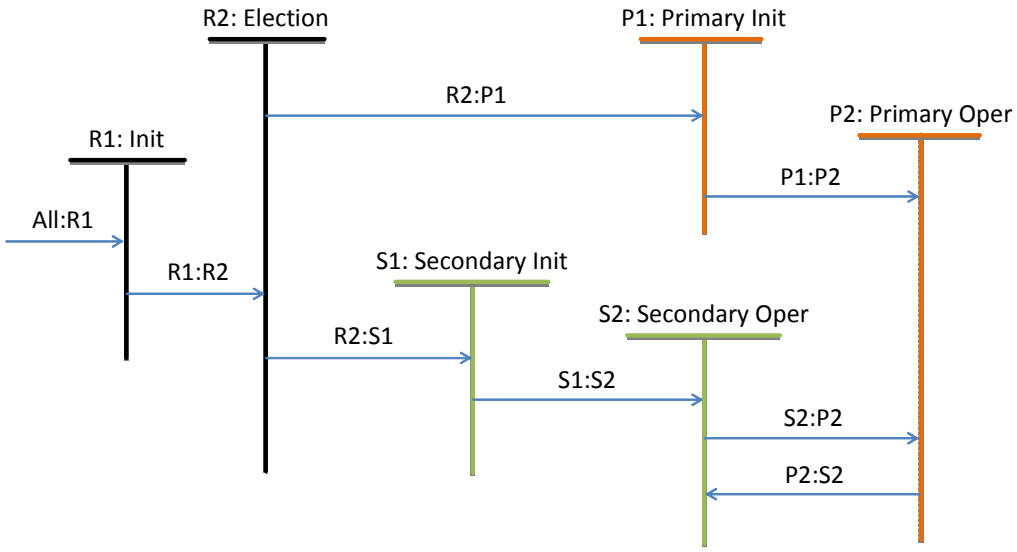

Figure 3 – Redundancy Protocol State Machine

State R1:Init. In this state a Controlling Switch waits to begin the processing for the redundancy protocol.

Transition R1:R2. Occurs when processing for the redundancy protocol begins. The redundancy protocol processing begins when:

a) the redundancy protocol is enabled;

b) the Controlling Switch Set and the FCDF Set are configured; and

c) Fabric configuration is completed.

Transition All:R1. Occurs when the redundancy protocol is disabled.

State R2:Election. In this state a Controlling Switch determines if it operates as Primary or Secondary. On entering this state the Controlling Switch sets the Down_Timer to Down_Interval. While in this state the Controlling Switch sends RHellos every RHello_Interval over each of its Augmented ISLs (AISLs) and waits to receive RHellos from at least one AISL. The Controlling Switch exits this state when:

a) the AISL Set is NULL;

b) the AISL Set is not NULL and Down_Timer expires without having received RHellos from any AISL; or

c) the AISL Set is not NULL and an RHello is received from an AISL.

If an RHello is received and the local Controlling Switch Priority is 01h and the remote Controlling Switch Priority is 01h (i.e., both Controlling Switches are manually configured to be Primary) then the Redundancy protocol is disabled and an error is logged.

Transition R2:P1. Occurs when:

a) the AISL Set is NULL;

b) the AISL Set is not NULL and Down_Timer expires without having received RHellos from any AISL;

c) the AISL Set is not NULL, an RHello is received from an AISL, and the local Controlling Switch Priority is lower than the remote Controlling Switch Priority; or

d) the AISL Set is not NULL, an RHello is received from an AISL, the local Controlling Switch Priority is equal to the remote Controlling Switch Priority, and the local Switch_Name is lower than the remote Switch_Name.

Transition R2:S1. Occurs when:

a) the AISL Set is not NULL, an RHello is received from an AISL, and the local Controlling Switch Priority is higher than the remote Controlling Switch Priority; or

b) the AISL Set is not NULL, an RHello is received from an AISL, the local Controlling Switch Priority is equal to the remote Controlling Switch Priority, and the local Switch_Name is greater than the remote Switch_Name.

State P1:Primary Initialization. In this state a Controlling Switch performs the operations to become the Primary Controlling Switch of the Distributed Switch. To this end the Controlling Switch sets its Priority to 02h and obtains an additional Domain_ID value (the Virtual Domain_ID) from the Principal Switch of the fabric by generating an RDI Request (see FC-SW-5) on behalf of the Distributed Switch Switch_Name. While in this state the Controlling Switch sends RHellos every RHello_Interval over each of its AISLs, if any.

Transition P1:P2. Occurs when the Virtual Domain_ID is available.

State P2:Primary Operational. In this state the Controlling Switch is operational as Primary. On entering this state the Controlling Switch:

a) sets its Priority to 02h;

b) sets the Down_Timer to Down_Interval;

c) sends a DFMD SW_ILS to all reachable FCDF of the FCDF Set declaring itself as Primary Controlling Switch;

d) on native Fibre Channel links that were Isolated because connected to FCDFs, if any, it performs an ELP; and

e) on FCoE interfaces, it establishes VA_Port to VA_Port Virtual Links with neighbor FDFs belonging to the FDF Set to which no VA_Port to VA_Port Virtual Links has been established, if any.

While in this state, the Controlling Switch:

a) performs the VA_Port Protocols (see T11/11-225v0);

b) sends RHellos every RHello_Interval over each of its AISLs;

c) resets the Down_Timer to Down_Interval everytime an RHello is received over at least one AISL;

d) on receiving an SSA SW_ILS (i.e., when the Secondary Controlling Switch completed its state synchronization) sends a DFMD SW_ILS to all reachable FCDFs of the FCDF Set declaring itself as Primary and the Secondary as Secondary; and

e) when Down_Timer expires (i.e., when the Secondary Controlling Switch is not anymore available) sends a DFMD SW_ILS to all reachable FCDFs of the FCDF Set declaring itself as Primary.

State S1:Secondary Initialization. In this state a Controlling Switch performs the operations to become the Secondary Controlling Switch of the Distributed Switch. The Controlling Switch has to synchronize its state with the one of the Primary Controlling Switch. To this end the Controlling Switch:

1) Requests to the Primary the Virtual Domain_ID and the FCDF topology through the GFSS (Get FCDF Set Status) SW_ILS;

2) Requests to the Primary the N_Port_ID Allocation state in the Distributed Switch through the GNAS (Get N_Port_ID Allocation State) SW_ILS;

3) Obtains the information associated with each N_Port_ID in the Name Server through the GE_ID CT Request; and

4) Communicates the achieved state synchronization to the Primary through the SSA (Secondary Synchronization Achieved) SW_ILS.

While in this state, the Controlling Switch:

a) sends RHellos every RHello_Interval over each of its AISLs; and

b) processes possible FDUN, FDRN, and NPZD Requests coming from the Primary.

Transition S1:S2. Occurs when the Secondary Controlling Switch has synchronized its state with the Primary.

State S2:Secondary Operational. In this state the Controlling Switch is operational as Secondary. On entering this state the Controlling Switch:

a) sets its Priority to its configured value;

b) sends RHellos every RHello_Interval over each of its AISLs;

c) on native Fibre Channel links that were Isolated because connected to FCDFs, if any, it performs an ELP; and

d) on FCoE interfaces, it establishes VA_Port to VA_Port Virtual Links with neighbor FDFs belonging to the FDF Set to which no VA_Port to VA_Port Virtual Links has been established, if any.

While in this state, the Secondary Controlling Switch:

a) performs the VA_Port Protocols (see T11/11-225v0); and

b) resets the Down_Timer to Down_Interval everytime an RHello is received over at least one AISL;

Transition S2:P2. Occurs when the Secondary Controlling Switch becomes Primary. This occurs when:

a) The Down_Timer expires (i.e., no RHellos have been received over any AISL for Down_Interval). This indicates the failure of the Primary Controlling Switch; or

b) The Priority field in the received RHellos has a value of FFh. This is an indication that the Primary Controlling Switch determined to become Secondary.

Transition P2:S2. Occurs when the Primary Controlling Switch determines to become Secondary by setting its Priority to FFh. This may happen when the Primary determines that the number of FCDFs reachable from the Secondary is greater than the number of FCDFs reachable from the Primary (e.g., after the failure of an ASL with the Primary). The Primary should set its Priority to FFh after ASL_Fail_Interval, in order to avoid unnecessary role transitions in presence of a flapping link.

1.4 Redundancy Protocol Messages

1.4.1 Redundancy Hello (RHello)

The Redundancy Hello (RHello) SW_ILS is used by the redundancy protocol as described in 1.3. The Rhello SW_ILS is transmitted in a unidirectional Exchange (i.e., it does not have an Reply Sequence).

RHello Request Sequence

Addressing: the S_ID field shall be set to FFFFFDh, indicating the originating VE_Port, and the D_ID field shall be set to FFFFFDh, indicating the destination VE_Port.

Payload: The format of the RHello Request Sequence Payload is shown in table 2.

Table 2 – RHello Request Payload

Item	Size (bytes)
SW_ILS Code	4
Originating Controlling Switch Switch_Name	8
RHello_Interval	4
Reserved	3
Originating Controlling Switch Priority	1

Originating Controlling Switch Switch_Name: contains the Switch_Name of the originating Controlling Switch.

RHello_Interval: contains the RHello_Interval value expressed in ms.

Originating Controlling Switch Priority: contains the operational Priority of the originating Controlling Switch (see table 1).

1.4.2 Get FCDF Set Status (GFSS)

The Get FCDF Set Status (GFSS) SW_ILS is used by the Secondary Controlling Switch to request to the Primary the Virtual Domain_ID value and the current FCDF Set topology, in order to synchronize its state with the one of the Primary.

GFSS Request Sequence

Addressing: the S_ID field shall be set to FFFFFDh, indicating the originating VE_Port, and the D_ID field shall be set to FFFFFDh, indicating the destination VE_Port.

Payload: The format of the GFSS Request Sequence Payload is shown in table 3.

Table 3 – GFSS Request Payload

Item	Size (bytes)
SW_ILS Code	4
Originating Controlling Switch Switch_Name	8

Originating Controlling Switch Switch_Name: contains the Switch_Name of the originating Controlling Switch.

GFSS Reply Sequence

SW_ACC: SW_ACC indicates the acceptance of the GFSS Request Sequence for processing. The format of the GFSS SW_ACC Payload is shown in table 4.

Table 4 – GFSS SW_ACC Payload

Item	Size (bytes)
SW_ILS Code = 0200 0000h	4
Originating Controlling Switch Switch_Name	8
Reserved	3
Virtual Domain_ID Value	1
Distributed Switch Switch_Name	8
Number of FCDF Connectivity Records (n)	4
FCDF Connectivity Record #1	
FCDF Connectivity Record #2	
...	
FCDF Connectivity Record #n	

Originating Controlling Switch Switch_Name: contains the Switch_Name of the originating Controlling Switch.

Virtual Domain_ID Value: contains the Virtual Domain_ID for the Distributed Switch.

Distributed Switch Switch_Name: contains the Switch_Name for the Distributed Switch.

Number of FCDF Connectivity Records: contains the number of FCDF Connectivity Records that follow. The format of the FCDF Connectivity Record is shown in table 5.

Table 5 – FCDF Connectivity Record Format

Item	Size (bytes)
FCDF Switch_Name	8
Number of ASL Neighbors (m)	4
Switch_Name of Neighbor #1	8
ASL cost to Neighbor #1	4
Switch_Name of Neighbor #2	8
ASL cost to Neighbor #2	4
...	
Switch_Name of Neighbor #m	8
ASL cost to Neighbor #m	4

FCDF Switch_Name: contains the Switch_Name of the FCDF being described.

Number of ASL Neighbors: contains the number of ASLs instantiated by the FCDF.

Switch_Name of Neighbor: contains the Switch_Name of the FCDF or Controlling Switch at the other end of the described ASL.

ASL cost to Neighbor: contains the link cost of the described ASL.

1.4.3 Get N_Port_ID Allocation State (GNAS)

The Get N_Port_ID Allocation State (GNAS) SW_ILS is used by the Secondary Controlling Switch to request to the Primary the current allocation of N_Port_IDs to each FCDF of the FCDF Set, in order to synchronize its state with the one of the Primary.

GNAS Request Sequence

Addressing: the S_ID field shall be set to FFFFFDh, indicating the originating VE_Port, and the D_ID field shall be set to FFFFFDh, indicating the destination VE_Port.

Payload: The format of the GNAS Request Sequence Payload is shown in table 6.

Table 6 – GNAS Request Payload

Item	Size (bytes)
SW_ILS Code	4
Originating Controlling Switch Switch_Name	8
FCDF Switch_Name	8

Originating Controlling Switch Switch_Name: contains the Switch_Name of the originating Controlling Switch.

FCDF Switch_Name: contains the Switch_Name of the FCDF whose N_Port_IDs allocation is requested.

GNAS Reply Sequence

SW_ACC: SW_ACC indicates the acceptance of the GNAS Request Sequence for processing. The format of the GNAS SW_ACC Payload is shown in table 7.

Table 7 – GNAS SW_ACC Payload

Item	Size (bytes)
SW_ILS Code = 0200 0000h	4
Originating Controlling Switch Switch_Name	8
FCDF Switch_Name	8
Number of Allocated N_Port_IDs (q)	4
Allocated N_Port_ID #1	4
Allocated N_Port_ID #2	4
...	
Allocated N_Port_ID #q	4

Originating Controlling Switch Switch_Name: contains the Switch_Name of the originating Controlling Switch.

FCDF Switch_Name: contains the Switch_Name of the FCDF whose N_Port_IDs allocation is provided.

Number of Allocated N_Port_IDs: contains the number of allocated N_Port_IDs that follow.

Allocated N_Port_ID: contains a reserved byte in the most significant byte and an N_Port_ID in the least significant bytes.

1.4.4 Secondary Synchronization Achieved (SSA)

The Secondary Synchronization Achieved (SSA) SW_ILS is used by the Secondary Controlling Switch to communicate to the Primary that it achieved state synchronization.

SSA Request Sequence

Addressing: the S_ID field shall be set to FFFFFDh, indicating the originating VE_Port, and the D_ID field shall be set to FFFFFDh, indicating the destination VE_Port.

Payload: The format of the SSA Request Sequence Payload is shown in table 8.

Table 8 – SSA Request Payload

Item	Size (bytes)
SW_ILS Code	4
Originating Controlling Switch Switch_Name	8

Originating Controlling Switch Switch_Name: contains the Switch_Name of the originating Controlling Switch.

SSA Reply Sequence

SW_ACC: SW_ACC indicates the acceptance of the SSA Request Sequence for processing. The format of the SSA SW_ACC Payload is shown in table 9.

Table 9 – SSA SW_ACC Payload

Item	Size (bytes)
SW_ILS Code = 0200 0000h	4
Originating Controlling Switch Switch_Name	8

Originating Controlling Switch Switch_Name: contains the Switch_Name of the originating Controlling Switch.