

*Claudio DeSanti (Cisco), Fred Knight (NetApp),
Craig Carlson (QLogic), Ed McGlaughlin (QLogic)*

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. See T11/11-223v1 for the terminology used in this document.

1.1 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, including the two Principal Domains and the Virtual Domain.

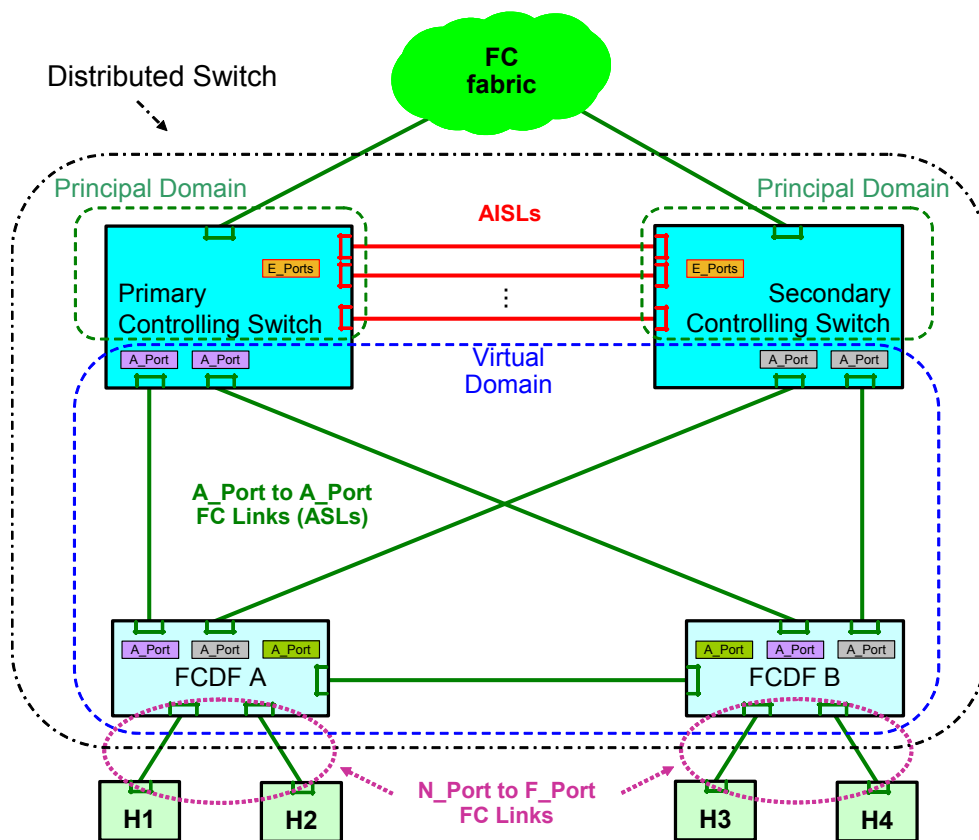


Figure 1 – Example of Redundant Distributed Switch

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 in the Distributed Switch. In addition, both Primary and Secondary Controlling Switch list the Virtual Domain as a directly attached Domain in their LSR. The resulting FSPF topology is de-

picted in figure 2, where Z is the Domain_ID of the Virtual Domain and X and Y are the Domain_IDs of the Principal Domains of the two Controlling Switches. X and Y result also connected between themselves by virtue of the AISLs.

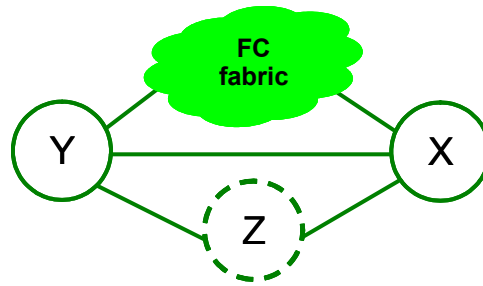


Figure 2 – Distributed Switch FSPF Topology

1.2 Redundancy Protocol State Machine

The redundancy protocol state machine reacts to AISLs failures in a timed fashion.

When AISLs are FCoE Virtual Links (i.e., they are Augmented VE_Port to VE_Port Virtual Links), the redundancy protocol state machine relies on the Virtual Link maintenance protocol to determine if a Virtual Link failed. To achieve a timely response to a Virtual AISL failure the default value for the FKA_ADV_PERIOD should be 1 000 ms for a Controlling FCF.

When AISLs are native FC links, the redundancy protocol state machine relies on indications from the physical layer to determine if a link failed. To avoid hyper-reacting to a flapping link, the redundancy protocol state machine reacts to a link failure after a Down_Interval timeout to a native FC link failure. To ensure a consistent behavior, the Down_Interval is computed as $2.5 * FKA_ADV_PERIOD$.

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.

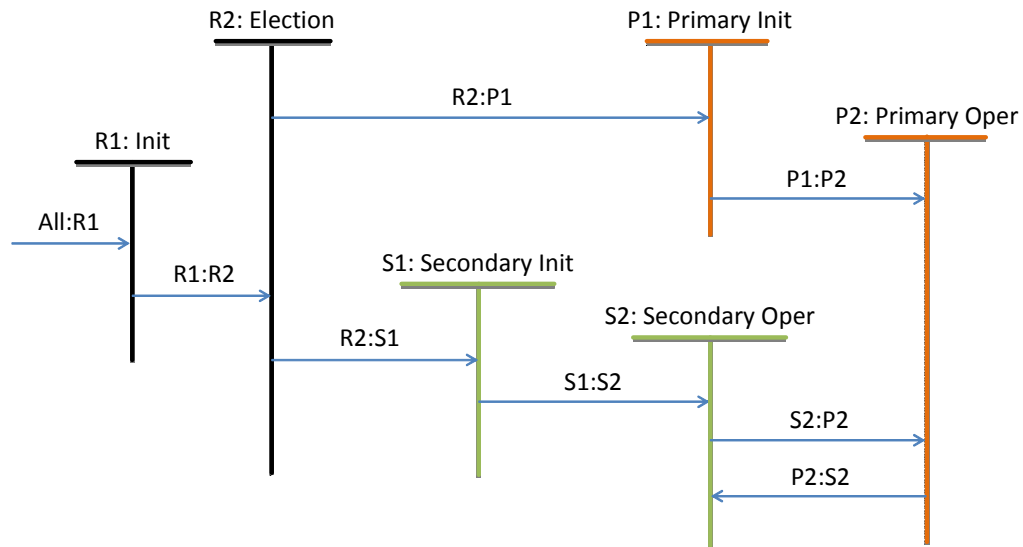


Figure 3 – Redundancy Protocol State Machine

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. If the AISL Set is NULL, then the Controlling Switch exits this state. If the AISL Set is not NULL, then an ERP Exchange is performed.

If the ERP Exchange shows that 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 the ERP Exchange showed that the local Controlling Switch Priority is lower than the remote Controlling Switch Priority; or

- c) the AISL Set is not NULL, the ERP Exchange showed that 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 and the ERP Exchange showed that the local Controlling Switch Priority is higher than the remote Controlling Switch Priority; or
- b) the AISL Set is not NULL, the ERP Exchange showed that 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 Virtual Domain Switch_Name.

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) initiates an ERP Exchange with the Secondary Controlling Switch, if available;
- 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 T11/11-225v1);
- b) 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;
- c) when the Secondary Controlling Switch is not anymore available (i.e., when all Virtual AISLs are down and when all native AISLs are down for more than Down_Interval) 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 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) initiates an ERP Exchange with the Primary Controlling Switch;
- 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 performs the VA_Port Protocols (see T11/11-225v1).

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

- a) when the Primary Controlling Switch is not anymore available (i.e., when all Virtual AISLs are down and when all native AISLs are down for more than Down_Interval); or
- b) The Priority field in a received ERP Request 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 as result of an administrative action.

1.3 Redundancy Protocol Messages

1.3.1 Exchange Redundancy Parameters (ERP)

The Exchange Redundancy Parameter (ERP) SW_ILS is used by the redundancy protocol to determine which Controlling Switch behaves as Primary and which one behaves as Secondary.

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 ERP Request Sequence Payload is shown in table 2.

Table 2 – ERP Request Payload

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

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

FKA_ADV_PERIOD: contains the FKA_ADV_PERIOD value expressed in ms.

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

ERP Reply Sequence

SW_ACC: SW_ACC indicates the acceptance of the ERP Request Sequence for processing. The format of the ERP SW_ACC Payload is shown in table 3.

Table 3 – ERP SW_ACC Payload

Item	Size (bytes)
SW_ILS Code = 0200 0000h	4
Originating Controlling Switch Switch_Name	8
FKA_ADV_PERIOD	4
Reserved	3
Originating Controlling Switch Priority	1

1.3.2 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 4.

Table 4 – 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 5.

Table 5 – 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 6.

Table 6 – 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.

1.3.3 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 FFFFFFFDh, indicating the originating VE_Port, and the D_ID field shall be set to FFFFFFFDh, indicating the destination VE_Port.

Payload: The format of the GNAS Request Sequence Payload is shown in table 7.

Table 7 – 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 8.

Table 8 – 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.

1.3.4 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 FFFFFFFDh, indicating the originating VE_Port, and the D_ID field shall be set to FFFFFFFDh, indicating the destination VE_Port.

Payload: The format of the SSA Request Sequence Payload is shown in table 9.

Table 9 – 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 10.

Table 10 – 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.