Specification

Mellanox IB DDR Auto-negotiation Protocol

Rev 1.0

Mellanox IB DDR Auto-Negotiation Protocol Specification

Document Number: 2429

Mellanox Technologies, Inc.
350 Oakmead Parkway, Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com

Tel: (408) 970-3400
Fax: (408) 970-3403

Mellanox Technologies Ltd
PO Box 586 Hermon Building
Yokneam 20692
Israel

Tel: +972-4-909-7200
Fax: +972-4-959-3245

Contents

List of Tables

List of Figures

1 Introduction

Mellanox provides HCA and switch devices having InfiniBand ports which are capable of transmitting/receiving data at SDR (Single Data Rate) or DDR (Double Data Rate). That is, their InfiniBand SERDESes (links) can operate at 2.5Gb/s (SDR) or 5Gb/s (DDR) when communicating with a DDR-capable device.

To set SDR or DDR SERDES operation for a device's InfiniBand (IB) port, both hardware and firmware configuration is required. A SERDES can be configured to perform one of the following:

1. Operate at SDR only
2. Operate at DDR only
3. Auto-negotiate with the SERDES at the other end of the link to operate at DDR

This document describes the specification of Mellanox Technologies' IB DDR Auto-negotiation for devices compliant with the *InfiniBand Architecture Specification, Release 1.1*.

Note that a 'link' in this document refers to a pair of SERDES lanes – one TX and one RX.

2 Mellanox IB DDR Auto-negotiation

2.1 Auto-negotiation Protocol

2.1.1 Outline

The following is an outline of the Mellanox IB DDR auto-negotiation protocol for the speed of the link connecting between two IB ports on two different devices. For reference, the two ends of the link are termed *Node 0* and *Node 1*.

1. ***Link @ SDR:*** Node 0 and Node 1, both configured to auto-negotiate for DDR operation, rise initially operating at SDR.
2. ***Capabilities Advertisement:*** Each of the nodes sends on the link a Vendor Specific MAD (see "Auto-negotiation Vendor Specific MAD" on page 14) to figure out whether the other node is DDR capable. If a positive acknowledgement is received, then a parameter check procedure is initiated.

   Note: The above is true for both nodes even if only one of the two gets acknowledged.

3. ***Capabilities Advertisement + ACK:*** The nodes transmit to each other how many SERDES TX and RX configurations each node wishes to test. The nodes transmit to each other the duration for testing each configuration.
4. ***Sweepmode @ DDR:*** Both nodes transfer to Sweep Mode state where each node transmits on its TX one configuration after the other, while the RX verifies whether the received configuration matches one of its options (set in the firmware configuration file), and that the received data is free of symbol errors.

   Note: Transmission and testing in Sweep Mode are performed at DDR rate. This means that the physical link (at SDR) becomes down and is represented as failing.

   Note: *This note applies only to a connection between a no-sweep HCA device and an InfiniScale III based Switch.* A node may choose to skip Sweep Mode entirely. This is requested in the ***Capabilities Advertisement*** and ***Capabilities Advertisement + ACK*** stages above, and the ***Sweepmode @ DDR*** stage will not be reached. See the ***NoSweep*** field in Table 1, "Vendor Specific MAD Layout," on page 14.

5. ***Return to SDR:*** Both nodes return to SDR state after Sweep Mode.
6. ***Result Advertisement:*** Both nodes transmit to each other the desired configuration ("my desired RX is your TX configuration # X").
7. ***Result Advertisement + ACK:*** Each node acknowledges the configuration request of the other and configures its SERDES TX configuration according to what the requested, and its SERDES RX configuration according to what it found to suit itself.

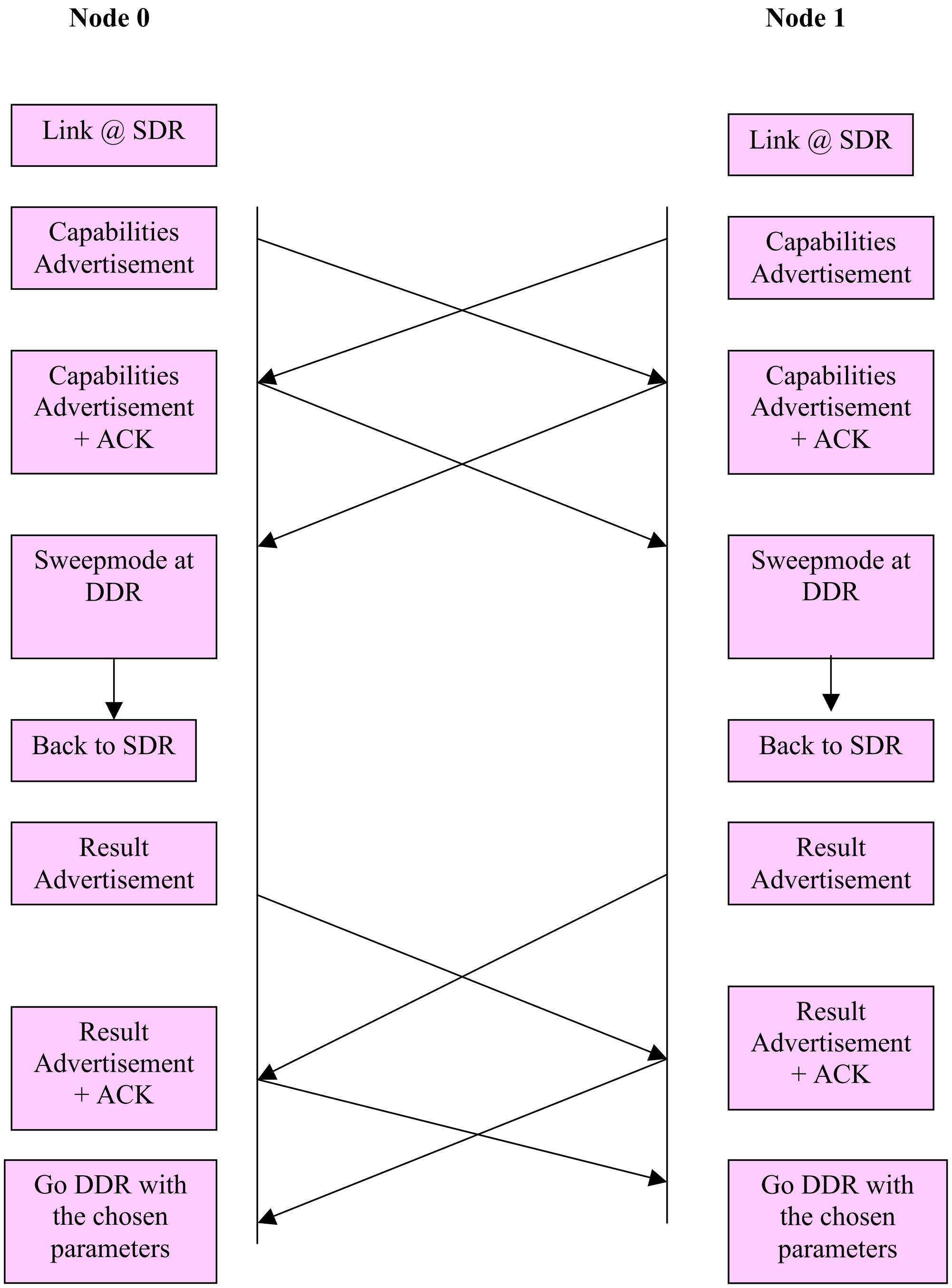
   Note: The above is true for both nodes even if only one of the two gets acknowledged.

8. ***GO DDR/SDR:*** If a suitable DDR configuration is agreed upon, both nodes will go to DDR operation with the chosen configuration parameters. If no suitable configuration was found, then both nodes will operate at SDR.

2.1.2 Flow Diagrams

The following three figures show flow diagrams of auto-negotiation between Node 0 and Node 1. The first two cases (Figure 1 on page 11 and Figure 2 on page 12) yield a successful transition to DDR operation upon finding appropriate configurations for the two nodes. The third case (Figure 3 on page 13) yields operation at SDR only due to failure in finding configurations suitable for the two nodes to operate together at DDR.

Figure 1: SDR to DDR Auto-negotiation (Successful Case 1)


Node 0
Node 1
Link @ SDR
Link @ SDR
Capabilities Advertisement
Capabilities Advertisement
Capabilities Advertisement + ACK
Capabilities Advertisement + ACK
Sweepmode at DDR
Sweepmode at DDR
Back to SDR
Back to SDR
Result Advertisement
Result Advertisement
Result Advertisement + ACK
Result Advertisement + ACK
Go DDR with the chosen parameters
Go DDR with the chosen parameters

Note that in Figure 2, Node 1 does not receive an ACK for its Capabilities Advertisement. Nevertheless, it carries along with the protocol since it re-issued the capabilities with its Capabilities Advertisement + ACK to Node 0. Similarly, Node 0 does not receive an ACK for its Result Advertisement. Nevertheless, it carries along with the protocol since it re-issued the result with its Result Advertisement + ACK to Node 1.

Figure 2: SDR to DDR Auto-negotiation (Successful Case 2)

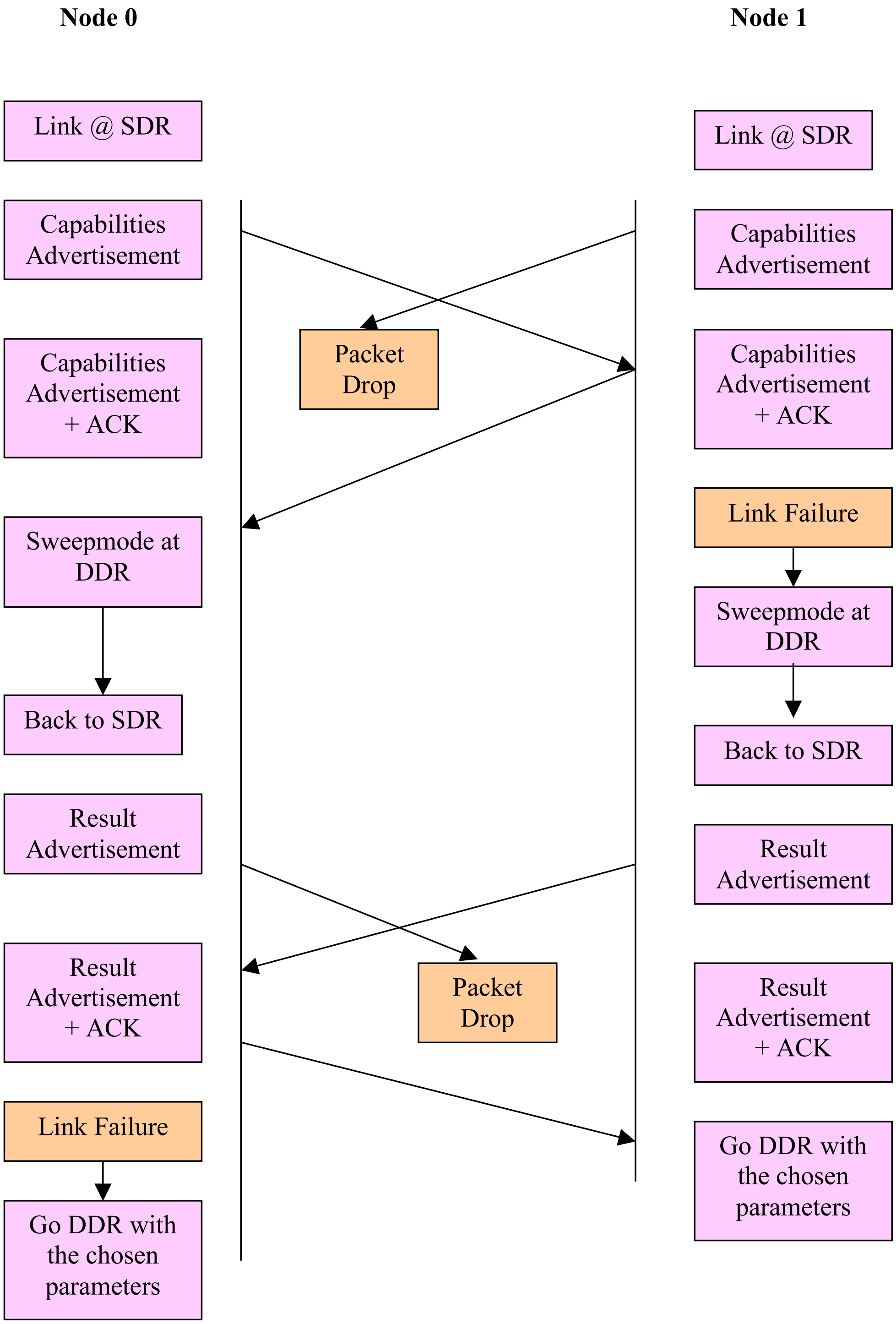
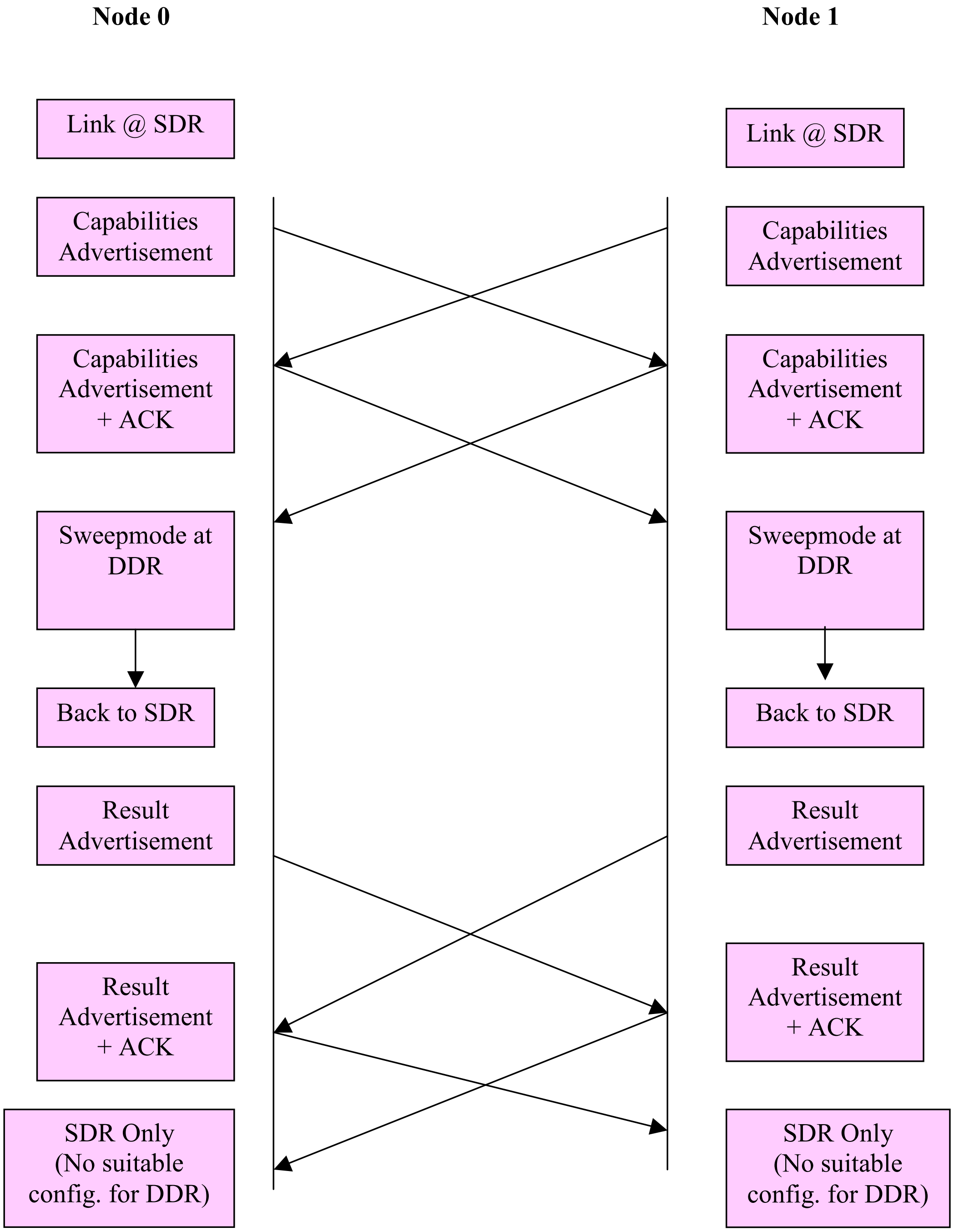

Node 0
Node 1
Link @ SDR
Link @ SDR
Capabilities Advertisement
Capabilities Advertisement
Packet Drop
Capabilities Advertisement + ACK
Capabilities Advertisement + ACK
Link Failure
Sweepmode at DDR
Sweepmode at DDR
Back to SDR
Back to SDR
Result Advertisement
Result Advertisement
Result Advertisement + ACK
Packet Drop
Result Advertisement + ACK
Link Failure
Go DDR with the chosen parameters
Go DDR with the chosen parameters

Figure 3: SDR to DDR Auto-negotiation (Failure Case)


Node 0
Node 1
Link @ SDR
Link @ SDR
Capabilities Advertisement
Capabilities Advertisement
Capabilities Advertisement + ACK
Capabilities Advertisement + ACK
Sweepmode at DDR
Sweepmode at DDR
Back to SDR
Back to SDR
Result Advertisement
Result Advertisement
Result Advertisement + ACK
Result Advertisement + ACK
SDR Only (No suitable config. for DDR)
SDR Only (No suitable config. for DDR)

2.1.3 Auto-negotiation Vendor Specific MAD

The Request and Ack packets exchanged between two link nodes auto-negotiating for DDR operation use the Vendor Specific MAD format described in Table 1 and Table 2.

The MAD should use the following attributes:

- MgmtClass = 0x81 (Direct Route)
- Permissive LID
- VL15
- QP0
- Method: Send (0x3)
- Attribute ID: 0x02C9
- AttributeModifier: 0x000002C9

Table 1 - Vendor Specific MAD Layout

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	Offset
A	SerDesConf	NoSweep	Reserved																												5G	00000h
Sustained DDR Receive BW																																00004h
Iteration Duration																																00004h
Number of Iterations / Chosen Iteration Number																																00004h
Reserved																																00008h - 000FCh

Table 2 - Vendor Specific MAD Field Descriptions

Offset	Bits	Name	Description
00000h	31	A	Ack bit - see Figure 4 on page 18 for details.
	30	SerDesConf	SERDES is Configured. See State machine diagram (Figure 4 on page 18) for details.
	29	NoSweep	If set, skip the sweep phase and configure according to the Chosen Iteration Number (see below). NoSweep should be set by the side that requested to skip the sweep phase for every packet it sends and for each Ack message it returns **Note**: This bit should be turned on by no-sweep HCA devices connected to a Mellanox InfiniScale III Switch device.
	28:1	Reserved	
	0	5G	1 - Supports DDR operation (5GHz) 0 - Does not support DDR operation (2.5GHz)
00004h	31:0	Sustained DDR Receive BW	This field must be set to 5000. This is the maximum sustained Bandwidth in DDR operation that each link (i.e. SerDes) can receive in 1Mbit/sec units. (Mbit = 2^20 bits)
00008h	31:0	Iteration Duration (IT)	If NoSweep=1 this field has no function. **Otherwise**, this is the minimum time required by receiver for adapting to a transmitter configuration. The transmitter must transmit IDLE data for at least this period of time prior to moving to the next configuration. Since each node may require a different IT, the higher number between the two will be used by both of them during Sweep Mode.
0000Ch	31:0	Number of Iterations / Chosen Iteration Number	If NoSweep=1 this field is Chosen Iteration Number. **Otherwise**, if (SerDesConf == 0) then this field is Number of Iterations (NOI); if (SerDesConf == 1) then this field is Chosen Iteration Number Number of Iterations (NOI) - Number of iterations to be executed in the DDRSWEEP (Sweep Mode) state. Since each node may require a different NOI, the higher number between the two will be used by both of them during Sweep Mode. (0 is not supported as NOI.) Chosen Iteration Number - The iteration number that the receiver declares as optimal. This is a number between 0 and 8, where 0 means the first iteration, 2 the second iteration, etc.
00010h 000FCh		Reserved	

2.1.4 Port Info for DDR Support

If an IB port is configured (via hardware and firmware) to auto-negotiate for DDR speed, the following link-speed fields of the PortInfo attribute should be updated as described in Table 3.

Table 3 - Affected PortInfo Attribute Fields

PortInfo Field	Value	Description
LinkSpeedSupported	3	2.5 or 5.0 Gbps
LinkSpeedActive	1 or 2	Reflects the actual port speed: - 1 for SDR - 2 for DDR
LinkSpeedEnabled	3	2.5 or 5.0 Gbps

2.2 SDR to DDR Auto-Negotiation State Machine

Figure 4 on page 18 shows the Mellanox SDR to DDR Auto-Negotiation (i.e., Link Training) state machine diagram. The state machine uses the following two variables:

- RetryCnt0 - number of transmissions of request packets
- RetryCntAck - number of transmissions of Ack packets

Also, the following parameters are used to configure the state machine:

- BurstLen - number of Request/Ack packets to be sent in each transmission. Default value = 4.
- MaxRetry - Maximum number of retries for Request/Ack transmission. Default value =4.
- XmitTimeOut - timeout for receiving Ack. Default value = 1ms.
- DDRTimeOut - Timeout before falling back to SDR. Default value = 400ms.

Once the IB ports are up, the auto-negotiation protocol is executed as described in the state machine diagram. Subnet Manager queries to the ports should be responded to with *LinkPhyState=Polling* and *LinkState=Down* until the protocol is completed (i.e., either SDR state or DDR state is reached).

The DDRSWEEP (Sweep Mode) sub-state-machine is shown in Figure 5 on page 19. See also the example following the figure.

Figure 4: SDR to DDR Auto-negotiation State Machine

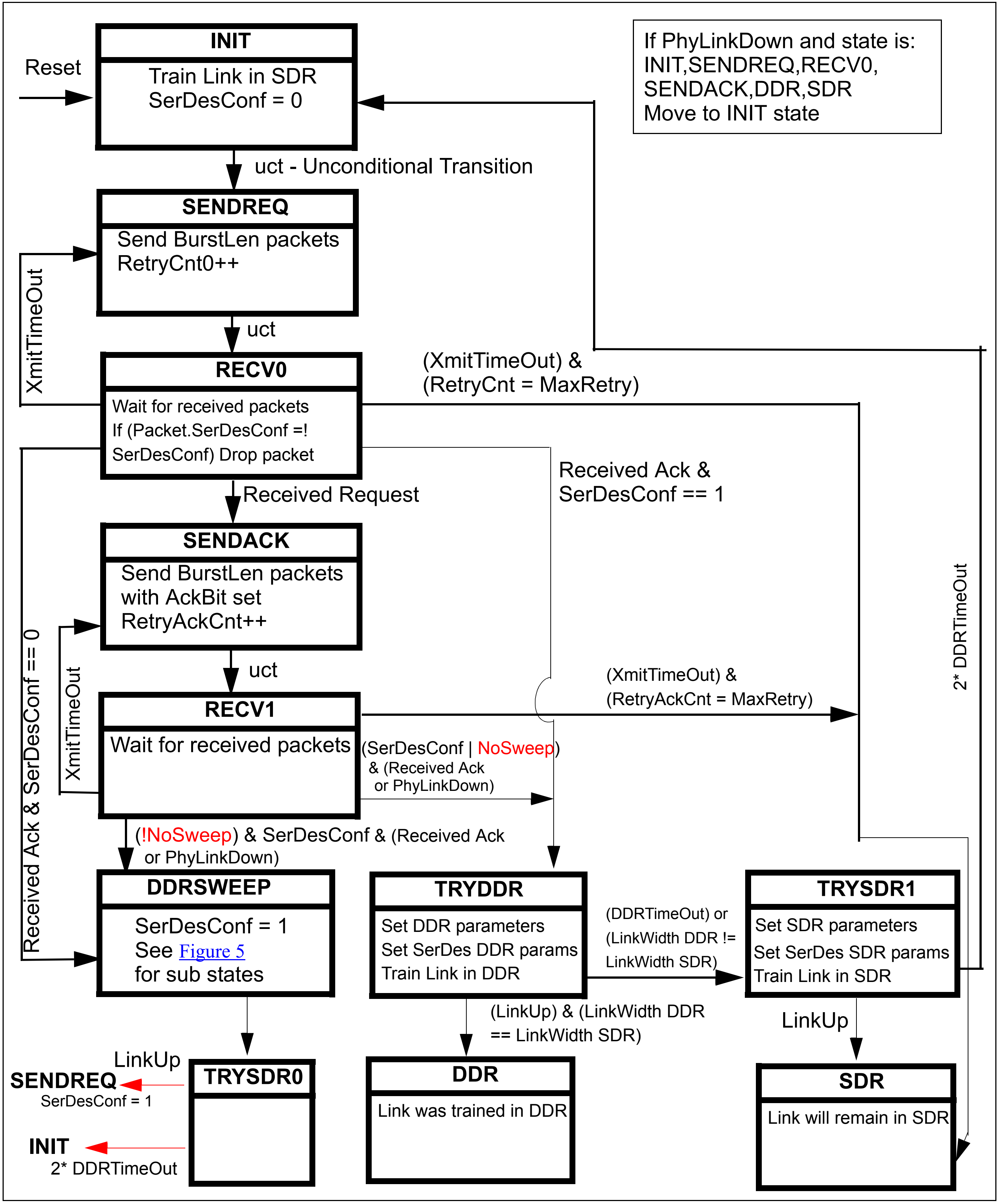
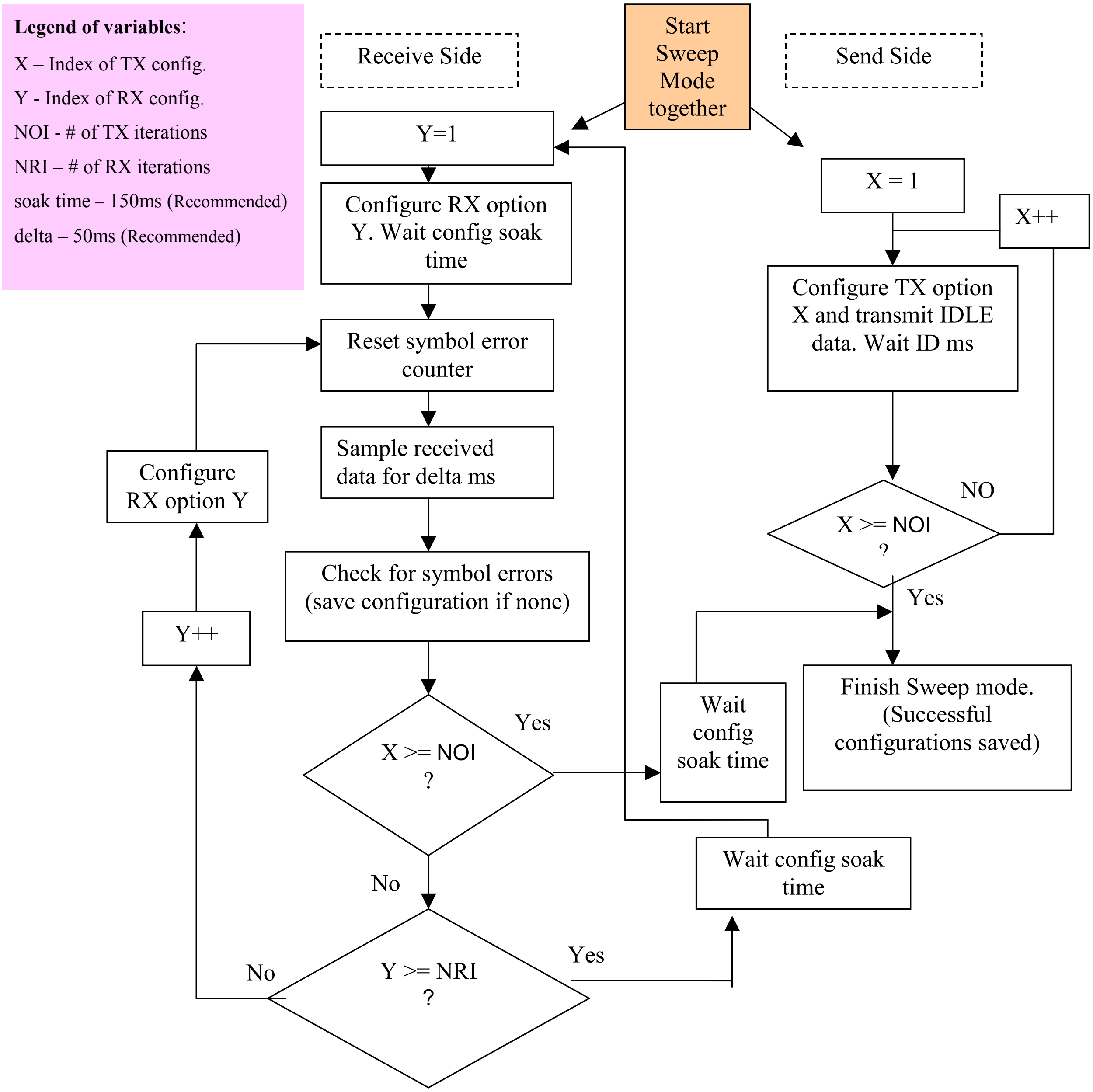


Figure 5: DDRSWEEP (Sweep Mode) State


Legend of variables:
X – Index of TX config.
Y - Index of RX config.
NOI - # of TX iterations
NRI – # of RX iterations
soak time – 150ms (Recommended)
delta – 50ms (Recommended)
Receive Side
Start Sweep Mode together
Send Side
Y=1
Configure RX option Y. Wait config soak time
Reset symbol error counter
Sample received data for delta ms
Check for symbol errors (save configuration if none)
X >= NOI ?
Yes
No
Y >= NRI ?
No
Yes
Configure RX option Y
Y++
Wait config soak time
Wait config soak time
X = 1
X++
Configure TX option X and transmit IDLE data. Wait ID ms
NO
X >= NOI ?
Yes
Finish Sweep mode. (Successful configurations saved)

Example: Consider a case of auto-negotiation on a link between an InfiniScale III switch device (Node 0) and an InfiniHost III Lx HCA device (Node 1). Table 4 on page 20 provides a summary of Sweep Mode parameters assuming that: delta=50ms and soak time=150ms.

Table 4 - Example of Configuration Parameters Used in Sweep Mode

Parameter	Node 0	Node 1	Value Used in Sweep Mode
NOI	3	5	5 (by both)[1]
NRI	9	4	9 by Node 0 4 by Node 1
Iteration Duration (IT)	soak + NRI*delta + soak = 150 + 9*50 +150 = 750ms	soak + NRI*delta + soak = 150 + 4*50 +150 = 500ms	750ms (by both)

[1] Since Node 0 has less transmit configurations to test, it can re-transmit two out of the three configurations it has.

With the above settings, 3 SERDES TX configurations are transmitted by Node 0, and will be checked against 4 SERDES RX configurations by Node 1, and 5 SERDES TX configurations transmitted by Node 1 will be checked against 9 SERDES RX configurations by Node 0. Each iteration lasts for 750ms.

3 Revision History

Rev 1.0 – May 7, 2009:

- Initial release