T10/07-392 revision 1

Date: October 02, 2007

To: T10 Committee (SCSI)

From: George Penokie (IBM)

Subject: SAS-2: Remove AWT reset on receipt of OPEN_REJECT (RETRY)

1 Overview

The fairness algorithms in SAS-2 have a flaw in that when a SAS device receives receipt an OPEN_REJECT (RETRY) it is required to reset the AWT to zero before sending another OPEN address frame to the SAS device that rejected the open. In a large SAS topology running a heavy workload this behavior can result in SAS devices not being able to get an open through in a reasonable length of time.

This proposal adds a NO_AWT_REZERO bit to Protocol-Specific Port mode page SAS-2.

Revision 1 - Changed name to CONTINUE AWT bit and moved it to the Shared Port Control mode page. Added to the port layer specifically the PL_OC state machine.

2 Proposed SAS-2 changes

7.12.3 Arbitration fairness

SAS supports least-recently used arbitration fairness for connection requests.

Each SAS port and expander port shall include an Arbitration Wait Time timer which counts the time from the moment when the port makes a connection request until the request is accepted or rejected. The Arbitration Wait Time timer is in the port layer state machine (see 8.2.2). The Arbitration Wait Time timer shall count in microseconds from 0 µs to 32 767 µs and in milliseconds from 32 768 µs to 32 767 ms + 32 768 µs. The Arbitration Wait Time timer shall stop incrementing when its value reaches 32 767 ms + 32 768 µs.

A SAS port (i.e., SAS initiator ports and SAS target ports) shall start the Arbitration Wait Time timer when it transmits the first OPEN address frame (see 7.8.3) for the connection request. When the SAS port retransmits the OPEN address frame (e.g., after losing arbitration and handling an inbound OPEN address frame), it shall set the ARBITRATION WAIT TIME field to the current value of the Arbitration Wait Time timer.

A SAS port should set the Arbitration Wait Time timer to zero when it transmits the first OPEN address frame for the connection request. A SAS initiator port or SAS target port may be unfair by setting the ARBITRATION WAIT TIME field in the OPEN address frame (see 7.8.3) to a higher value than its Arbitration Wait Time timer indicates. However, an unfair SAS port shall not set the ARBITRATION WAIT TIME field to a value greater than or equal to 8000h; this limits the amount of unfairness and helps prevent livelocks.

The expander port that receives an OPEN address frame shall set the Arbitration Wait Time timer to the value of the incoming ARBITRATION WAIT TIME field and start the Arbitration Wait Time timer as it arbitrates for internal access to the outgoing expander port. When the expander device transmits the OPEN address frame out another expander port, it shall set the outgoing ARBITRATION WAIT TIME field to the current value of the Arbitration Wait Time timer maintained by the incoming expander port.

A SAS port shall stop the Arbitration Wait Time timer and set it to zero when it has no more frames to send.

A SAS port shall stop the Arbitration Wait Time timer and set it to zero when it receives one of the following connection responses:

a) OPEN_ACCEPT;
b) OPEN_REJECT (PROTOCOL NOT SUPPORTED);
c) OPEN_REJECT (RESERVED ABANDON 1);
d) OPEN_REJECT (RESERVED ABANDON 2);
e) OPEN_REJECT (RESERVED ABANDON 3);
f) OPEN_REJECT (STP RESOURCES BUSY);
g) OPEN_REJECT (WRONG DESTINATION);
h) OPEN_REJECT (RESERVED CONTINUE 0); or

i) OPEN_REJECT (RESERVED CONTINUE 1).~~; or~~
~~j) OPEN_REJECT (RETRY).~~

NOTE 1 - Connection responses that are conclusively from the destination phy (see table 112 and table 113 in 7.2.5.13) are included in the list. Except for OPEN_REJECT (RETRY), connection responses that are only from or may be from expander phys are not included.

If the CONTINUE AWT bit in the Shared Port Control mode page (see 10.2.7.6) set to one, then a connection response of OPEN_REJECT (RETRY) shall not stop the Arbitration Wait Time timer and shall not set it to zero. If the CONTINUE AWT bit is set to zero, then a SAS port shall stop the Arbitration Wait Time timer and set it to zero.

A SAS port should not stop the Arbitration Wait Time timer and set it to zero when it receives an incoming OPEN address frame that has priority over the outgoing OPEN address frame according to table 1, regardless of whether it replies with an OPEN_ACCEPT or an OPEN_REJECT.

When arbitrating for access to an outgoing expander port, the expander device shall select the connection request based on the rules described in 7.12.4.

If two connection requests pass on a physical link, the phy shall determine the winner by comparing OPEN address frame field contents using the arbitration priority described in table 1.

Table 1 — Arbitration priority for OPEN address frames passing on a physical link

Bits 79-64 (79 is MSB)	Bits 63-0 (0 is LSB)
ARBITRATION WAIT TIME field value	SOURCE SAS ADDRESS field value

See 7.8.3 for details on the OPEN address frame and the ARBITRATION WAIT TIME field.

8.2.2.3 PL_OC2:Overall_Control state

8.2.2.3.1 PL_OC2:Overall_Control state overview

This state may receive Transmit Frame requests from the transport layers (i.e., SSP and SMP) and Retry frame messages from PL_PM state machines. This state shall create a pending Tx Frame message for each received Transmit Frame request and Retry Frame message. There may be more than one pending Tx Frame message at a time for each SSP transport layer. There shall be only one pending Tx Frame message at a time for each SMP transport layer.

This state selects PL_PM state machines through which connections are established. This state shall only attempt to establish connections through PL_PM state machines whose phys are enabled. In a vendor-specific manner, this state selects PL_PM state machines on which connections are established to transmit frames. This state shall receive a response to a message from a PL_PM state machine before sending another message to that PL_PM state machine.

This state also:

a) receives connection management requests from the transport layers;
b) sends connection management messages to PL_PM state machines;
c) receives connection management messages from PL_PM state machines; and
d) sends connection management confirmations to the transport layers.

After receiving a Transmit Frame request for a destination SAS address for which there is no connection established and for which no I_T Nexus Loss timer has been created, this state shall create an I_T Nexus Loss timer for that SAS address if:

a) the protocol is SSP, the port is an SSP target port, the Protocol-Specific Port mode page is implemented, and the I_T NEXUS LOSS TIME field in the Protocol-Specific Port mode page (see 10.2.7.4) is not set to 0000h;
b) the protocol is STP, the port is an STP target port, and the STP SMP I_T NEXUS LOSS TIME field in the SMP CONFIGURE GENERAL function is not set to 0000h; or

c) the protocol is SMP, the port is an SMP initiator port, and the STP SMP I_T NEXUS LOSS TIME field in the SMP CONFIGURE GENERAL function is not set to 0000h.

This state may create an I_T Nexus Loss timer for that SAS address if:

a) the protocol is SSP and the port is an SSP initiator port; or;
b) the protocol is STP and the port is an STP initiator port.

When this state creates an I_T Nexus Loss timer it shall:

1) initialize the I_T Nexus Loss timer as specified in table 151 (see 8.2.2.1); and
2) not start the I_T Nexus Loss timer.

If this state machine is in an SSP initiator port, then this state may create an I_T Nexus Loss timer for the SAS address. If a state machine in an SSP initiator port and creates an I_T Nexus Loss timer, then the state machine should use the value in the I_T NEXUS LOSS TIME field in the Protocol-Specific Port mode page for the SSP target port (see 10.2.7.4) as the initial value for its I_T Nexus Loss timer.

If there are no pending Tx Frame messages for a destination SAS address and an I_T Nexus Loss timer has been created for that destination SAS address, then this state shall delete the I_T Nexus Loss timer for that destination SAS address.

If this state receives a HARD_RESET Received confirmation, then this state shall discard all pending Tx Frame messages and delete all I_T Nexus Loss timers and send a HARD_RESET Received confirmation to the transport layer.

If this state receives a Notify Received (Power Loss Expected) confirmation, then this state shall:

a) discard all pending Tx Frame messages, if any;
b) delete all I_T Nexus Loss timers, if any;
c) send a Close Connection message to all the PL_PM state machines;
d) send a Cancel Open message to all the PL_PM state machines; and
e) send a Notify Received (Power Loss Expected) confirmation to the transport layer.

8.2.2.3.2 PL_OC2:Overall_Control state establishing connections

This state receives Phy Enabled confirmations indicating when a phy is available.

This state receives Retry Open messages from a PL_PM state machine.

This state creates pending Tx Open messages based on pending Tx Frame messages and Retry Open messages. Pending Tx Open messages are sent to a PL_PM state machine as Tx Open messages.

If this state receives a Retry Open (Retry) message, then this state shall process the Retry Open message.

If this state receives a Retry Open (No Destination) or a Retry Open (Open Timeout Occurred) message and an I_T Nexus Loss timer has not been created for the destination SAS address (e.g., an SSP target port does not support the I_T NEXUS LOSS TIME field in the Protocol-Specific Port mode page or the field is set to 0000h), then this state shall process the Retry Open message as either a Retry Open message or an Unable To Connect message. This selection is vendor-specific.

If this state receives a Retry Open (Pathway Blocked) message and an I_T Nexus Loss timer has not been created for the destination SAS address, then this state shall process the Retry Open message.

If this state receives a Retry Open (No Destination), Retry Open (Open Timeout Occurred), or Retry Open (Pathway Blocked) message, and an I_T Nexus Loss timer has been created for the destination SAS address with an initial value of FFFFh, then this state shall process the Retry Open message (i.e., the Retry Open message is never processed as an Unable to Connect message).

If this state receives a Retry Open (No Destination) or a Retry Open (Open Timeout Occurred) message, an I_T Nexus Loss timer has been created for the destination SAS address, and there is no connection established with the destination SAS address, then this state shall check the I_T Nexus Loss timer, and:

a) if the I_T Nexus Loss timer is not running, the I_T nexus loss time is not set to FFFFh, and the CONFIGURING bit is set to zero in the REPORT GENERAL response (see 10.4.3.3) for each expander

device between this port and the destination port that is two or more levels away from this port, then this state shall start the timer;

b) if the I_T Nexus Loss timer is not running and the I_T nexus loss time is not set to FFFFh, then this state shall start the timer;
c) if the I_T Nexus Loss timer is running, then this state shall not stop the timer; and
d) if the I_T Nexus Loss timer has expired, then this state shall process the Retry Open message as if it were an Unable To Connect message (see 8.2.2.3.4).

If this state receives a Retry Open (Pathway Blocked) message, an I_T Nexus Loss timer has been created for the destination SAS address, and there is no connection established with the destination SAS address, then this state shall check the I_T Nexus Loss timer, and:

a) if the I_T Nexus Loss timer is running, then this state shall not stop the timer; and
b) if the I_T Nexus Loss timer has expired, then this state shall process the Retry Open message as if it were an Unable To Connect message (see 8.2.2.3.4).

If this state receives a Retry Open (Retry) and an I_T Nexus Loss timer is running for the destination SAS address, then this state shall:

a) stop the I_T Nexus Loss timer (if the timer has been running); and
b) initialize the I_T Nexus Loss timer.

This state shall create a pending Tx Open message if:

a) this state has a pending Tx Frame message or has received a Retry Open message;
b) this state has fewer pending Tx Open messages than the number of PL_PM state machines (i.e., the number of phys in the port);
c) there is no pending Tx Open message for the destination SAS address; and
d) there is no connection established with the destination SAS address.

This state may create a pending Tx Open message if:

a) this state has a pending Tx Frame message, or this state has received a Retry Open message and has not processed the message by sending a confirmation; and
b) this state has fewer pending Tx Open messages than the number of PL_PM state machines.

This state shall have no more pending Tx Open messages than the number of PL_PM state machines.

If this state receives a Retry Open message and there are pending Tx Frame messages for which pending Tx Open messages have not been created, then this state should create a pending Tx Open message from the Retry Open message.

If this state does not create a pending Tx Open message from a Retry Open message (e.g., the current number of pending Tx Open messages equals the number of phys), then this state shall discard the Retry Open message. This state may create a new pending Tx Open message at a later time for the pending Tx Frame message that resulted in the Retry Open message.

If this state receives a Retry Open (Opened By Destination) message and the initiator port bit and protocol arguments match those in the Tx Open messages that resulted in the Retry Open message, then this state may discard the Retry Open message and use the established connection to send pending Tx Frame messages as Tx Frame messages to the destination SAS address. If this state receives a Retry Open (Opened By Destination) message, then, if this state has a pending Tx Open slot available, this state may create a pending Tx Open message from the Retry Open message.

NOTE 2 - If a connection is established by another port as indicated by a Retry Open (Opened By Destination) message, credit may not be granted for frame transmission. In this case this state may create a pending Tx Open message from a Retry Open message in order to establish a connection where credit is granted.

This state shall send a pending Tx Open message as a Tx Open message to a PL_PM state machine that has an enabled phy and does not have a connection established. If there is more than one pending Tx Open message, this state should send a Tx Open message for the pending Tx Open message that has been pending for the longest time first.

If this state creates a pending Tx Open message from one of the following messages:

a) a Retry Open (Opened By Destination);
b) a Retry Open (Opened By Other);
c) a Retry Open (Collided); ~~or~~
d) a Retry Open (Pathway Blocked); or
e) a Retry Open (Retry) and the CONTINUE AWT bit in the Shared Port Control mode page (see 10.2.7.6) set to one;

then this state shall:

1) create an Arbitration Wait Time timer for the pending Tx Open message;
2) set the Arbitration Wait Time timer for the pending Tx Open message to the arbitration wait time argument from the Retry Open message; and
3) start the Arbitration Wait Time timer for the pending Tx Open message.

When a pending Tx Open message is sent to a PL_PM state machine as a Tx Open message, the Tx Open message shall contain the following arguments to be used in an OPEN address frame:

a) initiator port bit from the Transmit Frame request;
b) protocol from the Transmit Frame request;
c) connection rate from the Transmit Frame request;
d) initiator connection tag from the Transmit Frame request;
e) destination SAS address from the Transmit Frame request;
f) source SAS address from the Transmit Frame request;
g) pathway blocked count; and
h) arbitration wait time.

If this state creates a pending Tx Open message from one of the following:

a) a Transmit Frame request;
b) a Retry Open (No Destination) message;
c) a Retry Open (Open Timeout Occurred) message; or
d) a Retry Open (Retry) message and the CONTINUE AWT bit in the Shared Port Control mode page (see 10.2.7.6) set to zero,

then this state shall:

a) set the pathway blocked count argument in the Tx Open message to zero; and
b) set the arbitration wait time argument in the Tx Open message to zero or a vendor-specific value less than 8000h (see 7.12.3).

If a pending Tx Open message was created as the result this state receiving a Retry Open (Pathway Blocked) message, then this state shall set the pathway blocked count argument in the Tx Open message to the value of the pathway blocked count argument received with the message plus one, unless the pathway blocked count received with the argument is FFh.

If a pending Tx Open message was created as the result of this state receiving one of the following:

a) a Retry Open (Opened By Destination) message;
b) a Retry Open (Opened By Other) message;
c) a Retry Open (Collided) message; or
d) a Retry Open (Pathway Blocked) message;

then this state shall set the arbitration wait time argument in the Tx Open message to be the value from the Arbitration Wait Time timer created as a result of the Retry Open message.

After this state sends a Tx Open message, this state shall discard the pending Tx Open message from which the Tx Open messages was created. After this state discards a pending Tx Open message, this state may create a new pending Tx Open message.

If this state receives a Connection Opened message and the initiator port bit and protocol arguments match those in any pending Tx Frame messages, then this state may use the established connection to send pending Tx Frame messages as Tx Frame messages to the destination SAS address.

8.2.2.3.3 PL_OC2:Overall_Control state connection established

If this state receives a Connection Opened message or a Retry Open (Opened By Destination) message for a SAS address, and an I_T Nexus Loss timer has been created for the SAS address, then this state shall:

a) stop the I_T Nexus Loss timer for the SAS address, if the timer has been running; and
b) initialize the I_T Nexus Loss timer.

10.2.7.6 Shared Port Control mode page

The Shared Port Control mode page contains parameters that affect SSP target port operation. If the mode page is implemented by one logical unit in a SCSI target device, it shall be implemented by all logical units in the SCSI target device that support the MODE SELECT or MODE SENSE commands.

The mode page policy (see SPC-4) for this mode page shall be shared.

Table 2 defines the format of this mode page.

Table 2 — Shared Port Control mode page

Byte\Bit	7	6	5	4	3	2	1	0
0	PS	SPF (1b)	PAGE CODE (19h)					
1	SUBPAGE CODE (02h)							
2	(MSB)	PAGE LENGTH (000Ch)						
3								(LSB)
4	Reserved							
5	CONTINUE AWT	Reserved			PROTOCOL IDENTIFIER (6h)			
6	(MSB)	POWER LOSS TIMEOUT						
7								(LSB)
8		Reserved						
15								

The PARAMETERS SAVEABLE (PS) bit is defined in SPC-4.

The SUBPAGE FORMAT (SPF) bit shall be set to one to access this mode page.

The PAGE CODE field shall be set to 19h.

The SUBPAGE CODE field shall be set to 02h.

The PAGE LENGTH field shall be set to the number of bytes in the page after the PAGE LENGTH field (i.e., 000Ch).

The PROTOCOL IDENTIFIER field shall be set to 6h indicating this is a SAS SSP specific mode page.

A CONTINUE AWT bit set to one specifies that the a SAS port shall not stop the Arbitration Wait Time timer and set it to zero when it receives an OPEN_REJECT (RETRY). A CONTINUE AWT bit set to zero specifies that the a SAS port shall stop the Arbitration Wait Time timer and set it to zero when it receives an OPEN_REJECT (RETRY).

The POWER LOSS TIMEOUT field contains the maximum time, in one millisecond increments, that a target port shall respond to connection requests with OPEN_REJECT (RETRY) after receiving NOTIFY (POWER LOSS EXPECTED) (see 7.2.5.11.3). A POWER LOSS TIMEOUT field set to 0000h is vendor-specific. The power loss timeout shall be restarted on each NOTIFY (POWER LOSS EXPECTED) that is received.