To: T10 Technical Committee
From: Rob Elliott, HP (elliott@hp.com)
Date: 12 April 2007
Subject: 07-178r0 SAS-2 Connection request livelock avoidance

Revision history

Revision 0 (12 April 2007) First revision

Related documents

sas2r09 - Serial Attached SCSI - 2 (SAS-2) revision 9

Overview

1. To avoid deadlocks between SSP ports, SAS-2 section 7.16.4 (SSP flow control) requires that an SSP initiator not refuse to provide RRDY credit because it needs to transmit a frame itself. Other reasons are OK; as long as they go away on their own accord, no deadlock will result.

Section 7.16.1 (Opening SSP connections) requires that an initiator (or target) provide at least one RRDY credit for each incoming connection request that it accepts.

When considered together, these rules are intended to mean that an initiator shall not reject a connection request with OPEN_REJECT (RETRY) because it needs to transmit a frame itself. Temporary, self-clearing reasons are OK. However, this is not obvious when reading the sections individually. Section 7.16.1 should specifically mention this case.

2. Phys supporting both initiator and target roles must take care not to create livelocks.

Figure 1 shows an example of SSP ports supporting both target and initiator roles.

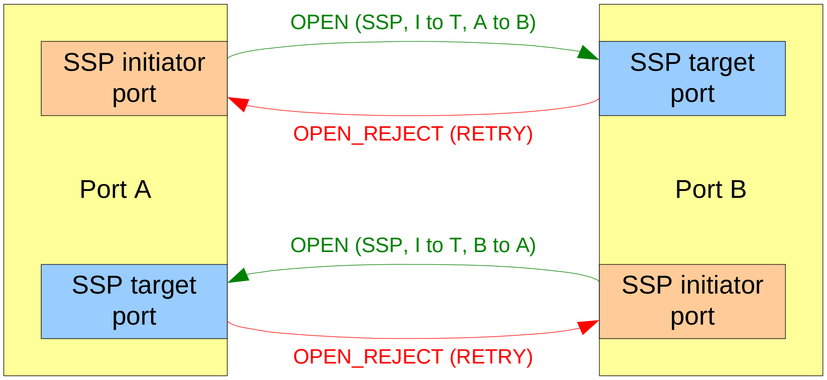

Figure 1 — SSP-only connection livelock example

Figure 2 shows an example of ports with initiators and targets of different protocols.

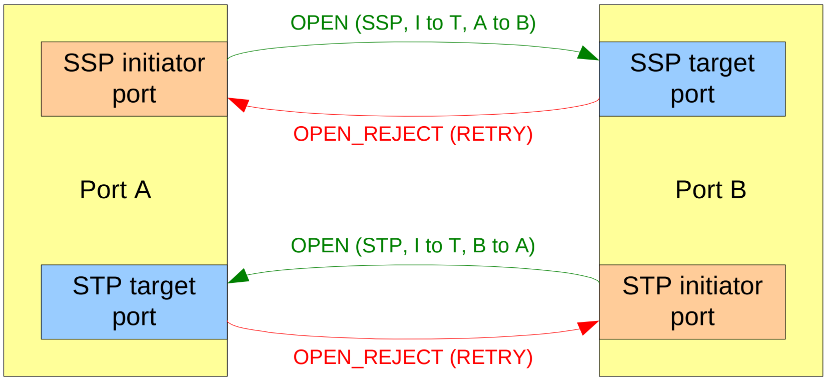

Figure 2 — SSP/STP connection livelock example

Assume phy A’s STP target port is sharing a buffer with phy A’s SSP initiator port, and the STP target port is using that buffer and needs to send a frame to phy B’s STP initiator port to free the buffer. Section 7.15.8 (Opening STP connections) allows STP target ports to reject incoming connection requests while trying to establish their own outgoing connections.

Assume phy B’s SSP target port is sharing a buffer with phy B’s STP initiator port, and the SSP target port is using that buffer and needs to send a frame to phy A’s SSP initiator port to free the buffer. Section 7.16.1 (per issue #1) allows SSP target ports allows to reject incoming connection requests while trying to establish their own outgoing connections.

These combine into a livelock, where neither phy A nor phy B is able to free its buffer as they continually reject each others’ connection requests.

To avoid this, the target allowance for sending OPEN_REJECT (RETRY) because it needs to win an outgoing connection needs to be further constrained. If an initiator and target are sharing a phy (regardless of protocol), the phy must accept an incoming connection request destined to its initiator role without waiting for any of its outgoing connection requests to win.

3. In the XL state machine, Arb Reject (Retry) needs to be included in the list of confirmations from the ECM that are handled, since self-configuring and zoning expander devices can now generate it.

Suggested changes to SAS-2

4.6.6.3 ECM interface

...

Table 1 describes the confirmations from the ECM to an expander logical phy. These confirmations are sent in confirmation of a Request Path request. See 7.12.4 for specific definitions for when each confirmation is sent.

Table 1 — ECM to expander logical phy confirmations (part 1 of 2)

Confirmation	Description
Arbitrating (Normal)	Confirmation that the ECM has received the Request Path request.
Arbitrating (Waiting On Partial)	Confirmation that the ECM is waiting on a partial pathway (see 4.1.11).
Arbitrating (Blocked On Partial)	Confirmation that the ECM is waiting on a blocked partial pathway (see 4.1.11).
Arbitrating (Waiting On Connection)	Confirmation that the ECM is waiting for a connection to complete (see 4.1.12).
Arb Won	Confirmation that an expander logical phy has won path arbitration.
Arb Lost	Confirmation that an expander logical phy has lost path arbitration.

Table 1 — ECM to expander logical phy confirmations (part 2 of 2)

Confirmation	Description
Arb Reject (No Destination)	Confirmation that the request is rejected because the expander device is not configuring (see 4.8) and there is no path to the destination.
Arb Reject (Bad Destination)	Confirmation that the request is rejected because the path to the destination maps back to the requesting expander port.
Arb Reject (Connection Rate Not Supported)	Confirmation that the request is rejected because there is a destination port capable of routing to the requested destination SAS address but no phys within the destination port are configured to support the requested connection rate.
Arb Reject (Zone Violation)	Confirmation that the request is rejected because the expander device is not locked and there is a zoning violation (see 4.9.3).
Arb Reject (Pathway Blocked)	Confirmation that the request is rejected because the requesting expander logical phy needs to back off according to SAS pathway recovery rules.
Arb Reject (Retry)	Confirmation that the request is rejected because: a) the expander device is configuring (see 4.7.2) and the ECM would otherwise have returned Arb Reject (No Destination); ~~or~~ b) the expander device is locked (see 4.9.6.2) and the ECM would otherwise have returned Arb Reject (Zone Violation); or c) the expander device has reduced functionality (see 4.6.8 and 7.12.4.2.5).

Editor's Note 1: reordered a) and b) above to match the order used in the OPEN_REJECT table

7.2.5.13 OPEN_REJECT

OPEN_REJECT specifies that a connection request has been rejected and specifies the reason for the rejection. The result of some OPEN_REJECTs is to abandon (i.e., not retry) the connection request and the result of other OPEN_REJECTs is to retry the connection request.

...

All of the OPEN_REJECT versions defined in table 107 shall result in the originating port retrying the connection request.

Table 107 — Retry-class OPEN_REJECT primitives (part 1 of 2)

Primitive	Originator	Description
OPEN_REJECT (NO DESTINATION) [a]	Expander phy	An expander device in the pathway is not configuring and determines that: a) there is no such destination phy; b) the connection request routes to a destination expander phy in the same expander port as the source expander phy and the expander port is using the subtractive routing method; or c) the SAS address is valid for an STP target port in an STP/SATA bridge, but the initial Register - Device to Host FIS has not been successfully received (see 10.4.3.9).

Table 107 — Retry-class OPEN_REJECT primitives (part 2 of 2)

Primitive	**Originator**	**Description**
OPEN_REJECT (PATHWAY BLOCKED) [b]	Expander phy	An expander device determined the pathway was blocked by higher priority connection requests.
OPEN_REJECT (RESERVED CONTINUE 0) [c]	Unknown	Reserved. Process the same as OPEN_REJECT (RETRY).
OPEN_REJECT (RESERVED CONTINUE 1) [c]		
OPEN_REJECT (RESERVED INITIALIZE 0) [a]	Unknown	Reserved. Process the same as OPEN_REJECT (NO DESTINATION).
OPEN_REJECT (RESERVED INITIALIZE 1) [a]		
OPEN_REJECT (RESERVED STOP 0) [b]	Unknown	Reserved. Process the same as OPEN_REJECT (PATHWAY BLOCKED).
OPEN_REJECT (RESERVED STOP 1) [b]		
OPEN_REJECT (RETRY) [c]	Destination phy or zoning expander phy	Either: a) a phy with destination SAS address exists but is temporarily not able to accept connections (see 7.16.1, 7.17.5, and 7.18.3); b) an expander device in the pathway is configuring and would otherwise have returned OPEN_REJECT (NO DESTINATION) (see 4.7.1 and 7.12.4.2.5); c) an expander device in the pathway is locked and would otherwise have returned OPEN_REJECT (ZONE VIOLATION) (see 4.7.2 and 7.12.4.2.5); or d) an expander device in the pathway has reduced functionality ~~(see 4.6.8)~~(see 4.6.8 and 7.12.4.2.5).

[a] If the I_T Nexus Loss timer is already running, it continues running; if it is not already running, it is initialized and started. Stop retrying the connection request if the I_T Nexus Loss timer expires.

[b] If the I_T Nexus Loss timer is already running, it continues running. Stop retrying the connection request if the I_T Nexus Loss timer expires.

[c] If the I_T Nexus Loss timer (see 8.2.2) is already running, it is stopped.

NOTE 1 - Some SAS logical phys compliant with earlier versions of this standard also transmit OPEN_REJECT (RETRY) if they receive an OPEN address frame while their SL_CC state machines are in the SL_CC5:BreakWait state (see 7.14.4.7).

When a SAS logical phy detects more than one reason to transmit an OPEN_REJECT, the SL_CC state machine determines the priority in the SL_CC2:Selected state (see 7.14.4.4).

When an expander logical phy detects more than one reason to transmit an OPEN_REJECT, the ECM determines the priority (see 7.12.4).

See 7.12 for details on connection requests.

7.12 Connections

7.12.4 Arbitration inside an expander device

7.12.4.2.5 Arb Reject confirmation

The ECM shall generate the following Arb Reject confirmation when any of the following conditions are met and all the Arb Won conditions (see 7.12.4.2.3) are not met:

1) Arb Reject (Bad Destination) if the source expander logical phy and destination expander logical phy(s) are in the same expander port and are using the direct routing method;
2) Arb Reject (Retry) if the expander device is unable to process the connection request because it has reduced functionality (see 4.6.8);
3) if the source expander logical phy and destination expander logical phy(s) are in the same expander port and are using the table routing method or the subtractive routing method:
   A) Arb Reject (No Destination) if the expander device is not configuring; and
   B) Arb Reject (Retry) if the expander device is configuring;
4) if there are no destination expander logical phys (i.e., there is no direct routing or table routing match and there is no subtractive phy):
   A) Arb Reject (No Destination) if the expander device is not configuring; and
   B) Arb Reject (Retry) if the expander device is configuring;
5) Arb Reject (Connection Rate Not Supported) if none of the destination expander logical phys supports the connection rate;
6) if access to the destination expander logical phy(s) is prohibited by zoning (see 4.9.3):
   A) Arb Reject (Zone Violation) if the zoning expander device is unlocked; and
   B) Arb Reject (Retry) if the zoning expander device is locked;

   and

7) Arb Reject (Pathway Blocked) if all the destination expander logical phys that support the connection rate contain blocked partial pathways (i.e., are all returning Phy Status (Blocked Partial Pathway)) and pathway recovery rules require this Request Path request be rejected to release path resources (see 7.12.4.5).

7.12.4.5 Pathway recovery

Pathway recovery provides a means to abort connection requests in order to prevent deadlock using pathway recovery priority comparisons. Pathway recovery priority comparisons compare the PATHWAY BLOCKED COUNT fields and SOURCE SAS ADDRESS fields of the OPEN address frames of the blocked connection requests as described in table 108.

Table 108 — Pathway recovery priority

Bits 71-64 (71 is MSB)	Bits 63-0 (0 is LSB)
PATHWAY BLOCKED COUNT field value	SOURCE SAS ADDRESS field value

When the Partial Pathway Timeout timer for an arbitrating expander phy expires (i.e., reaches a value of zero), the ECM shall determine whether to continue the connection request or to abort the connection request.

The ECM shall reply to a connection request with Arb Reject (Pathway Blocked) when:

a) the Partial Pathway Timeout timer expires; and
b) the pathway recovery priority of the arbitrating expander phy (i.e., the expander phy requesting the connection) is less than or equal to the pathway recovery priority of any of the expander phys within the destination port that are sending Phy Status (Blocked Partial Pathway) responses to the ECM.

The pathway blocked count and source SAS address values used to form the pathway recovery priority of a destination phy are those of the Request Path request if the phy sent a Request Path request to the ECM or those of the Forward Open indication if the phy received a Forward Open indication from the ECR.

7.15.4.5 Transition XL1:Request_Path to XL5:Forward_Open

This transition shall occur if a Forward Open indication is received and none of the following confirmations have been received:

a) Arbitrating (Normal);
b) Arbitrating (Waiting On Partial);
c) Arbitrating (Blocked On Partial);
d) Arbitrating (Waiting On Connection);
e) Arb Won;
f) Arb Lost;
g) Arb Reject (No Destination);
h) Arb Reject (Bad Destination);
i) Arb Reject (Connection Rate Not Supported);
j) Arb Reject (Zone Violation); ~~or~~
k) Arb Reject (Pathway Blocked)~~.~~; or
l) Arb Reject (Retry).

This transition shall include:

a) an OPEN Address Frame Received argument containing the arguments received in the Forward Open indication; and
b) a BREAK Received argument if a BREAK Received message was received.

7.15.7 XL4:Open_Reject state

7.15.7.1 State description

This state is used to reject a connection request.

This state shall send one of the following messages to the XL transmitter (see 7.15.7.2):

a) a Transmit OPEN_REJECT (No Destination) message when an Arb Reject (No Destination) argument is received with the transition into this state;
b) a Transmit OPEN_REJECT (Bad Destination) message when an Arb Reject (Bad Destination) argument is received with the transition into this state;
c) a Transmit OPEN_REJECT (Connection Rate Not Supported) message when an Arb Reject (Connection Rate Not Supported) argument is received with the transition into this state;
d) a Transmit OPEN_REJECT (Zone Violation) message when an Arb Reject (Zone Violation) argument is received with the transition into this state; ~~or~~
e) a Transmit OPEN_REJECT (Pathway Blocked) message when an Arb Reject (Pathway Blocked) argument is received with the transition into this state~~.~~or
f) a Transmit OPEN_REJECT (Retry) message when an Arb Reject (Retry) argument is received with the transition into this state.

7.15.7.2 Transition XL4:Open_Reject to XL0:Idle

This transition shall occur after sending a Transmit OPEN_REJECT message to the XL transmitter.

7.15.7.3 Transition XL4:Open_Reject to XL5:Forward_Open

This transition shall occur if a Forward Open indication is received. This transition shall include an OPEN Address Frame Received argument containing the arguments received in the Forward Open indication.

7.16 SSP link layer

7.16.1 Opening an SSP connection

An SSP phy that accepts a connection request (i.e., an OPEN address frame) shall transmit at least one RRDY in that connection within 1 ms of transmitting an OPEN_ACCEPT. If the SSP phy is not able to grant credit, it shall respond with OPEN_REJECT (RETRY) and not accept the connection request.

To prevent livelocks (e.g., where ports are waiting on each other to accept a connection request):

a) a SAS phy shall not reject an incoming connection request to an SSP initiator port with OPEN_REJECT (RETRY) because the SAS port containing that SAS phy needs an outgoing connection request to be accepted (e.g., if the SAS phy is used by an SSP initiator port and an SSP target port, they share a buffer, that buffer is being used by the SSP target port, and the SSP target port needs to transmit a frame to another SSP initiator port before it is able to free that buffer);
b) a SAS phy may reject an incoming connection request to an SSP initiator port with OPEN_REJECT (RETRY) for any reason that is not dependent on the SAS port containing that SAS phy having an outgoing connection request accepted (e.g., a temporary buffer full condition); and
c) a SAS phy may reject incoming connection requests to an SSP target port with OPEN_REJECT (RETRY) for any reason, including because the SAS port containing that SAS phy needs an outgoing connection request to be accepted (e.g., to transmit a frame and empty a buffer).

7.16.4 SSP flow control

An SSP phy uses RRDY to grant credit for permission for the other SSP phy in the connection to transmit frames. Each RRDY increments credit by one frame. Frame transmission decrements credit by one frame. Credit of zero frames is established at the beginning of each connection.

SSP phys shall not increment credit past 255 frames.

To prevent deadlocks where an SSP initiator port and SSP target port are both waiting on each other to provide credit, an SSP initiator port shall ~~never~~not refuse to provide credit by withholding RRDY because it needs to transmit a frame itself. It may refuse to provide credit for other reasons (e.g., temporary buffer full conditions).

An SSP target port may refuse to provide credit for any reason, including because it needs to transmit a frame itself.

If credit is zero, SSP phys that are going to be unable to provide credit for 1 ms may send CREDIT_BLOCKED. The other phy may use this to avoid waiting 1 ms to transmit DONE (CREDIT TIMEOUT) (see 7.16.8).

If credit is nonzero, SSP phys that are going to be unable to provide additional credit for 1 ms, even if they receive frames per the existing credit, may transmit CREDIT_BLOCKED.

After sending CREDIT_BLOCKED, an SSP phy shall not transmit any additional RRDYs in the connection.

7.15.8 Opening an STP connection

When the SATA host port in an STP/SATA bridge receives a SATA_X_RDY from the attached SATA device, the STP target port in the STP/SATA bridge shall establish an STP connection to the appropriate STP initiator port.

A wide STP initiator port shall not request more than one connection at a time to a specific STP target port.

While a wide STP initiator port is waiting for a response to a connection request to an STP target port, ~~it~~a SAS phy in the STP initiator port shall not reject an incoming connection request from that STP target port with OPEN_REJECT (RETRY) because ~~of its outgoing connection request.~~the SAS port containing the SAS phy needs an outgoing connection request to be accepted. ~~It~~The SAS phy may reject an incoming connection request~~s~~ from that STP target port with OPEN_REJECT (RETRY) for ~~other reasons (see 7.2.5.13)~~any reason that is not dependent on the SAS port containing that SAS phy having an outgoing connection request accepted (e.g., a temporary buffer full condition).

If a wide STP initiator port receives an incoming connection request from an STP target port while it has a connection established with that STP target port, it shall reject the request with OPEN_REJECT (RETRY).

A wide STP target port shall not request more than one connection at a time to a specific STP initiator port.

While a wide STP target port is waiting for a response to a connection request or has established a connection to an STP initiator port, it shall:

a) reject incoming connection requests from that STP initiator port with OPEN_REJECT (RETRY); and

b) if affiliations are supported and the maximum number of affiliations has been established (i.e., all affiliation contexts are in use), reject incoming connection requests from other STP initiator ports that do not have affiliations with OPEN_REJECT (STP RESOURCES BUSY).

A SAS phy may reject incoming connection requests to an STP target port with OPEN_REJECT (RETRY) for any reason, including because the SAS port containing that SAS phy needs an outgoing connection request to be accepted (e.g., to transmit a frame and empty a buffer).

An expander device should not allow its STP ports (e.g., the STP target ports in STP/SATA bridges and any STP initiator ports in the expander device) to attempt to establish more connections to a specific destination port than the destination port width or the width of the narrowest physical link on the pathway to the destination port. This does not apply to connection requests being forwarded by the expander device.

An expander device should not allow its STP ports (e.g., the STP target ports in STP/SATA bridges and any STP initiator ports in the expander device) to attempt to establish more connections than the width of the narrowest common physical link on the pathways to the destination ports of those connections. This does not apply to connection requests being forwarded by the expander device.

...

7.18.3 Opening an SMP connection

An SMP target port shall not attempt to establish an SMP connection.

A SAS phy may reject incoming connection requests to an SMP target port with OPEN_REJECT (RETRY) for any reason, including because the SAS port containing that SAS phy needs an outgoing connection request to be accepted (e.g., to transmit a frame and empty a buffer).