To:      T10 Technical Committee
From:   Rob Elliott, HP (elliott@hp.com)
Date:   22 February 2004
Subject: 04-032r1 SAS-1.1 ALIGNs through expanders

## Revision history

Revision 0 (31 December 2003) First revision
Revision 1 (22 February 2004) Corrected rate matching minimum numbers in tables - from "2048 or 2050" to "2049" for non-STOP, and from "2064 or 2066" to "2065" for STP. There is still toggling going on in every other 2K window, but a 4K window always has one short 2K sections and one long 2K section, so they add together ("1024 + 1025" for non-STP and "1032+1033" for STP).

## Related documents

sas1r03 - Serial Attached SCSI 1.1 revision 3
03-334r2 SAS-1.1 ALIGN insertion clarifications (incorporated into sas1r03). This proposal did not *change* any rule except relaxing the consecutiveness requirement for STP throttling ALIGNs; mostly it clarified existing rules.

## Overview

When an expander is forwarding dwords from a 1.5 Gbps link to a 3.0 Gbps link and inserting rate matching ALIGNs, the clock skew management ALIGN frequency requirement of 1 ALIGN every 2048 dwords cannot be met. Figure 1 shows the scenario.
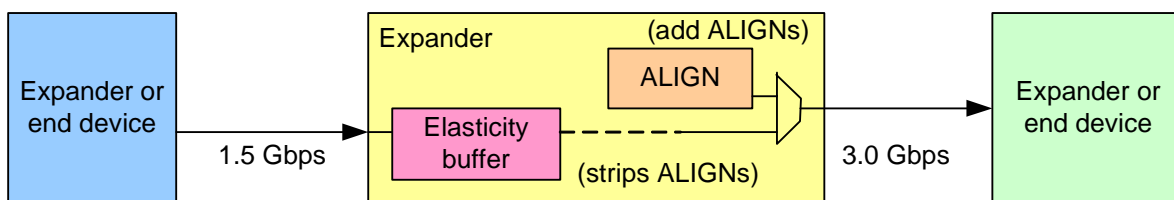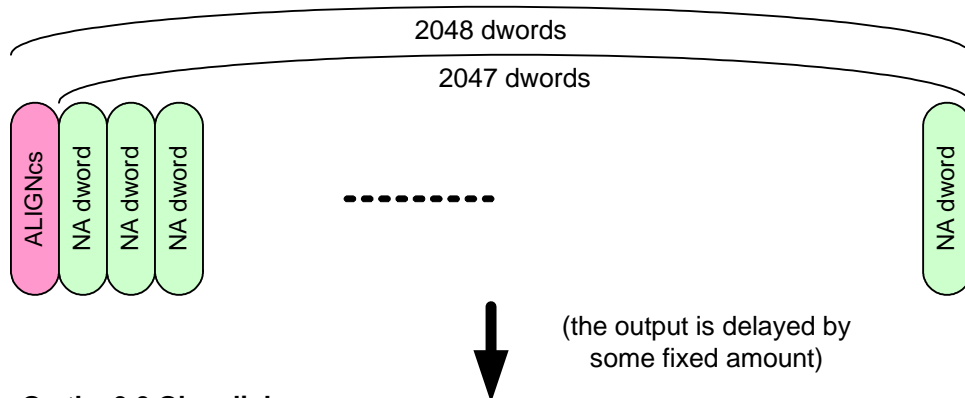


**Figure 1 — 1.5 Gbps to 3.0 Gbps connection**

As the expander forwards the dword stream, it has to insert rate matching ALIGNs as required on the 3.0 Gbps link to make up for underflows in the input 1.5 Gbps stream - 1 ALIGN every 2 dwords. Essentially, the transmitter underflows a lot. It is not allowed to buffer up N dwords and burst them out together, followed by N ALIGNs - it is required to dole out the ALIGNs every other dword so the receiver of the data is not required to have lots of extra buffering.

Additionally, The 1.5 Gbps source transmits 1 ALIGN every 2048 dwords for clock skew management (i.e., clock frequency tolerance compensation). Each of these ALIGNs creates additional underflows in the expander that consume two 3.0 Gbps dword widths of time. The expander probably sends the two 3.0 Gbps ALIGNs back-to-back. It could also scatter them, but that requires extra buffering in the expander.

Figure 2 shows the dword stream on each physical link.



NA dword = Non-ALIGN dword
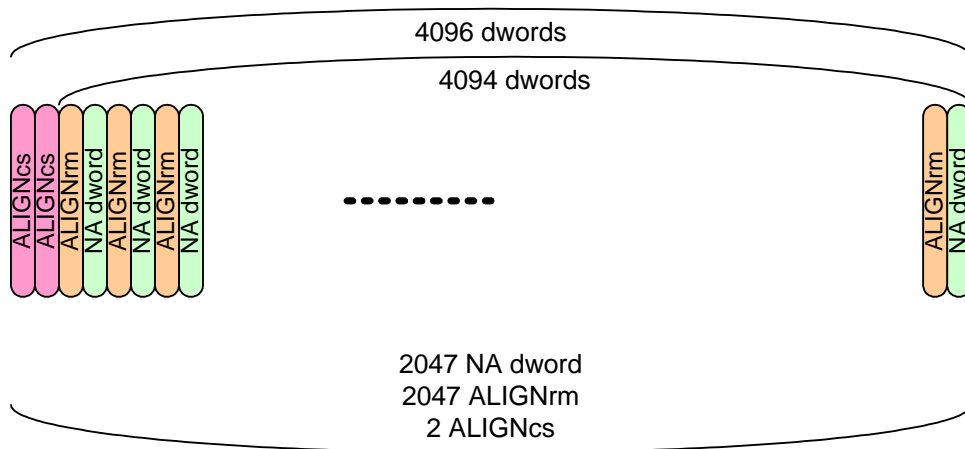ALIGNcs = ALIGN inserted for clock skew management
ALIGNrm = ALIGN inserted for rate matching

**Figure 2 — ALIGNs on the links**

Of the 4094 dwords on the 3.0 Gbps link that are not the two clock skew ALIGNs, 2047 are the non-ALIGN dwords from the 1.5 Gbps link and 2047 are the rate-matching ALIGNs. There are no clock skew ALIGNs in that group - they appeared at the beginning (of the 4096 dword window).

So, on the 3.0 Gbps link, all that can be guaranteed is 2 ALIGNs every 4096 dwords, not 1 ALIGN every 2048 dwords. There are many 2048 dword windows where those ALIGNs might not both appear (inside the 4094 dword window). For 6.0 Gbps (in SAS-2), the ratio will grow to 4 ALIGNs every 8192 dwords.

These ratios are all the same when considered over at least the time it takes to transmit the original 2047 non-ALIGN dwords (e.g. the 4096 window on a 3.0 Gbps link). Over any shorter term (e.g. a 2048 window on a 3.0 Gbps link), however, the ratio cannot be guaranteed.

This problem is not related to whether clock frequency differences actually exist in the system; it occurs even when the clock frequency difference is 0 ppm. The standard already allows links to not comply with the 1/2048 rule if clock frequency differences exist, but does not allow any violations if there is no need.

The impact of 2/4096 rather than 1/2048 is that the 3.0 Gbps receiver must have an extra dword buffer in its elasticity buffer than a 1.5 Gbps receiver. Most existing designs already include a few extra dwords, so this should not break any of them.

For SAS-2, 6.0 Gbps receivers will have to tolerate 4/8192 and will require a total of 3 additional dword buffers vs a 1.5 Gbps design.

Since 1/2048 is .000488, 100 ppm is .0001, and +/- 100 ppm is .0002, SAS currently requires 2.4x as many ALIGNs to be transmitted as should be needed. This may be enough to ensure that existing 3.0 Gbps receivers work with fewer ALIGNs even if they don't buffer an extra dword, provided they are not going out of their way to check the 1/2048 rule.

Mandating the expander guarantee 1/2048 would probably require it provide 1028 dwords of buffering (for 3.0 Gbps), which violates the original SAS design goal of expanders to have minimal buffers.

**Suggested changes to SAS-1.1**

**7.3 Clock skew management**

The internal clock for a device is typically based on a PLL with its own clock generator and is used when transmitting dwords on the physical link. When receiving, however, dwords need to be latched based on a clock derived from the input bit stream itself. Although the input clock is nominally a fixed frequency, it may differ slightly from the internal clock frequency, up to the physical link rate tolerance defined in table 25 (see 5.3.2). Over time, if the input clock is faster than the internal clock, the device may receive a dword and not be able to forward it to an internal buffer; this is called an overrun. If the input clock is slower than the internal clock, the device may not have a dword when needed in an internal buffer; this is called an underrun.

To solve this problem, transmitting devices insert ALIGNs or NOTIFYs in the dword stream. Receivers may pass ALIGNs and NOTIFYs through to their internal buffers, or may strip them out when an overrun occurs. Receivers add ALIGNs or NOTIFYs when an underrun occurs. The internal logic shall ignore all ALIGNs and NOTIFYs that arrive in the internal buffers.

Elasticity buffer circuitry, as shown in figure 76, is required to absorb the slight differences in frequencies between the SAS initiator phy, SAS target phy, and expander phys. The frequency tolerance for a phy is specified in 5.3.2. The depth of the elasticity buffer is vendor-specific but shall accomodate the clock skew management ALIGN insertion requirements in table 1.

**Figure 76 - Elasticity buffers**

[no change]

A phy that is the original source for the dword stream (i.e., a phy that is not an expander phy forwarding dwords from another expander phy) shall insert one ALIGN or NOTIFY for clock skew management ~~within every 2 048 dwords (i.e., every overlapping window of 2 048 dwords)~~as described in table 1.

**Table 1 — Clock skew management ALIGN insertion requirements [new table]**

| Physical link rate | Requirement |
|---|---|
| 1,5 Gbps | One ALIGN or NOTIFY within every 2 048 dwords |
| 3,0 Gbps | Two ALIGNs or NOTIFYs within every 4 096 dwords |

ALIGNs and NOTIFYs inserted for clock skew management are in addition to ALIGNs and NOTIFYs inserted for rate matching (see 7.13) and STP initiator throttling (see 7.17.x). See Annex L for a summary of their combined requirements.

An expander device that is forwarding dwords (i.e., is not the original source) is allowed to insert or delete as many ALIGNs or NOTIFYs as required to match the transmit and receive connection rates. It is not required to transmit ~~any particular~~ the number of ALIGNs and/or NOTIFYs for clock skew management described in table 1 when forwarding to a SAS physical link. It may increase or reduce that number based on clock frequency differences between the phy transmitting the dwords to the expander device and the expander device's receiving phy.

> NOTE 1 - One possible implementation for expander devices forwarding dwords is for the expander device to delete all ALIGNs and NOTIFYs received and to insert ALIGNs/NOTIFYs at the transmit port whenever its elasticity buffer is empty.

The STP target port of an STP/SATA bridge is allowed to insert or delete as many ALIGNs or NOTIFYs as required to match the transmit and receive connection rates. It is not required to transmit any particular number of ALIGNs and/or NOTIFYs for clock skew management when forwarding to a SAS physical link. and is not required to ensure that any ALIGNs and/or NOTIFYs it transmits are in pairs.

> NOTE 2 - Due to clock skew ALIGN and NOTIFY removal, the STP target port may not receive a pair of ALIGNs and/or NOTIFYs every 256 dwords, even though the STP initiator port transmitted at least one pair. However, the rate of the dword stream allows for ALIGN or NOTIFY insertion by the STP/SATA bridge. One possible implementation is for the STP/SATA bridge to delete all ALIGNs and NOTIFYs received by the STP target port and to insert two consecutive ALIGNs at the SATA host port when its elasticity buffer is empty or when 254 non-ALIGN dwords have been transmitted. It may need to buffer up to 2 dwords concurrently being received by the STP target port while it does so.

## Annex G

### (informative)

### ALIGN and/or NOTIFY insertion summary

Table G.1 shows all the possible combinations of ALIGN and/or NOTIFY insertion rates for clock skew management (see 7.3), rate matching (see 7.13), and STP initiator port throttling (see 7.17.2).

**Table 2 — ALIGN and NOTIFY insertion rate examples**

| Physical link rate | Connection rate | Type of dword stream | ALIGN/NOTIFY insertion rate (per dword) | Minimum number of ALIGNs/ NOTIFYs ~~within any 2 048 dword window~~ within the specified window |
|---|---|---|---|---|
| 3.0 Gbps | 3.0 Gbps | all but to STP target phy | ~~1 per 2 048~~ 2 per 4 096 | ~~1~~ 2 per 4 096 |
| 3.0 Gbps | 3.0 Gbps | to STP target phy | ~~1 per 2 048~~ 2 per 4 096 + 2 per 256 | ~~17~~ 34 per 4 096 |
| 3.0 Gbps | 1.5 Gbps | all but to STP target phy | ~~1 per 2 048~~ 2 per 4 096 + 1 per 2 | ~~1 024 or 1 025 (a)~~ 2 049 per 4 096 |
| 3.0 Gbps | 1.5 Gbps | to STP target phy | ~~1 per 2 048~~ 2 per 4 096 + 1 per 2 + 2 per 256 | ~~1032 or 1033 (b)~~ 2 065 per 4 096 |
| 1.5 Gbps | 1.5 Gbps | all but to STP target phy | 1 per 2 048 | 1 per 2 048 |
| 1.5 Gbps | 1.5 Gbps | to STP target phy | 1 per 2 048 + 2 per 256 | 17 per 2 048 |

~~(a) There are 2 047 dwords left after the clock skew management ALIGN/NOTIFY. These alternate between rate matching ALIGNs and other dwords. These requirements alternate every 2 048 running windows.~~

~~(b) There are 2 047 dwords left after the clock skew management ALIGN/NOTIFY. These alternate between rate matching ALIGNs/NOTIFYs and other dwords, leaving 1 024 or 1 025 dwords that are neither clock skew management nor rate matching ALIGNs/NOTIFYs. Of these, 2 per 256 (i.e., 8 per 1 024) is an STP initiator phy throttling ALIGN/NOTIFY. These requirements alternate every 2 048 running windows of 2 048 dwords.~~