Tsi620 Flow Control Application Note

80D7000_AN001_02

August 7, 2009

6024 Silver Creek Valley Road San Jose, California 95138
Telephone: (408) 284-8200 • FAX: (408) 284-3572
Printed in U.S.A.

GENERAL DISCLAIMER

Integrated Device Technology, Inc. ("IDT") reserves the right to make changes to its products or specifications at any time, without notice, in order to improve design or performance. IDT does not assume responsibility for use of any circuitry described herein other than the circuitry embodied in an IDT product. Disclosure of the information herein does not convey a license or any other right, by implication or otherwise, in any patent, trademark, or other intellectual property right of IDT. IDT products may contain errata which can affect product performance to a minor or immaterial degree. Current characterized errata will be made available upon request. Items identified herein as "reserved" or "undefined" are reserved for future definition. IDT does not assume responsibility for conflicts or incompatibilities arising from the future definition of such items. IDT products have not been designed, tested, or manufactured for use in, and thus are not warranted for, applications where the failure, malfunction, or any inaccuracy in the application carries a risk of death, serious bodily injury, or damage to tangible property. Code examples provided herein by IDT are for illustrative purposes only and should not be relied upon for developing applications. Any use of such code examples shall be at the user's sole risk.

The IDT logo is registered to Integrated Device Technology, Inc. IDT and CPS are trademarks of Integrated Device Technology, Inc.

.

“Accelerated Thinking” is a service mark of Integrated Device Technology, Inc.

1. Tsi620 Flow Control Application Note

This document describes how the Tsi620 feature called “Bridge Buffer Release Management (BRM)” can avoid or limit priority-based starvation that may occur during congestion conditions. It discusses the following topics:

Revision History

80D7000_AN001_02, Formal, August 2009

There are no technical changes made to this version.

80D7000_AN001_01, Formal, January 2009

This is the first version of the *Tsi620 Flow Control Application Note*.

1.1 Tsi620 Buffer Release Management

In priority-based protocols, reordering is required to avoid deadlock situations. Deadlock occurs when buffers are occupied by transactions that cannot make forward progress. Reordering helps prevent deadlock situations by allowing higher priority transactions to complete ahead of lower priority transactions.

To use Tsi620 BRM, refer to the Tsi620_BRM_config.txt script that is distributed with the JTAG Register Access Software for the IDT’s “Tsi” RapidIO devices.

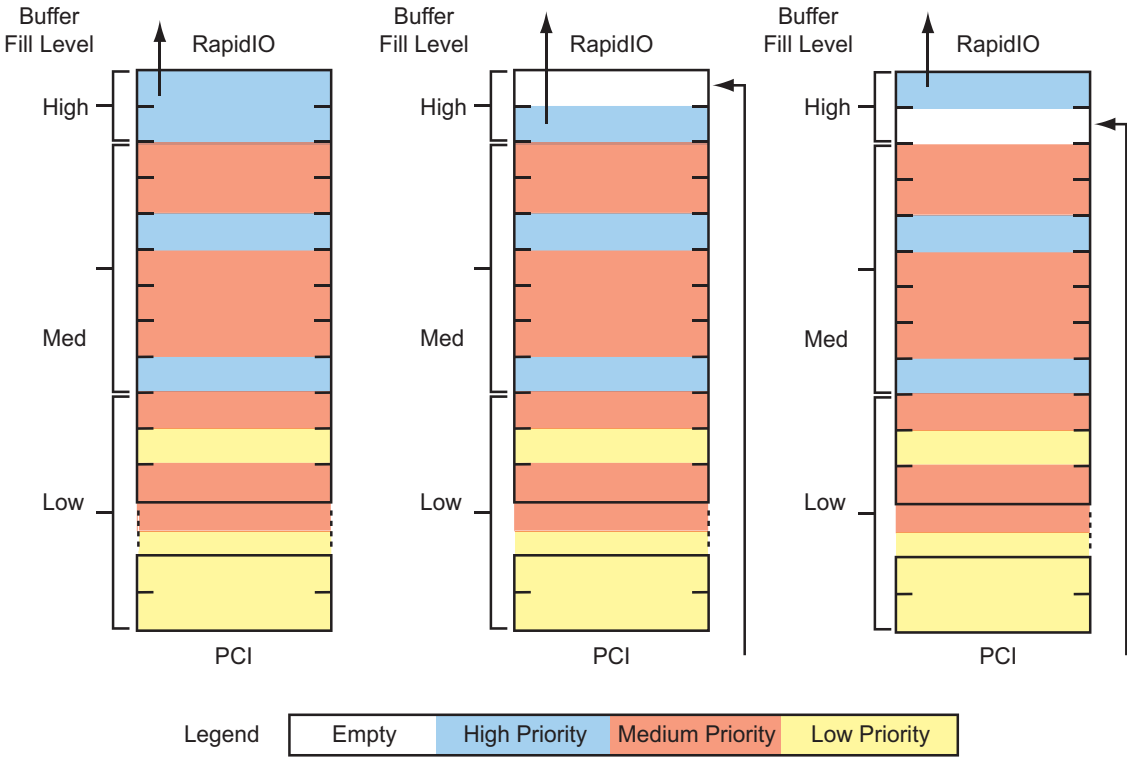
RapidIO and PCI both allow transaction reordering based on priority. With the PCI protocol, priority is associated with transaction type: writes can be sent ahead of read responses, and both writes and read responses can be sent ahead of read requests. RapidIO uses a numeric priority scheme, with 3 as the highest priority and 0 as the lowest. Higher priority packets can be sent ahead of lower priority packets. PCI transactions map to RapidIO priorities as follows:

- PCI writes – RapidIO priority 2
- PCI read responses – RapidIO priority 1
- PCI read requests – RapidIO priority 0

The PCI/RapidIO priority mapping preserves the PCI reordering necessary to avoid deadlocks.

A side effect of reordering is that low rates of higher priority transactions can starve lower priority packets during congestive conditions. As shown in Figure 1, as higher priority transactions are completed, they free up buffers that can be occupied only by other higher priority transactions. At the far left, the buffer is completely full and a high priority packet is being transferred out to RapidIO. In the middle, another high priority packet is being transferred out to RapidIO while a new high priority packet is being received into the buffer emptied in the previous step. At the far right, again only high priority packets are sent and received. The “ping pong” behavior causes starvation of lower priority packets.

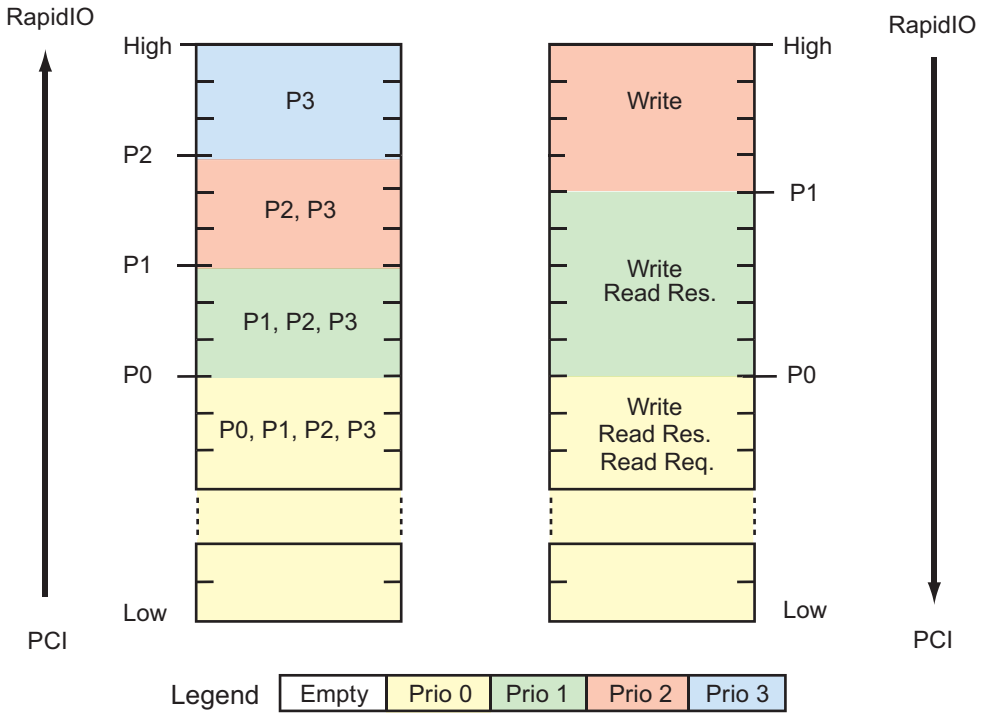
Figure 1: Buffer Management — High and Low Priority Packets



1.2 The Basic Mechanism

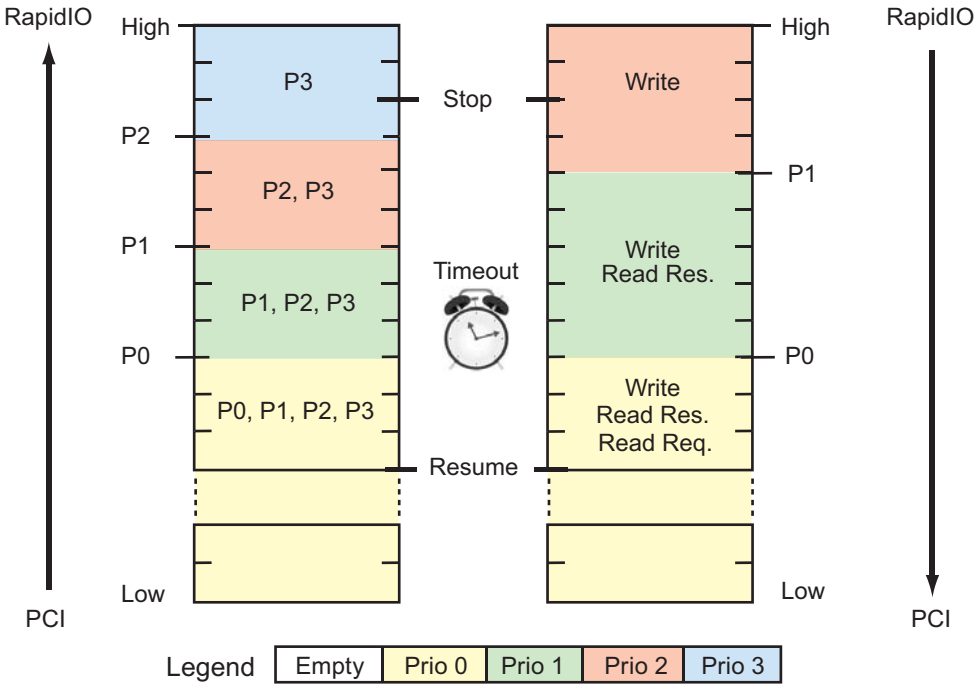
The Serial RapidIO EndPoint (SREP) in the Tsi620 allocates buffer space based on priority. Watermarks are the buffer fill levels that determine how many buffers can be used for packets of a given priority and above. Figure 2 shows how buffers are allocated for different RapidIO packet priorities and different types of PCI transactions. Note that PCI transactions have three priorities, while RapidIO packets have four.

Figure 2: I2R and R2I Watermarks



The Tsi620 BRM feature forces multiple transactions to complete before allowing more transactions to be accepted. This creates a temporary congestion-free situation whereby reordering behavior is prevented. The BRM feature is based on two buffer fill level settings, known as STOP and RESUME (see Figure 3).

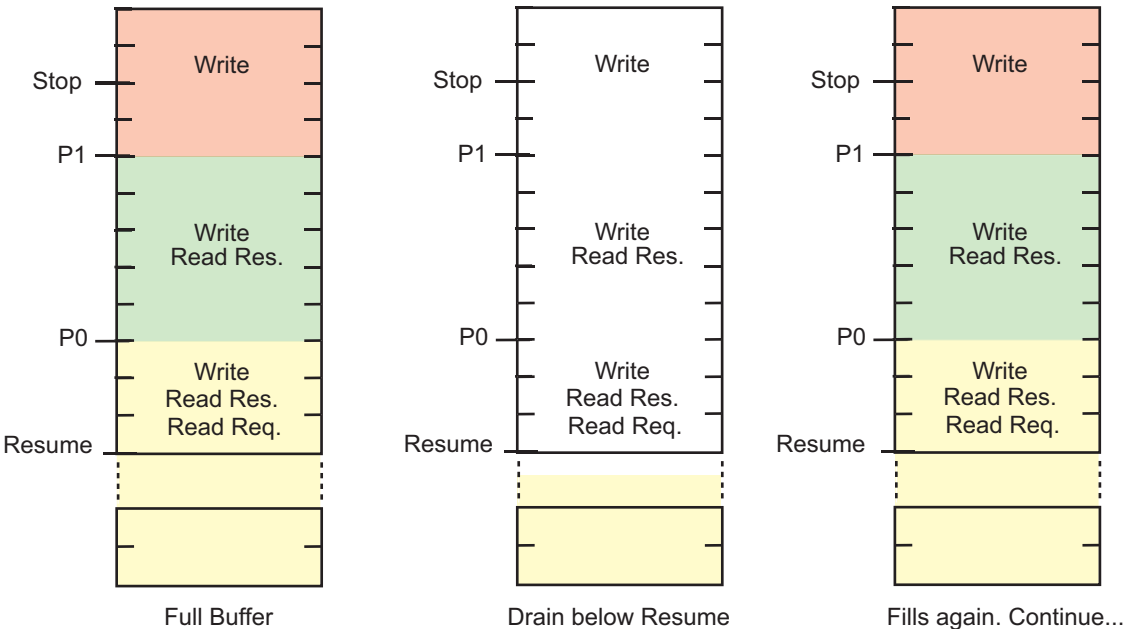
Figure 3: BRM Resume and Stop Levels Relationship to Watermarks



When the buffer fill level reaches the STOP point, the SREP stops informing the Bridge ISF/Switch ISF of buffers that are freed by completed transactions (see Figure 4). The Bridge ISF/Switch ISF stops forwarding packets, and the buffer fill level eventually drops to the RESUME point. Since the STOP setting is above the watermark for High priority packets, and the RESUME setting is below the watermark for Low priority packets, packets of all priorities can make forward progress as the buffer fill level drops from the STOP point to the RESUME point.

Once the RESUME point is reached, the Bridge ISF/Switch ISF is informed of the actual buffer fill level, and packets of all priorities can begin to flow into the buffers. Since the RESUME point is below the watermark for Low priority packets, and many buffers are now available, packets of all priorities can flow into the buffer. As a result, this buffer mechanism helps prevent priority-based starvation.

Figure 4: Buffer Release Management Operation



Under rare traffic conditions, the BRM mechanism may cause a deadlock by preventing the forward progress of higher priority packets required to complete outstanding transactions. To avoid deadlocks, the BRM sets a maximum time to be in the STOP condition. Once this timeout expires, two possible behaviors can be selected:

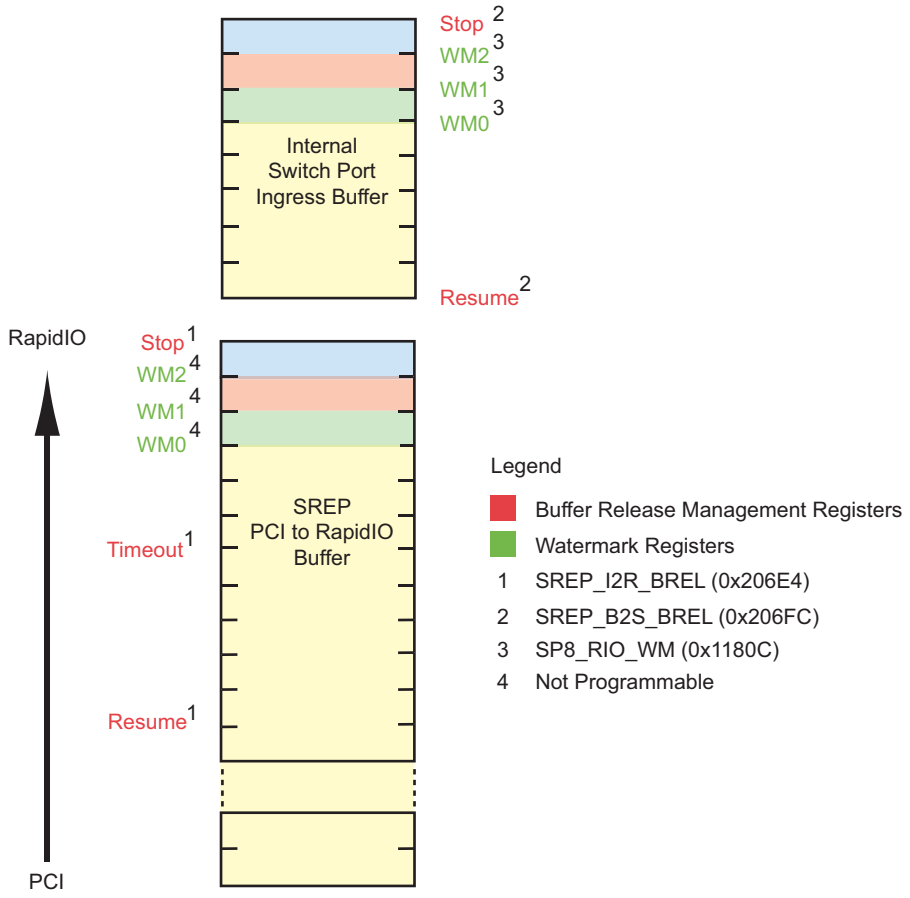
- Do not engage BRM until the RESUME value is reached — This disables BRM until the congestive condition no longer exists. This is the preferred mode of operation when periods of congestion are short and/or the probability of deadlock is high. This can lead to long periods of priority-based starvation, but prevents long periods where no packets are forwarded due to BRM.
- Re-engage BRM if the STOP level is reached again — This is the preferred mode of operation when periods of congestion are long and the probability of deadlock is low. This avoids priority-based starvation at the expense of long periods where no packets are forwarded when a deadlock occurs.

1.3 PCI-to-RapidIO Buffer Release Management

In general, the PCI bus cannot congest the Tsi620 Switch. The PCI bus total bandwidth is approximately 2 Gbps, which can be handled by at most two RapidIO 1x links. However, in combination with RapidIO-to-RapidIO traffic through the Tsi620 Switch, the switch can become congested and result in starvation of PCI reads and read responses.

Figure 5 outlines the registers that are responsible for various settings for PCI-to-RapidIO buffer release management. The recommended register values are found in the “Tsi620_BRM_config.txt Script Contents”.

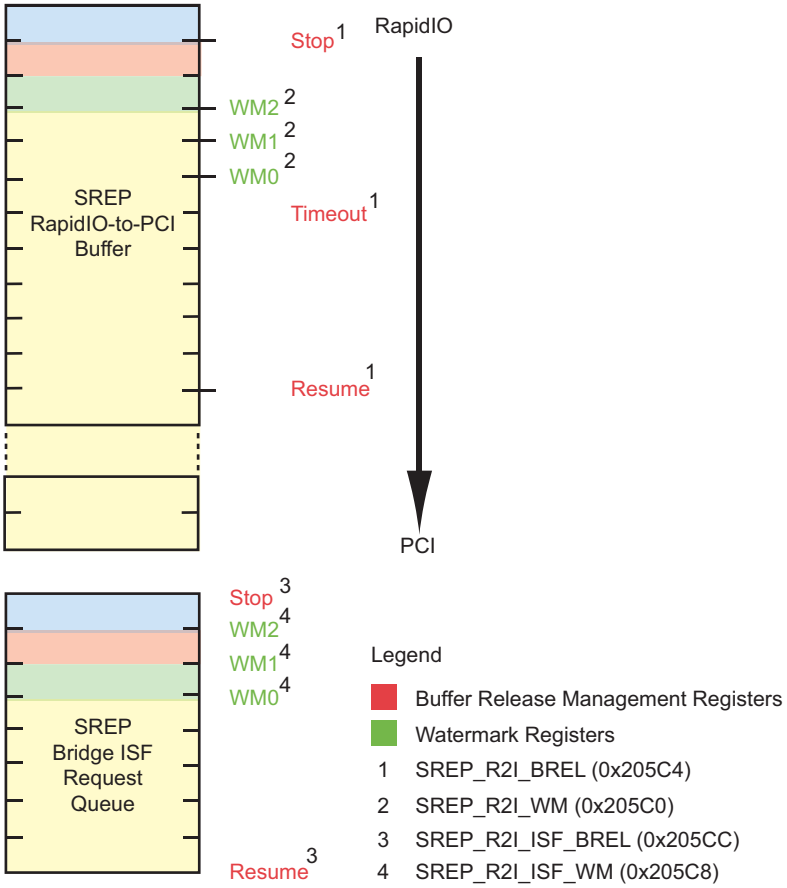
Figure 5: PCI-to-RapidIO Watermark and BRM Registers



1.4 RapidIO-to-PCI Buffer Management

Figure 6 describes the registers that are responsible for various settings for RapidIO-to-PCI buffer release management. The recommended register values are found in the "Tsi620_BRM_config.txt Script Contents".

Figure 6: RapidIO-to-PCI Watermark and BRM Registers

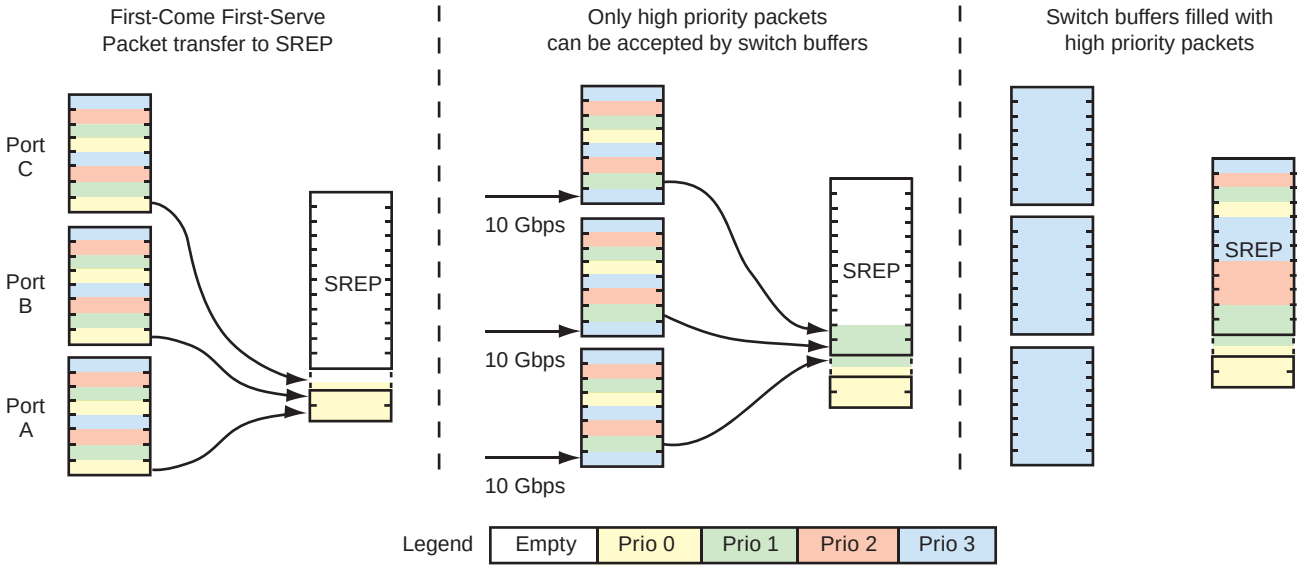


Note that short-term bursts of high priority packets can temporarily starve low priority packets, even if RapidIO-to-PCI buffer management is active. The mechanism is illustrated in Figure 7.

The basic cause of the short-term starvation is that the switch's ports always try to keep their ingress buffers full, which is an expected behavior. The left-most panel shows that the switch buffers are full of packets with a variety of priorities. Assume that SREP has just reached the RESUME threshold, and so packets are allowed to flow into SREP from the switch buffers. In the left-most panel of the diagram, the bottom switch buffer forwards a packet to SREP followed by the other switch buffers in round-robin order.

As shown in the center panel, once the first packet has been forwarded, only a high priority packet can be received into the switch buffer from the RapidIO link. This behavior is duplicated in the other two buffers so that the Switch buffers fill with priority 3 packets. The right-most panel shows SREP full of packets with a variety of priorities, but now the switch buffers have only high-priority packets available to fill the SREP buffers.

Figure 7: Receiver Based Flow Control Only Accepts High Priority Packets

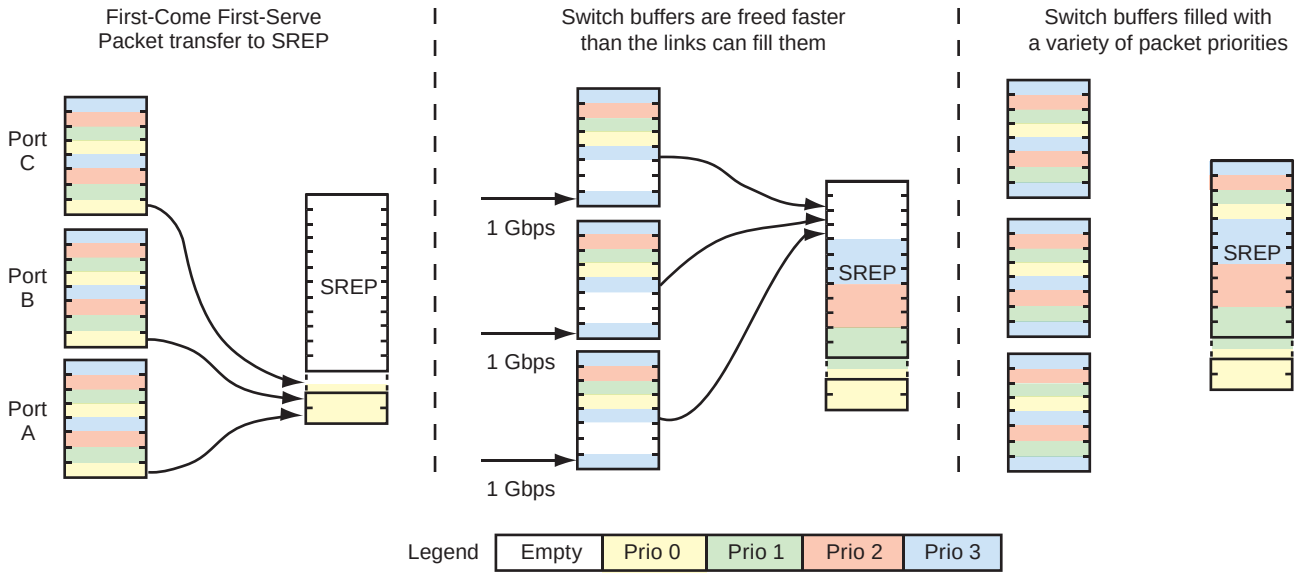


To avoid temporary starvation of low priority packets, the SREP must always accept packets faster than the input ports can receive and forward them (see Figure 8), or the amount of high priority traffic routed to SREP must be less than 10 Gbps.

On average, these conditions must be true; otherwise, the traffic sent to the PCI bus exceeds the capacity of the PCI bus.

These conditions must be true in general, otherwise the bandwidth mismatch causes the switch to be permanently congested due to too much data being set to the PCI bus. If the combined bandwidth forwarded to SREP is less than 10 Gbps, then SREP can accept packets faster than packets can be accepted from the link; therefore, starvation of lower priority packets can be avoided (see Figure 8). The starting conditions in the left-most panel are the same as in Figure 7, but the rate at which the switch buffers are filled is now 1 Gbps. This means that packets can be transferred to SREP buffers over three times faster than packets can be received by the Switch buffers, as shown in the middle panel. The result is that SREP and the Switch buffers can be filled with packets that have a variety of priorities, which avoids starvation of low priority packets.

Figure 8: Low Bandwidth Links Avoid Short Term Starvation of Low Priority Packets



If any 4x ports send packets to the SREP, the SREP may be able to accept lower priority packets faster than the 4x ports can receive them. If the total bandwidth of the ports that can send to the SREP is greater than 10 Gbps, receiver-based flow control will select higher priority packets for acceptance into the Switch buffers. Short-term starvation of low priority packets may occur in this situation.

1.5 Tsi620_BRM_config.txt Script Contents

Regardless of port width and lane speed, the following configuration is recommended for all RapidIO ports that forward traffic to SREP:

- TRANS_MODE to 1 in SPx_CTL_INDEP (Store-and-forward mode)
- SPx_RIO_WM to 0x00010203 (Minimal buffers reserved for high-priority packets)

In addition, the FAB_CTL[IN_ARB_MODE] bit must be set to 0 (first come, first serve), as this mode presents packets in an order that optimizes the benefits of BRM.

The configuration recommendations above are not implemented in the Tsi620_BRM_config.txt script, as this would override user-specific settings of other bits within the SPx_CTL_INDEP and FAB_CTL registers.

```
// This script configures the Tsi620 Buffer Release Management
// functions in both PCI-to-RapidIO, and RapidIO-to-PCI directions,
// using default settings.
//
// PCI to RapidIO Buffer Release Management Configuration
//
// Assumes standard PCI-to-RapidIO priority mapping:
// SREP_I2R_LUT_TA_LOWER.RD_PRIO = 0,
// SREP_I2R_LUT_TA_LOWER.WR_PRIO = 2
//
// Sets timeouts to maximum, and re-engages BRM after timing out if
// the buffer fill level hits STOP. These settings assume that the
// probability of deadlocks is low, and periods of congestion are long.
//

w 206e4 0x80FF1C04 // SREP_I2R_BREL
w 206FC 0x80FF0702 // SREP_B2S_BREL
w 1180c 0x00010203 // SP8_RIO_WM

// RapidIO to PCI Buffer Release Management Configuration
//
// Assumes the following:
// One buffer is sufficient to deal with decomposed PCI transactions.
// - This means that PCI reads must be less than 256 bytes in size.
// Sets timeouts to maximum, and re-engages BRM after timing out if
// the buffer fill level hits STOP. These settings assume that the
// probability of deadlocks is low, and periods of congestion are
// long.
//

w 205b0 0x00000080 // SREP_R2I_ISF_REQ_PRIO_CSR
w 205b4 0x11220000 // SREP_R2I_ISF_RESP_PRIO_CSR
w 205c0 0x01010203 // SREP_R2I_WM
w 205C4 0x80FF1D02 // SREP_R2I_BREL
w 205C8 0x00010203 // SREP_R2I_ISF_WM
```

```
w 205CC 0x80FF0802 // SREP_R2I_ISF_BREL
w 206F0 0x00010203 // SREP_NWR_ERR_WM
w 206F4 0x80FF1D02 // SREP_NWR_ERR_BREL
```

IMPORTANT NOTICE AND DISCLAIMER

RENESAS ELECTRONICS CORPORATION AND ITS SUBSIDIARIES (“RENESAS”) PROVIDES TECHNICAL SPECIFICATIONS AND RELIABILITY DATA (INCLUDING DATASHEETS), DESIGN RESOURCES (INCLUDING REFERENCE DESIGNS), APPLICATION OR OTHER DESIGN ADVICE, WEB TOOLS, SAFETY INFORMATION, AND OTHER RESOURCES “AS IS” AND WITH ALL FAULTS, AND DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT OF THIRD-PARTY INTELLECTUAL PROPERTY RIGHTS.

These resources are intended for developers who are designing with Renesas products. You are solely responsible for (1) selecting the appropriate products for your application, (2) designing, validating, and testing your application, and (3) ensuring your application meets applicable standards, and any other safety, security, or other requirements. These resources are subject to change without notice. Renesas grants you permission to use these resources only to develop an application that uses Renesas products. Other reproduction or use of these resources is strictly prohibited. No license is granted to any other Renesas intellectual property or to any third-party intellectual property. Renesas disclaims responsibility for, and you will fully indemnify Renesas and its representatives against, any claims, damages, costs, losses, or liabilities arising from your use of these resources. Renesas' products are provided only subject to Renesas' Terms and Conditions of Sale or other applicable terms agreed to in writing. No use of any Renesas resources expands or otherwise alters any applicable warranties or warranty disclaimers for these products.

(Disclaimer Rev.1.01)

Corporate Headquarters

TOYOSU FORESIA, 3-2-24 Toyosu,
Koto-ku, Tokyo 135-0061, Japan
www.renesas.com

Contact Information

For further information on a product, technology, the most up-to-date version of a document, or your nearest sales office, please visit www.renesas.com/contact-us/.

Trademarks

Renesas and the Renesas logo are trademarks of Renesas Electronics Corporation. All trademarks and registered trademarks are the property of their respective owners.