Performance Tuning for the SG2010 PCI-to-StarFabric Bridge

Revision Information: Revision 1.0

StarGen, Inc., 225 Cedar Hill Street, Suite 22, Marlborough, MA 01752
www.stargen.com

July 2003

StarGen, Inc. believes the information in this publication is correct; however, the information is subject to change without notice.

StarGen, Inc. does not claim that the use of its products in the manner described in this publication will not infringe on any existing or future patent rights, nor do the descriptions contained in this publication imply the granting of licenses to make, use, or sell equipment or software in accordance with the description.

Contents

Introduction

Performance for Write Traffic

Performance for Read Traffic

Appendix A

Introduction

StarFabric performance tuning is generally done in the edge nodes like the SG2010 PCI to StarFabric Bridge. Switches forward frames as they are received, and primarily affect performance through credit allocation and frame arbitration. In StarGen's SG1010 switch device, credit allocation is programmable but frame arbitration is not.

In the SG2010 PCI-to-StarFabric bridge, key performance parameters include the following:

- PCI Bus width (32bit vs. 64bit)
- PCI Bus speed (33Mhz vs. 66Mhz)
- The burst length of the transactions on the PCI bus
- How the resulting data payload is packed and unpacked into StarFabric frames
- How incoming frame data is driven onto the PCI bus.

This document describes SG2010 tuning bits that affect some of these parameters for reads and writes. Register space in StarFabric devices can be accessed by the following three methods:

- PCI BAR0 memory access
  - an offset from the BAR0 memory address assigned by the BIOS or O/S
- Channel 255 access (via a StarFabric "connection")
  - an offset from a memory "handle" after a connection is made
  - this offset is the same as the PCI BAR0 offset
- PCI Configuration access for the SG2010 Gateway function

This document provides the register offset of the tuning bits for all three of the methods listed.

References and Additional Information

If you need additional information, please contact StarGen at support@stargen.com or refer to one or more of the following reference documents:

PCI Special Interest Group (PCISIG) Specifications

PCI Local Bus Specification, Revision 2.2

PCI-to-PCI Bridge Architecture Specification, Rev 1.1

StarGen Specifications

SG2010 Hardware Reference Manual

SG2010 Data Sheet

StarFabric Protocol Reference Manual

Fabric Programmer's Manual

Revision History

Revision Number	Date mm/dd/yy	Description
0.1	06/10/02	Initial Revision
1.0	07/17/03	1. Fixed grammatical errors throughout 2. Added to the Read Data Retention section 3. Changed the General section to Introduction

2

Performance for Write Traffic

2.1 Write Traffic from PCI to the Fabric

StarFabric write frames have a maximum payload of 128 bytes (32 dwords). StarFabric link/protocol overhead is minimized if 128 byte, or larger, frames are used. In terms of packing write data from the PCI bus into frames, matching the burst length to the frame size should optimize performance results. Burst size is generally controlled by the PCI bus master. For applications that have the ability to control burst size, 32 dword or larger bursts will produce the best fabric performance.

2.2 Write Traffic from the Fabric to PCI

Burst length for writes going from the SG2010 fabric interface to the PCI bus are controlled by the following:

- Write Combining - Combines two or more write frames into a single PCI burst. This is only effective when two or more contiguous write frames are received. Write combining is enabled by setting the Write Combine Enable (Bit 12) in the Chip Control register. This is located at Gateway configuration offset 94h. It is also mapped at offset 5994h from either the PCI BAR0 address or the Channel 255 StarFabric memory “handle”.
- Master Latency Timer (MLT)- 8-bit value generally set by the BIOS to a reasonably large value. This only has an affect when the PCI grant is removed from the master. A typical setting is 40h.
  - To set the MLT for address routed traffic
    - When the SG2010 is in Root mode: Bridge MLT register located at configu ration offset 0Dh. It is also located at Channel 255/PCI BAR0 offset 580Dh.
    - When the SG2010 is in Leaf mode: Bridge Secondary MLT register located at configuration offset 1Bh. It is also located at Channel 255/PCI BAR0 offset 581Bh.
  - To set the MLT for path-routed/multicast traffic:
    - The Gateway MLT register is located at Gateway configuration offset 0Dh and at Channel 255/PCI BAR0 offset 590Dh.

- Fast Back-to-Back Enable. This does not extend burst length, but can remove the idle cycle between consecutive PCI transactions initiated by the SG2010.
  - To set Fast Back-to-back for address routed traffic
    - When the SG2010 is in Root mode: Bridge Command register, (Bit 9) located at bridge configuration offset 04h. It is also located at Channel 255/PCI BAR0 offset 5804h.
    - When the SG2010 is in Leaf mode: Bridge Control register, (Bit 7) located at bridge configuration offset 3Eh. It is also located at Channel 255/PCI BAR0 offset 583Eh.
  - To set Fast Back-to-back for path routed traffic
    - Gateway Command register, Bit 9 located at Gateway configuration offset 04h. Also located at Channel 255/PCI BAR0 offset 5904h.

3

Performance for Read Traffic

3.1 Read Performance Overview

Writes always perform better than Reads and are highly recommended when there is an opportunity to choose. Reads have lower performance for the following reasons:

- Reads require a round trip to complete (a request from the origin to the terminus, and a completion from the terminus to the origin).
- PCI protocol does not have a mechanism for a PCI bus master to specify how much read data it wants. There are three different read commands that provide "hints" to the target for how much read data to get. These are Memory Read, Memory Read Line and Memory Read Multiple.

3.2 Read Data Retention Feature

The SG2010’s *read data retention* feature can reduce the number of round trips through the fabric for read transactions initiated by a PCI bus master. This feature enables the SG2010 to retain prefetched read data in it’s read data buffers even if the master that initiated the read disconnects the transaction. The retained data is thrown away if the read data queue or the delayed transaction queue fill up. Read data retention is enabled by setting bit 11 in the Chip Control register located at Gateway configuration offset 94h. This register can also be accessed at Channel 255/PCI BAR0 offset 5994h.

There may be negative side effects when using this feature. Since prefetched read data can be held for an indeterminate amount of time, it can become “stale” if the source of the data is updated after it has been prefetched but before before it is consumed by the master. System designers must understand this and determine if issues exist before using Read Data Retention.

3.3 Cache Line Size Register

The Cache Line Size register is at offset 0Ch in Gateway configuration space. It is also located at Channel 255/PCI BAR0 offset 590Ch for path-routed reads. In addition, access is available in the bridge function configuration space for address-routed reads at

Channel 255/PCI BAR0 offset 580Ch. The Cache Line Size register is generally programmed by the PCI BIOS, but is mentioned here because it has an impact on read performance as discussed in Section 3.4.1.

3.4 Programming Read Data Prefetch Amounts

The SG2010 provides the ability to program read prefetch amounts for each of the three PCI Read Commands.

Note that it is not always better to increase the amount of read data prefetched. If the read data retention feature is turned off, or if the PCI bus master does not consume a large amount of read data from a contiguous address space, the extra prefetched data is thrown away. The PCI bus time spent to read the data from the target is wasted in these cases.

3.4.1 Read Requests from the PCI Bus into the Fabric

The SG2010 maps a PCI read command into a StarFabric "Read Request" frame that specifies the number of DWORDS to read from the target. The PCI read command is not carried along with the frame. By default, the number of dwords read is based on the PCI read command type, the cache line size, and the SG2010 programmable prefetch amounts. Each read command type has a corresponding programmable prefetch amount. These values are found in the SG2010 Chip Control register located at Gateway configuration offset 94h. They are also located at Channel 255/PCI BAR0 offset 5994h. Default values for each command are 00b. The programmable settings are described below.

- Amount prefetched for memory read commands, bit field [5:4]
  - 00: 1 cache line
  - 01: 2 cache lines
  - 10: 4 cache lines
  - 11: 1 dword
- Amount prefetched for memory read line commands, bit field [7:6]
  - 00: 1 cache line
  - 01: 2 cache lines
  - 10: 4 cache lines
  - 11: 8 cache lines
- Amount prefetched for memory read multiple commands, bit field [9:8]
  - 00: 2 cache line
  - 01: 4 cache lines
  - 10: 8 cache lines
  - 11: 16 cache lines

3.5 Prescriptive Reads

The SG2010's *prescriptive read* feature enables a user/master device to specify the exact number of Dwords to be read from a target. This feature requires StarFabric path-routing. When a PCI transaction is mapped into a prescriptive read, 1 to 512 Dwords are specified at the time the StarFabric path connection is set up (in single-dword granularity). The SG2010 at the terminus end of the fabric will issue the number of read commands necessary to obtain the "prescribed" number of Dwords. To use a prescriptive read, the following must be done:

- The Segment Table entry used to create the read request frame must have the Source Channel Enable set, and must specify a Source Channel Number
  - The Source Channel Enable is bit 0 of Byte 3 in each 8-byte segment table entry
  - The Source Channel Number is bit [2:0] of Byte 4 in each 8-byte segment table entry
- The corresponding Source Channel must have prescriptive reads enabled, and specify a prescriptive read amount.
  - The Prescriptive Read Enable is bit 10 at Channel 255/PCI BAR0 offset 5CxEh (where x is Source Channel number from 0-7)
  - The Prescriptive Read Amount is specified by bits [9:0] at Channel 255/PCI BAR0 offset 5CxEh (where x is Source Channel number from 0-7)

Care must be taken when using prescriptive reads. When used, the SG2010 holds onto the read data regardless of the state of the read request and read completion queues. It will only throw the data away if the master time-out timer expires (2^^15 or 2^^10 cycles). This can lead to performance degradation unless it is used carefully.

For more information on the SG2010 Segment table, refer to section 4.6.13 of the SG2010 Hardware Reference Manual. See section 3.3.2.4 of the same document for more information on prescriptive reads.

3.6 Managing Delayed Transaction Request Queues

The SG2010 has an eight entry delayed transaction buffer to hold information for incoming PCI transactions. By default, three of the eight entries are reserved for Addr/Async, Isoc/Hp-Isoc, and Prov/Hp-Async Classes-of-service respectively. The other five entries are first-come, first-served regardless of the CoS. If it's possible for an application to have more than five outstanding reads at one time, and they all use the same CoS, a user can change the reservation mode so that all eight entries can be used by any CoS. This is done by setting bit 22 in the Chip Control register located at Gateway configuration offset 94h, Channel 255/PCI BAR0 offset 5994h.

3.7 Requesting Data from the PCI Target

The SG2010 maps the dwords requested into a memory read command. The specific command type is not programmable. The main factor that the SG2010 uses for selecting the specific read command is the number of Dwords specified in the StarFabric frame. The Master Latency timer (MLT) and Fast back-to-back enable (FBB) are programmable values that can affect the SG2010's read performance as a master. Both of these settings are typically programmed by PCI BIOS. The Primary MLT is located at configuration offset 0Dh for each SG2010 function (Bridge and Gateway). The bridge function Secondary MLT is located at configuration offset 1Bh. These are dual-mapped at Channel 255/BAR0 offsets 590Ch and 591Bh respectively. A minimum MLT setting of 40h is recommended.

If a read is speculative, the SG2010 performs one read transaction, and whatever read data is received, it returns it and stops there. Therefore, it is possible to request 64 dwords but only get 8!

If a read is prescriptive, the SG2010 performs as many read transactions as it needs to in order to collect the amount of read data requested.

3.8 Returning Data Back to the PCI Master

Read data retention has the biggest performance impact. This was described in Section 3.2. Prescriptive reads always retain data. In order to enable read data retention for speculative reads, set the Read Data Retention bit in the Chip Control register (Gateway configuration offset 94h, Channel 255/PCI BAR0 offset 5994h).

If read data retention causes problems with master time-outs, one can reduce the master time-out value from 2^^15 to 2^^10 cycles. For address-routed reads this is done by settting bit 8 for a root and bit 9 for a leaf in the Bridge Control register (Bridge Configuration offset 3Eh. Channel 255/PCI BAR0 offset 583Eh). For path-routed reads this is done by setting bit 10 in the Chip Control register (Gateway configuration offset 94h, Channel 255/PCI BAR0 offset 5994h).

4

Appendix A

2.1 Performance Tuning SROM Files

The following SROM files can be used to set read and write performance tuning bits during the SG2010 SROM preload. In addition to performance tuning, all of the example SROM files set BARS 2-5 of the SG2010. BAR 2 is set as prefetchable. If this is not desired, the bold text can be extracted from the file to create a new SROM file with just performance tuning. The StarGen's software application PVX and StarView can be used to program the SROMs. The following PVX commands should be used.

bar0_address: BAR0 address of the SG2010's Other Bridge (as seen in PVX) funtion.

```
srom -e bar0_address                  //Erases all data currently in SROM

srom -b bar0_address filename.txt     //Reprograms SROM
```

The following StarView commands should be used.

FID: The Fabric ID of the SG2010 you are programming. Provided by the StarView interface (Ex. 0/0/R).

```
srom -e FID                           //Erases all data currently in SROM

srom -b FID filename.txt              //Reprograms SROM with filename.txt
```

Please refer to the StarGen Software Development Kit for more details. If PVX is being used, please note that all the comments should be removed prior to use.

2.1.1 Read Performance Tuning

2.1.1.1 Read Data Retention

The following SROM file can be used to enable read data retention (Bit 11 -> 1 Channel 255/BAR0 offset 5994) in the SG2010.

2010_rdr.txt

```
//Triggers the start of the SROM preload
0x86,

//PCI BAR2 Gateway Setup Register
0x80,0x59,
0x04,
0x08,0x00,0x80,0xff,

//PCI BAR3 Gateway setup register
0x84,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR4 Gateway setup register
0x88,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR5 Gateway setup register
0x8C,0x59,
0x04,
0x00,0x00,0x80,0xff,

//Gateway Chip Control Register. Enables Read Data Retention.
0x95,0x59,          //Address -> 5995
0x01,               //# of bytes to program = 1
0x08,               //data to write

//Triggers the end of the SROM Preload
0x00,0x00,0x00,0x00,
0x00,0x00,0x00,0x00
```

2.1.1.2 Read Data Prefetch Amounts

2.1.1.2.1 Setting Prefetch Amounts to Maximum Values

The following SROM file can be used to set the read prefetch amounts to their maximum values. The bits below can be found at Channel 255/BAR0 offset 5994.

- Bits 4:5 -> 10b Memory Read Prefetch Amount for speculative memory reads is set to 4 cache lines.
- Bits 6:7 -> 11b Memory Read Line Prefetch Amount for speculative memory reads is set to 8 cache lines.
- Bits 8:9 -> 11b Memory Read Multiple Prefetch Amount for speculative memory reads is set to 16 cache lines.

2010_pmx.txt

```
//Triggers the start of the SROM preload
0x86,

//PCI BAR2 Gateway Setup Register
0x80,0x59,
0x04,
0x08,0x00,0x80,0xff,
//PCI BAR3 Gateway setup register
0x84,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR4 Gateway setup register
0x88,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR5 Gateway setup register
0x8C,0x59,
0x04,
0x00,0x00,0x80,0xff,

//Gateway Chip Control Register. Sets Prefetch Amounts.
0x94,0x59,              //Address -> 5994
0x02,                   //# of bytes -> 2
0xe0,0x03,              //first byte -> e0, second byte -> 03

//Triggers the end of the SROM Preload
0x00,0x00,0x00,0x00,
0x00,0x00,0x00,0x00
```

2.1.1.2.2 Setting Prefetch Amounts to Intermediate Values

The following SROM file can be used to set the following bits of Channel 255/BAR0 offset 5994 accordingly.

- Bits 4:5 -> 01b Memory Read Prefetch Amount for speculative memory reads is set to 2 cache lines.
- Bits 6:7 -> 01b Memory Read Line Prefetch Amount for speculative memory reads is set to 2 cache lines.
- Bits 8:9 -> 01b Memory Read Multiple Prefetch Amount for speculative memory reads is set to 4 cache lines.

2010_pmd

```
//Triggers the start of the SROM preload
0x86,

//PCI BAR2 Gateway Setup Register
0x80,0x59,
0x04,
0x08,0x00,0x80,0xff,

//PCI BAR3 Gateway setup register
0x84,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR4 Gateway setup register
0x88,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR5 Gateway setup register
0x8C,0x59,
0x04,
0x00,0x00,0x80,0xff,

//Gateway Chip Control Register. Sets Prefetch Amounts.
0x94,0x59,
0x02,
0x50,0x01,

//Triggers the end of the SROM Preload
0x00,0x00,0x00,0x00,
0x00,0x00,0x00,0x00
```

2.1.2 Write Performance Tuning

2.1.2.1 Write Combining

The following SROM file can be used to enable write combining in the SG2010 (Bit 12 -> 1 Channel 255/BAR0 offset 5994).

2010_wc

```
//Triggers the start of the SROM preload
0x86,

//PCI BAR2 Gateway Setup Register
0x80,0x59,
0x04,
0x08,0x00,0x80,0xff,

//PCI BAR3 Gateway setup register
0x84,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR4 Gateway setup register
0x88,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR5 Gateway setup register
0x8C,0x59,
0x04,
0x00,0x00,0x80,0xff,

//Gateway Chip Control Register. Enables write combining.
0x95,0x59,
0x01,
0x10,

//Triggers the end of the SROM Preload
0x00,0x00,0x00,0x00,
0x00,0x00,0x00,0x00
```

2.1.2.2 Fast-Back-to-Back

The following SROM files can be used to enable fast back-to back in the SG2010. Three example SROM files are provided because the required settings are dependent upon the type of traffic and whether the SG2010 is set as a ROOT.

Address Routed Traffic

- SG2010 in ROOT mode: Bit 9 -> 1 Channel 255/BAR0 offset 5804h.
- SG2010 in LEAF mode: Bit 7-> 1 Channel 255/BAR0 offset 583Eh.

Path Routed Traffic

- Bit 9 -> 1 Channel 255/BAR0 offset 5904h.

2.1.2.2.1 Address Routed in ROOT Mode

The following SROM file should be used to enable fast back-to-back if the SG2010 is set as a ROOT and the traffic being sent is address-routed.

2010fbar

```
//Triggers the start of the SROM preload
0x86,

//PCI BAR2 Gateway Setup Register
0x80,0x59,
0x04,
0x08,0x00,0x80,0xff,

//PCI BAR3 Gateway setup register
0x84,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR4 Gateway setup register
0x88,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR5 Gateway setup register
0x8C,0x59,
0x04,
0x00,0x00,0x80,0xff,
//Bridge Command Register. Enables fast back-to-back ROOT address routed traffic.
0x05,0x59,
0x01,
0x02,

//Triggers the end of the SROM Preload
0x00,0x00,0x00,0x00,
0x00,0x00,0x00,0x00
```

2.1.2.2.2 Address Routed in LEAF Mode

The following SROM file should be used to enable fast back-to-back if the SG2010 is set as a LEAF and the traffic being sent is address-routed.

2010fbal

```
//Triggers the start of the SROM preload
0x86,

//PCI BAR2 Gateway Setup Register
0x80,0x59,
0x04,
0x08,0x00,0x80,0xff,

//PCI BAR3 Gateway setup register
0x84,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR4 Gateway setup register
0x88,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR5 Gateway setup register
0x8C,0x59,
0x04,
0x00,0x00,0x80,0xff,
//Bridge Control Register. Enables fast back-to-back for LEAF address routed traffic.
0x3E,0x59,
0x01,
0x80,

//Triggers the end of the SROM Preload
0x00,0x00,0x00,0x00,
0x00,0x00,0x00,0x00
```

2.1.2.2.3 Path Routed Traffic

The following SROM file should be used to enable fast back-to-back if the traffic being sent is path routed.

2010fbal

```
//Triggers the start of the SROM preload
0x86,

//PCI BAR2 Gateway Setup Register
0x80,0x59,
0x04,
0x08,0x00,0x80,0xff,

//PCI BAR3 Gateway setup register
0x84,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR4 Gateway setup register
0x88,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR5 Gateway setup register
0x8C,0x59,
0x04,
0x00,0x00,0x80,0xff,
//Gateway Command Register. Enables fast back-to-back for path routed traffic.
0x05,0x59,
0x01,
0x02,

//Triggers the end of the SROM Preload
0x00,0x00,0x00,0x00,
0x00,0x00,0x00,0x00
```

2.1.3 Write & Read Performance Tuning

The following SROM file enables all the performance tuning features.

2010allp

```
//Triggers the start of the SROM preload
0x86,

//PCI BAR2 Gateway Setup Register
0x80,0x59,
0x04,
0x08,0x00,0x80,0xff,

//PCI BAR3 Gateway setup register
0x84,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR4 Gateway setup register
0x88,0x59,
0x04,
0x00,0x00,0x80,0xff,

//PCI BAR5 Gateway setup register
0x8C,0x59,
0x04,
0x00,0x00,0x80,0xff,
//Gateway Command Register. Enables fast back-to-back for path routed traffic.
0x05,0x59,
0x01,
0x02,

//Bridge Control Register. Enables fast back-to-back for LEAF address routed
traffic.
0x3E,0x59,
0x01,
0x80,

//Bridge Command Register. Enables fast back-to-back ROOT address routed traffic.
0x05,0x59,
0x01,
0x02,

//Gateway Chip Control Register. Set all prefetch->MAX and RDR & WrComb
0x94,0x59,
0x02,
0xe0,0x1b,

//Triggers the end of the SROM Preload
0x00,0x00,0x00,0x00,
0x00,0x00,0x00,0x00
```