2-ch RAID0 Design (NVMe-IP) reference design manual

Rev1.0 6-Oct-17

1 Introduction

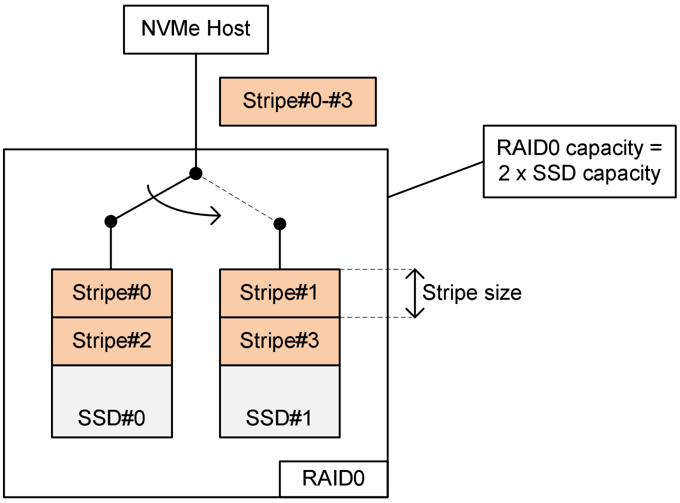

Figure 1-1 RAID0 by 2 SSDs Data Format

RAID0 system uses multiple storages to extend total storage capacity and increase write/read performance. Assumed that total number of device is N, total storage capacity is equal to N multiply by amount of storage and write and read speed are almost equal to N multiply by speed of one SSD.

Data format of RAID0 is shown in Figure 1-1. Data stream of the host side are split into a small stripe and transfer to one SSD at a time. Stripe size is the data size to store in one SSD before switching to other SSDs.

In the reference design, two SSDs are applied to run RAID0 system. Stripe size is equal to 512 bytes (one sector unit). Two SSDs connecting in the system should be same model to get the best performance and correct capacity. By using RAID0, the total capacity is equal to two times of SSDs and the performance for both write and read are almost two times. In our test system, Write speed of RAID0 NVMe is about 4200 MB/s and Read speed is about 6200 MB/s. (Performance from NVMe-IP demo by using one SSD are 2100 MB/s for write command and 3200 MB/s for read command).

User can modify RAID0 reference design to increase the numbers of NVMe SSD to achieve the better performance and bigger disk capacity.

2 Hardware overview

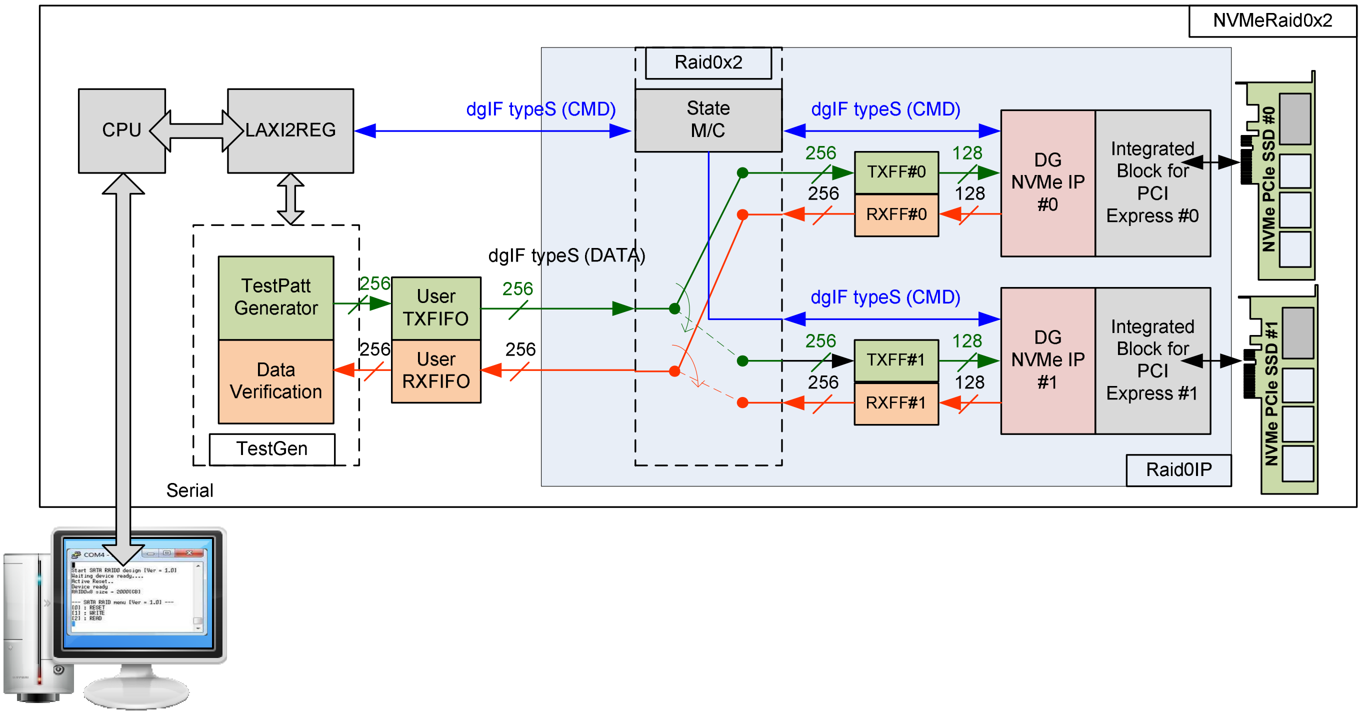

Figure 2-1 RAID0x2 Demo System by using NVMe-IP

RAID0x2 demo is modified from NVMe-IP standard reference design. Please see more details of standard reference design from following link.
http://www.dgway.com/products/IP/NVMe-IP/dg_nvmeip_refdesign_en.pdf
http://www.dgway.com/products/IP/NVMe-IP/dg_nvmeip_instruction_en.pdf

To support RAID0 operation, Raid0x2 module is designed to be the interface block between user logic and two NVMe-IPs. To support higher bandwidth, data bus size of RAID0 is increased to 256-bit (two times of 128-bit which is used in NVMe-IP standard demo). To compatible with DG storage standard, the interface of Raid0x2 module is dgIF typeS. The user interface of Raid0x2 module connects to LAXI2REG and TestGen, same as NVMe-IP standard demo, but data bus size is bigger.

Two sets of two FIFOs are connected between Raid0x2 and DG NVMe-IP. They are used to be data buffer and also used to convert data bus size between 256-bit and 128-bit. For RAID0 operation, 256-bit data stream of UserFIFO is transferred to FIFO#0 or FIFO#1, selected by the logic inside Raid0x2 module. Raid0 logic switches the active SSD every 1-sector data transferring.

State machine of Raid0x2 is designed to receive and decode user interface from the user. From user input, the address and length of each NVMe-IP are calculated and forward to DG NVMe-IP through dgIF types (CMD) interface. State machine generates command request to both NVMe-IP and monitors busy flag until end of transfer.

User can modify 2-ch RAID0 reference design to support more than two SSDs in the system. The numbers of NVMe-IP, Integrated Block for PCIe, and FIFOs must be increased. Also, the bus size between user logic and Raid0 module must be extended to N x 128-bit to increase data bandwidth at user side. The SSD model in every channel should be the same.

3 RAID0IP

Table 1 shows user interface of RAID0 module for both control interface and data interface. The interface is designed to dgIF typeS style. Comparing to NVMe-IP, the status signals and data bus size are double to support two channels.

Signal description of NVMe-IP is described in NVMe-IP datasheet.
http://www.dgway.com/products/IP/NVMe-IP/dg_nvme_ip_data_sheet_en.pdf

3.1 Port Description

Table 1 Signal Description of Raid0 IP (only control interface)

Signal	Dir	Description
User Interface		
RstB	In	Reset signal. Active low. Please use same reset signal as NVMe-IP.
Clk	In	System clock for running NVMe IP. The frequency must be more than or equal to PCIeClk which is output from Integrated Block for PCI Express (125 MHz for PCIe Gen2, 250 MHz for PCIe Gen3).
dgIF typeS		
UserCmd[1:0]	In	User Command. “00”: Identify command, “10”: Write PCIe SSD, ”11”: Read PCIe SSD.
UserAddr[47:0]	In	Start address to write/read SSD in sector unit (512 byte). **From SSD characteristic, it is recommended to set bit[3:0]=”0000” to align 8 Kbyte size which is 2xSSD page size**. Write/Read performance in most SSD are reduced when start addrss is not aligned to 4 Kbyte unit.
UserLen[47:0]	In	Total transfer size in the request in sector unit (512 byte). Valid from 1 to (LBASize-UserAddr).
UserReq	In	Request the new command. Can be asserted only when the IP is Idle (UserBusy=’0’). Asserted with valid value on UserCmd/UserAddr/UserLen signals.
UserBusy	Out	IP Busy status. New request will not be allowed if this signal is asserted to ‘1’.
LBASize[47:0]	Out	Total capacity of PCIe SSD in sector unit (512 byte). Default value is 0. This value is equal to two times of LBASize value output from IP#0.
UserError	Out	Error flag. Assert when UserErrorType is not equal to 0. The flag can be cleared by asserting RstB signal.
UserErrorType[0-1][31:0]	Out	Error status which are mapped from status in each NVMe-IP. [0]-IP#0, [1]-IP#1
UserFifoWrCnt[15:0]	In	Write data counter of User received FIFO. Used to check FIFO space size. If total size is less than 16-bit, please fill ‘1’ to upper bit. UserFifoWrEn can be asserted when UserFifoWrCnt[15:5] is not equal to all 1.
UserFifoWrEn	Out	Write data valid of User received FIFO
UserFifoWrData[255:0]	Out	Write data bus of User received FIFO. Synchronous to UserFifoWrEn.
UserFifoRdCnt[15:0]	In	Read data counter of User transmit FIFO. Used to check data available size in FIFO. If total FIFO size is less than 16-bit, please fill ‘0’ to upper bit. UserFifoRdEn can be asserted when UserFifoRdCnt[15:4] is not equal to 0.
UserFifoEmpty	In	FIFO empty flag of User transmit FIFO. This signal is unused in the design.
UserFifoRdEn	Out	Read valid of User transmit FIFO
UserFifoRdData[255:0]	In	Read data returned from User transmit FIFO. Valid in the next clock after UserFifoRdEn is asserted.

Signal	Dir	Description
Other Interface		
TestPin[0-1][31:0]	Out	Direct mapped from TestPin in each NVMe-IP. [0]-IP#0, [1]-IP#1
TimeOutSet[31:0]	Out	Timeout value to wait completion from SSD. Time unit is equal to 1/(Clk frequency).
LinkSpeed[0-1][1:0]	Out	PCIe speed in each NVMe-IP. Bit[0]-IP#0, [1]-IP#1
AdmCompStatus[0-1][15:0]	Out	Direct mapped from AdmCompStatus in each NVMe- IP. [0]-IP#0, [1]-IP#1
IOCompStatus[0-1]15:0]	Out	Direct mapped from IOCompStatus in each NVMe- IP. [0]-IP#0, [1]-IP#1
NVMeCAPReg[0-1][31:0]	Out	Direct mapped from NVMeCAPReg in each NVMe- IP. [0]-IP#0, [1]-IP#1.
IdenCtrlWrEn[1:0]	Out	Direct mapped from IdenCtrlWrEn in each NVMe- IP. [0]-IP#0, [1]-IP#1.
IdenCtrlWrAddr[0-1][7:0]	Out	Direct mapped from IdenCtrlWrAddr in each NVMe- IP. [0]-IP#0, [1]-IP#1.
IdenCtrlWrData[0-1][127:0]	Out	Direct mapped from IdenCtrlWrData in each NVMe- IP. [0]-IP#0, [1]-IP#1.
IdenNameWrEn[1:0]	Out	Direct mapped from IdenNameWrEn in each NVMe- IP. [0]-IP#0, [1]-IP#1.
IdenNameWrAddr[0-1][7:0]	Out	Direct mapped from IdenNameWrAddr in each NVMe- IP. [0]-IP#0, [1]-IP#1.
IdenNameWrData[0-1][127:0]	Out	Direct mapped from IdenNameWrData in each NVMe- IP. [0]-IP#0, [1]-IP#1.

3.2 Timing Diagram

Timing diagram of RAID user interface and Identify device interface are similar to NVMe-IP, so user can check more details from IP datasheet. For RAID FIFO interface, the details are described as follows.

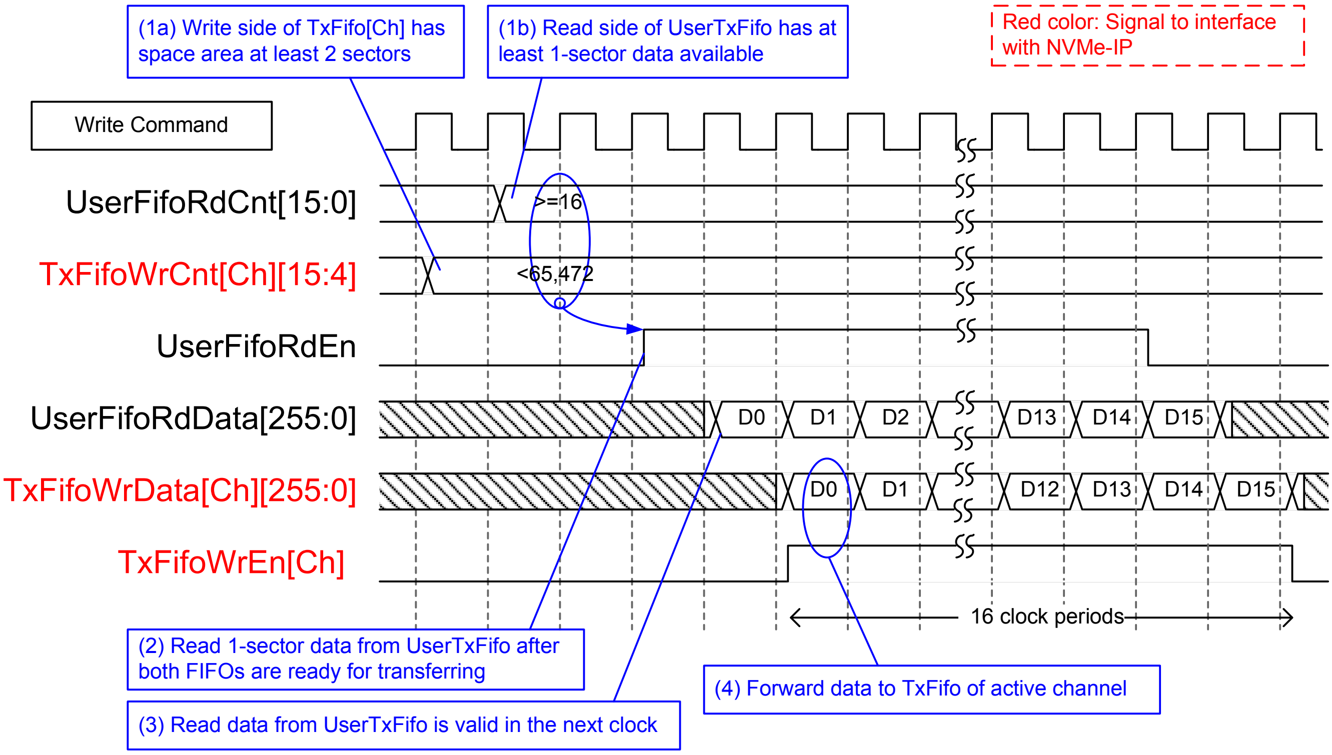

Figure 3-1 RAID FIFO Timing Diagram of Write Command

When user sends write command to RAID system, data stream are forwarded from UserTxFifo to TxFifo[0]-[1]. Only one TxFifo is active to transfer one sector data and the active NVMe channel is switched in the next sector transfer, following RAID0 behavior. Before forwarding data, UserFifoRdCnt and TxFifoWrCnt of active channel are monitored to confirm that at least 1 sector data is stored in UserTxFifo and at least 2-sector free space is available in TxFifo of active channel. UserFifoRdEn is asserted for 16 clock periods to transfer 512-byte data.

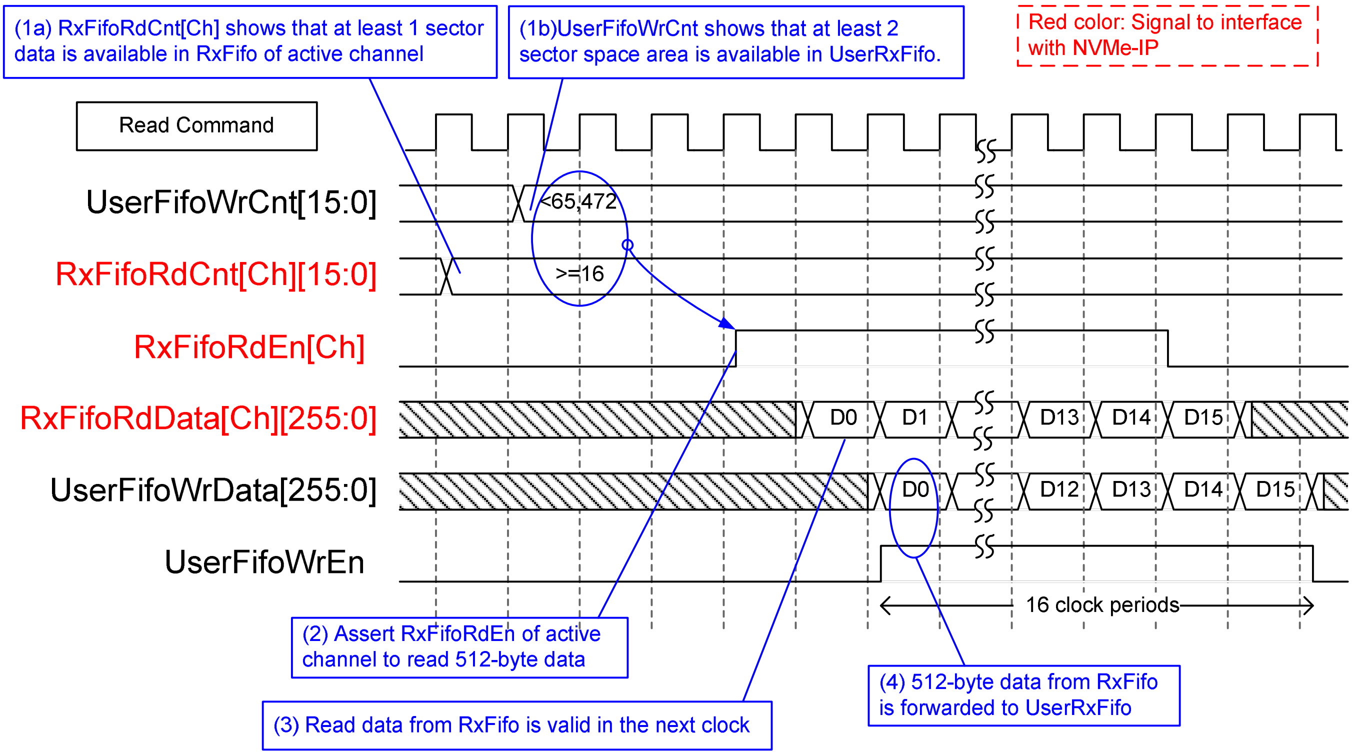

Figure 3-2 RAID FIFO Timing Diagram of Read Command

When user sends read command to RAID system, data stream are forwarded from RxFifo[0]-[1] to UserRxFifo, as shown in Figure 3-2. Similar to write command, only one RxFifo is active to transfer each 512-byte data. The active NVMe channel is switched before transferring the next sector. Before forwarding data, UserFifoWrCnt and RxFifoRdCnt of active channel are monitored to confirm that at least 1 sector data is stored in RxFifo of active channel and at least 2-sector free space is available in UserRxFifo. UserFifoWrEn is asserted for 16 clock periods to transfer 512-byte data.

4 CPU

CPU system in RAID0 design is almost same as NVMe-IP standard demo. But register map for expected pattern and read pattern are extended from 128-bit to 256-bit and the status signals are extended to support two channels, as shown in Table 2

Table 2 Register Map

Address Rd/Wr	Register Name (Label in "nvmeipraid0x2test.c")	Description
BA+0x00 Wr	User Address (Low) Reg (USRADRL_REG)	[31:0]: Input to be start sector address (UserAddr[31:0] of RAID0 following dgIF typeS)
BA+0x04 Wr	User Address (High) Reg (USRADRH_REG)	[15:0]: Input to be start sector address (UserAddr[47:32] of RAID0 following dgIF typeS)
BA+0x08 Wr	User Length (Low) Reg (USRLENL_REG)	[31:0]: Input to be transfer length in sector unit (UserLen[31:0] of RAID0 following dgIF typeS)
BA+0x0C Wr	User Length (High) Reg (USRLENH_REG)	[15:0]: Input to be transfer length in sector unit (UserLen[47:32] of RAID0 following dgIF typeS)
BA+0x10 Wr	User Command Reg (USRCMD_REG)	[1:0]: Input to be user command (UserCmd of RAID0 following dgIF typeS) "00"-Identify, "10"-Write SSD, "11"-Read SSD, When this register is written, the design generates command request to RAID0IP to start new command operation.
BA+0x14 Wr	Test Pattern Reg (PATTSEL_REG)	[2:0]: Test pattern select "000"-Increment, "001"-Decrement, "010"-All 0, "011"-All 1, "100"-LFSR
BA+0x100 Rd	User Status Reg (USRSTS_REG)	[0]: UserBusy of RAID0 following dgIF typeS ('0': Idle, '1': Busy) [1]: UserError of RAID0 following dgIF typeS ('0': Normal, '1': Error) [2]: Data verification fail ('0': Normal, '1': Error) [4:3]: PCIe speed from IP#0 [6:5]: PCIe speed from IP#1 ("00": No linkup, "01": PCIe Gen1, "10": PCIe Gen2, "11": PCIe Gen3)
BA+0x104 Rd	Total device size (Low) Reg (LBASIZEL_REG)	[31:0]: Total capacity of RAID0 in sector unit (LBASize[31:0] of RAID0 following dgIF typeS)
BA+0x108 Rd	Total device size (High) Reg (LBASIZEH_REG)	[15:0]: Total capacity of RAID0 in sector unit (LBASize[47:32] of RAID0 following dgIF typeS)
BA+0x180 Rd	User Error Type CH#0 Reg (USRERRTYPE0_REG)	[31:0]: Mapped to UserErrorType of NVMe-IP#0
BA+0x184 Rd	User Error Type CH#1 Reg (USRERRTYPE1_REG)	[31:0]: Mapped to UserErrorType of NVMe-IP#1
BA+0x190 Rd	Completion Status CH#0 Reg (COMPSTS0_REG)	[15:0]: Mapped to AdmCompStatus[15:0] of NVMe-IP#0 [31:16]: Mapped to IOCompStatus[15:0] of NVMe-IP#0
BA+0x194 Rd	Completion Status CH#1 Reg (COMPSTS1_REG)	[15:0]: Mapped to AdmCompStatus[15:0] of NVMe-IP#1 [31:16]: Mapped to IOCompStatus[15:0] of NVMe-IP#1
BA+0x1A0 Rd	NVMe CAP CH#0 Reg (NVMCAP0_REG)	[31:0]: Mapped to NVMeCAPReg[31:0] of NVMe-IP#0
BA+0x1A4 Rd	NVMe CAP CH#1 Reg (NVMCAP1_REG)	[31:0]: Mapped to NVMeCAPReg[31:0] of NVMe-IP#1
BA+0x1B0 Rd	Test pin of NVMe-IP#0 Reg (NVMTESTPIN0_REG)	[31:0]: Mapped to TestPin of NVMe-IP#0
BA+0x1B4 Rd	Test pin of NVMe-IP#1 Reg (NVMTESTPIN1_REG)	[31:0]: Mapped to TestPin of NVMe-IP#1

Address Rd/Wr	Register Name (Label in the "nvmeipraid0x2test.c")	Description
BA+0x200 Rd	Data Failure Address (Low) Reg (RDFAILNOL_REG)	[31:0]: Latch value of failure address[31:0] in byte unit from read command
BA+0x204 Rd	Data Failure Address (High) Reg (RDFAILNOH_REG)	[24:0]: Latch value of failure address [56:32] in byte unit from read command
BA+0x240 Rd	Expected value Word0 Reg (EXPPATW0_REG)	[31:0]: Latch value of expected data [31:0] from read command
BA+0x244 Rd	Expected value Word1 Reg (EXPPATW1_REG)	[31:0]: Latch value of expected data [63:32] from read command
BA+0x248 Rd	Expected value Word2 Reg (EXPPATW2_REG)	[31:0]: Latch value of expected data [95:64] from read command
BA+0x24C Rd	Expected value Word3 Reg (EXPPATW3_REG)	[31:0]: Latch value of expected data [127:96] from read command
BA+0x250 Rd	Expected value Word4 Reg (EXPPATW4_REG)	[31:0]: Latch value of expected data [159:128] from read command
BA+0x254 Rd	Expected value Word5 Reg (EXPPATW5_REG)	[31:0]: Latch value of expected data [191:160] from read command
BA+0x258 Rd	Expected value Word6 Reg (EXPPATW6_REG)	[31:0]: Latch value of expected data [223:192] from read command
BA+0x25C Rd	Expected value Word7 Reg (EXPPATW7_REG)	[31:0]: Latch value of expected data [255:224] from read command
BA+0x280 Rd	Read value Word0 Reg (RDPATW0_REG)	[31:0]: Latch value of read data [31:0] from read command
BA+0x284 Rd	Read value Word1 Reg (RDPATW1_REG)	[31:0]: Latch value of read data [63:32] from read command
BA+0x288 Rd	Read value Word2 Reg (RDPATW2_REG)	[31:0]: Latch value of read data [95:64] from read command
BA+0x28C Rd	Read value Word3 Reg (RDPATW3_REG)	[31:0]: Latch value of read data [127:96] from read command
BA+0x290 Rd	Read value Word4 Reg (RDPATW4_REG)	[31:0]: Latch value of read data [159:128] from read command
BA+0x294 Rd	Read value Word5 Reg (RDPATW5_REG)	[31:0]: Latch value of read data [191:160] from read command
BA+0x298 Rd	Read value Word6 Reg (RDPATW6_REG)	[31:0]: Latch value of read data [223:192] from read command
BA+0x29C Rd	Read value Word7 Reg (RDPATW7_REG)	[31:0]: Latch value of read data [255:224] from read command
BA+0x2C0 Rd	Current test byte (Low) Reg (CURTESTSIZEL_REG)	[31:0]: Current test data size of TestGen module in byte unit (bit[31:0])
BA+0x2C4 Rd	Current test byte (High) Reg (CURTESTSIZEH_REG)	[24:0]: Current test data size of TestGen module in byte unit (bit[56:32])
BA+0x2000 – 0x2FFF	Identify Device Command Data (IDENCTRL0_REG)	4Kbyte Identify Controller Data Structure from NVMe CH#0
BA+0x3000 – 0x3FFF	Identify Namespace Data (IDENNAME0_REG)	4Kbyte Identify Namespace Data Structure NVMe CH#0
BA+0x4000 – 0x4FFF	Identify Device Command Data (IDENCTRL1_REG)	4Kbyte Identify Controller Data Structure from NVMe CH#1
BA+0x5000 – 0x5FFF	Identify Namespace Data (IDENNAME1_REG)	4Kbyte Identify Namespace Data Structure NVMe CH#1

5 TestGen

Comparing to NVMe-IP single channel demo, data bus of test pattern is extended from 128-bit to 256-bit, as shown in Figure 5-1.

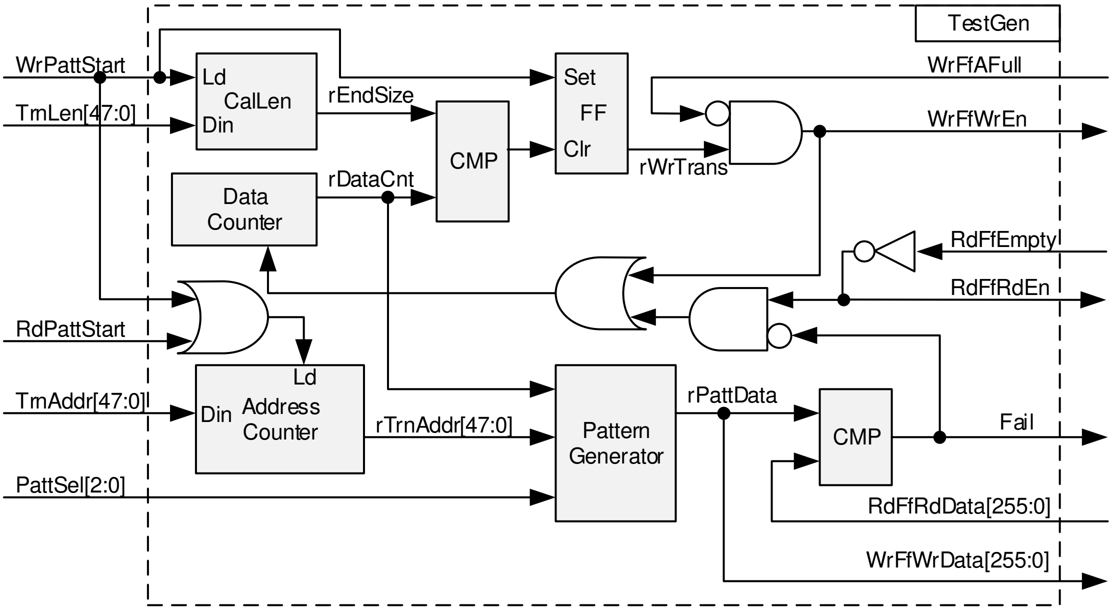

Figure 5-1 Logic design in TestGen

6 Example Test Result

The example test result when running RAID0 demo system by using two 512 GB Samsung 960 Pro SSDs is shown in Figure 6-1.

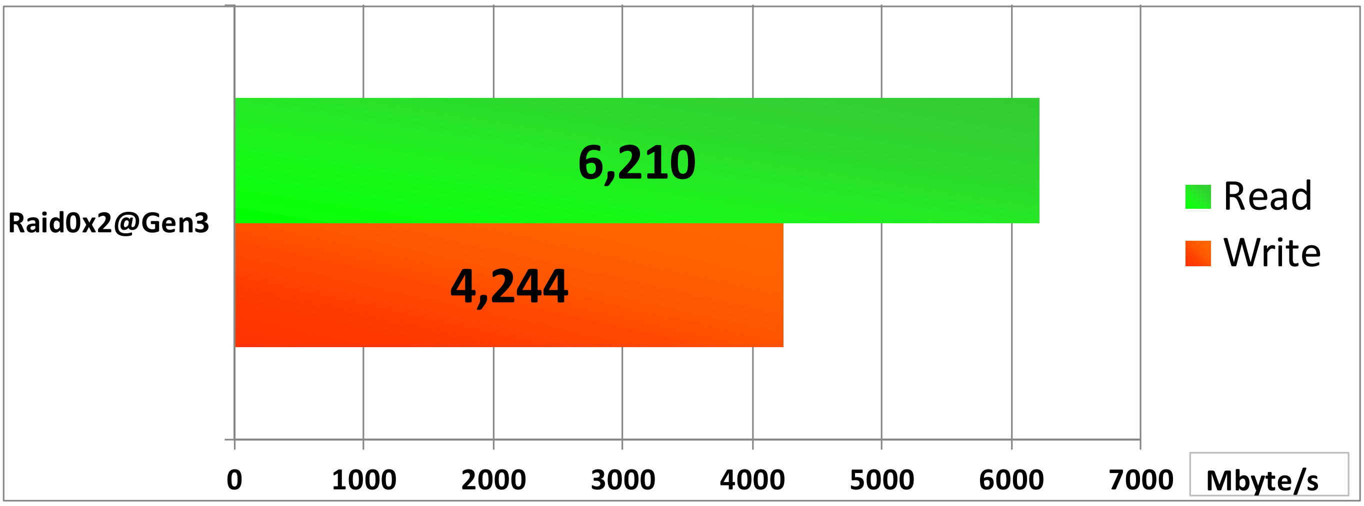

Figure 6-1 Performance of RAID0x2 IP demo by using Samsung 960 Pro SSD

When running 2-ch RAID0 with 2 PCIe Gen3, write performance is about 4200 Mbyte/sec and read performance is about 6200 Mbyte/sec.

7 Revision History

Revision	Date	Description
1.0	6-Oct-17	Initial version release