POWER8 Memory Buffer

User's Manual

Advance

22 April 2014
Version 1.1

Printed in the United States of America April 2014

IBM Systems and Technology Group
2070 Route 52, Bldg. 330
Hopewell Junction, NY 12533-6351

The IBM home page can be found at **ibm.com®**.

Version 1.1
22 April 2014

Contents

List of Tables

Revision Log

Version	Revision Date	Pages	Description
1.1	22 April 2014	–	First public release.
1.0	31 January 2014	–	Initial release.

1. Introduction

This document describes the IBM® POWER8™ Memory Buffer. Throughout this document, the term "memory buffer" refers to the POWER8 Memory Buffer.

To ensure you have the most current version of this document, contact your IBM technical representative or visit the IBM Customer Connect website.

2. Overview

2.1 Architectural and System Overview

The IBM POWER8 Memory Buffer is a memory buffer chip that supports multiple system configurations. The POWER8 Memory Buffer utilizes a high-speed differential interface to communicate with a processor chip using a memory-agnostic protocol. The memory controller and associated memory interface maintenance and calibration functions are initiated and contained within the memory buffer chip. It also contains a 16 MB on-board cache to support prefetching and improve system performance.

2.2 POWER8 Memory Buffer Features

The POWER8 Memory Buffer is a synchronous memory interface chip, manufactured using the 22 nm CMOS 14S technology. It has four memory ports and 16 MB of cache. A high-speed differential memory interface connects the memory buffer to the processor chip. The memory buffer supports DDR3 DRAM technologies.

Table 2-1. Feature Summary (Page 1 of 2)

Feature Type	Description
Basic features	• A single differential memory interface running at 9.6 Gbps • Four DDR3 DRAM command and address ports • Four 8-byte DDR3 DRAM data ports with a 9th byte for error correction code (ECC) • 32 addressable ranks (16 per port pair) • 16 clock enables (CKEs) and on-die termination (ODT) (four per port) • 16 differential memory clock pairs (four per port) • 16 MB 16-way associative cache employing eDRAM
Power planes (at module)	• V_{DD} core voltage ..... 0.969 V • V_{CS} eDRAM supply ..... 1.090 V • V_{IO} differential memory interface (DMI) and pervasive I/O ..... 1.127 V • V_{DDR} DDR PHY DDR3 ..... 1.35 V • AV_{DD} analog supply ..... 1.502 V • V_{SB} service interface ..... 1.208 V
Technology	• 22 nm CMOS SOI (14S) • Reliability grade 3 (72 KPOH, 2555 on/off cycles) • Chip image: 9.262 mm × 11.136 mm (103.14 sq. mm) • C4 pitch: 185.6 µm X, 148.48/185.6 µm Y (variable) • Package: – 4 port, 27.5 mm × 42 mm 3-2-3 FC-PBGA with 1.0 mm interstitial pitch – 3240 controlled collapse chip connections (C4s) with 867 signal I/Os
Operational maximum frequency targets	Memory channel..... 9.6 Gbps DDR3 ..... 1600 Mbps
Operational modes	• 128-byte (BL8) cache line read and write operations • Partial 128-byte cache line write operations (read-modify-write) • Return to high-impedance (Hi-Z) state on command/address ports • Cache disable • 2N addressing mode (2 bits per cycle)
Memory channel	• 9.6 Gbps differential ended unidirectional signals • Downstream link: 14+2+1+1 (data, spares, calibration, clock) • Upstream link: 21+2+1+1 (data, spares, calibration, clock) • Spare lane signaling enables dynamic link training and repair

Table 2-1. Feature Summary (Page 2 of 2)

Feature Type	Description
Power saving	• DRAM power down • Advanced run-time power management policies • Support for fast and slow exit power down of DRAMs • Aggressive dynamic clock gating
Manufacturing, test, and bringup	• Maintenance engine for initial program load (IPL) memory diagnostics • Memory Card Built-In Self Test (MCBIST) • Hard and soft error injection stations for DDR data and command/address ports • Memory channel link self test diagnostics • IEEE 1149.6 boundary scan wire test capability • Source selectable trace array • Two scope trigger pins with configurable events and measurement sources • Digital temperature and voltage sensor (TVSENSE) • On-product clock generation Logic Built-In Self Test (LBIST) • Array Built-In Self Test (ABIST) for eDRAMs and SRAMs with corresponding eFUSE repair structure • Rank Status Register for tracking reset, power down, and self-refresh state of DRAM devices • Fully accessible configuration registers through Flexible Service Interface (FSI) and IEEE 1149.1 Standard Test Access Port and Boundary-Scan Architecture (JTAG) scan communications (SCOM), millicode, and level-sensitive scan design (LSSD) scanning • Performance Monitor interface for frequency measurements of critical events, and measurements of power events • Electronic Chip Identification (ECID) for manufacturing part tracking • Debug bus interface
Reliability, availability, and serviceability (RAS)	• Slave interface of the service processor powered by V_{SB} standby voltage employing FSI protocol • Hierarchical Fault Isolation register (FIR) error-detection structure with configurable severity masks • Memory chip kill and additional error ECC on 64-byte data boundaries, with double error detect capabilities. • Chip kill support for x4 and x8 DRAM devices • Memory scrubbing • Bidirectional fault signal to respond to, or drive out, checkstop requests • Attention indicator that can be polled by the service interface • Memory channel protected by strong cyclic redundancy check (CRC) with automatic retransmission (replay) of packet errors • Configurable fast clock stop on fault • Dynamic lane sparing on memory channel to repair faulty lanes concurrently during mainline operation • All critical data flow and control elements are protected by either ECC, parity, or a soft error (SER) hardened latch

2.2.1 Memory Channels

Features of the memory channels include:

- Downstream write command/data interface, up to 1 TB of memory addressing
- Downstream link contains a differential bus clock pair, 14 command/data signals, one calibration signal, and two spare signals
- Commands and memory store data are interspersed within synchronous packets, four of which constitute a frame
- Upstream tagged read data/status interface
- The upstream bus comprises a differential bus clock with a 21-bit fetch data bus, one calibration lane, and two spare lanes
- CRC protection
- Error status and logging interface

- Automatic hardware retry of soft channel failures
- Automatic reconfiguration to use spare channel lanes to repair faulty lanes.
- Hardware and software channel initialization capability

2.2.2 Memory Ports

The POWER8 Memory Buffer ports include the following features:

- Four 9-byte ports comprising eight bytes of data plus one ECC byte
- Maximum of 16 ranks per port pair
- Two dual inline memory modules (DIMMs) slots per port
- Four pairs of differential memory clocks per port
- DDR3 interface calibration requirements include fine write leveling, initial and periodic ZQ calibration, clock alignment, coarse read/write leveling, and read calibration for optimizing read eye centering, data deskew and gate delay

2.2.3 Supported Memory

2.2.3.1 Attached Memory Rules and Restrictions

The following rules for attaching memory to the memory buffer ports must be followed at all times:

- All memory attached to a port pair must be of the same speed, type, size, and technology.
- Port pairs can contain different sizes and types of the same SDRAM generation. For example, port A/B (0/1) can attach to 2 GB industry standard (IS) DIMMs while port C/D (2/3) is attached to 4 GB IS DIMMs.
- All memory across all the ports must operate at the same frequency.
- Configurations that use a single port pair (two-port applications) must use ports A and B.
- All ranks attached to a port pair must be programmed to have the same latency.

2.2.4 Cache

- Reduces average read data latency and DRAM power by reducing DRAM read accesses and optimizing write accesses
- 16 MB eDRAM data cache
- 16-way set-associative
- Cache disabled mode supported

2.2.5 Memory Data Flow and Controls

- Independent command sequencing per port pair
- Retry of read accesses for memory ECC errors
- Support for 2N addressing mode

2.2.6 Power Saving

- Programmable command execution rate.
- Number of ranks powered-on is controlled through programmable minimum and maximum power domain thresholds.
- LP1 and LP2 low-power modes are supported.
- Support for both fast exit and slow exit power modes as defined by the JEDEC DDR3 definitions.
- Aggressive clock and data gating are employed throughout the chip to reduce switching power

2.2.7 Reliability, Availability and Serviceability (RAS)

- Memory ECC on 64-byte data boundaries
- Background memory scrubbing
- Hierarchical Fault Isolation Register (FIR) and Who's on First (WOF) structure
- Any error checker can be programmed to enact firmware attention-based intervention or system check-stop.
- Bidirectional fault line enables the memory buffer to respond to checkstop requests from the processor chips or to drive checkstop requests to the system.
- Monitoring of register clock driver (RCD) parity errors and DIMM thermal errors.
- All critical data flow and control elements within the chip are protected by either ECC, parity, or an SER-tolerant latch circuit.
- All critical address flow elements within the chip are protected by either parity or redundant signals to the two DRAM buses.
- The memory channel link provides a robust CRC protection mechanism that tolerates failures ranging from intermittent bit flips to complete lane failure
- Automated memory channel link retry recovery facility to retransmit lost frames and resynchronize the host.
- Dynamic lane sparing on the DMI link to repair fault lanes.
- Clock controls enable fast stopping on error to facilitate engineering debug.

2.2.8 Manufacturing, Test, and Bringup Support

Maintenance engine provides capability for various functions including:

- IPL memory diagnostics
- MCBIST
- LBIST support for sequential logic
- ABIST for all on-board RAMs
- Differential boundary scan wire test for the DMI interface
- Hardware error injection stations enable recovery testing of the DMI and DDR interfaces.
- Scope trigger pins to enable multiplexing a plurality of internal sources onto an external scoping device
- Temperature and voltage sensors provide thermal and power monitoring

- Electronic chip identification (ECID) to enable manufacturing part tracking
- Trace arrays that can be accessed by scanning or firmware access
- Fully accessible configuration and status registers via FSI and JTAG SCOM, millicode inband, or LSSD scanning
- Performance monitor interface (through SCOM) to allow frequency measurements of critical events and measurements of power events
- SCOM interface
- Debug bus interface

3. DIMM Configuration And Addressing

The memory buffer chip contains two independent memory controllers (MC0 and MC1). Each of these memory controllers drives a pair of 8-byte industry standard (IS) dual inline memory module (DIMM) channels for a total of four memory channels. Channels are populated and accessed in pairs by the memory controllers in the memory buffer chip. Each memory channel can have one or two DIMMs installed for a total supported memory configuration of eight DIMMs.

3.1 Supported DIMM Configurations

Table 3-1 illustrates the various DIMM types supported by the memory buffer. DIMMs are broken into groups based on the type of electrical structure of the DIMM. The memory buffer supports DDR3 registered DIMMs (RDIMMs) as described in the following sections.

Note: Only the DIMM types supported by the POWER8 Memory Buffer chip are described.

3.1.1 Type 1 DIMMs

Type 1 DIMMs are standard RDIMMs ranging from a single rank to quad rank DIMMs. These contain a register that re-drives the address and control signals to the DRAMs to reduce external loading at the system level. The DIMMs contain 9, 18, or 36 DRAM packages that can be planar (single-die package) or dual-die package (DDP) devices.

Table 3-1. Bus Topology Type 1 DIMMs

Type	DIMM Type	DIMM Config	DIMM Type	DIMM Height	DIMM Capacity			DRAM V_{DD}	Number of Devices	Stacking Type
					2 Gbit	4 Gbit	8 Gbit			
1A	DDR3	1Rx8	RDIMM	30 mm	2 GB	4 GB		1.35	9	Planar
	DDR3	1Rx4	RDIMM	30 mm	4 GB	8 GB		1.35	18	Planar
1B	DDR3	2Rx8	RDIMM	30 mm	4 GB	8 GB		1.35	18	Planar
	DDR3	2Rx4	RDIMM	30 mm	8 GB	16 GB		1.35	36	Planar
1C	DDR3	4Rx4	RDIMM	30 mm		32 GB		1.35	36 stacks	DDP

4. Memory Buffer Cache

This section describes the memory buffer cache (MBC).

4.1 Overview

The POWER8 Memory Buffer contains a 16 MB cache that is used by the chip as a large write buffer and read prefetch buffer. This additional function lowers the average latency of the memory read operations and enables more efficient scheduling of write commands.

The MBC reduces DRAM bus bandwidth and power consumption by allowing the DRAM command scheduler to have greater flexibility when it must send write commands, which are lower priority than read commands. The memory write data is delayed from having to be sent to DRAM by sitting in the cache in a modified state. This makes the cache behave as a very large write reorder buffer, which is possible due to the proximity of the cache to the DRAM command scheduler. In addition to enabling bursty read-write behavior to be leveled out, a very large write buffer increases the odds of the DRAM command scheduler finding page mode write operations, which lowers power consumption and improves bus utilization.

A third benefit of the MBC is that it helps reduce the time needed for read-modify-write memory operations. For example, many partial write operations to the same 128-byte line enable improved performance as the line is gathered into the MBC for the first partial write and future partial writes experiences a cache hit.

5. DDR Interface

5.1 Overview

The POWER8 Memory Buffer supports the DDR3 JEDEC specifications. The DDR unit is responsible for:

- Transporting and mapping of command, control, address, and data signals onto the memory buffer chip I/O.
- Providing all necessary configuration registers, state machines, control logic, and status monitoring to execute all required DDR calibration functions (for example, read calibration, fine and coarse write leveling, and I/O impedance [ZQ] calibration)
- Providing elastic interface style first in, first out (FIFO) physical layers (PHYs) for the purpose of sampling, deskewing, and bit aligning incoming data, as well as buffering and launching outgoing data.

Each DDR unit is self contained and comprises one physical entity that is replicated four times to create the four memory ports. The A and B ports always run as a synchronized pair, while the C and D ports form the other port pair.

5.2 Mainline Operation

The DDR unit must support all mainline functions initiated by the memory controller. This includes:

- Clock enable (CKE) controls for powering off and powering on ranks
- Bank activate commands
- Burst length 8 read and write operations
- Periodic refreshes

It is the memory controller's responsibility to ensure proper spacing and timing of all command, control, and data signals, as well as adherence to the JEDEC specifications. The primary responsibility of the DDR unit is to propagate all command, address, data, and control signals from the memory controller to the memory interface.

5.2.1 Rank-to-Rank Switching Delays

The DDR unit supports programmable rank switching delays depending on the speed and operational sequence. It also supports the opportunistic low-latency scheme, wherein the delay incurred between ranks can be minimized in situations where two logical ranks share the same delay settings (for example, two ranks adjoined with a 3DS stack).

5.2.2 2N Addressing Mode

For systems employing DIMMs that require using 2N addressing to close timing, the DDR unit contains a configuration bit that firmware can program. When running in this mode, all commands sent to the DRAMs are extended by one memory clock. The assertion of chip select, on-die termination (ODT), CKE, and data are delayed by one cycle to line up with the second cycle of command.

5.2.3 Address Mirroring

For systems employing DIMMs that incorporate address mirroring, the DDR unit contains configuration bits that result in the remapping of select address bits upon issuance of MRS commands. *Table 5-1* shows the that bits can be mirrored.

Table 5-1. Bit Mirror Remapping

A_m	A_n
A3	A4
A5	A6
A7	A8
A11	A13
BA0	BA1
BG0	BG1

The DDR unit contains multiplexing of the address bits to enable rerouting of the signals based on static configuration.

5.3 Supported Memory Topologies

Each DDR unit comprises one memory port. The memory buffer has a total of four memory ports, each of which accommodates up to 16 system-addressable ranks. Each port has at most two sockets, with each socket carrying a maximum of eight ranks.

5.3.1 Chip Select, ODT, and CKE Mapping

Regardless of the number of physically attached ranks, data is always accessed through a pair of ports. Therefore, the N physical ranks are treated as N/2 logical ranks by the memory controller. The port pairs are denoted as m[ab] and m[cd].

5.4 Memory Interface Calibration

The DDR unit is responsible for performing all initial and periodic calibration operations. Each step in a sequence can be disabled or skipped by use of a configuration bit. To simplify the design, the memory buffer powers on all ranks attached to the port before invoking any initial or periodic calibration operations. After all ranks are calibrated, dynamic CKE power management resumes.

5.4.1 Read Calibration

The Read Calibration algorithm optimizes gating of the arriving strobe, aligns the data bits, and centers the strobe within the data eye. The implementation algorithms combine data deskew (alignment) and strobe centering into a single operation. Internally, each data bit has its own strobe thereby enabling the phase rotator to align each DQ individually.

5.4.2 Write Calibration

Write calibration is comprised of fine write leveling for aligning the strobe with respect to the memory clock, and coarse write leveling to adjust the strobe into the correct logical write cycle. The memory buffer employs enhanced preskew and centering algorithms for precise alignment of the DQ bits with the strobe. The DDR unit also contains separate programmable per-nibble windage registers that apply a desired offset to the calculated delays.

5.4.3 DDR I/O Impedance (ZQ) Calibration

The POWER8 Memory Buffer contains a single ZQ calibration engine that uses one external precision resistor to calibrate the DDR I/Os on all four ports. Each DDR unit contains logic that interfaces to the central core and initiates a calibration operation.

5.4.4 DRAM I/O Impedance (ZQ) Calibration

JEDEC requires a DRAM device I/O to be initially and periodically calibrated. The DDR unit is responsible for issuing ZQ calibration commands to each rank on the port.

5.4.5 Initial (POR) Calibration

Initial power-on reset (POR) calibration consists of running a read, write, and ZQ (long) calibration on each rank, as well as calibrating the memory buffer DDR I/O circuits. In addition, the DDR unit contains configuration bits that enable or disable the various atomic operations that comprise the overall initialization sequence.

6. Power and Thermal Management

This section describes the power and thermal management of the POWER8 Memory Buffer chip.

6.1 Power Controls

6.1.1 SDRAM Power Down

Both fast- and slow-exit power down are supported by the POWER8 Memory Buffer, but this selection must be made at power up. The memory buffer does not support switching between fast- and slow-exit power down during run time. In general, the power down behavior is the same for both power modes. There are greater powered-down power savings in the slow-exit power-down mode, but there is a greater performance penalty to exit power down, access the powered-down domain, and power up the terminating ranks/domains.

6.1.2 Power Limiting Control

There are two approaches to reducing the power consumption of the POWER8 Memory Buffer chip by throttling the number of concurrent commands or events and limiting resource utilization:

- Maximum and minimum power domains
- Programmable N commands over M window

6.1.2.1 Maximum/Minimum Number of Power Domains Powered On

In the POWER8 Memory Buffer, a power domain consists of the group of CKEs that must be powered up to service a read, write, or refresh to a rank. A rank's power domain includes the rank's wired CKE, as well as the CKEs required for any terminating ranks. Power control dynamically controls the number of powered-on power domains to fit within the defined minimum and maximum power domain constraints and still provide reasonable bus utilization.

6.1.2.2 Programmable N Commands Over M Window

N and M refer to *N* read/write (address) operations permitted within a window of *M* DRAM clocks. This gives the sequencer a specific budget for how many read/write operations it can send out before it must stall execution. Scheduling resumes only after the M DRAM clocks have passed.

6.1.3 Emergency Throttle Mode

The POWER8 Memory Buffer receives an emergency throttle command from the processor when the processor detects a power supply or thermal issue that requires memory power to be immediately reduced. When this command is received, the memory buffer switches the state of the command sequencer to a preset value to reduce memory subsystem power. When the emergency throttle mode is reset, these parameters return to the values they had before entering the emergency throttle mode.

6.1.4 Return to Hi-Z Mode

The POWER8 Memory Buffer supports return to Hi-Z mode during mainline operation. Whenever the memory controller detects periods of inactivity, it signals the DDR PHYs to place the majority of the command/address bus into a high-impedance state to conserve I/O power. Ideally, the bus is in a Hi-Z mode during every idle cycle, but it is permissible for the command/address bus to be driven for one extra t_{CK}.

7. Glossary

ABIST	Array built-in self test
ACK	Acknowledge
CAC	Command/address/control
CAS	Column-address select
CE	Correctable error
CKE	Clock enable
CRC	Cyclic redundancy check
DDP	Dual die package
DED	Double error detected
DIMM	Dual in-line memory module
DMI	Differential memory interface
DPD	DRAM power domain
ECC	Error correction code
ECID	Electronic chip identification
FC-PBGA	Flip chip plastic ball grid array
FIFO	First in, first out
FIR	Fault Isolation Register
FPGA	Fine ball grid array
FSI	Flexible service interface
IPL	Initial program load
ISDIMM	Industry standard DIMM
KPOH	1,000 power-on hours
LBIST	Logic built-in self test
LRDIMM	Load reduction dual in-line memory module
LSSD	Level-sensitive scan design
MBA	Memory buffer asynchronous unit
MBC	Memory buffer cache
MBI	Memory buffer interface

MBS	Memory buffer synchronous unit
MCBIST	Memory card built-in self test
MPR	Multipurpose Register
ODT	On-die termination
PMU	Performance monitor unit
POR	Power-on reset
RAS	Row-address select
RCD	Register clock driver
RDIMM	Registered dual in-line memory module
SCOM	Scan communications
SDRAM	Synchronous dynamic random-access memory
SEC	Single error corrected
SEPD	Slow-exit power down
SER	Soft error
SRAM	Static random access memory
SUE	Special uncorrectable error
TVSENSE	Temperature and voltage sensor
UE	Uncorrectable error
WOF	Who's on first
WRQ	Write reorder queue
ZQ	I/O impedance