Overview of 3D-RAM And Its Functional Blocks

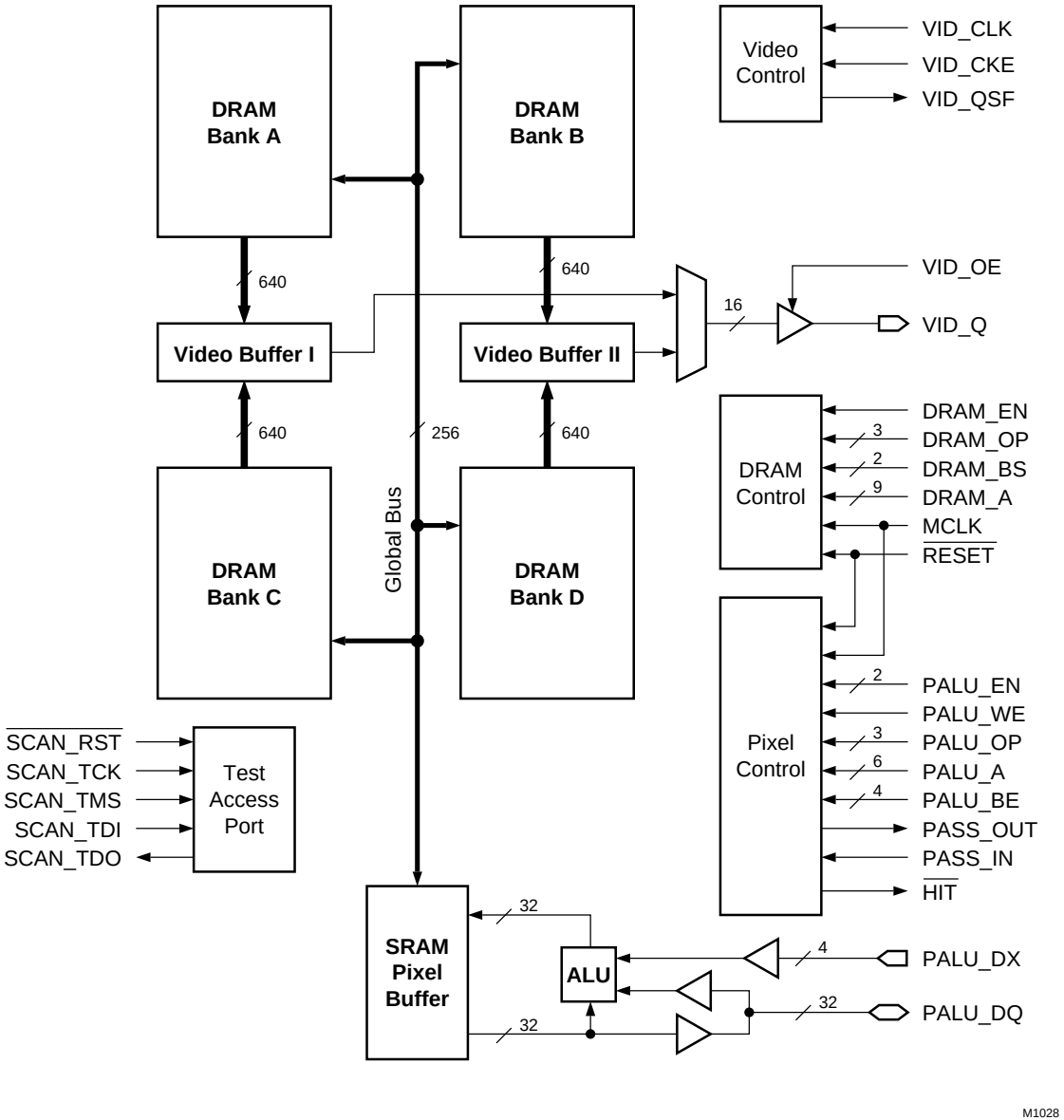
1.1 Simplified 3D-RAM Block Diagram

The 3D-RAM block diagram is shown in Figure 1-1. There are five major functional blocks in 3D-RAM: DRAM banks, Video Buffers, Pixel Buffer, Global Bus, and Pixel ALU.

The DRAM array is partitioned into four independent banks of 2.5 Mbits each. Together, these four banks can support a screen resolution of 1280 x 1024 x 8. The independent banks can be interleaved to facilitate an almost uninterrupted frame buffer update and, at the same time, can transfer pixel data to the dual Video Buffer for screen refresh. Each DRAM bank has 256 pages with 10,240 bits per page for a total storage of 2,621,440 bits. An additional 257th page can be accessed for special functions or used to hold off-screen data. A row decoder takes a 9-bit page address signals to generate 257 word lines, one for each page. The word lines select which page is connected to the sense amplifiers. The sense amplifiers read and write the page selected by the row decoder. Because the sense amplifiers retain data after the read/write operations, they function like a direct-mapped level-two pixel cache.

Data from the DRAM banks is transferred over the 256-bit Global Bus to the triple-ported SRAM Pixel Buffer. The Pixel Buffer consists of eight blocks, each of which is 256 bits and is updated in a single transfer on the Global Bus. Hence, the memory size of the Pixel Buffer is 2 Kbits. The ALU uses two of the Pixel Buffer ports to read and write data in the same clock cycle. Each Video Buffer is 80 x 8 bits and is loaded in a single DRAM operation. One Video Buffer can be loaded while the other is sending out video data.

Figure 1-1 The simplified 3D-RAM block diagram



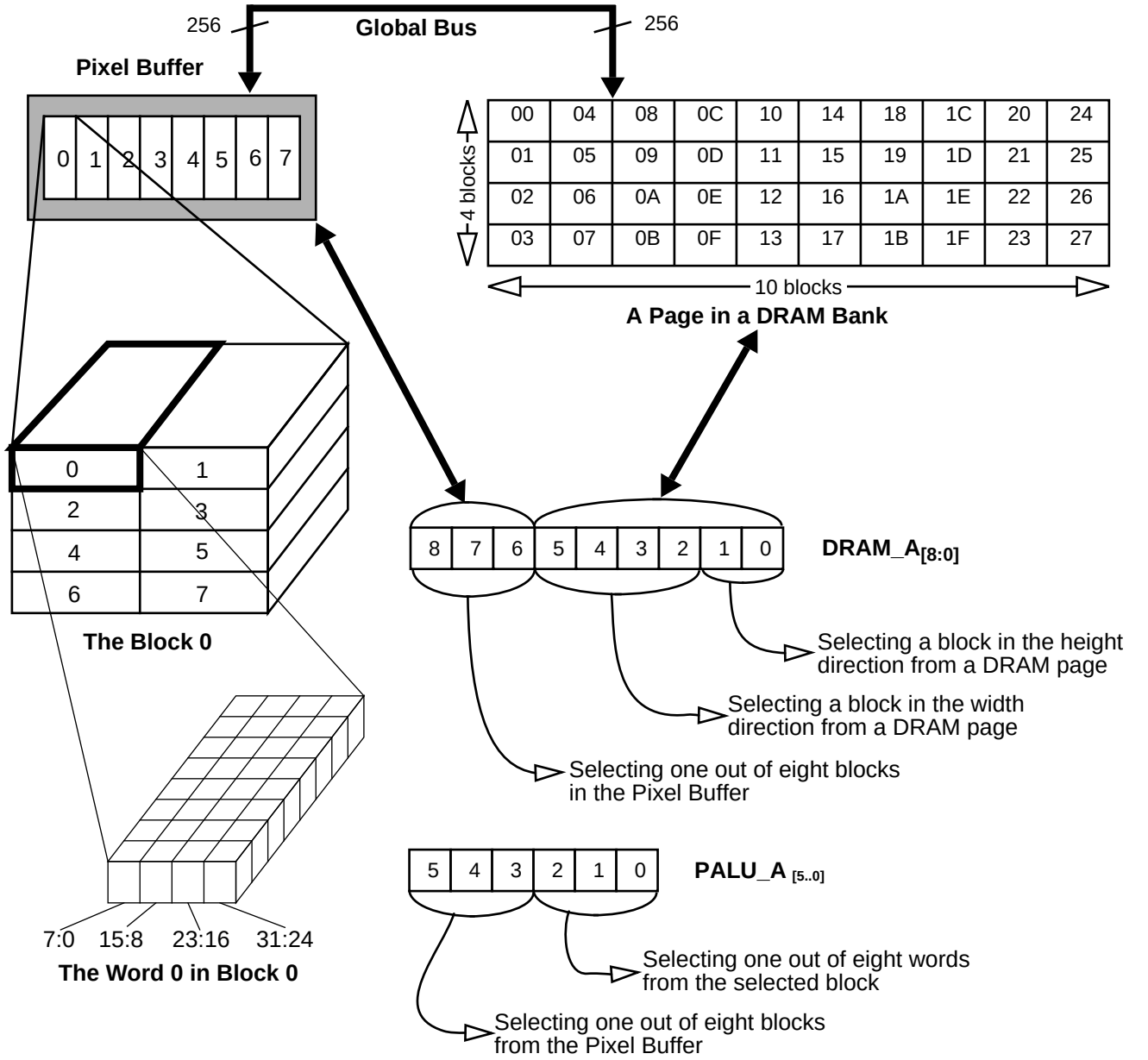
1.2 Block, Page, and Page Group

A word has 32 bits and is the unit of data operations within the Pixel ALU and between the Pixel ALU and Pixel Buffer. In an ALU write operation operates, each of the four bytes in a word may be individually masked. A block has 256 bits and is the unit of memory operations between a DRAM bank and the Pixel Buffer over the Global Bus, for example the Umasked Write Block (UWB), Masked Write Block (MWB), and Read Block (RDB). A page in a DRAM bank is organized as 10 x 4 blocks. Since a block has 256 bits, a page has 10,240 bits. There are four DRAM banks in a 3D-RAM chip, and the pages of the same page address from all four DRAM banks consist a page group. Therefore, a page group has 20 x 8 blocks.

Note in Figure 1-2, the block and page are purposely drawn in rectangular shape. The user may relate these to a tiled frame buffer memory organization. The advantage of such a frame buffer memory organization is the minimization of page miss penalty. 3D objects frequently occupy portions of multiple scan lines. Since in this case a page contains 80 x 16 pixels instead of 10,240 x 1 pixels, page miss is reduced. When an object extends beyond a page boundary, bank interleaving allows hidden precharge and uninterrupted memory access.

On the other hand, to support screen refresh, the Video Buffer must output pixel data one scan line at a time. The internal organization of a page also allows data to be transferred from a page to the Video Buffer one out of the sixteen scan lines of 80 bytes long each at a time.

Figure 1-2 *The relations and addressing scheme of blocks and words in the Pixel Buffer and in the DRAM page.*



2.1 DRAM Operations Basics

Table 2-1 lists all the DRAM operations. One operation can be launched every cycle. However, the sequence of these DRAM operations is bounded by the resource interlocks. The Access Page can only be issued after Precharge Bank, and the only operation after Precharge Bank is Access Page. The full specification contains the specific timing interlocks for DRAM operations in the same bank and between different banks.

Table 2-1 DRAM operation encoding

Operation	DRAM_OP	DRAM_BS	DRAM_A
Unmasked Write Block (UWB)	000	Bank	Pixel Buffer Block(3 pins), DRAM Block(6 pins)
Masked Write Block (MWB)	001	Bank	Pixel Buffer Block(3 pins), DRAM Block(6 pins)
Precharge Bank (PRE)	010	Bank	-
Video Transfer (VDX)	011	Bank	Control (2pins), Line (4pins)
Duplicate Page (DUP)	100	Bank	Page (9 pins)
Read Block (RDB)	101	Bank	Pixel Buffer Block(3 pins), DRAM Block(6 pins)
Access Page (ACP)	110	Bank	Page (9 pins)
No Operation (NOP)	111	-	-

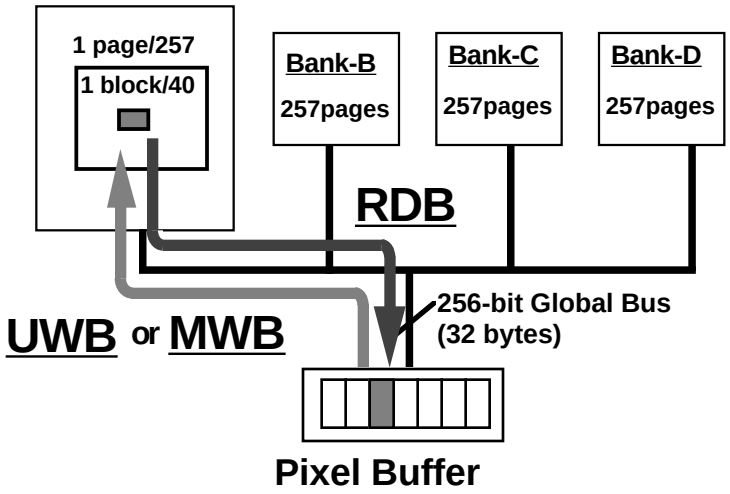
2.1.1 Unmasked Write Block (UWB)

The UWB operation copies 32 bytes from the specified Pixel Buffer block over the Global Bus to the specified block in the sense amplifiers of the currently active DRAM page of a selected DRAM bank. The 32-bit Plane Mask register has no effect on Unmasked Write Block operation. The 32-bit Dirty Tag still controls which bytes of the block are updated.

2.1.2 Masked Write Block (MWB)

The MWB operation functions like the UWB operation except that both the 32-bit Dirty Tag and the 32-bit Plane Mask register control which bytes of the block are updated.

Figure 2-3 Unmasked Write Block, Masked Write Block, and Read Block on the Global Bus



2.1.3 Precharge Bank (PRE)

The PRE operation first deactivates the word line corresponding to the most recently accessed DRAM page of a selected DRAM bank and then equalizes the bit lines of the sense amplifiers for a subsequent Access Page operation. After a Precharge Bank operation has been done to a certain DRAM bank, the operations that can be done to that DRAM bank are Access Page, Precharge Bank, and NOP. Other operations after a Precharge Bank operation are illegal, and the resulting data is undefined.

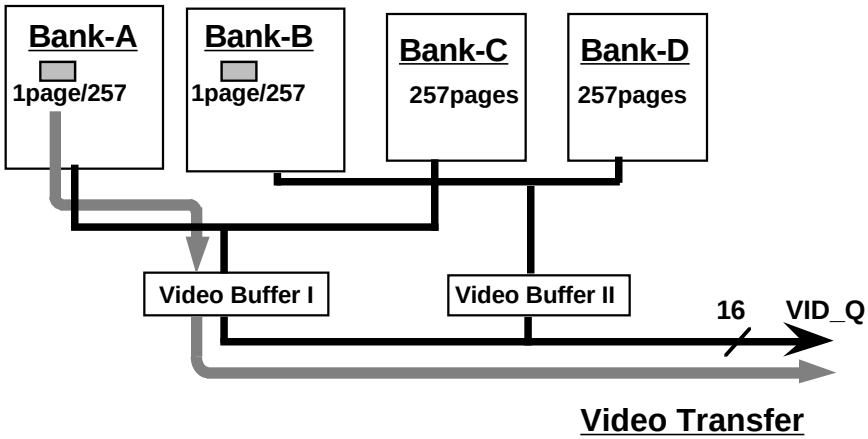
2.1.4 Video Transfer (VDX)

There are two parts to the VDX operation: video buffer load and video output. Video Buffer load relates to the transfer from the sense amplifiers of a selected DRAM bank to a corresponding Video Buffer. Video output relates to the transfer from a Video Buffer to the VID_Q pins.

2.1.4.1 *Video Buffer Load*

There are two video buffers available for interleave transfer. Video Buffer I is for Bank A and Bank C. Video Buffer II is for Bank B and Bank D. Figure 2-4 illustrates a Video Transfer example from a page in Bank A to Video Buffer I.

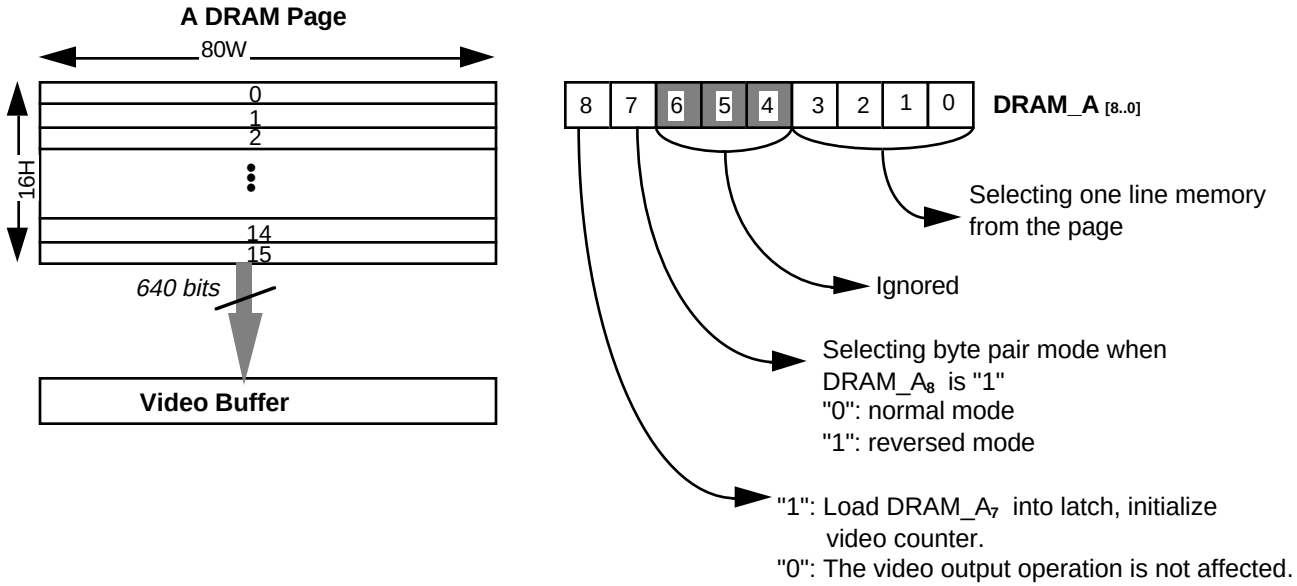
Figure 2-4 The Video Transfer from a Bank A page to Video Buffer I



In this paragraph, the addressing scheme for the Video Transfer operation is described in details. A DRAM page has a fixed organization of 10 blocks wide by 4 blocks high. For VDX operation, a 32-byte block is always considered as having 4 rows high (therefore either as 8W x 4H x 8 or as 2W x 4H x 32). That is, for VDX operation, a DRAM page is always viewed as containing 16 rows of 80 bytes each. In the case of 8 bits per pixel, the Video Transfer operation transfers a 80W x 1H x 8 line of pixel data from the sense amplifiers of a DRAM page to the corresponding Video Buffer. Since there are 16 VID_Q pins, one may think of the Video Buffer as 40 double-bytes.

There are two byte order formats for the video output pins VID_Q: normal mode and reversed mode. These 16-bit VID_Q bus output schemes are illustrated in Figure 2-6 for a 80W x 1H x 8 Video Buffer.

Figure 2-5 The addressing scheme for video transfer

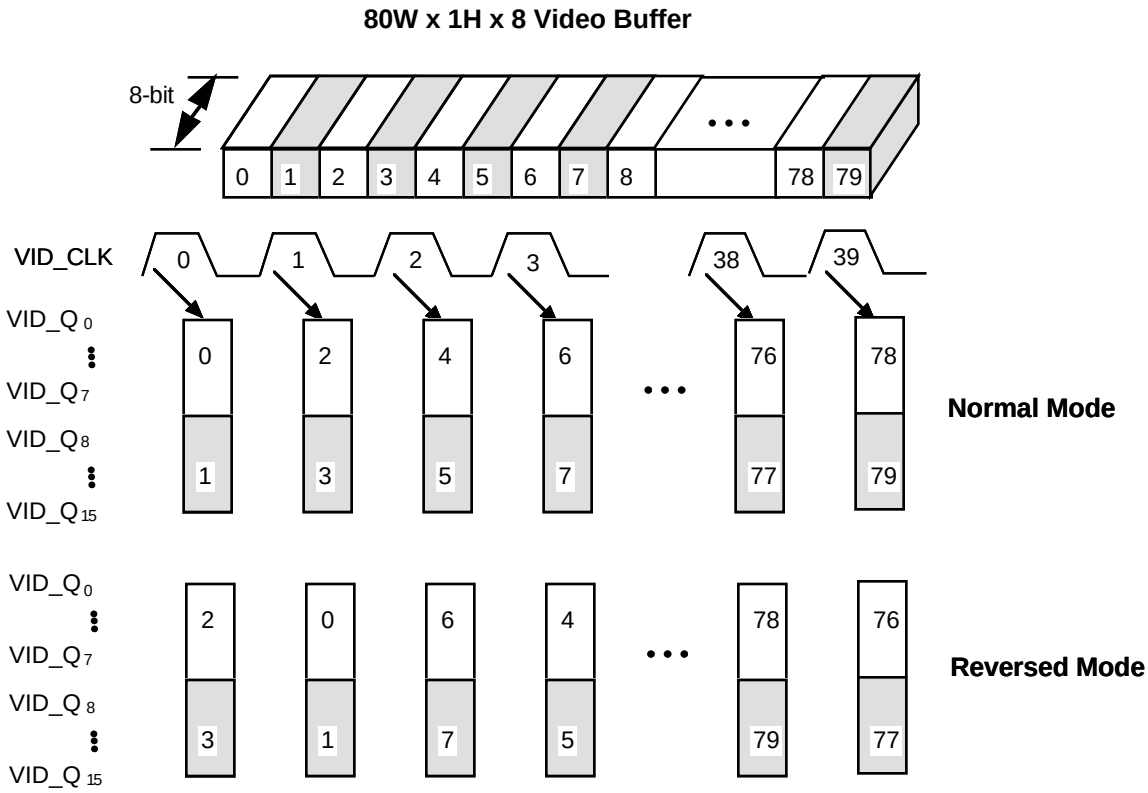


2.1.4.2 Video Output Operation

When the DRAM_A_8 pin is “1”, besides loading the byte pair mode latch, the current Video Buffer output operation will be aborted. The VID_Q bus starts to be driven from the Video Buffer indicated by the DRAM_BS_0 pin. Also, the modulo-40 Video Counter is initialized. If DRAM_A_8 is “0”, the Video Counter is not affected. The video output from the current Video Buffer continues until this buffer is exhausted. Then, the Video Buffer is automatically switched and the Video Counter is initialized. Note that VID_QSF settles from an unknown state to a known state after the initial Video Transfer with DRAM_A_8 = 1. Except this initial Video Transfer, the clean edge transition on VID_QSF is guaranteed for every occurrence of Video Buffer interleave.

To avoid data corruption in the Video Buffer, the user should not start a Video Transfer operation to the particular Video Buffer which is outputting data to the VID_Q bus.

Figure 2-6 The 16-bit VID_Q bus output scheme from a 80W x 1H x 8 Video Buffer



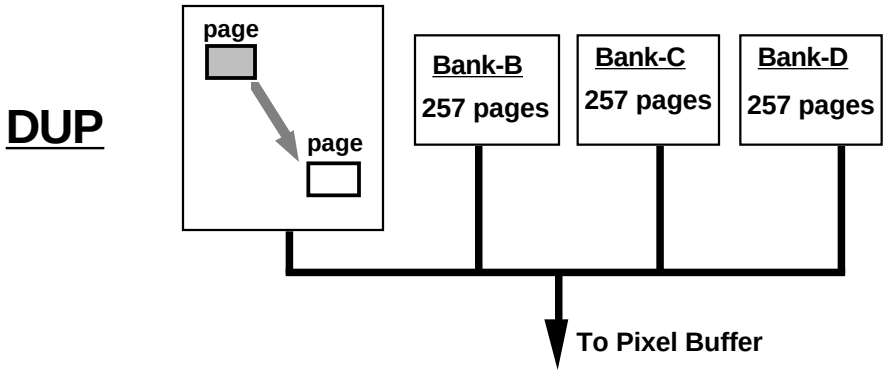
2.1.5 Duplicate Page (DUP)

All 10,240 bits of the data in the sense amplifiers of a selected DRAM bank can be transferred to any specified page in the same bank within one Duplicate Page operation. The data in the sense amplifiers is not affected by this operation. If the $DRAM\_A_8$ pin is 0, then the $DRAM\_A_{[7:0]}$ pins select one of the 256 normal pages. If $DRAM\_A_8$ is 1, then $DRAM\_A_{[7:0]}$ are ignored and the extra page is written. The Plane Mask register does not apply to this operation.

It may be helpful to point out that it is not necessary to use the DUP operation to write back the data in the sense amplifiers, because they function as a level-two write-through pixel cache. DUP is a special performance function that offers ultra-fast data movement in frame buffer. Consider the task of clearing the entire frame buffer of 1280 x 1024 x 32. Using only the MWB operations for this task, the 256-bit Global Bus and four bank interleaving plus parallel operations to the four 3D-RAM chips offer very good bandwidth. The data rate is 5.8 GB/s for

the -10 grade of 3D-RAM, and the entire screen is cleared in 860 µs, without considering the interruptions of video refresh. However, with the DUP performance function, the data rate is multiplied 10 times to 58.6 GB/s, and the entire screen is cleared in only 85 µs, with the same -10 grade of 3D-RAM.

Figure 2-7 Duplicate Page in DRAM Bank A



2.1.6 Read Block (RDB)

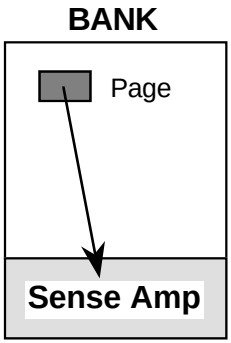
The RDB operation Copies 32 bytes from the sense amplifiers of a selected DRAM bank over the Global Bus to the specified block in the Pixel Buffer. The corresponding 32-bit Dirty Tag is cleared. The $DRAM\_A_{[5:0]}$ pins select one of the 40 blocks in a DRAM page. The $DRAM\_A_{[8:6]}$ pins select one of the eight Pixel Buffer blocks. The Read Block operation is also illustrated in Figure 2-3.

2.1.7 Access Page (ACP)

The ACP operation activates the word line corresponding to the specified DRAM page of a selected DRAM bank and transfers the data in the DRAM array to the sense amplifiers. If the $DRAM\_A_8$ pin is 0, then the $DRAM\_A_{[7:0]}$ pins select one of the 256 normal pages. If the $DRAM\_A_8$ pin is 1, then the $DRAM\_A_{[7:0]}$ pins are ignored and the extra page is transferred.

Before an Access Page can be done in a certain DRAM bank, a Precharge Bank must have been performed to that DRAM bank. After an Access Page operation, a number of the DRAM read and write operations, such as UWB, MWB, RDB, DUP, and VDX, may be performed.

Figure 2-8 Access Page means transferring a specified page to the sense amplifiers.



2.1.8 No Operation (NOP)

The NOP operation may be freely inserted between the ACP operation and the PRE operation on the same bank. NOPs are issued when the DRAM arrays are idle, no read or write is required by the Pixel Buffer, and no Video Buffer load is necessary. More importantly, NOPs are required to satisfy the timing interlocks of the various DRAM operations; for this application, each NOP operation simply takes one clock period of time.

3.1 Pixel Buffer

The Pixel Buffer is a 2048-bit SRAM organized as eight 256-bit blocks, as seen in Figure 1-2, and functions as a level-one write-back pixel cache. It has a 256-bit read/write port, a 32-bit read port, and a 32-bit write port. Referring to Figure 4-9, the 256-bit read/write port is connected to the Global Bus via a Write Buffer, and the two 32-bit ports are connected to the Pixel ALU and the pixel data pins. All three ports can be used simultaneously as long as the same memory cell is not accessed. If the two 32-bit ports access the same cell, the write operation will be successful but the read data will be undefined.

A 1-bit Dirty Tag bit is assigned to each byte data in the Pixel Buffer. Therefore, each block in the Pixel Buffer is associated with a 32-bit Dirty Tag in the dual-port Dirty Tag RAM. When a block is transferred from the sense amplifiers to the Pixel Buffer through the 256-bit port, the corresponding 32-bit Dirty Tag is cleared. When a block is transferred from the Pixel Buffer to a DRAM bank, the Dirty Tag determines which bytes are actually written. This feature can save as much as 50% of the power consumed by a 256-bit block write operation without the Dirty Tag.

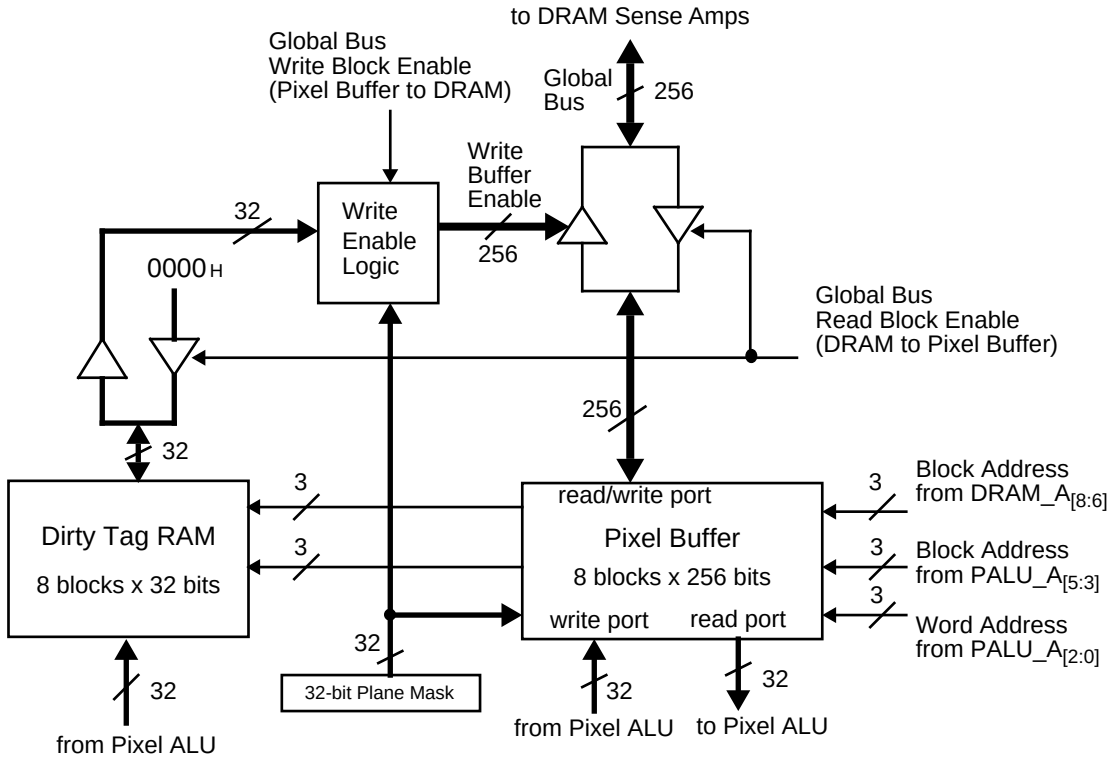
The cache set associativity is determined external to the 3D-RAM, thereby permitting optimal cache design tailored to the particular graphics system.

4.1 Global Bus

The Global Bus connects the Pixel Buffer to the sense amplifiers of all four DRAM banks. The Global Bus consists of 256 data lines. Referring to Figure 4-9, during a transfer from the Pixel Buffer to DRAM, the 256 bits are conditionally written depending on the 32-bit Dirty Tag and the 32-bit Plane Mask. When a data block is transferred from the Pixel Buffer to the sense amplifiers, the Dirty Tag and Plane Mask control which bits of the sense amplifiers are changed via the Write Buffer.

It is helpful to clarify that all read/write operations are referred to from the perspective of the rendering controller. In other words, a read operation across the Global Bus always means a read by the Pixel ALU; that is, data is transferred from a DRAM bank into the Pixel Buffer. Similarly, a write operation across the Global Bus means data is updated from the Pixel Buffer to a DRAM bank. This is also specifically noted in Figure 4-9 by the signals Global Bus Write Block Enable and Global Bus Read Block Enable.

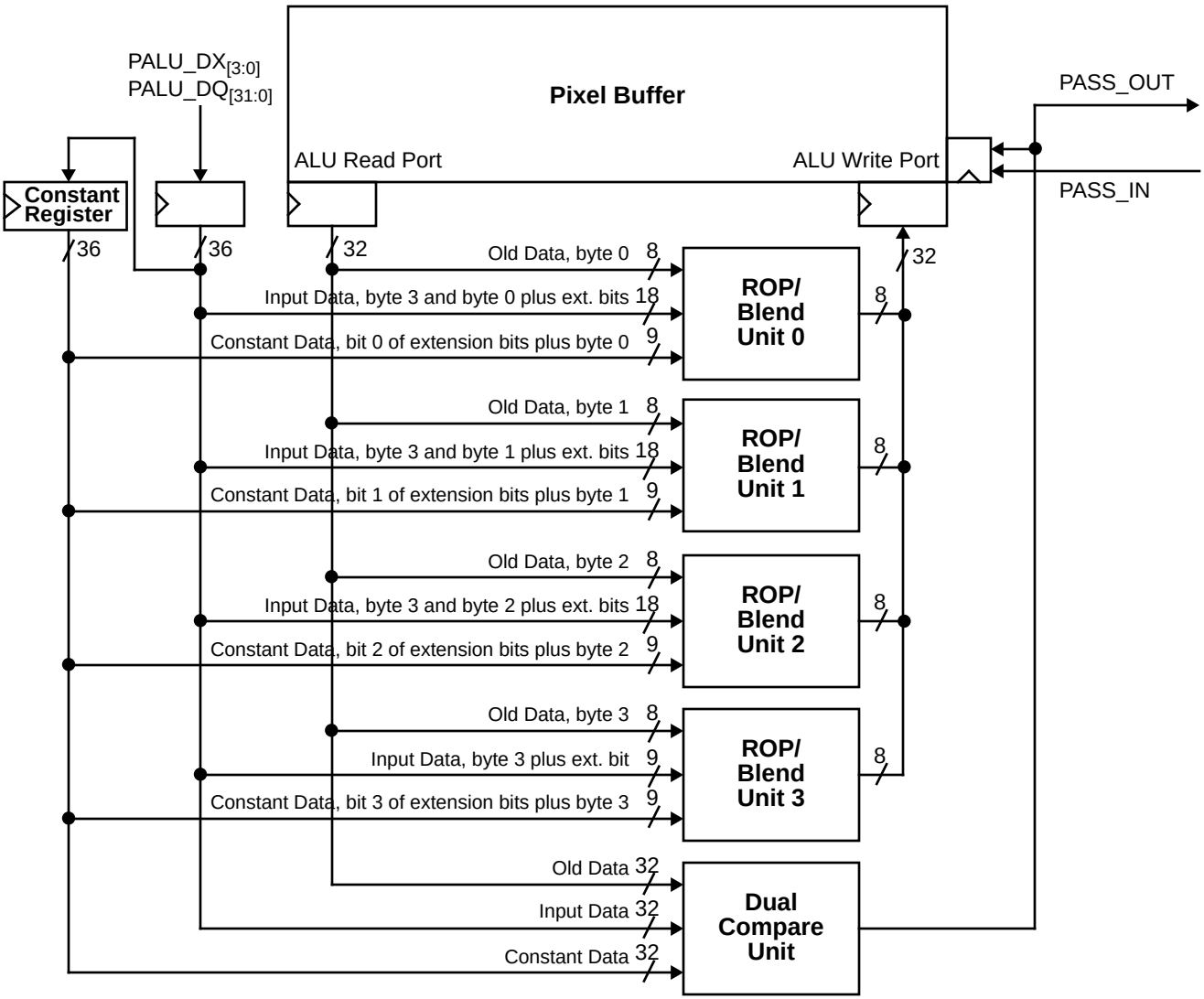
Figure 4-9 The tri-port Pixel Buffer, Global Bus and dual-port Dirty Tag RAM



5.1 Pixel ALU Basics

The Pixel ALU consists of four 8-bit ROP/Blend units, which may be independently programed to perform either a raster operation or a blending function, one 32-bit Match Compare unit, and one 32-bit Magnitude Compare unit. The two Compare units are also commonly referred to as the Dual Compare units. The motivation for including the Pixel ALU on chip is to convert the interface from a read-modify-write interface to a write-mostly interface. This logic integration with memory arrays greatly improves rendering throughput by avoiding time consuming reads and direction changes on the data bus.

Figure 5-10 Pixel ALU (Pipeline stages are not shown.)



The ROP/Blend units and the Dual Compare units are highly pipelined. Section 5.1.3 contains a brief discussion of the ALU pipeline. The output of a ROP/Blend unit is conditionally written to the Pixel Buffer, depending on the comparison results from the on-chip Dual Compare units and from the Dual Compare units of the preceding 3D-RAM chips. For example, for a 1280 x 1024 x 32 double-buffered graphics system with 32-bit Z buffer, there are effectively 96 bits per pixel. In this case, eight 3D-RAMs are used as color chips and four as Z chips. The Pixel ALUs of the Z chips perform magnitude comparisons and feed the comparison results via their PASS_OUT pins to the corresponding color chips. It is important to note that due to the pipelining, the color chips do not wait for the magnitude comparison results from the Z chips; rather, the results of the ROP/blending operations and comparison operations on the color chips, and the results of the magnitude comparison on the Z chips all are presented to the Pixel Buffer of the color chips in the same clock cycle. In this sense, the rendering controller can accomplish a pixel blending operation with Z compare and window ID compare all in one single clock cycle. Furthermore, because of the pipelining and the tri-ported architecture of the Pixel Buffer, the read and write operations may be performed on the Pixel Buffer of the 3D-RAM during the same clock cycle.

5.1.1 ROP/Blend Units

The ROP/Blend units can be configured as either a ROP unit or a Blend unit by setting a register bit. Each ROP unit can perform any of the 16 standard ROP functions. These functions are listed in Chapter 3. One of the operands of the ROP functions is the old data from the Pixel Buffer, and the other operand may be either the data from the primary I/O pins or the data from an internal register (called the Constant register). For the blending operation, the general equation is as following:

Write data to Pixel Buffer = New Term + (Old Data x Old Fraction)
= (New Data x New Fraction) + (Old Data x Old Fraction)

The 3D-RAM Blend units accomplish what is called destination blending, that is, the addition and the second multiplication in the above equation. The rendering controller must perform the multiplication of New Data with New Fraction (i.e. the source blending) and present the result as the New Term to 3D-RAM. It is important to note that the implementation of ROP/Blend units is sufficient to achieve a write-mostly interface on 3D-RAM.

The mathematics performed in the Blend unit is summarized in Table 3-2. The Clamped Result is written to the Pixel Buffer, depending (1) on the PASS_OUT pin which is the result of internal Compare units and (2) on the PASS_IN pin which is the PASS_OUT signal from the preceding 3D RAM.

Table 5-2 Mathematical operations in Blend unit n

Operand	Range	Sources	Comments
Old Fraction	0.00h~0.FFh (8-bit unsigned)	NX_n,$N_{[8n+7:8n]}$	Source is from PALU_DX_n and PALU_$DQ_{[8n+7:8n]}$ pins.
	0.00h~0.FFh (8-bit unsigned)	NX_3,$N_{[31:24]}$	Source is from PALU_DX_3 and PALU_$DQ_{[31:24]}$ pins.
	0.00h~0.FFh (8-bit unsigned)	KX_n,$K_{[8n+7:8n]}$	Source is from the internal Constant Register.
	1.00h (9-bit constant 1.00h)	1.00h	Fractions greater than 1.00h are clamped to 1.00h
Old Data	0~255 (8-bit unsigned)	$O_{[8n+7:8n]}$	Source is from the SRAM Pixel Buffer.
New Term	-256~255 (9-bit signed)	NX_n,$N_{[8n+7:8n]}$	Source is from PALU_DX_n and PALU_$DQ_{[8n+7:8n]}$ pins.
		KX_n,$K_{[8n+7:8n]}$	Source is from the internal Constant Register.
Intermediate Result	-256~510 (10-bit signed)	Old Fraction * Old Data + New Term	
Clamped Result	0~255 (8-bit unsigned)	Intermediate Result	The Clamped Result is written to the Pixel Buffer if the pass condition is valid. If source > 255, the result is clamped to 255. If source < 0, the result is clamped to 0.

5.1.2 Dual Compare Unit

Physically, the Dual Compare units consist of one 32-bit Match Compare unit and one 32-bit Magnitude Compare unit. Both Match Compare and Magnitude Compare are done in parallel. One of the sources is always the old data from the Pixel Buffer. The other source is

independently selectable between the data from the PALU_DQ pins and the data from the Constant register. There are also two mask registers, namely Match Mask and Magnitude Mask, that define which bits of the 32-bit words will be compared and which will be “don’t care”.

One application of the Match Compare unit is Window ID comparison, and the Magnitude Compare unit is typically used in the depth comparison of a Z-buffer algorithm for hidden surface removal. When these Compare units are used together, the system can achieve hidden surface removal for only a specific window on the display in one single cycle. Furthermore, since the data to be written into the Pixel Buffer always come through the ROP/Blend units, a system with 3D-RAM can achieve a pixel update with a raster or blending operation specifically on only the new objects in the selected window that are closer to the viewer than the existing objects in the frame buffer.

The results of both Match Compare and Magnitude Compare operations are logically ANDed together to generate the PASS_OUT pin. The PASS_IN signal (fed from another 3D-RAM chip) and the internally generated PASS_OUT signal are then logically ANDed together to produce a Write Enable signal to the Pixel Buffer. Thus, PASS_IN and PASS_OUT pins offer hardware support for display resolutions where multiple 3D-RAM chips are required, such as in the cases of 1280x 1024 x 32 (single color buffer plus Z buffer) and 1280 x 1024 x 96 (double color buffer plus Z buffer).

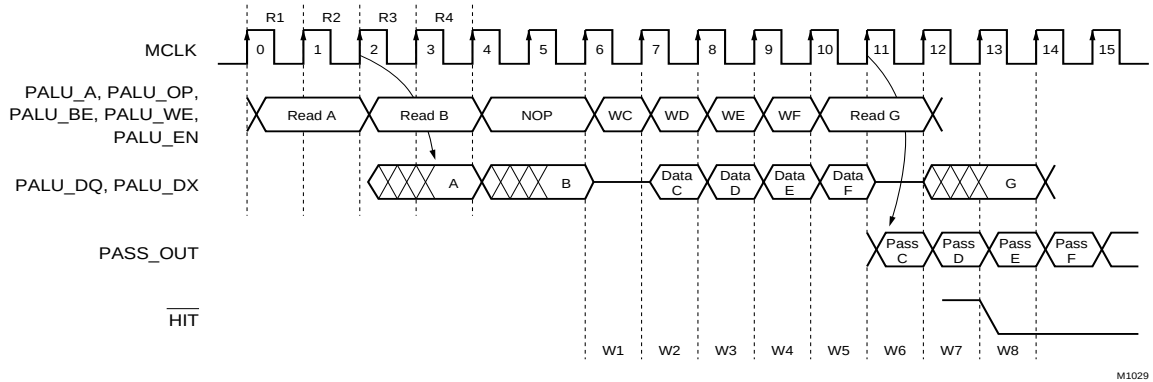
5.1.3 Pipelining

The 3D-RAM Pixel ALU pipeline is designed such that read and write operations can be done with minimal delay. This is achieved by having all operations adhere to a uniform 7-stage pipeline.

Figure 5-11 is an example to illustrate the efficiency afforded by the pipeline flow of Pixel ALU read/write operations. A stage of the pipeline begins with a rising edge of the MCLK and ends before the next rising edge of MCLK. (As a matter, in the 3D-RAM all references to MCLK are relative to the rising edge except for some boundary scan test operations.) For clarity, separate stage counts are provided for the first read and first write operations and are labeled as R1 through R4 and W1 through W7, respectively. The “Read A” operation is asserted for two cycles; “Read A” is first presented in Stage R1 and latched into the 3D-RAM by Clock 1 in Stage R2. The “Data A” is piped out by Clock 2 in Stage R3 and becomes stable for sampling in about Stage R4. Between “Read B” and “WC” (for Write C), two single-cycle NOPs are inserted to guarantee an idle cycle for data bus to turn around. On the other hand, a read operation can follow immediately after a write operation, as shown by “Read G” following “WF”. To allow maximum bandwidth for the rendering controller, a write operation may be started everything cycle. In this example, we start with the “WC” operation. The address and write

instruction are presented in Stage W1 and latched into the 3D-RAM by Clock 7 in Stage W2; “Data C” and “WD” are presented in Stage W2 and latched into the 3D-RAM by Clock 8 in Stage W3. Then, after three cycles for internal processing the valid PASS_OUT “Pass C” is piped out by Clock 11 in Stage W6. The actual updating of the Pixel Buffer takes place in Stage W7. Thus, *n* consecutive write operations take only *7 + n – 1 = n + 6* cycles to complete, including the all internal activities. It is important to point out, though, the effective write cycle time from the perspective of the rendering controller interface is only *n + 1* cycles for *n* consecutive write operations, as shown by “WC” through “WF”.

Figure 5-11 An example of Pixel Port read/write operations to satisfy the pipeline flow



5.1.4 The Picking Logic

From the user’s point of view, a common experience of picking function in 2D computer graphics may be using the mouse and the associated cursor to select an icon on the display screen and resulting in the selected icon to be highlighted by some different color. This is a basic function in interactive computer graphics, and 3D-RAM provides the Picking Logic and the $\overline{\text{HIT}}$ pin to support such picking function for selection of objects in a 3D scene.

A picking function may involve redrawing the objects into the frame buffer and returning a list of objects that intersect with some predefined selection volume. When multiple 3D-RAMs are used in a frame buffer design, to determine if a pixel data is successfully written by an ALU operation with plane masking and register effects enabled during the redraw process, the comparison result on the PASS_OUT pin from each chip needs to be logically ANDed. If this logical operation is left to off-chip glue logic between the 3D-RAM frame buffer and the rendering controller, excessive delay is unavoidable in this critical timing path. If the rendering

controller is to perform this logical operation, extra pins must be provided by both the 3D-RAM and the rendering controller, while delay is still significant. The Picking Logic brings the glue logic on chip and provides an open-drain $\overline{HIT}$ pin to interface with the rendering controller.

The block diagram of Picking Logic is shown in Figure 5-12.

Figure 5-12 The block diagram of Picking Logic