



© 2013 Intel Corporation

**Intel Open Source Graphics Programmer's Reference
Manual (PRM) for the 2013 Intel® Core™ Processor
Family, including Intel HD Graphics, Intel Iris™
Graphics and Intel Iris Pro Graphics**

Volume 5: Memory Views (Haswell)

12/18/2013



Copyright

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Implementations of the I2C bus/protocol may require licenses from various entities, including Philips Electronics N.V. and North American Philips Corporation.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others.

Copyright © 2013, Intel Corporation. All rights reserved.



Memory Views

Table of Contents

Introduction	9
GPU Memory Interface	10
GFX MMIO – MCHBAR Aperture	11
Graphics Memory Interface Functions	12
Address Tiling Function Introduction	13
Linear vs. Tiled Storage.....	14
Tile Formats.....	17
W-Major Tile Format.....	19
Tiling Algorithm	20
Tiled Channel Select Decision	23
Address Swizzling for Tiled-Y Surfaces	24
Address Swizzling For Tiled-X Surfaces.....	25
When and Where to use Tiled Address Swizzling.....	26
Tiling Support.....	27
Tiled (Fenced) Regions	28
Tiled Surface Parameters.....	29
Tiled Surface Restrictions	30
Per-Stream Tile Format Support	32
Graphics Translation Tables	33
Virtual Memory	34
Global Virtual Memory	35
Graphics Translation Table (GTT) Range (GTTADR).....	35
GTT Page Table Entries (PTEs).....	35
GTT Page Table Entries (PTEs)	35
Per Process GTT.....	38
Page Table Format	38
Page Walk.....	39
Two-Level Per-Process Virtual Memory	40
PPGTT Directory Entries (PDEs)	42
PPGTT Table Entries (PTEs).....	43



PPGTT Table Entries (PTEs)	43
PPGTT vs GGTT Limitation	45
Global GTT	46
Faulting.....	47
Page Faulting Support	47
Memory Types and Cache Interface	49
Memory Object Control State (MOCS)	50
MOCS Registers	51
Graphics Cache and Memory Interface	52
Features Added to Memory Interface for DevHSW	53
Memory Type: WB/UC/WT	53
Cache type: L3/LLC/eDRAM(eLLC)	53
Cache Replacement Age Allocation	53
Push Writes.....	54
Poorman's LIP25	54
Feature Set per DevHSW SKU	55
Enabling DevHSW Memory Interface Cache	56
GFX Translation Table (GTT) Entry.....	56
Memory Object Control State (MOCS).....	57
How are MOCS and PTE combined in DevHSW?	58
eDRAM (eLLC) Enabling Requirement on GFX Driver	59
Class of Service Usage for GFX	60
Driver Hints	62
Common Surface Formats	63
Non-Video Surface Formats	64
Surface Format Naming	65
Intensity Formats	66
Luminance Formats	67
R1_UNORM (same as R1_UINT) and MONO8.....	68
Palette Formats	69
P4A4_UNORM	69
A4P4_UNORM	69
P8A8_UNORM	70
A8P8_UNORM	70



P8_UNORM.....	71
P2_UNORM.....	71
Compressed Surface Formats	72
FXT Texture Formats.....	73
Overview of FXT1 Formats	74
FXT1 CC_HI Format.....	75
CC_HI Block Encoding	75
CC_HI Block Decoding.....	75
FXT1 CC_CHROMA Format	76
CC_CHROMA Block Encoding	77
CC_CHROMA Block Decoding.....	77
FXT1 CC_MIXED Format	79
CC_MIXED Block Encoding	79
CC_MIXED Block Decoding	79
FXT1 CC_ALPHA Format.....	83
CC_ALPHA Block Encoding.....	83
CC_ALPHA Block Decoding.....	85
DXT Texture Formats.....	87
Opaque and One-bit Alpha Textures (DXT1/BC1)	88
Opaque Textures (DXT1_RGB)	90
Compressed Textures with Alpha Channels (DXT2-5 / BC2-3)	91
BC4	94
BC5	96
BC6H.....	98
Field Definition.....	98
Endpoint Computation	115
Palette Color Computation	116
Texel Selection	117
ONE Mode	117
TWO Mode.....	118
BC7	120
Field Definition.....	121
Endpoint Computation	128
Palette Color Computation	129



Texel Selection	129
ONE Mode	129
TWO Mode.....	130
THREE Mode.....	132
Video Pixel/Texel Formats	134
Packed Memory Organization.....	135
Planar Memory Organization.....	136
Raw Format.....	138
Surface Memory Organizations.....	139
Display, Overlay, Cursor Surfaces	140
2D Render Surfaces	141
2D Monochrome Source.....	142
2D Color Pattern.....	143
3D Color Buffer (Destination) Surfaces.....	144
3D Depth Buffer Surfaces.....	145
3D Separate Stencil Buffer Surfaces.....	146
Surface Layout.....	147
Buffers	148
Structured Buffers	149
1D Surfaces	150
2D Surfaces	151
Computing MIP Level Sizes	151
Base Address for LOD Calculation	152
Minimum Pitch for MIPLAYOUT_RIGHT and Other Maps	154
Alignment Unit Size	154
Cartesian to Linear Address Conversion.....	154
Compressed Mipmap Layout.....	155
Surface Arrays.....	155
For All Surface Other Than Separate Stencil Buffer.....	155
Multisampled Surfaces	156
Compressed Multisampled Surfaces	156
Physical MSS Surface.....	157
Uncompressed Multisampled Surfaces	157
Cube Surfaces	158



Hardware Cube Map Layout.....	158
Restrictions.....	158
Cube Arrays.....	158
3D Surfaces	159
Minimum Pitch.....	161
Surface Padding Requirements	162
Sampling Engine Surfaces.....	163
Render Target and Media Surfaces	164





Introduction

The hardware supports three engines:

- The Render command streamer interfaces to 3D/IE and display streams.
- The Media command streamer interfaces to the fixed function media.
- The Blitter command streamer interfaces to the blit commands.

Software interfaces of all three engines are very similar and should only differ on engine-specific functionality.



GPU Memory Interface

GPU memory interface functions are divided into 4 different major sections:

- Global Arbitration
- Memory Interface Functions
- Page Translations (GFX Page Walker)
- Ring Interface Functions (GTI)

GT Interface functions are covered at a different chapter/HAS and not part of this documentation. The following documentation is meant for GFX arbitration paths in accessing to memory/cache interfaces and page translations and page walker functions.



GFX MMIO – MCHBAR Aperture

Address: 140000h – 17FFFFh

Default Value: Same as MCHBAR

Access: Aligned Word, Dword or Qword Read/Write

This range defined in the graphics MMIO range is an alias with which graphics driver can read and write registers defined in the MCHBAR MMIO space claimed through Device #0. Attributes for registers defined within the MCHBAR space are preserved when the same registers are accessed via this space. Registers that the graphics driver requires access to are Rank Throttling, GMCH Throttling, Thermal Sensor etc.

The Alias functions works for MMIO access from the CPU only. A command stream load register immediate will drop the data and store register immediate will return all Zeros.

Graphics MMIO registers can be accessed through MMIO BARs in function #0 and function #1 in Device #2. The aliasing mechanism is turned off if memory access to the corresponding function is turned off via software or in certain power states.



Graphics Memory Interface Functions

The major role of an integrated graphics device's Memory Interface (MI) function is to provide various client functions access to "graphics" memory used to store commands, surfaces, and other information used by the graphics device. This chapter describes the basic mechanisms and paths by which graphics memory is accessed.

Information not presented in this chapter includes:

- Microarchitectural and implementation-dependent features (e.g., internal buffering, caching and arbitration policies).
- MI functions and paths specific to the operation of external (discrete) devices attached via external connections.
- MI functions essentially unrelated to the operation of the internal graphics devices, .e.g., traditional "chipset functions" .
- GFX Page Walker and GT interface functions are covered in different chapters.



Address Tiling Function Introduction

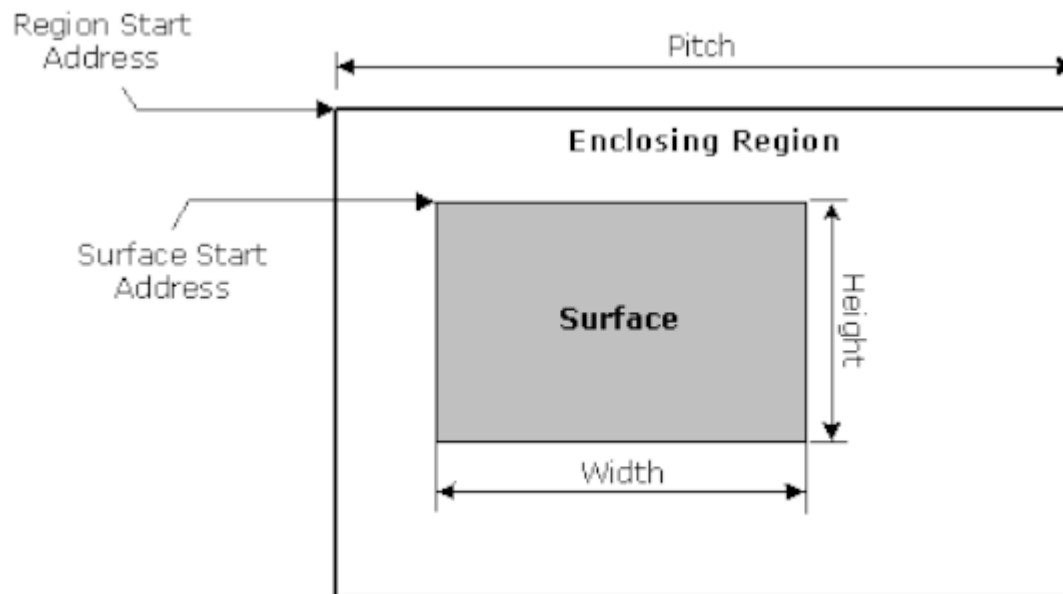
When dealing with memory operands (e.g., graphics surfaces) that are inherently rectangular in nature, certain functions within the graphics device support the storage/access of the operands using alternative (tiled) memory formats to increase performance. This section describes these memory storage formats, why and when they should be used, and the behavioral mechanisms within the device to support them.



Linear vs. Tiled Storage

Regardless of the memory storage format, "rectangular" memory operands have a specific *width* and *height*, and are considered as residing within an enclosing rectangular region whose width is considered the *pitch* of the region and surfaces contained within. Surfaces stored within an enclosing region must have widths less than or equal to the region pitch (indeed the enclosing region may coincide exactly with the surface). *Linear vs. Tiled Storage* shows these parameters.

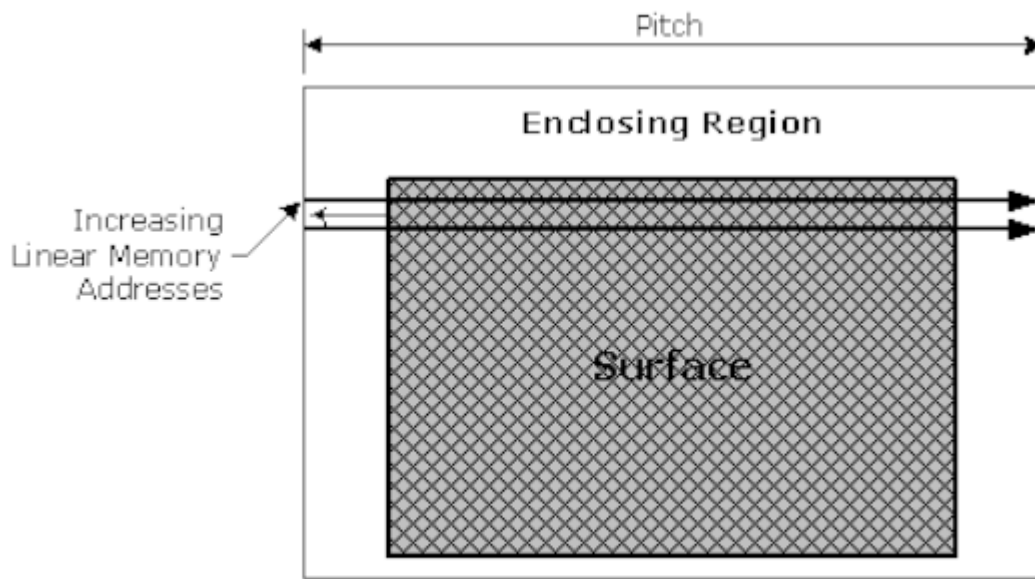
Rectangular Memory Operand Parameters



B 6690-01

The simplest storage format is the *linear* format (see *Linear vs. Tiled Storage*), where each row of the operand is stored in sequentially increasing memory locations. If the surface width is less than the enclosing region's pitch, there will be additional memory storage between rows to accommodate the region's pitch. The pitch of the enclosing region determines the distance (in the memory address space) between vertically-adjacent operand elements (e.g., pixels, texels).

Linear Surface Layout



B6691-01

The linear format is best suited for 1-dimensional row-sequential access patterns (e.g., a display surface where each scanline is read sequentially). Here the fact that one object element may reside in a different memory page than its vertically-adjacent neighbors is not significant; all that matters is that horizontally-adjacent elements are stored contiguously. However, when a device function needs to access a 2D subregion within an operand (e.g., a read or write of a 4x4 pixel span by the 3D renderer, a read of a 2x2 texel block for bilinear filtering), having vertically-adjacent elements fall within different memory pages is to be avoided, as the page crossings required to complete the access typically incur increased memory latencies (and therefore lower performance).

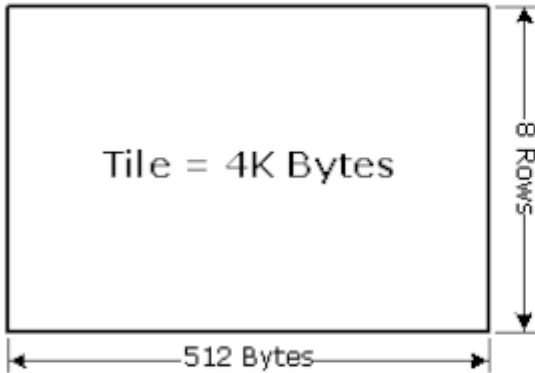
One solution to this problem is to divide the enclosing region into an array of smaller rectangular regions, called memory *tiles*. Surface elements falling within a given tile will all be stored in the same physical memory page, thus eliminating page-crossing penalties for 2D subregion accesses within a tile and thereby increasing performance.

Tiles have a fixed 4KB size and are aligned to physical DRAM page boundaries. They are either 8 rows high by 512 bytes wide or 32 rows high by 128 bytes wide (see *Linear vs. Tiled Storage*). Note that the dimensions of tiles are irrespective of the data contained within – e.g., a tile can hold twice as many 16-bit pixels (256 pixels/row x 8 rows = 2K pixels) than 32-bit pixels (128 pixels/row x 8 rows = 1K pixels).

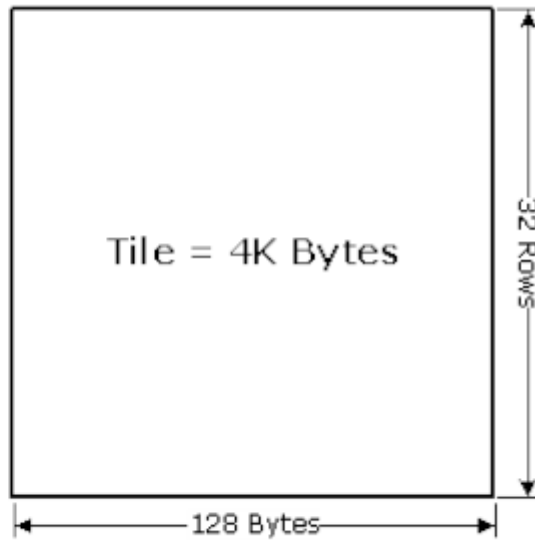
Memory Tile Dimensions



X Tile Dimensions



Y Tile Dimensions

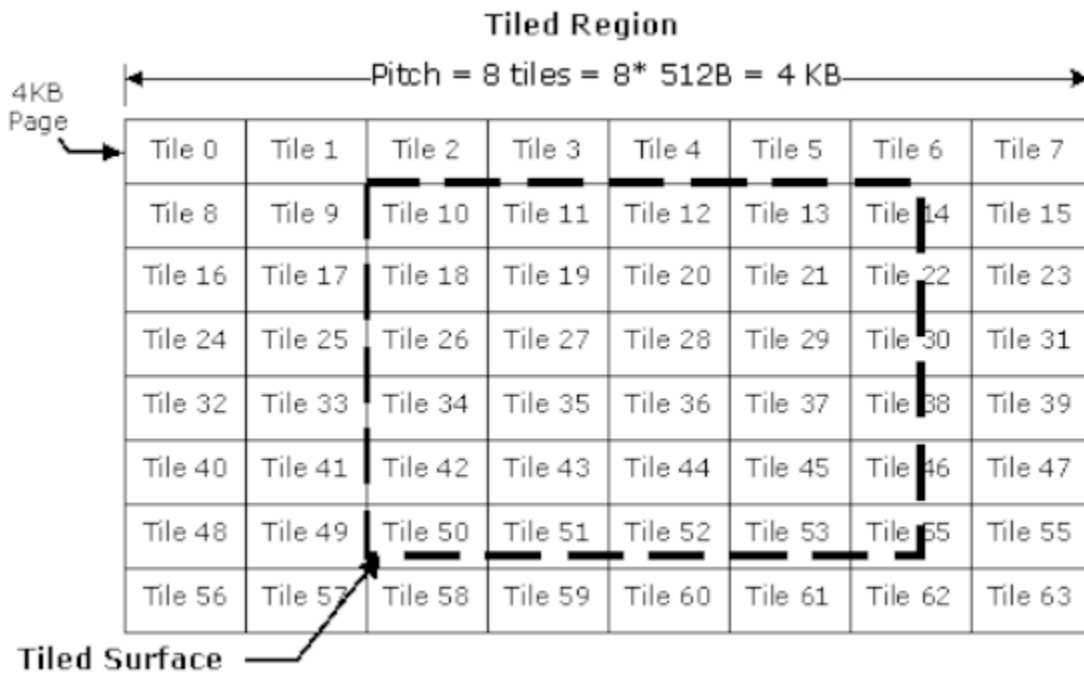


B6692-01

The pitch of a tiled enclosing region must be an integral number of tile widths. The 4KB tiles within a tiled region are stored sequentially in memory in row-major order.

The *Linear vs. Tiled Storage* figure shows an example of a tiled surface located within a tiled region with a pitch of 8 tile widths (512 bytes * 8 = 4KB). Note that it is the *enclosing region* that is divided into tiles – the surface is not necessarily aligned or dimensioned to tile boundaries.

Tiled Surface Layout

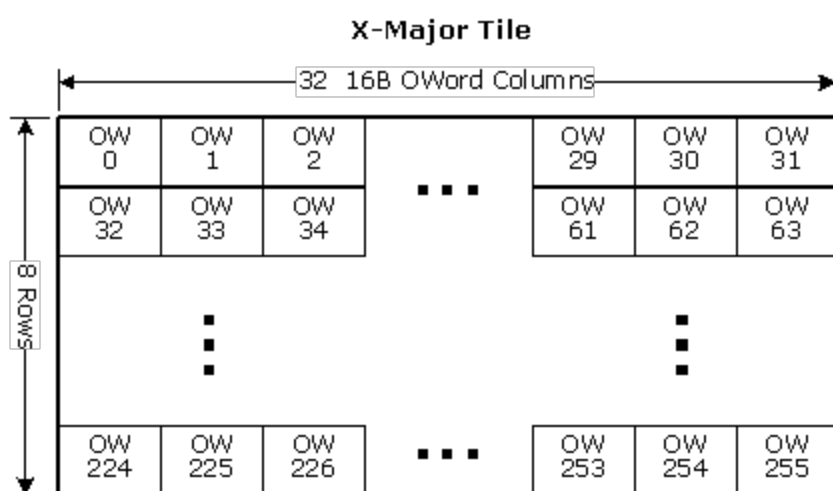


B6693-01

Tile Formats

The device supports both *X-Major* (row-major) and *Y-Major* (column major) storage of tile data units, as shown in the following figures. A 4KB tile is subdivided into an 8-high by 32-wide array of 16-byte OWords for X-Major Tiles (X Tiles for short), and 32-high by 8-wide array of OWords for Y-Major Tiles (Y Tiles). The selection of tile direction only impacts the internal organization of tile data, and does not affect how surfaces map onto tiles. Note that the diagrams are not to scale – the first format defines the contents of an 8-high by 512-byte wide tile, and the second a 32-high by 128-byte wide tile. The storage of tile data units in X-Major or Y-Major fashion is sometimes refer to as the *walk* of the tiling.

X-Major Tile Layout

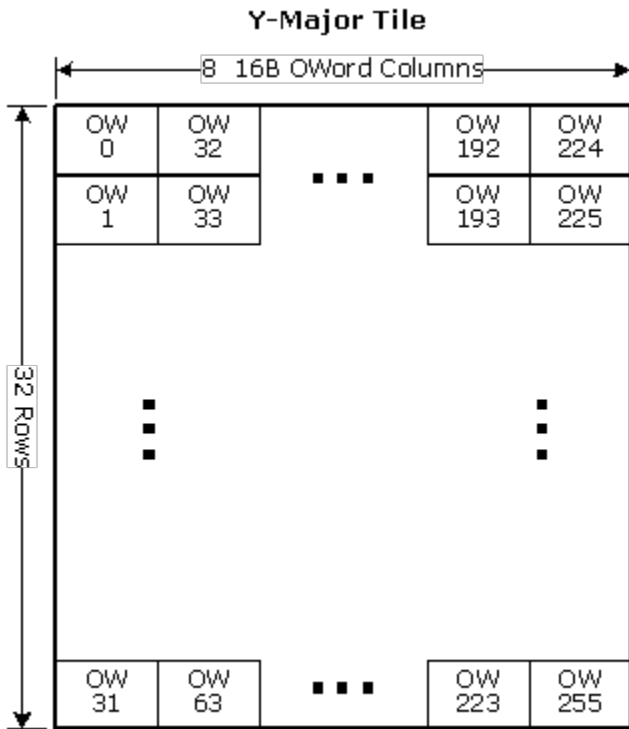


B6694-01

Note that an X-major tiled region with a tile pitch of 1 tile is actually stored in a linear fashion.



Y-Major Tile Layout

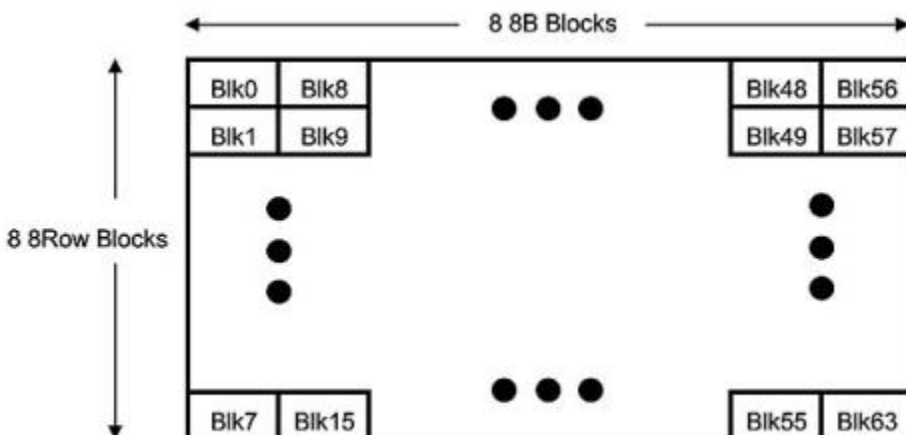


B6695-01

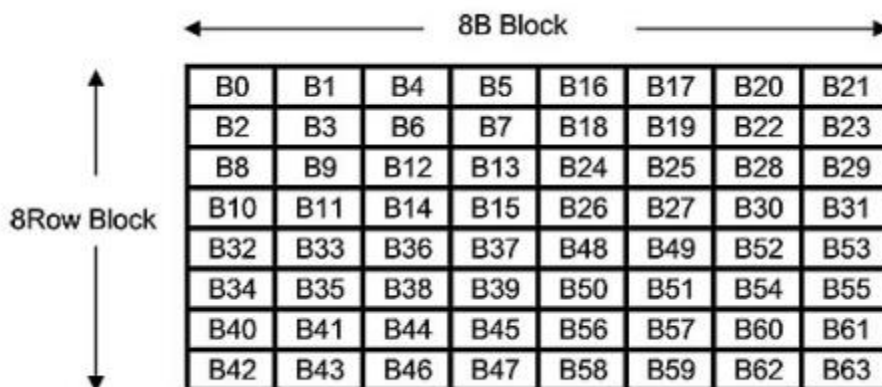
W-Major Tile Format

The device supports additional format *W-Major* storage of tile data units, as shown in the following figures. A 4 KB tile is subdivided into 8-high by 8-wide array of Blocks for W-Major Tiles (W Tiles). Each Block is 8 rows by 8 bytes. The selection of tile direction only impacts the internal organization of tile data, and does not affect how surfaces map onto tiles. W-Major Tile Format is used for separate stencil.

W-Major Tile Layout



W-Major Block Layout





Tiling Algorithm

The following pseudocode describes the algorithm for translating a tiled memory surface in graphics memory to an address in logical space.

```
Inputs: LinearAddress(offset into regular or LT aperture in terms of
bytes),
    Pitch_programmed(in Surface State or Stencil Buffer State),
    WalkY (1 for Y and 0 for the rest)
    WalkW (1 for W and 0 for the rest)

Static Parameters:
TileH (Height of tile, 8 for X, 32 for Y and 64 for W),
    TileW (Width of Tile in bytes, 512 for X, 128 for Y and 64 for W)
TileSize = TileH * TileW;
Pitch_in_Bytes = WalkW ? (Pitch_programmed+1) div 2
    : Pitch_programmed+1;
Pitch_in_Tiles = Pitch_in_Bytes div TileW;
RowSize = Pitch_in_Tiles * TileSize;
If (Fenced) {
    LinearAddress = LinearAddress - FenceBaseAddress
    LinearAddrInTileW = LinearAddress div TileW;
    Xoffset_inTile = LinearAddress mod TileW;
    Y = LinearAddrInTileW div Pitch_in_Tiles;
    X = LinearAddrInTileW mod Pitch_in_Tiles + Xoffset_inTile;
}
// Internal graphics clients that access tiled memory already have the
X, Y
// coordinates and can start here
YOff_Within_Tile = Y mod TileH;
XOff_Within_Tile = X mod TileW;
TileNumber_InY = Y div TileH;
TileNumber_InX = X div TileW;
    TiledOffsetY = RowSize * TileNumber_InY + TileSize *
    TileNumber_InX + TileH * 16 * (XOff_Within_Tile div 16) +
    YOff_Within_Tile * 16 + (XOff_Within_Tile mod 16);
    TiledOffsetW = RowSize * TileNumber_InY + TileSize *
    TileNumber_InX +
    TileH * 8 * (XOff_Within_Tile div 8) +
```



```
64* (YOff_Within_Tile div 8)+
32*((YOff_Within_Tile div 4) mod 2) +
16* ((XOff_Within_Tile div 4) mod 2) +
8 * ((YOff_Within_Tile div 2) mod 2) +
4* ((XOff_Within_Tile div 2) mod 2) +
2 * (YOff_Within_Tile mod 2) +
(XOff_Within_Tile mod 2);

    TiledOffsetX = RowSize * TileNumber_InY + TileSize *
    TileNumber_InX + TileW * YOff_Within_Tile +
    XOff_Within_Tile;

TiledOffset = WalkW? TiledOffsetW : (WalkY? TiledOffsetY :
TiledOffsetX);

    TiledAddress = Tiled? (BaseAddress + TiledOffset):
    (BaseAddress + Y*LinearPitch + X);TiledAddress = (Tiled &&
    (Address Swizzling for Tiled-Surfaces == 01)) ?
    (WalkW || WalkY) ?
    (TiledAddress div 128) * 128 +
    (((TiledAddress div 64) mod 2) ^
    ((TiledAddress div 512) mod 2)) +
    (TiledAddress mod 32)
    :
    (TiledAddress div 128) * 128 +
    (((TiledAddress div 64) mod 2) ^
    ((TiledAddress div 512) mod 2)
    ((TiledAddress Div 1024) mod2) +
    (TiledAddress mod 32)
    :
    TiledAddress;
}
```



For Address Swizzling for Tiled-Surfaces see ARB_MODE – Arbiter Mode Control register, ARB_CTL— Display Arbitration Control 1 and TILECTL - Tile Control register

The Y-Major tile formats have the characteristic that a surface element in an even row is located in the same aligned 64-byte cacheline as the surface element immediately below it (in the odd row). This spatial locality can be exploited to increase performance when reading 2x2 texel squares for bilinear texture filtering, or reading and writing aligned 4x4 pixel spans from the 3D Render pipeline.

On the other hand, the X-Major tile format has the characteristic that horizontally-adjacent elements are stored in sequential memory addresses. This spatial locality is advantageous when the surface is scanned in row-major order for operations like display refresh. For this reason, the Display and Overlay memory streams only support linear or X-Major tiled surfaces (Y-Major tiling is not supported by these functions). This has the side effect that 2D- or 3D-rendered surfaces must be stored in linear or X-Major tiled formats if they are to be displayed. Non-displayed surfaces, e.g., "rendered textures", can also be stored in Y-Major order.



Tiled Channel Select Decision

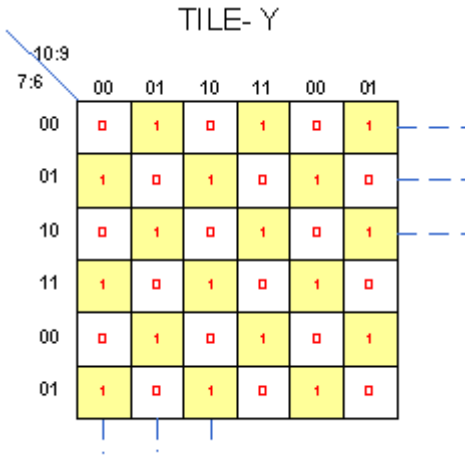
In order to spread DRAM accesses between multiple channels in the most efficient way, address bits are used to select the channel. The most common DRAM configuration is 2-channels with 64B interleaving where address bit[6] is used as channel select. However for tiled accesses, using bit[6] as is can be an issue due to back to back accesses to have different patterns compared to liner streams.

For linear stream (no-X/Y tiling) address bit[6] has no modification.



Address Swizzling for Tiled-Y Surfaces

The following board re-defines the address bit[6] after tiling.

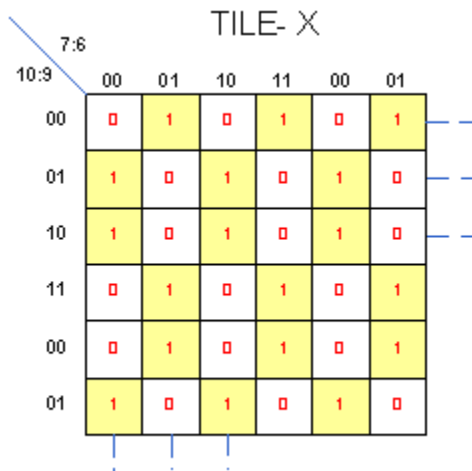


As shown in the tiling algorithm, The new address bit[6] becomes :

Address bit[6] <= TiledAddr bit[6] XOR TiledAddr bit[9]

Address Swizzling For Tiled-X Surfaces

Similar to Tiled-Y, we need to redefine the address bit[6] but for Tile-X, field mode needs to be taken into account.



As shown in the tiling algorithm, the new address bit[6] should be:

Address bit[6] <= TiledAddr bit[6] XOR TiledAddr bit[9] XOR TiledAddr bit[10]



When and Where to use Tiled Address Swizzling

Address swizzling for tiled surfaces will be used for certain DRAM channel configurations with 64B (or more) interleaving. However it is up to the discretion of the GFX driver to set up the system to enable address swizzling.

The need for address swizzling on Tiled Surfaces is communicated by GFX Driver to HW via MMIO register updates. The following register present in three places of GFX MMIO region and driver is responsible for consistent programming of all these values

[1:0]	R/W	00h	Address Swizzling for Tiled-Surfaces: This register location is updated via GFX Driver before enabling DRAM accesses. Driver needs to obtain the need for memory address swizzling via DRAM configuration registers and set the following bits (in Display Engine and Render/Media access streams): 00b: No address Swizzling 01b: Address bit[6] needs to be swizzled for tiled surfaces 10b: Reserved 11b: Reserved
-------	-----	-----	---

Note: the bit positions can vary due to assignments.

For GT => offset 0x0000_4030[5:4] ARB_MODE register

For DE => offset 0x0004_5000[14:13] Arbiter Control register



Tiling Support

The rearrangement of the surface elements in memory must be accounted for in device functions operating upon tiled surfaces. (Note that not all device functions that access memory support tiled formats). This requires either the modification of an element's linear memory address or an alternate formula to convert an element's X,Y coordinates into a tiled memory address.

However, before tiled-address generation can take place, some mechanism must be used to determine whether the surface elements accessed fall in a linear or tiled region of memory, and if tiled, what the tile region pitch is, and whether the tiled region uses X-Major or Y-Major format. There are two mechanisms by which this detection takes place: (a) an implicit method by detecting that the pre-tiled (linear) address falls within a "fenced" tiled region, or (b) by an explicit specification of tiling parameters for surface operands (i.e., parameters included in surface-defining instructions).

The following table identifies the tiling-detection mechanisms that are supported by the various memory streams.

Access Path	Tiling-Detection Mechanisms Supported
Processor access through the Graphics Memory Aperture	Fenced Regions
3D Render (Color/Depth Buffer access)	Explicit Surface Parameters
Sampled Surfaces	Explicit Surface Parameters
Blt operands	Explicit Surface Parameters
Display and Overlay Surfaces	Explicit Surface Parameters



Tiled (Fenced) Regions

The only mechanism to support the access of surfaces in tiled format by the host or external graphics client is to place them within "fenced" tiled regions within Graphics Memory. A fenced region is a block of Graphics Memory specified using one of the sixteen FENCE device registers. (See *Memory Interface Registers* for details). Surfaces contained within a fenced region are considered tiled from an external access point of view. Note that fences cannot be used to untile surfaces in the PGM_Address space since external devices cannot access PGM_Address space. Even if these surfaces (or any surfaces accessed by an internal graphics client) fall within a region covered by an enabled fence register, that enable will be effectively masked during the internal graphics client access. Only the explicit surface parameters described in the next section can be used to tile surfaces being accessed by the internal graphics clients.

Each FENCE register (if its Fence Valid bit is set) defines a Graphics Memory region ranging from 4KB to the aperture size. The region is considered rectangular, with a pitch in tile widths from 1 tile width (128B or 512B) to 256 tile X widths ($256 * 512B = 128KB$) and 1024 tile Y widths ($1024 * 128B = 128KB$). Note that fenced regions must not overlap, or operation is UNDEFINED.

Also included in the FENCE register is a Tile Walk field that specifies which tile format applies to the fenced region.



Tiled Surface Parameters

Internal device functions require explicit specification of surface tiling parameters via information passed in commands and state. This capability is provided to limit the reliance on the fixed number of fence regions.

The following table lists the surface tiling parameters that can be specified for 3D Render surfaces (Color Buffer, Depth Buffer, Textures, etc.) via SURFACE_STATE.

Surface Parameter	Description
Tiled Surface	If ENABLED, the surface is stored in a tiled format. If DISABLED, the surface is stored in a linear format.
Tile Walk	If Tiled Surface is ENABLED, this parameter specifies whether the tiled surface is stored in Y-Major or X-Major tile format.
Base Address	Additional restrictions apply to the base address of a Tiled Surface vs. that of a linear surface.
Pitch	Pitch of the surface. Note that, if the surface is tiled, this pitch must be a multiple of the tile width.



Tiled Surface Restrictions

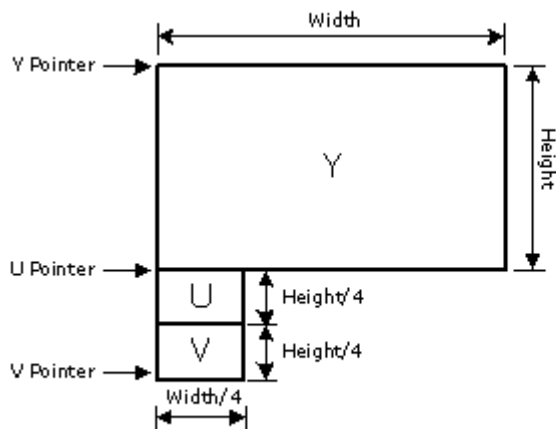
Additional restrictions apply to the Base Address and Pitch of a surface that is tiled. In addition, restrictions for tiling via SURFACE_STATE are subtly different from those for tiling via fence regions. The most restricted surfaces are those that will be accessed both by the host (via fence) and by internal device functions. An example of such a surface is a tiled texture that is initialized by the CPU and then sampled by the device.

The tiling algorithm for internal device functions is different from that of fence regions. Internal device functions always specify tiling in terms of a surface. The surface must have a base address, and this base address is not subject to the tiling algorithm. Only *offsets* from the base address (as calculated by X, Y addressing within the surface) are transformed through tiling. The base address of the surface must therefore be 4KB-aligned. This forces the 4KB tiles of the tiling algorithm to exactly align with 4KB device pages once the tiling algorithm has been applied to the offset. The width of a surface must be less than or equal to the surface pitch. There are additional considerations for surfaces that are also accessed by the host (via a fence region).

Fence regions have no base address per se. Host linear addresses that fall in a fence region are translated in their entirety by the tiling algorithm. It is as if the surface being tiled by the fence region has a base address in graphics memory equal to the fence base address, and all accesses of the surfaces are (possibly quite large) offsets from the fence base address. Fence regions have a virtual "left edge" aligned with the fence base address, and a "right edge" that results from adding the fence pitch to the "left edge". Surfaces in the fence region must not straddle these boundaries.

Base addresses of surfaces that are to be accessed both by an internal graphics client and by the host have the tightest restrictions. In order for the surface to be accessed without GTT re-mapping, the surface base address (as set in SURFACE_STATE) must be a "Tile Row Start Address" (TRSA). The first address in each tile row of the fence region is a Tile Row Start Address. The first TRSA is the fence base address. Each TRSA can be generated by adding an integral multiple of the row size to the fence base address. The row size is simply the fence pitch in tiles multiplied by 4KB (the size of a tile.)

Tiled Surface Placement



B6685-01

The pitch in SURFACE_STATE must be set equal to the pitch of the fence that will be used by the host to access the surface if the same GTT mapping will be used for each access. If the pitches differ, a different GTT mapping must be used to eliminate the "extra" tiles (4KB memory pages) that exist in the excess rows at the right side of the larger pitch. Obviously no part of the surface that will be accessed can lie in pages that exist only in one mapping but not the other. The new GTT mapping can be done manually by SW between the time the host writes the surface and the device reads it, or it can be accomplished by arranging for the client to use a different GTT than the host (the PPGTT -- see **Logical Memory Mapping**).

The width of the surface (as set in SURFACE_STATE) must be less than or equal to both the surface pitch and the fence pitch in any scenario where a surface will be accessed by both the host and an internal graphics client. Changing the GTT mapping will not help if this restriction is violated.

Surface Access	Base Address	Pitch	Width	Tile "Walk"
Host only	No restriction	Integral multiple of tile size $\leq 128\text{KB}$	Must be \leq Fence Pitch	No restriction
Client only	4KB-aligned	Integral multiple of tile size $\leq 256\text{KB}$	Must be \leq Surface Pitch	Restrictions imposed by the client (see <i>Per-Stream Tile Format Support</i>)
Host and Client, No GTT Remapping	Must be TRSA	Fence Pitch = Surface Pitch = integral multiple of tile size $\leq 256\text{KB}$	Width \leq Pitch	Surface Walk must meet client restriction, Fence Walk = Surface Walk
Host and Client, GTT Remapping	4KB-aligned for client (will be tile-aligned for host)	Both must be Integral multiple of tile size $\leq 128\text{KB}$, but not necessarily the same	Width \leq Min(Surface Pitch, Fence Pitch)	Surface Walk must meet client restriction, Fence Walk = Surface Walk



Per-Stream Tile Format Support

MI Client	Tile Formats Supported	
CPU Read/Write	All	
Display/Overlay	Y-Major not supported. X-Major required for Async Flips	
Blt	Linear and X-Major only No Y-Major support	
3D Sampler	All Combinations of TileY, TileX and Linear are supported. TileY is the fastest, Linear is the slowest.	
3D Color,Depth	Rendering Mode Color-vs-Depth bpp	Buffer Tiling Supported
	Classical Same Bpp	Both Linear Both TileX Both TileY Linear & TileX Linear & TileY TileX & TileY
	Classical Mixed Bpp	Both Linear Both TileX Both TileY Linear & TileX Linear & TileY TileX & TileY
NOTE: 128 BPP format color buffer (render target) supports Linear, TiledX and TiledY.		



Graphics Translation Tables

The Graphics Translation Tables GTT (Graphics Translation Table, sometimes known as the global GTT) and PPGTT (Per-Process Graphics Translation Table) are memory-resident page tables containing an array of DWord Page Translation Entries (PTEs) used in mapping logical Graphics Memory addresses to physical memory addresses, and sometimes snooped system memory "PCI" addresses.

The graphics translation tables must reside in (unsnooped) system memory.

The base address (MM offset) of the GTT and the PPGTT are programmed via the PGTBL_CTL and PGTBL_CTL2 MI registers, respectively. The translation table base addresses must be 4KB aligned. The GTT size can be either 128KB, 256KB, or 512KB (mapping to 128MB, 256MB, and 512MB aperture sizes respectively) and is physically contiguous. The global GTT should only be programmed via the range defined by GTTADR. The PPGTT is programmed directly in memory. The per-process GTT (PPGTT) size is controlled by the PGTBL_CTL2 register. The PPGTT can, in addition to the above sizes, also be 64KB in size (corresponding to a 64MB aperture). Refer to the GTT Range chapter for a bit definition of the PTE entries.

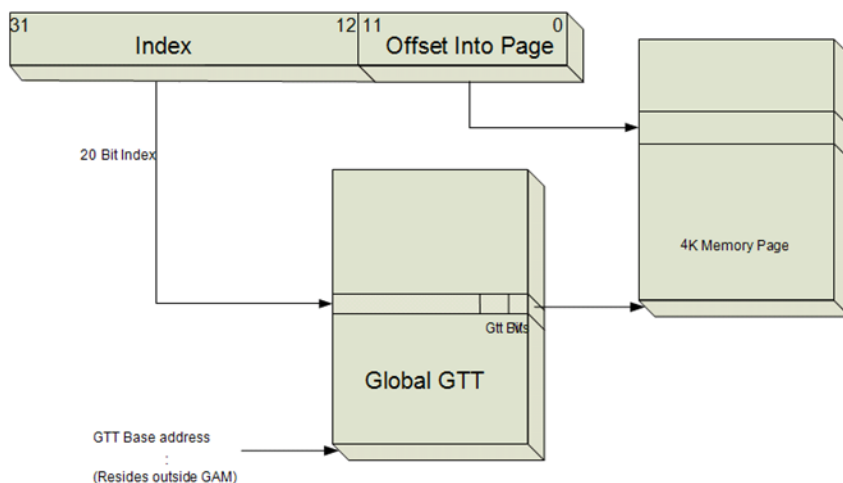


Virtual Memory

GT supports standard virtual memory models as defined by the IA programmer's guide. This section describes the different paging models, their behaviors and the page-table formats.

Global Virtual Memory

Global Virtual Memory is the default target memory if a PPGTT is not enabled. If a PPGTT is also present, the method to choose which is targeted by memory and rendering operations varies by product. See the sections on Per-Process Virtual Memory for more information. High priority graphics clients such as Display and Cursor always access global virtual memory.



Graphics Translation Table (GTT) Range (GTTADR)

Address: GTTADR in CPU Physical Space

Access: Aligned DWord Read/Write

The GTTADR memory BAR defined in graphics device config space is an alias for the Global GTT.

Programming Notes: It is recommended that the driver map all graphics memory pages in the GGTT to some physical page, if only a dummy page.

GTT Page Table Entries (PTEs)

This topic is currently under development.

GTT Page Table Entries (PTEs)

Page Table Entry: 1 DWord per 4KB Graphics Memory page.

OLD Page Table Entry (PTE)

31	12	11:04	3	2	1	0
Physical Page Address 31:12		Physical Page Address 39:32	GFDT	Cacheability Control		Valid



NEW Page Table Entry (PTE)

31:12	11	10:04	3:1	0
Physical Page Address 31:12	Cacheability Control[3]	Physical Page Address 38:32	Cacheability Control[2:0]	Valid

Bits	Description
31:12	Physical Page Address 31:12: If the Valid bit is set, This field provides the page number of the physical memory page backing the corresponding Graphics Memory page.
11	Cacheability Control[3] With the addition of new modes for LLC and addition of eDRAM the use of the cacheability controls via GTT has been updated – see below table
10:4	Physical Start Address Extension: This field specified Bits 38:32 of the page table entry. This field must be zero for 32 bit addresses.
3:1	Cacheability Control[2:0] With the addition of new modes for LLC and addition of eDRAM the use of the cacheability controls via GTT has been updated – see below table
0	Valid PTE: This field indicates whether the mapping of the corresponding Graphics Memory page is valid. 1: Valid 0: Invalid. An access (other than a CPU Read) through an invalid PTE will result in Page Table Error (Invalid PTE).

Cacheability Control field

Cacheability Control [3:0]	Description
0000	UC (LLC/eLLC) – allocation age is don't care
0111	WT in LLC/eLLC – Aged "3"
0110	WT in LLC/eLLC – Aged "0"
0101	WT in eLLC (UC in LLC) – Aged "3"
0100	WT in eLLC (UC in LLC) – Aged "0"
1000	WB in LLC/eLLC – Aged "3"
1001	WB in LLC/eLLC – Aged "2"

1010	WB in LLC/eLLC – Aged "1"
1011	WB in LLC/eLLC – Aged "0"
1100	WB in eLLC (UC in LLC) – Aged "3"
1101	WB in eLLC (UC in LLC) – Aged "2"
1110	WB in eLLC (UC in LLC) – Aged "1"
1111	WB in eLLC (UC in LLC) – Aged "0"
0010	WB in LLC Only – Aged "3"
0011	WB in LLC Only – Aged "0"

Note: LLC Only has been added to support MSAA color compression scheme where higher order plans are not getting victimized to eLLC where there has been to have poor spatial localization. Such case would lead to poor superline utilization in terms of number of 64Bs valid.

Surface State Treatment

Surface state treatment is separated for L3 clients vs non-L3 clients. All L3 clients are exposed to GAM with L3 SourceID, hence they can be identified.

L3 Clients: All Data Port accesses (HDC), Textures, State, Instruction and Constant buffers.

- L3 Clients: Control field bit[0] is already used for L3 cacheability, hence needs to be ignored by GAM. Control bit[1] can be used for
 - a. "0": Use PTE values
 - b. "1": Cache in LLC/eLLC as WB with age of "3"

Note: For L3 clients, it is strongly suggested for GFX driver to use PTE to set the caching attributes and keep the Control Field bit[1]=0

- Non-L3 Clients: Both control bits are available
 - a. "00": Use PTE values
 - b. "01" UC – uncacheable
 - c. "10" LLC/eLLC WB cacheable
 - d. "11" eLLC WB cacheable (UC in LLC)

Note: Use age of "3" when overridden via surface state.

Note that GFDT is always passed through the SurfaceState and there is no control of it through PTE. HSW GFDT support is not guaranteed via h/w.



Per Process GTT

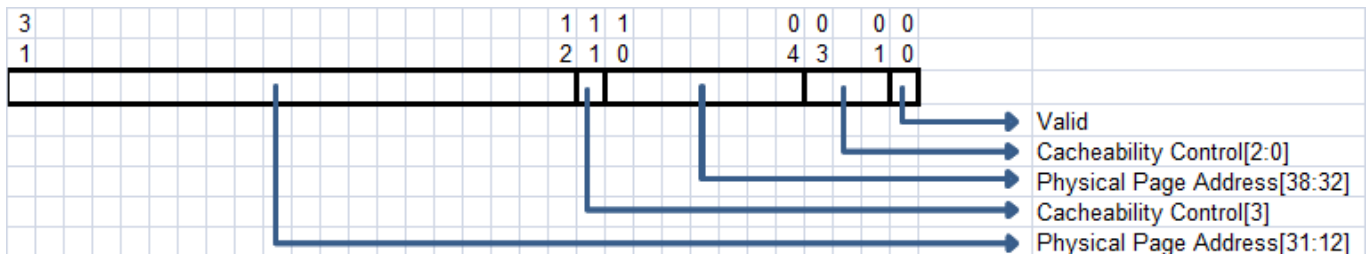
The Gen6-Gen7.5 family of GPUs supports a 2-level mapping scheme for PPGTT, consisting of a second-level page directory containing page table base addresses, and the page tables themselves on the first level, consisting of page addresses. The motivation for the 2-level scheme is simple – it allows for the lookup table (the collection of page tables) to exist in discontinuous memory, making allocation of memory for these structures less problematic for the OS. The directory and each page table fit within a single 4K page of memory that can be located anywhere in physical memory.

If a PPGTT is enabled, all rendering operations (including blit commands) target Per-process virtual memory. This means *all* commands *except* the Memory Interface Commands (MI_*). Certain Memory Interface Commands have a flag to choose global virtual memory (mapped via the GGTT) instead of per-process memory. Global Virtual Memory can be thought of as "privileged" memory in this case. Commands that elect to access privileged memory must have sufficient access rights to do so. Commands executing directly from a ring buffer or from a "secure" batch buffer (see the MI_BATCH_BUFFER_START command in Memory Interface Commands) have these access rights; other commands do not and are not permitted to access global virtual memory.

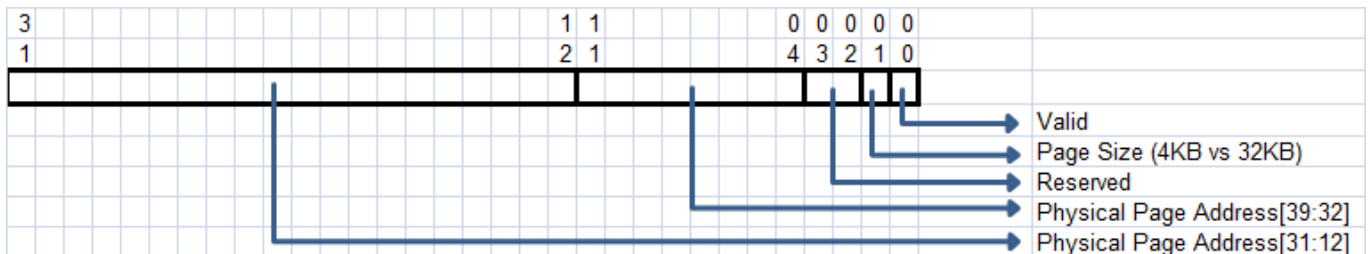
Page Table Format

Two levels have slightly different format, both have 32bits allowing 1024 entries per page mapping 2^{10} pages.

PTE (1st Level):

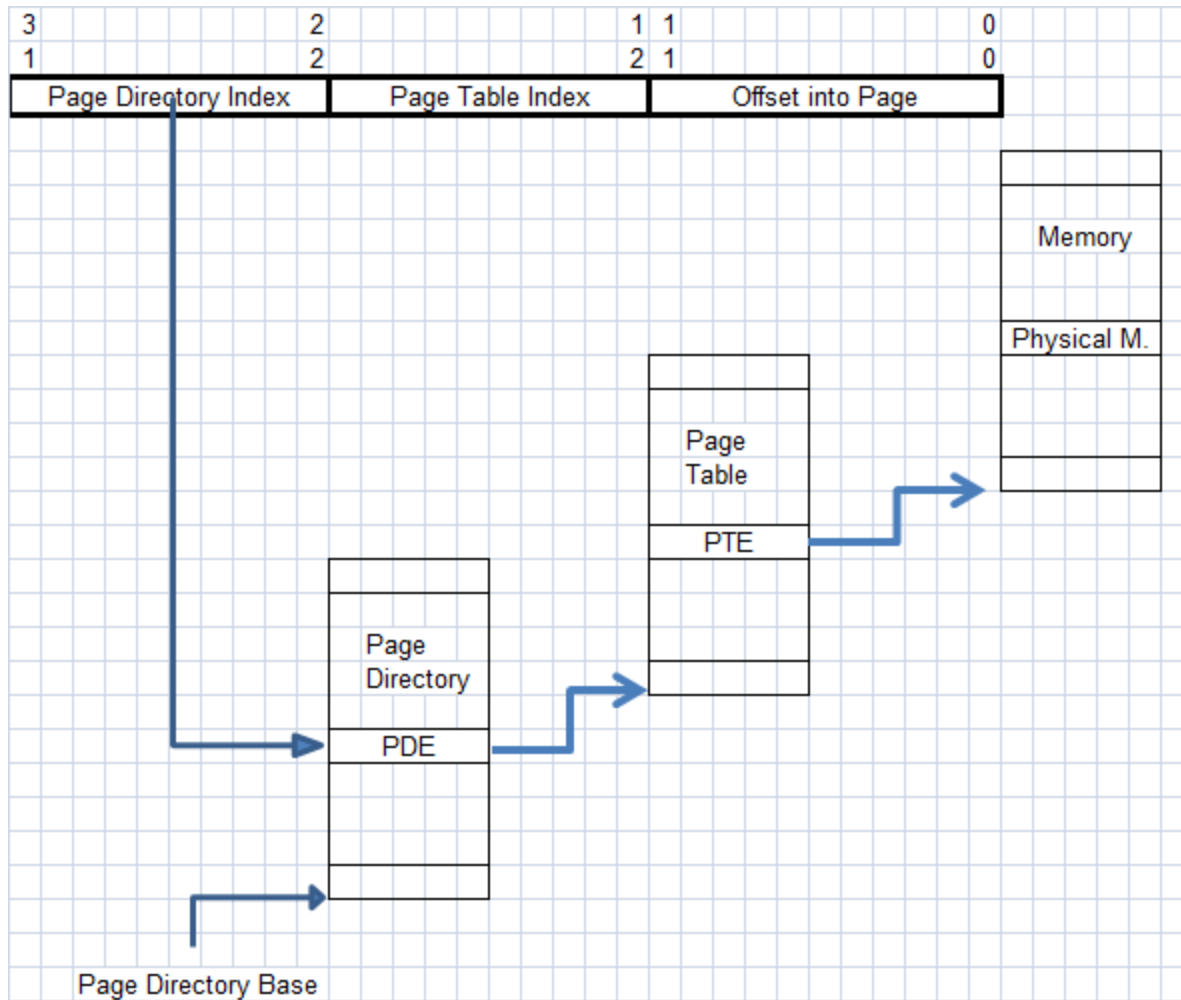


PDE (2nd Level):



Page Walk

Page walk starts with the pointer to Page Directory Base and using the Page Directory index, this is for the 2nd level page table where PDE is determined. Using the PDE and page table index, the PTE is determined which points to the page in physical memory. Using the offset to the page, the actual line is accessed.



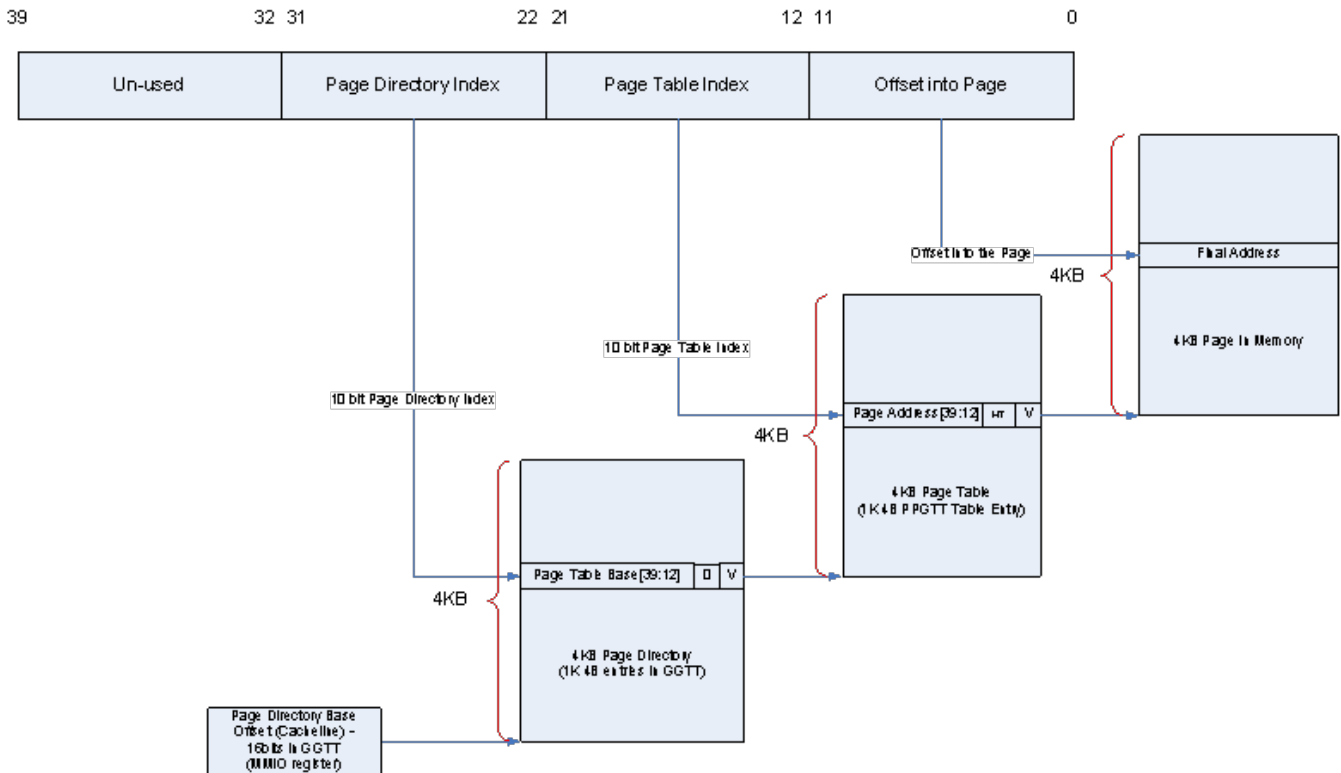


Two-Level Per-Process Virtual Memory

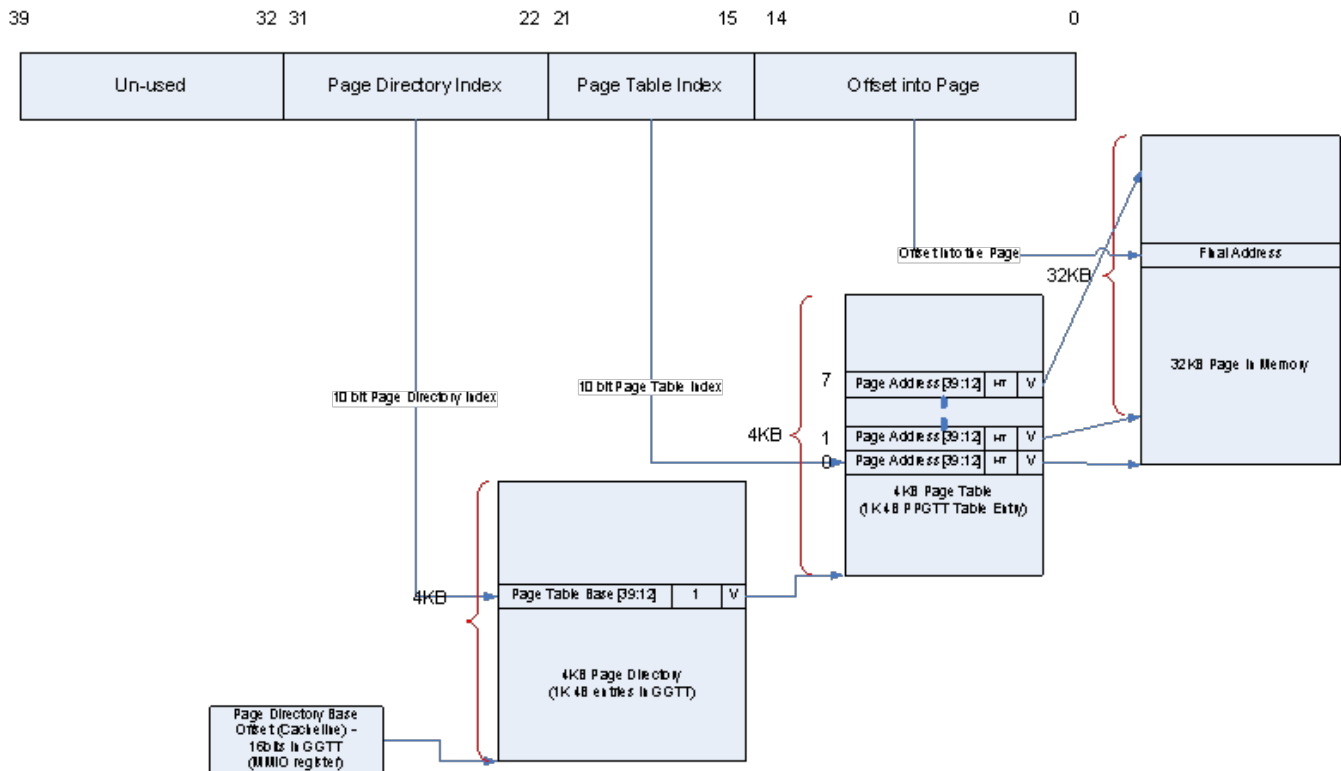
The motivation for the 2-level scheme is simple – it allows for the lookup table (the collection of page tables) to exist in discontinuous memory, making allocation of memory for these structures less problematic for the OS. The directory and each page table fit within a single 4K page of memory that can be located anywhere in physical memory.

If a PPGTT is enabled, all rendering operations (including blit commands) target per-process virtual memory. This means *all* commands *except* the Memory Interface Commands (MI_*). Certain Memory Interface Commands have a specifier to choose global virtual memory (mapped via the GGTT) instead of per-process memory. Global Virtual Memory can be thought of as "privileged" memory in this case. Commands that elect to access privileged memory must have sufficient access rights to do so. Commands executing directly from a ring buffer or from a "secure" batch buffer (see the MI_BATCH_BUFFER_START command in Memory Interface Commands) have these access rights; other commands do not and are not permitted to access global virtual memory. See the Memory Interface Commands chapters for details on command access to privileged memory.

The PPGTT is disabled by resetting the **Per-Process GTT Enable** bit.



In case of a big page client, the following access path is used:



Note: The starting address of a 32KB page needs to be natively aligned to a 32KB boundary in memory. Hardware uses certain bits (15 and up) to check for the TLB lookups.

Note: [HSW]: 32KB big page implementation has been expanded for all surfaces. 8 consecutive PTEs within the 32KB Big page need to be programmed to point to their respective 4KB pages.

If a PDE is marked 32KB, then:

1. The entire PT pointed to by that PDE must use 32KB pages.
2. PTEs for each 4KB need to be programmed to their respective pages. There is no difference in programming PTEs for PDE that is marked for 32KB pages vs 4KB pages.
3. The physical addresses of each 32KB page have to be natively 32KB aligned.



PPGTT Directory Entries (PDEs)

Directory Entry: 1 DWord per 4KB page table (4MB Graphics Memory). Page directories must be located entirely within the GGTT (the table itself.) Directory entries should be updated only for non-active contexts. If a directory entry update is done for a running context, it is unknown when that update will take effect since the device caches directory entries. Directory entries can only be modified using GTTADDR (see *Memory Interface Commands for Rendering Engine*).

31	12	11	4	3	2	1	0
Physical Page Address 31:12		Physical Page Address 39:32		Reserved		Size ("0":4KB, "1":32KB)	Valid

Bits	Description
31:12	Physical Page Address 31:12: If the Valid bit is set, This field provides the page number of the physical memory page backing the corresponding Graphics Memory page.
11:4	Physical Page Address Extension: This field specifies bits 39:32 of the directory entry.
3:2	Reserved: MBZ
1	Page Size: Two page sizes are supported through PDE. 0: 4KB pages 1: 32KB pages
0	Valid PDE: This field indicates whether this directory entry is valid. 1: Valid 0: Invalid. An access through an invalid PDE will result in a page fault.



PPGTT Table Entries (PTEs)

This topic is currently under development.

PPGTT Table Entries (PTEs)

Page Table Entry: 1 DWord per 4KB Graphics Memory page. Page Tables must be located in main memory (not snooped). They can be updated directly in memory if proper precautions are taken, or from the command stream by using the MI_UPDATE_GTT command (see *Memory Interface Commands for Rendering Engine*).

31:12	11	10:4	3:1	0
Physical Page Address 31:12	Cacheability Control[3]	Physical Page Address 38:32	Cacheability Control[2:0]	Valid

Bits	Description
31:12	Physical Page Address 31:12: If the Valid bit is set, This field provides the page number of the physical memory page backing the corresponding Graphics Memory page.
11	Cacheability Control[3] With the addition of new modes for LLC and addition of eDRAM, the use of the cacheability controls via GTT has been updated – see table that follows
10:4	Physical Start Address Extension: This field specified Bits 38:32 of the page table entry. This field must be zero for 32 bit addresses.
3:1	Cacheability Control[2:0] With the addition of new modes for LLC and addition of eDRAM, the use of the cacheability controls via GTT has been updated – see below table
0	Valid PTE: This field indicates whether the mapping of the corresponding Graphics Memory page is valid. 1: Valid 0: Invalid. An access (other than a CPU Read) through an invalid PTE will result in Page Table Error (Invalid PTE).

With the HSWpage table format, bit[39] of physical address is re-purposed to be part of cacheability controls given client systems are only required to support 39-bit physical addressing. Also GFDT bit controls are removed and re-assigned to be part of the cacheability controls.

4-bit of cacheability controls provides 16 individual groups of surface classifications which could be programmed to behave differently. It is up to software to decide what the group specifications would be (see following for options) and which pages map to which groups.

Cacheability Control Options:



1. **Memory Type: UC/WB/WT**

Memory Type is to decide whether a page needs to be cached in LLC and eLLC/eDRAM (if GT4). If the page is tagged as UC, any memory access will not be allocated to LLC/eDRAM caches and it will invalidate if same line already exists in these caches on the way to memory.

If WB, the allocation will be done in the LLC/eDRAM cache w/o the need to go to memory.

WT is a new mode of operation added and can be used for both LLC and eLLC. Reads are no different than WB mode however writes are both allocated into LLC/eDRAM caches as well as the final result is updated in memory. The cache allocation in this case is stated as non-modified (given memory already has most up-to-date copy). It is meant to be used primarily for caching display surfaces.

2. **Cache Type: LLC vs eDRAM**

Only valid if the memory type is defined as WT or WB. It allows the page to be allocated in LLC and/or eDRAM.

LLC: Allocate the line in LLC

eDRAM: Allocate the line in eDRAM

Allows GFX driver to choose which cache to be used for different surfaces.

3. **Age Allocation:**

LLC and eDRAM use(s) quad-aging to assist the replacement algorithm. If a particular allocation is done at youngest age ("3") it tends to stay longer in the cache as compared to older age allocations ("2", "1", or "0"). This option is given to driver to be able to decide which surfaces are more likely to generate HITS, hence need to be replaced least often in caches.

Programming:

Below is a sample programming for a given cache control group which is going to be used by the driver:

Cacheability Control

[3:0] Description

0000: UC (LLC/eLLC) – allocation age is don't care

0111: WT in LLC/eLLC – Aged "3"

0110: WT in LLC/eLLC – Aged "0"

0101: WT in eLLC (UC in LLC) – Aged "3"

0100: WT in eLLC (UC in LLC) – Aged "0"

1000: WB in LLC/eLLC – Aged "3"



1001: WB in LLC/eLLC – Aged "2"

1010: WB in LLC/eLLC – Aged "1"

1011: WB in LLC/eLLC – Aged "0"

1100: WB in eLLC (UC in LLC) – Aged "3"

1101: WB in eLLC (UC in LLC) – Aged "2"

1110: WB in eLLC (UC in LLC) – Aged "1"

1111: WB in eLLC (UC in LLC) – Aged "0"

0010: WB in LLC Only – Aged "3"

0011: WB in LLC Only – Aged "0"

Note: LLC Only has been added to support MSAA color compression scheme where higher order plans are not getting victimized to eDRAM where there has been to have poor spatial localization. Such case would lead to poor superline utilization in terms of number of 64Bs valid.

Note: WT mode is added for display surfaces.

It is highly recommended to use eDRAM only when present besides the exceptions above. Also GFX driver should use Page Table controls as the primary means of memory type control and use `memory_object_control_state` for override when needed.

PPGTT vs GGTT Limitation

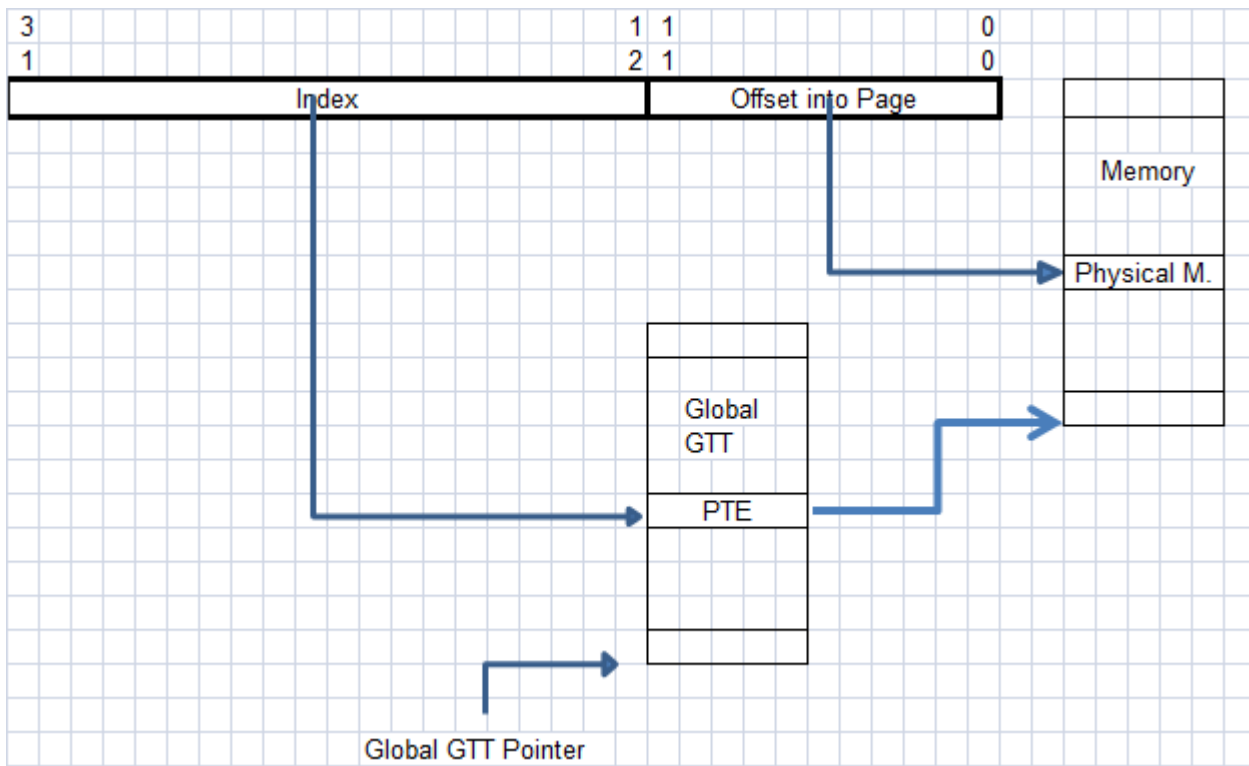


Global GTT

Global Virtual Memory is the default target memory if a PPGTT is not enabled. If a PPGTT is also present, the method to choose which is targeted by memory and rendering operations varies by product. As Display and Aperture path will use Global GTT even if GT is mapped via per-process GTT.

The PTE format for global GTT is identical to the format that is used in PPGTT, the difference is that the walk is always single level. Driver is assumed to program the GSM (GTT Stolen Memory) with the page pointers and all hardware does is to access to corresponding GSM location (Offset is acquired from virtual address indexing) and get a pointer to physical page.

For pre-Gen8, the GSM was limited to 2MB allowing indexing for 2GB of physical memory.

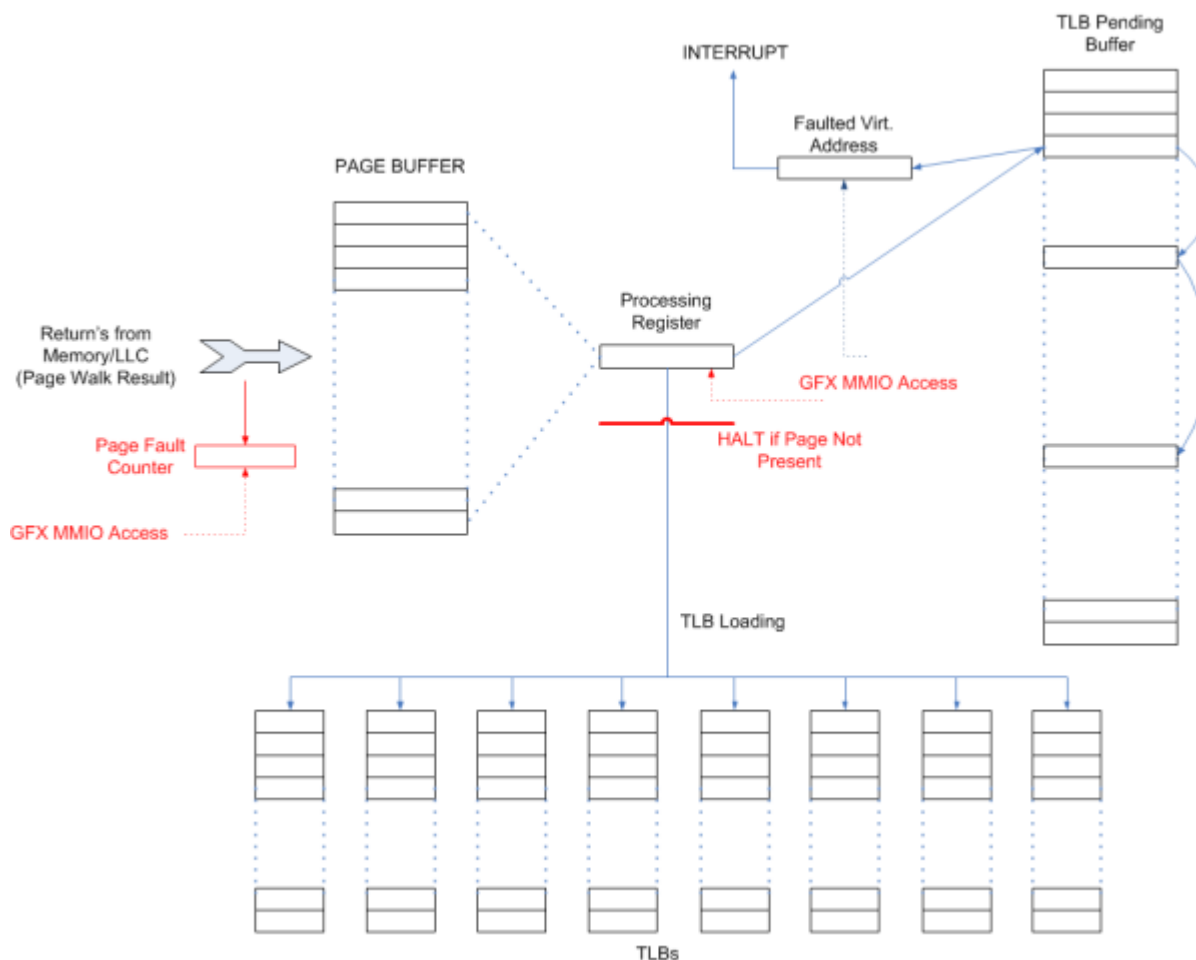


Faulting

This topic covers the following subtopics:

- Errors

Page Faulting Support



[Ed. Note: where is the

Page Faulting via fault halt is default disabled in GFX hardware and needs to be enabled via a global bit at x4090[14] (when set to "1").

Red pieces on this block diagram are the additions to Page Walker hardware for DevHSW (Gen7.5).

- Counter block is visible to s/w via GFX MMIO: Counts the number of page faulted responses from system memory and it increments the value hold in the register. It will decrement as faults are being serviced.
- HALT on Page Fault: Once the a page faulted entry moves from the "PAGE BUFFER" to processing register, the hardware process will halt (interrupting s/w) and expect s/w to come and update the physical address and clear the fault bit.



- GFX MMIO Access to faulted address and processing stage: Engine Fault register is picked to present the virtual address to s/w that was faulted. Bit[0] sets to "1" indicating a fault has been captured. S/W will override the contents this register with the new physical address and clear the bit[0] to "0" indicating to h/w that fault has been repaired.
- There is one register defined for all parallel engines in GFX when page fault repair mode is enabled, i.e. write of 1 on 0x4090 bit [14]:
 - GFX Engine Fault register: x4094

S/W sequence (w/o counter usage)

- Hardware will detect the fault & Halt processing misses, signal an interrupt to s/w
- Software will override the contents of the Engine Fault register with the physical address (i.e. PTE format is used), however it will clear bit [0].

S/W sequence (with counter usage)

- Hardware will detect the fault & Halt processing misses, signal an interrupt to s/w.
- Software will mask the interrupts and read the contents of the fault-count.
- Software will override the contents of the Engine Fault register with the physical address (i.e. PTE format is used), however it will clear bit [0].
- And poll for the same register to get bit[0] changed to "1".
- Software will keep fixing the faults as long as the count >1. Just before fixing the last fault, s/w needs to re-enable interrupts.
- Software can update its count by re-reading the fault counter register while processing page faults
 - Page Fault count is reported to Main GFX Arbiter Error register: 0x40A0 bits [15:10].

Note: Gen7.5/DevHSW cannot handle command buffers getting page faults.

Gen7.5/DevHSW page faulting is only applicable to Render Engine.

Gen7.5/DevHSW page faulting on graphics interface has limitation when it comes to interacting with global flows around power management events. There are known cases in h/w where geyserville transitions with page faulting may lead to h/w lock-ups which can only be cleared with system reset. All these issues are addressed in future generations of GFX paging functions.

Gen7.5/DevHSW can't handle faults on GGTT.



Memory Types and Cache Interface

This section has additional information on the types of memory which are accessible via the various GT mechanisms. It includes discussion on how the various paging models are used and accessed. See the Graphics Translation Tables for more detailed discussions on paging models.

This section also includes descriptions of how different surface types (MOCS) can be cached in the L3 and the different behaviors which can be enabled.



Memory Object Control State (MOCS)

The memory object control state defines the behavior of memory accesses beyond the graphics core, graphics data types that allow selective flushing of data from outer caches, and controlling cacheability in the outer caches.

This control uses several mechanisms. Control state for all memory accesses can be defined page by page in the GTT entries. Memory objects that are defined by state per surface generally have additional memory object control state in the state structure that defines the other surface attributes. Memory objects without state defining them have memory object state control defined per class in the STATE_BASE_ADDRESS command, with class divisions the same as the base addresses. Finally, some memory objects only have the GTT entry mechanism for defining this control. The table below enumerates the memory objects and the location of the control state for each:

Memory Object	Location of Control State
surfaces defined by SURFACE_STATE: sampling engine surfaces, render targets, media surfaces, pull constant buffers, streamed vertex buffers	SURFACE_STATE
depth, stencil, and hierarchical depth buffers	corresponding state command that defined the buffer attributes
stateless buffers accessed by data port	STATE_BASE_ADDRESS
indirect state objects	STATE_BASE_ADDRESS
kernel instructions	STATE_BASE_ADDRESS
push constant buffers	3DSTATE_CONSTANT_(VS GS PS)
index buffers	3DSTATE_INDEX_BUFFER
vertex buffers	3DSTATE_VERTEX_BUFFERS
indirect media object	STATE_BASE_ADDRESS
generic state prefetch	GTT control only
ring/batch buffers	GTT control only
context save buffers	GTT control only
store DWord	GTT control only



MOCS Registers

These registers provide the detailed format of the MOCS table entries, that need to be programmed to define each surface state.

MEMORY_OBJECT_CONTROL_STATE



Graphics Cache and Memory Interface

This document is only meant to highlight the new hooks added to DevHSW for memory interface and large caches (LLC/eDRAM(eLLC)).

Note that main memory arbiter has a known inefficiency where in case of streaming reads from execution units to the same address which is marked as uncacheable in L3\$, memory arbiter parks at the read stream where writes from other resources are not permitted to make progress towards memory. It is required for s/w to avoid heavy read streaming while keeping the line UC in L3\$.

The case would lead to live-locks.



Features Added to Memory Interface for DevHSW

Memory Type: WB/UC/WT

Memory Type Enhancements are new for DevHSW.

- **UnCacheable (UC).** A memory surface/page defined as UC translates to any memory access within the surface/page is not going to be cached in LLC and eLLC/eDRAM (eLLC). Once a page is tagged as UC, the memory access is not allocated to caches and it will invalidate if the same line already exists in the caches on the way towards memory. Note that UC capability is not new for DevHSW.
- **Writeback (WB).** The allocation is done in the cache without updating memory. This is the behavior when the cycle is tagged as cacheable in LLC. A memory surface/page defined as WB is cached in LLC/eDRAM (eLLC) (depending on the destination cache selection). Note that WB capability is not new for HSW.
- **WriteThrough (WT).** This is a new mode of memory type added to GFX and can be used for both LLC and eDRAM (eLLC). Reads are no different than WB mode however writes are both allocated into caches as well as the final memory location is updated in DRAM. The cache allocation in this case is stated as non-modified (given memory already has most up-to-date copy). It is meant to be used primarily for displayable surfaces. Note that WT capability is new for DevHSW.
 - Write Through memory type is only available in Crystwell SKU for eDRAM enabled systems with REV ID of "8h"

Cache type: L3/LLC/eDRAM(eLLC)

Only valid if the memory type is defined as WT or WB. It allows the page to be allocated in L3, LLC, and/or eLLC.

- L3: Allocate the line in L3.
- LLC: Allocate the line in LLC.
- eDRAM(eLLC): Allocate the line in eLLC.

Allows GFX driver to choose which cache to be used for different surfaces.

Cache Replacement Age Allocation

Both LLC and eDRAM(eLLC) use LLC uses quad-aging to assist the replacement algorithm. If a particular cache allocation is done at youngest age ("3") it tends to stay longer in the cache as compared to older age allocations ("2", "1", or "0"). This option is given to driver to be able to decide which surfaces are more likely to generate HITS, hence need to be replaced least often in caches.



Push Writes

A new write semantics is defined for DevHSW where the strong ordering requirements for stores are reduced for the GFX-only allocated memory surfaces. It is up to hardware to direct the GFX-only streams (hardwired in DevHSW) once the feature is enabled.

The idea to provide a shorter loop in LLC for stores hence increase the throughput for writes without the need to increase ring frequency.

MMIO OFFSET: x9008

[8]	R/W	0b	PushWrite Enable: This is the enable control for the more optimized write flow which is added to DevHSW. This mode can only be used with DevHSW when attached to DevHSW:GT3e Uncore . 0: Push Write is disabled (Default). 1: Push Writes are enabled.
-----	-----	----	---

Poorman's LIP25

LIP25 prevents the trash of render to texture allocations in cache while limiting the allocations due to RO surfaces. This is achieved via "class of service" where render traffic is assigned to a higher class compared to L3 (texture) reads that are assigned to lower class. Such scheme leads to reduction in number of ways that can allocate textures where render accesses can allocate the entire cache.

The mechanism needs to be enabled via following configuration and only to be used when LLC is present and eDRAM is not present.

MMIO OFFSET: x9030

[23]	R/W	0b	Poorman's LIP25: The control to change CLOS between various streams of traffic. Once this bit is set, all L3 reads are directed to CLOS of "11". 0: LIP25 disabled (Default) 1: LIP25 is enabled
------	-----	----	--



Feature Set per DevHSW SKU

	Push Write	NT Bit	Non-inclusive	Write-through (LLC)	Write-through (eDRAM/eLLC)
HSW Client	No	Yes	No	No	No
HSW ULT	No	Yes	No	No	No
DevHSW:GT3e	Yes	Yes	Yes	Yes	Yes



Enabling DevHSW Memory Interface Cache

There are two main mechanisms to control the caching at DevHSW and these are similar to what was there in previous generations of GFX. However the indexing and definition has been changed slightly.

GFX Translation Table (GTT) Entry

PTE definition has been changed to enable more of an indexing rather than to be direct control of memory interface functions. This is to present the GFX s/w all the options that are available in DevHSW without creating too many combinations.

NEW PTE FORMAT:

31:12	11	10:04	3:1	0
Physical Page Address 31:12	Cacheability Control[3]	Physical Page Address 38:32	Cacheability Control[2:0]	Valid

The cacheability control[3:0] indexing is hardwired to the following set in the h/w. It is up to GFX driver to select which surface/page is mapped to one of the following settings in the hardware.

Cacheability Control [3:0]	Description
0000	UC (LLC/eLLC) – allocation age is don't care
0111	WT in LLC/eLLC – Aged "3"
0110	WT in LLC/eLLC – Aged "0"
0101	WT in eLLC (UC in LLC) – Aged "3"
0100	WT in eLLC (UC in LLC) – Aged "0"
1000	WB in LLC/eLLC – Aged "3"
1001	WB in LLC/eLLC – Aged "2"
1010	WB in LLC/eLLC – Aged "1"
1011	WB in LLC/eLLC – Aged "0"
1100	WB in eLLC (UC in LLC) – Aged "3"
1101	WB in eLLC (UC in LLC) – Aged "2"
1110	WB in eLLC (UC in LLC) – Aged "1"
1111	WB in eLLC (UC in LLC) – Aged "0"
0010	WB in LLC Only – Aged "3"
0011	WB in LLC Only – Aged "0"

Note: When eDRAM is present, GT traffic primarily cached in eLLC.

Memory Object Control State (MOCS)

Memory object control state provides a lighter control over the memory interface caches compared to PTE settings. However MOCS is the only way to manage L3\$ caching for a given surface.

Bits	Description
2:1	<p>LLC/eLLC Cacheability Control (LLCCC)</p> <p>This is the field used in GT interface block to determine what type of access needs to be generated to uncore. For the cases where the LLCCC is set, cacheable transaction are generated to enable LLC/eLLC usage for particular stream.</p> <p>00b: Use PTE values</p> <p>01b: UC – uncacheable</p> <p>10b: LLC/eLLC WB cacheable</p> <p>11b: eLLC WB cacheable (UC in LLC)</p> <p>Exceptions:</p> <ul style="list-style-type: none"> • If a surface is cacheable in L3, evictions from L3 would ignore LLC/eLLC cacheability control (LLCCC) field and assume the same access is cacheable in LLC/eLLC. • LLCCC field is a HINT, however there are edge cases where it may not be followed. The resulting line is not guaranteed to end up in the target cache.
0	<p>L3 Cacheability Control (L3CC)</p> <p>This field is used to control the L3 cacheability (allocation) of the stream.</p> <p>0: Not cacheable in L3.</p> <p>1: Cacheable in L3.</p> <p>Note: Even if the surface is not cacheable in L3, it is still kept coherent with L3 content.</p>

Note: There is nothing wrong with defining a surface to be L3 cacheable but not LLC/eLLC cacheable, however h/w can not guarantee the exact behavior. In the sense, the L3 stored modified lines (R/W surface) when evicted can land in LLC/eLLC. The same restriction does not apply to RO surfaces, where only L3 allocation can be achieved without allocating in LLC/eLLC.

Note: *When eDRAM is present, GT traffic primarily cached in eLLC.*



How are MOCS and PTE combined in DevHSW?

DevHSW restores the MOCS override of the PTE functionality. If MOCS is set anything but "00: Use PTE values", it will override the cacheability controls coming from the PTE. All cacheability options are listed below in the cacheability control[3:0] table.

The algorithm works as follows:

MOCS: "01" =>

Cacheability Control[3:0]= 0000	UC (LLC/eLLC) – allocation age is don't care
---------------------------------	--

MOCS: "10" =>

Cacheability Control[3:0]=1000	WB in LLC/eLLC – Aged "3"
--------------------------------	---------------------------

MOCS: "11" =>

Cacheability Control[3:0]= 1100	WB in eLLC (UC in LLC) – Aged "3"
---------------------------------	-----------------------------------

MOCS: "00" =>

The following table based on PTE cacheability control field is used.

Cacheability Control [3:0]	Description
0000	UC (LLC/eLLC) – allocation age is don't care
0111	WT in LLC/eLLC – Aged "3"
0110	WT in LLC/eLLC – Aged "0"
0101	WT in eLLC (UC in LLC) – Aged "3"
0100	WT in eLLC (UC in LLC) – Aged "0"
1000	WB in LLC/eLLC – Aged "3"
1001	WB in LLC/eLLC – Aged "2"
1010	WB in LLC/eLLC – Aged "1"
1011	WB in LLC/eLLC – Aged "0"
1100	WB in eLLC (UC in LLC) – Aged "3"
1101	WB in eLLC (UC in LLC) – Aged "2"
1110	WB in eLLC (UC in LLC) – Aged "1"
1111	WB in eLLC (UC in LLC) – Aged "0"
0010	WB in LLC Only – Aged "3"
0011	WB in LLC Only – Aged "0"

eDRAM (eLLC) Enabling Requirement on GFX Driver

For the SKUs where eDRAM(eLLC) is enabled (see following bit for determination), it is required that the RING HASH to be modified to comply with superpage concept. All driver has to do is set the following bits appropriately when eDRAM (eLLC) SKU is detected.

To determine eDRAM(eLLC) presence:

Table: MMIO OFFSET: x120010

[0]	R/W	0b	EDRAM (eLLC) Present. When eDRAM (eLLC) is present on the system, this bit is set by hardware to indicate the presence. It is up to GFX driver to read this register once loaded and arrange caching policies accordingly.
-----	-----	----	---

Once eDRAM (eLLC) is detected, driver needs to program the following register:

Table: MMIO OFFSET: x9008

[21:16]	R/W	0b	<p>IDI HASH MASK. When a corresponding bit is set, the address line going into HASH for CBO ID (Ring) calculation is forced to "logic0"</p> <p>21 => Address Bit[11] 20 => Address Bit[10] 19 => Address Bit[9] 18 => Address Bit[8] 17 => Address Bit[7] 16 => Address Bit[6]</p> <p>Note: It is required for GFX Driver to set [19:16] to "1" when eDRAM configuration is present.</p>
---------	-----	----	---



Class of Service Usage for GFX

Cache Quality/Class of Service (alternatively shortened as QOS/COS/CLOS) is maintained in the uncore by allowing allocations only into the cache ways specified for the given level of service (0, 1, 2, 3). The ways allocated to a given class of service are stored as 16-bit masks where bit being set indicates that the way is allowed for allocations. Through this method the caches are effectively partitioned between the various service levels for IA and graphics. However COS masks do not prevent hits in other ways of the cache, it provides an allocation MASK for the incoming accesses.

IA and Graphics each have separate way-masks for the LLC and eLLC, for a total of 16 masks. All IA requests will first allocate into the LLC; graphics requests will be allocated into either the LLC or eLLC, depending on traffic type and eLLC state. It is not possible to determine the original source of LLC victims, so they will always be treated as IA level 0 requests for purposes of COS in the eLLC.

Class of Service Controls

MMIO OFFSET: x900c

[22:21]	R/W	00b	<p>Class of Service for Cache Allocations:</p> <p>4 possible service classes are allowed for GPU. The detailed assignments for each service class is listed separately.</p>
---------	-----	-----	--

HSW/CRW Settings

LLC Masks

	Class0		Class1		Class2		Class3	
	GT	IA	GT	IA	GT	IA	GT	IA
16 ways	0xfc00	0xffff	0xffff	0x000f	0xff0	0x03ff	0xf000	0x0fff
12 ways	0xf80	0x0fff	0x0fff	0x0007	0xff8	0x007f	0xe00	0x01ff
8 ways	0x0e0	0x00ff	0x00ff	0x000f	0x00f	0x001f	0x00c0	0x003f
4 ways	0x00c	0x000f	0x000f	0x0007	0x000	0x0003	0x0008	0x000f

eLLC Masks

	Class0		Class1		Class2		Class3	
	GT	IA	GT	IA	GT	IA	GT	IA
16 ways	0xfc00	0xffff	0xffff	0xffff	0xff0	0x03ff	0xf000	0x0fff
11 ways	0x780	0x07ff	0x07ff	0x07ff	0x07f8	0x007f	0x0300	0x04ff



	Class0		Class1		Class2		Class3	
	GT	IA	GT	IA	GT	IA	GT	IA
8 ways	0x00e0	0x00ff	0x00ff	0x00ff	0x00fc	0x001f	0x00c0	0x003f
6 ways	0x000c	0x003f	0x003f	0x003f	0x0038	0x0007	0x0018	0x002f

HSW/CRW B0 Settings:

Few issues in CRW-B0 that will affect the usage of the COS masks. The current default masks are included below.

There is also an additional issue in the B0 that will lead to more Non-inclusive graphics accesses to be filled into the LLC more often if there are multiple transactions that miss on the same 1KB sector in flight at the same time. This case is not uncommon, but the initial SPL victim selection should follow the eLLC masks, but may end up with more LLC eviction traffic than expected.

LLC Masks

	Class0		Class1		Class2		Class3	
	GT	IA	GT	IA	GT	IA	GT	IA
16 ways	0xfc00	0xffff	0xffff	0x000f	0xff0	0x03ff	0x4000	0x0fff
12 ways	0xf80	0x0fff	0x0fff	0x0007	0xff8	0x007f	0x0400	0x01ff
8 ways	0x00e0	0x00ff	0x00ff	0x0003	0x00fc	0x001f	0x0040	0x003f
4 ways	0x000c	0x000f	0x000f	0x0007	0x000e	0x0003	0x0004	0x000f

eLLC Masks

	Class0		Class1		Class2		Class3	
	GT	IA	GT	IA	GT	IA	GT	IA
16 ways	0xffff	0xffff	0xffff	0x000f	0xff0	0x03ff	0x4000	0x0fff
11 ways	0xf80	0x0fff	0x0fff	0x0007	0xff8	0x007f	0x0400	0x01ff
8 ways	0x00e0	0x00ff	0x00ff	0x0003	0x00fc	0x001f	0x0040	0x003f
6 ways	0x000c	0x000f	0x000f	0x0007	0x000e	0x0003	0x0004	0x000f



Driver Hints

- LLC is not big enough to cache all surfaces. It is recommended for GFX driver to have an algorithm to detect active surface size usage that is getting cached in LLC. If active surface exceeds LLC limits, the surfaces that benefit the least should not be cached in LLC.

Over caching in LLC leads to performance degradation rather than performance boost due to amount of dirty data that gets collected and victimization gets ahead of DRAM b/w, reducing utilization. GFX Architecture team observed this phenomenon on many workloads where surfaces cached are exceeding the cache size many orders of magnitude.

- When eDRAM is present, almost all GFX surfaces should be moved to eDRAM with the exception of a few where there is weak cacheline locality within the 1KB linear memory. GFX architecture recommendation is to keep high order MSAA related surfaces out of eDRAM and still keeping them in LLC where weak line utilization within 1KB linear memory is not an issue.
- Another exception is where we have heavy concurrent use of b/w between surfaces and one surface is typically small enough to fit into LLC.

eDRAM is large enough (128MB) to contain almost all GFX surfaces, GFX driver must enable caching on all surfaces that have any reuse possibility.

- Some DevHSW SKUs (DevHSW:GT3e) offer Write-Through (WT) memory type support both in LLC and eDRAM. This is to provide a boost for displayable surfaces which have reuse possibilities (i.e. Blend...). GFX driver must tag displayable surfaces as WT cacheable (rather than marking them uncacheable as it was done in previous generations of GFX) when the DevHSW:GT3e SKU is used.
 - *Write Through memory type is only available in Crystwell SKU with eDRAM enabled systems with REV ID of "8h"*
- Cache Age assignments are a new feature for GFX and provide a hook for control over cache replacements over a temporal distance. For the lines that will benefit most from caching and their reuse temporal distance is high (e.g. dynamic textures), choosing a younger age provides a better chance of not getting victimized. Meanwhile surfaces that have a rapid reuse, but may tend to cause trash among less frequently reused surfaces should be cached at an older age. Or surfaces that are reused across a longer life (e.g. static textures) should be cached at an older age to prevent victimization during the intra-frame reuse.

As part of the tuning for arbitration, x4090[6] needs to be set to 1 to enable better alignment for memory accesses. This setting is applicable to Gen7.5/DevHSW only.

- As part of the memory interface programming another option is to re-allocate TLBs between different streams of GFX. The GFX TLBs are organized as assigned resources for dedicated ports which could be re-programmed based on the context that is being executed. This is especially critical for the L3 backed clients which are seeing one large TLB. The default programming favors 3D workloads (384 entry for Textures, 64 entry for data port, 64 entry for rest), however for GPGPU recommended setting is (64 for textures, 384 for dataport, 64 for rest). With this new setting there are certain GPGPU workloads which benefit significantly. Driver needs to make this setting part of the context that is submitted.



Common Surface Formats

This section documents surfaces and how they are stored in memory, including 3D and video surfaces, including the details of compressed texture formats. Also covered are the surface layouts based on tiling mode and surface type.



Non-Video Surface Formats

This section describes the lowest-level organization of a surfaces containing discrete "pixel" oriented data (e.g., discrete pixel (RGB,YUV) colors, subsampled video data, 3D depth/stencil buffer pixel formats, bump map values etc. Many of these pixel formats are common to the various pixel-oriented memory object types.



Surface Format Naming

Unless indicated otherwise, all pixels are **stored** in "**little endian**" byte order. i.e., pixel bits 7:0 are stored in byte n , pixel bits 15:8 are stored in byte $n+1$, and so on. The format labels include color components in little endian order (e.g., R8G8B8A8 format is physically stored as R, G, B, A).

The name of most of the surface formats specifies its format. Channels are listed in little endian order (LSB channel on the left, MSB channel on the right), with the channel format specified following the channels with that format. For example, R5G5_SNORM_B6_UNORM contains, from LSB to MSB, 5 bits of red in SNORM format, 5 bits of green in SNORM format, and 6 bits of blue in UNORM format.



Intensity Formats

All surface formats containing "I" include an intensity value. When used as a source surface for the sampling engine, the intensity value is replicated to all four channels (R,G,B,A) before being filtered. Intensity surfaces are not supported as destinations.



Luminance Formats

All surface formats containing "L" include a luminance value. When used as a source surface for the sampling engine, the luminance value is replicated to the three color channels (R,G,B) before being filtered. The alpha channel is provided either from another field or receives a default value. Luminance surfaces are not supported as destinations.



R1_UNORM (same as R1_UINT) and MONO8

When used as a texel format, the R1_UNORM format contains 8 1-bit Intensity (I) values that are replicated to all color channels. Note that T0 of byte 0 of a R1_UNORM-formatted texture corresponds to Texel[0,0]. This is different from the format used for monochrome sources in the BLT engine.

7	6	5	4	3	2	1	0
T7	T6	T5	T4	T3	T2	T1	T0

Bit	Description
T0	Texel 0 On texture reads, this (unsigned) 1-bit value is replicated to all color channels. Format: U1
...	...
T7	Texel 7 On texture reads, this (unsigned) 1-bit value is replicated to all color channels. Format: U1

MONO8 format is identical to R1_UNORM but has different semantics for filtering. MONO8 is the only supported format for the MAPFILTER_MONO filter. See the *Sampling Engine* chapter.

Palette Formats

Palette formats are supported by the sampling engine. These formats include an index into the palette (Px) that selects the actual channel values from the palette, which is loaded via the 3DSTATE_SAMPLER_PALETTE_LOAD0 command.

P4A4_UNORM

This surface format contains a 4-bit Alpha value (in the high nibble) and a 4-bit Palette Index value (in the low nibble).

7			4	3			0
Alpha				Palette Index			

Bit	Description
7:4	<p>Alpha</p> <p>Alpha value which will be replicated to both the high and low nibble of an 8-bit value, and then divided by 255 to yield a [0.0,1.0] Alpha value.</p> <p>Format: U4</p>
3:0	<p>Palette Index</p> <p>A 4-bit index which is used to lookup a 24-bit (RGB) value in the texture palette (loaded via 3DSTATE_SAMPLER_PALETTE_LOADx)</p> <p>Format: U4</p>

A4P4_UNORM

This surface format contains a 4-bit Alpha value (in the low nibble) and a 4-bit Color Index value (in the high nibble).

7			4	3			0
Palette Index				Alpha			

Bit	Description
7:4	<p>Palette Index</p> <p>A 4-bit color index which is used to lookup a 24-bit RGB value in the texture palette.</p> <p>Format: U4</p>
3:0	<p>Alpha</p> <p>Alpha value which will be replicated to both the high and low nibble of an 8-bit value, and then divided by 255</p>



Bit	Description
	to yield a [0.0,1.0] alpha value. Format: U4

P8A8_UNORM

This surface format contains an 8-bit Alpha value (in the high byte) and an 8-bit Palette Index value (in the low byte).

15			8	7			0
Alpha				Palette Index			

Bit	Description
15:8	Alpha Alpha value which will be divided by 255 to yield a [0.0,1.0] Alpha value. Format: U8
7:0	Palette Index An 8-bit index which is used to lookup a 24-bit (RGB) value in the texture palette (loaded via 3DSTATE_SAMPLER_PALETTE_LOADx) Format: U8

A8P8_UNORM

This surface format contains an 8-bit Alpha value (in the low byte) and an 8-bit Color Index value (in the high byte).

15			8	7			0
Palette Index				Alpha			

Bit	Description
15:8	Palette Index An 8-bit color index which is used to lookup a 24-bit RGB value in the texture palette. Format: U8
7:0	Alpha Alpha value which will be divided by 255 to yield a [0.0,1.0] alpha value. Format: U8

P8_UNORM

This surface format contains only an 8-bit Color Index value.

Bit	Description
7:0	Palette Index An 8-bit color index which is used to lookup a 32-bit ARGB value in the texture palette. Format: U8

P2_UNORM

This surface format contains only a 2-bit Color Index value.

Bit	Description
1:0	Palette Index A 2-bit color index which is used to lookup a 32-bit ARGB value in the texture palette. Format: U2



Compressed Surface Formats

This section contains information on the internal organization of compressed surface formats.

FXT Texture Formats

There are four different FXT1 compressed texture formats. Each of the formats compress two 4x4 texel blocks into 128 bits. In each compression format, the 32 texels in the two 4x4 blocks are arranged according to the following diagram:

FXT1 Encoded Blocks

t0	t1	t2	t3
t4	t5	t6	t7
t8	t9	t10	t11
t12	t13	t14	t15

t16	t17	t18	t19
t20	t21	t22	t23
t24	t25	t26	t27
t28	t29	t30	t31

B6682-01



Overview of FXT1 Formats

During the compression phase, the encoder selects one of the four formats for each block based on which encoding scheme results in best overall visual quality. The following table lists the four different modes and their encodings:

FXT1 Format Summary

Bit 127	Bit 126	Bit 125	Block Compression Mode	Summary Description
0	0	X	CC_HI	2 R5G5B5 colors supplied. Single LUT with 7 interpolated color values and transparent black
0	1	0	CC_CHROMA	4 R5G5B5 colors used directly as 4-entry LUT.
0	1	1	CC_ALPHA	3 A5R5G5B5 colors supplied. LERP bit selects between 1 LUT with 3 discrete colors + transparent black and 2 LUTs using interpolated values of Color 0,1 (t0-15) and Color 1,2 (t16-31).
1	x	x	CC_MIXED	4 R5G5B5 colors supplied, where Color0,1 LUT is used for t0-t15, and Color2,3 LUT used for t16-31. Alpha bit selects between LUTs with 4 interpolated colors or 3 interpolated colors + transparent black.

FXT1 CC_HI Format

In the CC_HI encoding format, two base 15-bit R5G5B5 colors (Color 0, Color 1) are included in the encoded block. These base colors are then expanded (using high-order bit replication) to 24-bit RGB colors, and used to define an 8-entry lookup table of interpolated color values (the 8th entry is transparent black). The encoded block contains a 3-bit index value per texel that is used to lookup a color from the table.

CC_HI Block Encoding

The following table describes the encoding of the 128-bit (DQWord) CC_HI block format:

FXT CC_HI Block Encoding

Bit	Description
127:126	Mode = '00'b (CC_HI)
125:121	Color 1 Red
120:116	Color 1 Green
115:111	Color 1 Blue
110:106	Color 0 Red
105:101	Color 0 Green
100:96	Color 0 Blue
95:93	Texel 31 Select
...	...
50:48	Texel 16 Select
47:45	Texel 15 Select
...	...
2:0	Texel 0 Select

CC_HI Block Decoding

The two base colors, Color 0 and Color 1 are converted from R5G5B5 to R8G8B8 by replicating the 3 MSBs into the 3 LSBs, as shown in the following table:



FXT CC_HI Decoded Colors

Expanded Color Bit	Expanded Channel Bit	Encoded Block Source Bit
Color 1 [23:19]	Color 1 Red [7:3]	[125:121]
Color 1 [18:16]	Color 1 Red [2:0]	[125:123]
Color 1 [15:11]	Color 1 Green [7:3]	[120:116]
Color 1 [10:08]	Color 1 Green [2:0]	[120:118]
Color 1 [07:03]	Color 1 Blue [7:3]	[115:111]
Color 1 [02:00]	Color 1 Blue [2:0]	[115:113]
Color 0 [23:19]	Color 0 Red [7:3]	[110:106]
Color 0 [18:16]	Color 0 Red [2:0]	[110:108]
Color 0 [15:11]	Color 0 Green [7:3]	[105:101]
Color 0 [10:08]	Color 0 Green [2:0]	[105:103]
Color 0 [07:03]	Color 0 Blue [7:3]	[100:96]
Color 0 [02:00]	Color 0 Blue [2:0]	[100:98]

These two 24-bit colors (Color 0, Color 1) are then used to create a table of seven interpolated colors (with Alpha = 0FFh), along with an eight entry equal to RGBA = 0,0,0,0, as shown in the following table:

FXT CC_HI Interpolated Color Table

Interpolated Color	Color RGB	Alpha
0	Color0.RGB	0FFh
1	$(5 * \text{Color0.RGB} + 1 * \text{Color1.RGB} + 3) / 6$	0FFh
2	$(4 * \text{Color0.RGB} + 2 * \text{Color1.RGB} + 3) / 6$	0FFh
3	$(3 * \text{Color0.RGB} + 3 * \text{Color1.RGB} + 3) / 6$	0FFh
4	$(2 * \text{Color0.RGB} + 4 * \text{Color1.RGB} + 3) / 6$	0FFh
5	$(1 * \text{Color0.RGB} + 5 * \text{Color1.RGB} + 3) / 6$	0FFh
6	Color1.RGB	0FFh
7	RGB = 0,0,0	0

This table is then used as an 8-entry Lookup Table, where each 3-bit Texel n Select field of the encoded CC_HI block is used to index into a 32-bit A8R8G8B8 color from the table completing the decode of the CC_HI block.

FXT1 CC_CHROMA Format

In the CC_CHROMA encoding format, four 15-bit R5B5G5 colors are included in the encoded block. These colors are then expanded (using high-order bit replication) to form a 4-entry table of 24-bit RGB colors. The encoded block contains a 2-bit index value per texel that is used to lookup a 24-bit RGB color from the table. The Alpha component defaults to fully opaque (0FFh).



CC_CHROMA Block Encoding

The following table describes the encoding of the 128-bit (DQWord) CC_CHROMA block format:

FXT CC_CHROMA Block Encoding

Bit	Description
127:125	Mode = '010'b (CC_CHROMA)
124	Unused
123:119	Color 3 Red
118:114	Color 3 Green
113:109	Color 3 Blue
108:104	Color 2 Red
103:99	Color 2 Green
98:94	Color 2 Blue
93:89	Color 1 Red
88:84	Color 1 Green
83:79	Color 1 Blue
78:74	Color 0 Red
73:69	Color 0 Green
68:64	Color 0 Blue
63:62	Texel 31 Select
...	
33:32	Texel 16 Select
31:30	Texel 15 Select
...	
1:0	Texel 0 Select

CC_CHROMA Block Decoding

The four colors (Color 0-3) are converted from R5G5B5 to R8G8B8 by replicating the 3 MSBs into the 3 LSBs, as shown in the following tables:



FXT CC_CHROMA Decoded Colors

Expanded Color Bit	Expanded Channel Bit	Encoded Block Source Bit
Color 3 [23:17]	Color 3 Red [7:3]	[123:119]
Color 3 [18:16]	Color 3 Red [2:0]	[123:121]
Color 3 [15:11]	Color 3 Green [7:3]	[118:114]
Color 3 [10:08]	Color 3 Green [2:0]	[118:116]
Color 3 [07:03]	Color 3 Blue [7:3]	[113:109]
Color 3 [02:00]	Color 3 Blue [2:0]	[113:111]
Color 2 [23:17]	Color 2 Red [7:3]	[108:104]
Color 2 [18:16]	Color 2 Red [2:0]	[108:106]
Color 2 [15:11]	Color 2 Green [7:3]	[103:99]
Color 2 [10:08]	Color 2 Green [2:0]	[103:101]
Color 2 [07:03]	Color 2 Blue [7:3]	[98:94]
Color 2 [02:00]	Color 2 Blue [2:0]	[98:96]
Color 1 [23:17]	Color 1 Red [7:3]	[93:89]
Color 1 [18:16]	Color 1 Red [2:0]	[93:91]
Color 1 [15:11]	Color 1 Green [7:3]	[88:84]
Color 1 [10:08]	Color 1 Green [2:0]	[88:86]
Color 1 [07:03]	Color 1 Blue [7:3]	[83:79]
Color 1 [02:00]	Color 1 Blue [2:0]	[83:81]
Color 0 [23:17]	Color 0 Red [7:3]	[78:74]
Color 0 [18:16]	Color 0 Red [2:0]	[78:76]
Color 0 [15:11]	Color 0 Green [7:3]	[73:69]
Color 0 [10:08]	Color 0 Green [2:0]	[73:71]
Color 0 [07:03]	Color 0 Blue [7:3]	[68:64]
Color 0 [02:00]	Color 0 Blue [2:0]	[68:66]

This table is then used as a 4-entry Lookup Table, where each 2-bit Texel n Select field of the encoded CC_CHROMA block is used to index into a 32-bit A8R8G8B8 color from the table (Alpha defaults to 0FFh) completing the decode of the CC_CHROMA block.

FXT CC_CHROMA Interpolated Color Table

Texel Select	Color ARGB
0	Color0.ARGB
1	Color1.ARGB
2	Color2.ARGB
3	Color3.ARGB

FXT1 CC_MIXED Format

In the CC_MIXED encoding format, four 15-bit R5G5B5 colors are included in the encoded block: Color 0 and Color 1 are used for Texels 0-15, and Color 2 and Color 3 are used for Texels 16-31.

Each pair of colors are then expanded (using high-order bit replication) to form 4-entry tables of 24-bit RGB colors. The encoded block contains a 2-bit index value per texel that is used to lookup a 24-bit RGB color from the table. The Alpha component defaults to fully opaque (0FFh).

CC_MIXED Block Encoding

The following table describes the encoding of the 128-bit (DQWord) CC_MIXED block format:

FXT CC_MIXED Block Encoding

Bit	Description
127	Mode = '1'b (CC_MIXED)
126	Color 3 Green [0]
125	Color 1 Green [0]
124	Alpha [0]
123:119	Color 3 Red
118:114	Color 3 Green
113:109	Color 3 Blue
108:104	Color 2 Red
103:99	Color 2 Green
98:94	Color 2 Blue
93:89	Color 1 Red
88:84	Color 1 Green
83:79	Color 1 Blue
78:74	Color 0 Red
73:69	Color 0 Green
68:64	Color 0 Blue
63:62	Texel 31 Select
...	...
33:32	Texel 16 Select
31:30	Texel 15 Select
...	...
1:0	Texel 0 Select

CC_MIXED Block Decoding

The decode of the CC_MIXED block is modified by Bit 124 (Alpha [0]) of the encoded block.



Alpha[0] = 0 Decoding

When Alpha[0] = 0 the four colors are encoded as 16-bit R5G6B5 values, with the Green LSB defined as per the following table:

FXT CC_MIXED (Alpha[0]=0) Decoded Colors

Encoded Color Bit	Definition
Color 3 Green [0]	Encoded Bit [126]
Color 2 Green [0]	Encoded Bit [33] XOR Encoded Bit [126]
Color 1 Green [0]	Encoded Bit [125]
Color 0 Green [0]	Encoded Bit [1] XOR Encoded Bit [125]

The four colors (Color 0-3) are then converted from R5G5B6 to R8G8B8 by replicating the 3 MSBs into the 3 LSBs, as shown in the following table:

FXT CC_MIXED Decoded Colors (Alpha[0] = 0)

Expanded Color Bit	Expanded Channel Bit	Encoded Block Source Bit
Color 3 [23:17]	Color 3 Red [7:3]	[123:119]
Color 3 [18:16]	Color 3 Red [2:0]	[123:121]
Color 3 [15:11]	Color 3 Green [7:3]	[118:114]
Color 3 [10]	Color 3 Green [2]	[126]
Color 3 [09:08]	Color 3 Green [1:0]	[118:117]
Color 3 [07:03]	Color 3 Blue [7:3]	[113:109]
Color 3 [02:00]	Color 3 Blue [2:0]	[113:111]
Color 2 [23:17]	Color 2 Red [7:3]	[108:104]
Color 2 [18:16]	Color 2 Red [2:0]	[108:106]
Color 2 [15:11]	Color 2 Green [7:3]	[103:99]
Color 2 [10]	Color 2 Green [2]	[33] XOR [126]
Color 2 [09:08]	Color 2 Green [1:0]	[103:100]
Color 2 [07:03]	Color 2 Blue [7:3]	[98:94]
Color 2 [02:00]	Color 2 Blue [2:0]	[98:96]
Color 1 [23:17]	Color 1 Red [7:3]	[93:89]
Color 1 [18:16]	Color 1 Red [2:0]	[93:91]
Color 1 [15:11]	Color 1 Green [7:3]	[88:84]
Color 1 [10]	Color 1 Green [2]	[125]
Color 1 [09:08]	Color 1 Green [1:0]	[88:86]
Color 1 [07:03]	Color 1 Blue [7:3]	[83:79]
Color 1 [02:00]	Color 1 Blue [2:0]	[83:81]
Color 0 [23:17]	Color 0 Red [7:3]	[78:74]
Color 0 [18:16]	Color 0 Red [2:0]	[78:76]

Expanded Color Bit	Expanded Channel Bit	Encoded Block Source Bit
Color 0 [15:11]	Color 0 Green [7:3]	[73:69]
Color 0 [10]	Color 0 Green [2]	[1] XOR [125]
Color 0 [09:08]	Color 0 Green [1:0]	[73:71]
Color 0 [07:03]	Color 0 Blue [7:3]	[68:64]
Color 0 [02:00]	Color 0 Blue [2:0]	[68:66]

The two sets of 24-bit colors (Color 0,1 and Color 2,3) are then used to create two tables of four interpolated colors (with Alpha = 0FFh). The Color0,1 table is used as a lookup table for texel 0-15 indices, and the Color2,3 table used for texels 16-31 indices, as shown in the following figures:

FXT CC_MIXED Interpolated Color Table (Alpha[0]=0, Texels 0-15)

Texel 0-15 Select	Color RGB	Alpha
0	Color0.RGB	0FFh
1	$(2 * \text{Color0.RGB} + \text{Color1.RGB} + 1) / 3$	0FFh
2	$(\text{Color0.RGB} + 2 * \text{Color1.RGB} + 1) / 3$	0FFh
3	Color1.RGB	0FFh

FXT CC_MIXED Interpolated Color Table (Alpha[0]=0, Texels 16-31)

Texel 16-31 Select	Color RGB	Alpha
0	Color2.RGB	0FFh
1	$(2/3) * \text{Color2.RGB} + (1/3) * \text{Color3.RGB}$	0FFh
2	$(1/3) * \text{Color2.RGB} + (2/3) * \text{Color3.RGB}$	0FFh
3	Color3.RGB	0FFh

Alpha[0] = 1 Decoding

When Alpha[0] = 1, Color0 and Color2 are encoded as 15-bit R5G5B5 values. Color1 and Color3 are encoded as RGB565 colors, with the Green LSB obtained as shown in the following table:

FXT CC_MIXED (Alpha[0]=0) Decoded Colors

Encoded Color Bit	Definition
Color 3 Green [0]	Encoded Bit [126]
Color 1 Green [0]	Encoded Bit [125]

All four colors are then expanded to 24-bit R8G8B8 colors by bit replication, as show in the following diagram.

FXT CC_MIXED Decoded Colors (Alpha[0] = 1)

Expanded Color Bit	Expanded Channel Bit	Encoded Block Source Bit
Color 3 [23:17]	Color 3 Red [7:3]	[123:119]
Color 3 [18:16]	Color 3 Red [2:0]	[123:121]



Expanded Color Bit	Expanded Channel Bit	Encoded Block Source Bit
Color 3 [15:11]	Color 3 Green [7:3]	[118:114]
Color 3 [10]	Color 3 Green [2]	[126]
Color 3 [09:08]	Color 3 Green [1:0]	[118:117]
Color 3 [07:03]	Color 3 Blue [7:3]	[113:109]
Color 3 [02:00]	Color 3 Blue [2:0]	[113:111]
Color 2 [23:19]	Color 2 Red [7:3]	[108:104]
Color 2 [18:16]	Color 2 Red [2:0]	[108:106]
Color 2 [15:11]	Color 2 Green [7:3]	[103:99]
Color 2 [10:08]	Color 2 Green [2:0]	[103:101]
Color 2 [07:03]	Color 2 Blue [7:3]	[98:94]
Color 2 [02:00]	Color 2 Blue [2:0]	[98:96]
Color 1 [23:17]	Color 1 Red [7:3]	[93:89]
Color 1 [18:16]	Color 1 Red [2:0]	[93:91]
Color 1 [15:11]	Color 1 Green [7:3]	[88:84]
Color 1 [10]	Color 1 Green [2]	[125]
Color 1 [09:08]	Color 1 Green [1:0]	[88:87]
Color 1 [07:03]	Color 1 Blue [7:3]	[83:79]
Color 1 [02:00]	Color 1 Blue [2:0]	[83:81]
Color 0 [23:19]	Color 0 Red [7:3]	[78:74]
Color 0 [18:16]	Color 0 Red [2:0]	[78:76]
Color 0 [15:11]	Color 0 Green [7:3]	[73:69]
Color 0 [10:08]	Color 0 Green [2:0]	[73:71]
Color 0 [07:03]	Color 0 Blue [7:3]	[68:64]
Color 0 [02:00]	Color 0 Blue [2:0]	[68:66]

The two sets of 24-bit colors (Color 0,1 and Color 2,3) are then used to create two tables of four colors. The Color0,1 table is used as a lookup table for texel 0-15 indices, and the Color2,3 table used for texels 16-31 indices. The color at index 1 is the linear interpolation of the base colors, while the color at index 3 is defined as Black (0,0,0) with Alpha = 0, as shown in the following figures:

FXT CC_MIXED Interpolated Color Table (Alpha[0]=1, Texels 0-15)

Texel 0-15 Select	Color RGB	Alpha
0	Color0.RGB	0FFh
1	(Color0.RGB + Color1.RGB) / 2	0FFh
2	Color1.RGB	0FFh
3	Black (0,0,0)	0



FXT CC_MIXED Interpolated Color Table (Alpha[0]=1, Texels 16-31)

Texel 16-31 Select	Color RGB	Alpha
0	Color2.RGB	0FFh
1	(Color2.RGB + Color3.RGB) / 2	0FFh
2	Color3.RGB	0FFh
3	Black (0,0,0)	0

These tables are then used as a 4-entry Lookup Table, where each 2-bit Texel n Select field of the encoded CC_MIXED block is used to index into the appropriate 32-bit A8R8G8B8 color from the table, completing the decode of the CC_CMIXED block.

FXT1 CC_ALPHA Format

In the CC_ALPHA encoding format, three A5R5G5B5 colors are provided in the encoded block. A control bit (LERP) is used to define the lookup table (or tables) used to dereference the 2-bit Texel Selects.

CC_ALPHA Block Encoding

The following table describes the encoding of the 128-bit (DQWord) CC_ALPHA block format:



FXT CC_ALPHA Block Encoding

Bit	Description
127:125	Mode = '011'b (CC_ALPHA)
124	LERP
123:119	Color 2 Alpha
118:114	Color 1 Alpha
113:109	Color 0 Alpha
108:104	Color 2 Red
103:99	Color 2 Green
98:94	Color 2 Blue
93:89	Color 1 Red
88:84	Color 1 Green
83:79	Color 1 Blue
78:74	Color 0 Red
73:69	Color 0 Green
68:64	Color 0 Blue
63:62	Texel 31 Select
...	...
33:32	Texel 16 Select
31:30	Texel 15 Select
...	...
1:0	Texel 0 Select

CC_ALPHA Block Decoding

Each of the three colors (Color 0-2) are converted from A5R5G5B5 to A8R8G8B8 by replicating the 3 MSBs into the 3 LSBs, as shown in the following tables:

FXT CC_ALPHA Decoded Colors

Expanded Color Bit	Expanded Channel Bit	Encoded Block Source Bit
Color 2 [31:27]	Color 2 Alpha [7:3]	[123:119]
Color 2 [26:24]	Color 2 Alpha [2:0]	[123:121]
Color 2 [23:17]	Color 2 Red [7:3]	[108:104]
Color 2 [18:16]	Color 2 Red [2:0]	[108:106]
Color 2 [15:11]	Color 2 Green [7:3]	[103:99]
Color 2 [10:08]	Color 2 Green [2:0]	[103:101]
Color 2 [07:03]	Color 2 Blue [7:3]	[98:94]
Color 2 [02:00]	Color 2 Blue [2:0]	[98:96]
Color 1 [31:27]	Color 1 Alpha [7:3]	[118:114]
Color 1 [26:24]	Color 1 Alpha [2:0]	[118:116]
Color 1 [23:17]	Color 1 Red [7:3]	[93:89]
Color 1 [18:16]	Color 1 Red [2:0]	[93:91]
Color 1 [15:11]	Color 1 Green [7:3]	[88:84]
Color 1 [10:08]	Color 1 Green [2:0]	[88:86]
Color 1 [07:03]	Color 1 Blue [7:3]	[83:79]
Color 1 [02:00]	Color 1 Blue [2:0]	[83:81]
Color 0 [31:27]	Color 0 Alpha [7:3]	[113:109]
Color 0 [26:24]	Color 0 Alpha [2:0]	[113:111]
Color 0 [23:17]	Color 0 Red [7:3]	[78:74]
Color 0 [18:16]	Color 0 Red [2:0]	[78:76]
Color 0 [15:11]	Color 0 Green [7:3]	[73:69]
Color 0 [10:08]	Color 0 Green [2:0]	[73:71]
Color 0 [07:03]	Color 0 Blue [7:3]	[68:64]
Color 0 [02:00]	Color 0 Blue [2:0]	[68:66]

LERP = 0 Decoding

When LERP = 0, a single 4-entry lookup table is formed using the three expanded colors, with the 4th entry defined as transparent black (ARGB=0,0,0,0). Each 2-bit Texel n Select field of the encoded CC_ALPHA block is used to index into a 32-bit A8R8G8B8 color from the table completing the decode of the CC_ALPHA block.



FXT CC_ALPHA Interpolated Color Table (LERP=0)

Texel Select	Color	Alpha
0	Color0.RGB	Color0.Alpha
1	Color1.RGB	Color1.Alpha
2	Color2.RGB	Color2.Alpha
3	Black (RGB=0,0,0)	0

LERP = 1 Decoding

When LERP = 1, the three expanded colors are used to create two tables of four interpolated colors. The Color0,1 table is used as a lookup table for texel 0-15 indices, and the Color1,2 table used for texels 16-31 indices, as shown in the following figures:

FXT CC_ALPHA Interpolated Color Table (LERP=1, Texels 0-15)

Texel 0-15 Select	Color ARGB
0	Color0.ARGB
1	$(2 * \text{Color0.ARGB} + \text{Color1.ARGB} + 1) / 3$
2	$(\text{Color0.ARGB} + 2 * \text{Color1.ARGB} + 1) / 3$
3	Color1.ARGB

FXT CC_ALPHA Interpolated Color Table (LERP=1, Texels 16-31)

Texel 16-31 Select	Color ARGB
0	Color2.ARGB
1	$(2 * \text{Color2.ARGB} + \text{Color1.ARGB} + 1) / 3$
2	$(\text{Color2.ARGB} + 2 * \text{Color1.ARGB} + 1) / 3$
3	Color1.ARGB

DXT Texture Formats

Note that non-power-of-2 dimensioned maps may require the surface to be padded out to the next multiple of four texels – here the pad texels are not referenced by the device.

An 8-byte (QWord) block encoding can be used if the source texture contains no transparency (is opaque) or if the transparency can be specified by a one-bit alpha. A 16-byte (DQWord) block encoding can be used to support source textures that require more than one-bit alpha: here the 1st QWord is used to encode the texel alpha values, and the 2nd QWord is used to encode the texel color values.

These three types of format are discussed in the following sections:

- Opaque and One-bit Alpha Textures (DXT1)
- Opaque Textures (DXT1_RGB)
- Textures with Alpha Channels (DXT2-5)

Notes:

- Any single texture must specify that its data is stored as 64 or 128 bits per group of 16 texels. If 64-bit blocks—that is, format DXT1—are used for the texture, it is possible to mix the opaque and one-bit alpha formats on a per-block basis within the same texture. In other words, the comparison of the unsigned integer magnitude of color_0 and color_1 is performed uniquely for each block of 16 texels.
- When 128-bit blocks are used, then the alpha channel must be specified in either explicit (format DXT2 or DXT3) or interpolated mode (format DXT4 or DXT5) for the entire texture. Note that as with color, once interpolated mode is selected then either 8 interpolated alphas or 6 interpolated alphas mode can be used on a block-by-block basis. Again the magnitude comparison of alpha_0 and alpha_1 is done uniquely on a block-by-block basis.



Opaque and One-bit Alpha Textures (DXT1/BC1)

Texture format DXT1 is for textures that are opaque or have a single transparent color. For each opaque or one-bit alpha block, two 16-bit R5G6B5 values and a 4x4 bitmap with 2-bits-per-pixel are stored. This totals 64 bits (1 QWord) for 16 texels, or 4-bits-per-texel.

In the block bitmap, there are two bits per texel to select between the four colors, two of which are stored in the encoded data. The other two colors are derived from these stored colors by linear interpolation.

The one-bit alpha format is distinguished from the opaque format by comparing the two 16-bit color values stored in the block. They are treated as unsigned integers. If the first color is greater than the second, it implies that only opaque texels are defined. This means four colors will be used to represent the texels. In four-color encoding, there are two derived colors and all four colors are equally distributed in RGB color space. This format is analogous to R5G6B5 format. Otherwise, for one-bit alpha transparency, three colors are used and the fourth is reserved to represent transparent texels. Note that the color blocks in DXT2-5 formats strictly use four colors, as the alpha values are obtained from the alpha block .

In three-color encoding, there is one derived color and the fourth two-bit code is reserved to indicate a transparent texel (alpha information). This format is analogous to A1R5G5B5, where the final bit is used for encoding the alpha mask.



The following piece of pseudo-code illustrates the algorithm for deciding whether three- or four-color encoding is selected:

```

if (color_0 > color_1){ // Four-color block: derive the other two colors. // 00
= color_0, 01 = color_1, 10 = color_2, 11 = color_3 // These two bit codes
correspond to the 2-bit fields // stored in the 64-bit block. color_2 = (2 *
color_0 + color_1) / 3; color_3 = (color_0 + 2 * color_1) / 3;} else{ // Three-
color block: derive the other color. // 00 = color_0, 01 = color_1, 10 =
color_2, // 11 = transparent. // These two bit codes correspond to the 2-bit
fields // stored in the 64-bit block. color_2 = (color_0 + color_1) /
2; color_3 = transparent; }

```

The following tables show the memory layout for the 8-byte block. It is assumed that the first index corresponds to the y-coordinate and the second corresponds to the x-coordinate. For example, Texel[1][2] refers to the texture map pixel at (x,y) = (2,1).

Here is the memory layout for the 8-byte (64-bit) block:

Word Address	16-bit Word
0	Color_0
1	Color_1
2	Bitmap Word_0
3	Bitmap Word_1

Color_0 and Color_1 (colors at the two extremes) are laid out as follows:

Bits	Color
15:11	Red color component
10:5	Green color component
4:0	Blue color component

Bits	Texel
1:0 (LSB)	Texel[0][0]
3:2	Texel[0][1]
5:4	Texel[0][2]
7:6	Texel[0][3]
9:8	Texel[1][0]
11:10	Texel[1][1]
13:12	Texel[1][2]
15:14	Texel[1][3]



Bitmap Word_1 is laid out as follows:

Bits	Texel
1:0 (LSB)	Texel[2][0]
3:2	Texel[2][1]
5:4	Texel[2][2]
7:6	Texel[2][3]
9:8	Texel[3][0]
11:10	Texel[3][1]
13:12	Texel[3][2]
15:14 (MSB)	Texel[3][3]

Example of Opaque Color Encoding

As an example of opaque encoding, we will assume that the colors red and black are at the extremes. We will call red color_0 and black color_1. There will be four interpolated colors that form the uniformly distributed gradient between them. To determine the values for the 4x4 bitmap, the following calculations are used:

00 ? color_001 ? color_110 ? $2/3 \text{ color}_0 + 1/3 \text{ color}_1$ 11 ? $1/3 \text{ color}_0 + 2/3 \text{ color}_1$

Example of One-bit Alpha Encoding

This format is selected when the unsigned 16-bit integer, color_0, is less than the unsigned 16-bit integer, color_1. An example of where this format could be used is leaves on a tree to be shown against a blue sky. Some texels could be marked as transparent while three shades of green are still available for the leaves. Two of these colors fix the extremes, and the third color is an interpolated color.

The bitmap encoding for the colors and the transparency is determined using the following calculations:

00 ? color_0

01 ? color_1

10 ? $1/2 \text{ color}_0 + 1/2 \text{ color}_1$

11 ? Transparent

Opaque Textures (DXT1_RGB)

Texture format DXT1_RGB is identical to DXT1, with the exception that the One-bit Alpha encoding is removed. Color 0 and Color 1 are not compared, and the resulting texel color is derived strictly from the Opaque Color Encoding. The alpha channel defaults to 1.0.

Note: The behavior of this format is not compliant with the OGL spec. As a workaround, the **Surface Format** should be set to BC1 and the **Shader Channel Select A** should be set to SCS_ONE. This workaround is only available for DevHSW+ due to lack of Shader Channel Select support on earlier products.

Compressed Textures with Alpha Channels (DXT2-5 / BC2-3)

There are two ways to encode texture maps that exhibit more complex transparency. In each case, a block that describes the transparency precedes the 64-bit block already described. The transparency is either represented as a 4x4 bitmap with four bits per pixel (explicit encoding), or with fewer bits and linear interpolation analogous to what is used for color encoding.

The transparency block and the color block are laid out as follows:

Word Address	64-bit Block
3:0	Transparency block
7:4	Previously described 64-bit block

Explicit Texture Encoding

For explicit texture encoding (DXT2 and DXT3 formats), the alpha components of the texels that describe transparency are encoded in a 4x4 bitmap with 4 bits per texel. These 4 bits can be achieved through a variety of means such as dithering or by simply using the 4 most significant bits of the alpha data. However they are produced, they are used just as they are, without any form of interpolation.

Note:

DirectDraw's compression method uses the 4 most significant bits.

The following tables illustrate how the alpha information is laid out in memory, for each 16-bit word.

This is the layout for Word 0:

Bits	Alpha
3:0 (LSB)	[0][0]
7:4	[0][1]
11:8	[0][2]
15:12 (MSB)	[0][3]

This is the layout for Word 1:

Bits	Alpha
3:0 (LSB)	[1][0]
7:4	[1][1]
11:8	[1][2]
15:12 (MSB)	[1][3]

This is the layout for Word 2:

Bits	Alpha
3:0 (LSB)	[2][0]
7:4	[2][1]
11:8	[2][2]



Bits	Alpha
15:12 (MSB)	[2][3]

This is the layout for Word 3:

Bits	Alpha
3:0 (LSB)	[3][0]
7:4	[3][1]
11:8	[3][2]
15:12 (MSB)	[3][3]

Three-Bit Linear Alpha Interpolation

The encoding of transparency for the DXT4 and DXT5 formats is based on a concept similar to the linear encoding used for color. Two 8-bit alpha values and a 4x4 bitmap with three bits per pixel are stored in the first eight bytes of the block. The representative alpha values are used to interpolate intermediate alpha values. Additional information is available in the way the two alpha values are stored. If alpha_0 is greater than alpha_1, then six intermediate alpha values are created by the interpolation. Otherwise, four intermediate alpha values are interpolated between the specified alpha extremes. The two additional implicit alpha values are 0 (fully transparent) and 255 (fully opaque).

The following pseudo-code illustrates this algorithm:

```
// 8-alpha or 6-alpha block?
if (alpha_0 > alpha_1) {
    // 8-alpha block: derive the other 6 alphas.
    // 000 = alpha_0, 001 = alpha_1, others are interpolated
    alpha_2 = (6 * alpha_0 + alpha_1) / 7; // bit code 010
    alpha_3 = (5 * alpha_0 + 2 * alpha_1) / 7; // Bit code 011
    alpha_4 = (4 * alpha_0 + 3 * alpha_1) / 7; // Bit code 100
    alpha_5 = (3 * alpha_0 + 4 * alpha_1) / 7; // Bit code 101
    alpha_6 = (2 * alpha_0 + 5 * alpha_1) / 7; // Bit code 110
    alpha_7 = (alpha_0 + 6 * alpha_1) / 7; // Bit code 111
}
else {
    // 6-alpha block: derive the other alphas.
    // 000 = alpha_0, 001 = alpha_1, others are interpolated
    alpha_2 = (4 * alpha_0 + alpha_1) / 5; // Bit code 010
    alpha_3 = (3 * alpha_0 + 2 * alpha_1) / 5; // Bit code 011
    alpha_4 = (2 * alpha_0 + 3 * alpha_1) / 5; // Bit code 100
    alpha_5 = (alpha_0 + 4 * alpha_1) / 5; // Bit code 101
    alpha_6 = 0; // Bit code 110
    alpha_7 = 255; // Bit code 111
}
```

The memory layout of the alpha block is as follows:

Byte	Alpha
0	Alpha_0
1	Alpha_1
2	[0][2] (2 LSBs), [0][1], [0][0]
3	[1][1] (1 LSB), [1][0], [0][3], [0][2] (1 MSB)
4	[1][3], [1][2], [1][1] (2 MSBs)
5	[2][2] (2 LSBs), [2][1], [2][0]
6	[3][1] (1 LSB), [3][0], [2][3], [2][2] (1 MSB)
7	[3][3], [3][2], [3][1] (2 MSBs)



BC4

These formats (BC4_UNORM and BC4_SNORM) compresses single-component UNORM or SNORM data. An 8-byte compression block represents a 4x4 block of texels. The texels are labeled as texel[row][column] where both row and column range from 0 to 3. Texel[0][0] is the upper left texel.

The 8-byte compression block is laid out as follows:

Bit	Description
7:0	red_0
15:8	red_1
18:16	texel[0][0] bit code
21:19	texel[0][1] bit code
24:22	texel[0][2] bit code
27:25	texel[0][3] bit code
30:28	texel[1][0] bit code
33:31	texel[1][1] bit code
36:34	texel[1][2] bit code
39:37	texel[1][3] bit code
42:40	texel[2][0] bit code
45:43	texel[2][1] bit code
48:46	texel[2][2] bit code
51:49	texel[2][3] bit code
54:52	texel[3][0] bit code
57:55	texel[3][1] bit code
60:58	texel[3][2] bit code
63:61	texel[3][3] bit code



There are two interpolation modes, chosen based on which reference color is larger. The first mode has the two reference colors plus six equal-spaced interpolated colors between the reference colors, chosen based on the three-bit code for that texel. The second mode has the two reference colors plus four interpolated colors, chosen by six of the three-bit codes. The remaining two codes select min and max values for the colors. The values of red_0 through red_7 are computed as follows:

```
red_0 = red_0; // bit code 000
red_1 = red_1; // bit code 001
if (red_0 > red_1) {
    red_2 = (6 * red_0 + 1 * red_1) / 7; // bit code 010
    red_3 = (5 * red_0 + 2 * red_1) / 7; // bit code 011
    red_4 = (4 * red_0 + 3 * red_1) / 7; // bit code 100
    red_5 = (3 * red_0 + 4 * red_1) / 7; // bit code 101
    red_6 = (2 * red_0 + 5 * red_1) / 7; // bit code 110
    red_7 = (1 * red_0 + 6 * red_1) / 7; // bit code 111
}
else {
    red_2 = (4 * red_0 + 1 * red_1) / 5; // bit code 010
    red_3 = (3 * red_0 + 2 * red_1) / 5; // bit code 011
    red_4 = (2 * red_0 + 3 * red_1) / 5; // bit code 100
    red_5 = (1 * red_0 + 4 * red_1) / 5; // bit code 101
    red_6 = UNORM ? 0.0 : -1.0; // bit code 110 (0 for UNORM, -1 for SNORM)
    red_7 = 1.0; // bit code 111
}
```



BC5

These formats (BC5_UNORM and BC5_SNORM) compresses dual-component UNORM or SNORM data. A 16-byte compression block represents a 4x4 block of texels. The texels are labeled as texel[row][column] where both row and column range from 0 to 3. Texel[0][0] is the upper left texel.

The 16-byte compression block is laid out as follows:

Bit	Description
7:0	red_0
15:8	red_1
18:16	texel[0][0] red bit code
21:19	texel[0][1] red bit code
24:22	texel[0][2] red bit code
27:25	texel[0][3] red bit code
30:28	texel[1][0] red bit code
33:31	texel[1][1] red bit code
36:34	texel[1][2] red bit code
39:37	texel[1][3] red bit code
42:40	texel[2][0] red bit code
45:43	texel[2][1] red bit code
48:46	texel[2][2] red bit code
51:49	texel[2][3] red bit code
54:52	texel[3][0] red bit code
57:55	texel[3][1] red bit code
60:58	texel[3][2] red bit code
63:61	texel[3][3] red bit code
71:64	green_0
79:72	green_1
82:80	texel[0][0] green bit code
85:83	texel[0][1] green bit code
88:86	texel[0][2] green bit code
91:89	texel[0][3] green bit code
94:92	texel[1][0] green bit code
97:95	texel[1][1] green bit code
100:98	texel[1][2] green bit code
103:101	texel[1][3] green bit code
106:104	texel[2][0] green bit code
109:107	texel[2][1] green bit code

Bit	Description
112:110	texel[2][2] green bit code
115:113	texel[2][3] green bit code
118:116	texel[3][0] green bit code
121:119	texel[3][1] green bit code
124:122	texel[3][2] green bit code
127:125	texel[3][3] green bit code

There are two interpolation modes, chosen based on which reference color is larger. The first mode has the two reference colors plus six equal-spaced interpolated colors between the reference colors, chosen based on the three-bit code for that texel. The second mode has the two reference colors plus four interpolated colors, chosen by six of the three-bit codes. The remaining two codes select min and max values for the colors. The values of red_0 through red_7 are computed as follows:

```

red_0 = red_0; // bit code 000
red_1 = red_1; // bit code 001
if (red_0 > red_1) {
    red_2 = (6 * red_0 + 1 * red_1) / 7; // bit code 010
    red_3 = (5 * red_0 + 2 * red_1) / 7; // bit code 011
    red_4 = (4 * red_0 + 3 * red_1) / 7; // bit code 100
    red_5 = (3 * red_0 + 4 * red_1) / 7; // bit code 101
    red_6 = (2 * red_0 + 5 * red_1) / 7; // bit code 110
    red_7 = (1 * red_0 + 6 * red_1) / 7; // bit code 111
}
else {
    red_2 = (4 * red_0 + 1 * red_1) / 5; // bit code 010
    red_3 = (3 * red_0 + 2 * red_1) / 5; // bit code 011
    red_4 = (2 * red_0 + 3 * red_1) / 5; // bit code 100
    red_5 = (1 * red_0 + 4 * red_1) / 5; // bit code 101
    red_6 = UNORM ? 0.0 : -1.0; // bit code 110 (0 for UNORM, -1 for SNORM)
    red_7 = 1.0; // bit code 111
}

```

The same calculations are done for green, using the corresponding reference colors and bit codes.



BC6H

These formats (BC6H_UF16 and BC6H_SF16) compresses 3-channel images with high dynamic range (> 8 bits per channel). BC6H supports floating point denorms but there is no support for INF and NaN, other than with BC6H_SF16 –INF is supported. The alpha channel is not included, thus alpha is returned at its default value.

The BC6H block is 16 bytes and represents a 4x4 block of texels. The texels are labeled as texel[row][column] where both row and column range from 0 to 3. Texel[0][0] is the upper left texel. BC6H has 14 different modes, the mode that the block is in is contained in the least significant bits (either 2 or 5 bits).

The basic scheme consists of interpolating colors along either one or two lines, with per-texel indices indicating which color along the line is chosen for each texel. If a two-line mode is selected, one of 32 partition sets is indicated which selects which of the two lines each texel is assigned to.

Field Definition

There are 14 possible modes for a BC6H block, the format of each is indicated in the 14 tables below. The mode is selected by the unique mode bits specified in each table. The first 10 modes use two lines ("TWO"), and the last 4 use one line ("ONE"). The difference between the various two-line and one-line modes is with the precision of the first endpoint and the number of bits used to store delta values for the remaining endpoints. Two modes (9 and 10) specify each endpoint as an original value rather than using the deltas (these are indicated as having no delta values).

The endpoints values and deltas are indicated in the tables using a two-letter name. The first letter is "r", "g", or "b" indicating the color channel. The second letter is "w", "x", "y", or "z" indicating which of the four endpoints. The first line has endpoints "w" and "x", with "w" being the endpoint that is fully specified (i.e. not as a delta). The second line has endpoints "y" and "z". Modes using ONE mode do not have endpoints "y" and "z" as they have only one line.

In addition to the mode and endpoint data, TWO blocks contain a 5-bit "partition" which selects one of the partition sets, and a 46-bit set of indices. ONE blocks contain a 63-bit set of indices. These are described in more detail below.

Mode 0: (TWO) Red, Green, Blue: 10-bit endpoint, 5-bit deltas

Bit	Description
1:0	mode = 00
2	gy[4]
3	by[4]
4	bz[4]
14:5	rw[9:0]
24:15	gw[9:0]
34:25	bw[9:0]
39:35	rx[4:0]
40	gz[4]
44:41	gy[3:0]
49:45	gx[4:0]
50	bz[0]
54:51	gz[3:0]
59:55	bx[4:0]
60	bz[1]
64:61	by[3:0]
69:65	ry[4:0]
70	bz[2]
75:71	rz[4:0]
76	bz[3]
81:77	partition
127:82	indices



Mode 1: (TWO) Red, Green, Blue: 7-bit endpoint, 6-bit deltas

Bit	Description
1:0	mode = 01
2	gy[5]
3	gz[4]
4	gz[5]
11:5	rw[6:0]
12	bz[0]
13	bz[1]
14	by[4]
21:15	gw[6:0]
22	by[5]
23	bz[2]
24	gy[4]
31:25	bw[6:0]
32	bz[3]
33	bz[5]
34	bz[4]
40:35	rx[5:0]
44:41	gy[3:0]
50:45	gx[5:0]
54:51	gz[3:0]
60:55	bx[5:0]
64:61	by[3:0]
70:65	ry[5:0]
76:71	rz[5:0]
81:77	partition
127:82	indices

Mode 2: (TWO) Red: 11-bit endpoint, 5-bit deltas

Green, Blue: 11-bit endpoint, 4-bit deltas

Bit	Description
4:0	mode = 00010
14:5	rw[9:0]
24:15	gw[9:0]
34:25	bw[9:0]
39:35	rx[4:0]
40	rw[10]
44:41	gy[3:0]
48:45	gx[3:0]
49	gw[10]
50	bz[0]
54:51	gz[3:0]
58:55	bx[3:0]
59	bw[10]
60	bz[1]
64:61	by[3:0]
69:65	ry[4:0]
70	bz[2]
75:71	rz[4:0]
76	bz[3]
81:77	partition
127:82	indices



Mode 3: (TWO) Red, Blue: 11-bit endpoint, 4-bit deltas

Green: 11-bit endpoint, 5-bit deltas

Bit	Description
4:0	mode = 00110
14:5	rw[9:0]
24:15	gw[9:0]
34:25	bw[9:0]
38:35	rx[3:0]
39	rw[10]
40	gz[4]
44:41	gy[3:0]
49:45	gx[4:0]
50	gw[10]
54:51	gz[3:0]
58:55	bx[3:0]
59	bw[10]
60	bz[1]
64:61	by[3:0]
68:65	ry[3:0]
69	bz[0]
70	bz[2]
74:71	rz[3:0]
75	gy[4]
76	bz[3]
81:77	partition
127:82	indices

Mode 4: (TWO) Red, Green: 11-bit endpoint, 4-bit deltas

Blue: 11-bit endpoint, 5-bit deltas

Bit	Description
4:0	mode = 01010
14:5	rw[9:0]
24:15	gw[9:0]
34:25	bw[9:0]
38:35	rx[3:0]
39	rw[10]
40	by[4]
44:41	gy[3:0]
48:45	gx[3:0]
49	gw[10]
50	bz[0]
54:51	gz[3:0]
59:55	bx[4:0]
60	bw[10]
64:61	by[3:0]
68:65	ry[3:0]
69	bz[1]
70	bz[2]
74:71	rz[3:0]
75	bz[4]
76	bz[3]
81:77	partition
127:82	indices



Mode 5: (TWO) Red, Green, Blue: 9-bit endpoint, 5-bit deltas

Bit	Description
4:0	mode = 01110
13:5	rw[8:0]
14	by[4]
23:15	gw[8:0]
24	gy[4]
33:25	bw[8:0]
34	bz[4]
39:35	rx[4:0]
40	gz[4]
44:41	gy[3:0]
49:45	gx[3:0]
50	bz[0]
54:51	gz[3:0]
59:55	bx[4:0]
60	bz[1]
64:61	by[3:0]
69:65	ry[4:0]
70	bz[2]
75:71	rz[4:0]
76	bz[3]
81:77	partition
127:82	indices

Mode 6: (TWO) Red: 8-bit endpoint, 6-bit deltas

Green, Blue: 8-bit endpoint, 5-bit deltas

Bit	Description
4:0	mode = 10010
12:5	rw[7:0]
13	gz[4]
14	by[4]
22:15	gw[7:0]
23	bz[2]
24	gy[4]
32:25	bw[7:0]
33	bz[3]
34	bz[4]
40:35	rx[5:0]
44:41	gy[3:0]
49:45	gx[4:0]
50	bz[0]
54:51	gz[3:0]
59:55	bx[4:0]
60	gz[1]
64:61	by[3:0]
70:65	ry[5:0]
76:71	rz[5:0]
81:77	partition
127:82	indices



Mode 7: (TWO) Red, Blue: 8-bit endpoint, 5-bit deltas

Green: 8-bit endpoint, 6-bit deltas

Bit	Description
4:0	mode = 10110
12:5	rw[7:0]
13	bz[0]
14	by[4]
22:15	gw[7:0]
23	gy[5]
24	gy[4]
32:25	bw[7:0]
33	gz[5]
34	bz[4]
39:35	rx[4:0]
40	gz[4]
44:41	gy[3:0]
50:45	gx[5:0]
54:51	gz[3:0]
59:55	bx[4:0]
60	bz[1]
64:61	by[3:0]
69:65	ry[4:0]
70	bz[2]
75:71	rz[4:0]
76	bz[3]
81:77	partition
127:82	indices

Mode 8: (TWO) Red, Green: 8-bit endpoint, 5-bit deltas

Blue: 8-bit endpoint, 6-bit deltas

Bit	Description
4:0	mode = 11010
12:5	rw[7:0]
13	bz[1]
14	by[4]
22:15	gw[7:0]
23	by[5]
24	gy[4]
32:25	bw[7:0]
33	bz[5]
34	bz[4]
39:35	rx[4:0]
40	gz[4]
44:41	gy[3:0]
49:45	gx[4:0]
50	bz[0]
54:51	gz[3:0]
60:55	bx[5:0]
64:61	by[3:0]
69:65	ry[4:0]
70	bz[2]
75:71	rz[4:0]
76	bz[3]
81:77	partition
127:82	indices



Mode 9: (TWO) Red, Green, Blue: 6-bit endpoints for all four, no deltas

Bit	Description
4:0	mode = 11110
10:5	rw[5:0]
11	gz[4]
12	bz[0]
13	bz[1]
14	by[4]
20:15	gw[5:0]
21	gy[5]
22	by[5]
23	bz[2]
24	gy[4]
30:25	bw[5:0]
31	gz[5]
32	bz[3]
33	bz[5]
34	bz[4]
40:35	rx[5:0]
44:41	gy[3:0]
50:45	gx[5:0]
54:51	gz[3:0]
60:55	bx[5:0]
64:61	by[3:0]
70:65	ry[5:0]
76:71	rz[5:0]
81:77	partition
127:82	indices

Mode 10: (ONE) Red, Green, Blue: 10-bit endpoints for both, no deltas

Bit	Description
4:0	mode = 00011
14:5	rw[9:0]
24:15	gw[9:0]
34:25	bw[9:0]
44:35	rx[9:0]
54:45	gx[9:0]
64:55	bx[9:0]
127:65	indices

Mode 11: (ONE) Red, Green, Blue: 11-bit endpoints, 9-bit deltas

Bit	Description
4:0	mode = 00111
14:5	rw[9:0]
24:15	gw[9:0]
34:25	bw[9:0]
43:35	rx[8:0]
44	rw[10]
53:45	gx[8:0]
54	gw[10]
63:55	bx[8:0]
64	bw[10]
127:65	indices



Mode 12: (ONE) Red, Green, Blue: 12-bit endpoints, 8-bit deltas

Bit	Description
4:0	mode = 01011
14:5	rw[9:0]
24:15	gw[9:0]
34:25	bw[9:0]
42:35	rx[7:0]
43	rw[11]
44	rw[10]
52:45	gx[7:0]
53	gw[11]
54	gw[10]
62:55	bx[7:0]
63	bw[11]
64	bw[10]
127:65	indices

Mode 13: (ONE) Red, Green, Blue: 16-bit endpoints, 4-bit deltas

Bit	Description
4:0	mode = 01111
14:5	rw[9:0]
24:15	gw[9:0]
34:25	bw[9:0]
38:35	rx[3:0]
39	rw[15]
40	rw[14]
41	rw[13]
42	rw[12]
43	rw[11]
44	rw[10]
48:45	gx[3:0]
49	gw[15]
50	gw[14]
51	gw[13]
52	gw[12]
53	gw[11]
54	gw[10]
58:55	bx[3:0]
59	bw[15]
60	bw[14]
61	bw[13]
62	bw[12]
63	bw[11]
64	bw[10]
127:65	indices

Undefined mode values (10011, 10111, 11011, and 11111) return zero in the RGB channels.

The "indices" fields are defined as follows:



TWO mode *indices* field with fix-up index [1] at texel[3][3]

Bit	Description
83:82	texel[0][0] index
86:84	texel[0][1] index
89:87	texel[0][2] index
92:90	texel[0][3] index
95:93	texel[1][0] index
98:96	texel[1][1] index
101:99	texel[1][2] index
104:102	texel[1][3] index
107:105	texel[2][0] index
110:108	texel[2][1] index
113:111	texel[2][2] index
116:114	texel[2][3] index
119:117	texel[3][0] index
122:120	texel[3][1] index
125:123	texel[3][2] index
127:126	texel[3][3] index

TWO mode *indices* field with fix-up index [1] at texel[0][2]

Bit	Description
83:82	texel[0][0] index
86:84	texel[0][1] index
88:87	texel[0][2] index
91:89	texel[0][3] index
94:92	texel[1][0] index
97:95	texel[1][1] index
100:98	texel[1][2] index
103:101	texel[1][3] index
106:104	texel[2][0] index
109:107	texel[2][1] index
112:110	texel[2][2] index
115:113	texel[2][3] index
118:116	texel[3][0] index
121:119	texel[3][1] index
124:122	texel[3][2] index
127:125	texel[3][3] index





TWO mode *indices* field with fix-up index [1] at texel[2][0]

Bit	Description
83:82	texel[0][0] index
86:84	texel[0][1] index
89:87	texel[0][2] index
92:90	texel[0][3] index
95:93	texel[1][0] index
98:96	texel[1][1] index
101:99	texel[1][2] index
104:102	texel[1][3] index
106:105	texel[2][0] index
109:107	texel[2][1] index
112:110	texel[2][2] index
115:113	texel[2][3] index
118:116	texel[3][0] index
121:119	texel[3][1] index
124:122	texel[3][2] index
127:125	texel[3][3] index

ONE mode *indices* field

Bit	Description
67:65	texel[0][0] index
71:68	texel[0][1] index
75:72	texel[0][2] index
79:76	texel[0][3] index
83:80	texel[1][0] index
87:84	texel[1][1] index
91:88	texel[1][2] index
95:92	texel[1][3] index
99:96	texel[2][0] index
103:100	texel[2][1] index
107:104	texel[2][2] index
111:108	texel[2][3] index
115:112	texel[3][0] index
119:116	texel[3][1] index
123:120	texel[3][2] index
127:124	texel[3][3] index



Endpoint Computation

The endpoints can be defined in many different ways, as shown above. This section describes how the endpoints are computed from the bits in the compression block. The method used depends on whether the BC6H format is signed (BC6H_SF16) or unsigned (BC6H_UF16).

First, each channel (RGB) of each endpoint is extended to 16 bits. Each is handled identically and independently, however in some modes different channels have different incoming precision which must be accounted for. The following rules are employed:

- If the format is BC6H_SF16 or the endpoint is a delta value, the value is sign-extended to 16 bits
- For all other cases, the value is zero-extended to 16 bits

If there are no endpoints that are delta values, endpoint computation is complete. For endpoints that are delta values, the next step involves computing the absolute endpoint. The "w" endpoint is always absolute and acts as a base value for the other three endpoints. Each channel is handled identically and independently.

$$\begin{aligned}x &= w + x \\y &= w + y \\z &= w + z\end{aligned}$$

The above is performed using 16-bit integer arithmetic. Overflows beyond 16 bits are ignored (any resulting high bits are dropped).



Palette Color Computation

The next step involves computing the color palette values that provide the available values for each texel's color. The color palette for each line consists of the two endpoint colors plus 6 (TWO mode) or 14 (ONE mode) interpolated colors. Again each channel is processed independently.

First the endpoints are unquantized, with each channel of each endpoint being processed independently. The number of bits in the original base "w" value represents the precision of the endpoints. The input endpoint is called "e", and the resulting endpoints are represented as 17-bit signed integers and called e' below.

For the BC6H_UF16 format:

- if the precision is already 16 bits, $e' = e$
- if $e = 0$, $e' = 0$
- if e is the maximum representable in the precision, $e' = 0xFFFF$
- otherwise, $e' = ((e \ll 16) + 0x8000) \gg \text{precision}$

For the BC6H_SF16 format, the value is treated as sign magnitude. The sign is not changed, e' and e refer only to the magnitude portion:

- if the precision is already 16 bits, $e' = e$
- if $e = 0$, $e' = 0$
- if e is the maximum representable in the precision, $e' = 0x7FFF$
- otherwise, $e' = ((e \ll 15) + 0x4000) \gg (\text{precision} - 1)$

Next, the palette values are generated using predefined weights, using the tables below:

`palette[i] = (w' * (64 - weight[i]) + x' * weight[i] + 32) >> 6`

TWO mode weights:

palette index	0	1	2	3	4	5	6	7
weight	0	9	18	27	37	46	55	64

ONE mode weights:

palette index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
weight	0	4	9	13	17	21	26	30	34	38	43	47	51	55	60	64

Note that the two end palette indices are equal to the two endpoints given that the weights are 0 and 64. In the above equation w' and x' represent the endpoints e' computed in the previous step corresponding to w and x, respectively. For the second line in TWO mode, w and x are replaced with y and z.



The final step in computing the palette colors is to rescale the final results. For BC6H_UF16 format, the values are multiplied by $31/64$. For BC6H_SF16, the values are multiplied by $31/32$, treating them as sign magnitude. These final 16-bit results are ultimately treated as 16-bit floats.

Texel Selection

The final step is to select the appropriate palette index for each texel. This index then selects the 16-bit per channel palette value, which is re-interpreted as a 16-bit floating point result for input into the filter. This procedure differs depending on whether the mode is TWO or ONE.

ONE Mode

In ONE mode, there is only one set of palette colors, but the "indices" field is 63 bits. This field consists of a 4-bit palette index for each of the 16 texels, with the exception of the texel at [0][0] which has only 3 bits, the missing high bit being set to zero.



TWO Mode

32 partitions are defined for TWO, which are defined below. Each of the 32 cases shows the 4x4 block of texels, and is indexed by adding its hexadecimal row number (00-1C) to its column number (0-3). Each texel in the 4x4 block indicates whether that texel is part of line 0 (endpoints w and x) or line 1 (endpoints y and z). Each case has one texel each of "[0]" and "[1]", the index that this is at is termed the "fix-up index". These texels have one less bit in the index.

	0				1				2				3			
00	[0]	0	1	1	[0]	0	0	1	[0]	1	1	1	[0]	0	0	1
	0	0	1	1	0	0	0	1	0	1	1	1	0	0	1	1
	0	0	1	1	0	0	0	1	0	1	1	1	0	0	1	1
	0	0	1	[1]	0	0	0	[1]	0	1	1	[1]	0	1	1	[1]
04	[0]	0	0	0	[0]	0	1	1	[0]	0	0	1	[0]	0	0	0
	0	0	0	1	0	1	1	1	0	0	1	1	0	0	0	1
	0	0	0	1	0	1	1	1	0	1	1	1	0	0	1	1
	0	0	1	[1]	1	1	1	[1]	1	1	1	[1]	0	1	1	[1]
08	[0]	0	0	0	[0]	0	1	1	[0]	0	0	0	[0]	0	0	0
	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	0
	0	0	0	1	1	1	1	1	0	1	1	1	0	0	0	1
	0	0	1	[1]	1	1	1	[1]	1	1	1	[1]	0	1	1	[1]
0C	[0]	0	0	1	[0]	0	0	0	[0]	0	0	0	[0]	0	0	0
	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
	1	1	1	[1]	1	1	1	[1]	1	1	1	[1]	1	1	1	[1]
10	[0]	0	0	0	[0]	1	[1]	1	[0]	0	0	0	[0]	1	[1]	1
	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1
	1	1	1	0	0	0	0	0	[1]	0	0	0	0	0	0	1
	1	1	1	[1]	0	0	0	0	1	1	1	0	0	0	0	0
14	[0]	0	[1]	1	[0]	0	0	0	[0]	0	0	0	[0]	1	1	1
	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1
	0	0	0	0	[1]	1	0	0	[1]	0	0	0	0	0	1	1
	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	[1]
18	[0]	0	[1]	1	[0]	0	0	0	[0]	1	[1]	0	[0]	0	[1]	1
	0	0	0	1	1	0	0	0	0	1	1	0	0	1	1	0
	0	0	0	1	[1]	0	0	0	0	1	1	0	0	1	1	0
	0	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0
1C	[0]	0	0	1	[0]	0	0	0	[0]	1	[1]	1	[0]	0	[1]	1

	0				1				2				3			
	0	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1
	[1]	1	1	0	[1]	1	1	1	1	0	0	0	1	0	0	1
	1	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0

The 46-bit "indices" field consists of a 3-bit palette index for each of the 16 texels, with the exception of the bracketed texels that have only two bits each. The high bit of these texels is set to zero.



BC7

These formats (BC7_UNORM and BC7_UNORM_SRGB) compresses 3-channel and 4-channel fixed point images.

The BC7 block is 16 bytes and represents a 4x4 block of texels. The texels are labeled as texel[row][column] where both row and column range from 0 to 3. Texel[0][0] is the upper left texel. BC7 has 8 different modes, the mode that the block is in is contained in the least significant bits (1-8 bits depending on mode).

The basic scheme consists of interpolating colors and alpha in some modes along either one, two, or three lines, with per-texel indices indicating which color/alpha along the line is chosen for each texel. If a two- or three-line mode is selected, one of 64 partition sets is indicated which selects which of the two lines each texel is assigned to, although some modes are limited to the first 16 partition sets. In the color-only modes, alpha is always returned at its default value of 1.0.

Some modes contain the following fields:

- **P-bits.** These represent shared LSB for all components of the endpoint, which increases the endpoint precision by one bit. In some cases both endpoints of a line share a P-bit.
- **Rotation bits.** For blocks with separate color and alpha, this 2-bit field allows selection of which of the four components has its own indexes (scalar) vs. the other three components (vector).
- **Index selector.** This 1-bit field selects whether the scalar or vector components uses the 3-bit index vs. the 2-bit index.

Field Definition

There are 8 possible modes for a BC7 block, the format of each is indicated in the 8 tables below. The mode is selected by the unique mode bits specified in each table. Each mode has particular characteristics described at the top of the table.

Mode 0: Color only, 3 lines (THREE), 4-bit endpoints with one P-bit per endpoint, 3-bit indices, 16 partitions

Bit	Description
0	mode = 0
4:1	partition
8:5	R0
12:9	R1
16:13	R2
20:17	R3
24:21	R4
28:25	R5
32:29	G0
36:33	G1
40:37	G2
44:41	G3
48:45	G4
52:49	G5
56:53	B0
60:57	B1
64:61	B2
68:65	B3
72:69	B4
76:73	B5
77	P0
78	P1
79	P2
80	P3
81	P4
82	P5
127:83	indices



Mode 1: Color only, 2 lines (TWO), 6-bit endpoints with one shared P-bit per line, 3-bit indices, 64 partitions

Bit	Description
1:0	mode = 10
7:2	partition
13:8	R0
19:14	R1
25:20	R2
31:26	R3
37:32	G0
43:38	G1
49:44	G2
55:50	G3
61:56	B0
67:62	B1
73:68	B2
79:74	B3
80	P0
81	P1
127:82	indices

Mode 2: Color only, 3 lines (THREE), 5-bit endpoints, 2-bit indices, 64 partitions

Bit	Description
2:0	mode = 100
8:3	partition
13:9	R0
18:14	R1
23:19	R2
28:24	R3
33:29	R4
38:34	R5
43:39	G0
48:44	G1
53:49	G2
58:54	G3
63:59	G4
68:64	G5
73:69	B0
78:74	B1
83:79	B2
88:84	B3
93:89	B4
98:94	B5
127:99	indices



Mode 3: Color only, 2 lines (TWO), 7-bit endpoints with one P-bit per endpoint, 2-bit indices, 64 partitions

Bit	Description
3:0	mode = 1000
9:4	partition
16:10	R0
23:17	R1
30:24	R2
37:31	R3
44:38	G0
51:45	G1
58:52	G2
65:59	G3
72:66	B0
79:73	B1
86:80	B2
93:87	B3
94	P0
95	P1
96	P2
97	P3
127:98	indices

Mode 4: Color and alpha, 1 line (ONE), 5-bit color endpoints, 6-bit alpha endpoints, 16 2-bit indices, 16 3-bit indices, 2-bit component rotation, 1-bit index selector

Bit	Description
4:0	mode = 10000
6:5	rotation
7	index selector
12:8	R0
17:13	R1
22:18	G0
27:23	G1
32:28	B0
37:33	B1
43:38	A0
49:44	A1
80:50	2-bit indices
127:81	3-bit indices

Mode 5: Color and alpha, 1 line (ONE), 7-bit color endpoints, 8-bit alpha endpoints, 2-bit color indices, 2-bit alpha indices, 2-bit component rotation

Bit	Description
5:0	mode = 100000
7:6	rotation
14:8	R0
21:15	R1
28:22	G0
35:29	G1
42:36	B0
49:43	B1
57:50	A0
65:58	A1
96:66	color indices
127:97	alpha indices



Mode 6: Combined color and alpha, 1 line (ONE), 7-bit endpoints with one P-bit per endpoint, 4-bit indices

Bit	Description
6:0	mode = 1000000
13:7	R0
20:14	R1
27:21	G0
34:28	G1
41:35	B0
48:42	B1
55:49	A0
62:56	A1
63	P0
64	P1
127:65	indices



Mode 7: Combined color and alpha, 2 lines (TWO), 5-bit endpoints with one P-bit per endpoint, 2-bit indices, 64 partitions

Bit	Description
7:0	mode = 10000000
13:8	partition
18:14	R0
23:19	R1
28:24	R2
33:29	R3
38:34	G0
43:39	G1
48:44	G2
53:49	G3
58:54	B0
63:59	B1
68:64	B2
73:69	B3
78:74	A0
83:79	A1
88:84	A2
93:89	A3
94	P0
95	P1
96	P2
97	P3
127:98	indices



Undefined mode values (bits 7:0 = 00000000) return zero in the RGB channels.

The indices fields are variable in length and due to the different locations of the fix-up indices depending on partition set there are a very large number of possible configurations. Each mode above indicates how many bits each index has, and the fix-up indices (one in ONE mode, two in TWO mode, and three in THREE mode) each have one less bit than indicated. However, the indices are always packed into the index fields according to the table below, with the specific bit assignments of each texel following the rules just given.

Bit	Description
LSBs	texel[0][0] index
	texel[0][1] index
	texel[0][2] index
	texel[0][3] index
	texel[1][0] index
	texel[1][1] index
	texel[1][2] index
	texel[1][3] index
	texel[2][0] index
	texel[2][1] index
	texel[2][2] index
	texel[2][3] index
MSBs	texel[3][0] index
	texel[3][1] index
	texel[3][2] index
MSBs	texel[3][3] index

Endpoint Computation

The endpoints can be defined with different precision depending on mode, as shown above. This section describes how the endpoints are computed from the bits in the compression block. Each component of each endpoint follows the same steps.

If a P-bit is defined for the endpoint, it is first added as an additional LSB at the bottom of the endpoint value. The endpoint is then bit-replicated to create an 8-bit fixed point endpoint value with a range from 0x00 to 0xFF.

Palette Color Computation

The next step involves computing the color palette values that provide the available values for each texel's color. The color palette for each line consists of the two endpoint colors plus 2, 6, or 14 interpolated colors, depending on the number of bits in the indices. Again each channel is processed independently.

The equation to compute each palette color with index i , given two endpoints is as follows, using the tables below to determine the weight for each palette index:

$$\text{palette}[i] = (E0 * (64 - \text{weight}[i]) + E1 * \text{weight}[i] + 32) \gg 6$$

2-bit index weights:

palette index	0	1	2	3
weight	0	21	43	64

3-bit index weights:

palette index	0	1	2	3	4	5	6	7
weight	0	9	18	27	37	46	55	64

4-bit index weights:

palette index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
weight	0	4	9	13	17	21	26	30	34	38	43	47	51	55	60	64

Note that the two end palette indices are equal to the two endpoints given that the weights are 0 and 64. In the above equation $E0$ and $E1$ represent the even-numbered and odd-numbered endpoints computed in the previous step for the component and line currently being computed.

Texel Selection

The final step is to select the appropriate palette index for each texel. This index then selects the 8-bit per channel palette value, which is interpreted as an 8-bit UNORM value for input into the filter (In BC7_UNORM_SRGB to UNORM values first go through inverse gamma conversion). This procedure differs depending on whether the mode is ONE, TWO, or THREE.

ONE Mode

In ONE mode, there is only one set of palette colors, thus there is only a single "partition set" defined, with all texels selecting line 0 and texel [0][0] being the "fix-up index" with one less bit in the index.



TWO Mode

64 partitions are defined for TWO, which are defined below. Each of the 64 cases shows the 4x4 block of texels, and is indexed by adding its hexadecimal row number (00-3C) to its column number (0-3). Each texel in the 4x4 block indicates whether that texel is part of line 0 (endpoints 0 and 1) or line 1 (endpoints 2 and 3). Each case has one texel each of "[0]" and "[1]", the index that this is at is termed the "fix-up index". These texels have one less bit in the index.

	0				1				2				3			
00	[0]	0	1	1	[0]	0	0	1	[0]	1	1	1	[0]	0	0	1
	0	0	1	1	0	0	0	1	0	1	1	1	0	0	1	1
	0	0	1	1	0	0	0	1	0	1	1	1	0	0	1	1
	0	0	1	[1]	0	0	0	[1]	0	1	1	[1]	0	1	1	[1]
04	[0]	0	0	0	[0]	0	1	1	[0]	0	0	1	[0]	0	0	0
	0	0	0	1	0	1	1	1	0	0	1	1	0	0	0	1
	0	0	0	1	0	1	1	1	0	1	1	1	0	0	1	1
	0	0	1	[1]	1	1	1	[1]	1	1	1	[1]	0	1	1	[1]
08	[0]	0	0	0	[0]	0	1	1	[0]	0	0	0	[0]	0	0	0
	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	0
	0	0	0	1	1	1	1	1	0	1	1	1	0	0	0	1
	0	0	1	[1]	1	1	1	[1]	1	1	1	[1]	0	1	1	[1]
0C	[0]	0	0	1	[0]	0	0	0	[0]	0	0	0	[0]	0	0	0
	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
	1	1	1	[1]	1	1	1	[1]	1	1	1	[1]	1	1	1	[1]
10	[0]	0	0	0	[0]	1	[1]	1	[0]	0	0	0	[0]	1	[1]	1
	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1
	1	1	1	0	0	0	0	0	[1]	0	0	0	0	0	0	1
	1	1	1	[1]	0	0	0	0	1	1	1	0	0	0	0	0
14	[0]	0	[1]	1	[0]	0	0	0	[0]	0	0	0	[0]	1	1	1
	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1
	0	0	0	0	[1]	1	0	0	[1]	0	0	0	0	0	1	1
	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	[1]
18	[0]	0	[1]	1	[0]	0	0	0	[0]	1	[1]	0	[0]	0	[1]	1
	0	0	0	1	1	0	0	0	0	1	1	0	0	1	1	0
	0	0	0	1	[1]	0	0	0	0	1	1	0	0	1	1	0
	0	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0
1C	[0]	0	0	1	[0]	0	0	0	[0]	1	[1]	1	[0]	0	[1]	1
	0	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1

	0				1				2				3			
	[1]	1	1	0	[1]	1	1	1	1	0	0	0	1	0	0	1
	1	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0
20	[0]	1	0	1	[0]	0	0	0	[0]	1	0	1	[0]	0	1	1
	0	1	0	1	1	1	1	1	1	0	[1]	0	0	0	1	1
	0	1	0	1	0	0	0	0	0	1	0	1	[1]	1	0	0
	0	1	0	[1]	1	1	1	[1]	1	0	1	0	1	1	0	0
24	[0]	0	[1]	1	[0]	1	0	1	[0]	1	1	0	[0]	1	0	1
	1	1	0	0	0	1	0	1	1	0	0	1	1	0	1	0
	0	0	1	1	[1]	0	1	0	0	1	1	0	1	0	1	0
	1	1	0	0	1	0	1	0	1	0	0	[1]	0	1	0	[1]
28	[0]	1	[1]	1	[0]	0	0	1	[0]	0	[1]	1	[0]	0	[1]	1
	0	0	1	1	0	0	1	1	0	0	1	0	1	0	1	1
	1	1	0	0	[1]	1	0	0	0	1	0	0	1	1	0	1
	1	1	1	0	1	0	0	0	1	1	0	0	1	1	0	0
2C	[0]	1	[1]	0	[0]	0	1	1	[0]	1	1	0	[0]	0	0	0
	1	0	0	1	1	1	0	0	0	1	1	0	0	1	[1]	0
	1	0	0	1	1	1	0	0	1	0	0	1	0	1	1	0
	0	1	1	0	0	0	1	[1]	1	0	0	[1]	0	0	0	0
30	[0]	1	0	0	[0]	0	[1]	0	[0]	0	0	0	[0]	0	0	0
	1	1	[1]	0	0	1	1	1	0	0	[1]	0	0	1	0	0
	0	1	0	0	0	0	1	0	0	1	1	1	[1]	1	1	0
	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
34	[0]	1	1	0	[0]	0	1	1	[0]	1	[1]	0	[0]	0	[1]	1
	1	1	0	0	0	1	1	0	0	0	1	1	1	0	0	1
	1	0	0	1	1	1	0	0	1	0	0	1	1	1	0	0
	0	0	1	[1]	1	0	0	[1]	1	1	0	0	0	1	1	0
38	[0]	1	1	0	[0]	1	1	0	[0]	1	1	1	[0]	0	0	1
	1	1	0	0	0	0	1	1	1	1	1	0	1	0	0	0
	1	1	0	0	0	0	1	1	1	0	0	0	1	1	1	0
	1	0	0	[1]	1	0	0	[1]	0	0	0	[1]	0	1	1	[1]
3C	[0]	0	0	0	[0]	0	[1]	1	[0]	0	[1]	0	[0]	1	0	0
	1	1	1	1	0	0	1	1	0	0	1	0	0	1	0	0
	0	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1
	0	0	1	[1]	0	0	0	0	1	1	1	0	0	1	1	[1]



THREE Mode

64 partitions are defined for THREE, which are defined below. Each of the 64 cases shows the 4x4 block of texels, and is indexed by adding its hexadecimal row number (00-3C) to its column number (0-3). Each texel in the 4x4 block indicates whether that texel is part of line 0 (endpoints 0 and 1), line 1 (endpoints 2 and 3), or line 2 (endpoints 4 and 5). Each case has one texel each of "[0]", "[1]", and "[2]", the index that this is at is termed the "fix-up index". These texels have one less bit in the index.

	0				1				2				3			
00	[0]	0	1	[1]	[0]	0	0	[1]	[0]	0	0	0	[0]	2	2	[2]
	0	0	1	1	0	0	1	1	2	0	0	1	0	0	2	2
	0	2	2	1	[2]	2	1	1	[2]	2	1	1	0	0	1	1
	2	2	2	[2]	2	2	2	1	2	2	1	[1]	0	1	1	[1]
04	[0]	0	0	0	[0]	0	1	[1]	[0]	0	2	[2]	[0]	0	1	1
	0	0	0	0	0	0	1	1	0	0	2	2	0	0	1	1
	[1]	1	2	2	0	0	2	2	1	1	1	1	[2]	2	1	1
	1	1	2	[2]	0	0	2	[2]	1	1	1	[1]	2	2	1	[1]
08	[0]	0	0	0	[0]	0	0	0	[0]	0	0	0	[0]	0	1	2
	0	0	0	0	1	1	1	1	1	1	[1]	1	0	0	[1]	2
	[1]	1	1	1	[1]	1	1	1	2	2	2	2	0	0	1	2
	2	2	2	[2]	2	2	2	[2]	2	2	2	[2]	0	0	1	[2]
0C	[0]	1	1	2	[0]	1	2	2	[0]	0	1	[1]	[0]	0	1	[1]
	0	1	[1]	2	0	[1]	2	2	0	1	1	2	2	0	0	1
	0	1	1	2	0	1	2	2	1	1	2	2	[2]	2	0	0
	0	1	1	[2]	0	1	2	[2]	1	2	2	[2]	2	2	2	0
10	[0]	0	0	[1]	[0]	1	1	[1]	[0]	0	0	0	[0]	0	2	[2]
	0	0	1	1	0	0	1	1	1	1	2	2	0	0	2	2
	0	1	1	2	[2]	0	0	1	[1]	1	2	2	0	0	2	2
	1	1	2	[2]	2	2	0	0	1	1	2	[2]	1	1	1	[1]
14	[0]	1	1	[1]	[0]	0	0	[1]	[0]	0	0	0	[0]	0	0	0
	0	1	1	1	0	0	0	1	0	0	[1]	1	1	1	0	0
	0	2	2	2	[2]	2	2	1	0	1	2	2	[2]	2	[1]	0
	0	2	2	[2]	2	2	2	1	0	1	2	[2]	2	2	1	0
18	[0]	1	2	[2]	[0]	0	1	2	[0]	1	1	0	[0]	0	0	0
	0	[1]	2	2	0	0	1	2	1	2	[2]	1	0	1	[1]	0
	0	0	1	1	[1]	1	2	2	[1]	2	2	1	1	2	[2]	1
	0	0	0	0	2	2	2	[2]	0	1	1	0	1	2	2	1
1C	[0]	0	2	2	[0]	1	1	0	[0]	0	1	1	[0]	0	0	0
	1	1	0	2	0	[1]	1	0	0	1	2	2	2	0	0	0

	0				1				2				3			
	[1]	1	0	2	2	0	0	2	0	1	[2]	2	[2]	2	1	1
	0	0	2	[2]	2	2	2	[2]	0	0	1	[1]	2	2	2	[1]
20	[0]	0	0	0	[0]	2	2	[2]	[0]	0	1	[1]	[0]	1	2	0
	0	0	0	2	0	0	2	2	0	0	1	2	0	[1]	2	0
	[1]	1	2	2	0	0	1	2	0	0	2	2	0	1	[2]	0
	1	2	2	[2]	0	0	1	[1]	0	2	2	[2]	0	1	2	0
24	[0]	0	0	0	[0]	1	2	0	[0]	1	2	0	[0]	0	1	1
	1	1	[1]	1	1	2	0	1	2	0	1	2	2	2	0	0
	2	2	[2]	2	[2]	0	[1]	2	[1]	[2]	0	1	1	1	[2]	2
	0	0	0	0	0	1	2	0	0	1	2	0	0	0	1	[1]
28	[0]	0	1	1	[0]	1	0	[1]	[0]	0	0	0	[0]	0	2	2
	1	1	[2]	2	0	1	0	1	0	0	0	0	1	[1]	2	2
	2	2	0	0	2	2	2	2	[2]	1	2	1	0	0	2	2
	0	0	1	[1]	2	2	2	[2]	2	1	2	[1]	1	1	2	[2]
2C	[0]	0	2	[2]	[0]	2	2	0	[0]	1	0	1	[0]	0	0	0
	0	0	1	1	1	2	[2]	1	2	2	[2]	2	2	1	2	1
	0	0	2	2	0	2	2	0	2	2	2	2	[2]	1	2	1
	0	0	1	[1]	1	2	2	[1]	0	1	0	[1]	2	1	2	[1]
30	[0]	1	0	[1]	[0]	2	2	[2]	[0]	0	0	2	[0]	0	0	0
	0	1	0	1	0	1	1	1	1	[1]	1	2	2	[1]	1	2
	0	1	0	1	0	2	2	2	0	0	0	2	2	1	1	2
	2	2	2	[2]	0	1	1	[1]	1	1	1	[2]	2	1	1	[2]
34	[0]	2	2	2	[0]	0	0	2	[0]	1	1	0	[0]	0	0	0
	0	[1]	1	1	1	1	1	2	0	[1]	1	0	0	0	0	0
	0	1	1	1	[1]	1	1	2	0	1	1	0	2	1	[1]	2
	0	2	2	[2]	0	0	0	[2]	2	2	2	[2]	2	1	1	[2]
38	[0]	1	1	0	[0]	0	2	2	[0]	0	2	2	[0]	0	0	0
	0	[1]	1	0	0	0	1	1	1	1	2	2	0	0	0	0
	2	2	2	2	0	0	[1]	1	[1]	1	2	2	0	0	0	0
	2	2	2	[2]	0	0	2	[2]	0	0	2	[2]	2	[1]	1	[2]
3C	[0]	0	0	[2]	[0]	2	2	2	[0]	1	0	[1]	[0]	1	1	[1]
	0	0	0	1	1	2	2	2	2	2	2	2	2	0	1	1
	0	0	0	2	0	2	2	2	2	2	2	2	[2]	2	0	1
	0	0	0	[1]	[1]	2	2	[2]	2	2	2	[2]	2	2	2	0



Video Pixel/Texel Formats

This section describes the "video" pixel/texel formats with respect to memory layout. See the Overlay chapter for a description of how the Y, U, V components are sampled.

Packed Memory Organization

Color components are all 8 bits in size for YUV formats. For YUV 4:2:2 formats each DWord will contain two pixels and only the byte order affects the memory organization.

The following four YUV 4:2:2 surface formats are supported, listed with alternate names:

- YCRCB_NORMAL (YUYV/YUY2)
- YCRCB_SWAPUVY (VYUY) (R8G8_B8G8_UNORM)
- YCRCB_SWAPUV(YVYU) (G8R8_G8B8_UNORM)
- YCRCB_SWAPY (UYVY)

The channels are mapped as follows:

Cr (V)	Red
Y	Green
Cb (U)	Blue

Memory layout of packed YUV 4:2:2 formats

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0																
V								Y								U								Y																							
Pixel N																Pixel N+1																Pixel N															

YUV 4:2:2 (Normal)

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0																
U								Y								V								Y																							
Pixel N																Pixel N+1																Pixel N															

YUV 4:2:2 (UV Swap)

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0																
Y								V								Y								U																							
Pixel N+1																Pixel N																Pixel N															

YUV 4:2:2 (Y Swap)

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0																
Y								U								Y								V																							
Pixel N+1																Pixel N																Pixel N															

YUV 4:2:2 (UV/Y Swap)



Planar Memory Organization

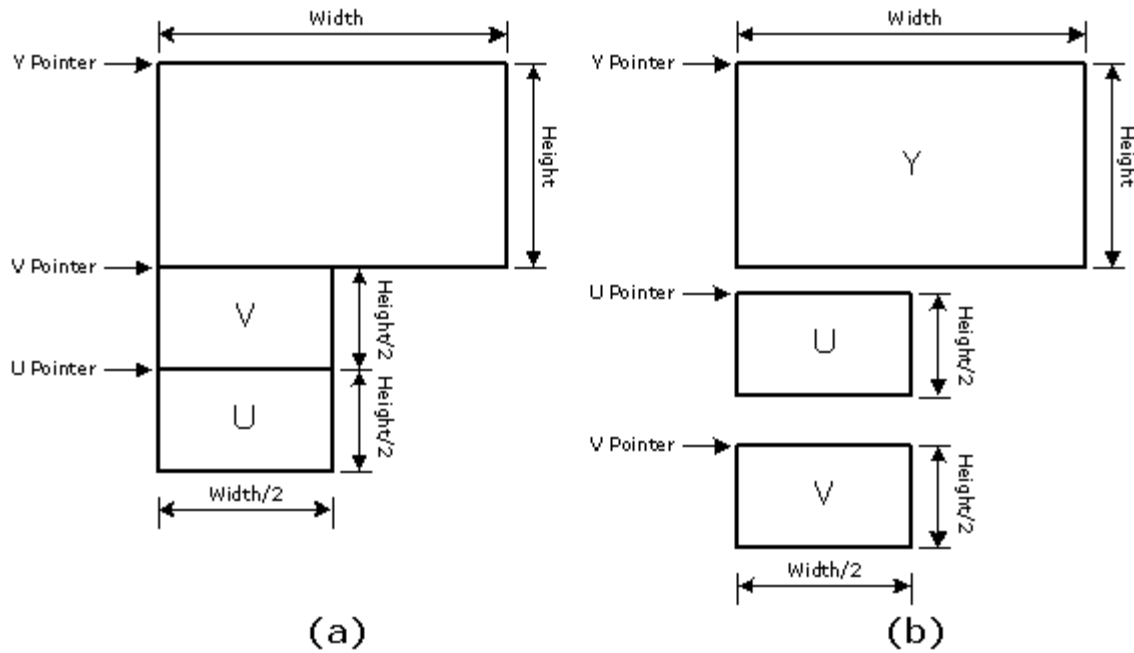
Planar formats use what could be thought of as separate buffers for the three color components. Because there is a separate stride for the Y and U/V data buffers, several memory footprints can be supported.

Note: There is no direct support for use of planar video surfaces as textures. The sampling engine can be used to operate on each of the 8bpp buffers separately (via a single-channel 8-bit format such as I8_UNORM). The U and V buffers can be written concurrently by using multiple render targets from the pixel shader. The Y buffer must be written in a separate pass due to its different size.

The following figure shows two types of memory organization for the YUV 4:2:0 planar video data:

1. The memory organization of the common YV12 data, where all three planes are contiguous and the strides of U and V components are half of that of the Y component.
2. An alternative memory structure that the addresses of the three planes are independent but satisfy certain alignment restrictions.

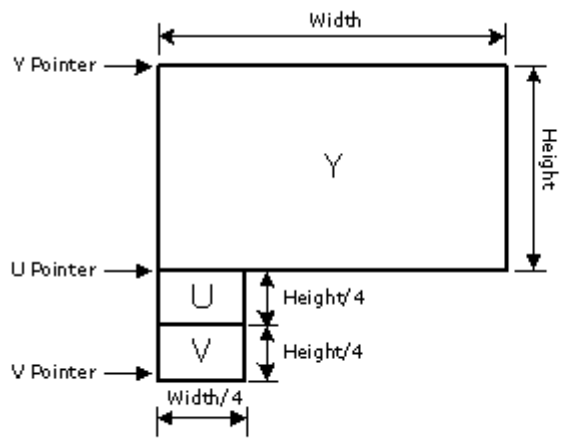
YUV 4:2:0 Format Memory Organization



B6684-01

The following figure shows memory organization of the planar YUV 4:1:0 format where the planes are contiguous.

YUV 4:1:0 Format Memory Organization



B.6685-01



Raw Format

A new surface format is added that is only supported with the untyped surface read/write and atomic operation data port messages. This new format is called simply RAW. It means that the surface has no inherent format. Surfaces of type RAW are addressed with byte-based offsets that must be DWord-aligned (multiple of 4). Data is returned in DWord quantities. The RAW surface format can be applied only to surface types of BUFFER and STRBUF.



Surface Memory Organizations

See *Memory Interface Functions* chapter for a discussion of tiled vs. linear surface formats.



Display, Overlay, Cursor Surfaces

These surfaces are memory image buffers (planes) used to refresh a display device in non-VGA mode. See the Display chapter for specifics on how these surfaces are defined/used.



2D Render Surfaces

These surfaces are used as general source and/or destination operands in 2D BLT operations.

Note that there is no coherency between 2D render surfaces and the texture cache. Software must explicitly invalidate the texture cache before using a texture that has been modified via the BLT engine.

See the 2D Instruction and 2D Rendering chapters for specifics on how these surfaces are used, restrictions on their size, placement, etc.



2D Monochrome Source

These 1 BPP (bit per pixel) surfaces are used as source operands to certain 2D BLT operations, where the BLT engine expands the 1 BPP source to the required color depth.

The texture cache stores any monochrome sources. There is no mechanism to maintain coherency between 2D render surfaces and texture-cached monochrome sources. Software must explicitly invalidate the texture cache before using a memory-based monochrome source that has been modified via the BLT engine. (Here the assumption is that SW enforces memory-based monochrome source surfaces as read-only surfaces.)

See the 2D Instruction and 2D Rendering chapters for specifics on how these surfaces are used, restrictions on their size, placement, coherency rules, etc.

2D Color Pattern

Color pattern surfaces are used as special pattern operands in 2D BLT operations.

The device uses the texture cache to store color patterns. There is no mechanism to maintain coherency between 2D render surfaces and (texture)-cached color patterns. Software is required to explicitly invalidate the texture cache before using a memory-based color pattern that has been modified via the BLT engine. (Here the assumption is that SW enforces memory-based color pattern surfaces as read-only surfaces.)

See the *2D Instruction* and *2D Rendering* chapters for specifics on how these surfaces are used, restrictions on their size, placement, etc.



3D Color Buffer (Destination) Surfaces

3D Color Buffer surfaces hold per-pixel color values for use in the 3D Pipeline. The 3D Pipeline always requires a Color Buffer to be defined.

See the Non-Video Pixel/Texel Formats section in this chapter for details on the Color Buffer pixel formats. See the 3D Instruction and 3D Rendering chapters for Color Buffer usage details.

The Color Buffer is defined as the `BUFFERID_COLOR_BACK` memory buffer via the `3DSTATE_BUFFER_INFO` instruction. That buffer can be mapped to LM, SM (snooped or unsnooped) and can be linear or tiled. When both the Depth and Color Buffers are tiled, the respective Tile Walk directions must match.

When a linear Color Buffer and a linear Depth Buffer are used together:

- The buffers may have different pitches, though both pitches must be a multiple of 32 bytes.
- The buffers must be co-aligned with a 32-byte region.

3D Depth Buffer Surfaces

Depth Buffer surfaces hold per-pixel depth values and per-pixel stencil values for use in the 3D Pipeline. The 3D Pipeline does not require a Depth Buffer in general, though a Depth Buffer is required to perform non-trivial Depth Test and Stencil Test operations.

The Depth Buffer is specified via the `3DSTATE_DEPTH_BUFFER` command. See the description of that instruction in *Windower* for restrictions.

See *Depth Buffer Formats* below for a summary of the possible depth buffer formats. See the Depth Buffer Formats section in this chapter for details on the pixel formats. See the *Windower* and *DataPort* chapters for details on the usage of the Depth Buffer.

Table: Depth Buffer Formats

DepthBufferFormat / DepthComponent	BPP (Bits Per Pixel)	Description
D32_FLOAT_S8X24_UINT	64	32-bit floating point Z depth value in first DWord, 8-bit stencil in lower byte of second DWord
D32_FLOAT	32	32-bit floating point Z depth value
D24_UNORM_S8_UINT	32	24-bit fixed point Z depth value in lower 3 bytes, 8-bit stencil value in upper byte
D16_UNORM	16	16-bit fixed point Z depth value



3D Separate Stencil Buffer Surfaces

Separate Stencil Buffer surfaces hold per-pixel stencil values for use in the 3D Pipeline. Note that the 3D Pipeline does not require a Stencil Buffer to be allocated, though a Stencil Buffer is required to perform non-trivial Stencil Test operations.

The Depth Buffer Formats table below summarizes Stencil Buffer formats. Refer to the Stencil Buffer Formats section in this chapter for details on the pixel formats. Refer to the *Windower* chapters for Stencil Buffer usage details.

The Stencil buffer is specified via the `3DSTATE_STENCIL_BUFFER` command. See that instruction description in *Windower* for restrictions.

Table: Depth Buffer Formats

DepthBufferFormat / DepthComponent	BPP (bits per pixel)	Description
R8_UNIT	8	8-bit stencil value in a byte



Surface Layout

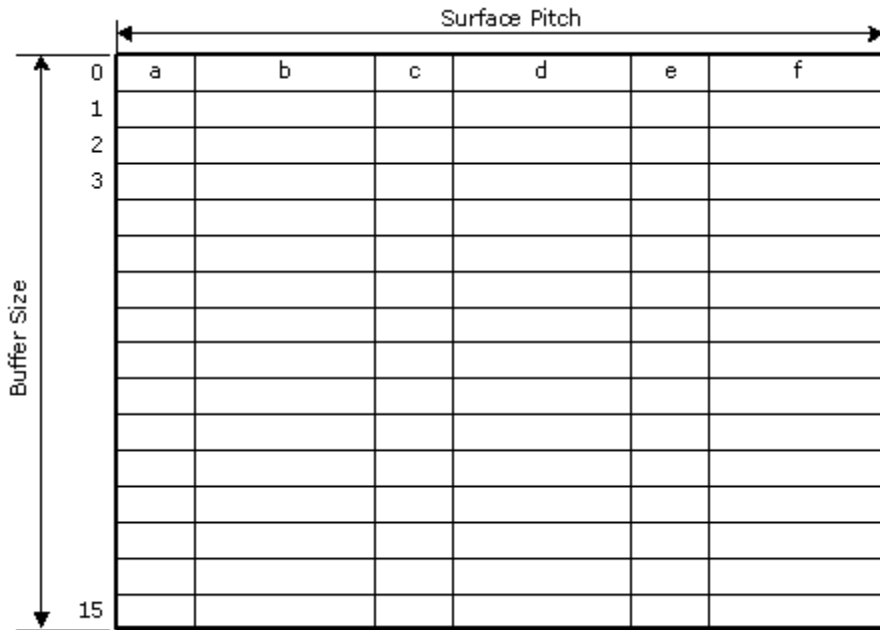
In addition to restrictions on maximum height, width, and depth, surfaces are also restricted to a maximum size in bytes. This maximum is 2 GB for all products and all surface types.



Buffers

A buffer is an array of structures. Each structure contains up to 2048 bytes of elements. Each element is a single surface format using one of the supported surface formats depending on how the surface is being accessed. The surface pitch state for the surface specifies the size of each structure in bytes.

The buffer is stored in memory contiguously with each element in the structure packed together, and the first element in the next structure immediately following the last element of the previous structure. Buffers are supported only in linear memory.



B.6686-01

Structured Buffers

A structured buffer is a surface type that is accessed by a 2-dimensional coordinate. It can be thought of as an array of structures, where each structure is a predefined number of DWords in size. The first coordinate (U) defines the array index, and the second coordinate (V) is a byte offset into the structure which must be a multiple of 4 (DWord-aligned). A structured buffer must be defined with **Surface Format RAW**.

The structured buffer has only one dimension programmed in SURFACE_STATE which indicates the array size. The byte offset dimension (V) is assumed to be bounded only by the **Surface Pitch**.



1D Surfaces

One-dimensional surfaces are identical to 2D surfaces with height of one. Arrays of 1D surfaces are also supported. Please refer to the 2D Surfaces section for details on how these surfaces are stored.

2D Surfaces

Surfaces that comprise texture mip-maps are stored in a fixed "monolithic" format and referenced by a single base address. The base map and associated mipmaps are located within a single rectangular area of memory identified by the base address of the upper left corner and a pitch. The base address references the upper left corner of the base map. The pitch must be specified at least as large as the widest mip-map. In some cases it must be wider; see the section on Minimum Pitch below.

These surfaces may be overlapped in memory and must adhere to the following memory organization rules:

- For non-compressed texture formats, each mipmap must start on an even row within the monolithic rectangular area. For 1-texel-high mipmaps, this may require a row of padding below the previous mipmap. This restriction does not apply to any compressed texture formats; each subsequent (lower-res) compressed mipmap is positioned directly below the previous mipmap.
- Vertical alignment restrictions vary with memory tiling type: 1 DWord for linear, 16-byte (DQWord) for tiled. (Note that tiled mipmaps are *not* required to start at the left edge of a tile row.)

Computing MIP Level Sizes

Map width and height specify the size of the largest MIP level (LOD 0). Less detailed LOD level (i+1) sizes are determined by dividing the width and height of the current (i) LOD level by 2 and truncating to an integer (floor). This is equivalent to shifting the width/height by 1 bit to the right and discarding the bit shifted off. The map height and width are clamped on the low side at 1.

In equations, the width and height of an LOD "L" can be expressed as:

$$W_L = ((width \gg L) > 0 ? width \gg L : 1)$$

$$H_L = ((height \gg L) > 0 ? height \gg L : 1)$$

[HSW]: If the surface is multisampled and it is a depth or stencil surface or **Multisampled Surface StorageFormat** in SURFACE_STATE is MSFMT_DEPTH_STENCIL, W_L and H_L must be adjusted as follows before proceeding:

Number of Multisamples	$W_L =$	$H_L =$
2	$\text{ceiling}(W_L / 2) * 4$	H_L [no adjustment]
4	$\text{ceiling}(W_L / 2) * 4$	$\text{ceiling}(H_L / 2) * 4$
8	$\text{ceiling}(W_L / 2) * 8$	$\text{ceiling}(H_L / 2) * 4d$
16	$\text{ceiling}(W_L / 2) * 8$	$\text{ceiling}(H_L / 2) * 8$



Base Address for LOD Calculation

It is conceptually easier to think of the space that the map uses in Cartesian space (x, y) , where x and y are in units of texels, with the upper left corner of the base map at $(0, 0)$. The final step is to convert from Cartesian coordinates to linear addresses as documented at the bottom of this section.

It is useful to think of the concept of "stepping" when considering where the next MIP level will be stored in rectangular memory. We either step down or step right when moving to the next higher LOD.

- for MIPLAYOUT_RIGHT maps:
 - step right when moving from LOD 0 to LOD 1
 - step down for all of the other MIPs
- for MIPLAYOUT_BELOW maps:
 - step down when moving from LOD 0 to LOD 1
 - step right when moving from LOD 1 to LOD 2
 - step down for all of the other MIPs

To account for the cache line alignment required, we define i and j as the width and height, respectively, of an *alignment unit*. This alignment unit is defined below. We then define lower-case w_L and h_L as the padded width and height of LOD "L" as follows:

$$w_L = i * \text{ceil}\left(\frac{W_L}{i}\right)$$

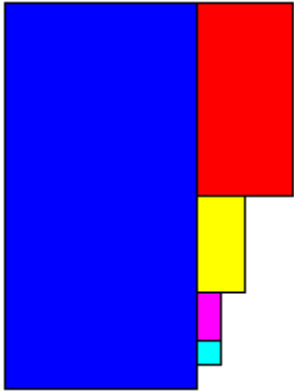
$$h_L = j * \text{ceil}\left(\frac{H_L}{j}\right)$$

For separate stencil buffer, the width must be multiplied by 2 and height divided by 2 as follows:

$$w_s = 2 * i * \text{ceil}\left(\frac{W_L}{i}\right)$$

$$h_s = 1/2 * j * \text{ceil}\left(\frac{H_L}{j}\right)$$

Equations to compute the upper left corner of each MIP level are then as follows:



for *MIPLAYOUT_RIGHT* maps:

$$LOD_0 = (0,0)$$

$$LOD_1 = (w_0,0)$$

$$LOD_2 = (w_0,h_1)$$

$$LOD_3 = (w_0,h_1 + h_2)$$

$$LOD_4 = (w_0,h_1 + h_2 + h_3)$$

...

for *MIPLAYOUT_BELOW* maps:

$$LOD_0 = (0,0)$$

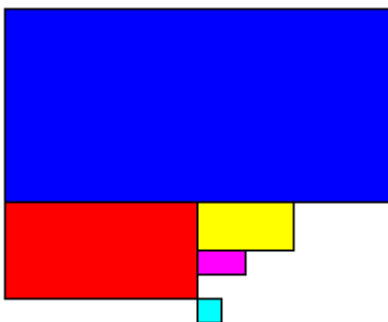
$$LOD_1 = (0,h_0)$$

$$LOD_2 = (w_1,h_0)$$

$$LOD_3 = (w_1,h_0 + h_2)$$

$$LOD_4 = (w_1,h_0 + h_2 + h_3)$$

...





Minimum Pitch for MIPLAYOUT_RIGHT and Other Maps

For MIPLAYOUT_RIGHT maps, the minimum pitch must be calculated before choosing a fence to place the map within. This is approximately equal to 1.5x the pitch required by the base map, with possible adjustments made for cache line alignment. For MIPLAYOUT_BELOW and MIPLAYOUT_LEGACY maps, the minimum pitch required is equal to that required by the base (LOD 0) map.

A safe but simple calculation of minimum pitch is equal to 2x the pitch required by the base map for MIPLAYOUT_RIGHT maps. This ensures that enough pitch is available, and since it is restricted to MIPLAYOUT_RIGHT maps, not much memory is wasted. It is up to the driver (hardware independent) whether to use this simple determination of pitch or a more complex one.

Alignment Unit Size

This section documents the alignment parameters *i* and *j* that are used depending on the surface.

Surface Defined By	Surface Format	Alignment Unit Width "i"	Alignment Unit Height "j"
3DSTATE_DEPTH_BUFFER	D16_UNORM	8	4
	not D16_UNORM	4	4
3DSTATE_STENCIL_BUFFER	N/A	8	8
SURFACE_STATE	BC*, ETC*, EAC*	4	4
	FXT1	8	4
	all others	set by Surface Horizontal Alignment	set by Surface Vertical Alignment

Cartesian to Linear Address Conversion

A set of variables are defined in addition to the *i* and *j* defined above.

- *b* = bytes per texel of the native map format (0.5 for DXT1, FXT1, and 4-bit surface format, 2.0 for YUV 4:2:2, others aligned to surface format)
- *t* = texel rows / memory row (4 for DXT1-5 and FXT1, 1 for all other formats)
- *p* = pitch in bytes (equal to pitch in dwords * 4)
- *B* = base address in bytes (address of texel 0,0 of the base map)
- *x*, *y* = cartesian coordinates from the above calculations in units of texels (assumed that *x* is always a multiple of *i* and *y* is a multiple of *j*)
- *A* = linear address in bytes

$$A = B + \frac{yp}{t} + xbt$$

This calculation gives the linear address in bytes for a given MIP level (taking into account L1 cache line alignment requirements).

Compressed Mipmap Layout

Mipmaps of textures using compressed (DXTn, FXT) texel formats are also stored in a monolithic format. The compressed mipmaps are stored in a similar fashion to uncompressed mipmaps, with each block of source (uncompressed) texels represented by a 1 or 2 QWord compressed block. The compressed blocks occupy the same logical positions as the texels they represent, where each row of compressed blocks represent a 4-high row of uncompressed texels. The format of the blocks is preserved, i.e., there is no "intermediate" format as required on some other devices.

The following exceptions apply to the layout of compressed (vs. uncompressed) mipmaps:

- Mipmaps are not required to start on even rows, therefore each successive mip level is located on the texel row immediately below the last row of the previous mip level. Pad rows are neither required nor allowed.
- The dimensions of the mip maps are first determined by applying the sizing algorithm presented in Non-Power-of-Two Mipmaps above. Then, if necessary, they are padded out to compression block boundaries.

Surface Arrays

Arrays of 1D and 2D surfaces can be treated as a single surface. This section covers the layout of these composite surfaces.

For All Surface Other Than Separate Stencil Buffer

Both 1D and 2D surfaces can be specified as an array. The only difference in the surface state is the presence of a depth value greater than one, indicating multiple array "slices".

A value $QPitch$ is defined which indicates the worst-case height for one slice in the texture array. This $QPitch$ is multiplied by the array index to and added to the vertical component of the address to determine the vertical component of the address for that slice. Within the slice, the map is stored identically to a `MIPLAYOUT_BELOW` 2D surface. *MIPLAYOUT_BELOW is the only format supported by 1D non-arrays and both 2D and 1D arrays, the programming of the MIP Map Layout Mode state variable is ignored when using a TextureArray.*

The following equation is used for surface formats other than compressed textures:

$$QPitch = (h_0 + h_1 + 11j)$$

The input variables in this equation are defined in sections above.

The equation for compressed textures (BC* and FXT1 surface formats) follows:

$$QPitch = \frac{(h_0 + h_1 + 11j)}{4}$$



Multisampled Surfaces

Multisampled render targets and sampling engine surfaces are supported. There are three types of multisampled surface layouts designated as follows:

- **IMS** Interleaved Multisampled Surface
- **CMS** Compressed Multisampled Surface
- **UMS** Uncompressed Multisampled Surface

These surface layouts are described in the following sections.

Compressed Multisampled Surfaces

Multisampled render targets can be compressed. If **MCS Enable** is enabled in SURFACE_STATE, hardware handles the compression using a software-invisible algorithm. However, performance optimizations in the multisample resolve kernel using the sampling engine are possible if the internal format of these surfaces is understood by software. This section documents the formats of the Multisample Control Surface (MCS) and Multisample Surface (MSS).

The MCS surface consists of one element per pixel, with the element size being an 8-bit unsigned integer value for 4x multisampled surfaces and a 32-bit unsigned integer value for 8x multisampled surfaces. Each field within the element indicates which sample slice (SS) the sample resides on.

The 4x MCS is 8 bits per pixel. The 8 bits are encoded as follows:

4x MCS [HSW]

7:6	5:4	3:2	1:0
sample 3 SS	sample 2 SS	sample 1 SS	sample 0 SS

Each 2-bit field indicates which sample slice (SS) the sample's color value is stored. An MCS value of 0x00 indicates that all four samples are stored in sample slice 0 (thus all have the same color). This is the fully compressed case. An MCS value of 0xff indicates that all samples in the pixel are in the clear state, and none of the sample slices are valid. The pixel's color must be replaced with the surface's clear value.

Extending the mechanism used for the 4x MCS to 8x requires 3 bits per sample times 8 samples, or 24 bits per pixel. The 24-bit MCS value per pixel is placed in a 32-bit footprint, with the upper 8 bits unused as shown below.

8x MCS [HSW]

31:24	23:21	20:18	17:15	14:12	11:9	8:6	5:3	2:0
reserved	sample 7 SS	sample 6 SS	sample 5 SS	sample 4 SS	sample 3 SS	sample 2 SS	sample 1 SS	sample 0 SS

Other than this, the 8x algorithm is the same as the 4x algorithm. The MCS value indicating clear state is 0x00ffffff.



Physical MSS Surface

The physical MSS surface is stored identically to a 2D array surface, with the height and width matching the *pixel* dimensions of the logical multisampled surface. The number of array slices in the physical surface is 2, 4, 8, or 16 times that of the logical surface (depending on the number of multisamples). Sample slices belonging to the same logical surface array slice are stored in adjacent physical slices. The sampling engine *ld2dss* message gives direct access to a specific sample slice.

Uncompressed Multisampled Surfaces

UMS surfaces similar to CMS, except that the MCS is disabled, and there is no MCS surface. UMS contains only an MSS surface, where each sample is stored on its sample slice (SS) of the same index.



Cube Surfaces

The 3D Pipeline supports *cubic environment maps*, conceptually arranged as a cube surrounding the origin of a 3D coordinate system aligned to the cube faces. These maps can be used to supply texel (color/alpha) data of the environment in any direction from the enclosed origin, where the direction is supplied as a 3D "vector" texture coordinate. These cube maps can also be mipmapped.

Each texture map level is represented as a group of six, square *cube face* texture surfaces. The faces are identified by their relationship to the 3D texture coordinate system. The subsections below describe the cube maps as described at the API as well as the memory layout dictated by the hardware.

Hardware Cube Map Layout

The cube face textures are stored in the same way as 2D array surfaces are stored (see section *2D Surfaces* for details). For cube surfaces, the depth (array instances) is equal to 6. The array index "q" corresponds to the face according to the following table:

"q" coordinate	face
0	+x
1	-x
2	+y
3	-y
4	+z
5	-z

Restrictions

- The cube map memory layout is the same whether or not the cube map is mip-mapped, and whether or not all six faces are "enabled", though the memory backing disabled faces or non-supplied levels can be used by software for other purposes.
- The cube map faces all share the same **Surface Format**

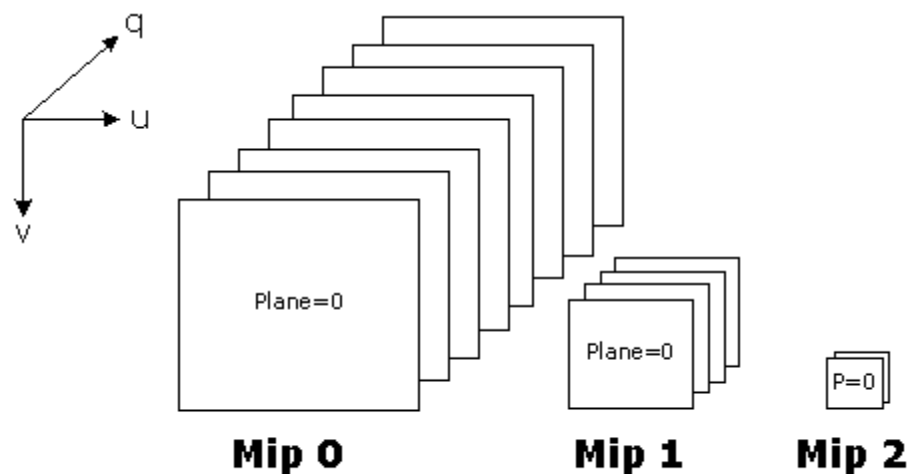
Cube Arrays

Cube arrays are stored identically to 2D surface arrays. A group of 6 consecutive array elements makes up a single cube map. A cube array with N array elements is stored identically to a 2D array with 6N array elements.

3D Surfaces

Multiple texture map surfaces (and their respective mipmap chains) can be arranged into a structure known as a Texture3D (volume) texture. A volume texture map consists of many *planes* of 2D texture maps. See *Sampler* for a description of how volume textures are used.

Volume Texture Map



B.6688-01

Note that the number of planes defined at each successive mip level is halved. Volumetric texture maps are stored as follows. All of the LOD=0 q-planes are stacked vertically, then below that, the LOD=1 q-planes are stacked two-wide, then the LOD=2 q-planes are stacked four-wide below that, and so on.

The width, height, and depth of LOD "L" are as follows:

$$W_L = ((width \gg L) > 0 ? width \gg L : 1)$$

$$H_L = ((height \gg L) > 0 ? height \gg L : 1)$$

This is the same as for a regular texture. For volume textures we add:

$$D_L = ((depth \gg L) > 0 ? depth \gg L : 1)$$

Cache-line aligned width and height are as follows, with *i* and *j* being a function of the map format as shown in [Alignment Unit Size](#).



LOD 0 (Mip 0)		
q=0		
q=1		
q=2		
q=3		
q=4		
q=5		
q=6		
q=7		
LOD 1 (Mip 1)	q=0	q=1
	q=2	q=3
LOD 2 (Mip 2)	q=0	q=1
LOD 3 (Mip 3)	q=0	

$$w_L = i * \text{ceil}\left(\frac{W_L}{i}\right)$$

$$h_L = j * \text{ceil}\left(\frac{H_L}{j}\right)$$

Note that it is not necessary to cache-line align in the "depth" dimension (i.e. lowercase "d").

The following equations for $LOD_{L,q}$ give the base address Cartesian coordinates for the map at LOD L and depth q.

$$LOD_{0,q} = (0, q * h_0)$$

$$LOD_{1,q} = ((q \gg 2) * w_1, D_0 * h_0 + (q \gg 1) * h_1)$$

$$LOD_{2,q} = ((q \gg 4) * w_2, D_0 * h_0 + \text{ceil}\left(\frac{D_1}{2}\right) * h_1 + (q \gg 2) * h_2)$$

$$LOD_{3,q} = ((q \gg 8) * w_3, D_0 * h_0 + \text{ceil}\left(\frac{D_1}{2}\right) * h_1 + \text{ceil}\left(\frac{D_2}{4}\right) * h_2 + (q \gg 3) * h_3)$$

...

These values are then used as "base addresses" and the 2D MIP Map equations are used to compute the location within each LOD/q map.

Minimum Pitch

The minimum pitch required to store the 3D map may in some cases be greater than the minimum pitch required by the LOD=0 map. This is due to cache line alignment requirements that may impact some of the MIP levels requiring additional spacing in the horizontal direction.



Surface Padding Requirements

This section covers the requirements for padding around surfaces stored in memory, as there are cases where the device will overfetch beyond the bounds of the surface due to implementation of caches and other hardware structures.

Sampling Engine Surfaces

The sampling engine accesses texels outside of the surface if they are contained in the same cache line as texels that are within the surface. These texels will not participate in any calculation performed by the sampling engine and will not affect the result of any sampling engine operation, however if these texels lie outside of defined pages in the GTT, a GTT error will result when the cache line is accessed. In order to avoid these GTT errors, "padding" at the bottom and right side of a sampling engine surface is sometimes necessary.

It is possible that a cache line will straddle a page boundary if the base address or pitch is not aligned. All pages included in the cache lines that are part of the surface must map to valid GTT entries to avoid errors. To determine the necessary padding on the bottom and right side of the surface, refer to the table in Section *Alignment Unit Size* for the i and j parameters for the surface format in use. The surface must then be extended to the next multiple of the alignment unit size in each dimension, and all texels contained in this extended surface must have valid GTT entries.

For example, suppose the surface size is 15 texels by 10 texels and the alignment parameters are $i=4$ and $j=2$. In this case, the extended surface would be 16 by 10. Note that these calculations are done in texels, and must be converted to bytes based on the surface format being used to determine whether additional pages need to be defined.

[HSW]: For buffers, which have no inherent "height," padding requirements are different. A buffer must be padded to the next multiple of 256 array elements, with an additional 16 bytes added beyond that to account for the L1 cache line.

For cube surfaces, an additional two rows of padding are required at the bottom of the surface. This must be ensured regardless of whether the surface is stored tiled or linear. This is due to the potential rotation of cache line orientation from memory to cache.

For compressed textures (BC*, FXT1, ETC*, EAC*), padding at the bottom of the surface is to an even compressed row. This is equivalent to a multiple of $2q$, where q is the compression block height in texels. Thus, for padding purposes, these surfaces behave as if $j = 2q$ only for surface padding purposes. The value of j is still equal to q for mip level alignment and QPitch calculation.

For packed YUV, 96 bpt, 48 bpt, and 24 bpt surface formats, additional padding is required. These surfaces require an extra row plus 16 bytes of padding at the bottom in addition to the general padding requirements.

For linear surfaces, additional padding of 64 bytes is required at the bottom of the surface. This is in addition to the padding required above.



Render Target and Media Surfaces

The data port accesses data (pixels) outside of the surface if they are contained in the same cache request as pixels that are within the surface. These pixels will not be returned by the requesting message, however if these pixels lie outside of defined pages in the GTT, a GTT error will result when the cache request is processed. In order to avoid these GTT errors, "padding" at the bottom of the surface is sometimes necessary.

If the surface contains an odd number of rows of data, a final row below the surface must be allocated. If the surface will be accessed in field mode (**Vertical Stride** = 1), enough additional rows below the surface must be allocated to make the extended surface height (including the padding) a multiple of 4.