

UG10159

i.MX Graphics User's Guide

Rev. 10.2 — 31 March 2025

User guide

Document information

Information	Content
Keywords	i.MX, Linux, Android, Graphics, UG10159
Abstract	The purpose of this document is to provide information on graphic APIs and driver support.



1 Introduction

The purpose of this document is to provide information on graphic APIs and driver support. Each chapter describes a specific set of APIs or driver integration as well as specific hardware acceleration customization. The target audiences for this document are developers writing graphics applications or video drivers.

1.1 i.MX full GPU line

The whole family of GPUs are listed in the following table. On i.MX 6 boards, only 6Quad and 6QuadPlus support OpenCL. The theoretical number of GFLOPS, the key performance indicator of OpenCL, is also shown in the table. Some benchmarks such as Clpeak, can be used to verify it.

i.MX 8QuadMax supports OpenVX, which will be introduced in later chapter.

Product	i.MX 6SoloX	i.MX 6Solo 6DualLite	i.MX 6Quad	i.MX 6DualPlus 6QuadPlus	i.MX 7ULP	i.MX 8ULP	i.MX 8M Mini	i.MX 8M Nano	i.MX 8X 8DualXPlus 8QuadXPlus	i.MX 8M Quad, Dual QuadLite	i.MX 8M Plus	i.MX 8 8QuadMax	i.MX 9 i.MX 95
GPU 2D	GC400T (2D)	GC320	GC355 (VG) GC320	GC355 (VG) GC328	GC328	GC520L	GC520L	N/A	High Perf 2D Blit Engine	N/A	GC520L	High Perf 2D Blit Engine	High Perf 2D Blit Engine
GPU 3D	GC400T (3D)	GC880	GC2000	GC2000+	GC700 NanoUltra	GC7000 NanoUltra31	GC7000 NanoUltra	GC7000 UltraLite	GC7000 Lite	GC7000 Lite	GC7000 UltraLite	x2 GC7000 XSVX	G310 V2
# Shaders (Vec4)	1	1	4	4	1	1	1	2	4	4	2	8 + 8	1
Clock (MHz) Core [Shader]	360 [720]	264 [528]	528 [594]	594 [720]	400 [400]	317 [317]	1000	500 [600]	700 [850]	800 [800]	1000[1000]	800 [1000]	1000
Pixel Rate (Mpix/s)	180	264	1056	1188	200	296	500	500	1400	1600	1000	1600 + 1600 (dual) 3200 (bridged)	4000
Geom. Rate (MTris/s)	36	81	176	198	40	52	50	83	234	267	166	267 + 267 (dual) 267 (bridged)	400
GFLOPS(Theoretical) Med/High Precision	2.9 (high)	4.2 (high)	19 (high)	46 / 23	3.2/1.6	4.8/2.4	16/8	19.2/9.6	55.2 / 27.6	51.2 / 25.6	32/16	256 / 128	120/60
2D API	OpenVG 1.1 [†] , G2D	OpenVG 1.1 [†] , G2D	OpenVG 1.1 G2D	OpenVG 1.1, G2D	OpenVG 1.1 [†] , G2D	OpenVG 1.1 [†] , G2D	OpenVG 1.1 [†] , G2D	OpenVG 1.1 [†] , G2D	OpenVG 1.1 [†] , G2D	OpenVG 1.1 [†]	OpenVG 1.1 [†] , G2D	OpenVG 1.1 [†] , G2D	G2D
3D API	OGL ES 2.0	OGL ES 3.0	OGL ES 3.0	OGL ES 3.0	OGL ES 2.0	OGL ES 3.1 Vulkan	OGL ES 2.0	OGL ES 3.1, Vulkan	OGL ES 3.1, Vulkan	OGL ES 3.1, Vulkan	OGL ES 3.1, Vulkan	OGL ES 3.2, Vulkan	OGL ES 3.2, Vulkan
Compute	N/A	N/A	OCL 1.2 EP	OCL 1.2 FP	N/A	OCL 3.0	N/A	OCL 3.0	OCL 3.0	OCL 3.0	OCL 3.0	OCL 3.0	OCL 3.0
Other	2D / 3D Multithreaded	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	OpenVX 1.2 (NPU)	OpenVX 1.2	No

Figure 1. GPU Scalability across i.MX processors

Note: † OpenVG on 3D GPU with software tessellation.

2 i.MX G2D API

2.1 Overview

The G2D Application Programming Interface (API) is designed to be easy to understand and to use the 2D Bit blit (BLT) function. It allows the user to implement the customized applications with simple interfaces. It is hardware and platform independent for i.MX 2D Graphics.

G2D API supports the following features but is not limited to these:

- Simple BLT operation from source to destination
- 16/32bit RGB(alpha) and YUV color format conversions
- Alpha blending for source and destination with Porter-Duff rules
- High-performance memory copy from source to destination
- Up-scaling and down-scaling from source to destination
- 90/180/270 degrees rotation from source to destination

- Horizontal and vertical flip from source to destination
- Enhanced visual quality with dither for pixel precision-loss (*)
- High performance memory clear for destination
- Pixel-level cropping for source surface
- Global alpha blending for source only
- Asynchronous mode and sync
- Contiguous memory allocator
- Support cacheable memory (*)
- Support VG engine (*)
- Multi source blit (*)

Note: The features with (*) are available on specific devices. Applications can query G2D for available features.

The G2D API document includes a detailed interface description and sample code for reference.

The API is designed with C-Style coding and can be used in both C and C++ applications.

2.2 Enumerations and structures

This chapter describes all enumerations and structure definitions in G2D.

2.2.1 g2d_format enumeration

This enumeration describes the pixel format for source and destination.

Table 1. g2d_format enumeration

Name	Numeric	Description
G2D_RGB565	0	RGB565 pixel format
G2D_RGBA8888	1	32-bit RGBA pixel format
G2D_RGBX8888	2	32-bit RGBX without alpha blending
G2D_BGRA8888	3	32-bit BGRA pixel format
G2D_BGRX8888	4	32-bit BGRX without alpha blending
G2D_BGR565	5	16-bit BGR565 pixel format
G2D_ARGB8888	6	32-bit ARGB pixel format
G2D_ABGR8888	7	32-bit ABGR pixel format
G2D_XRGB8888	8	32-bit XRGB without alpha
G2D_XBGR8888	9	32-bit XBGR without alpha
G2D_RGB888	10	24-bit RGB
G2D_BGR888	11	24-bit BGR
G2D_RGBA5551	12	16-bit RGBA5551 pixel format
G2D_RGBX5551	13	16-bit RGBX5551 without alpha
G2D_BGRA5551	14	16-bit BGRA5551 pixel format
G2D_BGRX5551	15	16-bit BGRX5551 without alpha
G2D_RGBA1010102	16	16-bit RGBA1010102 pixel format
G2D_GRAY8	19	8-bit GRAY8 pixel format
G2D_NV12	20	Y plane followed by interleaved U/V plane

Table 1. g2d_format enumeration...continued

Name	Numeric	Description
G2D_I420	21	Y, U, V are within separate planes
G2D_YV12	22	Y, V, U are within separate planes
G2D_NV21	23	Y plane followed by interleaved V/U plane
G2D_YUYV	24	Interleaved Y/U/Y/V plane
G2D_YVYU	25	Interleaved Y/V/Y/U plane
G2D_UYVY	26	Interleaved U/Y/V/Y plane
G2D_VYUY	27	Interleaved V/Y/U/Y plane
G2D_NV16	28	Y plane followed by interleaved U/V plane
G2D_NV61	29	Y plane followed by interleaved V/U plane

2.2.2 g2d_blend_func enumeration

This enumeration describes the blend factor for source and destination.

Table 2. g2d_blend_func enumeration

Name	Numeric	Description
G2D_ZERO	0	Blend factor with 0
G2D_ONE	1	Blend factor with 1
G2D_SRC_ALPHA	2	Blend factor with source alpha
G2D_ONE_MINUS_SRC_ALPHA	3	Blend factor with 1 - source alpha
G2D_DST_ALPHA	4	Blend factor with destination alpha
G2D_ONE_MINUS_DST_ALPHA	5	Blend factor with 1 - destination alpha
G2D_PRE_MULTIPLIED_ALPHA	0x10	Extensive blend as pre-multiplied alpha
G2D_DEMULTIPLY_OUT_ALPHA	0x20	Extensive blend as demultiply out alpha

2.2.3 g2d_cap_mode enumeration

This enumeration describes the alternative capability in 2D BLT.

Table 3. g2d_cap_mode enumeration

Name	Numeric	Description
G2D_BLEND	0	Enable alpha blend in 2D BLT
G2D_DITHER	1	Enable dither in 2D BLT
G2D_GLOBAL_ALPHA	2	Enable global alpha in blend
G2D_BLEND_DIM	3	Enable blend dim effect
G2D_BLUR	4	Enable blur effect
G2D_YUY_BT_601	5	Enable YUV BT.601 mode
G2D_YUY_BT_709	6	Enable YUV BT.709 mode
G2D_YUY_BT_601FR	7	Enable YUV BT.601 full range mode
G2D_YUY_BT_709FR	8	Enable YUV BT.709 full range mode

Table 3. g2d_cap_mode enumeration...continued

Name	Numeric	Description
G2D_WARPING	9	Enable Warp/Dewarp

Note: G2D_GLOBAL_ALPHA is only valid when G2D_BLEND is enabled.

2.2.4 g2d_rotation enumeration

This enumeration describes the rotation mode in 2D BLT.

Table 4. g2d_rotation enumeration

Name	Numeric	Description
G2D_ROTATION_0	0	No rotation
G2D_ROTATION_90	1	Rotation with 90 degrees
G2D_ROTATION_180	2	Rotation with 180 degrees
G2D_ROTATION_270	3	Rotation with 270 degrees
G2D_FLIP_H	4	Horizontal flip
G2D_FLIP_V	5	Vertical flip

2.2.5 g2d_cache_mode enumeration

This enumeration describes the cache operation mode.

Table 5. g2d_cache_mode enumeration

Name	Numeric	Description
G2D_CACHE_CLEAN	0	Clean the cacheable buffer
G2D_CACHE_FLUSH	1	Clean and invalidate cacheable buffer
G2D_CACHE_INVALIDATE	2	Invalidate the cacheable buffer

2.2.6 g2d_hardware_type enumeration

This enumeration describes the supported hardware type.

Table 6. g2d_hardware_type enumeration

Name	Numeric	Description
G2D_HARDWARE_2D	0	GPU 2D hardware type
G2D_HARDWARE_VG	1	GPU VG hardware type
G2D_HARDWARE_DPU_V1	2	DPU V1 hardware type
G2D_HARDWARE_DPU_V2	3	DPU V2 hardware type
G2D_HARDWARE_PXP	4	PXP hardware type

2.2.7 g2d_surface structure

This structure describes the surface with operation attributes.

Table 7. g2d_surface structure

g2d_surface member	Type	Description
format	g2d_format	Pixel format of surface buffer
planes[3]	unsigned int	Physical addresses of surface buffer
left	Int	Left offset in blit rectangle
top	Int	Top offset in blit rectangle
right	Int	Right offset in blit rectangle
bottom	Int	Bottom offset in blit rectangle
stride	Int	RGB/Y stride of surface buffer
width	Int	Surface width in pixel unit
height	Int	Surface height in pixel unit
blendfunc	g2d_blend_func	Alpha blend mode
global_alpha	Int	Global alpha value 0~255
clrcolor	Int	Clear color is 32bit RGBA
rot	g2d_rotation	Rotation mode

Note: RGB and YUV formats conversion, Y(*) means feature available on i.MX 6Quad Plus, i.MX 7ULP and i.MX 8 family devices.

<div> <div>DST</div> <div>SRC</div> </div>	G2D_RGBs	G2D_YV12	G2D_I420	G2D_NV12	G2D_NV21	G2D_YUYV	G2D_NV16	G2D_NV61
G2D_RGBs	Y	N	N	N	N	Y(*)	N	N
G2D_NV12	Y	N	N	N	N	Y(*)	N	N
G2D_I420	Y	N	N	N	N	Y(*)	N	N
G2D_YV12	Y	N	N	N	N	Y(*)	N	N
G2D_NV21	Y	N	N	N	N	Y(*)	N	N
G2D_YUYV	Y	N	N	Y(*)	Y(*)	Y(*)	Y(*)	Y(*)
G2D_VYU	Y	N	N	N	N	Y(*)	N	N
G2D_UYVY	Y	N	N	N	N	Y(*)	N	N
G2D_VYUY	Y	N	N	N	N	Y(*)	N	N
G2D_NV16	Y	N	N	N	N	Y(*)	N	N
G2D_NV61	Y	N	N	N	N	Y(*)	N	N

- RGB pixel buffer only uses planes [0], buffer address is with 16 bytes alignment on i.MX 6 (except i.MX 6Quad Plus), 1 pixel alignment on i.MX 6Quad Plus, i.MX 7ULP and i.MX 8 family devices.
- NV12: Y in planes [0], UV in planes [1], with 64bytes alignment,
- I420: Y in planes [0], U in planes [1], V in planes [2], with 64 bytes alignment
- The cropped region in source surface is specified with left, top, right and bottom parameters.
- RGB stride alignment is 16 bytes on i.MX 6 (except i.MX 6Quad Plus), 1 pixel alignment on i.MX 6Quad Plus, i.MX 7ULP and i.MX 8 family devices, both for source and destination surface.
- NV12 stride alignment is 8 bytes for source surface, UV stride = Y stride,
- I420 stride alignment is 8 bytes for source surface, U stride=V stride = ½ Y stride.
- G2D_ROTATION_0/G2D_FLIP_H/G2D_FLIP_V shall be set in source surface, and the clockwise rotation degree shall be set in destination surface.
- Application should calculate the rotated position and set it for destination surface.
- The geometry definition of surface structure is described as follows.

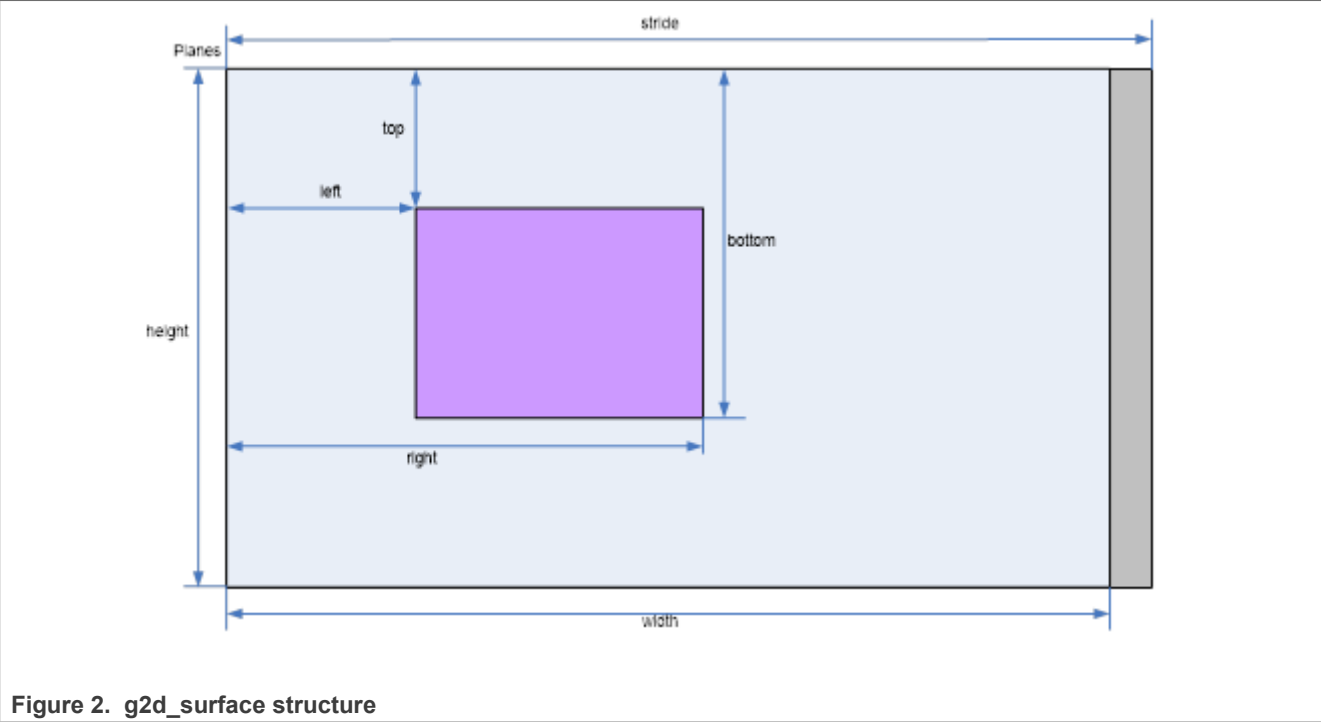


Figure 2. g2d_surface structure

2.2.8 g2d_buf structure

This structure describes the buffer used as G2D interfaces.

Table 8. g2d_buf structure

g2d_buf member	Type	Description
buf_handle	void *	The handle associated with buffer
buf_vaddr	void *	Virtual address of the buffer
buf_paddr	unsigned int	Physical address of the buffer
buf_size	int	The actual size of the buffer

2.2.9 g2d_surface_pair structure

This structure binds one source `g2d_surface` and one destination `g2d_surface` as a pair. When doing multi-source blit, they are one-to-one correspondent.

Table 9. g2d_surface_pair structure

g2d_surface_pair member	Type	Description
s	g2d_surface	Source g2d_surface
d	g2d_surface	Destination g2d_surface

2.2.10 g2d_feature enumeration

This enumeration describes the features in G2D BLT.

Table 10. g2d_feature enumeration

Name	Numeric	Description
G2D_SCALING	0	Scaling
G2D_ROTATION	1	Rotation
G2D_SRC_YUV	2	Source YUV format
G2D_DST_YUV	3	Destination YUV format
G2D_MULTI_SOURCE_BLT	4	Multisource blit
G2D_FAST_CLEAR	5	Support fast clear blit
G2D_WARP_DEWARP	6	Warp/Dewarp

2.2.11 g2d_tiling enumeration

This enumeration describes the tiling format for source and destination.

Table 11. g2d_tiling enumeration

Name	Numeric	Description
G2D_LINEAR	0x1	LINEAR tiling format
G2D_TILED	0x2	TILED tiling format
G2D_SUPERTILED	0x4	SUPERTILED tiling format
G2D_AMPHION_TILED	0x8	AMPHION_TILED tiling format
G2D_AMPHION_INTERLACED	0x10	AMPHION_INTERLACED tiling format
G2D_TILED_STATUS	0x20	TILED_STATUS tiling format
G2D_AMPHION_TILED_10BIT	0x40	AMPHION_TILED_10BIT tiling format

2.2.12 g2d_surfaceEx structure

This enumeration describes the surface with tiling attributes.

Table 12. g2d_surfaceEx structure

g2d_surfaceEx member	Type	Description
base	g2d_surface	Base g2d_surface
tiling	g2d_tiling	Tiling format
ts	g2d_tile_status	Not used
reserved[8]	int	Not used

2.2.13 g2d_warp_map_format enumeration

This enumeration describes the content of the pixel data in the coordinate buffer.

Table 13. g2d_warp_map_format enumeration

Name	Numeric	Description
G2D_WARP_MAP_PNT	0	x and y (sample points)
G2D_WARP_MAP_DPNT	1	dx and dy (vectors between adjacent sample points)

Table 13. g2d_warp_map_format enumeration...continued

Name	Numeric	Description
G2D_WARP_MAP_DDPNT	2	ddx and ddy (deltas between adjacent vectors)

2.2.14 g2d_warp_coordinates structure

This enumeration describes the coordinates buffer with operation attributes.

Table 14. g2d_warp_coordinates structure

g2d_warp_coordinates member	Type	Description
addr	g2d_phys_addr_t	Physical address of the coordinates buffer
format	g2d_warp_map_format	Pixel format of of the coordinates buffer
bpp	int	Bits per pixel of the coordinates buffer
width	int	Width in pixel unit of the coordinates buffer
height	int	Height in pixel unit of the coordinates buffer
arb_start_x	unsigned int	Start point for the sample-point interpolation (X coordinate)
arb_start_y	unsigned int	Start point for the sample-point interpolation (Y coordinate)
arb_delta_xx	unsigned int	X coordinate of vector between the first and second sample point
arb_delta_xy	unsigned int	Y coordinate of vector between the first and second sample point
arb_delta_yx	unsigned int	X coordinate of vector between the start and first sample point
arb_delta_yy	unsigned int	Y coordinate of vector between the start and first sample point

2.3 G2D function description

2.3.1 g2d_open

Description	Open a G2D device and return a handle.
Syntax	<pre>int g2d_open (void **handle);</pre>
Parameters	handle: Pointer to receive G2D device handle
Returns	Success with 0, fail with -1

2.3.2 g2d_close

Description	Close G2D device with the handle.
Syntax	<pre>int g2d_close (void *handle);</pre>
Parameters	handle: G2D device handle
Returns	Success with 0, fail with -1

2.3.3 g2d_make_current

Description	Set the specific hardware type for the current context.
Syntax	<pre>int g2d_make_current (void *handle, enum g2d_hardware_type type);</pre>
Parameters	handle: G2D device handle
Returns	Success with 0, fail with -1

2.3.4 g2d_clear

Description	Clear a specific area.
Syntax	<pre>int g2d_clear (void *handle, struct g2d_surface *area);</pre>
Parameters	handle: G2D device handle area: The area to be cleared
Returns	Success with 0, fail with -1

2.3.5 g2d_blit

Description	G2D blit from source to destination with alternative operation (Blend, Dither, etc.).
Syntax	<pre>int g2d_blit (void *handle, struct g2d_surface *src, struct g2d_surface *dst);</pre>
Parameters	handle: G2D device handle src: source surface dst: destination surface
Returns	Success with 0, fail with -1

2.3.6 g2d_copy

Description	G2D copy with specified size.
Syntax	<pre>int g2d_copy (void *handle, struct g2d_buf *d, struct g2d_buf* s, int size);</pre>
Parameters	handle: G2D device handle d: destination buffer s: source buffer size: copy bytes

Limitations	If the destination buffer is cacheable, it must be invalidated before g2d_copy due to the alignment limitation of G2D driver.
Returns	Success with 0, fail with -1

2.3.7 g2d_query_cap

Description	Query the alternative capability enablement.
Syntax	<pre>int g2d_query_cap (void *handle, enum g2d_cap_mode cap, int *enable);</pre>
Parameters	handle: G2D device handle cap: G2D capability to query enable: Pointer to receive G2D capability enablement
Returns	Success with 0, fail with -1

2.3.8 g2d_enable

Description	Enable G2D capability with the specific mode.
Syntax	<pre>int g2d_enable (void *handle, enum g2d_cap_mode cap);</pre>
Parameters	handle: G2D device handle cap: G2D capability to enable
Returns	Success with 0, fail with -1

2.3.9 g2d_disable

Description	Disable G2D capability with the specific mode.
Syntax	<pre>int g2d_disable (void *handle, enum g2d_cap_mode cap);</pre>
Parameters	handle: G2D device handle cap: G2D capability to disable
Returns	Success with 0, fail with -1

2.3.10 g2d_cache_op

Description	Perform cache operations for the cacheable buffer allocated through the G2D driver.
Syntax	<pre>int g2d_cache_op (struct g2d_buf *buf, enum g2d_cache_mode op);</pre>
Parameters	buf: the buffer to be handled with cache operations

	op: cache operation type
Returns	Success with 0, fail with -1

2.3.11 g2d_alloc

Description	Allocate a buffer through G2D device
Syntax	<pre>struct g2d_buf *g2d_alloc (int size, int cacheable);</pre>
Parameters	size: allocated bytes cacheable: 0, non-cacheable; 1, cacheable attribute defined by system
Returns	Success with valid G2D buffer pointer, fail with 0

2.3.12 g2d_free

Description	Free the buffer through G2D device.
Syntax	<pre>int g2d_free (struct g2d_buf *buf);</pre>
Parameters	buf: G2D buffer to free
Returns	Success with 0, fail with -1

2.3.13 g2d_flush

Description	Flush G2D command and return without completing pipeline.
Syntax	<pre>int g2d_flush (void *handle);</pre>
Parameters	handle: G2D device handle
Returns	Success with 0, fail with -1

2.3.14 g2d_finish

Description	Flush G2D command and then return when pipeline is finished.
Syntax	<pre>int g2d_finish (void *handle);</pre>
Parameters	handle: G2D device handle
Returns	Success with 0, fail with -1

2.3.15 g2d_multi_blit

Description	Blit multiple sources to one destination.
Syntax	<pre>int g2d_multi_blit (void *handle, struct g2d_surface_pair *sp[], int layers);</pre>

Parameters	handle: G2D device handle sp: array in which elements point to g2d_surface_pair layers: number of the source layers that need to be blited
Returns	Success with 0, fail with -1

Note:

There are some restrictions for this API that we should be aware of.

- This API only works on the i.MX 6DualPlus/QuadPlus platform.
- The maximum number of the source layers that can be blited one time is 8.
- Although g2d_surface_pair binds one source g2d_surface and one destination g2d_surface as a pair, it only supports one destination surface. The relationship between the source and destination is many to one, but each source surface can be set separately and differently, and its dimension, stride, rotation, and format can differ with that of the destination surface.
- The rotation of the destination surface is set to 0 degrees by default, and cannot be changed.
- The key restriction is that the destination rectangle cannot be set, which means that the destination rectangle must be the same as the source rectangle. Therefore, if the source rectangle is set to (l, t, r, b), the destination rectangle should also be set to (l, t, r, b) by hardware. In the chapter on multi source blit (Section 2.5.4), as it makes no sense to set the destination rectangles, we just set all of them to (0, 0, width, height) for future extension.

2.3.16 g2d_query_hardware

Description	Query whether g2d_hardware_type is available in the current G2D.
Syntax	<pre>int g2d_query_hardware (void *handle, enum g2d_hardware_type type, int *available);</pre>
Parameters	handle: G2D device handle type: G2D hardware type available: Pointer to receive G2D hardware type availability
Returns	Success with 0, fail with -1

2.3.17 g2d_query_feature

Description	Query if the features are available in G2D BLT.
Syntax	<pre>int g2d_query_feature (void *handle, enum g2d_feature feature, int *available);</pre>
Parameters	handle: G2D device handle feature: G2D feature in g2d_blit available: Pointer to receive G2D feature availability
Returns	Success with 0, fail with -1

2.3.18 g2d_blitEx

Description	G2D blit from the source to destination (with g2d_tiling) with alternative operation (Blend, Dither, etc.).
Syntax	<pre>int g2d_blitEx(void *handle, struct g2d_surfaceEx *srcEx, struct g2d_surfaceEx *dstEx);</pre>
Parameters	<p>handle: G2D device handle</p> <p>srcEx: Source surfaceEx</p> <p>dstEx: Destination surfaceEx</p>
Returns	Success with 0, fail with -1

2.3.19 g2d_set_clipping

Description	Set a rectangular clipping window for the destination.
Syntax	<pre>int g2d_set_clipping(void *handle, int left, int top, int right, int bottom);</pre>
Parameters	<p>handle: G2D device handle</p> <p>left: Left offset of the clipping rectangle</p> <p>top: Top offset of the clipping rectangle</p> <p>right: Right offset of the clipping rectangle</p> <p>bottom: Bottom offset of the clipping rectangle</p>
Returns	Success with 0, fail with -1

2.3.20 g2d_set_csc_matrix

Description	Set the Color Space Conversion Matrix.
Syntax	<pre>int g2d_set_csc_matrix(void *handle, const unsigned *matrix);</pre>
Parameters	<p>handle: G2D device handle</p> <p>matrix: 4x4 matrix</p>
Returns	Success with 0, fail with -1

2.3.21 g2d_buf_from_fd

Description	Get g2d_buf pointer from the buffer file descriptor.
Syntax	<pre>struct g2d_buf *g2d_buf_from_fd(int fd);</pre>

Parameters	fd: Buffer file descriptor
Returns	Success with g2d_buf pointer, fail with NULL

2.3.22 g2d_buf_export_fd

Description	Get the buffer file descriptor from the g2d_buf pointer.
Syntax	<pre>int g2d_buf_export_fd(struct g2d_buf * buf);</pre>
Parameters	buf: g2d_buf pointer
Returns	Success with buffer fd, fail with -EINVAL

2.3.23 g2d_buf_from_virt_addr

Description	Get the g2d_buf pointer from a virtual address.
Syntax	<pre>struct g2d_buf *g2d_buf_from_virt_addr(void *vaddr, int size);</pre>
Parameters	vaddr: Virtual address size: Size of the buffer
Returns	Success with g2d_buf pointer, fail with NULL

2.3.24 g2d_create_fence_fd

Description	Create G2D fence file descriptor.
Syntax	<pre>int g2d_create_fence_fd(void *handle);</pre>
Parameters	handle: G2D device handle
Returns	Success with valid fence fd (>=0), fail or not supported with -1

2.3.25 g2d_set_warp_coordinates

Description	Set the coordinate buffer when doing warping.
Syntax	<pre>int g2d_set_warp_coordinates(void *handle, struct g2d_warp_coordinates *coord);</pre>
Parameters	handle: G2D device handle

	coord: G2D warp coordinate buffer
Returns	Success with 0, fail with -1

Note:

There is a restriction for this API:

This API only works on the i.MX 8QuadXPlus/8QuadMax/95 platform.

2.4 Support of new operating system in G2D

G2D code is independent on operating system (OS) except of buffer allocation. Allocating the memory for buffer is made by mechanism that is offered by each OS differently. The code for allocation is located in [G2D repository copy]/source/os/[OS name]. Therefore, supporting new OS includes the following steps:

1. Create a new folder in **[G2D repository copy]/source/os/** with the name of the new OS and update implementation in the included source code according to the new OS allocation mechanism.
2. When creating new makefiles for the OS, include the files from the new folder.
3. The test named **overlay_test** contains the OS dependent code. For supporting the new OS in this test, create new folder in **[G2D repository copy]/test/overlay_test/os** and update the code according to the new OS mechanism for display initialization. Also update makefiles to include code from the new folder.

2.5 Sample code for G2D API usage

This chapter provides the brief prototype code with G2D API.

2.5.1 Color space conversion from YUV to RGB

```
g2d_open(&handle);
src.planes[0] = buf_y;
src.planes[1] = buf_u;
src.planes[2] = buf_v;
src.left = crop.left;
src.top = crop.top;
src.right = crop.right;
src.bottom = crop.bottom;
src.stride = y_stride;
    src.width = y_width;
    src.height = y_height;
src.rot = G2D_ROTATION_0;
src.format = G2D_I420;
dst.planes[0] = buf_rgba;
dst.left = 0;
dst.top = 0;
dst.right = disp_width;
dst.bottom = disp_height;
dst.stride = disp_width;
    dst.width = disp_width;
    dst.height = disp_height;
dst.rot = G2D_ROTATION_0;
dst.format = G2D_RGBA8888;
g2d_blit(handle, &src, &dst);
    g2d_finish(handle);
g2d_close(handle);
```

2.5.2 Alpha blend in source over mode

```
g2d_open(&handle);
src.planes[0] = src_buf;
src.left = 0;
src.top = 0;
src.right = test_width;
src.bottom = test_height;
src.stride = test_width;
src.width = test_width;
src.height = test_height;
src.rot = G2D_ROTATION_0;
src.format = G2D_RGBA8888;
src.blendfunc = G2D_ONE;
dst.planes[0] = dst_buf;
dst.left = 0;
dst.top = 0;
dst.right = test_width;
dst.bottom = test_height;
dst.stride = test_width;
dst.width = test_width;
dst.height = test_height;
dst.format = G2D_RGBA8888;
dst.rot = G2D_ROTATION_0;
dst.blendfunc = G2D_ONE_MINUS_SRC_ALPHA;
g2d_enable(handle, G2D_BLEND);
g2d_blit(handle, &src, &dst);
g2d_finish(handle);
g2d_disable(handle, G2D_BLEND);
g2d_close(handle);
```

2.5.3 Source cropping and destination rotation

```
g2d_open(&handle);
src.planes[0] = src_buf;
src.left = crop.left;
src.top = crop.left;
src.right = crop.right;
src.bottom = crop.bottom;
src.stride = src_stride;
src.width = src_width;
src.height = src_height;
src.format = G2D_RGBA8888;
src.rot = G2D_ROTATION_0; //G2D_FLIP_H or G2D_FLIP_V
dst.planes[0] = dst_buf;
dst.left = 0;
dst.top = 0;
dst.right = dst_width;
dst.bottom = dst_height;
dst.stride = dst_width;
dst.width = dst_width;
dst.height = dst_height;
dst.format = G2D_RGBA8888;
dst.rot = G2D_ROTATION_90;
g2d_blit(handle, &src, &dst);
g2d_finish(handle);
```

```
g2d_close(handle)
```

2.5.4 Multi source blit

```
const int layers = 8;
struct g2d_buf *d_buf;
struct g2d_buf *mul_s_buf[layers];
struct g2d_surface_pair *sp[layers];
g2d_open(&handle)
for(n = 0; n < layers; n++) {
sp[n] = (struct g2d_surface_pair *)malloc(sizeof(struct g2d_surface_pair));
}
d_buf = g2d_alloc(test_width * test_height * 4, 0);
for(n = 0; n < layers; n++) {
mul_s_buf[n] = g2d_alloc(test_width * test_height * 4, 0);
}
for(n = 0; n < layers; n++) {
sp[n]->s.left = img_info_ptr[n]->img_left;
sp[n]->s.top = img_info_ptr[n]->img_top;
sp[n]->s.right = img_info_ptr[n]->img_right;
sp[n]->s.bottom = img_info_ptr[n]->img_bottom;
sp[n]->s.stride = img_info_ptr[n]->img_width;
sp[n]->s.width = img_info_ptr[n]->img_width;
sp[n]->s.height = img_info_ptr[n]->img_height;
sp[n]->s.rot = img_info_ptr[n]->img_rot;
sp[n]->s.format = img_info_ptr[n]->img_format;
sp[n]->s.planes[0] = mul_s_buf[n]->buf_paddr;
}
sp[0]->d.left = 0;
sp[0]->d.top = 0;
sp[0]->d.right = test_width;
sp[0]->d.bottom = test_height;
sp[0]->d.stride = test_width;
sp[0]->d.width = test_width;
sp[0]->d.height = test_height;
sp[0]->d.format = G2D_RGBA8888;
sp[0]->d.rot = G2D_ROTATION_0;
sp[0]->d.planes[0] = d_buf->buf_paddr;
for(n = 1; n < layers; n++) {
sp[n]->d = sp[0]->d;
}
g2d_multi_blit(handle, sp, layers);
g2d_finish(handle);
for(n = 0; n < layers; n++)
g2d_free(mul_s_buf[n]);
g2d_free(d_buf);
g2d_close(handle);
```

2.5.5 Sharing Buffers between APIs using G2D Buffers:

The G2D buffers can be used to avoid memory copies between APIs. Create a buffer using `g2d_alloc` and then map it as an OpenGL ES texture or as an OpenVX buffer or an OpenCV Mat:

Allocate your buffer with:

```
struct g2d_buf * buffer0;
buffer0 = g2d_alloc(WIDTH*HEIGHT*4, 0);
```

For OpenCV, you map the buffer to the data field of the cv::Mat

```
cv::Mat buffer0Mat;
buffer0Mat.create (WIDTH, HEIGHT, CV_8UC4);
buffer0Mat.data = (uchar *) ((unsigned long) buffer0->buf_vaddr);
```

For OpenGL ES, you can make use of the DirectVIV extensions:

```
glGenTextures(1, &textureHandle[0]);
glBindTexture(GL_TEXTURE_2D, textureHandle[0]);
glTexParameteri(GL_TEXTURE_2D, GL_TEXTURE_MAG_FILTER, GL_LINEAR);
glTexParameteri(GL_TEXTURE_2D, GL_TEXTURE_MIN_FILTER, GL_LINEAR);
glTexDirectVIVMap(GL_TEXTURE_2D, WIDTH, HEIGHT, GL_RGBA,
                  &buffer0->buf_vaddr, (uint *)&buffer0->buf_paddr);
glTexDirectInvalidateVIV (GL_TEXTURE_2D);
glBindTexture(GL_TEXTURE_2D, 0);
```

For OpenVX you create vxImages from the buffer ranges:

```
vx_imagepatch_addressing_t patch0 = { (vx_uint32)WIDTH, (vx_uint32)HEIGHT,
(vx_int32)4, (vx_int32)HEIGHT*4, VX_SCALE_UNITY, VX_SCALE_UNITY, 1, 1 };
void *ptr0 = buffer0->buf_vaddr;
vxInputImage = vxCreateImageFromHandle(contextVX,
VX_DF_IMAGE_RGBX, &patch0, (void **)&ptr0, VX_MEMORY_TYPE_HOST);
```

With this scheme you can create a multi API pipeline, where you can post-process your OpenGL ES render result with CV or VX without the need of copying data.

2.5.6 Warp/Dewarp

```
g2d_open(&handle);
g2d_query_feature(handle, G2D_WARP_DEWARP, &support_warp);
if (!support_warp) {
    fprintf(stderr, "G2D device cannot perform warp/dewarp operations\n");
    return -1;
}

s_buf = g2d_alloc(s_buf_size, 0);
d_buf = g2d_alloc(d_buf_size, 0);
coord_buf = g2d_alloc(coord_buf_size, 0);

/* read src to s_buf
 * ...
 */

// copy warp_coord to coord_buf
coord_buf_size = width*height*coord_bpp;
memcpy(coord_buf->buf_vaddr, warp_coord, coord_buf_size);

src.left = 0;
src.top = 0;
src.right = width;
src.bottom = height;
src.width = width;
src.height = height;
src.format = in_format;
src.stride = width;
```

```
src.planes[0] = s_buf->buf_paddr;

dst.left = 0;
dst.top = 0;
dst.right = width;
dst.bottom = height;
dst.width = width;
dst.height = height;
dst.format = out_format;
dst.stride = width;
dst.planes[0] = d_buf->buf_paddr;
dst.planes[1] = d_buf->buf_paddr + dst_plane0_size;
dst.planes[2] = d_buf->buf_paddr + dst_plane0_size + dst_plane1_size;

coord.addr = coord_buf->buf_paddr;
coord.width = width;
coord.height = height;
coord.format = coord_format;
coord.bpp = coord_bpp;

coord.arb_start_x = start_x;
coord.arb_start_y = start_y;
coord.arb_delta_xx = delta_xx;
coord.arb_delta_xy = delta_xy;
coord.arb_delta_yx = delta_yx;
coord.arb_delta_yy = delta_yy;

g2d_enable(handle, G2D_WARPING);
g2d_set_warp_coordinates(handle, &coord);
g2d_blit(handle, &src, &dst);
g2d_disable(handle, G2D_WARPING);
g2d_finish(handle);

/* write d_buf to dst
 * ...
 */

g2d_free(ctx->s_buf);
g2d_free(ctx->d_buf);
g2d_free(ctx->coord_buf);
g2d_close(ctx->handle);
```

Note:

- When doing warping, the RGB SRC format is supported on the i.MX 8QuadXPlus/8QuadMax/95 platform. The packed YUV422 SRC format is only supported on the i.MX 95 platform.
- Some parameters and initial values are required from the Dewarp Calibreton Tool to do warping. BitsPerPixel (bpp) of the coordinate buffer and initial values are related to the warp algorithm in the Dewarp Calibreton Tool.

Table 15. Parameters and initial values for warping

G2D warp map format (coord.format)	Warp algorithm	BPP of coordinate buffer (coord.bpp)	Initial value required
G2D_WARP_MAP_PNT	absolute_32bpp	32	no
G2D_WARP_MAP_DPNT	delta_32bpp	32	start_x
	delta_16bpp	16	start_y

Table 15. Parameters and initial values for warping...continued

G2D warp map format (coord.format)	Warp algorithm	BPP of coordinate buffer (coord.bpp)	Initial value required
	delta_8bpp	8	
G2D_WARP_MAP_DDPNT	deltadelta_32bpp	32	start_x
	deltadelta_16bpp	16	start_y
	deltadelta_8bpp	8	delta_xx
	deltadelta_4bpp	4	delta_xy delta_yx delta_yy

2.6 Feature list on multiple platforms

This user guide is for multiple platforms, such as i.MX 6 and i.MX 8, and the hardware for the G2D implementation are different on those platforms, so some G2D features are also different.

For example, the G2D_YVYU and G2D_VYUY formats are not supported on the i.MX 8, and the g2d_multi_blit function only works on the i.MX 6DualPlus/QuadPlus. Therefore, we list those differences in the following feature table.

Table 16. Feature list on multiple platforms

Feature	i.MX 6		i.MX 7	i.MX 8	
	6Solo/6Dual/6Quad	6DualPlus/6QuadPlus	7ULP	8M Mini/ 8M Plus	8QuadMax/8QuadXPlus
G2D_YVYU	Yes	Yes	Yes	Yes	No
G2D_VYUY	Yes	Yes	Yes	Yes	No
G2D_HARDWARE_VG	Yes	Yes	No	No	No
G2D_MULTI_SOURCE_BLT	No	Yes	Yes	Yes	No
g2d_cache_op	Yes	Yes	Yes	Yes	No

2.7 Arbitrary Warping

Arbitrary Warping is useful for the applications like lens distortion removal or windshield correction for head-up displays. For example, the result of such a lens distortion removal is an image with straightened lines and with objects that look natural.



Arbitrary sample-points from a coordinate buffer allows any kind of static re-sampling pattern on the dynamic image content.

In the coordinate buffer, an x and y value for each pixel is stored. These values are used to calculate the coordinate of the current fetch sample-point.

Sizes from 4 bpp to 32 bpp for one x/y value pair in the coordinate buffer are supported. The values are stored as a signed fix-point value with integer and fractional part.

The fractional precision corresponds to sub-pixel precision of the sampling points on the source buffer pixel grid.

The coordinate buffer can be interpreted as:

- **Coordinate mode:** Sample points (x and y coordinate relative to source image).
- **Delta mode:** Deltas of adjacent sample points.
- **Delta increment mode:** Increments of adjacent deltas.

Storing deltas and even more storing delta increments allow using much smaller bpp formats for the same amount of distortion to save memory and bandwidth. On the other side, it results in some limitation on the sample pattern. Therefore, the recommended setup is delta or delta increment mode with a coordinate format of 8 bpp or below.

3 Vivante EGL and OGL Extension Support

3.1 Introduction

The following tables list the level of support for EGL and OES extensions available with i.MX hardware and software. Support levels are current as of the date of the document and subject to change.

Two tables are provided. The first table lists the EGL interface extensions. The second table lists extensions for OpenGL ES 1.1, OpenGL ES 2.0, and OpenGL ES 3.0.

Key:

- **Extension Name and Number:** Each listed extension is derived from the relevant khronos.org webpage list and includes the extension number as well as a hyperlink to the khronos description of the extension.
- **Yes:** Support is currently available.
- **No:** Support is not available. (Reasons for lack of support may vary: the extension may be proprietary or obsolete, or not applicable to the specified OES version.)
- **N/A:** Support is not provided as the extension is not applicable in this and subsequent versions of the specification.

3.2 EGL extension support

The following table includes the list of all current EGL Extensions and indicates their support level.

(list from www.khronos.org/registry/egl/ as of 1/24/2020)

Table 17. EGL extension support

EGL Extension Number, Name and hyperlink (2020)	Linux	Android	QNX
1. EGL_KHR_config_attribs			
2. EGL_KHR_lock_surface	YES	YES	YES
3. EGL_KHR_image	YES	YES	YES
4. EGL_KHR_vg_parent_image			
5. EGL_KHR_gl_texture_2D_image	YES	YES	YES
EGL_KHR_gl_texture_cubemap_image	YES	YES	YES

Table 17. EGL extension support...continued

EGL Extension Number, Name and hyperlink (2020)	Linux	Android	QNX
EGL_KHR_gl_texture_3D_image			
EGL_KHR_gl_renderbuffer_image	YES	YES	YES
6. EGL_KHR_reusable_sync	YES	YES	YES
7. EGL_KHR_image_base	YES	YES	YES
8. EGL_KHR_image_pixmap	YES	YES	YES
9. EGL_IMG_context_priority	YES	YES	
10. EGL_NOK_texture_from_pixmap			
11. EGL_KHR_lock_surface2			
12. EGL_NV_coverage_sample			
13. EGL_NV_depth_nonlinear			
14. EGL_NV_sync			
15. EGL_KHR_fence_sync	YES	YES	YES
16. EGL_NOK_swap_region2			
17. EGL_HI_clientpixmap			
18. EGL_HI_colorformats			
19. EGL_MESA_drm_image			
20. EGL_NV_post_sub_buffer			
21. EGL_ANGLE_query_surface_pointer			
22. EGL_ANGLE_surface_d3d_texture_2d_share_handle			
23. EGL_NV_coverage_sample_resolve			
24. EGL_NV_system_time			
25. EGL_KHR_stream			
EGL_KHR_stream_attrib			
26. EGL_KHR_stream_consumer_gltexture			
27. EGL_KHR_stream_producer_eglsurface			
28. EGL_KHR_stream_producer_aldatalocator			
29. EGL_KHR_stream_fifo			
30. EGL_EXT_create_context_robustness			
31. EGL_ANGLE_d3d_share_handle_client_buffer			
32. EGL_KHR_create_context	YES	YES	YES
33. EGL_KHR_surfaceless_context	YES	YES	YES
34. EGL_KHR_stream_cross_process_fd			
35. EGL_EXT_multiview_window			
36. EGL_KHR_wait_sync	YES	YES	YES
37. EGL_NV_post_convert_rounding			
38. EGL_NV_native_query			

Table 17. EGL extension support...continued

EGL Extension Number, Name and hyperlink (2020)	Linux	Android	QNX
39. EGL_NV_3dvision_surface			
40. EGL_ANDROID_framebuffer_target		YES	
41. EGL_ANDROID_blob_cache		YES	
42. EGL_ANDROID_image_native_buffer		YES	
43. EGL_ANDROID_native_fence_sync		YES	
44. EGL_ANDROID_recordable		YES	
45. EGL_EXT_buffer_age	YES	YES	YES
46. EGL_EXT_image_dma_buf_import	YES	YES	
47. EGL_ARM_pixmap_multisample_discard			
48. EGL_EXT_swap_buffers_with_damage	YES	YES	YES
49. EGL_NV_stream_sync			
50. EGL_EXT_platform_base	YES	YES	YES
51. EGL_EXT_client_extensions	YES	YES	YES
52. EGL_EXT_platform_x11	YES	YES	YES
53. EGL_KHR_cl_event			
54. EGL_KHR_get_all_proc_addresses	YES	YES	YES
EGL_KHR_client_get_all_proc_addresses	YES	YES	YES
55. EGL_MESA_platform_gbm			
56. EGL_EXT_platform_wayland	YES		
57. EGL_KHR_lock_surface3			
58. EGL_KHR_cl_event2			
59. EGL_KHR_gl_colorspace			
60. EGL_EXT_protected_surface	YES	YES	YES
61. EGL_KHR_platform_android		YES	
62. EGL_KHR_platform_gbm	YES	YES	YES
63. EGL_KHR_platform_wayland	YES		
64. EGL_KHR_platform_x11	YES		
65. EGL_EXT_device_base			
66. EGL_EXT_platform_device			
67. EGL_NV_device_cuda			
68. EGL_NV_cuda_event			
69. EGL_TIZEN_image_native_buffer			
70. EGL_TIZEN_image_native_surface			
71. EGL_EXT_output_base			
72. EGL_EXT_device_drm			
EGL_EXT_output_drm			

Table 17. EGL extension support...continued

EGL Extension Number, Name and hyperlink (2020)	Linux	Android	QNX
73. EGL_EXT_device_openwf			
EGL_EXT_output_openwf			
74. EGL_EXT_stream_consumer_egloutput			
75. EGL_KHR_partial_update	YES	YES	YES
76. EGL_KHR_swap_buffers_with_damage	YES	YES	YES
77. EGL_ANGLE_window_fixed_size			
78. EGL_EXT_yuv_surface			
79. EGL_MESA_image_dma_buf_export			
80. EGL_EXT_device_enumeration			
81. EGL_EXT_device_query			
82. EGL_ANGLE_device_d3d			
83. EGL_KHR_create_context_no_error			
84. EGL_KHR_debug			
85. EGL_NV_stream_metadata			
86. EGL_NV_stream_consumer_gltexture_yuv			
87. EGL_IMG_image_plane_attribs			
88. EGL_KHR_mutable_render_buffer			
89. EGL_EXT_protected_content			
90. EGL_ANDROID_presentation_time			
91. EGL_ANDROID_create_native_client_buffer			
92. EGL_ANDROID_front_buffer_auto_refresh			
93. EGL_KHR_no_config_context	YES	YES	YES
94. EGL_KHR_context_flush_control			
95. EGL_ARM_implicit_external_sync			
96. EGL_MESA_platform_surfaceless			
97. EGL_EXT_image_dma_buf_import_modifiers	YES	YES	
98. EGL_EXT_pixel_format_float			
99. EGL_EXT_gl_colorspace_bt2020_linear			
EGL_EXT_gl_colorspace_bt2020_pq			
100. EGL_EXT_gl_colorspace_scrgb_linear			
101. EGL_EXT_surface_SMPTE2086_metadata			
102. EGL_NV_stream_fifo_next			
103. EGL_NV_stream_fifo_synchronous			
104. EGL_NV_stream_reset			
105. EGL_NV_stream_frame_limits			
106. EGL_NV_stream_remote			

Table 17. EGL extension support...continued

EGL Extension Number, Name and hyperlink (2020)	Linux	Android	QNX
EGL_NV_stream_cross_object			
EGL_NV_stream_cross_display			
EGL_NV_stream_cross_process			
EGL_NV_stream_cross_partition			
EGL_NV_stream_cross_system			
107. EGL_NV_stream_socket			
EGL_NV_stream_socket_unix			
EGL_NV_stream_socket_inet			
108. EGL_EXT_compositor			
109. EGL_EXT_surface_CTA861_3_metadata			
110. EGL_EXT_gl_colorspace_display_p3			
111. EGL_EXT_gl_colorspace_display_p3_linear			
112. EGL_EXT_gl_colorspace_srgb (non-linear)			
113. EGL_EXT_image_implicit_sync_control			
114. EGL_EXT_bind_to_front			
115. EGL_ANDROID_get_frame_timestamps			
116. EGL_ANDROID_get_native_client_buffer			
117. EGL_NV_context_priority_realtime			
118. EGL_EXT_image_gl_colorspace			
119. EGL_KHR_display_reference			
120. EGL_NV_stream_flush			
121. EGL_EXT_sync_reuse			
122. EGL_EXT_client_sync			
123. EGL_EXT_gl_colorspace_display_p3_passthrough			
124. EGL_MESA_query_driver			
125. EGL_ANDROID_GLES_layers			
126. EGL_NV_n_buffer			
127. EGL_NV_stream_origin			
128. EGL_NV_stream_dma			
129. EGL_WL_bind_wayland_display	YES		
130. EGL_WL_create_wayland_buffer_from_image	YES		

3.3 OpenGL ES extension support

The following table includes the list of all current OpenGL ES Extensions and indicates their support level.

(list from www.khronos.org/registry/gles/ as of 6/14/2020)

Table 18. OpenGL ES extension support

Extension Number, Name and hyperlink	ES1.1	ES2.0/3.0/3.1/3.2
1. GL_OES_blend_equation_separate	YES	
2. GL_OES_blend_func_separate	YES	
3. GL_OES_blend_subtract	YES	
4. GL_OES_byte_coordinates	YES	
5. GL_OES_compressed_ETC1_RGB8_texture	YES	YES
6. GL_OES_compressed_paletted_texture	YES	YES
7. GL_OES_draw_texture	YES	
8. GL_OES_extended_matrix_palette	YES	
9. GL_OES_fixed_point	YES	
10. GL_OES_framebuffer_object	YES	
11. GL_OES_matrix_get	YES	
12. GL_OES_matrix_palette	YES	
13. GL_OES_point_size_array	YES	
14. GL_OES_point_sprite	YES	
15. GL_OES_query_matrix	YES	
16. GL_OES_read_format	YES	
17. GL_OES_single_precision	YES	
18. GL_OES_stencil_wrap	YES	
19. GL_OES_texture_cube_map	YES	
20. GL_OES_texture_env_crossbar		
21. GL_OES_texture_mirrored_repeat	YES	
22. GL_OES_EGL_image	YES	YES
23. GL_OES_depth24	YES	YES
24. GL_OES_depth32		YES
25. GL_OES_element_index_uint	YES	YES
26. GL_OES_fbo_render_mipmap	YES	YES
27. GL_OES_fragment_precision_high		YES
28. GL_OES_mapbuffer	YES	YES
29. GL_OES_rgb8_rgba8	YES	YES
30. GL_OES_stencil1		
31. GL_OES_stencil4		
32. GL_OES_stencil8	YES	
33. GL_OES_texture_3D		
34. GL_OES_texture_float_linear		
GL_OES_texture_half_float_linear		CORE
35. GL_OES_texture_float		CORE

Table 18. OpenGL ES extension support...continued

Extension Number, Name and hyperlink	ES1.1	ES2.0/3.0/3.1/3.2
GL_OES_texture_half_float		CORE
36. GL_OES_texture_npot	YES	YES
37. GL_OES_vertex_half_float	YES	YES
38. GL_AMD_compressed_3DC_texture		
39. GL_AMD_compressed_ATC_texture		
40. GL_EXT_texture_filter_anisotropic	CORE	CORE
41. GL_EXT_texture_type_2_10_10_10_REV		CORE
42. GL_OES_depth_texture		YES
43. GL_OES_packed_depth_stencil	YES	YES
44. GL_OES_standard_derivatives		YES
45. GL_OES_vertex_type_10_10_10_2		CORE
46. GL_OES_get_program_binary		YES
47. GL_AMD_program_binary_Z400		
48. GL_EXT_texture_compression_dxt1		YES
49. GL_AMD_performance_monitor		
50. GL_EXT_texture_format_BGRA8888	YES	YES
51. GL_NV_fence		
52. GL_IMG_read_format		
53. GL_IMG_texture_compression_pvrtc		
54. GL_QCOM_driver_control		
55. GL_QCOM_performance_monitor_global_mode		
56. GL_IMG_user_clip_plane		
57. GL_IMG_texture_env_enhanced_fixed_function		
58. GL_APPLE_texture_2D_limited_npot		
59. GL_EXT_texture_lod_bias	YES	
60. GL_QCOM_writeonly_rendering		
61. GL_QCOM_extended_get		
62. GL_QCOM_extended_get2		
63. GL_EXT_discard_framebuffer		YES
64. GL_EXT_blend_minmax	YES	YES
65. GL_EXT_read_format_bgra	YES	YES
66. GL_IMG_program_binary		
67. GL_IMG_shader_binary		
68. GL_EXT_multi_draw_arrays	YES	YES
GL_SUN_multi_draw_arrays	NO	
69. GL_QCOM_tiled_rendering		

Table 18. OpenGL ES extension support...continued

Extension Number, Name and hyperlink	ES1.1	ES2.0/3.0/3.1/3.2
70. GL_OES_vertex_array_object		YES
71. GL_NV_coverage_sample		
72. GL_NV_depth_nonlinear		
73. GL_IMG_multisampled_render_to_texture		
74. GL_OES_EGL_sync	YES	YES
75. GL_APPLE_rgb_422		
76. GL_EXT_shader_texture_lod		
77. GL_APPLE_framebuffer_multisample		
78. GL_APPLE_texture_format_BGRA8888		
79. GL_APPLE_texture_max_level		
80. GL_ARM_mali_shader_binary		
81. GL_ARM_rgba8		
82. GL_ANGLE_framebuffer_blit		
83. GL_ANGLE_framebuffer_multisample		
84. GL_VIV_shader_binary		
85. GL_EXT_frag_depth		YES
86. GL_OES_EGL_image_external	YES	YES
87. GL_DMP_shader_binary		
88. GL_QCOM_alpha_test		
89. GL_EXT_unpack_subimage		
90. GL_NV_draw_buffers		
91. GL_NV_fbo_color_attachments		
92. GL_NV_read_buffer		
93. GL_NV_read_depth_stencil		
94. GL_NV_texture_compression_s3tc_update		
95. GL_NV_texture_npot_2D_mipmap		
96. GL_EXT_color_buffer_half_float		CORE
97. GL_EXT_debug_label		
98. GL_EXT_debug_marker		
99. GL_EXT_occlusion_query_boolean		
100. GL_EXT_separate_shader_objects		
101. GL_EXT_shadow_samplers		
102. GL_EXT_texture_rg		YES
103. GL_NV_EGL_stream_consumer_external		
104. GL_EXT_sRGB		YES
105. GL_EXT_multisampled_render_to_texture		YES

Table 18. OpenGL ES extension support...continued

Extension Number, Name and hyperlink	ES1.1	ES2.0/3.0/3.1/3.2
106. GL_EXT_robustness		YES
107. GL_EXT_texture_storage		
108. GL_ANGLE_instanced_arrays		
109. GL_ANGLE_pack_reverse_row_order		
110. GL_ANGLE_texture_compression_dxt3		
GL_ANGLE_texture_compression_dxt1		
GL_ANGLE_texture_compression_dxt5		
111. GL_ANGLE_texture_usage		
112. GL_ANGLE_translated_shader_source		
113. GL_FJ_shader_binary_GCCSO		
114. GL_OES_required_internalformat		YES
115. GL_OES_surfaceless_context		YES
116. GL_KHR_texture_compression_astc_hdr		
GL_KHR_texture_compression_astc_ldr		YES
117. GL_KHR_debug		YES
118. GL_QCOM_binning_control		
119. GL_ARM_mali_program_binary		
120. GL_EXT_map_buffer_range		
121. GL_EXT_shader_framebuffer_fetch		CORE
GL_EXT_shader_framebuffer_fetch_non_coherent		
122. GL_APPLE_copy_texture_levels		
123. GL_APPLE_sync		
124. GL_EXT_multiview_draw_buffers		
125. GL_NV_draw_texture		
126. GL_NV_packed_float		
127. GL_NV_texture_compression_s3tc		
128. GL_NV_3dvision_settings		
129. GL_NV_texture_compression_latc		
130. GL_NV_platform_binary		
131. GL_NV_pack_subimage		
132. GL_NV_texture_array		
133. GL_NV_pixel_buffer_object		
134. GL_NV_bgr		
135. GL_OES_depth_texture_cube_map		YES
136. GL_EXT_color_buffer_float		CORE
137. GL_ANGLE_depth_texture		

Table 18. OpenGL ES extension support...continued

Extension Number, Name and hyperlink	ES1.1	ES2.0/3.0/3.1/3.2
138. GL_ANGLE_program_binary		
139. GL_IMG_texture_compression_pvrtc2		
140. GL_NV_draw_instanced		
141. GL_NV_framebuffer_blit		
142. GL_NV_framebuffer_multisample		
143. GL_NV_generate_mipmap_sRGB		
144. GL_NV_instanced_arrays		
145. GL_NV_shadow_samplers_array		
146. GL_NV_shadow_samplers_cube		
147. GL_NV_sRGB_formats		
148. GL_NV_texture_border_clamp		
149. GL_EXT_disjoint_timer_query		
150. GL_EXT_draw_buffers		
151. GL_EXT_texture_sRGB_decode		YES
152. GL_EXT_sRGB_write_control		
153. GL_EXT_texture_compression_s3tc		YES
154. GL_EXT_pvrtc_sRGB		
155. GL_EXT_instanced_arrays		
156. GL_EXT_draw_instanced		
157. GL_NV_copy_buffer		
158. GL_NV_explicit_attrib_location		
159. GL_NV_non_square_matrices		
160. GL_EXT_shader_integer_mix		
161. GL_OES_texture_compression_astc		
162. GL_NV_blend_equation_advanced		
GL_NV_blend_equation_advanced_coherent		
163. GL_INTEL_performance_query		
164. GL_ARM_shader_framebuffer_fetch		
165. GL_ARM_shader_framebuffer_fetch_depth_stencil		
166. GL_EXT_shader_pixel_local_storage		
167. GL_KHR_blend_equation_advanced		CORE
GL_KHR_blend_equation_advanced_coherent		
168. GL_OES_sample_shading		CORE
169. GL_OES_sample_variables		CORE
170. GL_OES_shader_image_atomic		CORE
171. GL_OES_shader_multisample_interpolation		CORE

Table 18. OpenGL ES extension support...continued

Extension Number, Name and hyperlink	ES1.1	ES2.0/3.0/3.1/3.2
172. GL_OES_texture_stencil8		CORE
173. GL_OES_texture_storage_multisample_2d_array		CORE
174. GL_EXT_copy_image		CORE
175. GL_EXT_draw_buffers_indexed		CORE
176. GL_EXT_geometry_shader		CORE
GL_EXT_geometry_point_size		CORE
177. GL_EXT_gpu_shader5		CORE
178. GL_EXT_shader_implicit_conversions		CORE
179. GL_EXT_shader_io_blocks		CORE
180. GL_EXT_tessellation_shader		CORE
GL_EXT_tessellation_point_size		CORE
181. GL_EXT_texture_border_clamp		CORE
182. GL_EXT_texture_buffer		CORE
183. GL_EXT_texture_cube_map_array		CORE
184. GL_EXT_texture_view		
185. GL_EXT_primitive_bounding_box		CORE
186. GL_ANDROID_extension_pack_es31a		CORE
187. GL_EXT_compressed_ETC1_RGB8_sub_texture		
188. GL_KHR_robust_buffer_access_behavior		YES
189. GL_KHR_robustness		YES
190. GL_KHR_context_flush_control		
GLX_ARB_context_flush_control		
WGL_ARB_context_flush_control		
191. GL_DMP_program_binary		
192. GL_APPLE_clip_distance		
193. GL_APPLE_color_buffer_packed_float		
194. GL_APPLE_texture_packed_float		
195. GL_NV_internalformat_sample_query		
196. GL_NV_bindless_texture		
197. GL_NV_conditional_render		
198. GL_NV_path_rendering		
199. GL_NV_image_formats		
200. GL_NV_shader_noperspective_interpolation		
201. GL_NV_viewport_array		
202. GL_EXT_base_instance		
203. GL_EXT_draw_elements_base_vertex		CORE

Table 18. OpenGL ES extension support...continued

Extension Number, Name and hyperlink	ES1.1	ES2.0/3.0/3.1/3.2
204. GL_EXT_multi_draw_indirect		CORE
205. GL_EXT_render_snorm		
206. GL_EXT_texture_norm16		
207. GL_OES_copy_image		CORE
208. GL_OES_draw_buffers_indexed		CORE
209. GL_OES_geometry_shader		CORE
210. GL_OES_gpu_shader5		CORE
211. GL_OES_primitive_bounding_box		CORE
212. GL_OES_shader_io_blocks		CORE
213. GL_OES_tessellation_shader		CORE
GL_OES_tessellation_point_size		CORE
214. GL_OES_texture_border_clamp		CORE
215. GL_OES_texture_buffer		CORE
216. GL_OES_texture_cube_map_array		CORE
217. GL_OES_texture_view		CORE
218. GL_OES_draw_elements_base_vertex		CORE
219. GL_OES_EGL_image_external_essl3		CORE
220. GL_EXT_texture_sRGB_R8		
221. GL_EXT_YUV_target		
222. GL_EXT_texture_sRGB_RG8		
223. GL_EXT_float_blend		
224. GL_EXT_post_depth_coverage		
225. GL_EXT_raster_multisample		
226. GL_EXT_texture_filter_minmax		
227. GL_NV_conservative_raster		
228. GL_NV_fragment_coverage_to_color		
229. GL_NV_fragment_shader_interlock		
230. GL_NV_framebuffer_mixed_samples		
231. GL_NV_fill_rectangle		
232. GL_NV_geometry_shader_passthrough		
233. GL_NV_path_rendering_shared_edge		
234. GL_NV_sample_locations		
235. GL_NV_sample_mask_override_coverage		
236. GL_NV_viewport_array2		
237. GL_NV_polygon_mode		
238. GL_EXT_buffer_storage		

Table 18. OpenGL ES extension support...continued

Extension Number, Name and hyperlink	ES1.1	ES2.0/3.0/3.1/3.2
239. GL_EXT_sparse_texture		
240. GL_OVR_multiview		
241. GL_OVR_multiview2		
242. GL_KHR_no_error		
243. GL_INTEL_framebuffer_CMAA		
244. GL_EXT_blend_func_extended		
245. GL_EXT_multisample_compatibility		
246. GL_KHR_texture_compression_astc_sliced_3d		
247. GL_OVR_multiview_multisampled_render_to_texture		
248. GL_IMG_texture_filter_cubic		
249. GL_EXT_polygon_offset_clamp		
250. GL_EXT_shader_pixel_local_storage2		
251. GL_EXT_shader_group_vote		
252. GL_IMG_framebuffer_downsample		
253. GL_EXT_protected_textures		
254. GL_EXT_clip_cull_distance		
255. GL_NV_viewport_swizzle		
256. GL_EXT_sparse_texture2		
257. GL_NV_gpu_shader5		
258. GL_NV_shader_atomic_fp16_vector		
259. GL_NV_conservative_raster_pre_snap_triangles		
260. GL_EXT_window_rectangles		
261. GL_EXT_shader_non_constant_global_initializers		
262. GL_INTEL_conservative_rasterization		
263. GL_NVX_blend_equation_advanced_multi_draw_buffers		
264. GL_OES_viewport_array		
265. GL_EXT_conservative_depth		

3.4 Extension GL_VIV_direct_texture

Name	VIV_direct_texture
Name strings	GL_VIV_direct_texture
IPStatus	Contact NXP Semiconductor regarding any intellectual property questions associated with this extension.
Status	Implemented: July, 2011
Version	Last modified: 29 July, 2011 Revision: 2
Number	Unassigned

Dependencies	OpenGL ES 1.1 is required. OpenGL ES 2.0/3.x support is available.
Overview	Create a texture with direct access support. This is useful when an application desires to use the same texture over and over while frequently updating its content. It could also be used for mapping live video to a texture. A video decoder could write its result directly to the texture and then the texture could be directly rendered onto a 3D shape. glTexDirectVIVMap is similar to glTexDirectVIV. The only difference is that it has two inputs, "Logical" and "Physical," which support mapping a user space memory or a physical address into the texture surface.

3.4.1 New Procedures and Functions

glTexDirectVIV

Syntax:

```
GL_API void GL_APIENTRY
glTexDirectVIV(
    GLenum Target,
    GLsizei Width,
    GLsizei Height,
    GLenum Format,
    GLvoid ** Pixels
);
```

Parameters

Target	Target texture. Must be GL_TEXTURE_2D.
Width Height	Size of LOD 0. Width must be 16 pixel aligned. The width and height of LOD 0 of the texture is specified by the Width and Height parameters. The driver may auto-generate the rest of LODs if the hardware supports high quality scaling (for non-power of 2 textures) and LOD generation. If the hardware does not support high quality scaling and LOD generation, the texture remains a single-LOD texture.
Format	Choose the format of the pixel data from the following formats: GL_VIV_YV12, GL_VIV_NV12, GL_VIV_NV21, GL_VIV_YUY2, GL_VIV_UYVY, GL_RGBA, and GL_BGRA_EXT. <ul style="list-style-type: none">• If the format is GL_VIV_YV12, glTexDirectVIV creates a planar YV12 4:2:0 texture and the format of the Pixels array is as follows: Yplane, Vplane, Uplane.• If the format is GL_VIV_NV12, glTexDirectVIV creates a planar NV12 4:2:0 texture and the format of the Pixels array is as follows: Yplane, UVplane.• If the format is GL_VIV_NV21, glTexDirectVIV creates a planar NV21 4:2:0 texture and the format of the Pixels array is as follows: Yplane, VUplane.• If the format is GL_VIV_YUY2 or GL_VIV_UYVY, glTexDirectVIV creates a packed 4:2:2 texture and the Pixels array contains only one pointer to the packed YUV texture.• If Format is GL_RGBA, glTexDirectVIV creates a pixel array with four GL_UNSIGNED_BYTE components: the first byte for red pixels, the second byte for green pixels, the third byte for blue, and the fourth byte for alpha.• If Format is GL_BGRA_EXT, glTexDirectVIV creates a pixel array with four GL_UNSIGNED_BYTE components: the first byte for blue pixels, the second byte for green pixels, the third byte for red, and the fourth byte for alpha.
Pixels	Stores the memory pointer created by the driver.

Output

If the function succeeds, it returns a pointer, or, for some YUV formats, it returns a set of pointers that directly point to the texture. The pointer(s) are returned in the user-allocated array pointed to by the Pixels parameter.

GLTexDirectVIVMap

Syntax:

```
GL_API void GL_APIENTRY
glTexDirectVIVMap (
    GLenum Target,
    GLsizei Width,
    GLsizei Height,
    GLenum Format,
    GLvoid ** Logical,
    const GLuint * Physical
);
```

Parameters

Target	Target texture. Must be GL_TEXTURE_2D.
Width Height	Size of LOD 0. Width must be 16 pixel aligned. See glTexDirectVIV.
Format	Same as glTexDirectVIV Format.
Logical	Pointer to the logical address of the application-defined texture buffer. Logical address must be 64 bit (8 byte) aligned.
Physical	Pointer to the physical address of the application-defined buffer to the texture, or ~0 if no physical address has been provided.

GLTexDirectInvalidateVIV

Syntax:

```
GL_API void GL_APIENTRY
glTexDirectInvalidateVIV (
    GLenum Target
);
```

Parameters

Target	Target texture. Must be GL_TEXTURE_2D.
--------	--

New Tokens

GL_VIV_YV12	0x8FC0
GL_VIV_NV12	0x8FC1
GL_VIV_YUY2	0x8FC2
GL_VIV_UYVY	0x8FC3
GL_VIV_NV21	0x8FC4

Error codes

GL_INVALID_ENUM	Target is not GL_TEXTURE_2D, or format is not a valid format.
GL_INVALID_VALUE	Width or Height parameter is less than 1.
GL_OUT_OF_MEMORY	A memory allocation error occurred.

GL_INVALID_OPERATION

Specified format is not supported by the hardware, or no texture is bound to the active texture unit, or some other error occurs during the call.

Example 1.

First, call `glTexDirectVIV` to get a pointer.

Second, copy the texture data to this memory address.

Then, call `glTexDirectInvalidateVIV` to apply the texture before drawing something with that texture.

```
... ..
glTexDirectVIV(GL_TEXTURE_2D, 512, 512, GL_VIV_YV12, &texels);
... ..
glTexDirectInvalidateVIV(GL_TEXTURE_2D);
... ..
glDrawArrays(...);
... ..
```

Example 2.

First, call `glTexDirectVIVMap` to map Logical and Physical address to the texture.

Second, modify Logical and Physical data.

Then, call `glTexDirectInvalidateVIV` to apply the texture before drawing something with that texture.

```
... ..
char *Logical = (char*) malloc (sizeof(char)*size);
GLuint physical = ~0U;
glTexDirectVIVMap(GL_TEXTURE_2D, 512, 512, GL_VIV_YV12, (void**)&Logical,
    &physical);
... ..
glTexDirectInvalidateVIV(GL_TEXTURE_2D);
... ..
glDrawArrays(...);
```

Issues

None

3.5 Extension GL_VIV_texture_border_clamp**Name**

VIV_texture_border_clamp

Name Strings

GL_VIV_texture_border_clamp

Status

Implemented September 2012.

Version

Last modified: 27 September 2012

Vivante revision: 1

Number

Unassigned

Dependencies

This extension is implemented for use with OpenGL ES 1.1 and OpenGL ES 2.0.

This extension is based on OpenGL ARB Extension #13: GL_ARB_texture_border_clamp: www.opengl.org/registry/specs/ARB/texture_border_clamp.txt. See also vendor extension GL_SGIS_texture_border_clamp: www.opengl.org/registry/specs/SGIS/texture_border_clamp.txt.

Overview

This extension was adapted from the OpenGL extension for use with OpenGL ES implementations. The OpenGL ARB Extension 13 description applies here as well:

“The base OpenGL provides clamping such that the texture coordinates are limited to exactly the range [0,1]. When a texture coordinate is clamped using this algorithm, the texture sampling filter straddles the edge of the texture image, taking 1/2 its sample values from within the texture image, and the other 1/2 from the texture border. It is sometimes desirable for a texture to be clamped to the border color, rather than to an average of the border and edge colors.

This extension defines an additional texture clamping algorithm. CLAMP_TO_BORDER_[VIV] clamps texture coordinates at all mipmap levels such that NEAREST and LINEAR filters return only the color of the border texels.”

The color returned is derived only from border texels and cannot be configured.

Issues

None

New Tokens

Accepted by the <param> parameter of TexParameteri and TexParameterf, and by the <params> parameter of TexParameteriv and TexParameterfv, when their <pname> parameter is TEXTURE_WRAP_S, TEXTURE_WRAP_T, or TEXTURE_WRAP_R:

CLAMP_TO_BORDER_VIV	0x812D
---------------------	--------

Errors

None.

New State

Only the type information changes for these parameters.

See OES 2.0 Specification Section 3.7.4, page 75-76, Table 3.10, “Texture parameters and their values.”

4 Vivante Framebuffer API

4.1 Overview

The graphics software includes i.MX Framebuffer (FB) API which enables users to easily create and port their graphics applications by using a framebuffer device without the need to expend additional effort handling platform-related tasks. i.MX Framebuffer API focuses on providing mechanisms for controlling display, window, and pixmap render surfaces.

The EGL Native Platform Graphics Interface provides mechanisms for creating rendering surfaces onto which client APIs can draw, creating graphics contexts for client APIs, and synchronizing drawing by client APIs as well as native platform rendering APIs. This enables seamless rendering using Khronos APIs such as OpenGL ES and OpenVG for high-performance, accelerated, mixed-mode 2D, and 3D rendering. For further information on EGL, see www.khronos.org/registry/egl. The API described in this document is compatible with EGL version 1.4 of the specification.

Note:

i.MX 8 and later on Linux OS supports Direct Rendering Manager (DRM) where the Linux framebuffer support is limited, recommended to use the Graphics Buffer Manager (GBM).

4.2 API data types and environment variables

4.2.1 Data types

The GPU software provides platform independent member definitions for the following EGL types:

```
typedef struct _FBDisplay      * EGLNativeDisplayType;
typedef struct _FBWindow * EGLNativeWindowType;
typedef struct _FBPixmap      * EGLNativePixmapType;
```

Types [2.1.1]

unsigned int	EGLBoolean
unsigned int	EGLenum
void	*EGLConfig
void	*EGLContext
void	*EGLDisplay
void	*EGLSurface
void	*EGLClientBuffer

The following types differ based on platform.

Windows platform:

HDC	EGLNativeDisplayType
HBITMAP	EGLNativePixmapType
HWND	EGLNativeWindowType

Linux/X11 platform:

Display	*EGLNativeDisplayType
Pixmap	EGLNativePixmapType
Window	EGLNativeWindowType

Android platform:

ANativeWindow*	EGLNativeWindowType
----------------	---------------------

Figure 4. Types as listed on EGL 1.4 API Quick Reference Card

(from www.khronos.org/files/egl-1-4-quick-reference-card.pdf)

4.2.2 Environment variables

Table 19. i.MX FB API environment variables

Environment Variables	Description
FB_MULTI_BUFFER	<p>To use multiple-buffer rendering, set the environment variable FB_MULTI_BUFFER to an unsigned integer value, which indicates the number of buffers required. The maximum is 8.</p> <p>Recommended values: 4.</p> <p>The FB_MULTI_BUFFER variable can be set to any positive integer value.</p> <ul style="list-style-type: none">• If set to 1, the multiple-buffer function is not enabled, and the VSYNC is also disabled, so there may be tearing on screen, but it is good for benchmark test.• If set to 2 or 3, VSYNC is enabled and there are double or triple frame buffer. Because of the hardware limitation of current IPU, there may be tearing on screen.• If set to 4 or more, VSYNC is enabled and no screen tearing appears.• If set to a value more than 8, the driver uses 8 as the buffer count.

Table 19. i.MX FB API environment variables...continued

Environment Variables	Description
FB_FRAMEBUFFER_0, FB_FRAMEBUFFER_1, FB_FRAMEBUFFER_2, FB_FRAMEBUFFER_n	To open a specified framebuffer device, set the environment variable FB_FRAMEBUFFER_n to a proper value (for example, FB_FRAMEBUFFER_0 = /dev/fb0). Allowed values for n: any positive integer. Note: If there are no environment variables set, the driver tries to use the default framebuffer devices (fb0 for index 0, fb1 for index 1, fb2 for index 2, fb3 for index 3, and so on).
FB_IGNORE_DISPLAY_SIZE	When set to a positive integer and a window's initial size request is greater than the display size, the window size is not reduced to fit within the display. Global. Allowed values: any positive integer. Note: The drivers read the value from this environment variable as a Boolean to check if the user wants to ignore the display size when creating a window. <ul style="list-style-type: none"> If the variable is set to value 0, or this environment variable is not set, when creating window, the driver uses display size to cut down the size of the window to ensure that the entire window area is inside the display screen. If the user sets this variable to 1, or any positive integer value, then the window area can be partly or entirely outside of the display screen area (see the image below in which the ignore display size is equal to 1).
GPU_VIV_DISABLE_CLEAR_FB	It turns off zero fill memory, so the content of FBDEV buffer is not cleared.
FB_LEGACY	If the board supports drm-fb, the GPU will render though DRM by default. If the user wants to render to framebuffer directly instead of through DRM, set this variable to 1 .

Below are some usage syntax examples for environment variables:

To create a window with its size different from the display size, use the environment variable FB_IGNORE_DISPLAY_SIZE. Example usage syntax:

```
export FB_IGNORE_DISPLAY_SIZE=1
```

To let the driver use multiple buffers to do swap work, use the environment variable FB_MULTI_BUFFER. Example usage syntax:

```
export FB_MULTI_BUFFER=2
```

To specify the display device, use the environment variable FB_FRAMEBUFFER_n, where n = any positive integer. Example usage syntax:

```
export FB_FRAMEBUFFER_0=/dev/fb0
```

```
export FB_FRAMEBUFFER_1=/dev/fb1
```

```
export FB_FRAMEBUFFER_2=/dev/fb2
```

```
export FB_FRAMEBUFFER_3=/dev/fb3
```

4.3 API description and syntax

fbGetDisplay:

Description	This function is used to get the default display of the framebuffer device. To open the framebuffer device, set an environment variable FB_FRAMEBUFFER_n to the framebuffer location.
Syntax	EGLNativeDisplayType

	<pre>fbGetDisplay (void * context);</pre>
Parameters	context: Pointer to the native display instance.
Return Values	The function returns a pointer to the EGL native display instance if successful; otherwise, it returns a NULL pointer.

fbGetDisplayByIndex:

Description	<p>This function is used to get a specified display within a multiple framebuffer environment by providing an index number.</p> <p>To use multiple buffers when rendering, set the environment variable <code>FB_MULTI_BUFFER</code> to an unsigned integer value, which indicates the number of buffers. Maximum is 3.</p> <p>To open a specific Framebuffer device, set environment variables to their proper values (e.g., set <code>FB_FRAMEBUFFER_0 = /dev/fb0</code>). If there are no environment variables set, the driver tries to use the default fb devices (fb0 for index 0, fb1 for index 1, fb2 for index 2, fb3 for index 3, and so on).</p>
Syntax	<pre>EGLNativeDisplayType fbGetDisplayByIndex (int DisplayIndex);</pre>
Parameters	<p>DisplayIndex:</p> <p>An integer value where the integer is associated with one of the following environment variables for framebuffer devices:</p> <pre>FB_FRAMEBUFFER_0 FB_FRAMEBUFFER_1 FB_FRAMEBUFFER_2 FB_FRAMEBUFFER_n</pre>
Return Value	The function returns a pointer to the EGL native display instance if successful; otherwise, it returns a NULL pointer.

fbGetDisplayGeometry:

Description	This function is used to get display width and height information.
Syntax	<pre>void fbGetDisplayGeometry (EGLNativeDisplayType Display, int * Width, int * Height</pre>

	<code>);</code>
Parameters	<p>Display: [in] Pointer to EGL native display instance created by <code>fbGetDisplay</code>.</p> <p>Width: [out] Pointer that receives the width of the display.</p> <p>Height: [out] Pointer that receives the height of the display.</p>

fbGetDisplayInfo:

Description	This function is used to get display information.
Syntax	<pre>void fbGetDisplayInfo (EGLNativeDisplayType Display, int * Width, int * Height, unsigned long * Physical, int * Stride, int * BitsPerPixel);</pre>
Parameters	<p>Display: [in] A pointer to the EGL native display instance created by <code>fbGetDisplay</code>.</p> <p>Width: [out] A pointer to the location that contains the width of the display.</p> <p>Height: [out] A pointer to the location that contains the height of the display.</p> <p>Physical: [out] A pointer to the location that contains the physical start address of the display.</p> <p>Stride: [out] A pointer to the location that contains the stride of the display.</p> <p>BitsPerPixel: [out] A pointer to the location that contains the pixel depth of the display.</p>

fbDestroyDisplay:

Description	This function is used to destroy a display.
Syntax	<pre>void fbDestroyDisplay (EGLNativeDisplayType Display);</pre>
Parameters	<p>Display: [in] Pointer to EGL native display instance created by <code>fbGetDisplay</code>.</p>

fbCreateWindow:

Description	This function is used to create a window for the framebuffer platform with the specified position and size. If width/height is 0, it uses the display width/height as its value.
-------------	--

	<p>Note: When either window X + width or the Y + height is larger than the display's width or height respectively, the API reduces the window size to force the whole window inside the display screen limits. To avoid reducing the window size in this scenario, users can set a value of "1" to the environment variable FB_IGNORE_DISPLAY_SIZE.</p>
Syntax	<pre>EGLNativeWindowType fbCreateWindow (EGLNativeDisplayType Display, int X, int Y, int Width, int Height);</pre>
Parameters	<p>Display: [in] Pointer to EGL native display instance created by fbGetDisplay.</p> <p>X: [in] Specifies the initial horizontal position of the window.</p> <p>Y: [in] Specifies the initial vertical position of the window.</p> <p>Width: [in] Specifies the width of the window.</p> <p>Height: [in] Specifies the height of the window in device units.</p>
Return Value	<p>The function returns a pointer to the EGL native window instance if successful; otherwise, it returns a NULL pointer.</p>

fbGetWindowGeometry:

Description	<p>This function is used to get window position and size information.</p>
Syntax	<pre>void fbGetWindowGeometry (EGLNativeWindowType Window, int * X, int * Y, int * Width, int * Height);</pre>
Parameters	<p>Window: [in] Pointer to EGL native window instance created by fbCreateWindow.</p> <p>X: [out] Pointer that receives the horizontal position value of the window.</p> <p>Y: [out] Pointer that receives the vertical position value of the window.</p> <p>Width: [out] Pointer that receives the width value of the window.</p> <p>Height: [out] Pointer that receives the height value of the window.</p>

fbGetWindowInfo:

Description	This function is used to get window position and size and address information.
Syntax	<pre> void fbGetWindowInfo (EGLNativeWindowType Window, int * X, int * Y, int * Width, int * Height, int * BitsPerPixel, unsigned int * Offset); </pre>
Parameters	<p>Window: [in] A pointer to the EGL native window instance created by <code>fbCreateWindow</code>.</p> <p>X: [out] A pointer to the location that contains the horizontal position value of the window.</p> <p>Y: [out] A pointer to the location that contains the vertical position value of the window.</p> <p>Width: [out] A pointer to the location that contains the width of the window.</p> <p>Height: [out] A pointer to the location that contains the height of the window.</p> <p>BitsPerPixel: [out] A pointer to the location that contains the pixel depth of the window.</p> <p>Offset: [out] A pointer to the location that contains the offset of the window.</p>

fbDestroyWindow:

Description	This function is used to destroy a window.
Syntax	<pre> void fbDestroyWindow (EGLNativeWindowType Window); </pre>
Parameters	Window: [in] Pointer to EGL native window instance created by <code>fbCreateWindow</code> .

fbCreatePixmap:

Description	This function is used to create a pixmap of a specific size on the specified framebuffer device. If either the width or height is 0, the function fails to create a pixmap and return NULL.
Syntax	<pre> EGLNativePixmapType fbCreatePixmap (EGLNativeDisplayType Display, int Width, int Height); </pre>

Parameters	Display: [in] Pointer to the EGL native display instance created by fbGetDisplay. Width: [in] Specifies the width of the pixmap. Height: [in] Specifies the height of the pixmap.
Return Value	The function returns a pointer to the EGL native pixmap instance if successful; otherwise, it returns a NULL pointer.

fbCreatePixmapWithBpp:

Description	This function is used to create a pixmap of a specific size and bit depth on the specified framebuffer device. If either the width or height is 0, the function fails to create a pixmap and return NULL.
Syntax	<pre>EGLNativePixmapType fbCreatePixmapWithBpp (EGLNativeDisplayType Display, int Width, int Height int BitsPerPixel);</pre>
Parameters	Display: [in]A pointer to the EGL native display instance created by fbGetDisplay. Width: [in] Specifies the width of the pixmap. Height: [in] Specifies the height of the pixmap. BitsPerPixel: [in] Specifies the bit depth of the pixmap.
Return Value	The function returns a pointer to the EGL native pixmap instance if successful; otherwise, it returns a NULL pointer.

fbGetPixmapGeometry:

Description	This function is used to get pixmap size information.
Syntax	<pre>void fbGetPixmapGeometry (EGLNativePixmapType Pixmap, int * Width, int * Height);</pre>
Parameters	Pixmap: [in] Pointer to the EGL native pixmap instance created by fbCreatePixmap. Width: [out] Pointer that receives a width value for pixmap. Height: [out] Pointer that receives a height value for pixmap.

fbGetPixmapInfo:

Description	This function is used to get pixmap size and depth information.
Syntax	<pre>void fbGetPixmapInfo (EGLNativePixmapType Pixmap, int * Width, int * Height int * BitsPerPixel int * Stride, void ** Bits);</pre>
Parameters	<p>Pixmap: [in] A pointer to the EGL native pixmap instance created by fbCreatePixmap.</p> <p>Width: [out] A pointer to the location that contains a width value for pixmap.</p> <p>Height: [out] A pointer to the location that contains a height value for pixmap.</p> <p>BitsPerPixel: [out] A pointer to the location that contains the pixel depth of the pixmap.</p> <p>Stride: [out] A pointer to the location that contains the stride of the pixmap.</p> <p>Bits: [out] A pointer to the location that contains the bit address of the pixmap.</p>

fbDestroyPixmap:

Description	This function is used to destroy a pixmap.
Syntax	<pre>void fbDestroyPixmap (EGLNativePixmapType Pixmap);</pre>
Parameters	<p>Pixmap: [in] Pointer to the EGL native pixmap instance created by fbCreatePixmap.</p>

5 OpenCL

5.1 Overview

5.1.1 General description

Open Computing Language (OpenCL) is an open industry standard application programming interface (API) used to program multiple devices including GPUs, CPUs, as well as other devices organized as part of a single computational platform. The OpenCL standard targets a wide range of devices from mobile phones, tablets, PCs, and consumer electronic (CE) devices, all the way to embedded applications such as automotive and image processing functions. The API takes advantage of all resources in a platform to fully utilize all compute capability and to efficiently process the growing complexity of incoming data streams from multiple I/O (input/output) sources. I/O streams can be camera inputs, images, scientific or mathematical data, and any other form of complex data that can make use of data or task parallelism.

OpenCL uses parallel execution SIMD (single instruction, multiple data) engines found in GPUs to enhance data computational density by performing massively parallel data processing on multiple data items, across multiple compute engines. Each compute unit has its own arithmetic logic units (ALUs), including pipelined floating point (FP), integer (INT) units and a special function unit (SFU) that can perform computations as well as transcendental operations. The parallel computations and associated series of operations are called a kernel, and the GPU cores can execute a kernel on thousands of work-items in parallel at any given time.

At a high level, OpenCL provides both a programming language and a framework to enable parallel programming. OpenCL includes APIs, libraries and a runtime system to assist and support software development. With OpenCL, it is possible to write general purpose programs that can execute directly on GPUs, without needing to know graphics architecture details or using 3D graphics APIs like OpenGL or DirectX. OpenCL also provides a low-level Hardware Abstraction Layer (HAL) as well as a framework that exposes many details of the underlying hardware layer and thus allows the programmer to take full advantage of the hardware.

For more details on all the capabilities of OpenCL, see the following specifications from the Khronos Group:

- OpenCL 3.0 Specification

<https://registry.khronos.org/OpenCL/specs/3.0-unified/pdf>

- OpenCL 3 C Language Specification

https://registry.khronos.org/OpenCL/specs/3.0-unified/pdf/OpenCL_C.pdf

5.1.2 OpenCL framework

The OpenCL framework has two principal parts, similar to OpenGL, the host C API and the device C-based language runtime. The host in OpenCL terminology corresponds to the client in OpenGL and the device corresponds to the server. Device programs are called kernels. Execution of an OpenCL program is preceded by a series of API calls that configure the system and Vivante OCL-compatible IP for execution.

OpenCL abstracts today's heterogeneous architectures using a hierarchical platform model. A host coordinates the execution and data transfers on, to and from one or several compute devices. Compute devices are comprised of compute units and each such unit contains an array of processing elements.

5.1.2.1 OpenCL execution model: kernels and work elements

The OpenCL execution model is defined by how the kernels are executed. When a kernel is submitted for execution by the host, an index space is defined. An instance of the kernel executes for each point in this index space. This kernel instance is called a **work-item**. Work-items are identified by their position in the index space that provides the global ID for the work-item. Each work-item executes the same code but the specific pathway through the code and the data operated upon varies by work-item.

Work-items are organized into **work-groups**. Work-groups provide a broader decomposition of the index space. Work-groups are each assigned a unique work-group ID with the same dimensionality as the index space used for the work-items. Work-items are assigned a unique local ID within a work-group so that a single work-item can be uniquely identified by its global ID or by a combination of its local ID and work-group ID. The work-items in a given work-group execute concurrently on the same compute device.

The index space supported in OpenCL is called an **NDRange**. An NDRange is an N-dimensional index space, where **N** is one (1), two (2) or three (3). An NDRange is defined by an integer array of length N specifying the extent of the index space in each dimension starting at an offset index **F** (zero by default). Each work-item's global ID and local ID are N-dimensional tuples. The global ID components are values in the range from F, to F plus the number of elements in that dimension minus one.

Work-groups are assigned IDs using a similar approach to that used for work-item global IDs. An array of length N defines the number of work-groups in each dimension. Work-items are assigned to a work-group and given

a local ID with components in the range from zero to the size of the work-group in that dimension minus one. Hence, the combination of a work-group ID and the local-ID within a work-group uniquely defines a work-item. Each work-item is identifiable in two ways; in terms of a global index, unique through the whole kernel index space, and in terms of a local index, unique within a work group.

5.1.2.2 OpenCL command queues

OpenCL provides both task and data parallelism. Data movements are coordinated via **command queues**, which provide a general means of specifying inter-task relationships and task execution orders that obey the dependencies in the computation. OpenCL may execute several tasks in parallel, if they are not order dependent. Tasks are composed of data-parallel kernels which, similarly to shaders, apply a single function to a range of elements in parallel. Only restricted synchronization and communication is allowed during kernel execution.

OpenCL kernels execute over a 1, 2 or 3 dimensional index space. All work-items execute the same program (kernel) but their execution may diverge, with branching dependent on the data or their index. For details regarding how many work groups are allowed within an index space see “Using clEnqueueNDRangeKernel”.

A kernel or a memory operation is first **enqueued** onto a command queue. Kernels are executed asynchronously and the host application execution may proceed right after the enqueue operation. The application may opt to wait for an operation to complete and an operation (kernel or memory) may be marked with a list of events that must occur before it executes.

Events are kernel completion and memory operations. OpenCL traverses the dependence graph between the kernels and memory transfers in a queue and ensures the correct execution order. Multiple command queues may be constructed, further enhancing parallelism control across platforms and multiple compute devices.

- **Command-queue barriers** are used to control the commands within the command queue. The command-queue barrier indicates which commands must be finished before proceeding. This allows for out-of-order command processing. The command queue barrier ensures that all previously enqueued commands finish execution before any following commands begin execution.

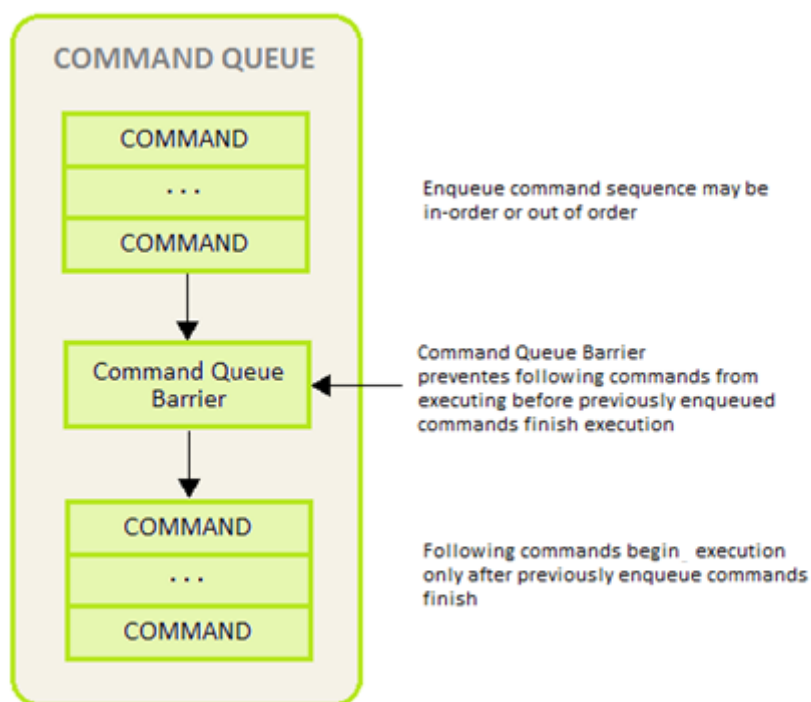


Figure 5. Command queue barrier

The work-group barrier built-in function provides control of the work-item flow within work-groups. All work-items must execute the barrier construct before any can continue execution beyond the barrier.

5.1.2.3 OpenCL memory model

The OpenCL memory model is divided into four different types of memory domains. These are:

- **Global Memory:** Each compute device has global memory space which can reside off-chip in system memory (DRAM) or inside the chip at the L1 or temporary register level. Global memory is accessible to all work-items executing in a context, as well as to the host (read, write, and map commands).
- **Constant Memory:** is also global memory, but it is read-only. Constant memory can be placed in any level of memory that the application programmer decides, making it an implementation dependent decision. This is the region for host-allocated and host-initialized objects that are not changed during kernel execution.
- **Local Memory:** Each compute unit has local memory which resides very near the processing elements. Access to local memory is very fast and the size of local memory is much smaller than global memory, making it a scarce resource that needs to be controlled for optimal communication of work-items inside a work-group. Local memory is specific to a work-group, and is accessible only by work-items belonging to that work group.
- **Private Memory:** Each processing element has another level of memory called private memory, which is only accessible to a single work-item. Private memory is specific to a work-item and is not visible to other work-items.

During run-time, each processing element is assigned a set of on-chip registers that are used for data storage of intermediate data. Data that cannot be stored in registers spills over to global memory which can be very costly in terms of performance and constant data movement to/from temporary registers. Software may emulate local and private memory using global memory. System Memory is often loaded to L1 cache, Temporary or

Local Storage Registers and the GPGPU reads from those locations. At every level of the application program, the programmer must be aware of the size and hierarchy of storage elements.

Table 20. Vivante memory structures mapped to Khronos OpenCL memory types

Khronos OpenCL Memory Model Name	Vivante GPGPU OpenCL Memory Structures Utilized	Definition
Private Memory	Registers, System Memory	Accessible only to an individual work-item; not visible to any other work-items
Local Memory	Local Storage Registers, System Memory	Accessible to all work-items within a specific work-group; accessible only by work-items belonging to that work-group
Global Memory	System Memory	Accessible to all-work-items executing in a context, as well as to the host (read, write, and map commands).
Constant Memory	Constant Registers, System Memory	Read only global memory region for host-allocated and initialized objects that are not changed during kernel execution
Host (CPU) Memory	Host Memory	Region for a kernel application's program data and structures

The OpenCL concurrent-read /concurrent-write (CRCW) memory model has so-called relaxed consistency which means that different work-items may see a different view of global memory as the computation proceeds. Within individual work-items reads and writes to all memory spaces are ordered. Synchronization between work-items in a work-group is necessary to ensure consistency. No mechanism for synchronization between work-groups is provided. Such a model assures parallel scalability by requiring explicit synchronization and communication.

For the highest throughput and computational speed, kernels should use high-speed on-chip memories and registers as much as possible. Instruction control flow and memory operations, including data gathering / scattering and direct memory access (DMA) should be automatically reorganized / re-ordered depending on data dependencies detected by the optimized compiler. The Vivante OpenCL compiler automatically maps dependencies and re-orders instructions for the best performance.

5.1.2.4 Host to Vivante compute device data transfers

The application running on the host uses the OpenCL API to create memory objects in global memory, and to enqueue memory commands that operate on these memory objects. The host and OpenCL device memory models are, for the most part, independent of each other. This is by necessity as the host is defined outside of OpenCL. They do, however, at times need to interact. This interaction occurs in one of two ways: by explicitly copying data from the host to the GPU compute device memory, or implicitly, by mapping and unmapping regions of a memory object.

- **Explicit** using `clEnqueueReadBuffer` and `clEnqueueWriteBuffer` (`clEnqueueReadImage`, `clEnqueueWriteImage`.)

To copy data explicitly, the host enqueues commands to transfer data between the memory object and host memory. These memory transfer commands may be blocking or non-blocking. The OpenCL function call for a blocking memory transfer returns once the associated memory resources on the host can be safely reused. For a non-blocking memory transfer, the OpenCL function call returns as soon as the command is enqueued regardless of whether host memory is safe to use.

- **Implicit** using `clEnqueueMapBuffer` and `clEnqueueUnMapMemObject`.

The mapping/unmapping method of interaction between the host and OpenCL memory objects allows the host to map a region from the memory object into its address space. The memory map command may be blocking

or non-blocking. Once a region from the memory object has been mapped, the host can read or write to this region. The host unmaps the region when accesses (reads and/or writes) to this mapped region by the host are complete.

The OpenCL specification does not explicitly state where each memory space will be mapped to on individual implementations. This provides great freedom for vendors on the one hand and some uncertainty for programmers on the other. Fortunately, kernels may be compiled just-in-time and possible differences may be tackled during run-time.

When using these interfaces, it is important to consider the amount of copying involved to/from system memory and the various levels within the compute device(s). There is a two-copy process: between host and AXI (or SoC internal bus), and between AXI (or SoC internal bus) and the Vivante GPGPU compute device. Double copying lowers overall system memory bandwidth and lowers performance. Because of variations in system architecture (both internal and external/memory), there is sometimes a large performance delta between the system or calculated GFLOPS and the kernel or GPGPU GFLOPS. GPGPU GFLOPS are based on the theoretical computational capability of the ALUs within the GPGPU, assuming the system architecture can deliver full data to the GPGPU. OpenCL APIs for buffers and images aid in avoiding double copy by allowing the mapping of host memory to device memory. With proper memory transfer management and the use of host/CPU memory remapped to the GPGPU memory space, copying between host memory and GPGPU memory can be skipped so data transfer becomes a one-copy process. The trade-off is that the programmer needs to be mindful of page boundaries and memory alignment issues.

5.1.3 OpenCL profiles

In addition to Full Profile, the OpenCL specification also includes an Embedded Profile, which relaxes the OpenCL compliance requirements for mobile and embedded devices. The main commons and differences between OpenCL 1.1/1.2 EP (Embedded Profile) and FP (Full Profile) come down to:

Commons:

- Both EP and FP significantly offload the CPU of parallel, multi-threaded tasks.
- For both EP and FP double precision and half-precision floating point are optional.

Difference:

- Full Profile is for highly complex, accurate, and real time computations, while Embedded Profile is a small subset targeting smaller devices (handheld, mobile, embedded) that perform GPGPU/OpenCL processing with relaxed data type and precision requirements (image processing, augmented reality, gesture recognition, and more).
- 64-bit integers are required for FP and optional for EP.
- EP requires either RTZ or RTE. FP requires both.
- Computational precision (units in the last place; i.e., ULP) requirements in EP are relaxed.
- Atomic instruction support is not required in EP.
- 3D Image support is not required in EP.
- Minimum requirements for constant buffer size, object allocation size, constant argument counts and local memory sizes are scaled down in EP.
- And more (in general EP is a scaled down version of FP).
- Die size and power increase with FP because of the higher requirements, features and memory sizes.

5.1.4 Vivante OpenCL embedded compatible IP

As of the date of this document, select Vivante GPGPU cores are compatible with OpenCL Embedded Profile version 1.1. The following table lists the hardware capability deltas.

Table 21. Vivante OpenCL embedded profile hardware

Hardware and revision	GC2000
Feature	5.1.0.rc8a
Compute Devices (GPGPU cores)	1
Compute Units per device (Shader cores)	4
Processing Elements per compute unit	4
Profile	Embedded
Preferred work-group/thread group size	16
Max count global work-items each dim	64K
Max count of work-items each dim per work-group	1K
Local Storage Registers On-chip	64
Instruction Memory	512
Texture Samplers	8 PS + 4 VS
Texture Samplers available to OCL (HW, unlimited via SW)	4
L1 Cache Size	4 KB
L1 Cache Banks	1
L1 Cache Sets/Bank	4
L1 Cache Ways/Set	16
L1 Cache Line Size	64B
L1 Cache MC ports	1

5.1.5 Vivante OpenCL full profile hardware model

As of the date of this document, select Vivante GPGPU cores are compatible with OpenCL Full Profile versions 1.1, 1.2, and 3.0. Hardware capability deltas are subject to change and includes:

Table 22. Vivante OpenCL full profile hardware

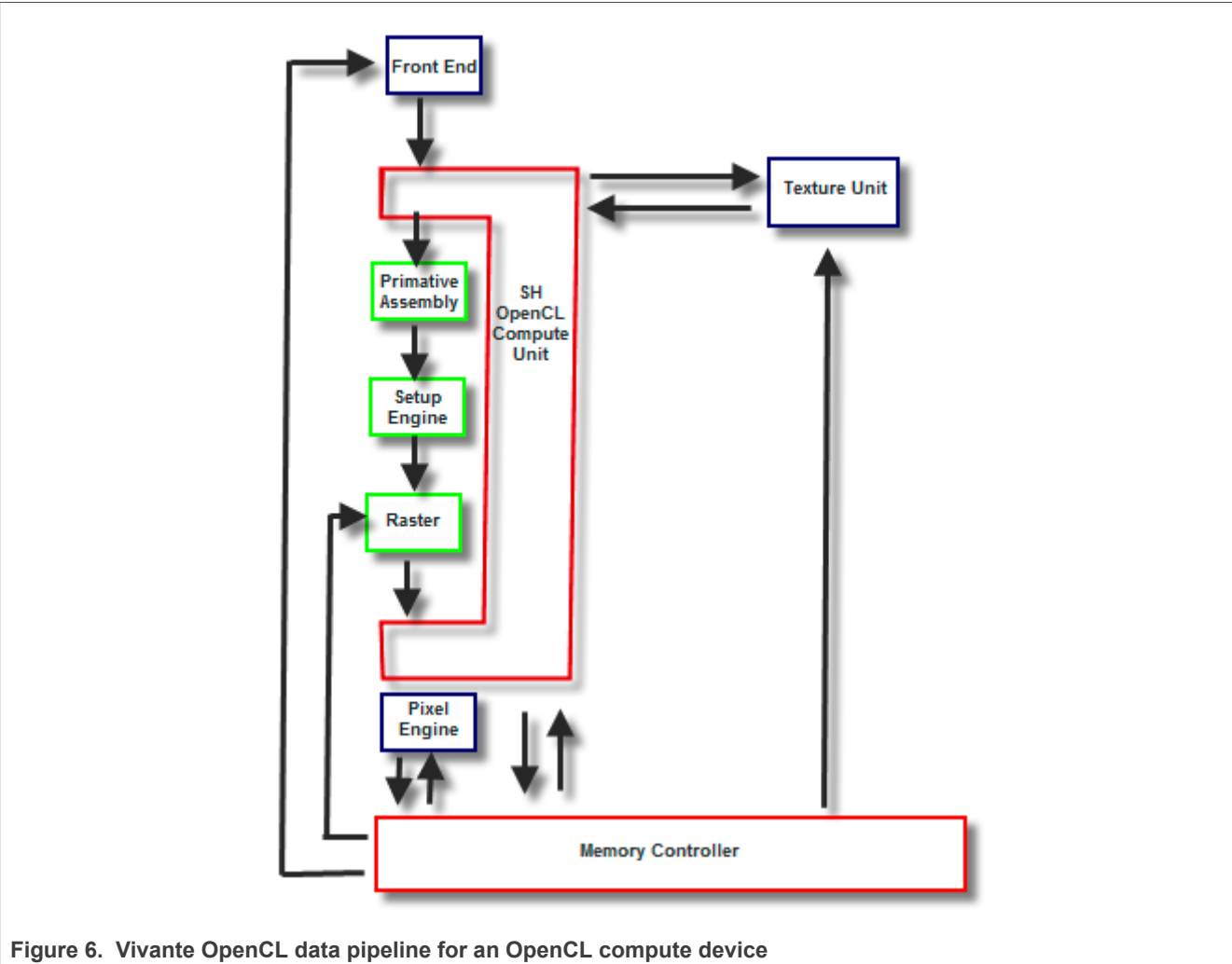
Hardware and revision	GC2000+	GC7000XSVX	GC7000L	GC7000UL
i.MX SoC	i.MX 6QuadPlus, i.MX 6DualPlus	i.MX 8 QuadMax	i.MX 8M Quad, i.MX 8QuadXPlus	i.MX 8M Nano i.MX 8M Plus
Compute Devices (GPGPU cores)	1	1	1	1
Compute Units per device (for sub-device)	1	1	1	1
Processing Elements per device	16	32	16	8
Profile	Full-Lite*	Full	Full	Full
Preferred work-group/ thread group size	16	32	16	8
Max count global work-items each dim (if 3D only 1 dim can be up to 4G, the others 64K)	4 G/64 K	4 G/64 K	4G	4G
Max count of work-items each dim per work-group	1 K	1 K	1K	1K

Table 22. Vivante OpenCL full profile hardware...continued

Local Storage Registers On-chip	0	2048 (32 K)	16 (KB)	
Instruction Memory	I\$:512/1 M	8K	8K	8K
Texture Samplers	32 undefined	32 undefined	32	32
Texture Samplers available to OCL	32	32	32	32
L1 Cache Size	4 KB	64 KB	16KB	8 KB
L1 Cache Banks	2	4	2	1
L1 Cache Sets/Bank	2	8	N/A	8
L1 Cache Ways/Set	16	8	8	8
L1 Cache Line Size	64 B	64 B	64 B	64 B
L1 Cache MC ports per GPGPU core	2	2	2	1

5.2 Vivante OpenCL implementation

5.2.1 OpenCL pipeline



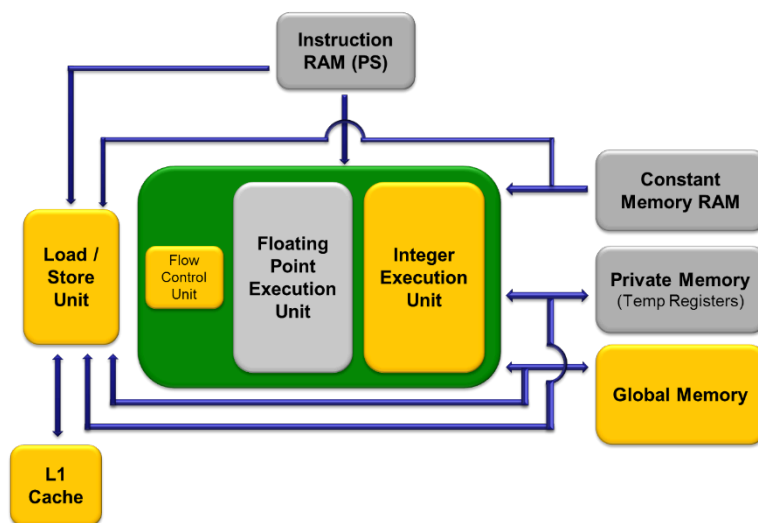


Figure 7. Vivante OpenCL compute device showing memory scheme

5.2.2 Front end

The front end passes the instructions and constant data as State Loads to the OpenCL Compute Unit (Shader) block. State Loads program instructions and constant data and work groups initiate execution on the instructions and the constants loaded.

5.2.3 OpenCL compute unit

All OpenCL executions occur in this block and all work-groups in a compute unit should belong to the same kernel. Threads from a work-group are grouped into internal “Thread-groups”. All the threads in a thread-group execute in parallel. Barrier instruction is supported to enforce synchronization within a work-group.

The compute unit contains Local Memory and the L1 Cache and is where the Load/Store instruction to access global memory originates. The compute unit can accommodate multiple work-groups (based on the temporary register and local memory usage) simultaneously.

5.2.4 Memory hierarchy

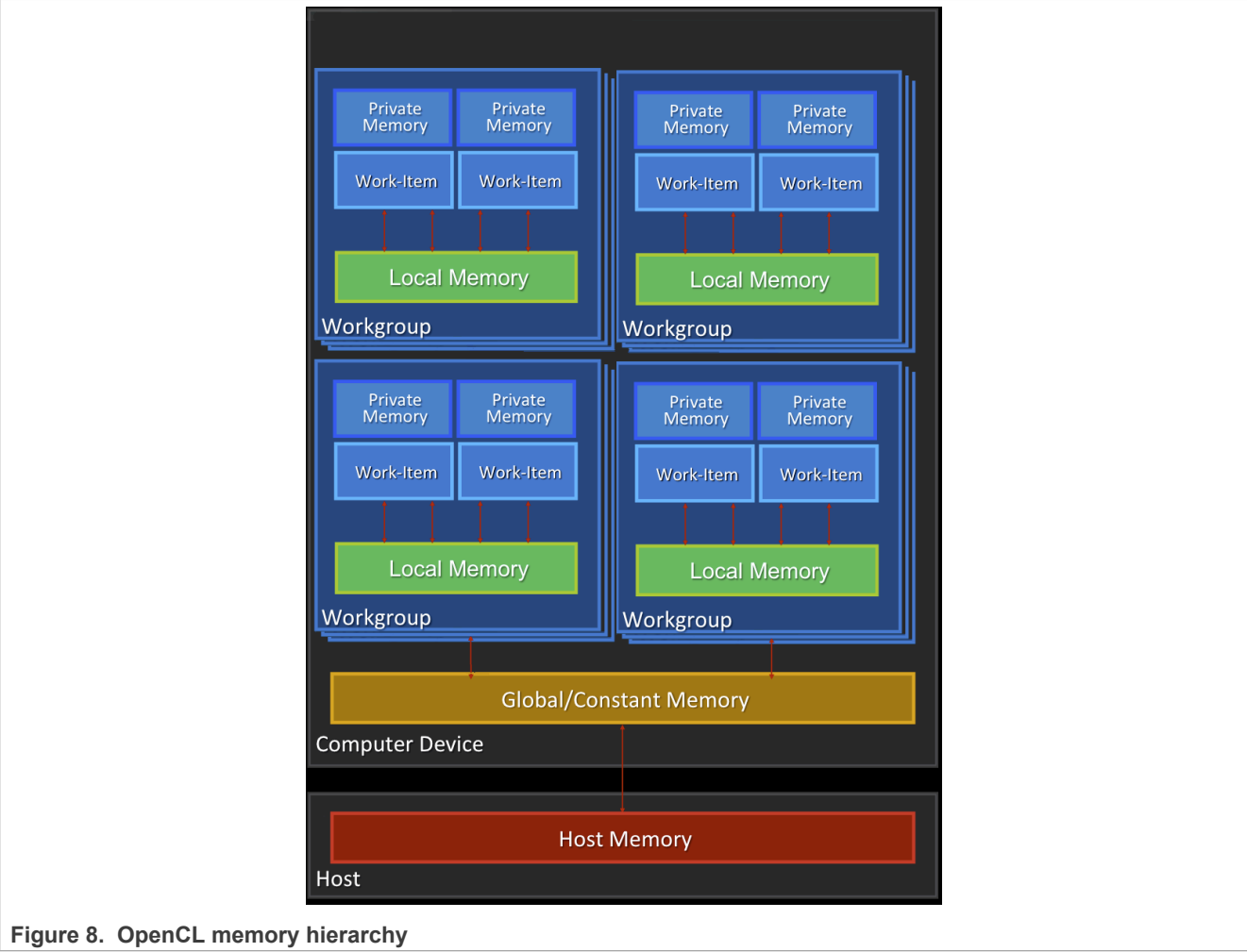


Figure 8. OpenCL memory hierarchy

5.2.5 CL Extension support

5.2.5.1 CL_DEVICE_EXTENSION support

The following table provides a list of CL_DEVICE_EXTENSIONS referenced in the OpenCL 1.2 specification (pp. 46-47). The support level for these device specific extensions is also indicated.

List from OpenCL 1.2 Specification <https://www.khronos.org/registry/OpenCL/specs/openscl-1.2.pdf> (version 1.2, document revision 19, revision date 11/14/12)

Table 23. Support level for these device specific extensions (1)

CL_DEVICE_EXTENSIONS OpenCL C 1.2 Extensions which must be returned (p. 47)	SW 6.2.x/6.4.x
cl_khr_byte_addressable_store	YES
cl_khr_fp64 (for backward compatibility if double precision is supported)	
cl_khr_global_int32_base_atomics	CORE

Table 23. Support level for these device specific extensions (1)...continued

CL_DEVICE_EXTENSIONS OpenCL C 1.2 Extensions which must be returned (p. 47)	SW 6.2.x/6.4.x
cl_khr_global_int32_extended_atomics	CORE
cl_khr_local_int32_base_atomics	CORE
cl_khr_local_int32_extended_atomics	CORE

Table 24. Support level for these device specific extensions (2)

CL_DEVICE_EXTENSIONS Device specific support for Khronos approved extension names (p.46) <i>A number after the extension name indicates the extension is also listed in the numbered extensions on the Khronos website.</i>	SW 6.2.x/6.4.x
cl_khr_3d_image_writes	
cl_khr_context_abort	
cl_khr_d3d10_sharing (#6)	
cl_khr_d3d11_sharing	
cl_khr_depth_images	
cl_khr_dx9_media_sharing	
cl_khr_fp16	
cl_khr_gl_depth_images	
cl_khr_gl_event	
cl_khr_gl_msaa_sharing	
cl_khr_gl_sharing (#1)	YES
cl_khr_image2d_from_buffer	
cl_khr_initialize_memory	
cl_khr_int64_base_atomics	
cl_khr_int64_extended_atomics	
cl_khr_spir	

5.2.5.2 Vivante OpenCL extension support

The following table provides a list of all current OpenCL Extensions and indicates their support level in Vivante software.

Table 25. CL extensions supported by Vivante with 6.2.x SW

OpenCL Extension Number, Name and hyperlink	SW 6.2.x
cl_khr_byte_addressable_store	YES
cl_khr_external_memory_dma_buf	YES (from 6.4.11)
cl_khr_command_buffer	YES (from 6.4.11)
cl_khr_gl_sharing	YES
cl_khr_icd	YES

Table 25. CL extensions supported by Vivante with 6.2.x SW...continued

OpenCL Extension Number, Name and hyperlink	SW 6.2.x
VIV_bitfield_extension	YES (from 6.2.2, revised in 6.2.3)
VIV_cmplx_extension	YES (from 6.2.3)
VIV_uncached_host_mem	YES (from 6.2.2)
VIV_vx_extension	YES, for VX/VIP hw (from 6.2.2)
cl_khr_fp16	YES (from 6.4.7)
cl_khr_il_program	YES (from 6.4.8)

5.3 Optimization for OpenCL embedded profile

OpenCL EP (Embedded Profile) is basically a scaled down version of OpenCL FP(Full Profile) and thus may require extra optimization. The guidelines below help with the optimization of Vivante OpenCL Embedded Profile GPGPU cores.

When optimizing code on Vivante hardware, it is important to remember a few key points to get the best performance from the hardware:

- Take advantage of algorithm and data parallelism
- Choose the correct execution configuration (more details below)
- Overlap memory transfer from different levels of the OpenCL memory hierarchy with simultaneous thread execution
- Maximize memory bandwidth and minimize data transfers (large transfers are more beneficial than many smaller transfers because of the impact of latency)
- Maximize instruction throughput and minimize instruction count

5.3.1 Using preferred multiple of work-group size

The work-group size should be a multiple of the thread group size. Otherwise, some threads remain idle and the application does not fully utilize all the compute resources. For example, if the work-group size is 8 and the Vivante core supports 16, only half the compute resources are used. For example, in some early Vivante GPGPU revisions, the work-group size limit is 192 and the thread group size is 16. See the Overview section on OpenCL Compatible IP for IP-specific capabilities.

5.3.2 Using multiple work-groups of reduced size

Multiple work groups need to be set to reduce synchronization penalties. To prevent stalls at barriers, it is recommended to have at least four (4) work-groups to keep the cores busy or as long as the number of work-groups is greater than or equal to two (2). One work-group is very inefficient; four or more is preferred and helps avoid latency.

5.3.3 Packing work-item data

It is important to pack data to extract the optimal performance from the SIMD ALU hardware and align the data into a format supported by the hardware. Efficient use of the Vivante GPGPU core requires that the kernel contains enough parallelism to fill all four vector units. Work-items in the same thread group have the same program counter and execute the same instruction for each cycle. Whenever possible, pack together work-items that follow the same direction (e.g., on branches) since the granularity is very close and there may be

less divergence and higher performance. If each work-item handles less than or equal to 8 bytes, it is better to combine two or more work-items into one to improve utilization of the SIMD ALU.

5.3.4 Improving locality

If the input data is an array-of-structs, and each work-item needs to access only a small part of the struct across many array elements at different stages, it may be better to convert and use a struct-of-arrays or several different arrays as input to improve data locality and avoid cache thrashing.

If each work-item needs to process a row of data without sharing any data with other work-items, it is better to check if the algorithm can be converted to make each work-item process a column of data so that data accessed by adjacent work-items can share the same cache lines.

5.3.5 Minimizing use of 1 KB local memory

The OpenCL Embedded Profile specification defines the minimum requirement for local memory to be 1KB to pass conformance testing. Based on algorithm analysis and profiling different image and computer vision algorithms, we found that a 1KB local memory size was too small to benefit those algorithms. In most instances, those algorithms actually slowed down when using 1KB local memory. To increase performance, we recommend not using local memory since it is more efficient to transfer larger chunks of data from system memory to keep the OpenCL pipeline full.

Note: If local memory type is CL_GLOBAL, the local memory is emulated using global memory, and the performance is the same as global memory. There is extra overhead on data copy from global to local, which slows down the performance.

5.3.6 Using 16 byte memory Read/Write size

When accessing memory, it is important to minimize the read/write count and to ensure L1 cache utilization is high to reduce outstanding read/write requests. Since the internal GPGPU read-write-request queue has a limit, if the queue and L1 cache are filled, then the GPGPU remains idle.

5.3.7 Using _RTZ rounding mode

Wherever possible, use _RTZ (round to zero) since it is natively supported in hardware with one instruction. Support for _RTE (round to nearest even) is optional in OpenCL EP and is only supported in Vivante GPGPU EP hardware from 2013. This function is handled in software for EP cores if necessary.

5.3.8 Using float4 for better performance on i.MX 8M Quad and i.MX 8QuadXPlus

Since both the i.MX 8M Quad and i.MX 8QuadXPlus boards have new RTL 6214, the CL kernel compiler generates GPU instructions using more registers on RTL6214. Float4 is recommended for real applications for better performance.

5.3.9 Using native functions

5.3.9.1 Using native_function() for increased performance

There are two types of runtime math libraries available to developers. Native_function() and regular function().

- Function(): slower, computationally expensive, higher instruction count, and greater accuracy
- Native_function(): faster, computationally inexpensive, lower instruction count (sometimes reduced to one instruction), and lower accuracy.

- If accuracy is not important but speed/performance is, use native math functions that map directly to the Vivante GPGPU hardware.

For image processing computations that do not require high accuracy, use native instructions to significantly lower the instruction count and speed up performance. Based on actual analysis and performance profiling with the Vivante GPGPU, we found that using `native_function()` instructions such as `sin`, `cos`, etc., reduces the instruction count from many instructions to one or two instructions. Use of native functions also sped performance by 3x-10x.

5.3.9.2 Using `native_divide` and `native_reciprocal` for faster floating point calculations

There are two use cases for floating point division which a user can select:

- Normal use of the division operator (/) in OpenCL has high precision and covers all corner use cases. This operator generates more instructions and runs slower.
- Native Divide: this use case uses the built-in function `native_divide` or `native_reciprocal`, which uses what the hardware supports. The Vivante OpenCL compiler generates one or two instructions for each `native_divide` or `native_reciprocal` instruction. If there are no corner use cases in applications, such as NaN, INF, or $(2^{127}) / (2^{127})$, it is better to use `native_divide` since it is faster.

5.3.9.3 Using compile option for native functions

Both the `function()` and `native_function()` methods are supported in the Vivante GPGPUs, so it is up to the developer to use whichever method makes sense for their application. If the OpenCL program uses the standard division operator and a developer wants to use `native_divide` or `native_reciprocal` without modifying their program, the Vivante OpenCL compiler has a simple option “-cl-fast-relaxed-math” that uses native built-in functions during compilation.

5.3.10 Using buffers instead of images

For the following image functions, it is better to use buffers instead of images.

- `read_image{f/i/ui/h}`
- `write_image{f/i/ui/h}`

`Write_image*` functions are implemented by software; it is better to use buffers to reduce the additional overhead involved in checking for size, format, etc. Since a few formats are not supported by Vivante GPGPU hardware, some built-in `read_image()` functions are implemented in software. The software implementation uses more instructions with many steps of “condition” checking. To improve performance, we recommend using buffers since it reduces instruction count.

5.4 OpenCL Debug messages

When writing OpenCL applications, it is important to check the code returned by the API. Since the return codes specified in the OpenCL specification may not be descriptive enough to isolate where the problem is located, the Vivante OpenCL driver provides an environment variable, `VIV_DEBUG`, to help debug problems. When `VIV_DEBUG` is set to `-MSG_LEVEL:ERROR`, the Vivante OpenCL driver prints onscreen error messages and returns the error code to the caller.

The following error code descriptions and suggested workarounds are provided.

5.4.1 OCL-007005: (`clCreateKernel`) cannot link kernel

One of the following “Not Enough” messages usually precedes this message. Issuer indicates the real reason for the problem which may be:

- Not Enough Register Memory (constant or temp)
- Not Enough Instruction Memory

5.4.2 Not enough register memory

Local variables, including arrays, are implemented using temp registers. If an array is larger than the number of available temp registers, a link-time failure occurs.

Workarounds:

1. If the array size is more than 64, use an array address to force the compiler to use private memory instead of temp registers.
2. If there are many variables, use variable addresses to force the compiler to use private memory to reduce register usage.

Note that there is performance degradation when using private memory instead of registers. It is better to change the algorithm to use a smaller array or less variables.

5.4.3 Not enough instruction memory

Workarounds:

1. Replace `sin/cos/tan/divide/powr/exp/exp2/exp10/log/log2/log10/sqrt/rsqrt/ recip` with `native_sin/native_divide, etc.`
2. Convert unrolled-loops back to loops.
3. Use buffer instead of image for write, and for reads which are not linear-filtered.
4. If the program is too long, it should be split into two or more programs with intermediate data saved from one program to next.

5.4.4 GlobalWorkSize over hardware limit

WORKAROUND:

1. Split one `clEnqueueNDRangeKernel` into several instances. Change the kernel source to compute real global/local/group ID using offset as a parameter.
2. Convert one dimension to two dimensions, or two dimensions to three. For example, one dimension of 1M work-items can be converted to a GlobalWorkSize of 64K x16 work-items. The kernel function needs modification to reflect the change of dimension.

5.5 Zero copy

A buffer object can be created with `clCreateBuffer(cl_context context, cl_mem_flags flags, size_t size, void* host_ptr, cl_int* error_code_ret)`. If memory flags contain `CL_MEM_USE_HOST_PTR`, GPU will map the memory pointed by host ptr for GPU to use to avoid copying data between CPU and GPU.

To make sure the results are correct, the size of buffer, the third parameter of `clCreateBuffer()`, needs to be aligned with 64-byte since Arm data cache operations are performed line by line, the unaligned bits will be cleared with cache line mask. A53, A57, A72 and A73 all have 64-byte cacheline size. If the size of the buffer doesn't meet this, GPU will use copy method instead.

Besides, the `host_ptr` should be aligned with 64-bit to meet the ARM cacheline mechanism.

At last, need to call `clEnqueueReadBuffer()` to make sure the data has been read back to CPU.

5.6 Instruction cache availability for i.MX graphics

This section describes the instruction cache (iCache) available in the Vivante graphics IP included in the selected i.MX products.

There is hardware support for iCache available for i.MX 6QuadPlus and all later IP including that used in i.MX 8 products. There is no SH (Shader) instruction limit for these newer chips beyond the ISA limitation of 2*20.

Only the older chips have a SH instruction limit.

Table 26. i.MX products with graphics IP with iCache

i.MX Product	GPU IP & rev	Instruction Limit	Description
i.MX 8 Series and later	various (from rev 5450)	none	HW supports iCache
i.MX 6QuadPlus	GC2000 Plus rev FFFF5450	none	HW supports iCache
S32V234	GC3000 rev 5451	none	HW supports iCache

The SH limitation for i.MX products is listed in the following table.

Table 27. i.MX products with instruction limited graphics IP

i.MX Product	GPU IP & rev	Instruction Limit	Description
i.MX 6SoloX	GC400 rev 4645	256 for VS, 256 for PS	Separate Instruction buffers for Vertex Shader and for Pixel Shader
i.MX 7ULP	GCNanoUltra rev 4653a	256 for VS, 256 for PS	Separate Instruction buffers for Vertex Shader and for Pixel Shader
i.MX 6DualLite	GC880 rev 5106	512	Instruction buffer shared by Vertex and Pixel Shaders
i.MX 6Quad	GC2000 rev 5108	512	Instruction buffer shared by Vertex and Pixel Shaders

6 OpenCV

6.1 Overview

OpenCV is a popular open-source computer vision library that provides functions for image and video processing tasks such as object detection, image segmentation, and feature extraction. It is widely used in various applications, including robotics, autonomous vehicles, medical imaging, remote sensing, and security systems.

This section describes how to accelerate OpenCV with Arm and VSI GPUs on the i.MX.

For more details on OpenCV, see the document: [OpenCV](#)

6.2 Acceleration with OpenCL

OpenCV is accelerated with GPU using OpenCL.

OpenCL is an open standard for writing code that runs across heterogeneous platforms including CPUs, GPUs, DSPs, etc. In particular, OpenCL provides applications with an access to GPUs for non-graphical computing (GPGPU) that in some cases results in significant speed-up. In Computer Vision, many algorithms can run

on a GPU much more effectively than on a CPU, such as image processing, matrix arithmetic, computational photography, and object detection.

If OpenCL is enabled (see [Section 6.3.1](#)), OpenCV is executed using OpenCL. If OpenCL is disabled, OpenCV uses CPU.

The following sections describe how to enable OpenCL acceleration and the functions in OpenCV that can be accelerated by OpenCL.

6.3 Usages of OpenCV Accelerator

6.3.1 How to enable/disable OpenCV Accelerator

To accelerate OpenCV with OpenCL, set the `WITH_OPENCL=ON` option when building OpenCV.

When the OpenCV libraries are built with OpenCL, the environment variable `export OPENCV_OPENCL_RUNTIME=0` should be used with runtime to disable OpenCL.

6.3.2 Requirements

When the OpenCL Accelerator is enabled, ensure that the input and output data type is UMat format.

By using UMat objects, OpenCV automatically uses the GPU computing available on the device that supports OpenCL, and falls back to CPU computing on the devices that do not support OpenCL, and thus avoids program failure and unifies the interface.

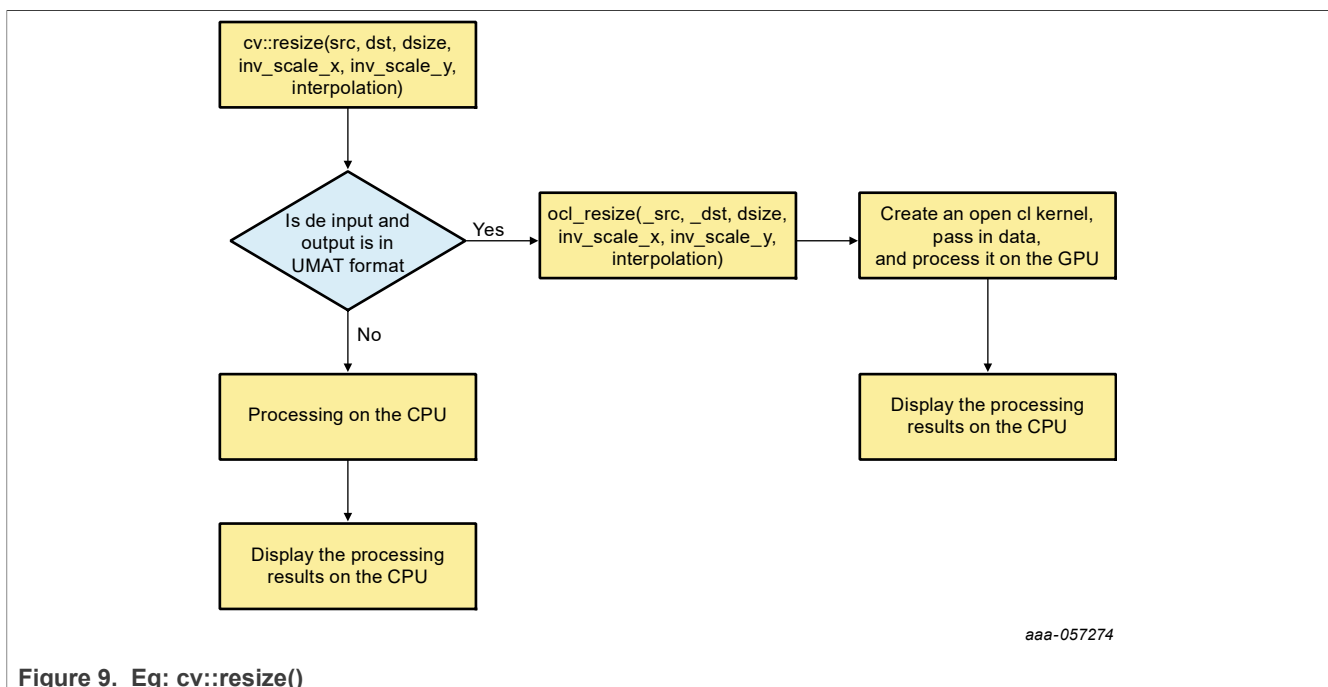


Figure 9. Eg: `cv::resize()`

Note: In OpenCV, Mat and UMat can be converted to each other.

6.4 OpenCV functions accelerated with OpenCL

This section describes the OpenCV functions that can be accelerated using OpenCL and the conditions for using them.

6.4.1 OpenCV function list

The following table lists the function APIs in OpenCV that can be accelerated using OpenCL.

Table 28. OpenCV function list

Function name	Functional description
pyrUP	Upsamples an image and then blurs it.
warpPerspective	Applies a perspective transformation to an image.
warpAffine	Applies an affine transformation to an image.
match Template	Compares a template against overlapped image regions.
Resize	Resizes an image.
Threshold	Applies a fixed-level threshold to each array element.
Sobel	Calculates the first, second, third, or mixed image derivatives using an extended Sobel operator.
filter2D	Convolves an image with the kernel. The function applies an arbitrary linear filter to an image. In-place operation is supported. When the aperture is partially outside the image, the function interpolates outlier pixel values according to the specified border mode.
morphologyEX	Performs advanced morphological transformations.
Erode	Erodes an image by using a specific structuring element.
Dilate	Dilates an image by using a specific structuring element.
GaussianBlur	Blurs an image using a Gaussian filter.
Blur	Blurs an image using the normalized box filter.
sqrBoxFilter	Calculates the normalized sum of squares of the pixel values overlapping the filter.
Remap	Applies a generic geometrical transformation to an image.
Laplacian	Calculates the Laplacian of an image.
Scharr	Calculates the first x- or y- image derivative using Scharr operator.
sepFilter2D	Applies a separable linear filter to an image.
calcHist	Calculates a histogram of a set of arrays.
accumulate	Adds an image to the accumulator image.
accumulateProduct	Adds the per-element product of two input images to the accumulator image.
accumulateWeighted	Updates a running average.
cornerMinEigenVal	Calculates the minimal eigenvalue of gradient matrices for corner detection.
cornerHarris	Harris corner detector.
preCornerDetect	Calculates a feature map for corner detection.
HoughLines	Finds lines in a binary image using the standard Hough transform.
HoughLinesP	Finds line segments in a binary image using the probabilistic Hough transform.
goodFeaturesToTrack	Determines strong corners on an image.

6.4.2 Conditions to use the accelerator

This section describes the conditions that an OpenCV function must meet to use OpenCL as an accelerator on the GPU.

6.4.2.1 pyrUP

Upsamples an image and then blurs it.

```
void cv::pyrUp (InputArray src, OutputArray dst, const Size & dstsize =
    Size(),int borderType = BORDER_DEFAULT)
```

Parameter	Requirement
src	Datatype: 8UC1/8UC3/8UC4/32FC1/32FC3/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Format: UMat

6.4.2.2 warpPerspective

Applies a perspective transformation to an image.

```
void cv::warpPerspective (InputArray _src, OutputArray _dst, InputArray _M0,
    Size dsize, int flags, int borderType, const Scalar& borderValue)
```

Parameter	Requirement
src	Datatype: 8UC1/8UC3/8UC4/32FC1/32FC3/32FC4 Sizelimit: 1080p (1920 x 1080) Format: UMat
dst	Output image that has the size dsize and the same type as SRC . Format: UMat
flags	INTER_NEAREST/INTER_LINEAR

6.4.2.3 warpAffine

Applies an affine transformation to an image.

```
void cv::warpAffine (InputArray _src, OutputArray _dst,
    InputArray _M0, Size dsize,
    int flags, int borderType, const Scalar& borderValue)
```

Parameter	Requirement
src	Datatype:8UC3/8UC4 Sizelimit:1080p(1920x1080) Format:UMat
dst	output image that has the size dsize and the same type as src . Format:UMat
flags	INTER_NEAREST/INTER_LINEAR/INTER_CUBIC

6.4.2.4 match Template

Compares a template against overlapped image regions.

```
void cv::matchTemplate (InputArray _img, InputArray _templ, OutputArray _result,
    int method, InputArray _mask)
```

Parameter	Requirement
_img	Datatype: 8UC1/8UC3/32FC1/32FC3 Sizelimit: 1280 x 1024 Format: UMat
_templ	Searched template. It must be not greater than the source image and have the same data type. Size: (11 x 11)/(41 x 41)
_result	If the image is W x H and the templ is w x h, the result is (W - w + 1) x (H - h + 1). Format: UMat
method	TM_SQDIFF, TM_SQDIFF_NORMED, TM_CCORR, TM_CCORR_NORMED, TM_CCOEFF, TM_CCOEFF_NORMED

6.4.2.5 resize

Resizes the image SRC down to or up to the specified size.

```
void cv::resize (InputArray _src, OutputArray _dst, Size dsize,
    double inv_scale_x, double inv_scale_y, int interpolation)
```

Parameter	Requirement
src	Datatype: 8UC4/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Output image. It has the size dsize (when it is non-zero) or the size computed from src.size(), fx, and fy. The type of dst is the same as of src. Format: UMat
dsize	Output image size. It is computed as: dsize = Size (round (fx * src.cols), round (fy * src.rows))
inv_scale_x	0.3/0.5/0.6/2.0
inv_scale_y	0.3/0.5/0.6/2.0
Interpolation	INTER_NEAREST/INTER_LINEAR/INTER_AREA

6.4.2.6 Threshold

Applies a fixed-level threshold to each array element.

```
cv::threshold (InputArray _src, OutputArray _dst, double thresh, double maxval,
    int type)
```

Parameter	Requirement
src	Datatype: 32FC4 Sizelimit: 1280 x 720 to 4k (3840 x 2160) Format: UMat
dst	Output array of the same size and type and the same number of channels as src. Format: UMat
type	THRESH_BINARY, THRESH_BINARY_INV, THRESH_TRUNC, THRESH_TOZERO_INV

6.4.2.7 Sobel

Calculates the first, second, third, or mixed image derivatives using an extended Sobel operator.

```
void cv::Sobel (InputArray src, OutputArray dst, int ddepth, int dx, int dy,
               int ksize, double scale, double delta, int borderType)
```

Parameter	Requirement
src	Datatype: 8UC4/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Output array of the same size and type and the same number of channels as src. Format: UMat

6.4.2.8 filter2D

Convolve an image with the kernel.

```
void cv::filter2D (InputArray src, OutputArray dst, int ddepth, InputArray
                  kernel,
                  Point anchor = Point(-1,-1), double delta = 0, int borderType = BORDER_DEFAULT)
```

Parameter	Requirement
src	Datatype: 8UC4/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Output array of the same size and type and the same number of channels as src. Format: UMat
kernel	Size: (3 x 3)/(5 x 5)

6.4.2.9 morphologyEX

Performs advanced morphological transformations.

```
void morphologyEx (InputArray _src, OutputArray _dst, int op,
                  InputArray _kernel, Point anchor, int iterations,
                  int borderType, const Scalar& borderValue)
```

Parameter	Requirement
src	Datatype: 32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Destination image of the same size and type as source image. Format: UMat
op	MORPH_OPEN, MORPH_CLOSE, MORPH_GRADIENT, MORPH_TOPHAT, MORPH_BLACKHAT
kernel	Size: (3 x 3)/(5 x 5)

6.4.2.10 erode

Erodes an image by using a specific structuring element.

```
void erode (InputArray src, OutputArray dst, InputArray kernel,
           Point anchor, int iterations,
           int borderType, const Scalar& borderValue)
```

Parameter	Requirement
src	Datatype: 32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Output image of the same size and type as src. Format: UMat
kernel	Size: (3 x 3)/(5 x 5)

6.4.2.11 dilate

Dilates an image by using a specific structuring element.

```
void dilate (InputArray src, OutputArray dst, InputArray kernel,
            Point anchor, int iterations,
            int borderType, const Scalar& borderValue)
```

Parameter	Requirement
src	Datatype: 32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Output image of the same size and type as src. Format: UMat
kernel	Size: (3 x 3)/(5 x 5)

6.4.2.12 GaussianBlur

Blurs an image using a Gaussian filter.

```
void GaussianBlur (InputArray src, OutputArray dst, Size ksize,
                  double sigmaX, double sigmaY = 0,
```

```
int borderType = BORDER_DEFAULT);
```

Parameter	Requirement
src	Datatype: 32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Output image of the same size and type as src. Format: UMat
ksize	Size: (3 x 3)/(5 x 5)/(7 x 7)

6.4.2.13 Blur

Blurs an image using the normalized box filter.

```
void blur (InputArray src, OutputArray dst,
           Size ksize, Point anchor, int borderType)
```

Parameter	Requirement
src	Datatype: 8UC1/8UC4/32FC1/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Output image of the same size and type as src. Format: UMat
ksize	Size: (3 x 3)/(5 x 5)

6.4.2.14 sqrBoxFilter

Calculates the normalized sum of squares of the pixel values overlapping the filter.

```
void sqrBoxFilter (InputArray _src, OutputArray _dst, int ddepth,
                  Size ksize, Point anchor,
                  bool normalize, int borderType)
```

Parameter	Requirement
src	Datatype: 8UC4/32FC1/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Output image of the same size and type as src. Format: UMat
ksize	Size: (3, 3)/(20, 3)/(3, 20)/(20, 20))

6.4.2.15 remap

Applies a generic geometrical transformation to an image.

```
void cv::remap (InputArray _src, OutputArray _dst,
                InputArray _map1, InputArray _map2,
```

```
int interpolation, int borderType, const Scalar& borderValue)
```

Parameter	Requirement
src	Datatype: 8UC3/8UC4/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Destination image. It has the same size as map1 and the same type as src . Format: UMat
interpolation	INTER_NEAREST, INTER_LINEAR

6.4.2.16 Laplacian

Calculates the Laplacian of an image.

```
void cv::Laplacian (InputArray _src, OutputArray _dst, int ddepth, int ksize,  
double scale, double delta, int borderType)
```

Parameter	Requirement
src	Datatype: 8UC4/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Destination image of the same size and the same number of channels as src. Format: UMat
ksize	Size:(3 x 3)/(5 x 5)

6.4.2.17 Scharr

Calculates the first x- or y- image derivative using Scharr operator.

```
void cv::Scharr (InputArray _src, OutputArray _dst, int ddepth, int dx, int dy,  
double scale, double delta, int borderType)
```

Parameter	Requirement
src	Datatype: 8UC4/32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Destination image of the same size and the same number of channels as src . Format: UMat

6.4.2.18 sepFilter2D

Applies a separable linear filter to an image.

```
void sepFilter2D (InputArray src, OutputArray dst, int ddepth,  
InputArray kernelX, InputArray kernelY,
```

```
Point anchor = Point(-1,-1),
double delta = 0, int borderType = BORDER_DEFAULT);
```

Parameter	Requirement
src	Datatype: 8UC1/8UC4/32FC1/32FC4 Sizelimit: 1080P (1920 x 1080) Format: UMat
dst	Destination image of the same size and the same number of channels as src. Format: UMat
kernel	(1, ksize, DATATYPE); ksize=3,5,7,9,11
kernelY	(1, ksize, DATATYPE); ksize=3,5,7,9,11

6.4.2.19 calcHist

Calculates a histogram of a set of arrays.

```
void cv::calcHist (InputArrayOfArrays images, const std::vector<int>& channels,
InputArray mask, OutputArray hist,
const std::vector<int>& histSize,
const std::vector<float>& ranges,
bool accumulate)
```

Parameter	Requirement
images	Datatype: 8UC1 Sizelimit: 1280 x 720 to 4k (3840 x 2160) Format: UMat
channels	std::vector<int> channels(1, 0) >
hist	hist(256, 1, CV_32FC1)
dst	Destination image of the same size and the same number of channels as src. Format: UMat
histSize	std::vector<int> histSize(1, 256);

6.4.2.20 accumulate

Adds an image to the accumulator image.

```
void cv::accumulate (InputArray _src, InputOutputArray _dst, InputArray _mask)
```

Parameter	Requirement
src	Datatype: 32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Accumulator image with the same number of channels as input image. Format: UMat

6.4.2.21 accumulateProduct

Adds the per-element product of two input images to the accumulator image.

```
void cv::accumulateProduct (InputArray _src1, InputArray _src2,  
                             InputOutputArray _dst, InputArray _mask)
```

Parameter	Requirement
src1	Datatype: 32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
src2	Datatype: 32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Accumulator image with the same number of channels as input images. Format: UMat

6.4.2.22 accumulateWeighted

Updates a running average.

```
void cv::accumulateWeighted (InputArray _src, InputOutputArray _dst,  
                              double alpha, InputArray _mask)
```

Parameter	Requirement
src	Datatype: 32FC4 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Accumulator image with the same number of channels as input image, 32-bit or 64-bit floating-point. Format: UMat
alpha	Weight of the input image. Value: 2.0

6.4.2.23 cornerMinEigenVal

Calculates the minimal eigenvalue of gradient matrices for corner detection.

```
void cv::cornerMinEigenVal (InputArray _src, OutputArray _dst, int blockSize,  
                             int ksize, int borderType)
```

Parameter	Requirement
src	Datatype: CV_8UC1, CV_32FC1 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Accumulator image with the same number of channels as input image, 32-bit or 64-bit floating-point. Format: UMat

Parameter	Requirement
blockSize	Value: 7

6.4.2.24 cornerHarris

Harris corner detector.

```
void cv::cornerHarris (InputArray _src, OutputArray _dst, int blockSize, int
    ksize, double k, int borderType)
```

Parameter	Requirement
src	Datatype: CV_8UC1, CV_32FC1 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Image to store the Harris detector responses. It has the same type and the same size as src. Format: UMat
blockSize	Value: 5
ksize	Value: 7

6.4.2.25 preCornerDetect

Calculates a feature map for corner detection.

```
void cv::preCornerDetect (InputArray _src, OutputArray _dst, int ksize, int
    borderType)
```

Parameter	Requirement
src	Datatype: CV_8UC1, CV_32FC1 Sizelimit: 4k (3840 x 2160) Format: UMat
dst	Image to store the Harris detector responses. It has the same type and the same size as src. Format: UMat
ksize	Value: 3

6.4.2.26 HoughLines

Finds lines in a binary image using the standard Hough transform.

```
void HoughLines (InputArray _image, OutputArray lines,
    double rho, double theta, int threshold,
    double srn, double stn, double min_theta, double max_theta)
```

Parameter	Requirement
image	Datatype: CV_8UC1 Sizelimit: 4k (3840 x 2160)

Parameter	Requirement
	Format: UMat
lines	Datatype: CV_32FC2 Format: UMat
rho	Value: 0.1, 1
theta	Value: CV_PI/180.0, 0.1

6.4.2.27 HoughLinesP

Finds line segments in a binary image using the probabilistic Hough transform.

```
void HoughLinesP (InputArray _image, OutputArray _lines,
                  double rho, double theta, int threshold,
                  double minLineLength, double maxGap)
```

Parameter	Requirement
image	Datatype: CV_32SC4 Format: UMat
lines	Datatype: CV_32SC4 Format: UMat
rho	Value: 0.1, 1
theta	Value: CV_PI/180.0, 0.1

6.4.2.28 goodFeaturesToTrack

Determines strong corners on an image.

```
void cv::goodFeaturesToTrack (InputArray image, OutputArray corners,
                              int maxCorners, double qualityLevel, double
                              minDistance,
                              InputArray mask, int blockSize, int gradientSize,
                              bool useHarrisDetector, double k)
```

Parameter	Requirement
image	Input 8-bit or floating-point 32-bit, single-channel image. Format: UMat
corners	Datatype: CV_32FC2 Format: UMat
minDistance	Value: 0.0, 3.0
harrisDetector	Value: True/False

6.5 Performance differences of OpenCV on Arm GPU and VSI GPU

Mali GPU has good performance in OpenCL program building/linking and is more suitable for OpenCV applications. It is recommended to use OpenCV on i.MX 95.

The way VSI GPU OpenCL program does building/linking, there is poor performance on some OpenCV applications.

7 OpenVX Introduction

7.1 Overview

OpenVX is a low-level programming framework domain to enable software developers to efficiently access computer vision hardware acceleration with both functional and performance portability. OpenVX has been designed to support modern hardware architectures, such as mobile and embedded SoCs as well as desktop systems. Many of these systems are parallel and heterogeneous: containing multiple processor types including multi-core CPUs, DSP subsystems, GPUs, dedicated vision computing fabrics as well as hardwired functionality. Additionally, vision system memory hierarchies can often be complex, distributed, and not fully coherent. OpenVX is designed to maximize functional and performance portability across these diverse hardware platforms, providing a computer vision framework that efficiently addresses current and future hardware architectures with minimal impact on applications.

OpenVX defines a C Application Programming Interface (API) for building, verifying, and coordinating graph execution, as well as for accessing memory objects. The graph abstraction enables OpenVX implementers to optimize the execution of the graph for the underlying acceleration architecture.

OpenVX also defines the vxu utility library, which exposes each OpenVX predefined function as a directly callable C function, without the need for first creating a graph. Applications built using the vxu library do not benefit from the optimizations enabled by graphs; however, the vxu library can be useful as the simplest way to use OpenVX and as first step in porting existing vision applications.

For more details of programming with OpenVX, see the following specification from Khronos Group,

OpenVX specification (<https://www.khronos.org/registry/vx>).

7.2 OpenVX extension implementation

VeriSilicon's VX Extensions for Vision Imaging provide additional functionality for Vision Image processing beyond the functions provided through the Khronos Group OpenVX API. These enhancements take advantage of the enhanced Vision capabilities available in VeriSilicon's Vision-capable hardware. VeriSilicon software provides a set of extensions which interface with OpenCL 1.2 and support higher level C language programming of VeriSilicon's custom EVIS (Enhanced Vision Instruction Set).

The VeriSilicon VX extension and enhancements includes three major components:

- An API level interface to the EVIS (Enhanced Vision Instruction Set)
- Extended C language features for Vision Processing
- Supported for a subset of Vision-compatible OpenCL built-in functions

7.2.1 Hardware requirements

Vision Imaging hardware capabilities are required to support full OpenVX. The following configurations are supported:

- GC7000XSVX (i.MX 8QuadMax)
- VIP8000NanoSI (i.MX 8M Plus)

7.2.2 EVIS instruction interface

Vivante's Vision Imaging capable IP have an Enhanced Vision Instruction Set (EVIS), which enhances the ability of the GPU or VIP (Vision Image Processor) to process complex vision operations. A single EVIS instruction can do a task which may require tens or even hundreds of normal ISA instructions to finish.

The following table shows the instructions supported as Intrinsic calls.

7.2.3 Extended language features

Vivante's OpenVX C programming Language corresponds closely to the OpenCL C programming language.

- Vivante's C language extensions for OpenVX C share many language facilities with OpenCL C 1.2. However, it can be considered a subset of OpenCL C 1.2, as it does not include OCL features which are useless for OpenVX and other Vision Imaging applications.
- Vivante's OpenVX C includes specific language facilities like Vision built-ins and data types specific for OpenVX.

Table 29. OPCODE EVIS instructions supported as intrinsic calls

EVIS OP_CODE	Description	Supported by Vivante VX
ABS_DIFF	Absolute difference between two values	Y
IADD	Adds two or three integer values	Y
IACC_SQ	Squares a value and adds it to an accumulator	Y
LERP	Linear interpolation between two values	Y
FILTER	Performs a filter on a 3x3 block	Y
MAG_PHASE	Computes magnitude and phase of 2 packed data values	Y
MUL_SHIFT	Multiplies two 8-or 16-bit integers and shifts	Y
DP16X1	1 Dot Product from 2 16 component values	Y
DP8X2	2 Dot Products from 2 8 component values	Y
DP4X4	4 Dot Products from 2 4 component values	Y
DP2X8	8 Dot Products from 2 2 component values	Y
CLAMP	Clamps up to 16 values to a max or min value	Y
BI_LINEAR	Computes a bilinear interpolation of 4 pixel values	Y
SELECT_ADD	Adds a pixel value or increments a counter inside bins	Y
ATOMIC_ADD	Adds a valid atomically to an address	Y
BIT_EXTRACT	Extracts up to 8 bitfields from a packed stream	Y
BIT_REPLACE	Replaces up to 8 bitfields from a packed stream	Y
DP32X1	1 Dot Product from 2 32 component values	Y
DP16X2	2 Dot Products from 2 16 component values	Y
DP8X4	4 Dot Products from 2 8 component values	Y
DP4X8	8 Dot Products from 2 4 component values	Y
DP2X16	16 Dot Products from 2 2 component values	Y

7.2.4 Packed types

Vivante's OpenCL compiler implements OpenCL C signed and unsigned char and short types in an unpacked format, such that a normal char4 occupies 128 bits (4 32-bit registers). This is undesirable for Vision applications, where packed data is the "natural" layout for almost all operations. To fully utilize the computing power of EVIS instructions, Vivante VX includes additional packed types, which can be identified by their **vxc_** prefix.

```
/* packed char2/4/8/16 */
```

```

typedef _viv_char2_packed vxc_char2;
typedef _viv_char4_packed vxc_char4;
typedef _viv_char8_packed vxc_char8;
typedef _viv_char16_packed vxc_char16;
/* packed uchar2/4/8/16 */
typedef _viv_uchar2_packed vxc_uchar2;
typedef _viv_uchar4_packed vxc_uchar4;
typedef _viv_uchar8_packed vxc_uchar8;
typedef _viv_uchar16_packed vxc_uchar16;
/* packed short2/4/8 */
typedef _viv_short2_packed vxc_short2;
typedef _viv_short4_packed vxc_short4;
typedef _viv_short8_packed vxc_short8;
/* packed ushort2/4/8 */
typedef _viv_ushort2_packed vxc_ushort2;
typedef _viv_ushort4_packed vxc_ushort4;
typedef _viv_ushort8_packed vxc_ushort8;

```

7.2.5 Initializing constants on load

Constant data in OpenCL requires compile-time initialization. There is also a need to initialize the data when the kernel is loaded/run, so that the application can control the behavior of a program by changing its constants at load-time. The VeriSilicon VX extended keyword **`_viv_uniform`** can be used to define load-time initialization constant data,

```
_viv_uniform vxc_512bits u512;
```

An application using VeriSilicon VX needs to set the proper values for `_viv_uniform` before the kernel program is run.

7.2.6 Inline assembly

A packed type cannot be used as an unpacked type in expressions or built-in functions. The programmer needs to convert packed type data to unpacked type data in order to perform these operations. The conversion negatively impacts performance in terms of both instruction count and register usage, so it is desirable to perform operations directly on packed data whenever possible. The Vivante Vision compiler accepts inline assembly for a wide range of operations to speed up packed data calculations.

For example, to add two packed char16 data, the programmer can use following inline assembly:

```

vxc_uchar16 a, b, c;
vxc_short8 b;
_viv_uniform vxc_512bits u512;
...
_viv_asm(ADD, c, a, b); /* c = a + b; */
where the syntax of inline assembly is:
_viv_asm(
  OP_CODE,
  dest,
  source0,
  source1
);

```

The following table lists the standard shader instructions that operate on packed data and are supported through inline assembly, keyword **`_viv_asm`**.

Table 30. OPCODES IR instructions supported by inline assembly

IR OP_CODE Instruction	Description	Supported by Vivante VX
ABS	Absolute value	Y
ADD	Add	Y
ADD_SAT	Integer add with saturation	Y
AND_BITWISE	Bitwise AND	Y
BIT_REVERSAL	Integer bit-wise reversal	ES31
BITEXTRACT	Extract Bits from src to dest	ES31
BITINSERT	Bit replacement	ES31
BITSEL	Bitwise Select	Y
BYTE_REVERSAL	Integer byte-wise reversal	ES31
CLAMP0MAX	clamp0max dest, value, max	Y
CMP	Compare each component	Y
CONV	Convert	Y
DIV	Divide	Y
FINDLSB	Find least significant bit	ES31
FINDMSB	Find most significant bit	ES31
LEADZERO	Detect Leading Zero	Y
LSHIFT	Left Shifter	Y
MADSAT	Integer multiple and add with saturation	Y
MOD	Modulus	Y
MOV	Move	Y
MUL	Multiply	Y
MULHI	Integer only	Y
MULSAT	Integer multiply with saturation	Y
NEG	neg(a) is similar to (0 - (a))	Y
NOT_BITWISE	Bitwise NOT	Y
OR_BITWISE	Bitwise OR	Y
POPCOUNT	Population Count	ES31/OCL1.2
ROTATE	Rotate	Y
RSHIFT	Right Shifter	Y
SUB	Substract	Y
SUBSAT	Integer subtraction with saturation	Y
XOR_BITWISE	Bitwise XOR	Y

Note: *ES31 = Supported by VivanteVX, but may not be needed for Vision processing

7.3 OpenCL functions compatible with Vivante vision

Vivante's VX extensions for Vision Image processing support most of the OpenCL 1.2 built-in functions for normal OCL data types. Packed types are not supported in these built-in functions.

For image read/write functions, only sample-less 1D/1D array/2D image read/write functions are supported.

7.3.1 Read_Imagef,i,ui

```
/* OCL image builtins can be used in VX kernel */
float4 read_imagef (image2d_t image, int2 coord);
int4 read_imagei (image2d_t image, int2 coord);
uint4 read_imageui (image2d_t image, int2 coord);
float4 read_imagef (image1d_t image, int coord);
int4 read_imagei (image1d_t image, int coord);
uint4 read_imageui (image1d_t image, int coord);
float4 read_imagef (image1d_array_t image, int2 coord);
int4 read_imagei (image1d_array_t image, int2 coord);
uint4 read_imageui (image1d_array_t image, int2 coord);
```

7.3.2 Write_Imagef,i,ui

```
void write_imagef (image2d_t image, int2 coord, float4 color);
void write_imagei (image2d_t image, int2 coord, int4 color);
void write_imageui (image2d_t image, int2 coord, uint4 color);
void write_imagef (image1d_t image, int coord, float4 color);
void write_imagei (image1d_t image, int coord, int4 color);
void write_imageui (image1d_t image, int coord, uint4 color);
void write_imagef (image1d_array_t image, int2 coord, float4 color);
void write_imagei (image1d_array_t image, int2 coord, int4 color);
void write_imageui (image1d_array_t image, int2 coord, uint4 color);
```

7.3.3 Query Image Dimensions

```
int2 get_image_dim (image2d_t image);
size_t get_image_array_size(image1d_array_t image);
/* Built-in Image Query Functions */
int get_image_width (image1d_t image);
int get_image_width (image2d_t image);
int get_image_width (image1d_array_t image);
int get_image_height (image2d_t image);
```

7.3.4 Channel Data Types Supported

```
/* Return the channel data type. Valid values are:
* CLK_SNORM_INT8
* CLK_SNORM_INT16
* CLK_UNORM_INT8
* CLK_UNORM_INT16
* CLK_UNORM_SHORT_565
* CLK_UNORM_SHORT_555
* CLK_UNORM_SHORT_101010
* CLK_SIGNED_INT8
* CLK_SIGNED_INT16
```

```
* CLK_SIGNED_INT32
* CLK_UNSIGNED_INT8
* CLK_UNSIGNED_INT16
* CLK_UNSIGNED_INT32
* CLK_HALF_FLOAT
* CLK_FLOAT
*/
int get_image_channel_data_type (image1d_t image);
int get_image_channel_data_type (image2d_t image);
int get_image_channel_data_type (image1d_array_t image);
```

7.3.5 Image Channel Orders Supported

```
/* Return the image channel order. Valid values are:
* CLK_A
* CLK_R
* CLK_Rx
* CLK_RG
* CLK_RGx
* CLK_RA
* CLK_RGB
* CLK_RGBx
* CLK_RGBA
* CLK_ARGB
* CLK_BGRA
* CLK_INTENSITY
* CLK_LUMINANCE
*/
int get_image_channel_order (image1d_t image);
int get_image_channel_order (image2d_t image);
int get_image_channel_order (image1d_array_t image);
```

8 Vulkan

8.1 Overview

Vulkan is a new generation graphics and compute API that provides high-efficiency, cross-platform access to modern GPUs used in a wide variety of devices from PCs and consoles to mobile phones and embedded platforms.

Vulkan defines as an API (Application Programming Interface) for graphics and compute hardware. The API consists of many commands that allow a programmer to specify shader programs, compute kernels, objects, and operations involved in producing high-quality graphical images, specifically color images of three-dimensional objects.

To the programmer, Vulkan is a set of commands that allow the specification of shader programs or shaders, kernels, data used by kernels or shaders, and state controlling aspects of Vulkan outside the scope of shaders. Typically, the data represents geometry in two or three dimensions and texture images, while the shaders and kernels control the processing of the data, rasterization of the geometry, and the lighting and shading of fragments generated by rasterization, resulting in the rendering of geometry into the framebuffer.

A typical Vulkan program begins with platform-specific calls to open a window or otherwise prepare a display device onto which the program will draw. Then, calls are made to open queues to which command buffers are submitted. The command buffers contain lists of commands which will be executed by the underlying hardware. The application can also allocate device memory, associate resources with memory and refer to these resources from within command buffers. Drawing commands cause application-defined shader programs

to be invoked, which can then consume the data in the resources and use them to produce graphical images. To display the resulting images, further platform-specific commands are made to transfer the resulting image to a display device or window.

For more details of programming with Vulkan, refer to the following specification from Khronos Group.

<https://www.khronos.org/registry/vulkan/>

8.2 Vulkan Validation Layers

Vulkan is an explicit API, enabling direct control over how GPUs actually work. By design, minimal error checking is done inside a Vulkan driver. Applications have full control and responsibility for correct operation. Any errors in how Vulkan is used can result in a crash. Vulkan validation layers that can be enabled to assist development by enabling developers to verify their applications correct use of the Vulkan API.

8.3 Window System Integration

Vulkan relies on a new mechanism to interact with the native Windowing System and present the rendered results to the user. This mechanism is called the Window System Integration and is provided via extensions outside of the core API.

In the i.MX BSPs where Vulkan is enabled, the default window manager is Weston, a Wayland compositor reference implementation.

When compiling a Vulkan application for Wayland make sure to define the `VK_USE_PLATFORM_WAYLAND_KHR` symbol, so all the proper includes and code paths are enabled.

GLFW and SDL can manage the surface creation and proper extension initializations, but when an application is newly developed without using any frameworks, require to enable the following instance extensions:

```
VK_KHR_SURFACE_EXTENSION_NAME
```

```
VK_KHR_WAYLAND_SURFACE_EXTENSION_NAME
```

Once there is a display connection to the Wayland server and a surface created, then start to use the `wl_display` and `wl_surface` pointers to populate the info structure required by `vkCreateWaylandSurfaceKHR`.

A word of advice, when querying the Physical Device Surface capabilities with `vkGetPhysicalDeviceSurfaceCapabilitiesKHR` before having created the Swapchain, the current extent width and height will return a value of `0xFFFFFFFF`, make sure to add checks for this in the code, when this happens, set the swapchain extent to the actual size of the surface want to render to, or a fallback extent size.

9 Vivante Multiple GPUs and Virtualization

9.1 Overview

Vivante multi-GPU implementations provide a variety of capabilities which can be managed through hardware and software controls. This chapter intends to summarize the software controls used for Vivante multi-GPU IP implementations.

Multi-GPU feature can be enabled with dual GC7000XSVX on i.MX 8QuadMax and the derived devices.

9.2 Multi-GPU configurations

Vivante Multi-GPU IP may be configured into one of the following behavior model through software:

- Combined Mode where two (or more) GPU cores in the multi-GPU design behave in concert. Driver presents multi-GPU to SW application as a single logical GPU. The multiple GPUs work in the same virtual address

space and share the same MMU page table. The multiple GPUs fetch and execute a shared Command Buffer.

- Independent Mode where each GPU in the multi-GPU design performs independently. The multiple GPUs work in different virtual address spaces but share the same MMU page table. Each GPU core fetches and executes its own Command Buffer. This enables different SW applications to run simultaneously on different GPU cores.
- OpenCL API allows application to handle the multi-GPU Independent Mode directly, as each GPU core in a multi-GPU design represents an independent OpenCL Compute Device.

9.3 GPU affinity configuration

In the multi-GPU Independent Mode, application can specify to run on a specific GPU among the multiple GPUs through an environment variable `VIV_MGPU_AFFINITY`. Once an application's GPU affinity is specified, the application will only run on the specified GPU and will not migrate to other GPUs even if those GPUs are idle.

`VIV_MGPU_AFFINITY` is the environment variable to control the application GPU affinity on multi-GPU platform. The client drivers will assume they are using a standalone GPU through a `gcoHARDWARE` object no matter how this variable is set. The possible values for the environment variable `VIV_MGPU_AFFINITY` include:

- Not defined or
- Defined as "0" `gcoHARDWARE` objects work in `gcvMULTI_GPU_COMBINED` mode (default)
 - "1:0" `gcoHARDWARE` objects work in `gcvMULTI_GPU_INDEPENDENT` mode and GPU0 is used
 - "1:1" `gcoHARDWARE` objects work in `gcvMULTI_GPU_INDEPENDENT` mode and GPU1 is used

On a single GPU device, setting `VIV_MGPU_AFFINITY` to 0 or 1 does not make any difference as all application processes/threads are bound to GPU0. But the application will fail the GPU context initialization if `VIV_MGPU_AFFINITY` is set to "1:1" (driver reports error).

9.4 OpenCL on multi-GPU device

OpenCL driver works in bridged mode as single logical compute device. In this configuration, multiple GPUs in the device operate as individual OpenCL Compute Devices. The OpenCL application is responsible to assign and dispatch the compute tasks to each GPU (Compute Device).

The following OpenCL APIs return the list of compute devices available on a platform, and the device information.

```
cl_int clGetDeviceIDs (cl_platform_id platform, cl_device_type device_type,
cl_uint num_entries,
cl_device_id *devices, cl_uint *num_devices)
cl_int clGetDeviceInfo (cl_device_id device, cl_device_info param_name, size_t
param_value_size,
void *param_value, size_t *param_value_size_ret)
```

9.5 GPU virtualization configuration

Multi-GPU also can be used on different OS systems as independent mode separately, this can be configured by overriding the `irq availability n DTS` entry for different OS implementation, in `arch/arm64/boot/dts/freescale/fsl-imx8qmxxx.dts`.

Guest OS 1 (GPU0 only)

```
&gpu_3d1 {
    status = "disable";
};
```

Guest OS 2 (GPU1 only)

```
&gpu_3d0 {  
    status = "disable";  
};
```

10 GBM - Generic Buffer Management

The GBM (Graphic Buffer Management) API is a thin layer over DRM KMS (Linux Direct Rendering Manager - Kernel ModeSetting API) that provides a mechanism for allocating buffers for graphics rendering. GBM is intended to be used as a native platform for EGL on DRM. The handle it creates can be used to initialize EGL and to create render target buffers. This can be resumed as a modern OpenGL ES FBDEV, because it permits full usage of the DRM KMS API with OpenGL ES acceleration.

Starting from i.MX 8, the DRM is supported and recommended to use GBM. GBM provides options of allocating modifier-abiding surfaces too, for Wayland compositors and the X11 server to render to.

10.1 Introduction to DRM Format Modifiers

A DRM format modifier is a 64-bit, vendor-prefixed, semi-opaque unsigned integer. Most modifiers represent a concrete, vendor-specific tiling format for images. Some exceptions are `DRM_FORMAT_MOD_LINEAR` (which is not vendor-specific); `DRM_FORMAT_MOD_NONE` (which is an alias of `DRM_FORMAT_MOD_LINEAR` due to historical accident); and `DRM_FORMAT_MOD_INVALID` (which does not represent a tiling format). The modifier's vendor prefix consists of the 8 most significant bits. The canonical list of modifiers and vendor prefixes is found in `drm_fourcc.h` in the Linux kernel source.

One goal of modifiers in the Linux ecosystem is to enumerate for each vendor a reasonably sized set of tiling formats that are appropriate for images shared across processes, APIs, and/or devices, where each participating component may possibly be from different vendors. A non-goal is to enumerate all tiling formats supported by all vendors. Some tiling formats used internally by vendors are inappropriate for sharing; no modifiers should be assigned to such tiling formats.

Modifier values typically do not describe memory layouts. More precisely, a modifier's lower 56 bits usually have no structure. Instead, modifiers name memory layouts; they name a small set of vendor-preferred layouts for image sharing. As a consequence, in each vendor namespace the modifier values are often sequentially allocated starting at 1.

Each modifier is usually supported by a single vendor and its name matches the pattern `{VENDOR}_FORMAT_MOD_*` or `DRM_FORMAT_MOD_{VENDOR}_*`. Examples are `DRM_FORMAT_MOD_VIVANTE_TILED` and `DRM_FORMAT_MOD_BROADCOM_VC4_T_TILED`. An exception is `DRM_FORMAT_MOD_LINEAR`, which is supported by most vendors.

Many APIs in Linux use modifiers to negotiate and specify the memory layout of shared images. For example, a Wayland compositor and Wayland client may, by relaying modifiers over the Wayland protocol `zwp_linux_dmabuf_v1`, negotiate a vendor-specific tiling format for a shared `wl_buffer`. The client may allocate the underlying memory for the `wl_buffer` with GBM, providing the chosen modifier to `gbm_bo_create_with_modifiers`. The client may then import the `wl_buffer` into Vulkan for producing image content, providing the resource's `dma_buf` to `VkImportMemoryFdInfoKHR` and its modifier to `VkImageDrmFormatModifierExplicitCreateInfoEXT`. The compositor may then import the `wl_buffer` into OpenGL for sampling, providing the resource's `dma_buf` and modifier to `eglCreateImage`. The compositor may also bypass OpenGL and submit the `wl_buffer` directly to the kernel's display API, providing the `dma_buf` and modifier through `drm_mode_fb_cmd2`.

11 Wayland and Weston

11.1 Overview

Wayland is a protocol for a compositor to talk to its clients as well as a C library implementation of that protocol. Wayland is intended as a simpler replacement for X, easier to develop and maintain. The compositor can be a standalone display server running on Linux kernel mode setting and evdev input devices, an X application, or a Wayland client itself. The clients can be traditional applications, X servers (rootless or full screen) or other display servers.

11.2 Wayland EGL

Wayland-EGL is the client side implementation of the Wayland that binds the EGL stack and buffer sharing mechanism to the generic Wayland API. Frontend of the wayland-egl is now part of the wayland and i.MX graphics driver supports the implementation of buffer sharing mechanism.

11.3 Weston compositor

Weston is reference implementation of a Wayland compositor. The Weston compositor is minimal and lightweight and is suitable for many embedded and mobile use cases. Weston support multiple renderers and backends which need to be chosen appropriately based on the processor configurations. This is usually preset in the i.MX image.

11.3.1 Weston backends

Weston have implementation to support different display APIs, which is called backend. i.MX 8 and i.MX 9 support KMS/DRM hence use DRM backend while the i.MX 6/7 uses FBDEV backend. i.MX graphics continues to support graphics acceleration with FBDEV backends.

11.3.1.1 RDP backend

RDP backend supports acceleration. Now the feature is available as a part of the i.MX release image where GStreamer is supported. The `librdp` library requires GStreamer. Perform the following steps to use RDP on the i.MX on the target device:

1. In the `/etc/xdg/weston/weston.ini` file, uncomment `start-on-startup=true`.
2. Generate RDP certificates.

```
mkdir -p /etc/freerdp/keys/  
winpr-makecert -rdp -path /etc/freerdp/keys
```

3. Rename the generated files to `server.crt` and `server.key`.

On the Windows PC, use the **Remote Desktop Connection** application and enter the target IP address.

Similarly, Linux and its tool also can be used.

11.3.2 Weston renderer

11.3.2.1 GL renderer

GL (GLES) renderer implementation is the default with Weston implementation. GL renderer takes the buffer passed from clone and maps as a texture. After the initial setup, the client only needs to tell the compositor which buffer to use and when and where it has rendered new content into it.

11.3.2.2 G2D renderer

G2D is the 2D API. See [Section 2](#) for full details of G2D APIs. G2D renderer provides mechanism to accelerate Weston with 2D GPU. The 2D Graphics Engine reduces the burden on the 3D GPU, saves power, and integrates well with the video capabilities of the SoC. G2D compositor can increase system bandwidth utilization, so the performance is better than the GL compositor in the complex usecase environment.

To enable the G2D compositor, open the file `/etc/xdg/weston/weston.ini` in the Linux image.

```
use-g2d=1
```

Note: When running benchmarks, set `WESTON_FORCE_RENDERER=1` to Weston for Mali GPU.

1. Add `WESTON_FORCE_RENDERER=1` in `/etc/environment`.
2. `systemctl` restarts Weston.

11.3.3 Weston shells

Weston supports multiple shells, each of these shells have its own public protocol interface for clients. This means that a client must be specifically written for a shell protocol. Otherwise, it will not work. Below are the currently supported shell.

Note: Weston 10 marked `wl_shell` as deprecated and has been removed by community since Weston 11, recommending to covert to `xdg-shell` for Wayland application developing.

11.3.3.1 Desktop shell

Desktop shell is like a typical X desktop environment, concentrating on traditional keyboard and mouse user interfaces and the familiar desktop-like window management. Desktop shell consists of the shell plugin `desktop-shell.so` and the special client `weston-desktop-shell` which provides the wallpaper, panel, and screen locking dialog.

11.3.3.2 Fullscreen shell

Fullscreen shell is intended for a client that needs to take over whole outputs, often all outputs. This is primarily intended for running another compositor on Weston. The other compositor does not need to handle any platform-specifics like DRM/KMS or `evdev/libinput`. The shell consists only of the shell plugin `fullscreen-shell.so`.

11.3.3.3 IVI-shell

In-vehicle infotainment shell is a special purpose shell that exposes a GENIVI Layer Manager compatible API to controller modules, and a very simple shell protocol towards clients. IVI-shell starts with loading `ivi-shell.so`, and then a controller module which may launch helper clients. This shell provides option of setting windowing position, which need to be programmed from the client application.

12 X Windowing Acceleration

X11 is accelerated on i.MX 8 through Xwayland. Support on i.MX 6 deprecated.

13 Advanced GPU Configuration

13.1 GPU Scaling Governor

i.MX 8QuadMax GPU design supports different running modes: overdrive, nominal, and underdrive. Nominal is the default, the overdrive is supposed to be performance/benchmark mode, and underdrive mode is expected as energy saving mode.

Switch among the 3 modes using command line without needing to recompile the GPU driver.

```
$ echo "overdrive" > /sys/bus/platform/drivers/galcore/gpu_govern
$ echo "nominal" > /sys/bus/platform/drivers/galcore/gpu_govern
$ echo "underdrive" > /sys/bus/platform/drivers/galcore/gpu_govern
```

To check the mode that is currently running, use the command line as follows:

```
$ cat /sys/bus/platform/drivers/galcore/gpu_govern
```

13.2 GPU Device Cooling

i.MX 6/7/8 devices support the thermal driver, which could signal the overheat event to the GPU driver. When the GPU driver receives the event, it can enable the GPU DFS feature to reduce the GPU frequency as N/64 of the original designated clock.

The default N factor is 1 in the original BSP release. The end-user can reconfigure it through the following command:

```
echo N >/sys/bus/platform/drivers/galcore/gpu3DMinClock
```

The user also can check the existing configuration as follows:

```
cat /sys/bus/platform/drivers/galcore/gpu3DMinClock
```

13.3 i.MX 95 GPU frequency scaling

The i.MX 95 and later platforms support frequency scaling based on the Linux devfreq framework. The clock rate varies among [500 MHz, 800 MHz, 1 GHz]. Currently, only `simple_ondemand` governor is supported. The frequency can be set to a certain clock by changing `max_freq` and `min_freq`. The following command can be used to make the GPU run at the highest clock rate 1 GHz.

```
$ echo 1000000000 > /sys/devices/platform/soc/4d900000.gpu/devfreq/4d900000.gpu/
min_freq
```

Or set the GPU to performance mode as follows:

```
$ echo performance > /sys/class/devfreq/4d900000.gpu/governor
```

Roll back the default governor `simple_ondemand` with the following command:

```
$ echo simple_ondemand > /sys/class/devfreq/4d900000.gpu/governor
```

13.3.1 simple_ondemand governor

If the percentage of the `busy_time/total_time` exceeds 90% (`upthreshold`), the frequency jumps to the maximum frequency. If the percentage fluctuates within 5% (`downthreshold`) each time, the frequency remains unchanged. Otherwise, set the desired frequency based on the percentage. More details can be found in `drivers/devfreq/governor_simpleondemand.c`.

Sometimes, the percent 90 cannot be achieved when running some applications. Consider the producer-consumer mode in multi-thread programming. Or the `upthreshold` can be replaced with a small value through DTS property `up threshold` and `down differential`.

Some AI machine learning use cases may require the GPU to handle the compute jobs as soon as possible. The `simple_ondemand` governor does not meet such demand, but the performance governor can meet such requirement.

14 Vivante IDE

14.1 VivanteIDE overview

The VivanteIDE provides a single interface to a set of applications designed to be used by graphics, compute, vision and neural network application developers to rapidly develop and port applications either stand alone or as part of an IDE. Vivante IDE is built on the top of Eclipse, CDT

VivanteIDE capabilities include the following key features.

- **Project Management**
The Project Manager supports individual compile options for each file. In addition, workspace options define project dependencies, removing the need for manual management of file builds.
- **Source code smart editing and analysis**
The VivanteIDE Editor provides timesaving editing features such as type ahead for structures, word completion and automatic code indentation for a readable, formatted code view.
- **Automatic code generation**
Kernel development wizard can automatically generate the kernel code basing on simple inputs.
- **Performance and bandwidth profiling**
The Profile tabbed window provides profiler information. Every time the profiler is suspected accumulated statistical information is updated. For OGL applications the VPD Analyzer is provided.
- **Post-mortem performance analysis**
VPD Analyzer visualized the hardware data recorded at GPU application runtime.
- **Texture browse and conversion**
Texture browser and converter support texture file preview and format conversion.
- **Command line tools for OGL, OCL and OVX** compile.
- **Command line tools for Vulkan** application development.
- **Command line tools for Texture** compression/decompression and tile/de-tiling.

14.1.1 VivanteIDE component overview

VivanteIDE provides both command line tools and GUI “Perspective” views for performing different activities. Some functionality is available through both GUI and command line, while tools such as vCompiler and vcCompiler are available only using command line syntax.

Table 31. VivanteIDE tool overview

Perspective/Tool	Key Functionality	GUI	Command Line
Debug	Debug projects	Yes	

Table 31. VivantelIDE tool overview...continued

Perspective/Tool	Key Functionality	GUI	Command Line
Profile	Configure projects	Yes	
vCompiler	Offline OGL compiler	No	Yes, vCompiler
vcCompiler	Offline OCL compiler	No	Yes, vcCompiler
VPD Analyzer	Performance analysis	Yes	No
vTexture, vTextureTools	Texture manipulations and viewing; Compress, Decompress, Tile, De-Tile	Yes Texture Viewer Texture Browser	Yes vTextureTools
SPIR-V Disassembly	Debug Vulkan apps	Yes	No
Shader Assistant	Shader programming	Yes	No

14.2 VivantelIDE Requirements

14.2.1 Operating system compatibility

VivantelIDE is available for both Linux and Windows environments. VivantelIDE has been verified to work in Windows 7, Windows 10, Ubuntu 18.04, and Ubuntu 16.04. It might work in other Windows or Linux systems but has not been verified for alternate environments.

Table 32. Operating System Tool Compatibility Summary

Components	Linux	Windows
VivantelIDE	GUI and command	GUI and command
Tools		
vCompiler, vcCompiler	command	command
vProfiler	Built part of i.MX unified driver (target)	Built part of i.MX unified in driver(target)
VPD Analyzer	GUI	GUI
Shader Assistant	GUI	GUI
Texture Viewer	GUI	GUI
Texture Browser	GUI	GUI
vTextureTools	GUI and command	GUI and command

14.2.2 Hardware requirements

VivantelIDE can be used in either a simulation environment or on i.MX processors supporting OpenGL ES, OpenCL, OpenVX, and Neural Networks capabilities in the tools assume compatible hardware capability in the running environment, which must be correctly profiled in the tool for accurate results.

14.2.3 VivantelIDE license

i.MX supported VivantelIDE release package contains with preloaded license and restricted only to use with NXP processors. For more information, read NXP EULA.

14.3 VivanteIDE installation

14.3.1 VivanteIDE package

Each release of VivanteIDE will be compatible with its companion driver version. Forward and backward compatibility is not tested and use of VivanteIDE with any different driver version other than its companion version is NOT RECOMMENDED.

The package is delivered as a compressed file from nxp.com as
`Verisilicon_Tool__IDE_<version>.tgz`.

Table 33. VivanteIDE package contents

Top level Directory and exe file	Description
VivanteIDE-<version>-Linux-x86_64-**-Install	Installation wizard for Linux 64-bit.
VivanteIDE-<version>-Windows-**-Setup.exe	Installation wizard for Windows 64-bit/32-bit
README	README with basic installation notes

After installation the following directories will be created in the installation directory

Table 34. VivanteIDE tools directory

Files and Directories	Description
ide/	Directory containing IDE executables and plugins
examples/	Directory containing examples (just for Windows)
cmdtools/	Directory containing Vivante command line tools: vCompiler, vcCompiler, vTextureTools
doc/	Directory containing documents
license/	Directory containing license tools and license files
jre/	Directory containing JRE binaries
mingw32/	Directory containing MinGW (just for Windows)
uninstall.exe	Uninstaller of VivanteIDE

14.3.2 Installation

Install the package to run both the GUI and command line tools. You must install the package even if you are only going to use the command line tools.

14.3.2.1 Linux GUI

Run `Vivante-<version>-Linux-x86_64-**-Install` to launch the installation wizard. Follow the installation steps guided by the installation wizard to finish the installation.

14.3.2.2 Windows GUI

Run `Vivante-<version>-Windows-**-Setup.exe` to launch the installation wizard. Follow the installation steps guided by the installation wizard to finish the installation.

14.3.2.3 Installation from command line

The VivantelIDE installer can also be launched from the command line. Options can be specified as follows:

```
installer [option1] [option2] [option3]
```

Example Usage for Windows:

```
installer /mode silent /prefix destination_location /license license_file_path
```

Example Usage for Linux:

```
installer --mode silent --prefix destination_location --license  
license_file_path
```

Table 35. Command line installer options

Option for Windows	Option for Linux	Description
/mode silent	--mode silent	Silent mode (without GUI, without prompting)
/license license_file_path	--license license_file_path	Specify a license file to be installed
/prefix destination_location	--prefix destination_location	Specify the folder where VivantelIDE will be installed

14.3.3 VivantelIDE launch

14.3.3.1 Linux launch of GUI tool

To launch the GUI tool,

- Double-click the desktop shortcut **VivantelIDE<version>**.
- Run `installation_dir/ide/vivanteide<version>` in a BASH.

14.3.3.2 Windows launch of GUI tool

To launch the GUI tool:

- Click **Start Menu->VeriSilicon->VivantelIDE <version>->VivantelIDE <version>**.
- Double-click the desktop shortcut **VivantelIDE <version>**.
- Run `installation_dir/ide/vivanteide<version>.bat`.

14.3.3.3 Command line tool launch

To launch the command line tools, use the following paths. For Linux OS, launch in a BASH.

Run `installation_dir/cmdtools/vCompiler`, `vcCompiler`, `vTextureTools`.

14.3.3.4 Basic launch path summary

Table 36. Basic launch instruction summary

Tool	Linux Basic Launch Instruction	Windows Basic Launch Instruction
VivantelIDE GUI	Run <code>installation_dir/ide/vivanteide<version></code> in a BASH.	Run <code>installation_dir/ide/vivanteide<version>.bat</code>

Table 36. Basic launch instruction summary...continued

Tool	Linux Basic Launch Instruction	Windows Basic Launch Instruction
vcCompiler	Run installation_dir/cmdtools/bin/vcCompiler in a BASH.	Run installation_dir/cmdtools/bin/vcCompiler.exe
vCompiler	Run installation_dir/cmdtools/bin/vcompiler in a BASH.	Run installation_dir/cmdtools/bin/vCompiler.exe
vTextureTools	Run installation_dir/cmdtools/bin/vtexturetools in a BASH.	Run installation_dir/cmdtools/bin/vTextureTools.exe

14.4 VivantelIDE GUI

The desktop development environment for VivantelIDE is referred to as the Workbench. The Workbench contains panes that may change depending on the current activity. Some key panes are indicated in the figure below.

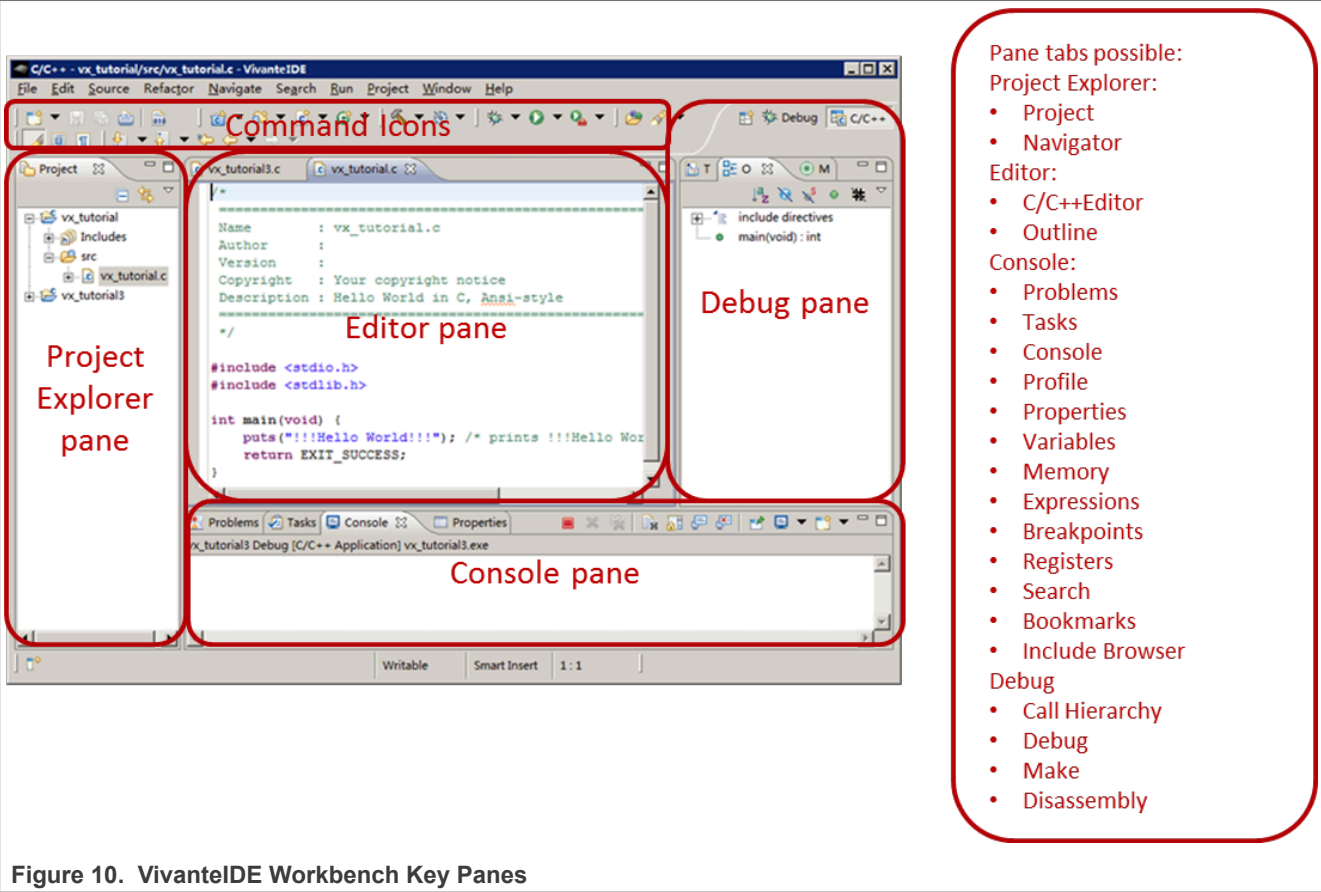


Figure 10. VivantelIDE Workbench Key Panes

The following examples provide users with basic simple steps to get started using VivantelIDE. The GUI is similar but not identical for each tool GUI: VPD Analyzer, Shader Assistant, Texture Browser, Texture Viewer.

14.4.1 Selecting a workspace

When VivantelIDE is opened, the **Workspace Launcher - Select a workspace** dialog box pops up by default. Click the **OK** button.

If the workspace is a new empty workspace, the **Welcome** dialog box is displayed.

If the workspace is not a new empty workspace, the workbench is displayed.

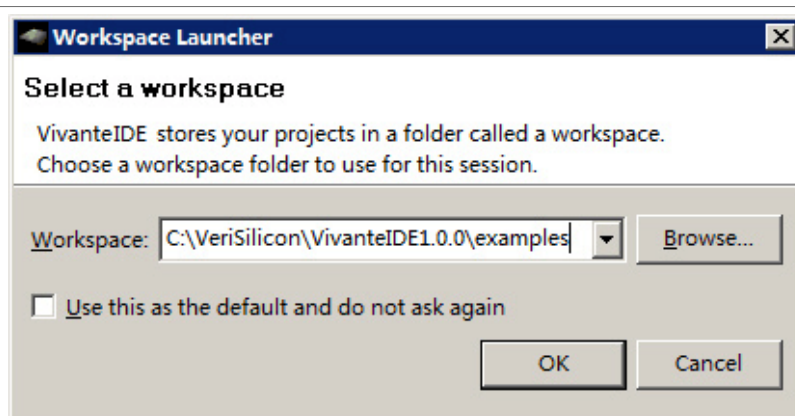


Figure 11. Figure 21. Workspace Launcher

14.4.2 Switching perspective

Click the pull-down menu items or click directly on the visible perspective name to switch perspective views.

Switch to the C/C++ perspective to manage projects and write source code. VivanteIDE will switch to the Debug perspective by default after a program is launched successfully in Debug mode.

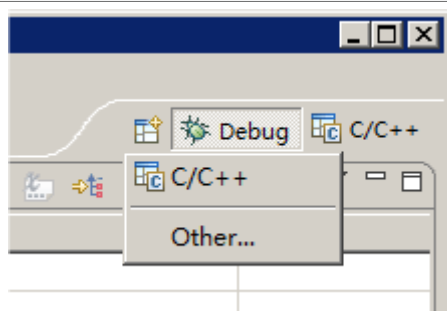


Figure 12. Switching perspective

14.4.3 Creating a new project

This section describes how to create an OpenVX project as an example.

New project creation is available from the main menu. Choose **File-->New-->Project...**

In the **New Project - Select a wizard** dialog box, open the **C/C++** folder in the **Wizards** list box and select **OpenVX C Project**.

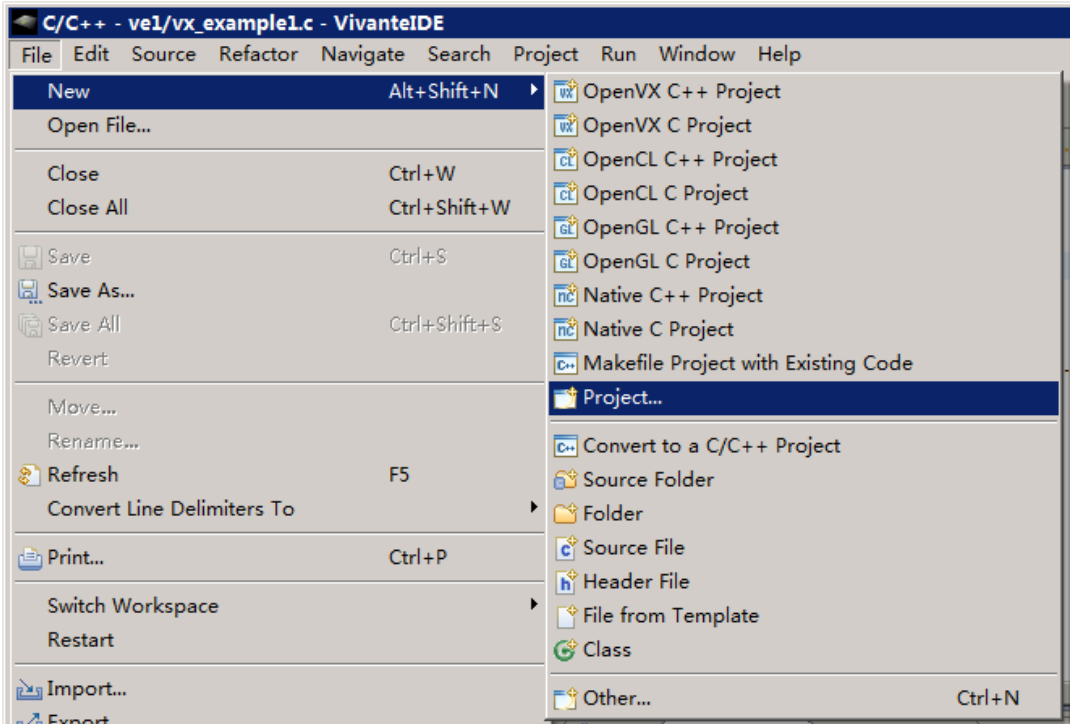
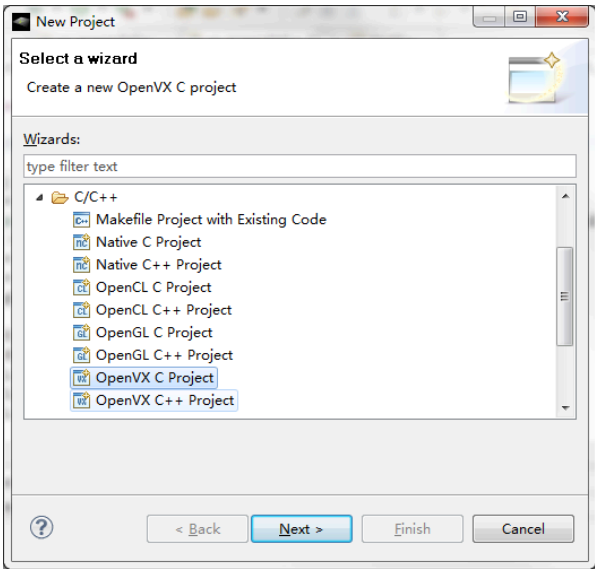


Figure 13. Creating a new project

14.4.4 Creating an OpenVX kernel wizard

1. To create an **OpenVX C(C++)** project, in the **OpenVX C(C++) Project** dialog box, enter the Project name, select **OpenVX Kernel Project(1.1)** under **Static Library** or **Shared Library**.



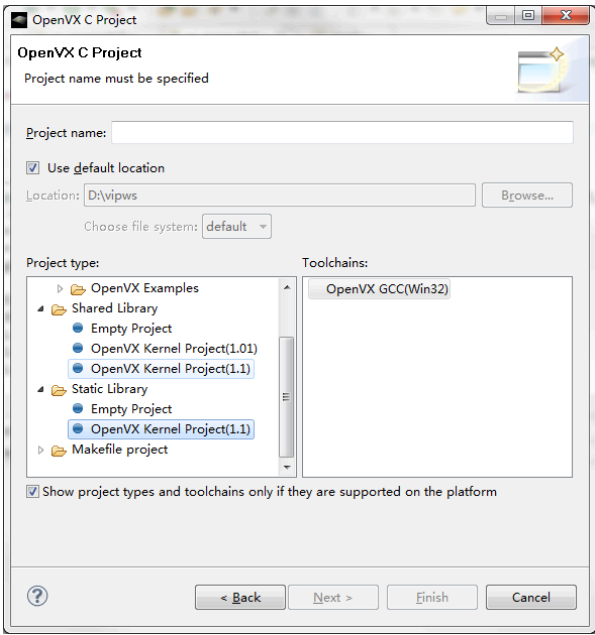


Figure 14. Creating a new project (1)

- 2. Press **Next** to input **Author** and **Copyright notice**, **Kernel ENUM offset** and **Kernel Name prefix** information in the following dialogs, and then add arguments for the kernel.

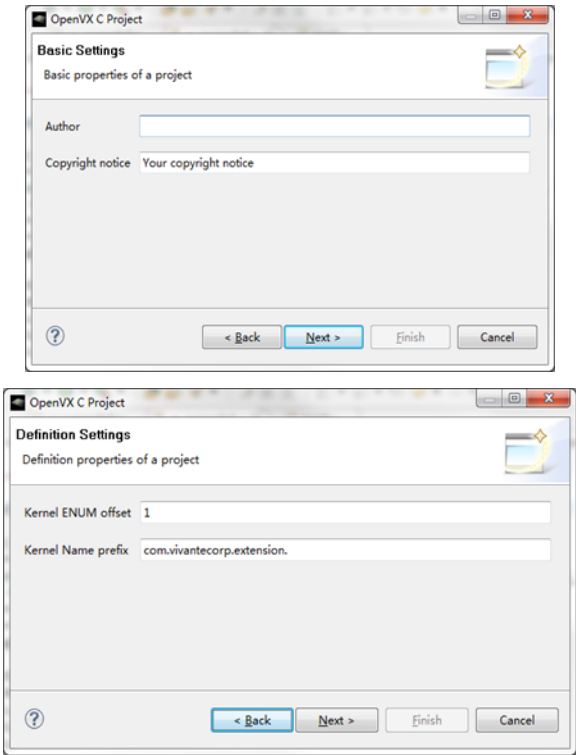


Figure 15. Creating a new project (2)

- 3. Click the **Finish** button, and the new kernel project will be created. Refer to the *VivanteIDE User Guide* for detailed information.

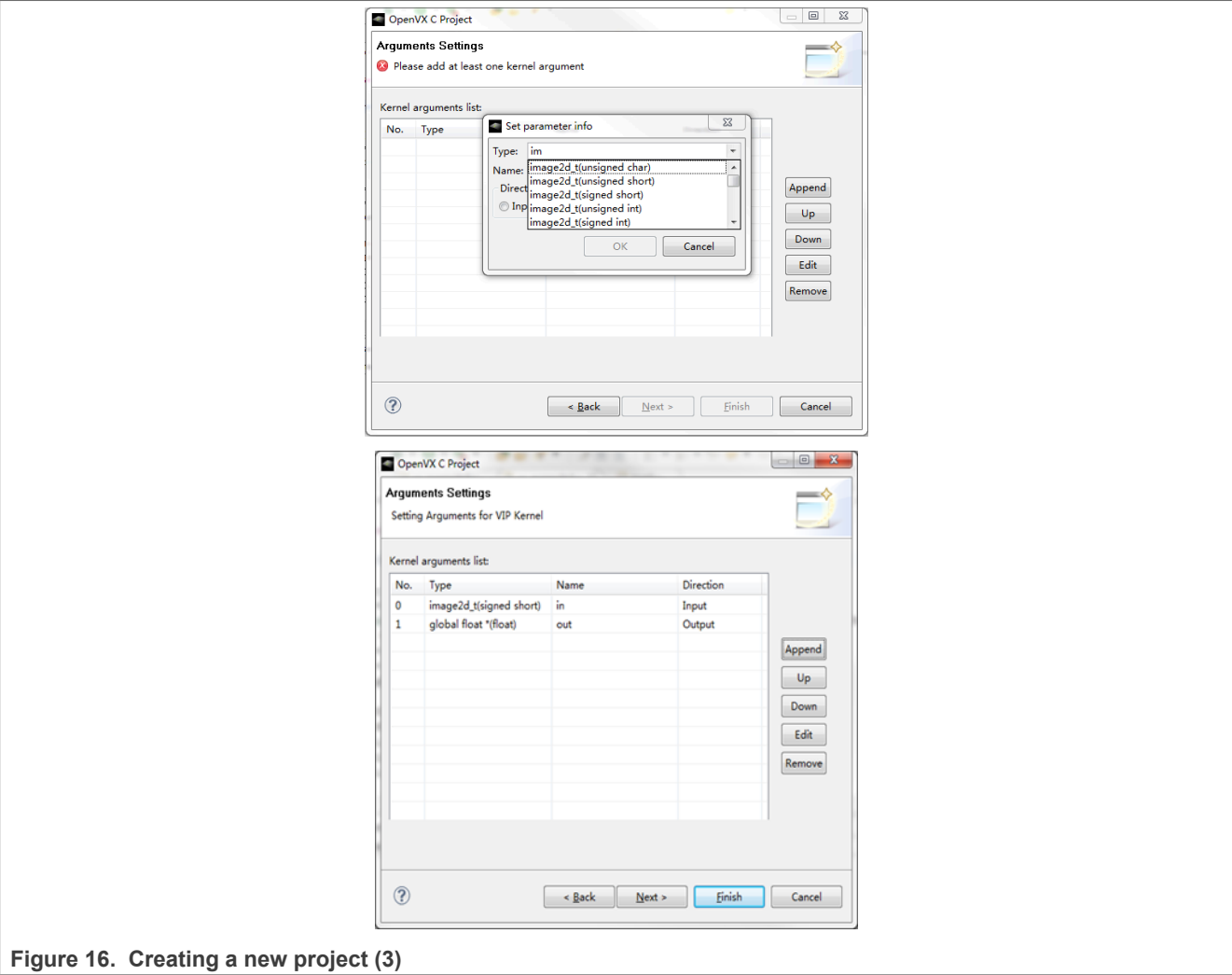


Figure 16. Creating a new project (3)

14.4.5 Source code smart editing for OpenVX and OpenCL

When a user edits a source file in VivanteIDE, the OpenVX/OpenCL keywords and predefined structure will be automatically highlighted. The Editor also supports keyword completion using keyboard combination "alt"+"/".

In addition, the **Outline** view tab will provide structured information and quick navigation for the source file currently being edited.

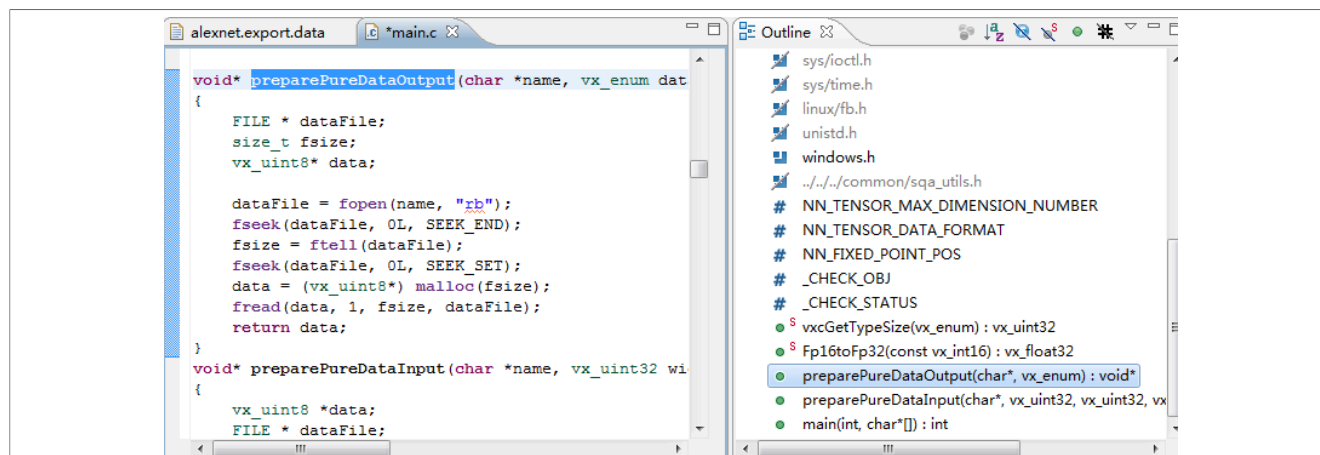


Figure 17. Source code smart editing for OpenVX and OpenCL (1)

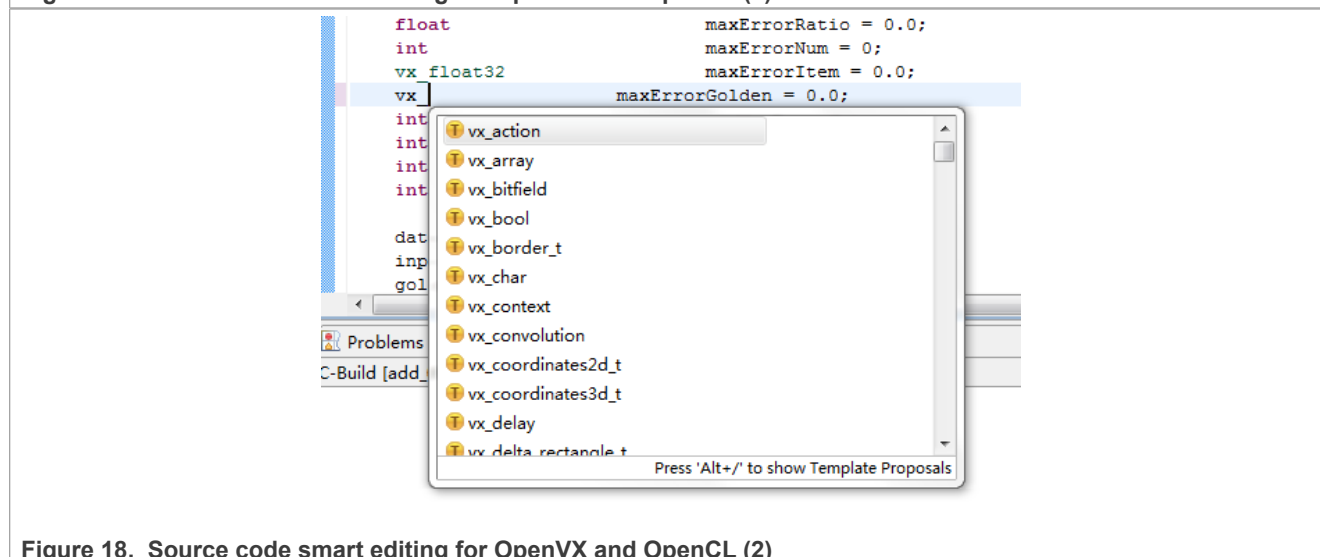


Figure 18. Source code smart editing for OpenVX and OpenCL (2)

14.4.6 Creating a Neural Network Inference Project from a model file

New project creation is available from the main menu.

1. Choose **File-->New-->Project...**

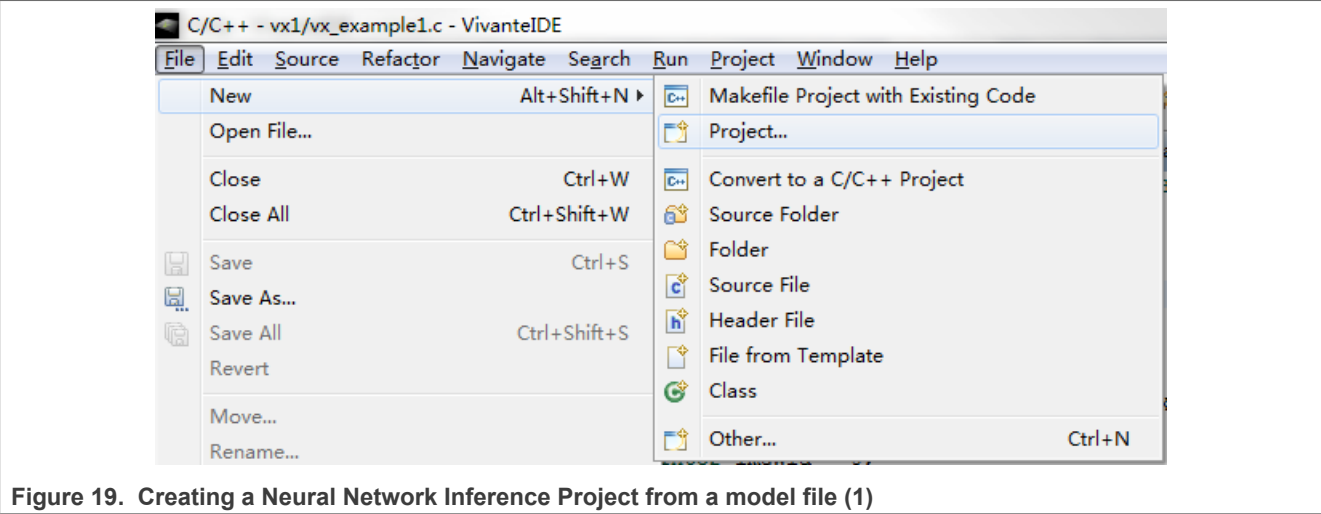


Figure 19. Creating a Neural Network Inference Project from a model file (1)

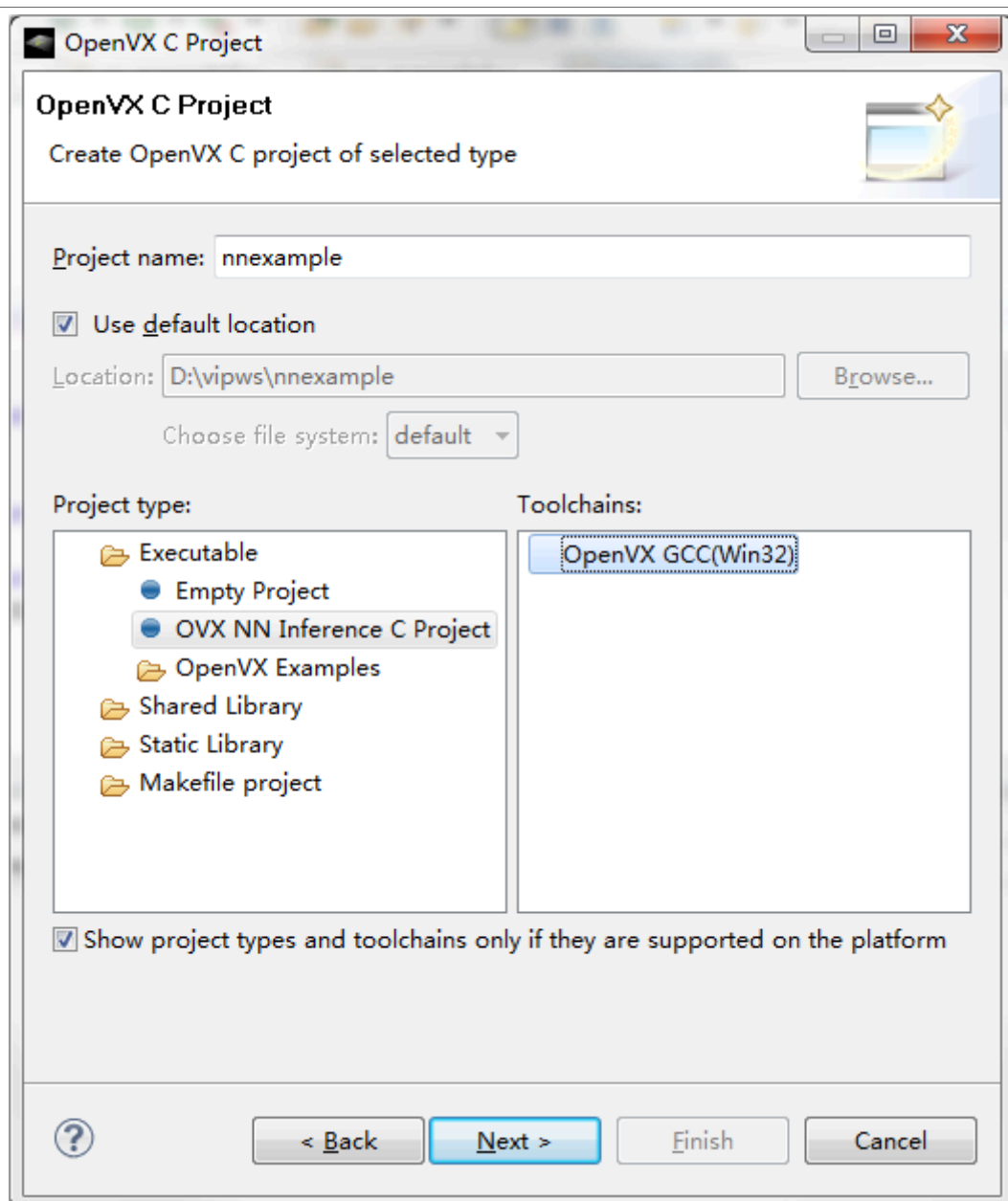


Figure 20. Creating a Neural Network Inference Project from a model file (2)

2. In the **New Project - Select a wizard** dialog box, open the **C/C++** folder in the **Wizards** list box and select **OpenVX C Project**.

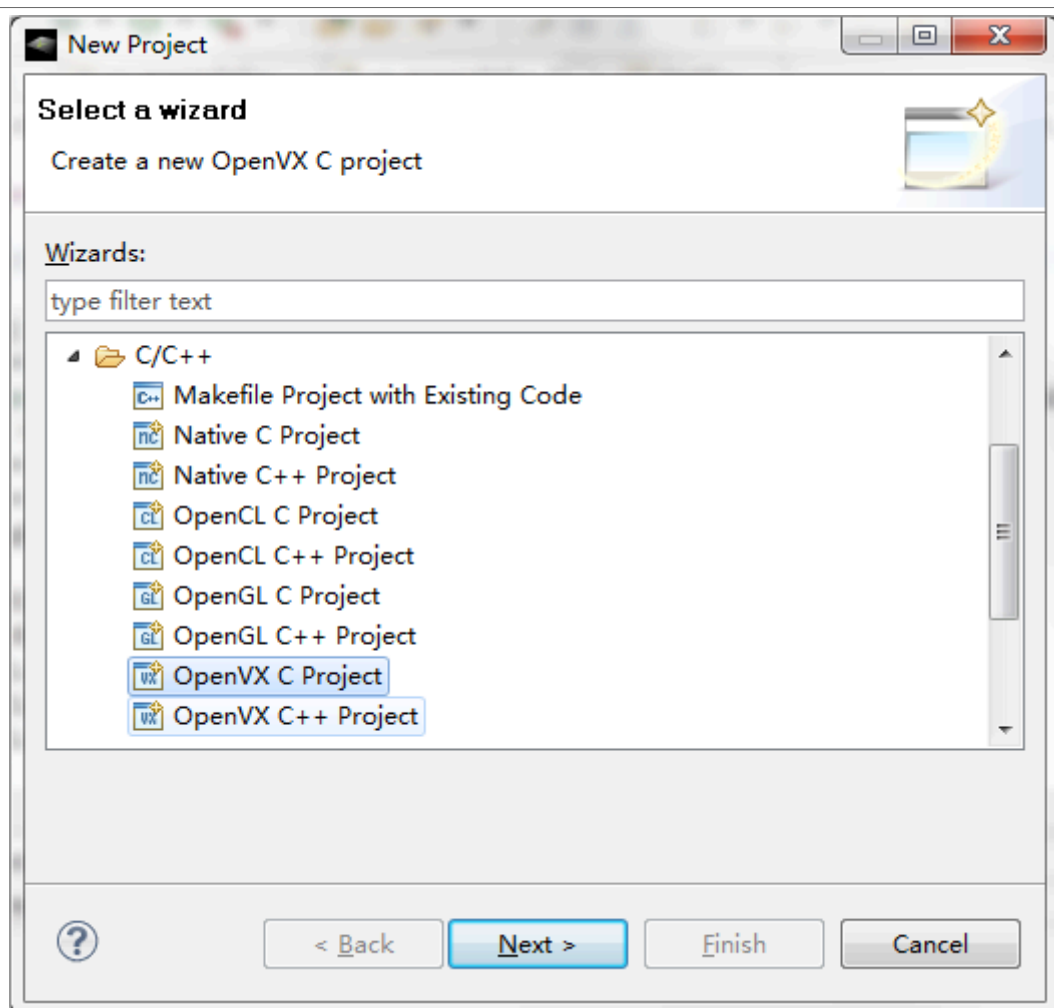


Figure 21. Creating a Neural Network Inference Project from a model file (3)

3. Click **Next** to continue.
4. In the **OpenVX C Project** dialog box, enter the Project name. Check the **Use default location** checkbox. This will cause our new directory to be created in our workspace. The directory path is displayed.
5. Select the Project type: **Executable -> OVX NN Inference C Project**.
6. Once the project name is entered, click **Next** to continue. The **OpenVX C Project - Basic Settings** dialog box is displayed.

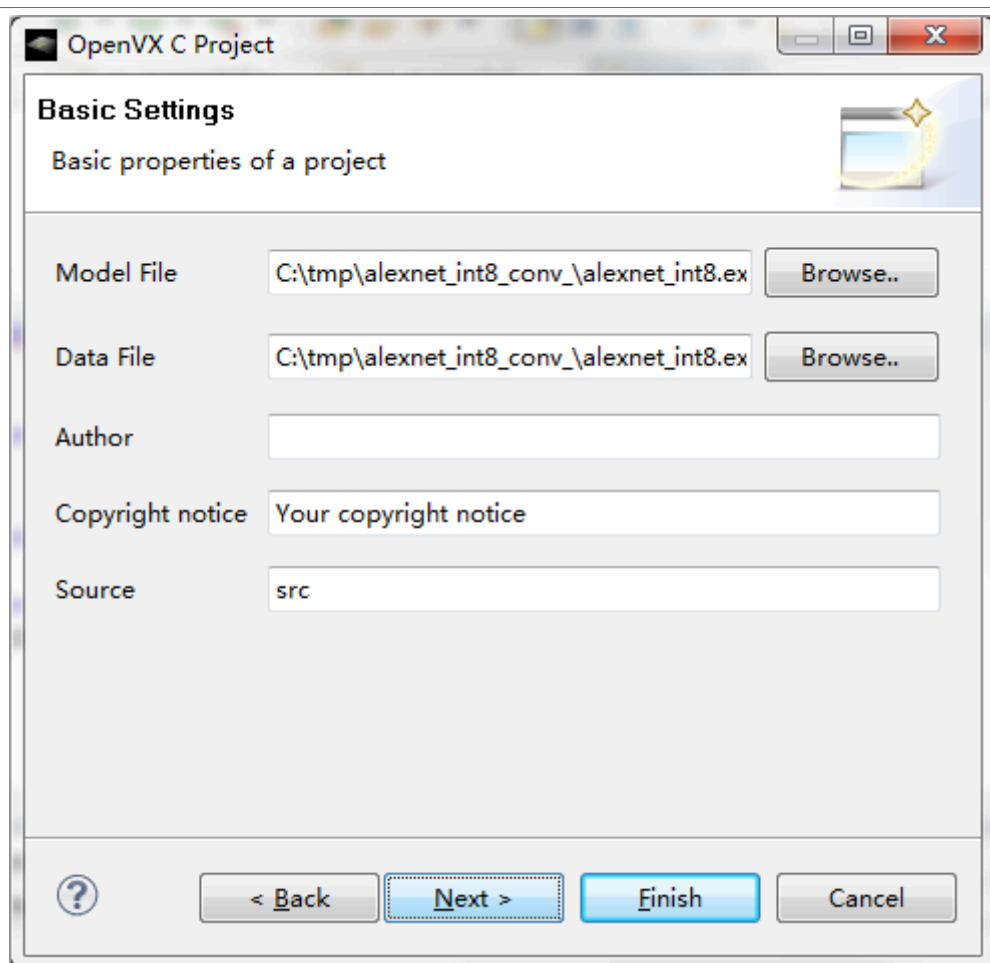


Figure 22. Figure 31. Creating a Neural Network Inference Project from a model file (4)

7. Browse or input the information to select a **Model file** and a **Data file**.
8. Click **Next** to continue. The **OpenVX C Project - Conversion Settings** dialog box is displayed. Make sure the **Add reference main.c** checkbox is checked.

Note:

*If **Add reference main.c** is checked, a **main.c** would be created by this wizard. If it is unchecked, **main.c** would not be created.*

*Function `main()` locates in **main.c**, which is just an application for testing the model.*

*Usually the NN model is a part of an OpenVX application, so writing function `main` to use the NN model is still necessary to execute the project if **Add reference main.c** is not checked.*

9. Click **Next** to continue. The **OpenVX C Project - Select Configurations** dialog box is now displayed.

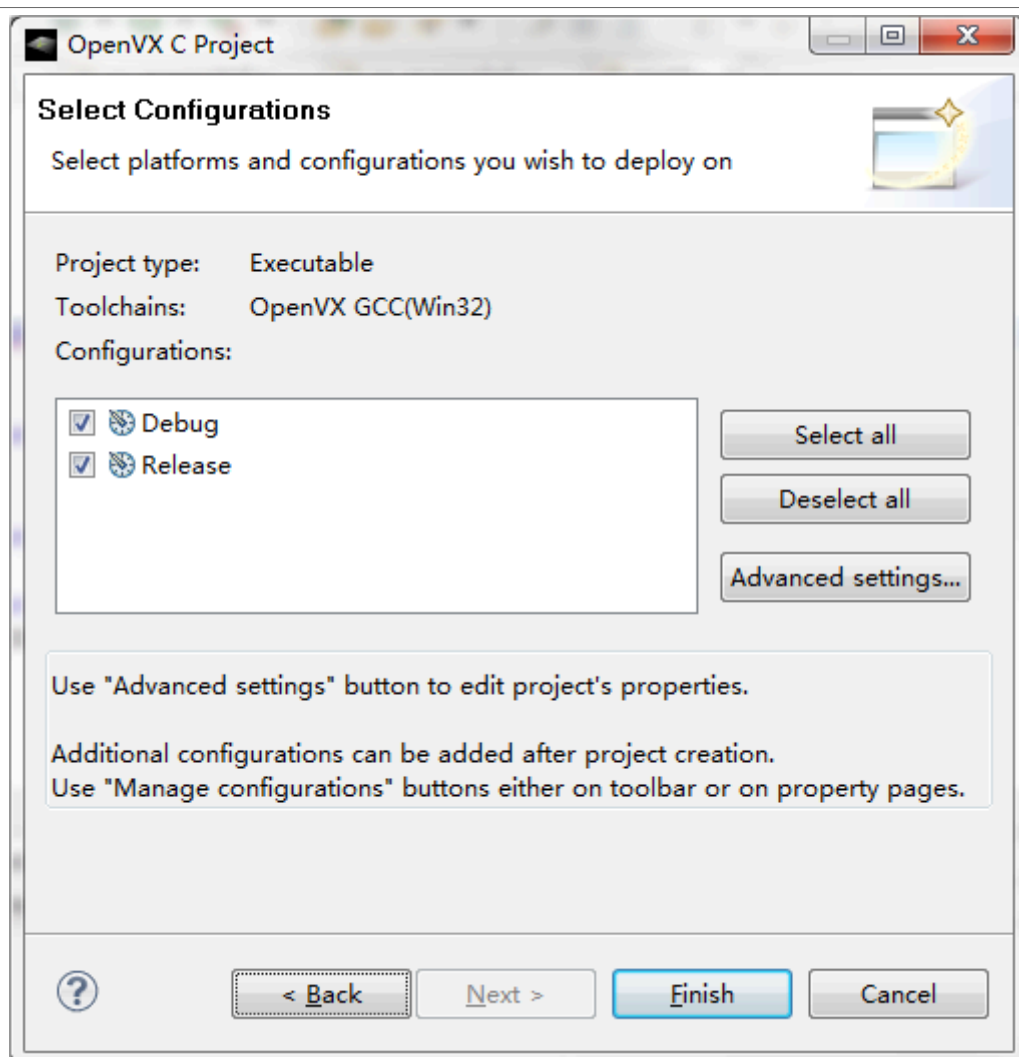


Figure 23. Creating a Neural Network Inference Project from a model file (5)

10. Click the **Finish** button. The new project is now created. The new Project is viewable in the **Project Explorer** pane.

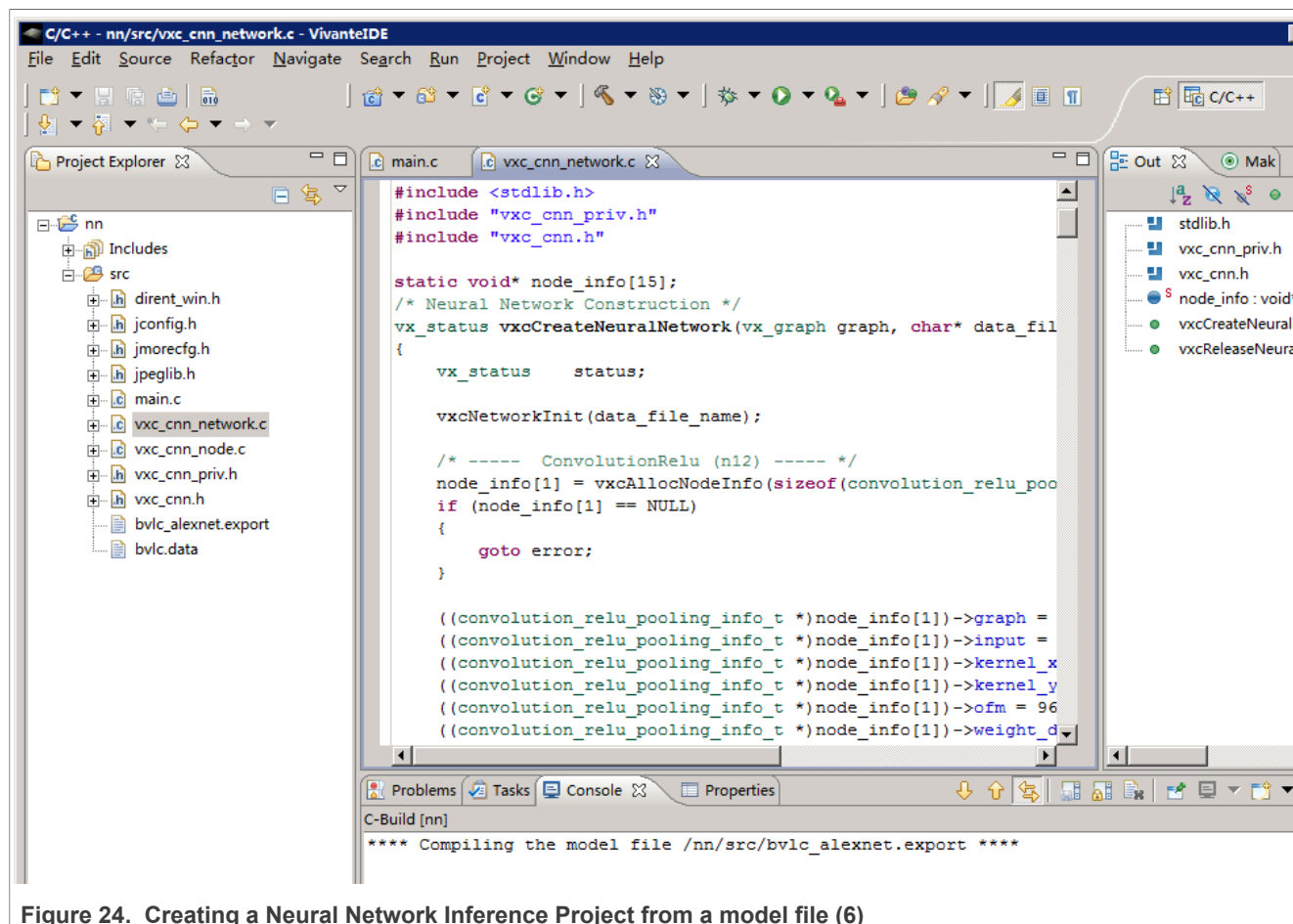


Figure 24. Creating a Neural Network Inference Project from a model file (6)

14.4.7 Building a sample project

1. On the **Project** tab, select **Properties** to open the **Properties Setting** dialog to modify the build settings.

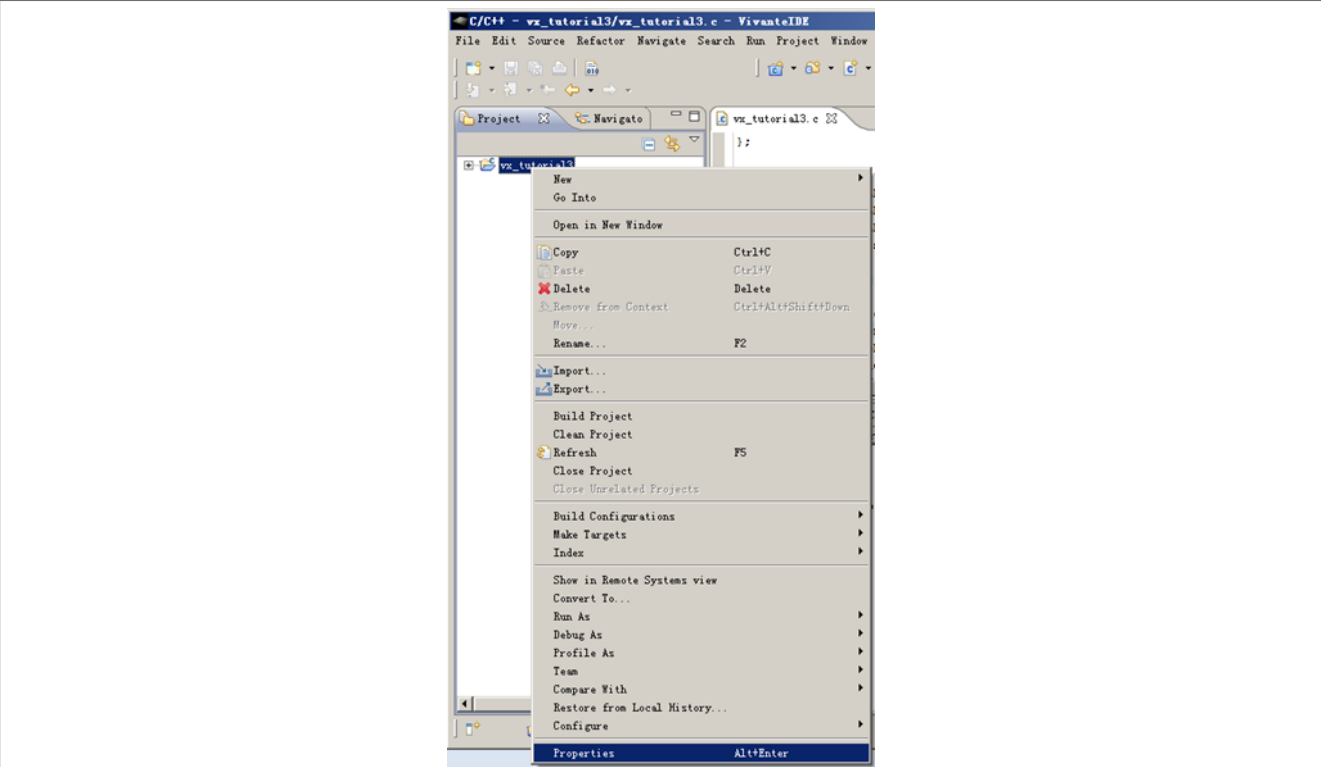


Figure 25. Building a sample project (1)

2. There are build tools available that can be set for C or C++ projects.

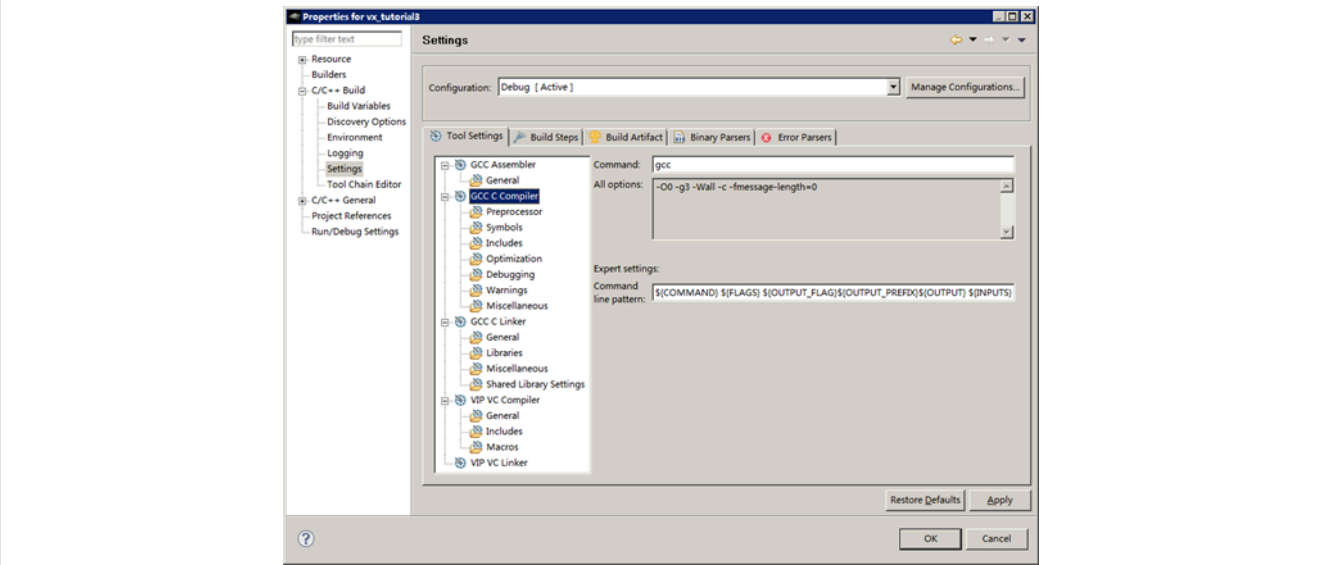


Figure 26. Building a sample project (2)

3. The sample project 'vx_tutorial3' is ready to build after the build settings are saved. You can build the 'vx_tutorial3' project by using one of following two methods, with the target project selected in the left pane:
- Choose from the main menu **Project->Build Project**.
 - Right-click the target project and select **Build Project**.

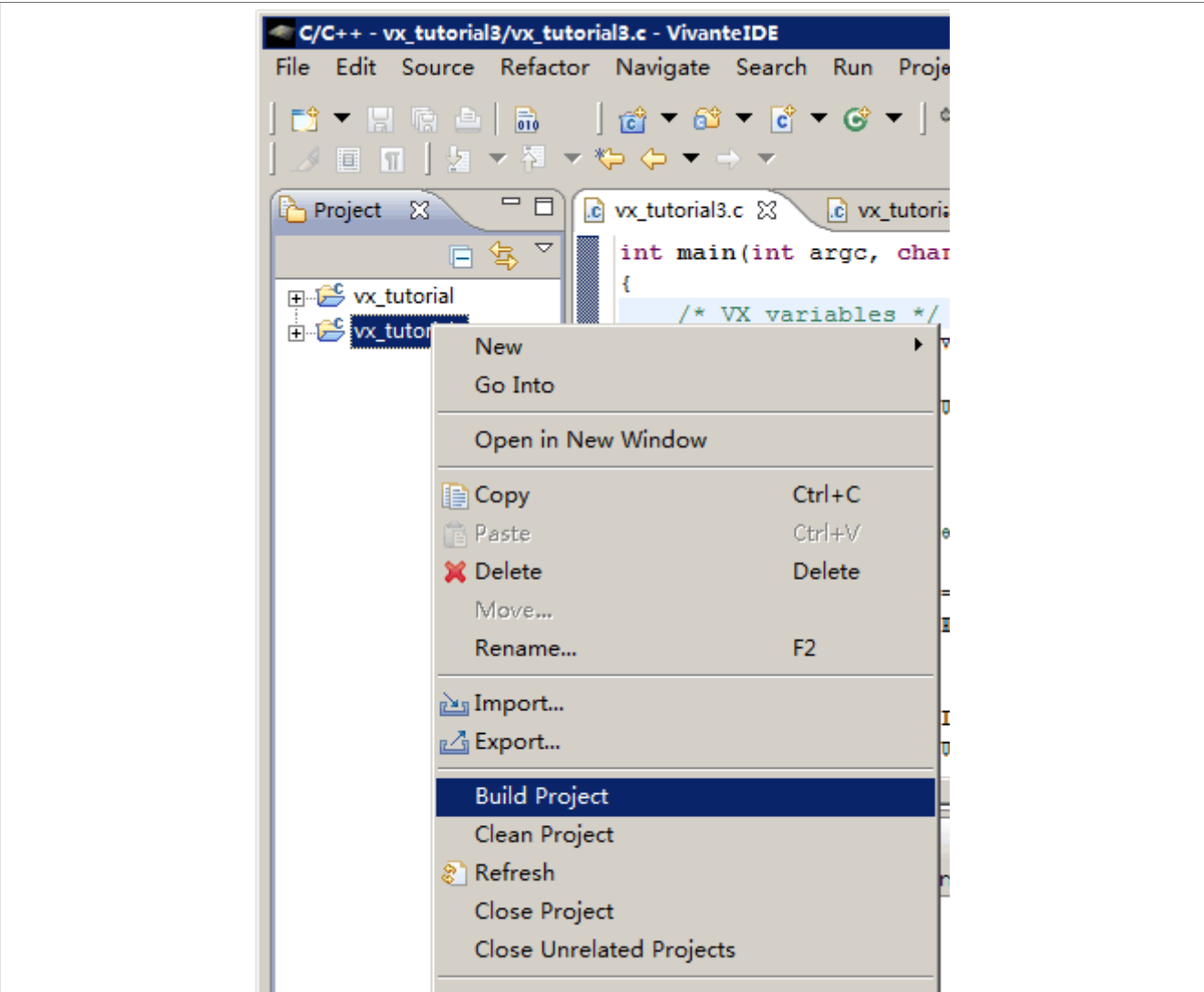


Figure 27. Building a sample project (3)

4. The build results are displayed on the **Console** and **Problems** tabs of the lower right pane of the application.

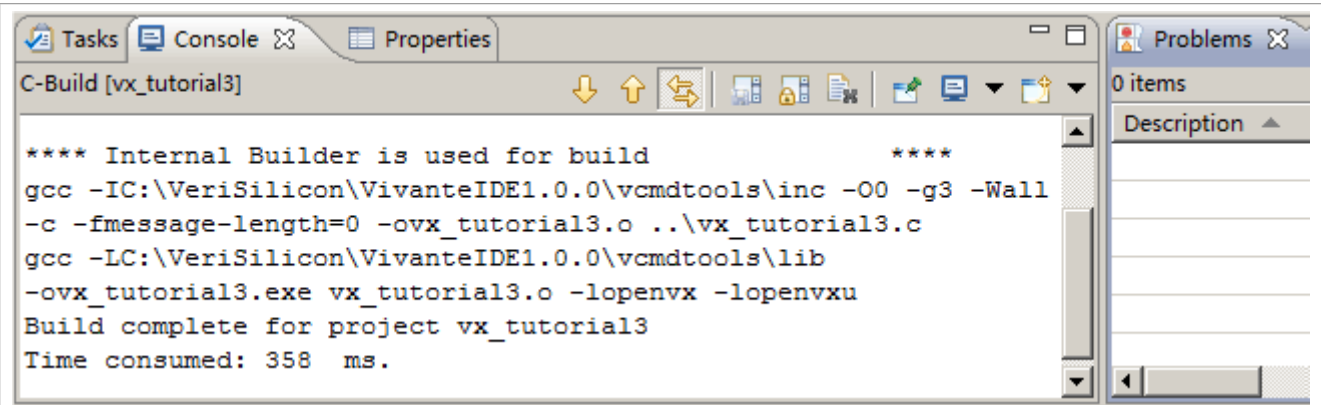


Figure 28. Building a sample project (4)

5. If **No error occurs**, build was successful, the executable file is displayed in the **Project Explorer** pane.

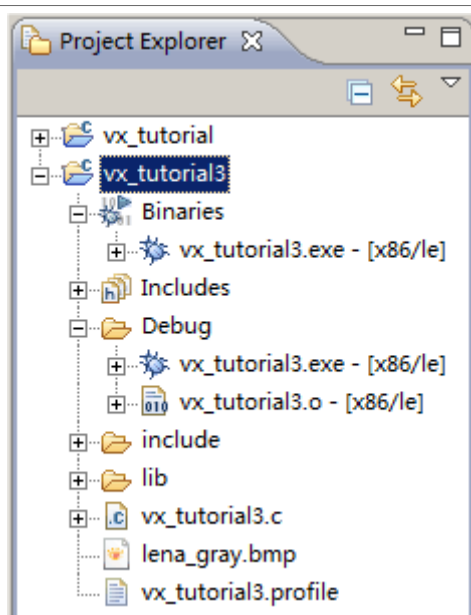


Figure 29. Building a sample project (5)

6. Use the **Build Steps** tab on the **Properties > C/C++ Build > Settings** dialog to customize the selected build configuration allowing for the specification of user defined build command steps, as well to enable displaying of descriptive messages in the build output, immediately before and after, normal build processing.

14.4.8 Debugging and profiling a project

1. To open the **Debug Configurations** dialog box, select **Run->Debug Configurations...** from the main menu.
2. Set the dialog options, and then click **Debug** to debug your project.

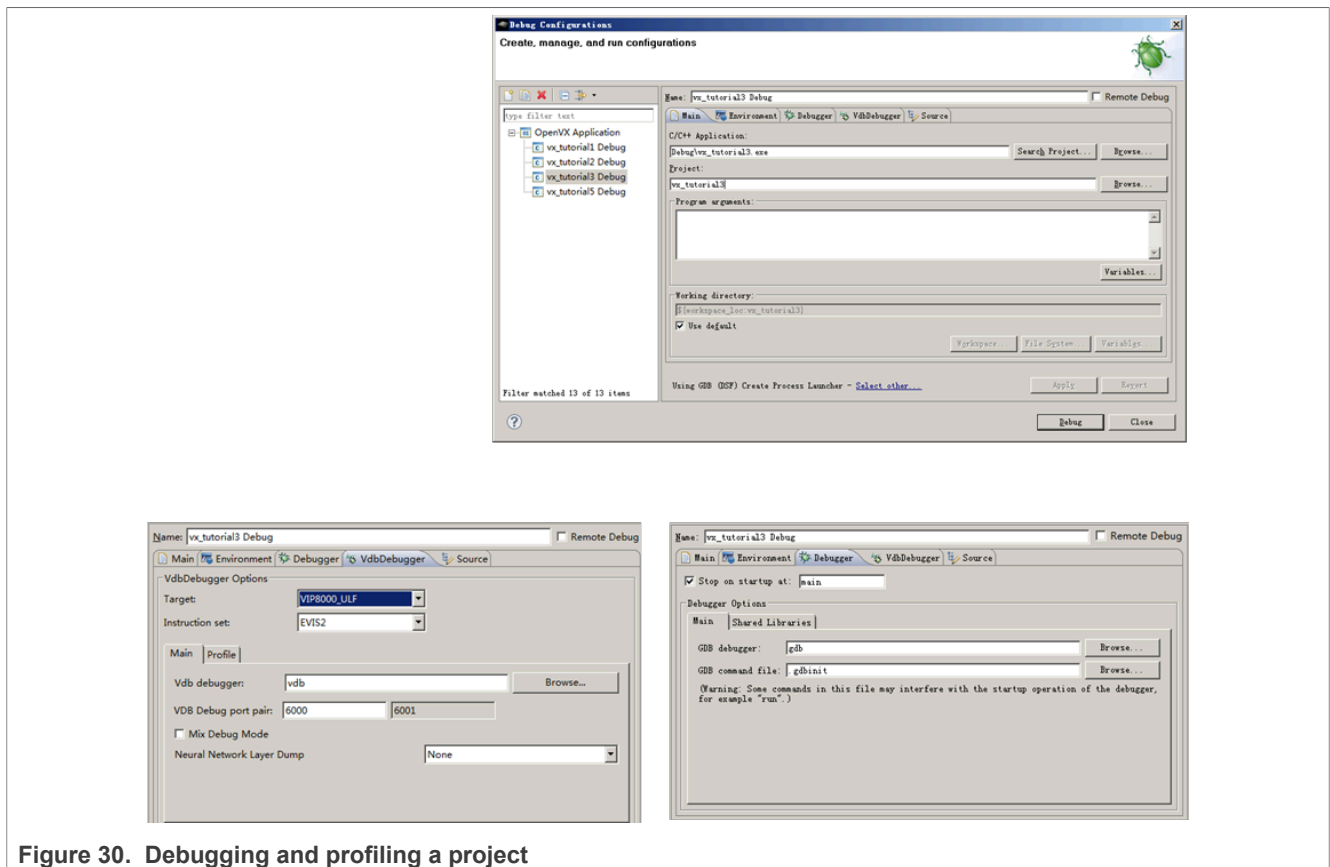


Figure 30. Debugging and profiling a project

14.5 VivanteIDE – Debug and Profiling

14.5.1 Fundamentals of performance optimization

Whenever an application runs on a computer, it makes use of one or more of the available resources. These compute resources include the CPU, the graphics processor, caches and memory, hard disks, and possibly even the network. Viewed simplistically, it is always true that one of these resources is the limiting factor in how quickly the application can finish its tasks. This limiting resource is the performance bottleneck. Remove this bottleneck, and application performance should be improved. Note, however, that removing one limiting factor always promotes something else to become the new performance bottleneck.

The goal of optimizing, or tuning application performance is to balance the use of resources so that none of them holds back the application more than any of the others. In practice, there is no single, simple way to tune an application. The whole system needs to be considered, including the size and speed of individual components as well as interactions and dependencies among components.

vProfiler collects information on GPU usage and on calls to Vivante functions within the graphics pipeline. It provides an excellent view into what is happening on the GCCORE graphics processor at any point in time, down to the individual frame. When the application performance is GPU-bound, vProfiler and VPD Analyser are the right tools to help determine why.

Note that the initial determination regarding which component of the computer system is the performance bottleneck – CPU, GPU, memory, and so on, which is the domain of system performance analyzers and is outside the scope of the GPU tools. A list of such performance analysis tools can be found at Wikipedia:

en.wikipedia.org/wiki/List_of_performance_analysis_tools

14.5.2 VPD Analyzer for Analyzing Performance Data

vProfiler is a run-time environment for collecting performance statistics of an application and the graphics pipeline. The VPD Analyzer perspective view is provided to facilitate graphically displaying the data gathered by vProfiler and aiding in visual analysis of graphics performance. Used together, these tools assist software developers in optimizing application performance on Vivante enabled platforms.

14.5.3 vProfiler

When building Vivante Graphics Drivers, the driver is built with vProfiler capability. vProfiler gathers data from these counters during runtime and can track data for a range of frames or a single frame from any graphics, compute application. vProfiler outputs performance data to binary files with a `.vpd` extension. These files can be using the VivanteIDE VPD Analyzer both in text lists and as line graphs. VPD Analyzer gives the user several ways to inspect any frame in a captured animation sequence.

14.5.4 Enabling vProfiler on Linux OS

When building Vivante Graphics Drivers in a Linux OS environment, the driver is built with vProfiler capability.

- vProfiler functionality can be enabled by `export VIV_PROFILE=1`.
- To enable OpenVX profile, use `export VIV_VX_PROFILE=1`.
- To enable OpenCL profile, use `export VIV_CL_PROFILE=1`.

Kernel module driver arguments are no longer needed.

14.5.4.1 Setting vProfiler property options for OpenGL ES

vProfiler property options are set using environment variables on Linux. The following table summarizes the environment variables that vProfiler supports.

Table 37. vProfiler property options

Environment Variable	Description
VIV_PROFILE	[0] Disable vProfiler (default), [1] Enable vProfiler, [2] Control via application call, [3] Allows control over which frames to profile with vProfiler
VP_OUTPUT	Specify the output file name of vProfiler (default is <code>vprofiler.vpd</code>)
VP_FRAME_NUM	When VIV_PROFILE=1, specify the number of frames dumped by vProfiler.
VP_FRAME_START	When VIV_PROFILE=3, specify the frame to start profiling with vProfiler.
VP_FRAME_END	When VIV_PROFILE=3, specify the frame to end profiling with vProfiler.
VP_USE_GLFINISH	Enable [1] or disable [0] the use of <code>glFinish()/glFlush()</code> APIs as the frame delimiter in addition to <code>eglSwapBuffers()</code> (default 0). This variable enables application thread which does not use <code>eglSwapBuffers()</code> to generate useful GPU profiling data for analysis.
VP_PERDRAW_MODE	Enable [1] or disable [0] (default). When enabled, vProfiler will collect a counter for each draw call.
VP_DISABLE_PROBE	Disables PROBE mode and makes vProfiler to use AHB counters for profiling.
VP_ENABLE_PRINT	Enable vProfiler to print out the counter information to the console.

14.5.5 Setting vProfiler property options for Vision, OpenVX Profiling

vProfiler for OpenVX Profiling (for use with Vision/VIP/VX IP) is similar to vProfiler for OpenGL, except that fewer environment variables and fewer supported values for those variables are available.

Table 38. vProfiler property options

Environment Variable	Description
VIV_VX_PROFILE	[0] Disable vProfiler for OpenVX(default), [1] Enable vProfiler for OpenVX
VIV_CL_PROFILE	[0] Disable vProfiler for OpenCL(default), [1] Enable vProfiler for OpenCL
VP_OUTPUT	Specify the output file name of vProfiler (default is vprofiler.vpd)

14.5.6 Enabling vProfiler Option for Android OS

i.MX Android release GPU drivers are built with vProfiler capability. To enable the vProfiler feature, boot the Android image, and then stop U-Boot by pressing a key on the serial terminal.

```
setenv append_bootargs galcore.powerManagement=0 galcore.gpuProfiler=1
boota
```

Perform the following steps to capture the VPD file using vProfiler on Android OS.

Note: For Android versions earlier than 11.0.0 2.x.y, remove the "vendor." prefix from the property name.

1. Set application name to be profiled, for example, nenamark2 application.

```
setprop vendor.VP_PROCESS_NAME se.nena.nenamark2
```

2. Set the profile output file path, for example, nenamark2 application.

```
setprop vendor.VP_OUTPUT /data/data/se.nena.nenamark2/
```

For Android Automotive, a path to the current user storage has to be used (default user ID is 10): /data/user/<user_id>/se.nena.nenamark2/.

3. Start profiling.

```
setprop vendor.VIV_PROFILE 1
```

4. Run application and check if the *.vpd file is generated in the path indicated by vendor.VP_OUTPUT, for example, nenamark2 application.

```
ls -l /data/data/se.nena.nenamark2/*.vpd
```

5. Stop profiling.

```
setprop vendor.VIV_PROFILE 0
```

14.5.7 Setting vProfiler property options for OpenGL ES Profiling with Android

The following table summarizes the property options that vProfiler supports through running the command adb shell setprop [OPTIONS]. These options are similar to the environment variables available for Linux.

Table 39. vProfiler property options

adb shell setprop OPTIONS	Description
setprop vendor.VIV_PROFILE 0	Run this command in adb shell to disable vProfiler in the drivers
setprop vendor.VIV_PROFILE 1	Run this command in adb shell to enable vProfiler in the drivers
setprop vendor.VIV_PROFILE 2	Run this command in adb shell to have vProfiler enable/disable controlled in the application by glEnable(GL_PROFILE_VIV) and glDisable(GL_PROFILE_VIV) calls.
setprop vendor.VIV_PROFILE 3	Run these commands in adb shell to have vProfiler start-stop at frames specified in vendor.VP_FRAME_START and vendor.VP_FRAME_END.

Table 39. vProfiler property options...continued

adb shell setprop OPTIONS	Description
<pre>setprop vendor.VIV_FRAME_START xxx setprop vendor.VP_FRAME_END xxx</pre>	
<pre>setprop vendor.VP_PROCESS_NAME appname</pre>	<p>Run this command in adb shell to specify the application you need to profile. Change the app name as needed to profile another application.</p> <p>Note: There may be different sub-case names used by an app. Be sure to accurately specify a case name to match the name that you saw on the command line when using <code>ps</code> command. This option is only available for Android, not available for Linux.</p>
<pre>setprop vendor.VP_OUTPUT newpath</pre>	<p>Run this command in adb shell to specify a new location for vProfiler output. By default, the vpd file will created under <code>/sdcard/</code>. If an application has no access to the SD card, you can specify another path where the application does have write permission.</p> <p>Note: For applications which initialize during Android system boot startup, such as launcher, you need to kill the process after you change to a new path. When the application automatically restarts, then your vpd will be accessible where you want it.</p>
<pre>setprop vendor.VP_FRAME_NUM xxx</pre>	<p>Run this command in adb shell to limit the number of frames to analyze. For example, to make vProfiler dump performance data for the first 100 frames:</p> <pre>setprop vendor.VP_FRAME_NUM 100.</pre> <p>Note: Only use when <code>vendor.VIV_PROFILER</code> is set to 1. When this option is not used, the profile file generated when running an application for a long time can be very large. This takes up a large amount of disk space and also makes it hard to view the data in vAnalyzer.</p>
<pre>setprop vendor.VP_USE_GLFINISH 0 setprop vendor.VP_USE_GLFINISH 1</pre>	<p>Run this command in adb shell to enable or disable use of <code>glFinish()</code> / <code>glFlush()</code> as the frame delimiter in addition to <code>eglSwapBuffers()</code> (default 0). By default, <code>eglSwapBuffers()</code> is used as the frame delimiter. This command will make application thread which does not use <code>eglSwapBuffers()</code> to generate useful GPU profiling data for analysis.</p>
<pre>setprop vendor.VP_PERDRAW_MODE 0 setprop vendor.VP_PERDRAW_MODE 1</pre>	<p>Run this command in adb shell to enable or disable per draw mode. When enabled, vProfiler will collect a counter for each draw call.</p>
<pre>setprop vendor.VP_DISABLE_PROBE 1</pre>	<p>Run this command in adb shell to disable PROBE mode and make vProfiler use AHB counters for profiling.</p>
<pre>setprop vendor.VP_ENABLE_PRINT 1</pre>	<p>Run this command in adb shell to enable vProfiler to print out the counter information to the console.</p>

14.5.8 vProfiler Set Property Options for Vision/OVX Profiling with Android

vProfiler for Vision Profiling (for use with Vision/VIP/VX IP) is similar to vProfiler for OpenGL, except that fewer property options and fewer supported values are available.

Table 40. vProfiler Set Property Options

adb shell setprop OPTIONS for VIP/VX/OVX	Description
<pre>setprop vendor.VIV_VX_PROFILE 0</pre>	Run this command in adb shell to disable vProfiler in the drivers

Table 40. vProfiler Set Property Options...continued

adb shell setprop OPTIONS for VIP/VX/OVX	Description
setprop vendor.VIV_VX_PROFILE 1	Run this command in adb shell to enable vProfiler in the drivers
setprop vendor.VP_PROCESS_NAME appname	Run this command in adb shell to specify the application you need to profile. Change the app name as needed to profile another application. Note: There may be different sub-case names used by an app. Be sure to accurately specify a case name to match the name that you saw on the command line when using ps command. This option is only available for Android, not available for Linux.
setprop vendor.VP_OUTPUT newpath	Run this command in adb shell to specify a new location for vProfiler output. By default, the vpd file will be created under /sdcard/. If an application has no access to the SD card, you can specify another path where the application does have write permission. Note: For applications that initialize during Android system boot startup, such as launcher, you need to kill the process after you change to a new path. When the application automatically restarts, then your vpd will be accessible where you want it.

14.5.9 Enabling vProfiler Option for QNX

When building the Vivante Graphics Drivers for QNX environment, build the driver with the vProfiler capability.

The `graphics.conf` file contains the configuration information for Screen and is found under the following directory:

SCREEN-DIR/usr/lib/graphics/TARGET-SPECIFIC

To activate the vProfiler functionality, add the **gpu-gpuProfiler=1** option into the *khronos* section of the corresponding `graphics.conf` file:

```
begin khronos
...
begin wfd device 1
...
gpu-gpuProfiler=1
...
end wfd device
...
end khronos
```

14.5.9.1 Setting vProfiler Environment Variables for OGL/OES Profiling

The following table summarizes the environment variables that vProfiler supports.

Table 41. vProfiler Environment Variables

Environment Variable	Description
VIV_PROFILE	[0] Disable vProfiler (default), [1] Enable vProfiler, [2] Control via application call, [3] Allows control over which frames to profile with vProfiler
VP_OUTPUT	Specify the output file name of vProfiler (default is vprofiler.vpd)
VP_FRAME_NUM	When VIV_PROFILE=1, specify the number of frames dumped by vProfiler.
VP_FRAME_START	When VIV_PROFILE=3, specify the frame to start profiling with vProfiler.

Table 41. vProfiler Environment Variables...continued

Environment Variable	Description
VP_FRAME_END	When VIV_PROFILE=3, specify the frame to end profiling with vProfiler.
VP_USE_GLFINISH	Enable [1] or disable [0] the use of glFinish()/glFlush() APIs as the frame delimiter in addition to eglSwapBuffers() (default 0). This variable enables application thread which does not use eglSwapBuffers() to generate useful GPU profiling data for analysis.
VP_PERDRAW_MODE	Enable [1] or disable [0] (default). When enabled, vProfiler will collect a counter for each draw call.
VP_DISABLE_PROBE	Disables PROBE mode and makes vProfiler to use AHB counters for profiling.
VP_ENABLE_PRINT	Enable vProfiler to print out the counter information to the console.

14.5.9.2 Setting vProfiler Environment Variables for Vision, OpenVX Profiling

vProfiler for OpenVX Profiling (for use with Vision/VIP/VX IP) is similar to vProfiler for OpenGL, except that fewer environment variables and fewer supported values for those variables are available.

Table 42. vProfiler Environment Variables

Environment Variable	Description
VIV_VX_PROFILE	[0] Disable vProfiler for OpenVX(default), [1] Enable vProfiler for OpenVX
VIV_CL_PROFILE	[0] Disable vProfiler for OpenCL(default), [1] Enable vProfiler for OpenCL
VP_OUTPUT	Specify the output file name of vProfiler (default is vprofiler.vpd)

14.5.10 Environment Variable Details

14.5.10.1 VIV_PROFILE

The environment variable VIV_PROFILE can be used to control enable/disable and set profiling modes for vProfiler.

- **VIV_PROFILE=0**

By default, vProfiler is disabled in the driver. If vProfiler has been enabled and you wish to disable it, set VIV_PROFILE to 0:

```
export VIV_PROFILE=0
```

- **VIV_PROFILE=1**

To enable vProfiler, set VIV_PROFILE to 1:

```
export VIV_PROFILE=1
```

To limit the number of frames to analyze, use the environment variable VP_FRAME_NUM. (This option is available only when VIV_PROFILE=1.) For example, this setting will make vProfiler dump performance data for the first 100 frames.

```
export VP_FRAME_NUM=100
```

- **VIV_PROFILE=2**

Mode VIV_PROFILE=2 provides support for glEnable(GL_PROFILE_VIV) and glDisable(GL_PROFILE_VIV), which are used to choose which frames are to be profiled. In this mode, vProfiler is disabled by default. It begins to do profiling only after a glEnable(GL_PROFILE_VIV) call from the application. And it will stop

profiling when `glDisable (GL_PROFILE_VIV)` is called. Note that the flag is only checked at every frame end, i.e., in `eglSwapBuffers()`. To use this mode, set `VIV_PROFILE` to 2:

```
export VIV_PROFILE=2
```

- **VIV_PROFILE=3**

Setting `VIV_PROFILE` to 3 provides support for two environment variables `VP_FRAME_START` and `VP_FRAME_END`, which are used to choose which frames are to be profiled. In this mode, vProfiler is disabled by default. It begins to do profiling starting at the frame number specified by `VP_FRAME_START`, and it ends the profiling after the frame number specified by `VP_FRAME_END`. For example to use this mode, set `VIV_PROFILE` to 3:

```
export VIV_PROFILE=3 export VP_FRAME_START=10 export VP_FRAME_END=90
```

Note:

*To get precise profiling data, the IP's Power Management (PM) functions need to be disabled. When kernel module **galcore** is inserted with `gpuProfiler=1`, the PM functions in the driver are not disabled. The PM functions are disabled when `VIV_PROFILE` is set to 1, 2, or 3, and the application starts. The PM functions are enabled when `VIV_PROFILE` is set to 0, and the application starts again.*

14.5.10.2 VP_OUTPUT

The output file of vProfiler is `vprofiler.vpd` by default. To specify an alternate filename use the environment variable `VP_OUTPUT`. For example,

```
export VP_OUTPUT=sample.vpd
```

14.5.10.3 VP_USE_GLFINISH

`glFinish()/glFlush()` will be treated as the frame delimiter in addition to `eglSwapBuffers()`. By default, vProfiler only uses `eglSwapBuffers()` as the delimiter to check hardware counters. The command below will enable vProfiler to use `glFinish()/glFlush()` as additional delimiters so an application thread which does not use `eglSwapBuffers()` can generate useful profiling data for analysis.

```
export VP_USE_GLFINISH=1
```

14.5.10.4 VP_DISABLE_PROBE

This variable only applies to IP with the PROBE feature support. It disables PROBE mode and makes vProfiler use AHB counters for profiling. This variable has no affect on hardware that only supports the AHB counter. The default value is off.


14.5.10.5 VP_ENABLE_PRINT

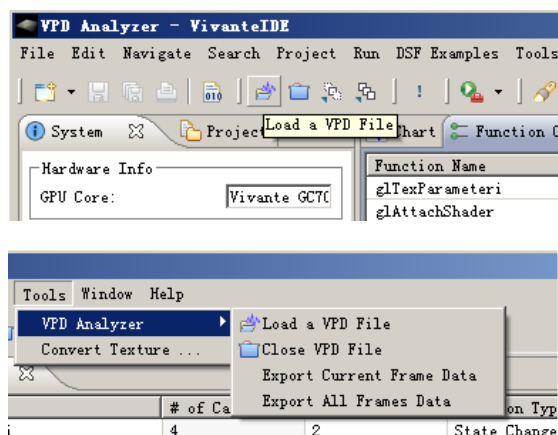
This variable provides a convenient way to check some critical profiling information without using the off-line vAnalyzer to open a VPD file. Once it is enabled, vProfiler prints out the counter information to the console. For the OpenVX and OpenCL drivers, the default value is on; for GLES and GL drivers, the default value is off.

14.6 VPD Analyzer

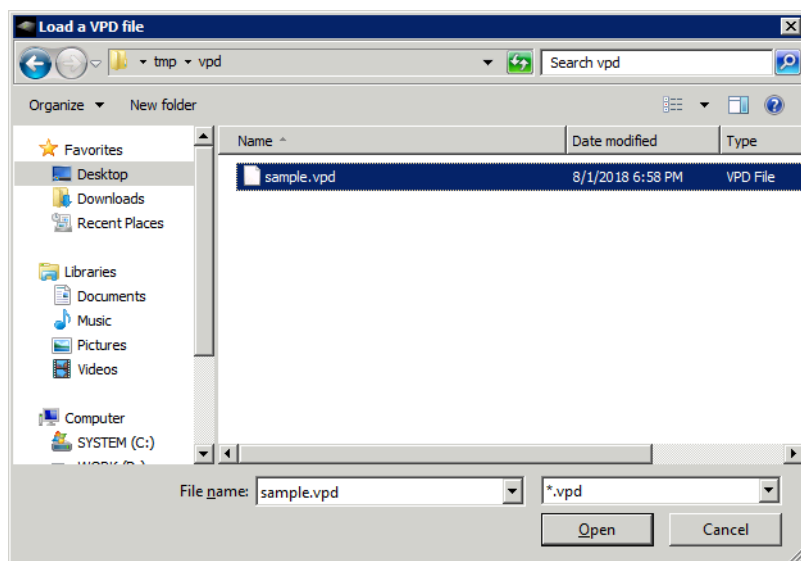
VPD Analyzer provides graphic displays of the data gathered by vProfiler and aids in the visual analysis of graphics, compute and vision performance. vProfiler outputs performance data to binary files with a `.vpd` extension. These files can be opened using the VivanteIDE VPD Analyzer both in text lists and as line graphs. VPD Analyzer gives the user several ways to inspect any frame in a captured animation sequence.

14.6.1 Loading a VPD File

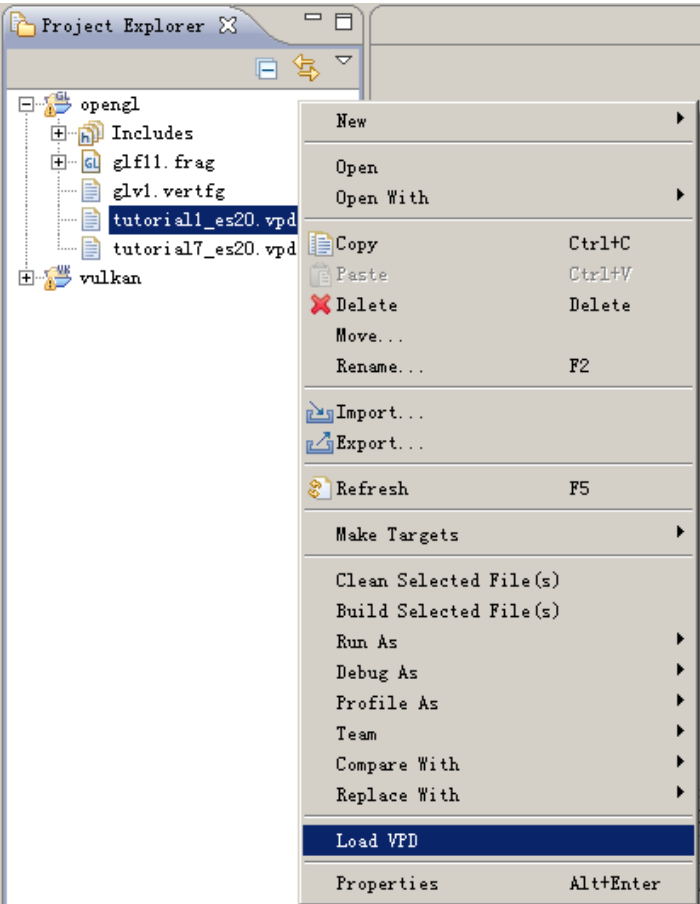
To open the VPD Analyzer perspective based on a VPD file, click the icon  from the toolbar or select **Tools->VPD Analyzer->Load VPD File ...**



The **Load a VPD file** dialog box appears. Select a VPD (.vpd) file, and click **Open**.



Or, in the Project Explorer view, right-click on a VPD file and select Load VPD.



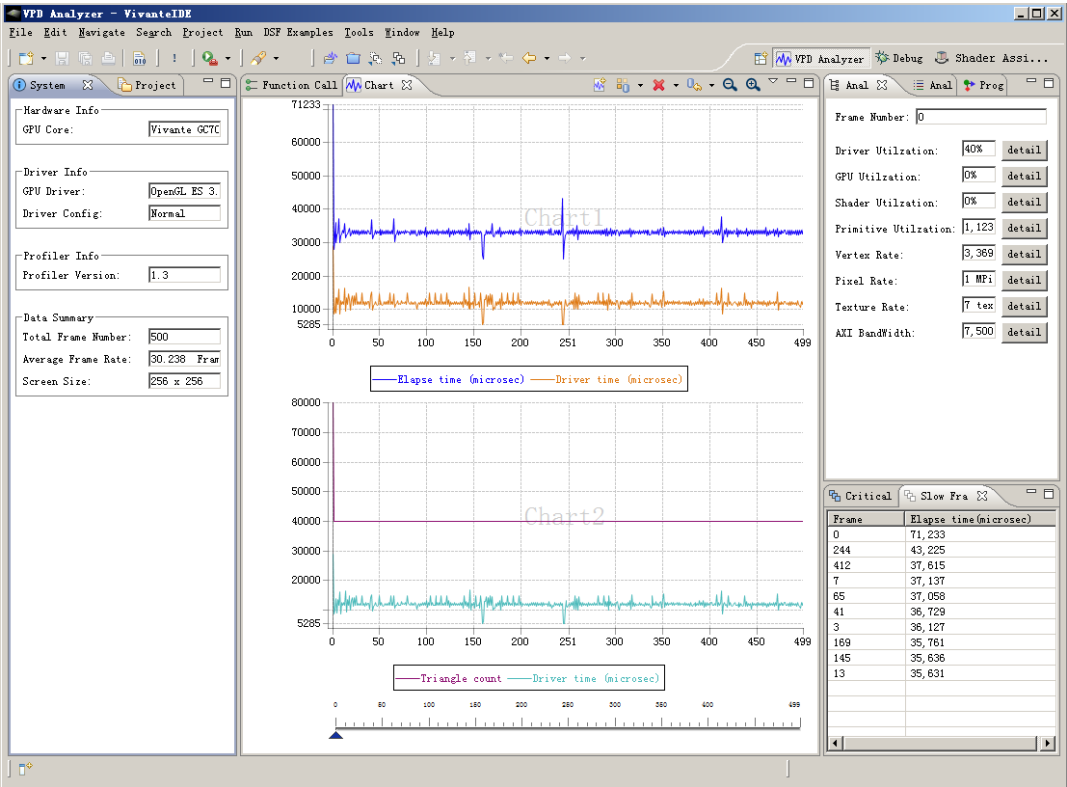
14.6.2 VPD Analyzer Perspective

Once the VPD file is loaded, the VivanteIDE workbench switches to the VPD Analyzer perspective view, and analyze data from the selected VPD file will be displayed on a series of tabs in chart or text format.

Available tabs (left to right) are:

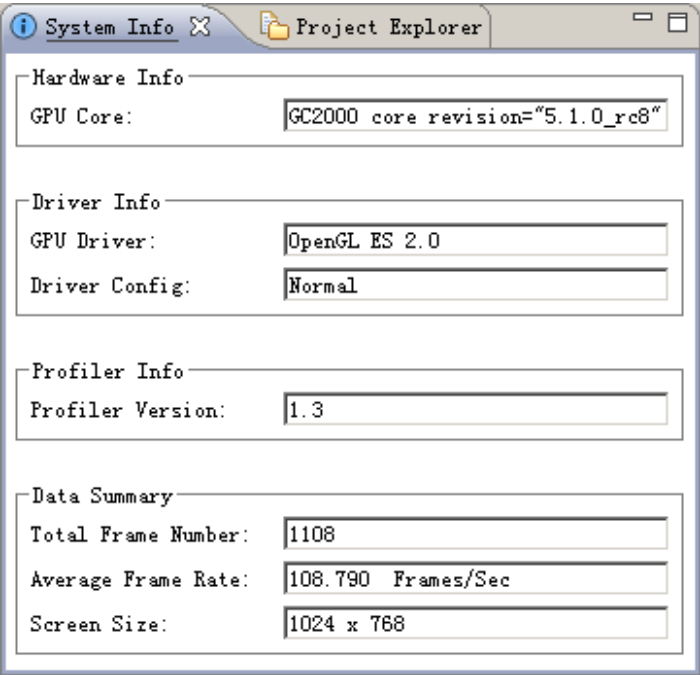
Table 43. Available tabs

VPD Analyzer Tab	Description
System Info	Shows hardware and software version information and Average Frame Rate
Project Explorer	Shows project files
Chart	Shows customizable graph views of various counters
Function Call	Three panes shows a table of functions called, a graph of Top 5 calls and properties of the selected call.
Analysis Summary	Shows data for the current frame
Analysis Detail	Shows analysis detail for the current frame
Program	Shows program counters and their value



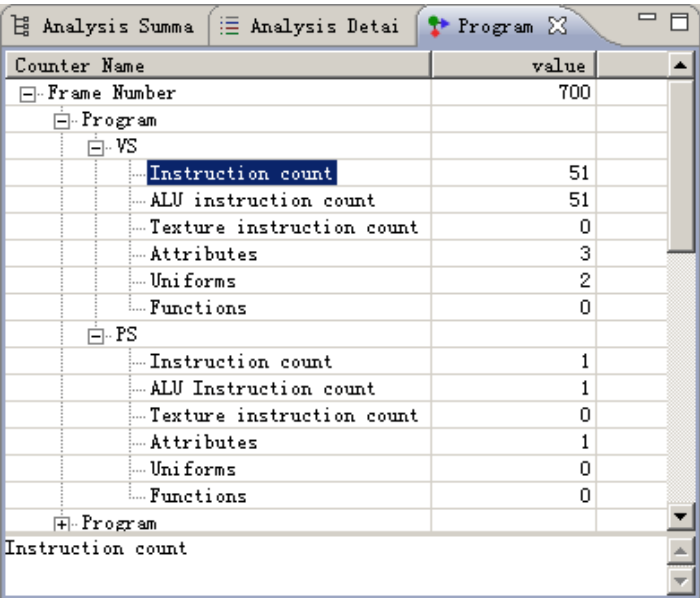
14.6.3 System Info View

The left most System Info tab shows the system information related to the VPD data under analysis, such as hardware, driver and vProfiler versions. The Average Frame Rate is also reported on this tab.




14.6.4 Program Counters View

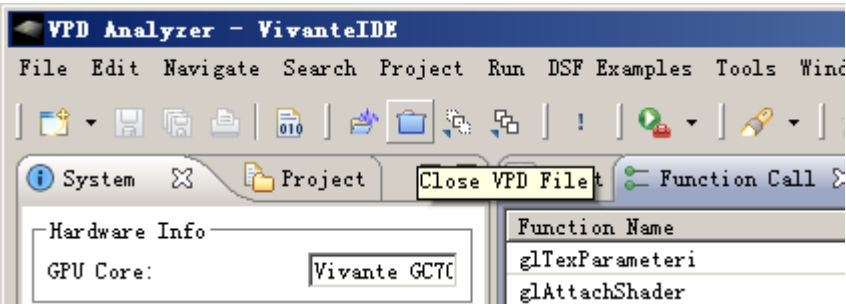
The rightmost tab in the rightmost pane is the Program tab which shows program counter information, such as Instruction counts and attribute counts.



Counter Name	value
[-] Frame Number	700
[-] Program	
[-] VS	
Instruction count	51
ALU instruction count	51
Texture instruction count	0
Attributes	3
Uniforms	2
Functions	0
[-] PS	
Instruction count	1
ALU Instruction count	1
Texture instruction count	0
Attributes	1
Uniforms	0
Functions	0
[+] Program	
Instruction count	

14.6.5 Closing the VPD File

Click the icon  from the toolbar or select **Tools->VPD Analyzer->Close VPD File** to close the current VPD file. The analysis data associated with the closed file will be cleared from all views.



14.7 SPIR-V Disassembler

A SPIR-V Disassembler tool is provided as an aid in debugging Vulkan applications. If a SPIR_V file is already located in a project, simply double click on it to disassemble. Otherwise use the main menu **File -> Open File...** to locate the SPIR-V. Options can be set via the **Window->Preferences** dialog box.

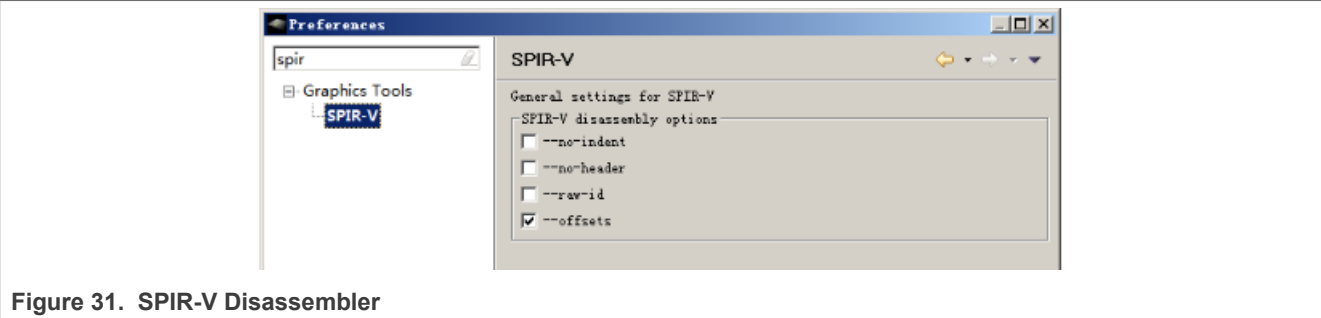


Figure 31. SPIR-V Disassembler

14.7.1 Shader Assistant

Shader Assistant perspective is provided for Shader program development for OpenGL, OpenCL and Vulkan projects. Shader Assistant provides an environment for editing, previewing, analyzing, and optimizing shader programs. Shader Assistant includes samples of shader programs, a number of standard meshes (sphere, cube, tea pot, pyramid, etc.) and a text editor. These extra features will help programmers get a quick start on creating their shader programs.

There are two ways to switch to the Shader Assistant perspective view. From the main menu, choose Window -> Open Perspective -> Shader Assistant, or in the C/C++ **Project Explorer** pane, right click and select **Develop Shader**. Using the table in the left pane **Preview Settings** tab, select items in the *Setting* column and configure project as well as header, shaders, attributes, etc.

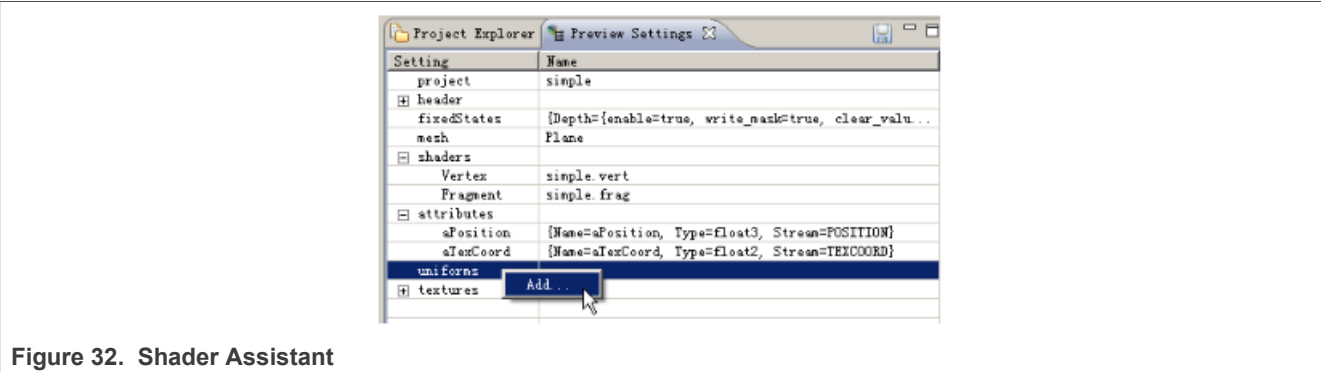


Figure 32. Shader Assistant

14.7.2 vTexture

Texture manipulation and viewing is available in four different areas of VivanteIDE:

- **Texture Editor** dialog boxes accessible from the Shader Assistant Preview Settings tab provides for texture customization, q.v. preceding Section 13.7.1 for launching Shader Assistant.
- **vTexture** Browser and Viewer panes are available from the main menu **Window -> Open Perspective -> VTexture**. It provides thumbnail and detail view of textures as well as the basic properties of the textures, such as image size and color depth.

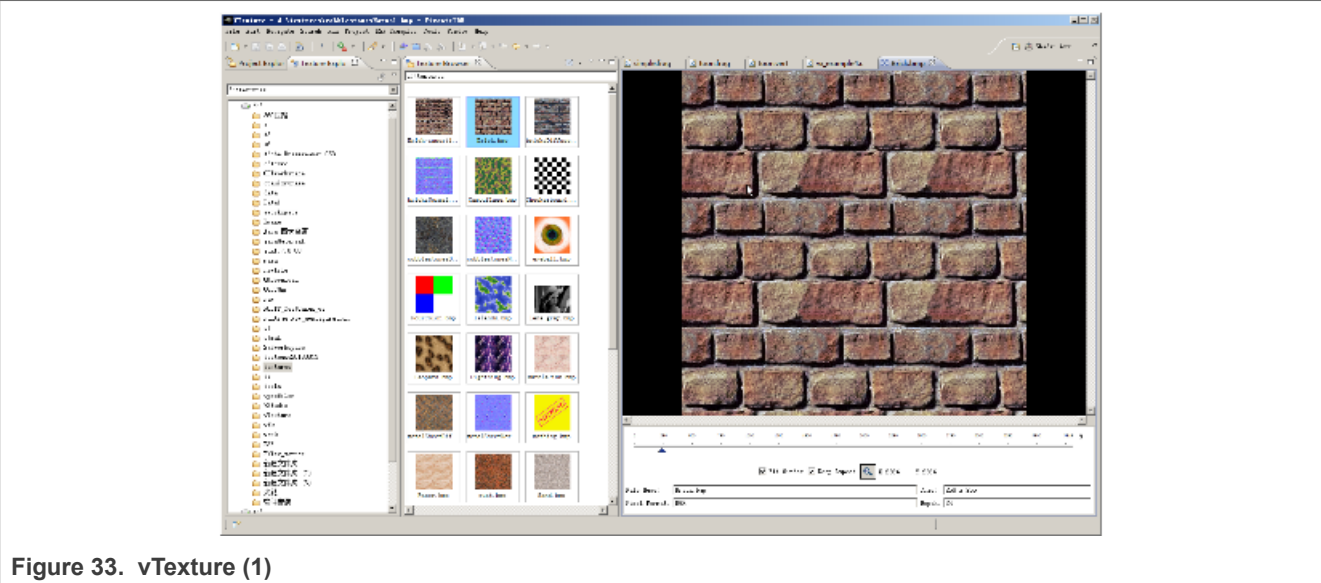


Figure 33. vTexture (1)

- Convert Texture provides a GUI for texture compression/decompression and tiling/de-tiling. It is accessible by clicking on the main menu **Tools->Convert Texture**. Note that **vTextureTools** is the command line tool version of this tool. Refer to Section 13.8.4 for details.

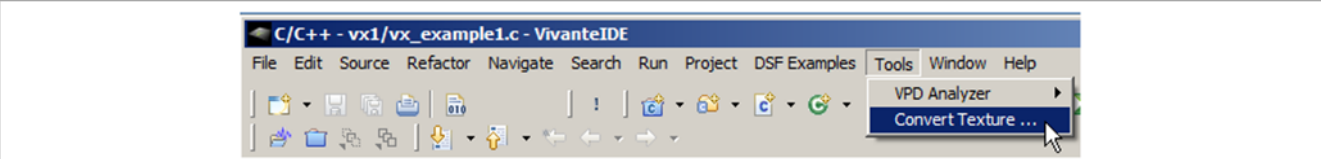


Figure 34. vTexture (2)

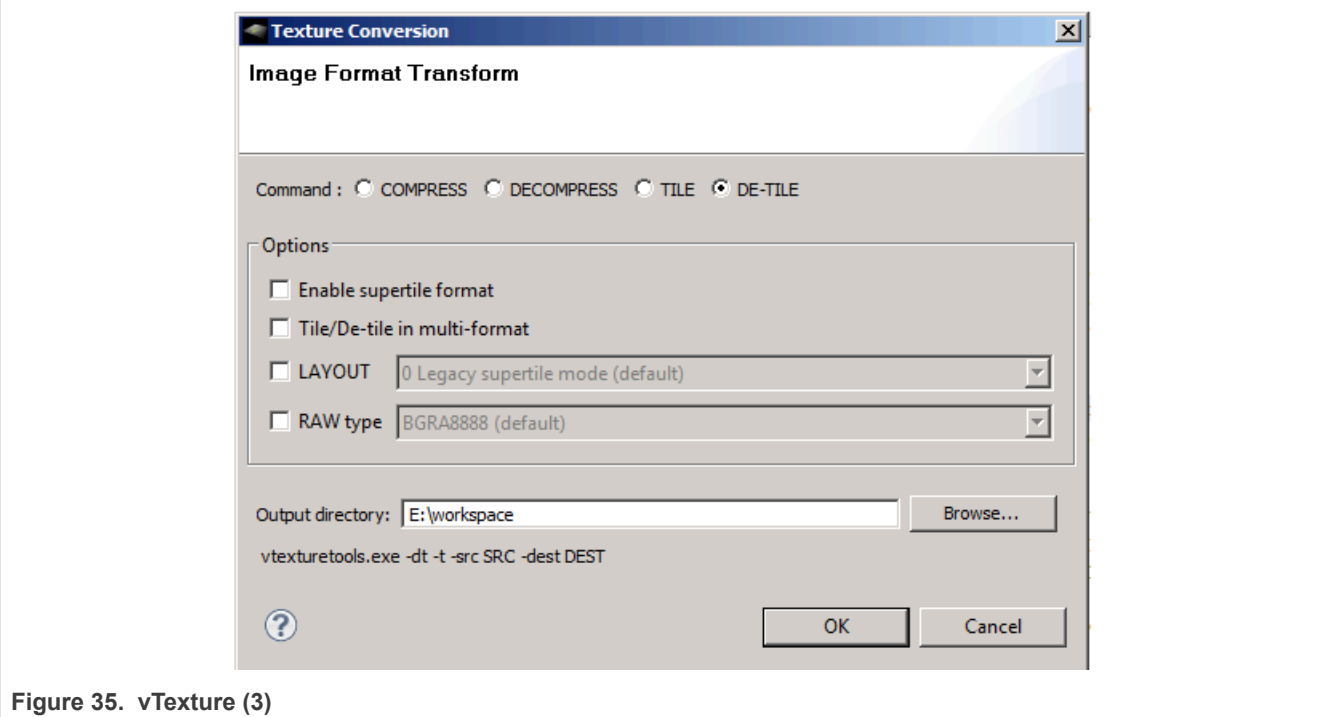


Figure 35. vTexture (3)

14.8 VivanteIDE command line tools

For easy reference, the syntax for the VivanteIDE command line tools are provided on the following pages. You can also refer to the *VivanteIDE User Guide* or inline `-h` (help) for syntax for these command line tools.

14.8.1 Preparing the environment

Before running command line tools, prepare the environment as in the examples below.

For Linux OS

- Launch a BASH
- `$ source installation_dir/ide/setenv-vivanteide<version> # initialize the environment`

For Windows OS

- Launch a Command Shell
- `> installation_dir/ide/setenv-vivanteide<version>.bat # initialize the environment`

14.8.2 vCompiler Command Line Syntax for OGL and OGLES

Open a Command prompt. Navigate to the folder, which contains the vTextureTools files (for example, `installation_dir/cmdtools/vCompiler`, and launch the `vCompiler` application executable using the command line syntax described below.

Make sure the configuration file is customized for your target environment.

14.8.2.1 Syntax

Windows and Linux command line syntax is the same.

Optional inputs are indicated by brackets. A fixed order for options in the command is not required.

```
vCompiler [-f <gpuConfigurationFile>] <shaderInputFileName>
[shaderInputFileName_2]
[ -c ] [ -h ] [ -l ] [ -o <outputFileName> ] [ -On ] [ -v ] [ -x <shaderType> ]
```

14.8.2.2 Input parameters (required)

shaderInoutFileName	shader input file name, which must contain one of the following file extensions: <ul style="list-style-type: none">• vert: vertex shader source file• frag: fragment shader source file• vgcSL: previously compiled vertex shader input/output file• pgcSL: previously compiled pixel shader input/output file
---------------------	--

14.8.2.3 Input parameters (optional)

shaderInputFileName_2	Up to two shader files can be specified. The second shader file is optional but must have one of the file extensions described above for shader InputFileName. If the first shader is a vertex shader, this second shader should be a fragment shader; conversely if the first shader is a fragment shader, the second should be a vertex shader.
-----------------------	---

	<p>Note: Pre-compiled and compiled shaders may be mixed, as long as one is a vertex shader and the other a fragment shader.</p>
-c	<p>Compile each vertex .vert file into a vgcSL file and/or fragment shader .frag file into a pgcSL only, with no merged result file of type .gcPGM.</p> <p>If the -c option is not specified:</p> <ul style="list-style-type: none"> • When only one shader is specified, that shader will be compiled into a [v/p]gcSL file. • When two shaders are specified, one is assumed to be a vertex shader and the other a fragment shader. Each shader can be either a previously compiled .vgcSL or .pgcSL file or a .vert or .frag still to be compiled. The two will be merged into a .gcPGM file after successful compilation.
-f <gpuConfigurationFile>	<p>Specifies a configuration file (from VTK 1.6.2). If -f is not specified, the file viv_gpu.config in the vCompiler working directory will be used as the default configuration file. Example syntax:</p> <pre>vCompiler -f viv_gpu_880.config foo.vert bar.frag</pre> <p>Note: vCompiler will not work correctly if the GPU configuration file cannot be found or contains incorrect content.</p>
-h	Shows a help message on all the command options.
-l	<p>Create a log file. The log file name is created by taking the first input file name, then replacing its file extension with ".log". If the input file name does not have a file extension, .log is appended, e.g.,</p> <pre>myvert.vert => myvert.log inputfrag => inputfrag.log</pre>
-o <outputFileName>	<p>Specify the output file name. If the path is other than the current directory, it must also be specified. Any extension can be specified. If the extension is not specified, the outputFileName supported default types are as follows:</p> <ul style="list-style-type: none"> • vgcSL: compiled vertex shader output file, usually compiled from a .vert input source file (default result for single file compile) • pgcSL: compiled pixel shader output file, usually compiled from a .frag source input file. • gcPGM: compiled file merging vertex shader and fragment/pixel shader into a single output file
-O<n>	<p>Optimization level. Default is -O2:</p> <ul style="list-style-type: none"> • -O0: Disable optimizations • -O1: Some optimizations are enabled. • -O2: All optimization levels are on (default).
-v	Verbose; prints compiler version and diagnostic messages to STDOUT.

-x<shaderType>	Explicitly specifies the type of shader instead of relying on the file extension. This option applies to all following input files until the next -x option. ShaderType: supported values for Shader type include: <ul style="list-style-type: none">• vert: vertex shader source file• frag: fragment shader source file• vgcSL: compiled vertex shader input/output file• pgcSL: compiled pixel shader input/output file
-x none	Revert back to recognizing shader type according to the file name extension.

14.8.2.4 vCompilerOutput

Output files are placed in the current directory, unless another directory is specified with the -o option. The files can be of the three types described above under outputFileFileName value of the -o option.

14.8.2.5 vCompiler Syntax examples

```
vcCompiler foo.vert produces foo. vgcSL.  
vcCompiler bar.frag produces bar.pgcSL.  
vcCompiler foo.vert bar.frag produces foo.gcPGM.  
vcCompiler -v -l -O1 foo.ver tbar.frag produces foo.gcPGM and foo.log.  
vcCompiler -v -l -O1 -o foo_bar foo.vert bar.frag produces foo_bar.gcPGM and  
foo_bar.log.
```

14.8.3 vcCompiler Command Line Syntax for OCL

Open a Command prompt. Navigate to the folder which contains the vTextureTools files (for example, installation_dir/cmdtools/vCompiler, and launch the vcCompiler application executable using the command line syntax described below.

Make sure the configuration file is customized for your target environment.

14.8.3.1 Syntax

Windows and Linux command line syntax is the same.
Optional inputs are indicated by brackets. A fixed order for options in the command is not required.

```
vcCompiler [-f <gpuConfigurationFile>] [-v] [-l] [-O0] [-D <MacroDefinition>] [-  
I <IncludeDirectory>]  
[-K <KernelName>] [-M] [-B] <OpenCLorOpenVXFileName>  
<OpenCLorOpenVXFileName_2> . . . [-allkernel]
```

14.8.3.2 Input parameters (required)

OpenCLorOpenVXFileName	Input file name, which must contain one of the following file extensions: <ul style="list-style-type: none">• cl: OpenCL source file
------------------------	---

- vx: OpenVX Vision source file
If an input file extension is not specified, vcCompiler will report a “wrong file extension” error.

14.8.3.3 Input parameters (optional)

OpenCLOrOpenVXFileName_2, _n	Multiple input files can be specified. The second and additional files are optional but must have the appropriate file extension as described above. All files must be of the same type (.cl or .vx).
-allkernel	Allows VX applications to create all kernels in one program and save them into one package.
-B	Support source level intrinsic built-in functions.
-D <MacroDefinition>	Predefined inline macro, as referenced in the input file.
-f <gpuConfigurationFile>	Specifies a configuration file. If -f is not specified, the file viv_gpu.config in the vcCompiler working directory will be used as the default configuration file. Syntax example: <pre>vcCompiler -f viv_gpu_gc7000.config foo.cl</pre> <p>Note: vcCompiler will not work correctly if the GPU configuration file cannot be found or contains incorrect content.</p>
-h	Shows a help message on all the command options.
-I <IncludeDirectory>	Specify the directory path for include files.
-K <KernelName>	Link with kernel name. Default is main .
-l	Create a log file. The log file name is created by taking the input file name, then replacing its file extension with “.log”. If there are multiple input files, the filename of the first input file will be used, <pre>inputcl.cl => inputcl.log myvx1.vx myvx2.vx => myvx1.log</pre>
-M	Merge all compiled output from each file into one file. The combined output will have the name of the last input file combined with the output extension .gcPGM.
-O<n>	Optimization level. Default is -O2 : <ul style="list-style-type: none"> • -O0: Disable optimizations • -O1: Some optimizations are enabled. • -O2 All optimization levels are on (default).

-v	Verbose; prints compiler version and diagnostic messages to STDOUT
----	--

14.8.3.4 vcCompiler Output

Output files are placed in the current directory. When compiled successfully, the supported output file extensions for vcCompiler are:

- .clgcSL: compiled CL output file, compiled from a .cl input source file.
- .vxgcSL: compiled VX output file, compiled from a .vx input source file.

14.8.3.5 vcCompiler Syntax Examples

```
vcCompiler [-f <gpuConfigurationFile>] [-v] [-l] [-O0] [-D <MacroDefinition>] [-I <IncludeDirectory>] [-K <KernelName>] [-M] [-B] <OpenCLorOpenVXFileName> <OpenCLorOpenVXFileName_2> [-allkernel] . . .
```

vcCompiler -v -O1 foo.cl: produces foo.clgcSL.
vcCompiler -v -l foo.vx: produces foo.vxgcSL and foo.log.

14.8.4 vTextureTools command line tool

Open a Command prompt. Navigate to the folder which contains the vTextureTools files, for example, installation_dir/cmdtools/vTextureTools, and launch the vTextureTools application executable using the command line syntax described below.

14.8.4.1 Syntax

The usage of the command line tool is as follows for compression/decompression:

```
vTextureTools -c TYPE [-s SPEED] -src FILE [-dest FILE]
```

or

```
vTextureTools -d TYPE -src FILE [-dest FILE]
```

The usage of the command line tool is as follows for tiling/de-tiling:

```
vTextureTools -t|-st [-2] [-r|--raw=FORMAT] [-m LAYOUT] -src FILE [-dest FILE]
```

or

```
vTextureTools -dt -t|-st [-2] [-r|--raw=FORMAT] [-m LAYOUT] -src FILE [-dest FILE]
```

14.8.4.2 General parameters

General parameters:

- **-h** show help
- **-src [FILE]** source file - input image path and filename. vTexture will use the file extension type as image type.
 - For option **-c** compress, the application expects an input filename with a .TGA extension.
 - For **-d** decompression, the application expects .DDS, .KTX or .PKM.
 - For **-t** tile, the application expects .BMP or .TGA.
 - For **-dt** detile, the application expects .BMP or .TGA.
- **-dest [FILE]** destination file - image path and filename.
 - The application expects a filename with a .TGA, .DDS, .KTX or .PKM extension for compress/uncompress or .BMP or .RAW for tile/detile.
 - If the **-dest** parameter is not set, vTexture will auto generate a name for the newly generated file, using the source file name as the prefix appending critical parameters and file type information.

14.8.4.3 Compression/Decompression parameters

These parameters are used for compression and decompression:

- **-c** compress a source image of format uncompressed TGA
- **[TYPE]** specify the target output compression format:
 - **-DXT1** compress image to DXT1 format (default format).
 - **-DXT3** compress image to DXT3 format.
 - **-DXT5** compress image to DXT5 format.
 - **-ETC1** compress image to ETC1 format
 - **-ETC2** compress image to ETC2 format
- **-d** decompress a source image of format specified by the value **[TYPE]**.
The resulting file type will be uncompressed TGA.
This option decompresses DXT1, DXT3, DXT5, ETC1 or ETC2 format image to TGA format.
- **-s** compression **[SPEED]** mode for ETCn images:
 - **slow**
 - **medium**
 - **fast** (default)

14.8.4.4 Tile/De-Tile parameters

The parameters listed in the following table are used for tiling and de-tiling between linear and tiled formats.

Table 44. Tile/De-Tile parameters

-t	Convert linear data to tiled texture output.
-st	Enable supertile format. This option is an alternate to -t . If -st and -t are used together, -st will be set.
-dt	De-tile: Convert tiled texture to linear texture output.
-2	Tile/de-tile in multi-format. Tile format is multi-tiled (when used with -t) or multi-supertiled (with -st).
-m	[LAYOUT] : layout mode for supertiled or multi-supertiled textures: <ul style="list-style-type: none"> • 0: Legacy supertile mode (default). • 1: Supertile mode when hardware has HZ. • 2: Supertile mode when hardware has NEW_HZ or FAST_MSAA.

Table 44. Tile/De-Tile parameters...continued

-r	Specify output data as raw pixel output instead of BMP. Use --raw=rgb565 to specify raw pixel [FORMAT]. Supported raw formats (8) are:
	rgba8888, bgra8888, rgb888, bgr888, rgb565, bgr565, argb1555, yuy2

14.8.4.5 vTexture Syntax Examples

COMPRESS:

```
vTextureTools -c dxt1 -src d:\myfile.png -dest c:\compress.dds
vTextureTools -c dxt1 -src d:\myfile.tga -dest c:\compress.dds
vTextureTools -c etc1 -s slow -src d:\myfile.png -dest c:\compress.pkm
vTextureTools -c etc1 -s slow -src d:\myfile.tga -dest c:\compress.pkm
vTextureTools -c etc2 -s slow -src d:\myfile.bmp -dest c:\compress.ktx
vTextureTools -c etc2 -s slow -src d:\myfile.tga -dest c:\compress.ktx
vTextureTools -c etc2 -src d:\myfile.bmp -dest c:\compress.ktx
vTextureTools -c etc2 -src d:\myfile.tga -dest c:\compress.ktx
vTextureTools -c etc2 -src d:\myfile.tga -dest c:\compress.pkm
```

DECOMPRESS:

```
vTextureTools -d etc1 -src c:/vtexin/myfile2.pkm -dest c:/vtextout/myfile2.tga
vTextureTools -d -src c:/vtexin/myfile3.dds -dest c:/vtextout/myfile3.tga
  (assumes DXT1)
vTextureTools -d tga -src d:\myfile.dds -dest c:\decompress.tga
vTextureTools -d tga -src d:\myfile.ktx -dest c:\decompress.tga
```

TILE: LINEAR TO TILE CONVERSION:

- Tile linear texture to standard tile texturev

```
TextureTools.exe -t -src 123.bmp
```

- Tile linear texture to multi-tiled texture

```
vTextureTools.exe -t -2 -src 123.bmp
```

- Tile linear texture to supertiled texture

```
vTextureTools.exe -st -src 123.bmp
```

- Tile linear texture to multi-supertiled texture

```
vTextureTools.exe -2 -st -src 123.bmp
```

- Tile linear texture to multi-supertiled texture and output rgb565

```
vTextureTools.exe -2 --raw=rgb565 -src 123.bmp
```

- Tile linear texture to multi-supertiled texture with layout mode 2

```
vTextureTools.exe -st -2 -m 2 -src 123.bmp
```

DE-TILE: TILED TO LINEAR CONVERSION:

- De-tile tiled texture to linear texture

```
vTextureTools.exe -dt -t -src 123-tiled.bmp
```

- De-tile supertiled texture to linear texture

```
vTextureTools.exe -dt -st -src 123-supertiled.bmp
```

- De-tile multi-supertiled texture to linear texture

```
vTextureTools.exe -dt -t -2 -src 123-tiled-multi-tiled.bmp
```

- De-tile multi-Super-tiled texture with layout mode 2 to linear texture

```
vTextureTools.exe -dt -st -2 -m 2 -src 123-multi-supertiled-2.bmp
```

15 GPU Tools

Note: All SoCs support this tool if not specified.

15.1 gputop tool

`gputop` monitors the GPU clients memory, hardware counters, occupancy state load on DMA engines, video memory, and and DDR memory bandwidth (only under Linux OS).

- The `gputop` tool is developed to trace the overall memory utilization in classification of memory pools.
- The available memory size is reported for the reserved pool.
- GPU idle time is reported from the last capture.

15.1.1 Synopsis

- `gputop [options]`
- `gputop -m [mode]`: Where the mode can be: `mem`, `counter_1`, `counter_2`, `occupancy`, `dma`, `vidmem`, and `ddr` (under Linux/Android). Use this option to start `gputop` directly in a mode as required. For `counter_1` and `counter_2`, a context is needed. See [Section 15.1.5](#) for why this is necessary.
- `gputop -c ctx_no`: Specifies a context to attach when display context-aware hardware counters.
- `gputop -b`: Displays in batch mode. For other modes than memory, this only takes an instantaneous sample. See `-f`.
- `gputop -f`: Use it when using `gputop` from a script.
- `gputop -x`: Useful to display contexts when used with `-b`.
- `gputop -i`: Ignores warnings about kernel mismatch.
- `gputop -h`: Displays usage and help.

Note:

Unsupported command options for i.MX 95: `gputop -m [mode]`, `gputop -c ctx_no`, and `gputop -b`.

15.1.2 Interactive mode

Normally, when starting up, `gputop` starts in interactive mode. The following are a list of useful commands:

- `h`: Displays the help page.
- `0-6/Left-Right` arrows: Switches between viewing pages.
- `x`: Displays application contexts.
- `SPACE`: Selects a context that you want to track. Useful for reading `counter_1` and `counter_2` values.
- `r`: Useful for hardware-counter pages to display different viewing modes (switches between different modes of aggregation: `MIN/MAX/AVERAGE/TIME`).
- `q/ESC`: Exits `gputop`.

- `p`: Stops reading counter values and displays only the current values. Useful to get a instantaneous values of the counters.

Note: *Unsupported command option for i.MX 95: `r`, `p`.*

15.1.3 Description

`gputop` can be used to determine the memory usage your application is using, or to read the hardware counters exposed by the GPU in real-time. Additionally, DMA engines and Occupancy states are displayed. `gputop` has multiple viewing pages: a memory usage page, two hardware counter pages, a DMA engine page, and an Occupancy page. When normally started, `gputop` is in interactive mode. Type `h` to get a list of the current keybindings.

15.1.4 Requirements

15.1.4.1 Linux OS

`gputop` requires access to `debugfs` sub-system on Linux OS to display the memory usage, used by clients submitting commands to the GPU. `gputop` tries to mount the `debugfs` pseudo-filesystem if it is not already mounted. To read hardware counters, the profiler must be activated in the driver. Usually this can be set by setting the environment variable `export VIV_PROFILE=1`.

15.1.4.2 QNX

Just like in Linux OS, to read the hardware counter values, `gpu-gpuProfiler` should be set to **1** in the `graphics.conf` file under the `$GRAPHICS_ROOT` directory. Other views like occupancy and DMA require `gpu-powerManagement` to be set to **0** (disabled).

15.1.5 Notes

15.1.5.1 Sampling hardware-counters

GPUPop samples the driver for hardware counter values. Internally the driver updates the values of the counters whenever the application submits a special type of command to the GPU. Depending on how fast that happens, GPUPop cannot foresee/adjust the values of the counters. Therefore, tweaking the amount of sample taken or the delay time does not really help. For dealing with situations where the application submits either too fast or too low commands to the GPU, several modes of viewing counters have been added. Cycle between them to understand or get a bird-eye view of the counter values. Empirically MAX/AVERAGE displays the closest values to the truth.

15.1.5.2 Context-aware counters

`counter_1` and `counter_2` are context-aware counters (for example, tied to an application). This is not supported on the i.MX 95.

Internally, the driver assigns various context IDs to the application submitting commands to the GPU. These contexts IDs are currently required to read those hardware counter values. Either use `-x` on the command line (together with `-b` option and choosing `-m mem` viewing mode), or for interactive mode, use `x` and then `SPACE` to show and select a context ID.

If you are getting zero'ed out values for `counter_1` and/or `counter_2` values, cycle through the available counter IDs.

Due to the way the driver is built, single-GPU core applications have two context-ids. Empirically the largest integer values holds the real context ID.

15.1.5.3 Unsupported GPUs

For GCV600 (i.MX 7ULP, i.MX 8M Mini, and i.MX 95), the IDLE/LOAD register is not available, so `gputop` displays incorrect (inversed) values.

15.1.6 Pages for VSI GPUs

15.1.6.1 Client attached page

When viewing the client attached page, the following head columns are displayed:

PID	RES	(KB)	CONT	(KB)	VIRT	(KB)	Non-PGD	(KB)	Total	(KB)	CMD
-----	-----	------	------	------	------	------	---------	------	-------	------	-----

- PID: process ID
- RES: reserved memory
- CONT: contiguous memory
- VIRT: virtual memory
- Non-PGD: Non-paged memory
- Total: sum of all above
- CMD: name of the application (trimmed)

These memory items correspond to memory pools in the driver.

15.1.6.2 Vidmem page

When viewing vidmem page, the following head columns are displayed for each process.

PID	IN	VE	TE	RT	DE	BM	TS	IM	MA	SC	HZ	IC	TD	FE	TFB
-----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----

- IN: index
- VE: vertex
- TE: texture
- RT: render target
- DE: depth
- BM: bitmap
- TS: tile status
- IM: image
- MA: mask
- SC: scissor
- HZ: hz
- IC: i_cache
- TD: tx_desc
- FE: fence
- TFB: tfb header

15.1.7 Pages for Mali GPU

GPUTop for i.MX 95 tool has total 6 pages.

0-6/Left-Right arrows: Switches between viewing pages.

15.1.7.1 Page0: Main Page

This page is a summary of GPUTop information, including GPU DDR memory read/write bandwidth in mega bytes based on Linux perf, GPU kernel and user space memory usage summary, and GPU utilization.

15.1.7.2 Page1: GPU INFO

This page lists the GPU kernel, GPU drive architecture information, and performance statistics.

15.1.7.3 Page2: Kernel Memory Usage

This page is memory consumption from kernel mode perspective. This page lists total kernel allocated memory and memories allocated based on each kernel context.

Note: From the kernel displayed in this page, there is always 7 page memory (28 KB) difference between the sum of the kernel memory for all kernel contexts and the total GPU Mali memory. This fixed 7 page is for internal kbase shared device usage.

15.1.7.4 Page3: PID-Based Process Memory Usage

Starting from GPU Mali DDK release r52, the memory profile has been changed from the class memory type to simple memory profile. The content of this page has been changed as follows since Mali DDK release r52:

This page lists the user space GPU memory allocation and freeing by using MALI CMEM allocators. It also provides a breakdown of the user space memory usage.

When viewing the client main page, the following head columns are displayed:

PID	Total (kB)	Explicit (kB)	Committed (kB)	UnCommitted (kB)	Implicit (kB)
Imported (kB)	CMD				

- PID: Process ID
- Total: Total allocated memory in user space
- Explicit: Total virtual address space that has been allocated explicitly
- Committed: Amount of memory that is physically backed for explicitly allocated memory
- Uncommitted: Amount of memory that has been allocated explicitly but not yet committed
- Implicit: Memory that is committed on the GPU page fault
- Imported: Memory allocated in other processes and imported in this process
- CMD: Name of the application

This is the example for this page:

```
PID 562 /usr/bin/weston
user space Mem: 60 kB total
user space Explicit VA Mem: 60 kB total
user space Implicit committed Mem: 0 kB total
user space Imported Mem: 0 kB total
Explicitly Committed GPU Memory:
- Allocated VA: 61440
- Committed: 61440
- Wasted for headers: 1184
- Wasted for footers: 2716
- Unused: 17160
- Used: 40380
- Usage Breakdown:
  - Base Internal: 4096
```

```
- Program: 27096
- Device Internal: 9188
Implicitly Committed GPU Memory:
- Usage Breakdown:
Imported GPU Memory:
- Usage Breakdown:
```

Before Mali GPU release r52, this page is memory usage statistics from user-mode driver perspective. It lists the detailed memory consumed including memory mapped to kernel space and memory of each class type for every application.

When viewing the client attached page, the following head columns are displayed:

PID	Class_MEM(KB)	Allocated_GPU_MEM(KB)	CMD
-----	---------------	-----------------------	-----

- PID: Process ID
- Class_MEM: All types of class memory allocated for this process
- Allocated_GPU_MEM: All kernel memory allocated for this process
- CMD: Name of the application

15.1.7.5 Page4: GPU Core Utilization

This page lists the GPU last render frequency, total GPU utilization, protect mode utilization, GPU fragment shader, GPU no-shader utilization, and GPU tiler utilization.

15.1.7.6 Page5: Perf DDR Memory Bandwidth

This page lists the GPU DDR read/write memory bandwidth in mega bytes.

15.1.8 Examples

When using -b option, gputop starts in interactive mode and executes its main loop only once. This is useful for various reasons, either to get an instantaneous view of a different viewing page or scripting.

- Get a list of processes attached to the GPU.

```
$ gputop -m mem -b
```

- Get a list of processes attached to the GPU, but also display the contexts IDs.

```
$ gputop -m mem -bx
```

- Display counters (counter_1) using context_id.

```
$ gputop -m counter_1 -b -c <context_id>
```

- Display counters (counter_2) using context_id.

```
$ gputop -m counter_2 -b -c <context_id>
```

- Get IDLE/USAGE

```
$ gputop -m occupancy -b | grep IDLE
```

15.1.9 See Also

- Under QNX, see graphics.conf for disabling powerManagement and enabling gpuProfiler.

- Under Linux, see `/sys/module/galcore/parameters/powerManagement`.

15.2 GPU clock information and debugging

GPU driver supports dynamic frequency scaling. Users can perform the following steps to query and update the GPU clock information, which is useful for GPU debugging.

1. Get the GPU clock. This is affected by the system RTC timer. Sometimes it varies between different boards.

```
root@imx8mpevk:/# mount -t debugfs none /sys/kernel/debug (optional, exec it
only if there is no gc dir)
root@imx8mpevk:/# cat /sys/kernel/debug/gc/clock
gpu0 mc clock: 1000018036 HZ.
gpu0 sh clock: 1000021374 HZ.
gpu1 mc clock: 1000002214 HZ.
gpu1 sh clock: 999986723 HZ.
gpu8 mc clock: 499991523 HZ.
```

2. Change the GPU clock.

Read the `gpu3DClockScale` as the denominator using the following command:

```
root@imx8mpevk:/# cat /sys/bus/platform/drivers/galcore/gpu3DClockScale
64
```

The GPU frequency can be changed to `numerator/gpu3DClockScale * clock` for different GPU instances. For example, the `gpu0`'s mc and sh clock can be change to 1/2 and 1/4 of the original frequency.

```
root@imx8mpevk:/# echo 0 32 16 > /sys/kernel/debug/gc/clock
[ 2625.977856] Change core:0 MC scale:32 SH scale:16
[ 2625.982610] Warning: Power management status will be changed forever!
root@imx8mpevk:/# cat /sys/kernel/debug/gc/clock
gpu0 mc clock: 499997481 HZ.
gpu0 sh clock: 249997541 HZ.
gpu1 mc clock: 999995540 HZ.
gpu1 sh clock: 999992141 HZ.
gpu8 mc clock: 499998453 HZ.
```

15.3 Apitrace user guide

15.3.1 Introduction

Apitrace is a set of tools enhanced from open source project apitrace, supported by i.MX 6, i.MX 7, and i.MX 8 with Vivante GPU IP. This tool can dump OpenGL/GLES1.1/GLES2.0/GLES3.0 API calls and replay on a wide range of other devices.

For more information, see apitrace.github.io/.

15.3.2 Install

15.3.2.1 Yocto

Apitrace source code release is part of the i.MX Yocto Project Linux BSP release. The source code have more patches added on top of official Apitrace release. The Yocto Project recipes pull the Apitrace source package and install as needed for supported backend.

15.3.2.2 PC

Apitrace have set of PC tools. Prebuilt binary packages can be directly downloaded from the Apitrace website. Currently supports Ubuntu 14.04 LTS, 64-bit.

```
sudo apt-get install libgles1-mesa libgles2-mesa libqt4-dev
```

15.3.3 Usage

15.3.3.1 Trace OpenGL ES1.1/2.0/3.0 application

```
apitrace trace --api=egl <app name and arguments>
```

e.g., `apitrace trace --api=egl es2gears_x11`

It generates trace file (`.trace`) under the current directory. To specify a new path, use `--output=<path_name>`.

15.3.3.2 Trace OpenGL ES 1.1/2.0/3.0 Java application on the Android platform

On the Android platform, a GLES application can be native (e.g., `frameworks/native/opengl/angeles`). This type of application can be traced as normal Linux application. Some other applications involving the Java virtual machine cannot run in this way. A script `apitrace_dalvik.sh` is provided to run this type of application. This is an example to trace `com.android.settings`:

```
sh /data/apitrace/bin/apitrace_dalvik.sh com.android.settings start
```

To stop tracing, run:

```
sh /data/apitrace/bin/apitrace_dalvik.sh com.android.settings stop
```

Because there is no “current” directory for a Java application, the trace file is stored under `/sdcard/`.

If Apitrace is installed in a different directory, update `apitrace_dalvik.sh` manually.

15.3.3.3 Trace OpenGL application

```
apitrace trace --api=glx <app name and arguments>
```

Only the X11 backend supports this feature.

15.3.3.4 Replay

This utility is also called `retrace`. It reads in the trace file and executes OpenGL (ES) APIs one by one. Each OpenGL (ES) API call is processed by a callback function. In that callback function, a hook can be inserted for debug or analysis purposes.

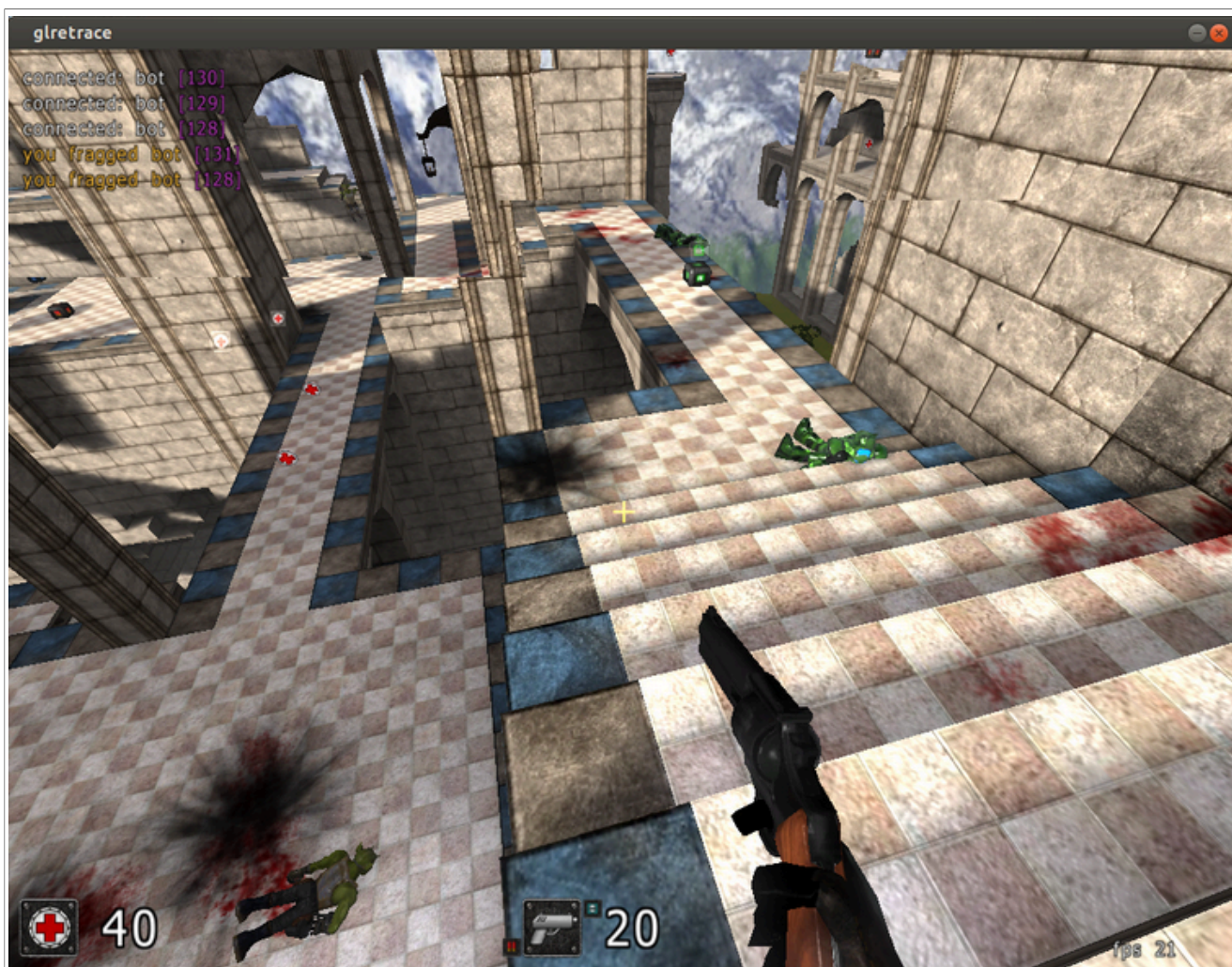


Figure 36. Replay

OpenGL ES 1.1/2.0/3.0 applications can be replayed with `eglretrace`; OpenGL applications can be replayed with **`glretrace`**:

```
eglretrace <trace file>
glretrace <trace file>
```

15.3.3.4.1 Analysis

`qapitrace` provides a detailed look at the trace file. It can only run on a PC. It was verified on Ubuntu 14.04 LTS 64-bit. The command is:

```
qapitrace <trace file name>
```

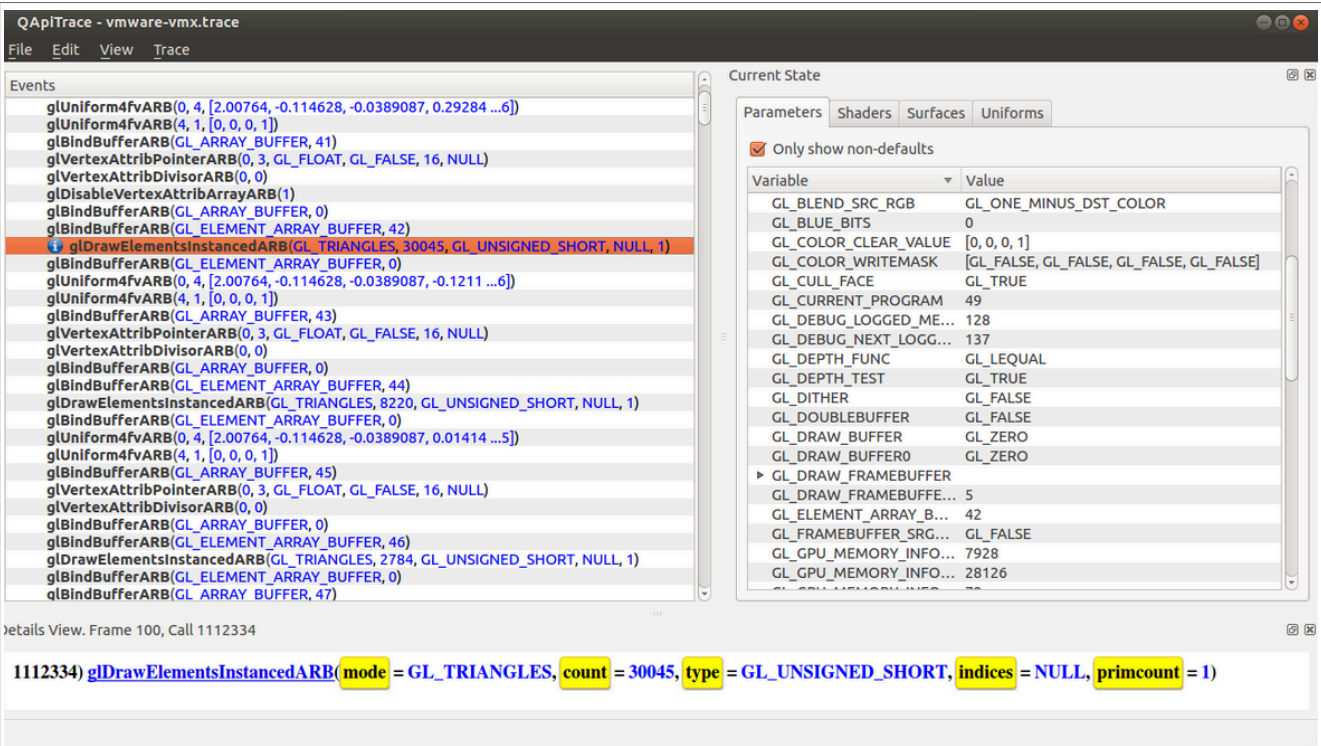


Figure 37. Checking state of every API call

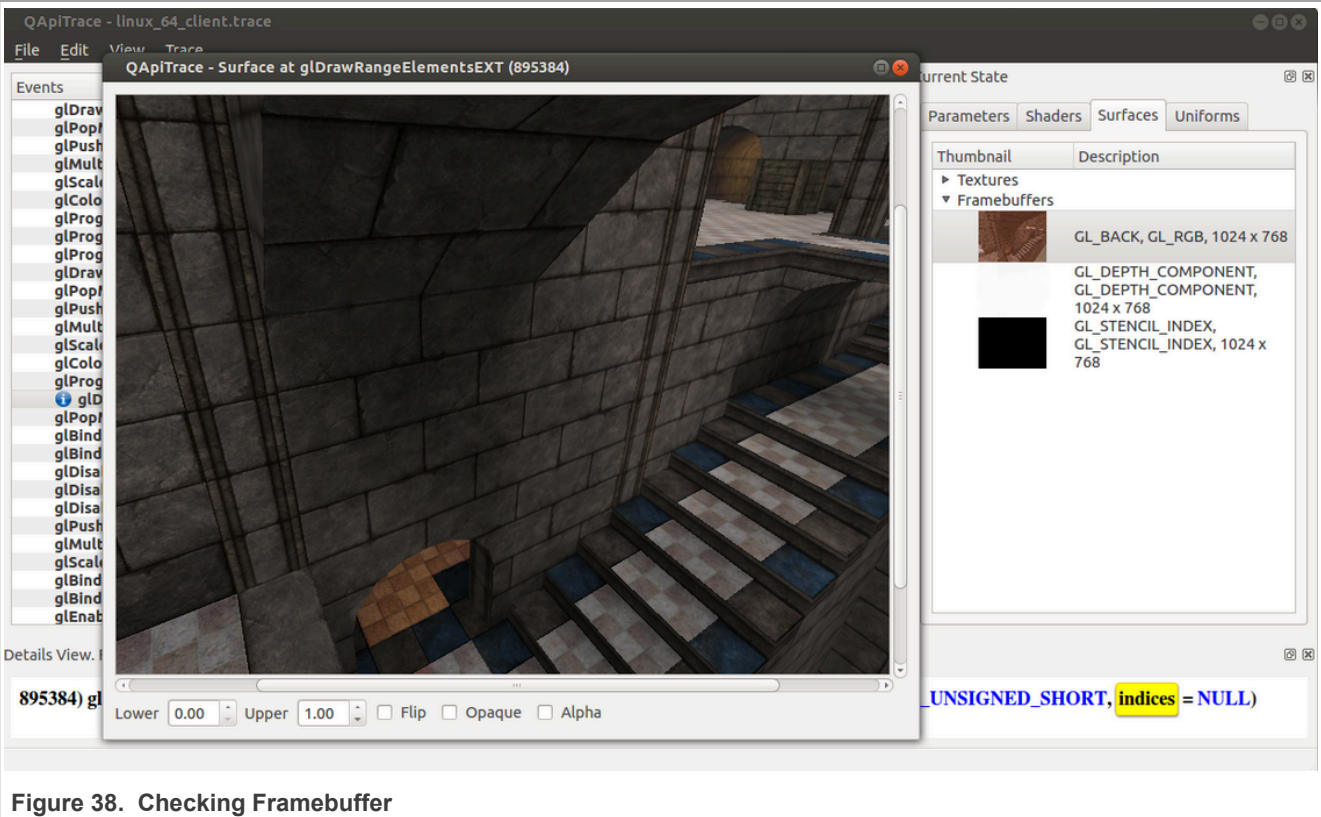


Figure 38. Checking Framebuffer

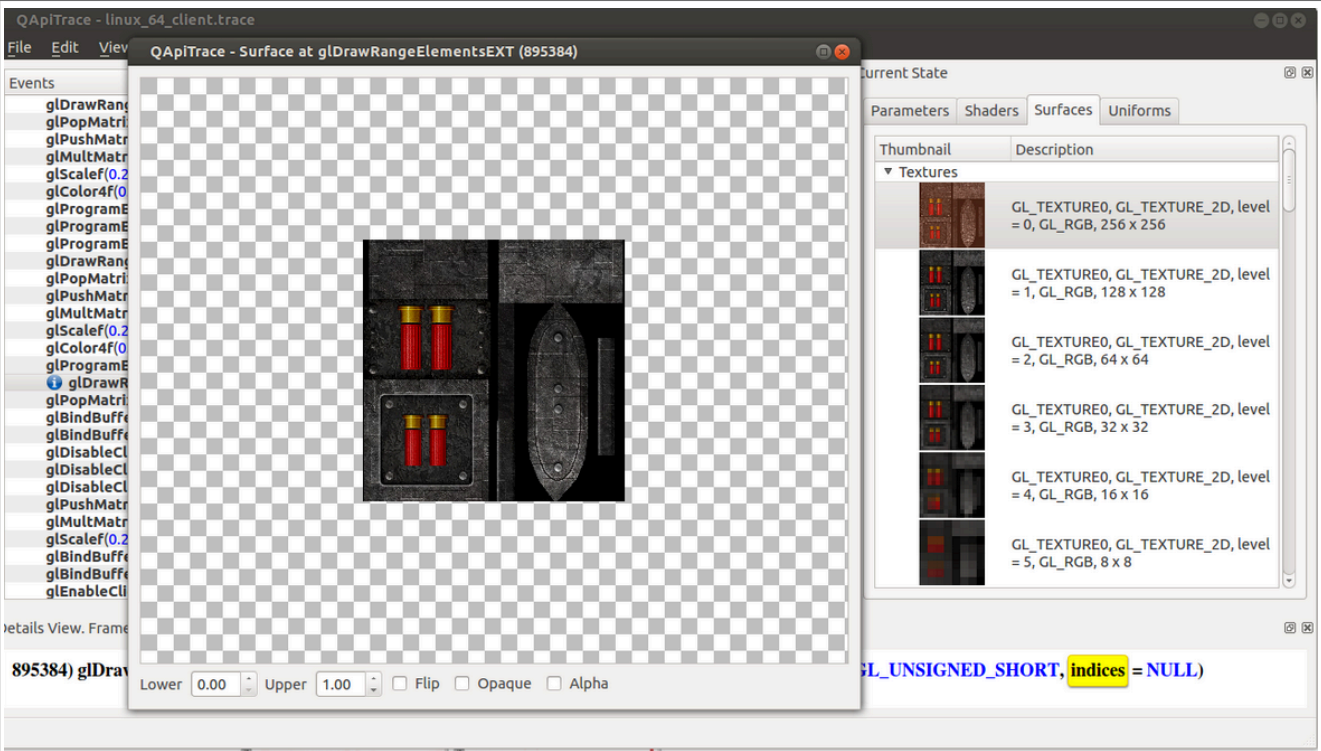


Figure 39. Checking Texture

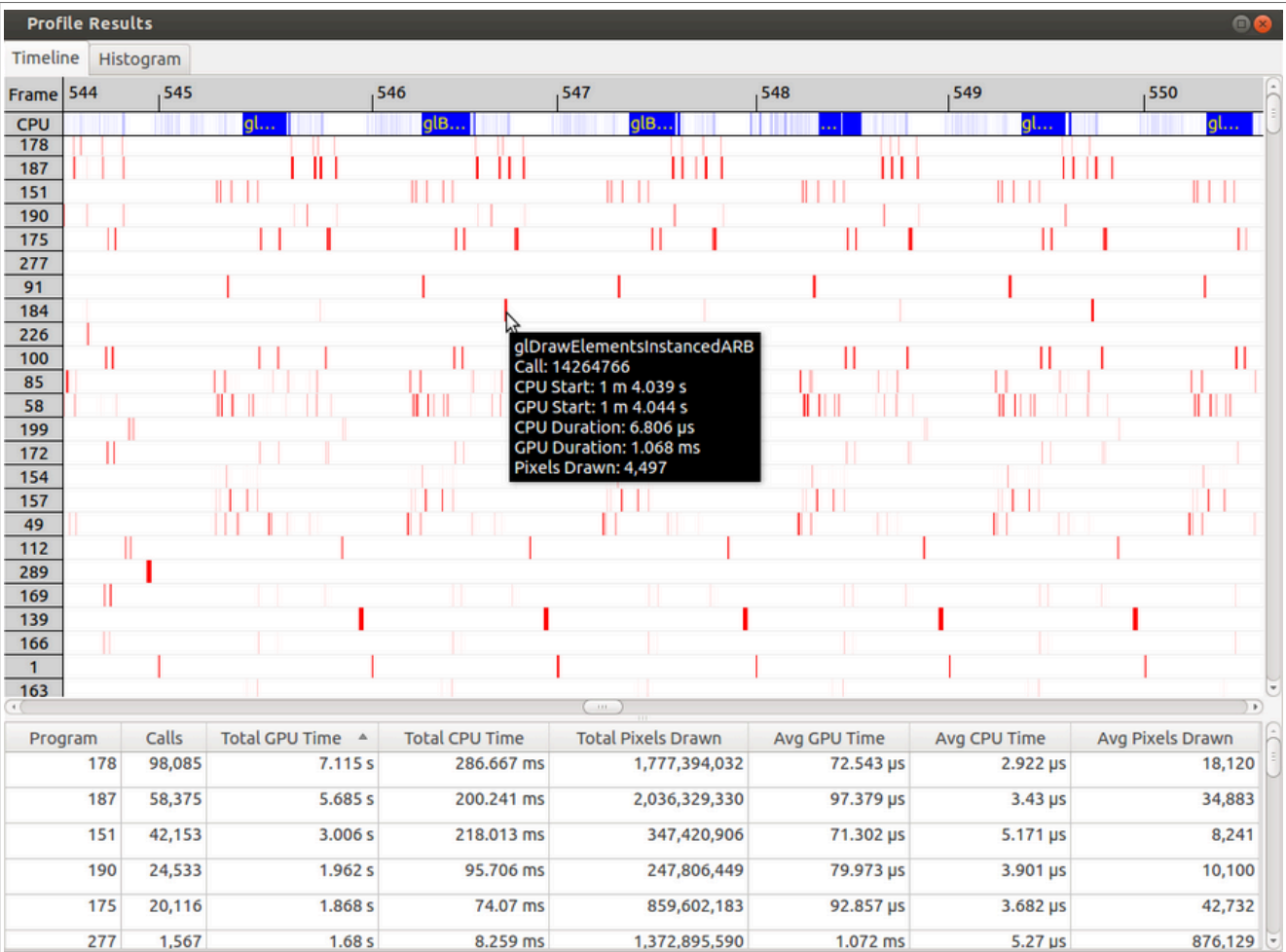


Figure 40. Checking performance

15.3.4 Reference

1. Apitrace introduction: apitrace.github.io/
2. More uses: github.com/apitrace/apitrace/blob/master/README.markdown

15.4 Renderdoc

Renderdoc is a frame-capture based graphics debugger, generally support for Vulkan, D3D11, D3D12, OpenGL, and OpenGL ES development. On i.MX, support is available only for Vulkan. RenderDoc provides tools for deep analysis and graphics inspection, as well as detailed examination of API usage - allowing developers to locate bugs and problems in their programs.

15.4.1 Renderdoc components

Renderdoc source code release is part of the i.MX Yocto Project Linux BSP release. The source code has more patches added on top of the official Renderdoc release. The Yocto Project recipes pull the renderdoccmd tool source package and install it as needed for the supported backend.

Renderdoc has a set of PC tools. Prebuilt binary packages can be directly downloaded from Renderdoc website.

The renderdoccmd tool will be available on the i.MX board for capturing frames and replaying locally, as for debugging purposes qrenderdoc needs to be used remotely on a host machine.

15.4.2 Running renderdoccmd on i.MX

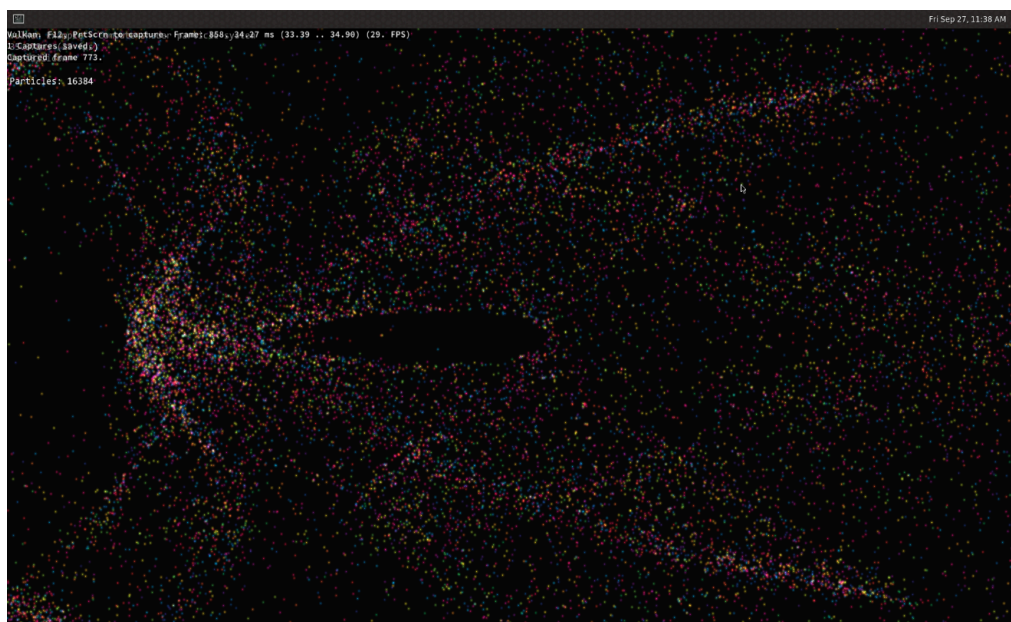
```
renderdoccmd capture <options> <app_name> <arguments>
```

Renderdoccmd usage example:

- For capturing a frame from a graphics application available in the SDK, run

```
renderdoccmd capture /opt/imx-gpu-sdk/Vulkan/Some_example/Some_example_Wayland
```

- Press F12 to capture frames:

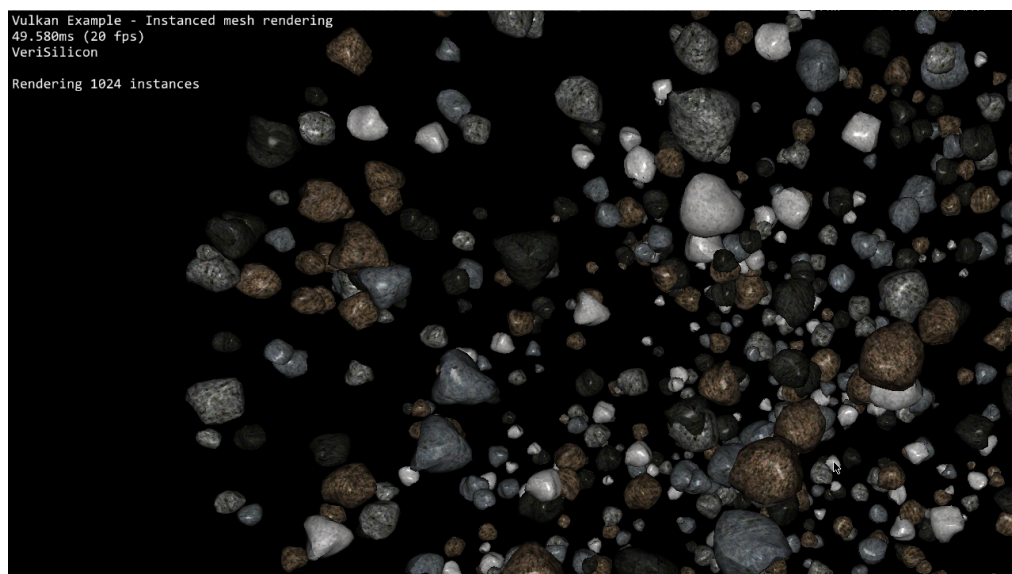


Frames will be written in /tmp/Renderdoc/ (run `renderdoccmd capture` to see all the options)

- For replaying a capture run

```
renderdoccmd replay /path/to/capture/file
```

(Run `renderdoccmd replay` for more options).

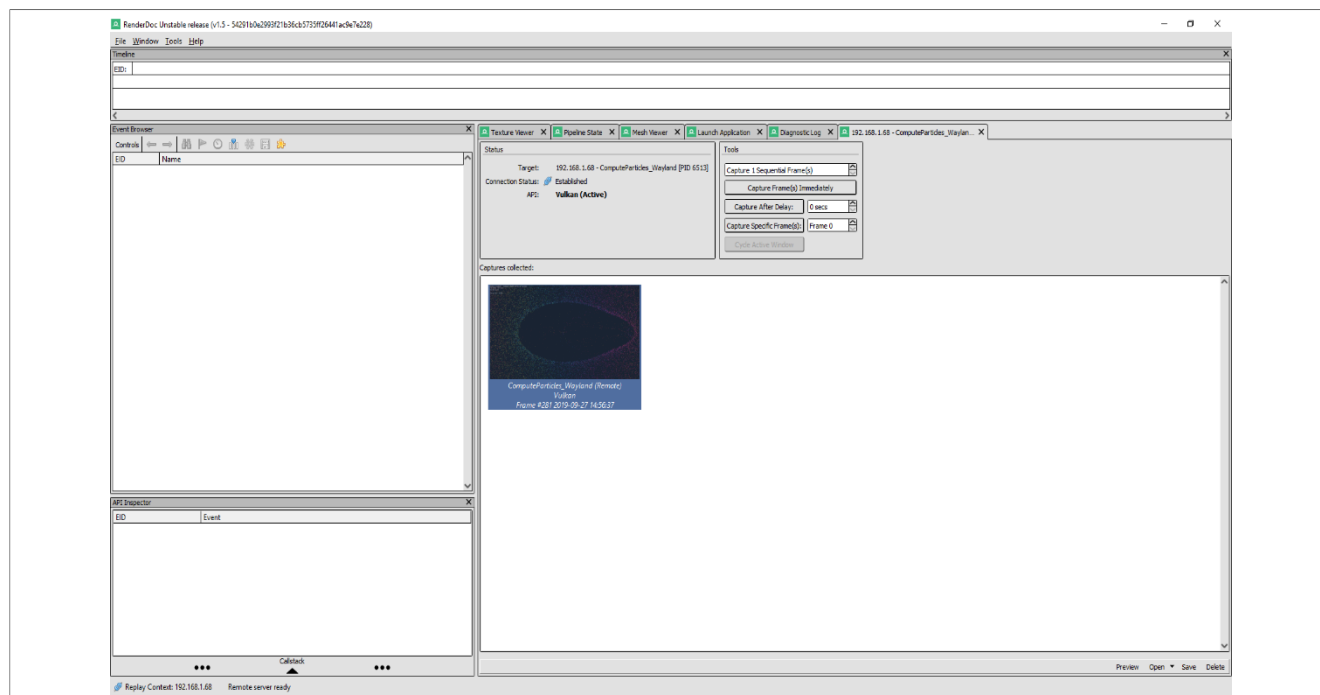


- Press F for full screen. Press F again to come back to the default window dimensions. Press ESC to quit replaying.

15.4.3 Capturing and replaying remotely

Usage:

1. Download a Renderdoc build from the website on your Windows/Linux host machine.
2. Set up a connection between the host and the board.
3. On the i.MX board, run `renderdoccmd remoteserver`.
4. On your machine, run `qrenderdoc`. Go to **File -> Attach to running instance**.
5. In the **Remote Host Manager Window**, add the target's IP address. Then `qrenderdoc` on your local machine should establish a connection with the `renderdoccmd` server instance.
6. In the left down corner of the screen, select **Replay Context** and change it from Local to the target's IP address.
7. Select **File -> Launch Application**. On **Executable Path**, insert the path of your Vulkan example from the target: `/opt/imx-gpu-sdk/Vulkan/Some_example/Some_example_Wayland`.
8. Press **Launch** and then capture. A new capture preview should appear.
9. You can save it by right clicking **Save** on the preview.



10. If you close the Vulkan application from the board, qrenderdoc will open the capture file.
11. To debug the capture, check the documentation available on the Renderdoc site.
12. To replay remotely, just use `renderdoccmd` on your local machine. Run `renderdoccmd replay --remote-host <target ip> <capture_file_on_you_local_machine>` and you should see exactly the same thing as when running on the target locally.

Notes for Android:

- Before starting the remote server and Vulkan application, Android HWUI renderer must be set to Vulkan renderer. In Android console: `setprop debug.hwui.renderer skiavk`.
- Remote server on the Android platform is started from qrenderdoc application. Connect the board to PC through the USB-C port. In qrenderdoc, go to **Tools -> Manages Remote Servers**, and select the connected board. For example, "npx MEK-MX8Q", and press the **Run Server** button.
- On the Android platform, add permission "Allow access to manage all files" to RenderDocCmd when it is launched for the first time.
- Launch an application from qrenderdoc. Be sure the correct Replay Context is selected in the left bottom corner. Select a Vulkan application in the **Executable path** field from the **Launch Application** tab. Click the **Launch** button.
- Capture frame from qrenderdoc.
- Capture is replayed automatically on the Android platform when the Vulkan application is closed.

15.4.4 Reference

<https://renderdoc.org/>

<https://github.com/baldurk/renderdoc/blob/v1.x/README.md>

16 VSI GPU Memory Introduction

16.1 VSI GPU memory overview

- OpenGL-ES
 - Texture buffer
 - Vertex buffer
 - Index buffer
 - PBuffer surface
 - Color buffer
 - Z/Stencil buffer
 - HZ depth buffer
 - Tiled status buffer
 - 3D Command buffer
 - 3D Context buffer
- OpenVG
 - Image buffer
 - Tessellation buffer
 - VG command buffer
 - VG context buffer
- 2D buffers
 - 2D command buffer
 - 2D temporary buffer

16.2 VSI GPU memory pools

- Reserved memory

In the Linux 6.6.y kernel, the memory is reserved from CMA implemented in the GPU kernel driver, the size can be changed through U-Boot args with `galcore.contiguousSize =xxx`.
The memory allocation and lock very fast, but cannot support cacheable attribute.
- Contiguous memory

The contiguous memory is from CMA or Normal or Highmem with `alloc_pages_exact`.
The GPU driver tries the CMA allocator for non-cacheable request first. If CMA memory is used up, it goes to system allocator.
The CMA allocator does not support the cacheable attribute, the system allocator supports cacheable attribute, but the memory performance is slow with the additional cache flush operations.
- Virtual memory pool

The virtual memory is from Normal or Highmem with multiple `page_alloc`.
The memory support cacheable attribute, but slow with GPU MMU and cache flush.
The GPU virtual command buffer is allocated from virtual memory pool directly.
- Nonpaged memory pool

In the 5.x GPU driver, this pool is not used any more.

16.3 VSI GPU memory allocators

Two kinds of allocators are implemented in i.MX GPU kernel driver, see `drivers/mxc/gpu-viv/`.

- The video memory allocator implementation is very complicated. The memory is from the reserved pool, system contiguous pool (supports CMA), or system virtual pool (enables GPU MMU).

- The CMA allocator supports non-cacheable contiguous memory. It is implemented as a part of contiguous pool. When the system requests contiguous memory, the allocator tries CMA first. If CMA is used up, it goes to allocate the system contiguous pages.
- GPU memory-killer is implemented for special requirement of force contiguous GPU memory.

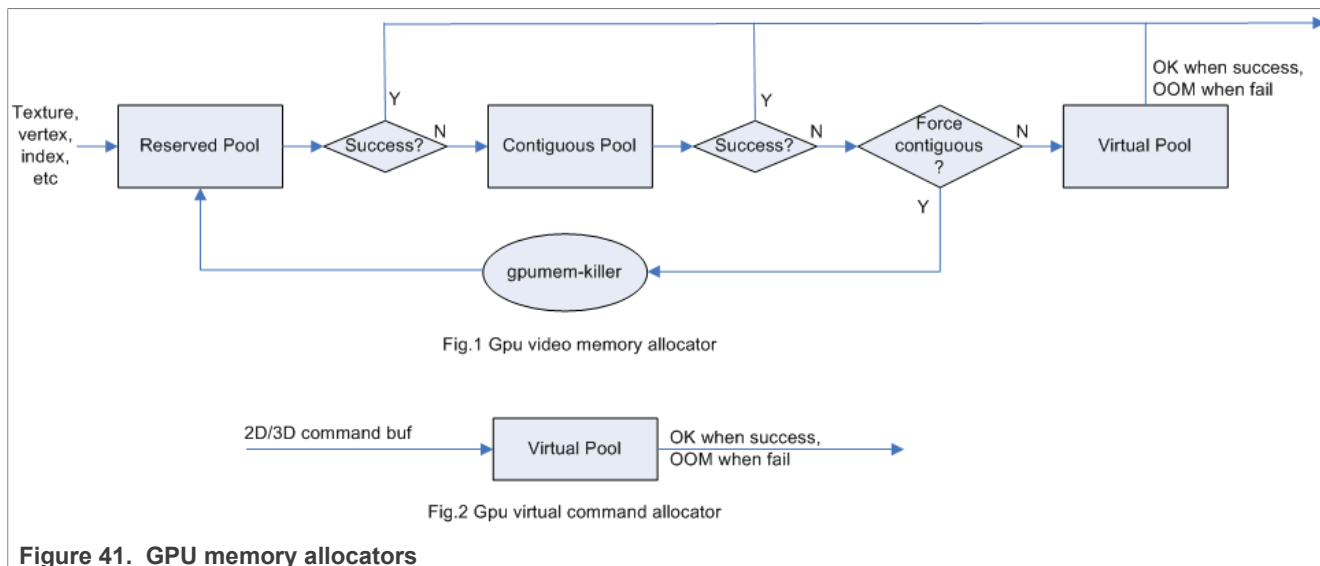


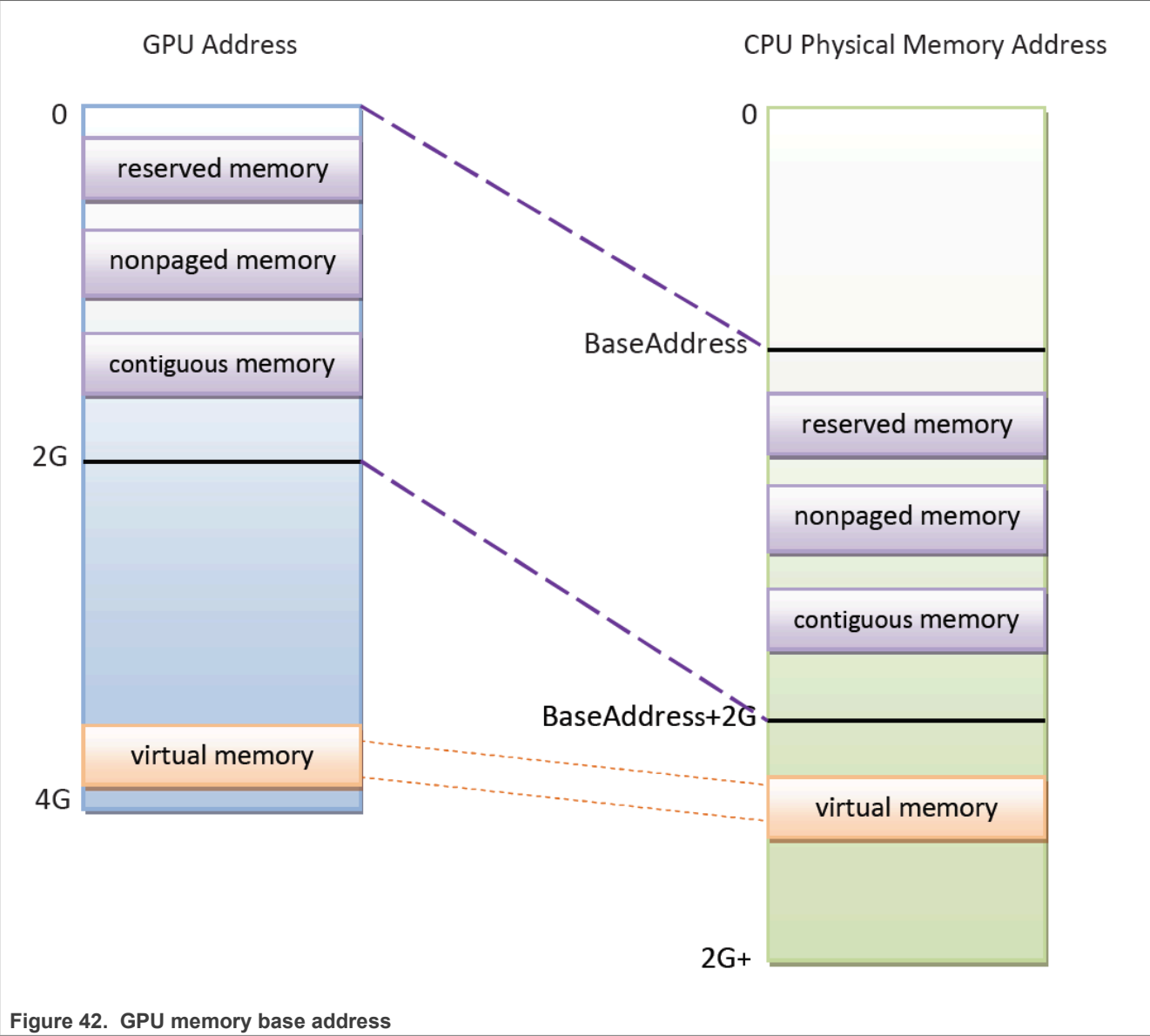
Figure 41. GPU memory allocators

16.4 VSI GPU reserved memory

- The reserved memory is managed by two dual linked lists, one is free list, and another is node list.
- When allocate the reserved memory, the free list is scanned from head to tail until a available node is selected, it is very fast but makes more memory fragments, under test, 10~20M of 128M is not available to use after a lot of allocate/free operations.
- When the available node is selected, it is removed from the free list, but it always keeps the dual linked nodes to merge the conjoint available memory when freed.
- The reserved memory is mapped once when application process is attached, during 3D application running, the memory map/un-map operations are very fast, the virtual address is just calculated with logical base and offset.

16.5 VSI GPU memory base address

- GPU support contiguous physical memory within (0-2G) address directly:
 - GPU address = CPU Physical address – GPU BaseAddress
- GPU MMU is enabled for two kinds of memory type as below:
 - Separated page memory from Virtual memory pool
 - Contiguous page memory with address out of (0-2G)
- BaseAddress should be set to RAM start address to achieve the better performance by reducing GPU MMU mapping.



17 Mali Valhall GPU

i.MX 95 integrates the Mali Vale V2 GPU, a significant change in the graphics from previous i.MX. It performs 32 FP32 FMAs, reads four bilinear filtered texture samples, blends two fragments, and writes two pixels per clock. For more details about Mali Vale shader core, see <https://developer.arm.com/documentation/102203/0100/?lang=en>.

The Vale GPU has a module named Command Stream Front (CSF), which replaces the job management in the Midgard and Bifrost architecture, and offloads some operation from CPU to GPU, so that the CPU can focus on general operations to increase the rendering FPS. It is more friendly to the newer graphics API vulkan.

17.1 Features

- Tile-Based Deferred Rendering (TBDR)
- OpenGL ES 1.1/2.0/3.0/3.1/3.2

- Vulkan 1.3
- OpenCL 3.0
- AFBC/AFRC

17.2 Mali Shader offline Compiler

malisc is a Mali offline shader compiler to compile the vertex shader and fragment shader. It is only for syntax checking when you are developing the shader. Its output is not ELF binaries. It is a specific Mali version called Mali Binaries Specification version2 (MBS2).

```
#version 320 es
//test.vert to show malisc usage
in vec4 position;
out vec4 color;

void main (void)
{
    gl_Position = position;
    color = vec4(1.0f, 0.0f, 0.0f, 1.0f);
}
```

You can modify the shader source above to learn the Malisc usage.

```
Usage: malisc --util [options] <a.vert> [<a.frag> <b.vert> ...]
local@imx95-19x19-lpddr5-evk:~# malisc test.vert --core=Mali-G310 --
revision=r0p0
```

More options can be found when the following command is executed:

```
local@imx95-19x19-lpddr5-evk:~# malisc --help
```

17.3 Mali OpenCL Offline Compiler

mali_clcc is Mali OpenCL C offline compiler. It can be used for syntax checking, and its output program binary can be used with `clCreateProgramWithBinary()`.

```
//test.cl
__kernel void vector_add(__global float* a, __global float* b, __global float*
c)
{
    //get the global ID
    const int i = get_global_id(0);

    //run the vector add
    c[i] = a[i] + b[i];
}
```

The kernel source file above can be compiled with the following command:

```
local@imx95-19x19-lpddr5-evk:~# mali_clcc test.cl -o test.bin
```

More options can be found when the following command is executed:

```
local@imx95-19x19-lpddr5-evk:~# mali_clcc -help
```

17.4 Capture the GLES application with ptrace

The ptrace has been installed into the i.MX 95 BSP rootfs since LF6.6.23_2.0.0. Otherwise, build it from the source code or copy it from the released BSP.

The [ptrace](#) is the software for capturing GLES calls of an application and replaying them on a different device, keeping the GPU workload the same. It is similar to the open source apitrace project, but optimized for performance measurements. See the [User Manual](#).

Capture Examples

On the Linux OS, some environment variables must be set.

- Set LD_LIBRARY_PATH to the path of the built fake driver.
- Set INTERCEPTOR_LIB to the location of the built interceptor [libegltrace.so](#).
- Set TRACE_LIBEGL to the location of DDK driver [libEGL.so](#).
- Set TRACE_LIBGLES1 to the location of DDK driver [libGLESv1_CM.so](#).
- Set TRACE_LIBGLES2 to the location of DDK driver [libGLESv2.so](#).
- Set OUT_TRACE_FILE variable as the path and filename of the captured trace.

The following are the recommended usages:

Example 1, capture the QT application:

```
LD_PRELOAD=/opt/ptrace/lib/libegltrace.so OUT_TRACE_FILE=qtexample ./qt-bin
```

Example 2, capture the glmark2-es-wayland:

```
LD_LIBRARY_PATH=/opt/ptrace/lib/ INTERCEPTOR_LIB=/opt/ptrace/lib/  
libegltrace.so TRACE_LIBEGL=/usr/lib/libEGL.so TRACE_LIBGLES1=/usr/lib/  
libGLESv1_CM.so TRACE_LIBGLES2=/usr/lib/libGLESv2.so LD_PRELOAD=/opt/ptrace/  
lib/libegltrace.so OUT_TRACE_FILE=glmark glmark2-es2-wayland -b build
```

Replay and check the EGL/GLES calls. The following command can be used to replay the ptrace capture:

```
export PATH=/opt/ptrace/bin/:$PATH  
paretrace [ -debugfull ] <tracefile.pat>
```

Note: *-debugfull is optional. Output all of the current invoked GL functions, with callNo, frameNo, and skipped or discarded information.*

17.5 Arm Performance Studio

[Arm Performance Studio](#) is the new name for Arm Mobile Studio. If the tool is upgraded to 2024.0 or later from a previous version of Arm Mobile Studio, its name changes in the download package, the tool installers, the web content, and the user documentation. There is no reduction in functionality and all the tools are still available for free. Additionally, it supports capturing from the Linux targets, and you now get RenderDoc for Arm GPUs in the bundle. Read the blogs for more information.

17.5.1 Tools included in Arm Performance Studio

Arm Performance Studio includes the following tools that each targets a different stage in the profiling workflow.

Table 45. Tools included in Arm Performance Studio

Tool	Description
Streamline	Captures a performance profile for deep-dive analysis, using all of the CPU, GPU, and memory system performance data in the system. Identifies the critical path hardware units for your application, as well as workload efficiency metrics, allowing you to target optimizations at the areas that matter most.
Performance Advisor	Part of the Streamline tool, Performance Advisor generates an easy-to-read performance report from an annotated Streamline profile. Gets actionable advice about how to optimize your application. These reports can be generated manually from a Streamline capture, but they are designed to ease the deployment of automated performance testing workflows.
Frame Advisor	Use Frame Advisor to analyze a problem frame from a mobile application. Captures the API calls and rendering and gets comprehensive geometry metrics to discover what might be slowing down your application or overheating the device.
Mali Offline Compiler	Compiles your shader programs and checks how they will perform across on any Mali GPUs. Performance reports give you information on the shader register usage and thread occupancy, an estimated cycle cost breakdown for the target GPU, and other stage-specific performance feedback.
RenderDoc for Arm GPUs	The industry-standard tool for debugging Vulkan graphics applications, including early support for Arm GPU extensions and Android features.

More details can be found in the help content or tools user manual, which are delivered in the Arm Performance Studio.

17.5.2 Streamline offline capture

The command for streamline offline capture is as follows. `gator` can be found in the installation directory of Arm Performance Studio.

```
gator --output ~/output --config-xml /etc/streamline/configuration.xml -s /
etc/streamline/session.xml
```

`session.xml` is as follows, and `configuration.xml` can be generated in the Arm Performance Studio.

```
<?xml version="1.0" encoding="UTF-8"?>
<session call_stack_unwinding="yes" filter_call_stacks="yes"
  parse_debug_info="yes" exclude_kernel_events="no" version="1"
  high_resolution="no" buffer_mode="streaming" sample_rate="normal" duration="60"
  target_address="" live_rate="100" stop_gator="no" capture_log="yes">
  <energy_capture version="1" command_path="C:\Program Files\Arm
\Development Studio 2023.1\sw\streamline\bin\win-64\caiman.exe" type="none">
    <channel id="0" resistance="20" power="yes"/>
  </energy_capture>
</session>
```

17.6 References and Useful links

- Tile-Based Rendering: <https://developer.arm.com/documentation/102662/0100/?lang=en>
- The Valhall shader core: <https://developer.arm.com/documentation/102203/0100/?lang=en>
- Arm Mali Offline Compiler User Guide: <https://developer.arm.com/documentation/101863/0803/?lang=en>

18 Application Programming Recommendations

The recommendations listed below take a holistic approach centered on overall system level optimizations that balance graphics and system resources.

18.1 Understanding the system configuration and target application

Knowing details about the application and use case allows developers to correctly utilize the hardware resources in an ideal access pattern. For example, an implementation for a 2D or 3D GUI could be rendered in a single pass instead of multiple passes if the draw call sequence is correctly ordered. In addition, knowing the most common graphics function calls allow developers to parallelize rendering to maximize performance.

Using Vivante and vendor-specific SoC profiling tools, you can determine bottlenecks in the GPU and CPU and make changes as needed. For example, in a 3D game, most CPU cycles may be spent on audio processing, AI, and physics and less on rendering or scene setup for the GPU. In this instance, the application is CPU-bound and configurations dealing with non-graphics tasks need to be reviewed and modified. If the system is GPU-bound, the profiler can point out where the GPU programming code bottlenecks are located and which sections to optimize to remove restrictions.

18.2 Optimizing off-chip data transfer such as accessing off-chip DDR memory/mobile DDR memory

Any data transfer off-chip takes bandwidth and resources from other functional blocks in the SoC, increases power, and causes additional cycles of latency and delay as the GPU pipeline needs to wait for data to return from memory. Using on-chip cache and writing the application to better take advantage of cache locality and coherency increase performance. In addition, accessing the GPU frame buffer from the CPU (not recommended) cause the driver to flush all queued render commands in the command buffer, slowing down performance as the GPU has to wait since the command queue is partially empty (inefficient use of resources) and CPU-GPU synchronization is not parallelized.

18.3 Avoiding W-clipping issue in the application program

The w-clipping overflow issue typically occurs with these three factors:

- **Objects with very large primitives.**
In a 3D scene, this is usually the sky, the outer world or a long road that expands far behind the camera and far in front of the camera. At the same time, the object may also expand far in either the x or y direction.
- **Near-plane with a very small value**
Usually this value is very close to zero. An example would be 10^{-4} .
- **Large screen resolution**

These three factors can cause the final window coordinate to overflow the 24-bit mantissa precision in IEEE single precision floating point format.

The following are suggested ways to modify an application to avoid overflow:

1. For draw calls with very large primitives such as sky or world, set the near-plane to 0.99 as an initial value.
2. If this removes the rendering error and the entire scene is rendered correctly, the issue can be considered resolved.
3. If the rendering error is still there and no desired objects are being culled (or there are no missing objects), increase the near-plane value until the rendering error disappears.
4. If the near-plane value is large (>10.0) already, the issue persists and some desired objects are being culled, reduce the near-plane value until the desired objects appear again then go to the next step.
5. Tessellate the large objects into smaller primitives until the rendering error disappears.

Please note that the suggested near plane adjustment can be done on a per draw call basis, and only needs to be modified for objects with very large primitives. Some applications scale the object by reducing the w value in vertex shader, as changing w value will finally affect the near plane, which is not recommended. A better way to scale the object is scale the x, y, z coordinate, not w.

18.4 Avoiding GPU hanging and data corruption when using occlusion query

Description:

On i.MX 6Dual/Quad GPU IP, both Hierarchical Depth (Hz) write and Occlusion Query (OQ) write share the same port. If HZ Fast Clear (FC) is enabled, and OQ uses the HZ port to perform a write, the HZ FC data may become corrupted, even leading to GPU hanging unexpectedly.

Software Workaround:

A software workaround is recommended for this issue and is available from L4.9 bsp release. Because the issue occurs very infrequently, a per-application work around is most efficient. Software will disable HZ with a per-app detection and also provide a new environment variable control (VIV_DISABLE_HZ).

18.5 Avoiding random cache or memory access

Cache thrashing, misses, and the need to access data in external memory causes performance hits. An example would be random texture cache access since it is expensive when performing per-pixel texture reads if the texture units need to access the cache randomly and go off-chip if there is a cache miss.

18.6 Optimizing your use of system memory

Memory is a valuable resource that needs to be shared between the GPU (frame buffer), CPU, system, and other applications. If you allocate too much memory for your OpenGL ES application, less memory is available for the rest of the system, which may impact system performance. Claim enough memory as needed for your application then deallocate it as soon as your application no longer needs it. For example, you can allocate a depth buffer only when needed or if your application only needs partial resources, load the necessary items initially and load the rest later.

18.7 Targeting a fixed frame rate that is visibly smooth

Smooth frame rate is achieved from a combination of a constant FPS and the lowest FPS (frames per second) that is visually acceptable. There is a trade-off between power and frame rates since the graphics engine loading increases with higher FPS. If the application is smooth at 30 FPS and no visual differences for the application are perceived at 50 FPS, then the developer should cap the FPS at 30 since the extra 20 FPS do not make a visual difference. The FPS limit also guarantees an achievable frame rate at all times. The savings in FPS help lower GPU and system power consumption.

18.8 Minimizing GL state changes

Setting up state values between draw calls adds significant overhead to application performance so they must be minimized. Most of these call setups are redundant since you are saving / restoring states prior to drawing. Try to avoid setting up multiple state calls between draw calls or setting the same values for multiple calls. Sometimes when a specific texture is used, it is better to sort draw calls around that texture to avoid texture thrashing which inhibits performance. Application developers should also try to group state changes.

18.9 Batch primitives to minimize the number of draw calls

When your application submits primitives to be processed by OpenGL ES, the CPU spends time preparing commands for the GPU hardware to execute. If you batch your draw calls into fewer calls, you reduce the CPU overhead and increase draw call efficiency. Batch processing allows a group of draw calls to be quickly executed without any intervention from the CPU (driver or application) in a fire-and-forget method.

Some examples of batching primitives are:

- Branching in shaders may allow better batching since each branch can be grouped together for execution.
- For primitives like triangle strips, the developer can combine multiple strips that share the same state to save successive draw calls (and state changes) into a single batch call that uses the same state (single setup) for many triangles.
- Developers can also consolidate primitives that are drawn in close proximity to take advantage of spatial relationships. If the batched primitives are too far apart, it is more difficult for the application to effectively cull if they are not visible in the frame.

18.10 Performing calculations per vertex instead of per fragment/pixel

Since the number of vertices is usually much less than the number of fragments/pixels, it is cheaper to do per vertex calculations to save processing power.

18.11 Enabling early-Z, hierarchical-Z, and back face culling

Hardware support of depth testing to determine if objects are in the user's field of view are used to save workload and processing on vertex and pixel processing. If the object is in view, then the vertices are sent down the pipeline for processing. If the object is hidden or not viewable, the triangles are culled and not sent to the pipeline. This improves graphics performance since computations are only spent on visible objects. If the application already knows details about the contents and relative position of objects in the scene or screen, the developer can use that information to automatically bound areas that never need to be touched (for example an automotive application that has multiple layers of dials where parts of the underlying dials are occluded can have the application avoid occluded areas from the beginning). Another optimization is to perform basic culling on the CPU since the CPU has first-hand information about the scene details and object positions so it knows what scene data to send to the GPU.

18.12 Using branching carefully

Static branches perform well since states are known but they tend to use many general purpose registers. An example is a long shader that combines multiple shaders into a single, large shader that reduces state changes and batch draw calls. Dynamic branching has non-constant overhead since it processes multiple pixels as one and everything executes whether a branch is taken or not. In other words, dynamic branching goes through different permutations/branches in parallel to reach the correct results. If all pixels take the same path, then performance is good. The more pixels processed translates to higher overhead and lower performance. For dynamic branching, smaller pixel sizes/groups are optimal for throughput. Developers need to be aware of branching in their code to make sure excessive calculations and branches are efficient. Profiling tools can help determine if certain parts of code are optimized or not.

18.13 Using VBOs instead of static or stack data as vertex data

A vertex buffer object (VBO) is a buffer object that provides the benefits of vertex array and display list and allows a substantial performance gain for uploading data (vertex position, color, normals, and texture coordinates) to the GPU. VBOs create buffer objects in memory and allow the GPU to directly access memory without CPU intervention (DMA). The memory manager can optimize buffer placement using feedback from the

application. VBOs can also handle static and dynamic data sets and are managed by the Vivante driver. The benefits of each are:

- A vertex array reduces the number of function calls and allows redundant data to be shared between related vertices, instead of re-sending all the data each time. Access to data can be referenced by the array index.
- The display list allows commands to be stored for later execution and can be used repeatedly over multiple frames without re-transmitting data, thus minimizing CPU cycles to transfer data. The display list can also be shared by multiple OpenGL / OpenGL ES clients so they can access the same buffer with the corresponding identifier. If you put computationally expensive operations (ex. lighting or material calculations) inside display lists, then these computations are processed once when the list is created and the final result can be re-used multiple times without needing to re-calculate again.

If you combine the benefits of both by using VBO, the performance is enhanced over static or stack data sets.

18.14 Using dynamic VBO when the data is changing frame by frame

Locking a static vertex buffer while the GPU is using it can create a performance penalty since the GPU needs to finish reading the vertex data from the buffer before it can return to the calling application. Locking and rendering from a static buffer many times per frame also prevents the GPU buffering render commands since it must finish commands before returning the lock pointer. Without buffered commands the GPU remains idle until the application finishes filling the vertex buffer and issues the draw commands.

If the scene data never changes from frame to frame then a static buffer may be sufficient. With newer applications (ex. games, maps) that have dynamic viewports where vertex data changes multiple times per frame or frame-to-frame, then a dynamic VBO is required to ensure performance is still met. If the *current* buffer is being used by the GPU when a lock is called, a pointer to a *new* buffer location is returned to the application to ensure updated data is written to the *new* buffer. The GPU can still access the old data (current buffer) while the application puts updated data into the new buffer. The Vivante memory management unit and driver automatically take care of allocating, re-allocating, or destroying buffers.

You can implement dynamic VBO depending on your preference, but one recommendation is to allocate a 1 MB dynamic VBO block and upload data to using different offsets for each dynamic buffer. If the buffer overflows you can loop back and use location offset 0 again.

18.15 Tessellating your data to make Hierarchical Z (HZ) work

We can break this into how OpenGL and OpenGL ES handle this use case.

OpenGL only renders simple convex polygons (edges only intersect at vertices with no duplicate vertices and only two edges meet at any vertex), in addition to points, lines, and triangles. If the application requires concave polygons (polygons with holes or intersecting edges), those polygons need to be subdivided into simple convex polygons, which is called tessellation (subdividing a polygon mesh into a bunch of smaller meshes). Once you have all the meshes in place our HZ hardware can automatically cull hidden polygons to efficiently process the frame, effectively breaking the frame into smaller chunks that can be processed very fast.

OpenGL ES only renders triangles, lines, and points. The same concepts apply as in OpenGL, which is to avoid very large polygons by breaking them down into smaller polygons where our internal GPU scheduler can distribute them into multiple threads to fully parallelize the process and remove hidden polygons.

18.16 Using dynamic textures as a texture cache (texture atlas)

The main reason for using dynamic textures as a cache is the application developer can create one larger texture that is subdivided into different regions (texture atlas). The application can upload data into each region and use an application side texture atlas to access the data. Each dynamic texture and sub-region can be locked, written to, and unlocked each frame, as needed. This method of allocating once is more efficient than using multiple smaller textures that need to be allocated, generated, and then destroyed each time.

18.17 Sticking small triangle strips together

It is better to combine several small, spatially related triangle strips together into a larger triangle strip to minimize overhead and increase performance. For each triangle strip, there are overhead and start up costs that are required by the CPU and GPU, including state loads. If there are too many small triangle strips that need to be loaded, this impacts performance. An application developer can combine multiple triangle strips by adding a degenerate triangle to join the strips together. The overhead to restart multiple new strips is much higher than adding the degenerate triangle.

18.18 Specifying EGL configuration attributes precisely

To obtain a 16 bit/pixel window buffer for rendering, the EGL config attributes need to be specified precisely according to the EGL spec. Specifying inaccurate EGL attributes may result in getting a 32-bit bit/pixel window buffer which doubles the bandwidth requirement for rendering which in turn leads to lower performance.

18.19 Using aligned texture/render buffers

The GPUs work on buffers with hardware-specific width/height alignment for better efficiency. Use the available API to query the GPU buffer alignment and allocate the texture / render buffers to satisfy these requirements, to avoid the cost of copies to aligned shadow memory.

18.20 Disabling MSAA rendering unless high quality is needed

Although MSAA rendering can achieve higher image quality with smoother lines and triangle edges, it requires much higher (4x, 8x) bandwidth because it has to render a single pixel 4x/8x times. So, if high rendering quality is not required, MSAA should be disabled.

18.21 Avoiding partial clears

Most GPUs have special hardware logic to do a fast clear of an entire buffer. So it is better to utilize the fast clear function to clear the entire buffer then render graphics again, instead of doing a partial clear to preserve a graphics region. If a partial clear is required by the application, make sure the clear area is aligned according to the GPU-specific requirements. Unaligned partial clears are expensive and should be avoided.

18.22 Avoiding mask operations

Do not use mask unless the mask is 0 (other than when you need a specific render quality). Clearing a surface with mask (color/depth stencil mask) could have a performance penalty. Pixel mask operations are normally pretty expensive on some GPUs as the mask operation has to be done on every single pixel.

18.23 Using MIPMAP textures

MIPMAP textures enable the application to sample a lower resolution texture image (1/2, 1/4, 1/8, 1/16, ... size of the original texture image) when the triangle is rendering further away from the view point. Thus, the bandwidth required to read the texture image is reduced which leads to better performance.

18.24 Using compressed textures if constricted by RAM/ROM budget

Compressed textures are normally only a fraction (up to 1/8) of the original texture size. Using compressed textures reduces the storage requirements in memory and can also reduce the required texture upload bandwidth, when using a format that is supported natively by the hardware.

Compressed textures should not be chosen, if only for the purposes of reducing the memory bandwidth required for sampling of the texture during rendering. This is because due to a fixed read request size from the GPU, the memory controller load is the same as for an uncompressed texture.

18.25 Drawing objects from near to far if possible

Drawing objects from near to far normally has better performance because the objects in the near foreground can block entire or partial objects in the background. Most GPUs have early Z rejection logic to reject the pixels that fail a Z compare. The GPU can skip fragment shader computations on these rejected pixels.

18.26 Avoiding indexed triangle strips

Index triangle strips can usually maximize the vertex cache utilization as each set of vertex data can be used in two triangles. There is however an errata in the GC2000 and GC880 GPUs which requires a SW conversion of indexed triangle strips to triangle lists in the driver. For small strips the conversion overhead is negligible, but for large geometries a different primitive type should be used.

18.27 Limiting vertex attribute stride within 256 bytes

Most Vivante GPUs provide native support for a 256 byte vertex attribute stride. If the vertex attribute stride is larger than 256 bytes, then the driver has to copy the vertex data around. Hardware versions v55 and higher (such as the GC7000L v55) support a 2048 byte vertex attribute stride as required in the OES3.1 spec.

18.28 Avoiding binding buffers to mixed index/vertex array

Most of Vivante GPUs do not natively support mixed index/vertex arrays. So the Vivante driver must copy the index and vertex data around to form separate vertex data streams for the GPU. Avoid mixing index and vertex data so the driver does not have to incur a performance hit while performing this task.

18.29 Avoiding using CPU to update texture/buffer contexts during render

Do not use the CPU to update texture/buffer contexts in the middle of rendering. Using the CPU to update texture/buffer causes the rendering pipeline to flush and stall, so that CPU can safely update the buffer contents. The pipeline flush/stall/resume causes significant performance impact.

18.30 Avoiding frequent context switching

Context switch is an inherently expensive operation as many GPU states need to be reset to start a new rendering context. Thus, frequent context switching has a negative impact on application performance.

18.31 Optimizing resources within a shader

Most GPUs have optimal support for a limited amount of resources (uniforms, varying, etc.). Using resources beyond the optimal working set causes the GPU to fetch/store resources from a lower performance memory pool and shader performance is negatively impacted.

18.32 Avoiding using glScissor Clear for small regions

glScissor Clear for small regions (less than 16x8 aligned window) fall back to CPU so the performance is not optimal.

18.33 Using PRE to accelerate data transfer

PRE is an optimized hardware that can transform tiled format image to linear framebuffer. With PRE, GPU can only output tiled render target and has no need to resolve it. To enable the PRE feature, set the environment GPU_VIV_EXT_RESOLVE variable to 1; otherwise, set it to 0. Its default value on the FB backend is 1, which means PRE is enabled by default on FB.

Warning:

VG use cases can only output the linear format image. It is impossible to render linear and tiled format target to the same framebuffer at the same time. Therefore, when running 3D use cases with PRE and VG use cases together, there is garbage on the display. Besides, when running 3D use cases with PRE, the framebuffer format is changed from linear to tiled. It is the user's responsibility to convert the format back after the use cases end, or the display is abnormal when showing the FB console.

18.34 i.MX 8QuadMax dual-GPU performance

For some legacy applications with small texture/rendering size and less shader complex, dual-GPU performance may become worse than single GPU mode, because the driver needs to take more CPU effort for dual-GPU programming, and the driver overhead is more significant than GPU load in the hardware pipeline.

For such kind of legacy case, the users can single-GPU to achieve better performance on the i.MX 8QuadMax.

19 Demo Framework

For detailed information, see the following links.

Introduction: <https://github.com/nxp-imx/gtec-demo-framework/tree/6.3.1>

Build guides:

- Yocto: https://github.com/nxp-imx/gtec-demo-framework/blob/6.3.1/Doc/Setup_guide_yocto.md
- Ubuntu: https://github.com/nxp-imx/gtec-demo-framework/blob/6.3.1/Doc/Setup_guide_ubuntu22.04.md
- Windows: https://github.com/nxp-imx/gtec-demo-framework/blob/6.3.1/Doc/Setup_guide_windows.md
- Android https://github.com/nxp-imx/gtec-demo-framework/blob/6.3.1/Doc/Setup_guide_android_sdk+ndk_on_windows.md
- Contributing: <https://github.com/nxp-imx/gtec-demo-framework/blob/6.3.1/CONTRIBUTING.md>
- Known issues: <https://github.com/nxp-imx/gtec-demo-framework/blob/6.3.1/KnownIssues.md>
- Additional documentation: <https://github.com/nxp-imx/gtec-demo-framework/tree/6.3.1/Doc>

20 Environment Variables Summary

The table below lists the environment variables (ENV) available in the GPU drivers.

The use of most environment variables remains static from driver version to driver version, but sometimes these variables need refinements to meet new, advanced conditions not present with the ENV initially introduced.

20.1 Environment variable for drivers and HAL

Table 46. Environment variables for drivers and HAL

ENV name	Backends supported	Note
FB_IGNORE_DISPLAY_SIZE	FB/WLD	0: Clip window to device display size. 1: Do not clip window to the device limits for width and height.

Table 46. Environment variables for drivers and HAL....continued

ENV name	Backends supported	Note
FB_MULTI_BUFFER	FB/WLD	Number of backend buffers of the framebuffer device. For WLD, define the multibuffer number of Weston.
FB_FRAMEBUFFER_N	FB/WLD	Define the Nth framebuffer device.
FB_LEGACY	FB	If board doesn't support drm-fb, ignore this variable. 0: GPU render through drm 1: GPU directly render to framebuffer.
VG_APITIME	FB/WLD/X11	Enable VG API function execution time print.
VIV_MGPU_AFFINITY	FB/WLD/X11	Control the multiple GPUs affinity configuration. Possible value: <ul style="list-style-type: none"> Not defined or defined as "0" GPUs work in GPU_COMBINED mode. 1:0 GPUs work in GPU_INDEPEDNENT mode, GPU0 is used. 1:1 GPUs work in GPU_INDEPEDNENT mode, GPU1 is used.
VIV_DEBUG	FB/WLD/X11	Define the user debug message level (-MSG_LEVEL: ERROR/WARNING).
VIV_FBO_PREFER_MEM	FB/WLD/X11	Renderbuffer is not freed after colorbuffer detaches from FBO (GL ES 2.0)
VIV_DISABLE_HZ	FB/WLD/X11	This variable can be specifically enabled for i.mx6d/q to avoid gpu hang with occlusion query in ES30, because of gpu hardware problem HBN1246
GPU_VIV_EXT_RESOLVE	FB/WLD/X11	Enable the external resolve mode (1 by default for FB).
GPU_VIV_DISABLE_SUPERTEXTURE	FB/WLD/X11	Disable supertiled texture (64x64 tiled texture is not used).
GPU_VIV_DISABLE_CLEAR_FB	FB/WLD/X11	Enable clear buffer when a new Window surface is created.
GPU_VIV_WL_MULTI_BUFFER	WLD	Define the client multibuffer number.
WL_EGL_SYNC_SWAP	WLD	0: Use asynchronous swap for better performance by default. 1: Enable synchronous swap with some performance impact.
DRI_IGNORE_DISPLAY_SIZE X_IGNORE_DISPLAY_SIZE	X11	0: Clip window to device display size. 1: Do not clip window to the device limits for width and height.
__GL_DEV_FB	X11	Set the path for framebuffer device like /dev/fb0.
LIBGL_ALWAYS_INDIRECT	X11	Make OGL go into indirect mode. All rendering is done by XserverSet.
LIBGL_DEBUG	X11	Print error messages to stderr if LIBGL_DEBUG env var is set. Print information messages to stderr if LIBGL_DEBUG env var is set to "verbose".
VIV_PROFILE	vProfiler	Enable profiler. Different level results generate different results.
VP_COUNTER_FILTER	vProfiler	Used to control profile different system resource like memory/CPU time usage.
VP_FRAME_END	vProfiler	When VIV_PROFILE=3, specify the frame to end profiling with vProfiler.

Table 46. Environment variables for drivers and HAL...continued

ENV name	Backends supported	Note
VP_FRAME_NUM	vProfiler	When VIV_PROFILE=1, used to specify the number of frames dumped by vProfiler.
VP_FRAME_START	vProfiler	When VIV_PROFILE=3, specify the frame to start profiling with vProfiler.
VP_OUTPUT	vProfiler	Specify the output file name of vProfiler (default is vprofiler.vpd).
VP_PROCESS_NAME	vProfiler	Choose profiler enable process (This option is only available for Android platform, not available for Linux OS).
VP_SYNC_MODE	vProfiler	Enable [1] or disable [0] the synchronous mode of vProfiler (default is synchronous enabled).
VP_USE_GLFINISH	vProfiler	Use glFinish as the frameEnd.
VIV_TRACE	vTracer	Enable tracer. Different levels could generate different logs.

20.2 Environment variable for compiler

Table 47. Environment variables for compiler

ENV NAME	Compiler	Note
VC_DUMP_SHADER_SOURCE	GLSLC/ VSC	Enable dumping the shader source code.

21 Note About the Source Code in the Document

Example code shown in this document has the following copyright and BSD-3-Clause license:

Copyright 2025 NXP Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

22 Revision History

Revision history

Document ID	Release date	Description
UG10159 v.10.2	31 March 2025	Updated Section 15.1.7.4 .
UG10159 v.10.1	16 December 2024	Updated Section 11 , Section 13 , and Section 15 .
UG10159 v.10	30 September 2024	Updated G2D API, Mali Valhall GPU, added OpenCV, RDP Backend, i.MX 95 GPU frequency scaling, removed gputool, etc.
UG10159 v.9.1	9 August 2024	Removed the subsections from Section 19 and added with reference links. Updated the document ID.
IMXGRAPHICUG_9	28 June 2024	Updated the Document ID according to the new convention.
IMXGRAPHICUG v.9	29 March 2024	Added Section "Mali Valhall GPU" and updated some section titles from "i.MX" to "Vivante".
IMXGRAPHICUG v.8.6	15 December 2023	Updated Figure 1 "GPU Scalability across i.MX processors".
IMXGRAPHICUG v.8.5.1	06/2023	Minor updates for the LF6.1.22_2.0.0 release.
IMXGRAPHICUG v.8.5	03/2023	Updated the OpenCL and Vivante IDE information.
IMXGRAPHICUG v.8.4.1	12/2022	Updated the VivanteIDE package name in Section 13.3.1.
IMXGRAPHICUG v.8.4	10/2022	Some minor updates for the android-12.1.0_1.0.0 release.
IMXGRAPHICUG v.8.3	09/2022	Updated Figure 1 and published the document in the new template.
IMXGRAPHICUG v.8.2	03/2022	Updated the back page (Legal information).
IMXGRAPHICUG v.8.2	10/2021	Added the i.MX 8ULP information to Section 1.1.
IMXGRAPHICUG v.8.1	09/2021	Removed the Section "Designing framework of OpenVX", and made minor updates for the Linux LF5.10.52_2.1.0 release.
IMXGRAPHICUG v.8	06/2021	Updated for the Linux LF5.10.35_2.0.0 and android-11.0.0_1.2.1 releases.
IMXGRAPHICUG v.7.1	03/2021	Updated Section 13.5.4 "Enabling vProfiler on Linux" as vProfiler no longer requires kernel module parameter, and made abundant changes to context description.
IMXGRAPHICUG v.7	12/2020	Updated for the Linux L5.4.70_2.3.0, android-11.0.0_1.0.0, and later release.
IMXGRAPHICUG v.6	06/2020	Updated for the Linux L5.4.24-2.1.0 and later release.
IMXGRAPHICUG v.5	04/2020	Updated for the Linux L5.4.3_2.0.0 and android-10.0.0_2.1.0 releases.
IMXGRAPHICUG v.4	11/2019	Updated the Vivante IDE information.
IMXGRAPHICUG v.3	08/2019	Added the i.MX 8M Nano information.
IMXGRAPHICUG v.2	06/2019	Made some grammatical updates.
IMXGRAPHICUG v.1	11/2018	Updated Chapter "OpenCL" with more precise information and also covered latest i.MX products.

Legal information

Definitions

Draft — A draft status on a document indicates that the content is still under internal review and subject to formal approval, which may result in modifications or additions. NXP Semiconductors does not give any representations or warranties as to the accuracy or completeness of information included in a draft version of a document and shall have no liability for the consequences of use of such information.

Disclaimers

Limited warranty and liability — Information in this document is believed to be accurate and reliable. However, NXP Semiconductors does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information and shall have no liability for the consequences of use of such information. NXP Semiconductors takes no responsibility for the content in this document if provided by an information source outside of NXP Semiconductors.

In no event shall NXP Semiconductors be liable for any indirect, incidental, punitive, special or consequential damages (including - without limitation - lost profits, lost savings, business interruption, costs related to the removal or replacement of any products or rework charges) whether or not such damages are based on tort (including negligence), warranty, breach of contract or any other legal theory.

Notwithstanding any damages that customer might incur for any reason whatsoever, NXP Semiconductors' aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms and conditions of commercial sale of NXP Semiconductors.

Right to make changes — NXP Semiconductors reserves the right to make changes to information published in this document, including without limitation specifications and product descriptions, at any time and without notice. This document supersedes and replaces all information supplied prior to the publication hereof.

Suitability for use — NXP Semiconductors products are not designed, authorized or warranted to be suitable for use in life support, life-critical or safety-critical systems or equipment, nor in applications where failure or malfunction of an NXP Semiconductors product can reasonably be expected to result in personal injury, death or severe property or environmental damage. NXP Semiconductors and its suppliers accept no liability for inclusion and/or use of NXP Semiconductors products in such equipment or applications and therefore such inclusion and/or use is at the customer's own risk.

Applications — Applications that are described herein for any of these products are for illustrative purposes only. NXP Semiconductors makes no representation or warranty that such applications will be suitable for the specified use without further testing or modification.

Customers are responsible for the design and operation of their applications and products using NXP Semiconductors products, and NXP Semiconductors accepts no liability for any assistance with applications or customer product design. It is customer's sole responsibility to determine whether the NXP Semiconductors product is suitable and fit for the customer's applications and products planned, as well as for the planned application and use of customer's third party customer(s). Customers should provide appropriate design and operating safeguards to minimize the risks associated with their applications and products.

NXP Semiconductors does not accept any liability related to any default, damage, costs or problem which is based on any weakness or default in the customer's applications or products, or the application or use by customer's third party customer(s). Customer is responsible for doing all necessary testing for the customer's applications and products using NXP Semiconductors products in order to avoid a default of the applications and the products or of the application or use by customer's third party customer(s). NXP does not accept any liability in this respect.

Terms and conditions of commercial sale — NXP Semiconductors products are sold subject to the general terms and conditions of commercial sale, as published at <https://www.nxp.com/profile/terms>, unless otherwise agreed in a valid written individual agreement. In case an individual agreement is concluded only the terms and conditions of the respective agreement shall apply. NXP Semiconductors hereby expressly objects to applying the customer's general terms and conditions with regard to the purchase of NXP Semiconductors products by customer.

Export control — This document as well as the item(s) described herein may be subject to export control regulations. Export might require a prior authorization from competent authorities.

Suitability for use in non-automotive qualified products — Unless this document expressly states that this specific NXP Semiconductors product is automotive qualified, the product is not suitable for automotive use. It is neither qualified nor tested in accordance with automotive testing or application requirements. NXP Semiconductors accepts no liability for inclusion and/or use of non-automotive qualified products in automotive equipment or applications.

In the event that customer uses the product for design-in and use in automotive applications to automotive specifications and standards, customer (a) shall use the product without NXP Semiconductors' warranty of the product for such automotive applications, use and specifications, and (b) whenever customer uses the product for automotive applications beyond NXP Semiconductors' specifications such use shall be solely at customer's own risk, and (c) customer fully indemnifies NXP Semiconductors for any liability, damages or failed product claims resulting from customer design and use of the product for automotive applications beyond NXP Semiconductors' standard warranty and NXP Semiconductors' product specifications.

HTML publications — An HTML version, if available, of this document is provided as a courtesy. Definitive information is contained in the applicable document in PDF format. If there is a discrepancy between the HTML document and the PDF document, the PDF document has priority.

Translations — A non-English (translated) version of a document, including the legal information in that document, is for reference only. The English version shall prevail in case of any discrepancy between the translated and English versions.

Security — Customer understands that all NXP products may be subject to unidentified vulnerabilities or may support established security standards or specifications with known limitations. Customer is responsible for the design and operation of its applications and products throughout their lifecycles to reduce the effect of these vulnerabilities on customer's applications and products. Customer's responsibility also extends to other open and/or proprietary technologies supported by NXP products for use in customer's applications. NXP accepts no liability for any vulnerability. Customer should regularly check security updates from NXP and follow up appropriately. Customer shall select products with security features that best meet rules, regulations, and standards of the intended application and make the ultimate design decisions regarding its products and is solely responsible for compliance with all legal, regulatory, and security related requirements concerning its products, regardless of any information or support that may be provided by NXP.

NXP has a Product Security Incident Response Team (PSIRT) (reachable at PSIRT@nxp.com) that manages the investigation, reporting, and solution release to security vulnerabilities of NXP products.

NXP B.V. — NXP B.V. is not an operating company and it does not distribute or sell products.

Trademarks

Notice: All referenced brands, product names, service names, and trademarks are the property of their respective owners.

NXP — wordmark and logo are trademarks of NXP B.V.

Contents

1	Introduction	2	3	Vivante EGL and OGL Extension Support	22
1.1	i.MX full GPU line	2	3.1	Introduction	22
2	i.MX G2D API	2	3.2	EGL extension support	22
2.1	Overview	2	3.3	OpenGL ES extension support	26
2.2	Enumerations and structures	3	3.4	Extension GL_VIV_direct_texture	34
2.2.1	g2d_format enumeration	3	3.4.1	New Procedures and Functions	35
2.2.2	g2d_blend_func enumeration	4	3.5	Extension GL_VIV_texture_border_clamp	37
2.2.3	g2d_cap_mode enumeration	4	4	Vivante Framebuffer API	38
2.2.4	g2d_rotation enumeration	5	4.1	Overview	38
2.2.5	g2d_cache_mode enumeration	5	4.2	API data types and environment variables	39
2.2.6	g2d_hardware_type enumeration	5	4.2.1	Data types	39
2.2.7	g2d_surface structure	5	4.2.2	Environment variables	39
2.2.8	g2d_buf structure	7	4.3	API description and syntax	40
2.2.9	g2d_surface_pair structure	7	5	OpenCL	46
2.2.10	g2d_feature enumeration	7	5.1	Overview	46
2.2.11	g2d_tiling enumeration	8	5.1.1	General description	46
2.2.12	g2d_surfaceEx structure	8	5.1.2	OpenCL framework	47
2.2.13	g2d_warp_map_format enumeration	8	5.1.2.1	OpenCL execution model: kernels and work elements	47
2.2.14	g2d_warp_coordinates structure	9	5.1.2.2	OpenCL command queues	48
2.3	G2D function description	9	5.1.2.3	OpenCL memory model	49
2.3.1	g2d_open	9	5.1.2.4	Host to Vivante compute device data transfers	50
2.3.2	g2d_close	9	5.1.3	OpenCL profiles	51
2.3.3	g2d_make_current	10	5.1.4	Vivante OpenCL embedded compatible IP	51
2.3.4	g2d_clear	10	5.1.5	Vivante OpenCL full profile hardware model	52
2.3.5	g2d_blit	10	5.2	Vivante OpenCL implementation	53
2.3.6	g2d_copy	10	5.2.1	OpenCL pipeline	53
2.3.7	g2d_query_cap	11	5.2.2	Front end	54
2.3.8	g2d_enable	11	5.2.3	OpenCL compute unit	54
2.3.9	g2d_disable	11	5.2.4	Memory hierarchy	55
2.3.10	g2d_cache_op	11	5.2.5	CL Extension support	55
2.3.11	g2d_alloc	12	5.2.5.1	CL_DEVICE_EXTENSION support	55
2.3.12	g2d_free	12	5.2.5.2	Vivante OpenCL extension support	56
2.3.13	g2d_flush	12	5.3	Optimization for OpenCL embedded profile	57
2.3.14	g2d_finish	12	5.3.1	Using preferred multiple of work-group size	57
2.3.15	g2d_multi_blit	12	5.3.2	Using multiple work-groups of reduced size	57
2.3.16	g2d_query_hardware	13	5.3.3	Packing work-item data	57
2.3.17	g2d_query_feature	13	5.3.4	Improving locality	58
2.3.18	g2d_blitEx	14	5.3.5	Minimizing use of 1 KB local memory	58
2.3.19	g2d_set_clipping	14	5.3.6	Using 16 byte memory Read/Write size	58
2.3.20	g2d_set_csc_matrix	14	5.3.7	Using _RTZ rounding mode	58
2.3.21	g2d_buf_from_fd	14	5.3.8	Using float4 for better performance on i.MX 8M Quad and i.MX 8QuadXPlus	58
2.3.22	g2d_buf_export_fd	15	5.3.9	Using native functions	58
2.3.23	g2d_buf_from_virt_addr	15	5.3.9.1	Using native_function() for increased performance	58
2.3.24	g2d_create_fence_fd	15	5.3.9.2	Using native_divide and native_reciprocal for faster floating point calculations	59
2.3.25	g2d_set_warp_coordinates	15	5.3.9.3	Using compile option for native functions	59
2.4	Support of new operating system in G2D	16	5.3.10	Using buffers instead of images	59
2.5	Sample code for G2D API usage	16	5.4	OpenCL Debug messages	59
2.5.1	Color space conversion from YUV to RGB	16	5.4.1	OCL-007005: (clCreateKernel) cannot link kernel	59
2.5.2	Alpha blend in source over mode	17	5.4.2	Not enough register memory	60
2.5.3	Source cropping and destination rotation	17			
2.5.4	Multi source blit	18			
2.5.5	Sharing Buffers between APIs using G2D Buffers:	18			
2.5.6	Warp/Dewarp	19			
2.6	Feature list on multiple platforms	21			
2.7	Arbitrary Warping	21			

5.4.3	Not enough instruction memory	60	7.3.4	Channel Data Types Supported	78
5.4.4	GlobalWorkSize over hardware limit	60	7.3.5	Image Channel Orders Supported	79
5.5	Zero copy	60	8	Vulkan	79
5.6	Instruction cache availability for i.MX graphics	61	8.1	Overview	79
6	OpenCV	61	8.2	Vulkan Validation Layers	80
6.1	Overview	61	8.3	Window System Integration	80
6.2	Acceleration with OpenCL	61	9	Vivante Multiple GPUs and Virtualization	80
6.3	Usages of OpenCV Accelerator	62	9.1	Overview	80
6.3.1	How to enable/disable OpenCV Accelerator	62	9.2	Multi-GPU configurations	80
6.3.2	Requirements	62	9.3	GPU affinity configuration	81
6.4	OpenCV functions accelerated with OpenCL	62	9.4	OpenCL on multi-GPU device	81
6.4.1	OpenCV function list	63	9.5	GPU virtualization configuration	81
6.4.2	Conditions to use the accelerator	64	10	GBM - Generic Buffer Management	82
6.4.2.1	pyrUP	64	10.1	Introduction to DRM Format Modifiers	82
6.4.2.2	warpPerspective	64	11	Wayland and Weston	83
6.4.2.3	warpAffine	64	11.1	Overview	83
6.4.2.4	match Template	65	11.2	Wayland EGL	83
6.4.2.5	resize	65	11.3	Weston compositor	83
6.4.2.6	Threshold	65	11.3.1	Weston backends	83
6.4.2.7	Sobel	66	11.3.1.1	RDP backend	83
6.4.2.8	filter2D	66	11.3.2	Weston renderer	83
6.4.2.9	morphologyEX	66	11.3.2.1	GL renderer	83
6.4.2.10	erode	67	11.3.2.2	G2D renderer	84
6.4.2.11	dilate	67	11.3.3	Weston shells	84
6.4.2.12	GaussianBlur	67	11.3.3.1	Desktop shell	84
6.4.2.13	Blur	68	11.3.3.2	Fullscreen shell	84
6.4.2.14	sqrBoxFilter	68	11.3.3.3	IVI-shell	84
6.4.2.15	remap	68	12	X Windowing Acceleration	84
6.4.2.16	Laplacian	69	13	Advanced GPU Configuration	85
6.4.2.17	Scharr	69	13.1	GPU Scaling Governor	85
6.4.2.18	sepFilter2D	69	13.2	GPU Device Cooling	85
6.4.2.19	calcHist	70	13.3	i.MX 95 GPU frequency scaling	85
6.4.2.20	accumulate	70	13.3.1	simple_ondemand governor	86
6.4.2.21	accumulateProduct	71	14	Vivante IDE	86
6.4.2.22	accumulateWeighted	71	14.1	VivanteIDE overview	86
6.4.2.23	cornerMinEigenVal	71	14.1.1	VivanteIDE component overview	86
6.4.2.24	cornerHarris	72	14.2	VivanteIDE Requirements	87
6.4.2.25	preCornerDetect	72	14.2.1	Operating system compatibility	87
6.4.2.26	HoughLines	72	14.2.2	Hardware requirements	87
6.4.2.27	HoughLinesP	73	14.2.3	VivanteIDE license	87
6.4.2.28	goodFeaturesToTrack	73	14.3	VivanteIDE installation	88
6.5	Performance differences of OpenCV on Arm GPU and VSI GPU	73	14.3.1	VivanteIDE package	88
7	OpenVX Introduction	74	14.3.2	Installation	88
7.1	Overview	74	14.3.2.1	Linux GUI	88
7.2	OpenVX extension implementation	74	14.3.2.2	Windows GUI	88
7.2.1	Hardware requirements	74	14.3.2.3	Installation from command line	89
7.2.2	EVIS instruction interface	74	14.3.3	VivanteIDE launch	89
7.2.3	Extended language features	75	14.3.3.1	Linux launch of GUI tool	89
7.2.4	Packed types	75	14.3.3.2	Windows launch of GUI tool	89
7.2.5	Initializing constants on load	76	14.3.3.3	Command line tool launch	89
7.2.6	Inline assembly	76	14.3.3.4	Basic launch path summary	89
7.3	OpenCL functions compatible with Vivante vision	78	14.4	VivanteIDE GUI	90
7.3.1	Read_Imagef,i,ui	78	14.4.1	Selecting a workspace	90
7.3.2	Write_Imagef,i,ui	78	14.4.2	Switching perspective	91
7.3.3	Query Image Dimensions	78	14.4.3	Creating a new project	91
			14.4.4	Creating an OpenVX kernel wizard	92
			14.4.5	Source code smart editing for OpenVX and OpenCL	94

14.4.6	Creating a Neural Network Inference Project from a model file	95	14.8.4.5	vTexture Syntax Examples	124
14.4.7	Building a sample project	101	15	GPU Tools	125
14.4.8	Debugging and profiling a project	104	15.1	gputop tool	125
14.5	VivanteIDE – Debug and Profiling	105	15.1.1	Synopsis	125
14.5.1	Fundamentals of performance optimization ...	105	15.1.2	Interactive mode	125
14.5.2	VPD Analyzer for Analyzing Performance Data	106	15.1.3	Description	126
14.5.3	vProfiler	106	15.1.4	Requirements	126
14.5.4	Enabling vProfiler on Linux OS	106	15.1.4.1	Linux OS	126
14.5.4.1	Setting vProfiler property options for OpenGL ES	106	15.1.4.2	QNX	126
14.5.5	Setting vProfiler property options for Vision, OpenVX Profiling	106	15.1.5	Notes	126
14.5.6	Enabling vProfiler Option for Android OS	107	15.1.5.1	Sampling hardware-counters	126
14.5.7	Setting vProfiler property options for OpenGL ES Profiling with Android	107	15.1.5.2	Context-aware counters	126
14.5.8	vProfiler Set Property Options for Vision/ OVX Profiling with Android	108	15.1.5.3	Unsupported GPUs	127
14.5.9	Enabling vProfiler Option for QNX	109	15.1.6	Pages for VSI GPUs	127
14.5.9.1	Setting vProfiler Environment Variables for OGL/OES Profiling	109	15.1.6.1	Client attached page	127
14.5.9.2	Setting vProfiler Environment Variables for Vision, OpenVX Profiling	110	15.1.6.2	Vidmem page	127
14.5.10	Environment Variable Details	110	15.1.7	Pages for Mali GPU	127
14.5.10.1	VIV_PROFILE	110	15.1.7.1	Page0: Main Page	128
14.5.10.2	VP_OUTPUT	111	15.1.7.2	Page1: GPU INFO	128
14.5.10.3	VP_USE_GLFINISH	111	15.1.7.3	Page2: Kernel Memory Usage	128
14.5.10.4	VP_DISABLE_PROBE	111	15.1.7.4	Page3: PID-Based Process Memory Usage ..	128
14.5.10.5	VP_ENABLE_PRINT	111	15.1.7.5	Page4: GPU Core Utilization	129
14.6	VPD Analyzer	111	15.1.7.6	Page5: Perf DDR Memory Bandwidth	129
14.6.1	Loading a VPD File	112	15.1.8	Examples	129
14.6.2	VPD Analyzer Perspective	113	15.1.9	See Also	129
14.6.3	System Info View	114	15.2	GPU clock information and debugging	130
14.6.4	Program Counters View	115	15.3	Aptirace user guide	130
14.6.5	Closing the VPD File	115	15.3.1	Introduction	130
14.7	SPIR-V Disassembler	115	15.3.2	Install	130
14.7.1	Shader Assistant	116	15.3.2.1	Yocto	130
14.7.2	vTexture	116	15.3.2.2	PC	131
14.8	VivanteIDE command line tools	118	15.3.3	Usage	131
14.8.1	Preparing the environment	118	15.3.3.1	Trace OpenGL ES1.1/2.0/3.0 application	131
14.8.2	vCompiler Command Line Syntax for OGL and OGLES	118	15.3.3.2	Trace OpenGL ES 1.1/2.0/3.0 Java application on the Android platform	131
14.8.2.1	Syntax	118	15.3.3.3	Trace OpenGL application	131
14.8.2.2	Input parameters (required)	118	15.3.3.4	Replay	131
14.8.2.3	Input parameters (optional)	118	15.3.4	Reference	135
14.8.2.4	vCompilerOutput	120	15.4	Renderdoc	135
14.8.2.5	vCompiler Syntax examples	120	15.4.1	Renderdoc components	135
14.8.3	vcCompiler Command Line Syntax for OCL ..	120	15.4.2	Running renderdoccmd on i.MX	136
14.8.3.1	Syntax	120	15.4.3	Capturing and replaying remotely	137
14.8.3.2	Input parameters (required)	120	15.4.4	Reference	138
14.8.3.3	Input parameters (optional)	121	16	VSI GPU Memory Introduction	139
14.8.3.4	vcCompiler Output	122	16.1	VSI GPU memory overview	139
14.8.3.5	vcCompiler Syntax Examples	122	16.2	VSI GPU memory pools	139
14.8.4	vTextureTools command line tool	122	16.3	VSI GPU memory allocators	139
14.8.4.1	Syntax	122	16.4	VSI GPU reserved memory	140
14.8.4.2	General parameters	122	16.5	VSI GPU memory base address	140
14.8.4.3	Compression/Decompression parameters	123	17	Mali Valhall GPU	141
14.8.4.4	Tile/De-Tile parameters	123	17.1	Features	141
			17.2	Mali Shader offline Compiler	142
			17.3	Mali OpenCL Offline Compiler	142
			17.4	Capture the GLES application with patrace ...	143
			17.5	Arm Performance Studio	143
			17.5.1	Tools included in Arm Performance Studio	143
			17.5.2	Streamline offline capture	144
			17.6	References and Useful links	144

18	Application Programming Recommendations	145	19	Demo Framework	151
18.1	Understanding the system configuration and target application	145	20	Environment Variables Summary	151
18.2	Optimizing off-chip data transfer such as accessing off-chip DDR memory/mobile DDR memory	145	20.1	Environment variable for drivers and HAL	151
18.3	Avoiding W-clipping issue in the application program	145	20.2	Environment variable for compiler	153
18.4	Avoiding GPU hanging and data corruption when using occlusion query	146	21	Note About the Source Code in the Document	153
18.5	Avoiding random cache or memory access ...	146	22	Revision History	154
18.6	Optimizing your use of system memory	146		Legal information	155
18.7	Targeting a fixed frame rate that is visibly smooth	146			
18.8	Minimizing GL state changes	146			
18.9	Batch primitives to minimize the number of draw calls	147			
18.10	Performing calculations per vertex instead of per fragment/pixel	147			
18.11	Enabling early-Z, hierarchical-Z, and back face culling	147			
18.12	Using branching carefully	147			
18.13	Using VBOs instead of static or stack data as vertex data	147			
18.14	Using dynamic VBO when the data is changing frame by frame	148			
18.15	Tessellating your data to make Hierarchical Z (HZ) work	148			
18.16	Using dynamic textures as a texture cache (texture atlas)	148			
18.17	Stitching small triangle strips together	149			
18.18	Specifying EGL configuration attributes precisely	149			
18.19	Using aligned texture/render buffers	149			
18.20	Disabling MSAA rendering unless high quality is needed	149			
18.21	Avoiding partial clears	149			
18.22	Avoiding mask operations	149			
18.23	Using MIPMAP textures	149			
18.24	Using compressed textures if constricted by RAM/ROM budget	149			
18.25	Drawing objects from near to far if possible ...	150			
18.26	Avoiding indexed triangle strips	150			
18.27	Limiting vertex attribute stride within 256 bytes	150			
18.28	Avoiding binding buffers to mixed index/vertex array	150			
18.29	Avoiding using CPU to update texture/buffer contexts during render	150			
18.30	Avoiding frequent context switching	150			
18.31	Optimizing resources within a shader	150			
18.32	Avoiding using glScissor Clear for small regions	150			
18.33	Using PRE to accelerate data transfer	151			
18.34	i.MX 8QuadMax dual-GPU performance	151			

Please be aware that important notices concerning this document and the product(s) described herein, have been included in section 'Legal information'.