



CUDA DRIVER API

TRM-06703-001 _vRelease Version | July 2017

API Reference Manual



TABLE OF CONTENTS

Chapter 1. Difference between the driver and runtime APIs.....	1
Chapter 2. API synchronization behavior.....	3
Chapter 3. Stream synchronization behavior.....	5
Chapter 4. Modules.....	7
4.1. Data types used by CUDA driver.....	8
CUDA_ARRAY3D_DESCRIPTOR.....	9
CUDA_ARRAY_DESCRIPTOR.....	9
CUDA_LAUNCH_PARAMS.....	9
CUDA_MEMCPY2D.....	9
CUDA_MEMCPY3D.....	9
CUDA_MEMCPY3D_PEER.....	9
CUDA_POINTER_ATTRIBUTE_P2P_TOKENS.....	9
CUDA_RESOURCE_DESC.....	9
CUDA_RESOURCE_VIEW_DESC.....	9
CUDA_TEXTURE_DESC.....	9
CUdevprop.....	9
CUeglFrame.....	9
CUipcEventHandle.....	9
CUipcMemHandle.....	9
CUstreamBatchMemOpParams.....	9
CUaddress_mode.....	9
CUarray_cubemap_face.....	10
CUarray_format.....	10
CUcomputemode.....	11
CUctx_flags.....	11
CUdevice_attribute.....	12
CUdevice_P2PAttribute.....	17
CUeglColorFormat.....	17
CUeglFrameType.....	22
CUeglResourceLocationFlags.....	22
CUevent_flags.....	23
CUfilter_mode.....	23
CUfunc_cache.....	23
CUfunction_attribute.....	24
CUgraphicsMapResourceFlags.....	25
CUgraphicsRegisterFlags.....	25
CUipcMem_flags.....	25
CUjit_cacheMode.....	25
CUjit_fallback.....	26
CUjit_option.....	26

CUjit_target.....	28
CUjitInputType.....	28
CULimit.....	29
CUMemAdvise.....	29
CUMemAttach_flags.....	30
CUMemoryType.....	30
CUOccupancy_flags.....	30
CUPointer_attribute.....	31
CUResourceType.....	31
CUResourceViewFormat.....	32
CUResult.....	33
CUshared_carveout.....	38
CUsharedconfig.....	38
CUstream_flags.....	39
CUstreamBatchMemOpType.....	39
CUstreamWaitValue_flags.....	39
CUstreamWriteValue_flags.....	40
CUArray.....	40
CUContext.....	40
CUDevice.....	40
CUDevicePtr.....	41
CUeglStreamConnection.....	41
CUEvent.....	41
CUFunction.....	41
CUgraphicsResource.....	41
CUmipmappedArray.....	41
CUmodule.....	41
CUOccupancyB2DSize.....	41
CUstream.....	41
CUstreamCallback.....	42
CUsurfObject.....	42
CUsurfref.....	42
CUtexObject.....	42
CUtexref.....	42
CU_DEVICE_CPU.....	42
CU_DEVICE_INVALID.....	42
CU_IPC_HANDLE_SIZE.....	42
CU_LAUNCH_PARAM_BUFFER_POINTER.....	42
CU_LAUNCH_PARAM_BUFFER_SIZE.....	43
CU_LAUNCH_PARAM_END.....	43
CU_MEMHOSTALLOC_DEVICEMAP.....	43
CU_MEMHOSTALLOC_PORTABLE.....	43
CU_MEMHOSTALLOC_WRITECOMBINED.....	43

CU_MEMHOSTREGISTER_DEVICEMAP.....	43
CU_MEMHOSTREGISTER_IOMEMORY.....	43
CU_MEMHOSTREGISTER_PORTABLE.....	44
CU_PARAM_TR_DEFAULT.....	44
CU_STREAM_LEGACY.....	44
CU_STREAM_PER_THREAD.....	44
CU_TRSA_OVERRIDE_FORMAT.....	44
CU_TRSF_NORMALIZED_COORDINATES.....	44
CU_TRSF_READ_AS_INTEGER.....	44
CU_TRSF_SRGB.....	44
CUDA_ARRAY3D_2DARRAY.....	45
CUDA_ARRAY3D_CUBEMAP.....	45
CUDA_ARRAY3D_DEPTH_TEXTURE.....	45
CUDA_ARRAY3D_LAYERED.....	45
CUDA_ARRAY3D_SURFACE_LDST.....	45
CUDA_ARRAY3D_TEXTURE_GATHER.....	45
CUDA_COOPERATIVE_LAUNCH_MULTI_DEVICE_NO_POST_LAUNCH_SYNC.....	45
CUDA_COOPERATIVE_LAUNCH_MULTI_DEVICE_NO_PRE_LAUNCH_SYNC.....	46
CUDA_VERSION.....	46
MAX_PLANES.....	46
4.2. Error Handling.....	46
cuGetErrorName.....	46
cuGetErrorString.....	47
4.3. Initialization.....	47
culnit.....	47
4.4. Version Management.....	48
cuDriverGetVersion.....	48
4.5. Device Management.....	49
cuDeviceGet.....	49
cuDeviceGetAttribute.....	50
cuDeviceGetCount.....	55
cuDeviceGetName.....	56
cuDeviceGetUuid.....	56
cuDeviceTotalMem.....	57
4.6. Device Management [DEPRECATED].....	58
cuDeviceComputeCapability.....	58
cuDeviceGetProperties.....	59
4.7. Primary Context Management.....	60
cuDevicePrimaryCtxGetState.....	60
cuDevicePrimaryCtxRelease.....	61
cuDevicePrimaryCtxReset.....	62
cuDevicePrimaryCtxRetain.....	63
cuDevicePrimaryCtxSetFlags.....	64

4.8. Context Management.....	65
cuCtxCreate.....	65
cuCtxDestroy.....	67
cuCtxGetApiVersion.....	68
cuCtxGetCacheConfig.....	69
cuCtxGetCurrent.....	70
cuCtxGetDevice.....	70
cuCtxGetFlags.....	71
cuCtxGetLimit.....	72
cuCtxGetSharedMemConfig.....	73
cuCtxGetStreamPriorityRange.....	74
cuCtxPopCurrent.....	75
cuCtxPushCurrent.....	75
cuCtxSetCacheConfig.....	76
cuCtxSetCurrent.....	77
cuCtxSetLimit.....	78
cuCtxSetSharedMemConfig.....	80
cuCtxSynchronize.....	81
4.9. Context Management [DEPRECATED].....	81
cuCtxAttach.....	81
cuCtxDetach.....	82
4.10. Module Management.....	83
cuLinkAddData.....	83
cuLinkAddFile.....	84
cuLinkComplete.....	85
cuLinkCreate.....	86
cuLinkDestroy.....	87
cuModuleGetFunction.....	87
cuModuleGetGlobal.....	88
cuModuleGetSurfRef.....	89
cuModuleGetTexRef.....	90
cuModuleLoad.....	91
cuModuleLoadData.....	92
cuModuleLoadDataEx.....	93
cuModuleLoadFatBinary.....	94
cuModuleUnload.....	95
4.11. Memory Management.....	95
cuArray3DCreate.....	95
cuArray3DGetDescriptor.....	99
cuArrayCreate.....	100
cuArrayDestroy.....	102
cuArrayGetDescriptor.....	103
cuDeviceGetByPCIBusId.....	104

cuDeviceGetPCIBusId.....	104
culpcCloseMemHandle.....	105
culpcGetEventHandle.....	106
culpcGetMemHandle.....	107
culpcOpenEventHandle.....	108
culpcOpenMemHandle.....	109
cuMemAlloc.....	110
cuMemAllocHost.....	111
cuMemAllocManaged.....	112
cuMemAllocPitch.....	115
cuMemcpy.....	116
cuMemcpy2D.....	117
cuMemcpy2DAsync.....	120
cuMemcpy2DUnaligned.....	123
cuMemcpy3D.....	126
cuMemcpy3DAsync.....	129
cuMemcpy3DPeer.....	132
cuMemcpy3DPeerAsync.....	133
cuMemcpyAsync.....	134
cuMemcpyAtoA.....	135
cuMemcpyAtoD.....	136
cuMemcpyAtoH.....	137
cuMemcpyAtoHAsync.....	138
cuMemcpyDtoA.....	139
cuMemcpyDtoD.....	140
cuMemcpyDtoDAsync.....	141
cuMemcpyDtoH.....	143
cuMemcpyDtoHAsync.....	144
cuMemcpyHtoA.....	145
cuMemcpyHtoAAsync.....	146
cuMemcpyHtoD.....	147
cuMemcpyHtoDAsync.....	148
cuMemcpyPeer.....	149
cuMemcpyPeerAsync.....	150
cuMemFree.....	151
cuMemFreeHost.....	152
cuMemGetAddressRange.....	153
cuMemGetInfo.....	154
cuMemHostAlloc.....	155
cuMemHostGetDevicePointer.....	157
cuMemHostGetFlags.....	158
cuMemHostRegister.....	159
cuMemHostUnregister.....	161

cuMemsetD16.....	161
cuMemsetD16Async.....	162
cuMemsetD2D16.....	163
cuMemsetD2D16Async.....	165
cuMemsetD2D32.....	166
cuMemsetD2D32Async.....	167
cuMemsetD2D8.....	168
cuMemsetD2D8Async.....	170
cuMemsetD32.....	171
cuMemsetD32Async.....	172
cuMemsetD8.....	173
cuMemsetD8Async.....	174
cuMipmappedArrayCreate.....	175
cuMipmappedArrayDestroy.....	178
cuMipmappedArrayGetLevel.....	179
4.12. Unified Addressing.....	180
cuMemAdvise.....	181
cuMemPrefetchAsync.....	185
cuMemRangeGetAttribute.....	186
cuMemRangeGetAttributes.....	188
cuPointerGetAttribute.....	189
cuPointerGetAttributes.....	192
cuPointerSetAttribute.....	193
4.13. Stream Management.....	194
cuStreamAddCallback.....	195
cuStreamAttachMemAsync.....	196
cuStreamCreate.....	198
cuStreamCreateWithPriority.....	199
cuStreamDestroy.....	200
cuStreamGetCtx.....	201
cuStreamGetFlags.....	202
cuStreamGetPriority.....	203
cuStreamQuery.....	203
cuStreamSynchronize.....	204
cuStreamWaitEvent.....	205
4.14. Event Management.....	206
cuEventCreate.....	206
cuEventDestroy.....	207
cuEventElapsedTime.....	207
cuEventQuery.....	209
cuEventRecord.....	209
cuEventSynchronize.....	210
4.15. Stream memory operations.....	211

cuStreamBatchMemOp.....	212
cuStreamWaitValue32.....	213
cuStreamWaitValue64.....	214
cuStreamWriteValue32.....	215
cuStreamWriteValue64.....	216
4.16. Execution Control.....	217
cuFuncGetAttribute.....	217
cuFuncSetAttribute.....	218
cuFuncSetCacheConfig.....	220
cuFuncSetSharedMemConfig.....	221
cuLaunchCooperativeKernel.....	222
cuLaunchCooperativeKernelMultiDevice.....	224
cuLaunchKernel.....	227
4.17. Execution Control [DEPRECATED].....	230
cuFuncSetBlockShape.....	230
cuFuncSetSharedSize.....	231
cuLaunch.....	232
cuLaunchGrid.....	233
cuLaunchGridAsync.....	234
cuParamSetf.....	235
cuParamSeti.....	236
cuParamSetSize.....	236
cuParamSetTexRef.....	237
cuParamSetv.....	238
4.18. Occupancy.....	239
cuOccupancyMaxActiveBlocksPerMultiprocessor.....	239
cuOccupancyMaxActiveBlocksPerMultiprocessorWithFlags.....	240
cuOccupancyMaxPotentialBlockSize.....	241
cuOccupancyMaxPotentialBlockSizeWithFlags.....	242
4.19. Texture Reference Management.....	243
cuTexRefGetAddress.....	244
cuTexRefGetAddressMode.....	244
cuTexRefGetArray.....	245
cuTexRefGetBorderColor.....	246
cuTexRefGetFilterMode.....	246
cuTexRefGetFlags.....	247
cuTexRefGetFormat.....	248
cuTexRefGetMaxAnisotropy.....	248
cuTexRefGetMipmapFilterMode.....	249
cuTexRefGetMipmapLevelBias.....	250
cuTexRefGetMipmapLevelClamp.....	250
cuTexRefGetMipmappedArray.....	251
cuTexRefSetAddress.....	252

cuTexRefSetAddress2D.....	253
cuTexRefSetAddressMode.....	254
cuTexRefSetArray.....	255
cuTexRefSetBorderColor.....	256
cuTexRefSetFilterMode.....	257
cuTexRefSetFlags.....	257
cuTexRefSetFormat.....	258
cuTexRefSetMaxAnisotropy.....	259
cuTexRefSetMipmapFilterMode.....	260
cuTexRefSetMipmapLevelBias.....	260
cuTexRefSetMipmapLevelClamp.....	261
cuTexRefSetMipmappedArray.....	262
4.20. Texture Reference Management [DEPRECATED].....	263
cuTexRefCreate.....	263
cuTexRefDestroy.....	263
4.21. Surface Reference Management.....	264
cuSurfRefGetArray.....	264
cuSurfRefSetArray.....	265
4.22. Texture Object Management.....	265
cuTexObjectCreate.....	265
cuTexObjectDestroy.....	271
cuTexObjectGetResourceDesc.....	271
cuTexObjectGetResourceViewDesc.....	272
cuTexObjectGetTextureDesc.....	272
4.23. Surface Object Management.....	273
cuSurfObjectCreate.....	273
cuSurfObjectDestroy.....	274
cuSurfObjectGetResourceDesc.....	274
4.24. Peer Context Memory Access.....	275
cuCtxDisablePeerAccess.....	275
cuCtxEnablePeerAccess.....	276
cuDeviceCanAccessPeer.....	277
cuDeviceGetP2PAttribute.....	278
4.25. Graphics Interoperability.....	279
cuGraphicsMapResources.....	279
cuGraphicsResourceGetMappedMipmappedArray.....	280
cuGraphicsResourceGetMappedPointer.....	281
cuGraphicsResourceSetMapFlags.....	282
cuGraphicsSubResourceGetMappedArray.....	283
cuGraphicsUnmapResources.....	284
cuGraphicsUnregisterResource.....	285
4.26. Profiler Control.....	286
cuProfilerInitialize.....	286

cuProfilerStart.....	287
cuProfilerStop.....	287
4.27. OpenGL Interoperability.....	288
OpenGL Interoperability [DEPRECATED].....	288
CUGLDeviceList.....	288
cuGLGetDevices.....	288
cuGraphicsGLRegisterBuffer.....	290
cuGraphicsGLRegisterImage.....	291
cuWGLGetDevice.....	292
4.27.1. OpenGL Interoperability [DEPRECATED].....	293
CUGLmap_flags.....	293
cuGLCtxCreate.....	293
cuGLInit.....	294
cuGLMapBufferObject.....	295
cuGLMapBufferObjectAsync.....	296
cuGLRegisterBufferObject.....	297
cuGLSetBufferObjectMapFlags.....	297
cuGLUnmapBufferObject.....	298
cuGLUnmapBufferObjectAsync.....	299
cuGLUnregisterBufferObject.....	300
4.28. Direct3D 9 Interoperability.....	301
Direct3D 9 Interoperability [DEPRECATED].....	301
CUd3d9DeviceList.....	301
cuD3D9CtxCreate.....	301
cuD3D9CtxCreateOnDevice.....	302
cuD3D9GetDevice.....	303
cuD3D9GetDevices.....	304
cuD3D9GetDirect3DDevice.....	305
cuGraphicsD3D9RegisterResource.....	306
4.28.1. Direct3D 9 Interoperability [DEPRECATED].....	308
CUd3d9map_flags.....	308
CUd3d9register_flags.....	308
cuD3D9MapResources.....	309
cuD3D9RegisterResource.....	310
cuD3D9ResourceGetMappedArray.....	312
cuD3D9ResourceGetMappedPitch.....	313
cuD3D9ResourceGetMappedPointer.....	314
cuD3D9ResourceGetMappedSize.....	315
cuD3D9ResourceGetSurfaceDimensions.....	316
cuD3D9ResourceSetMapFlags.....	318
cuD3D9UnmapResources.....	319
cuD3D9UnregisterResource.....	320
4.29. Direct3D 10 Interoperability.....	320

Direct3D 10 Interoperability [DEPRECATED].....	320
CUd3d10DeviceList.....	320
cuD3D10GetDevice.....	321
cuD3D10GetDevices.....	321
cuGraphicsD3D10RegisterResource.....	323
4.29.1. Direct3D 10 Interoperability [DEPRECATED].....	325
CUD3D10map_flags.....	325
CUD3D10register_flags.....	325
cuD3D10CtxCreate.....	325
cuD3D10CtxCreateOnDevice.....	326
cuD3D10GetDirect3DDevice.....	327
cuD3D10MapResources.....	328
cuD3D10RegisterResource.....	329
cuD3D10ResourceGetMappedArray.....	330
cuD3D10ResourceGetMappedPitch.....	331
cuD3D10ResourceGetMappedPointer.....	333
cuD3D10ResourceGetMappedSize.....	334
cuD3D10ResourceGetSurfaceDimensions.....	335
cuD3D10ResourceSetMapFlags.....	336
cuD3D10UnmapResources.....	337
cuD3D10UnregisterResource.....	338
4.30. Direct3D 11 Interoperability.....	338
Direct3D 11 Interoperability [DEPRECATED].....	339
CUd3d11DeviceList.....	339
cuD3D11GetDevice.....	339
cuD3D11GetDevices.....	340
cuGraphicsD3D11RegisterResource.....	341
4.30.1. Direct3D 11 Interoperability [DEPRECATED].....	343
cuD3D11CtxCreate.....	343
cuD3D11CtxCreateOnDevice.....	344
cuD3D11GetDirect3DDevice.....	345
4.31. VDPAU Interoperability.....	345
cuGraphicsVDPAURegisterOutputSurface.....	346
cuGraphicsVDPAURegisterVideoSurface.....	347
cuVDPAUCtxCreate.....	348
cuVDPAUGetDevice.....	349
4.32. EGL Interoperability.....	350
cuEGLStreamConsumerAcquireFrame.....	350
cuEGLStreamConsumerConnect.....	351
cuEGLStreamConsumerConnectWithFlags.....	351
cuEGLStreamConsumerDisconnect.....	352
cuEGLStreamConsumerReleaseFrame.....	353
cuEGLStreamProducerConnect.....	353

cuEGLStreamProducerDisconnect.....	354
cuEGLStreamProducerPresentFrame.....	355
cuEGLStreamProducerReturnFrame.....	356
cuEventCreateFromEGLSync.....	356
cuGraphicsEGLRegisterImage.....	357
cuGraphicsResourceGetMappedEglFrame.....	359
Chapter 5. Data Structures.....	361
CUDA_ARRAY3D_DESCRIPTOR.....	361
Depth.....	361
Flags.....	361
Format.....	362
Height.....	362
NumChannels.....	362
Width.....	362
CUDA_ARRAY_DESCRIPTOR.....	362
Format.....	362
Height.....	362
NumChannels.....	362
Width.....	362
CUDA_LAUNCH_PARAMS.....	362
blockDimX.....	363
blockDimY.....	363
blockDimZ.....	363
function.....	363
gridDimX.....	363
gridDimY.....	363
gridDimZ.....	363
hStream.....	363
kernelParams.....	363
sharedMemBytes.....	363
CUDA_MEMCPY2D.....	363
dstArray.....	364
dstDevice.....	364
dstHost.....	364
dstMemoryType.....	364
dstPitch.....	364
dstXInBytes.....	364
dstY.....	364
Height.....	364
srcArray.....	364
srcDevice.....	364
srcHost.....	364
srcMemoryType.....	365

srcPitch.....	365
srcXInBytes.....	365
srcY.....	365
WidthInBytes.....	365
CUDA_MEMCPY3D.....	365
Depth.....	365
dstArray.....	365
dstDevice.....	365
dstHeight.....	365
dstHost.....	365
dstLOD.....	366
dstMemoryType.....	366
dstPitch.....	366
dstXInBytes.....	366
dstY.....	366
dstZ.....	366
Height.....	366
reserved0.....	366
reserved1.....	366
srcArray.....	366
srcDevice.....	366
srcHeight.....	367
srcHost.....	367
srcLOD.....	367
srcMemoryType.....	367
srcPitch.....	367
srcXInBytes.....	367
srcY.....	367
srcZ.....	367
WidthInBytes.....	367
CUDA_MEMCPY3D_PEER.....	367
Depth.....	367
dstArray.....	368
dstContext.....	368
dstDevice.....	368
dstHeight.....	368
dstHost.....	368
dstLOD.....	368
dstMemoryType.....	368
dstPitch.....	368
dstXInBytes.....	368
dstY.....	368
dstZ.....	368

Height.....	369
srcArray.....	369
srcContext.....	369
srcDevice.....	369
srcHeight.....	369
srcHost.....	369
srcLOD.....	369
srcMemoryType.....	369
srcPitch.....	369
srcXInBytes.....	369
srcY.....	369
srcZ.....	370
WidthInBytes.....	370
CUDA_POINTER_ATTRIBUTE_P2P_TOKENS.....	370
CUDA_RESOURCE_DESC.....	370
devPtr.....	370
flags.....	370
format.....	370
hArray.....	370
height.....	370
hMipmappedArray.....	370
numChannels.....	371
pitchInBytes.....	371
resType.....	371
sizeInBytes.....	371
width.....	371
CUDA_RESOURCE_VIEW_DESC.....	371
depth.....	371
firstLayer.....	371
firstMipmapLevel.....	371
format.....	371
height.....	372
lastLayer.....	372
lastMipmapLevel.....	372
width.....	372
CUDA_TEXTURE_DESC.....	372
addressMode.....	372
borderColor.....	372
filterMode.....	372
flags.....	372
maxAnisotropy.....	372
maxMipmapLevelClamp.....	373
minMipmapLevelClamp.....	373

mipmapFilterMode.....	373
mipmapLevelBias.....	373
CUdevprop.....	373
clockRate.....	373
maxGridSize.....	373
maxThreadsDim.....	373
maxThreadsPerBlock.....	373
memPitch.....	373
regsPerBlock.....	374
sharedMemPerBlock.....	374
SIMDWidth.....	374
textureAlign.....	374
totalConstantMemory.....	374
CUipcEventHandle.....	374
CUipcMemHandle.....	374
CUstreamBatchMemOpParams.....	374
5.12. Difference between the driver and runtime APIs.....	374
Chapter 6. Data Fields.....	376
Chapter 7. Deprecated List.....	383

Chapter 1.

DIFFERENCE BETWEEN THE DRIVER AND RUNTIME APIS

The driver and runtime APIs are very similar and can for the most part be used interchangeably. However, there are some key differences worth noting between the two.

Complexity vs. control

The runtime API eases device code management by providing implicit initialization, context management, and module management. This leads to simpler code, but it also lacks the level of control that the driver API has.

In comparison, the driver API offers more fine-grained control, especially over contexts and module loading. Kernel launches are much more complex to implement, as the execution configuration and kernel parameters must be specified with explicit function calls. However, unlike the runtime, where all the kernels are automatically loaded during initialization and stay loaded for as long as the program runs, with the driver API it is possible to only keep the modules that are currently needed loaded, or even dynamically reload modules. The driver API is also language-independent as it only deals with cubin objects.

Context management

Context management can be done through the driver API, but is not exposed in the runtime API. Instead, the runtime API decides itself which context to use for a thread: if a context has been made current to the calling thread through the driver API, the runtime will use that, but if there is no such context, it uses a "primary context." Primary contexts are created as needed, one per device per process, are reference-counted, and are then destroyed when there are no more references to them. Within one process, all users of the runtime API will share the primary context, unless a context has been made current to each thread. The context that the runtime uses, i.e, either the current

context or primary context, can be synchronized with `cudaDeviceSynchronize()`, and destroyed with `cudaDeviceReset()`.

Using the runtime API with primary contexts has its tradeoffs, however. It can cause trouble for users writing plug-ins for larger software packages, for example, because if all plug-ins run in the same process, they will all share a context but will likely have no way to communicate with each other. So, if one of them calls `cudaDeviceReset()` after finishing all its CUDA work, the other plug-ins will fail because the context they were using was destroyed without their knowledge. To avoid this issue, CUDA clients can use the driver API to create and set the current context, and then use the runtime API to work with it. However, contexts may consume significant resources, such as device memory, extra host threads, and performance costs of context switching on the device. This runtime-driver context sharing is important when using the driver API in conjunction with libraries built on the runtime API, such as cuBLAS or cuFFT.

Chapter 2.

API SYNCHRONIZATION BEHAVIOR

The API provides memcpy/memset functions in both synchronous and asynchronous forms, the latter having an "Async" suffix. This is a misnomer as each function may exhibit synchronous or asynchronous behavior depending on the arguments passed to the function.

Memcpy

In the reference documentation, each memcpy function is categorized as synchronous or asynchronous, corresponding to the definitions below.

Synchronous

1. All transfers involving Unified Memory regions are fully synchronous with respect to the host.
2. For transfers from pageable host memory to device memory, a stream sync is performed before the copy is initiated. The function will return once the pageable buffer has been copied to the staging memory for DMA transfer to device memory, but the DMA to final destination may not have completed.
3. For transfers from pinned host memory to device memory, the function is synchronous with respect to the host.
4. For transfers from device to either pageable or pinned host memory, the function returns only once the copy has completed.
5. For transfers from device memory to device memory, no host-side synchronization is performed.
6. For transfers from any host memory to any host memory, the function is fully synchronous with respect to the host.

Asynchronous

1. For transfers from device memory to pageable host memory, the function will return only once the copy has completed.
2. For transfers from any host memory to any host memory, the function is fully synchronous with respect to the host.

3. For all other transfers, the function is fully asynchronous. If pageable memory must first be staged to pinned memory, this will be handled asynchronously with a worker thread.

Memset

The synchronous memset functions are asynchronous with respect to the host except when the target is pinned host memory or a Unified Memory region, in which case they are fully synchronous. The Async versions are always asynchronous with respect to the host.

Kernel Launches

Kernel launches are asynchronous with respect to the host. Details of concurrent kernel execution and data transfers can be found in the CUDA Programmers Guide.

Chapter 3.

STREAM SYNCHRONIZATION BEHAVIOR

Default stream

The default stream, used when `0` is passed as a `cudaStream_t` or by APIs that operate on a stream implicitly, can be configured to have either [legacy](#) or [per-thread](#) synchronization behavior as described below.

The behavior can be controlled per compilation unit with the `--default-stream` `nvcc` option. Alternatively, per-thread behavior can be enabled by defining the `CUDA_API_PER_THREAD_DEFAULT_STREAM` macro before including any CUDA headers. Either way, the `CUDA_API_PER_THREAD_DEFAULT_STREAM` macro will be defined in compilation units using per-thread synchronization behavior.

Legacy default stream

The legacy default stream is an implicit stream which synchronizes with all other streams in the same `CUcontext` except for non-blocking streams, described below. (For applications using the runtime APIs only, there will be one context per device.) When an action is taken in the legacy stream such as a kernel launch or `cudaStreamWaitEvent()`, the legacy stream first waits on all blocking streams, the action is queued in the legacy stream, and then all blocking streams wait on the legacy stream.

For example, the following code launches a kernel `k_1` in stream `s`, then `k_2` in the legacy stream, then `k_3` in stream `s`:

```
k_1<<<<1, 1, 0, s>>>();  
k_2<<<<1, 1>>>();  
k_3<<<<1, 1, 0, s>>>();
```

The resulting behavior is that `k_2` will block on `k_1` and `k_3` will block on `k_2`.

Non-blocking streams which do not synchronize with the legacy stream can be created using the `cudaStreamNonBlocking` flag with the stream creation APIs.

The legacy default stream can be used explicitly with the `CUstream` (`cudaStream_t`) handle `CU_STREAM_LEGACY` (`cudaStreamLegacy`).

Per-thread default stream

The per-thread default stream is an implicit stream local to both the thread and the `CUcontext`, and which does not synchronize with other streams (just like explicitly created streams). The per-thread default stream is not a non-blocking stream and will synchronize with the legacy default stream if both are used in a program.

The per-thread default stream can be used explicitly with the `CUstream` (`cudaStream_t`) handle `CU_STREAM_PER_THREAD` (`cudaStreamPerThread`).

Chapter 4.

MODULES

Here is a list of all modules:

- ▶ Data types used by CUDA driver
- ▶ Error Handling
- ▶ Initialization
- ▶ Version Management
- ▶ Device Management
- ▶ Device Management [DEPRECATED]
- ▶ Primary Context Management
- ▶ Context Management
- ▶ Context Management [DEPRECATED]
- ▶ Module Management
- ▶ Memory Management
- ▶ Unified Addressing
- ▶ Stream Management
- ▶ Event Management
- ▶ Stream memory operations
- ▶ Execution Control
- ▶ Execution Control [DEPRECATED]
- ▶ Occupancy
- ▶ Texture Reference Management
- ▶ Texture Reference Management [DEPRECATED]
- ▶ Surface Reference Management
- ▶ Texture Object Management
- ▶ Surface Object Management
- ▶ Peer Context Memory Access
- ▶ Graphics Interoperability
- ▶ Profiler Control

- ▶ OpenGL Interoperability
 - ▶ OpenGL Interoperability [DEPRECATED]
- ▶ Direct3D 9 Interoperability
 - ▶ Direct3D 9 Interoperability [DEPRECATED]
- ▶ Direct3D 10 Interoperability
 - ▶ Direct3D 10 Interoperability [DEPRECATED]
- ▶ Direct3D 11 Interoperability
 - ▶ Direct3D 11 Interoperability [DEPRECATED]
- ▶ VDPAU Interoperability
- ▶ EGL Interoperability

4.1. Data types used by CUDA driver

struct CUDA_ARRAY3D_DESCRIPTOR

struct CUDA_ARRAY_DESCRIPTOR

struct CUDA_LAUNCH_PARAMS

struct CUDA_MEMCPY2D

struct CUDA_MEMCPY3D

struct CUDA_MEMCPY3D_PEER

struct CUDA_POINTER_ATTRIBUTE_P2P_TOKENS

struct CUDA_RESOURCE_DESC

struct CUDA_RESOURCE_VIEW_DESC

struct CUDA_TEXTURE_DESC

struct CUdevprop

struct CUeglFrame

struct CUipcEventHandle

struct CUipcMemHandle

union CUstreamBatchMemOpParams

enum CUaddress_mode

Texture reference addressing modes

Values

CU_TR_ADDRESS_MODE_WRAP = 0

Wrapping address mode

CU_TR_ADDRESS_MODE_CLAMP = 1

Clamp to edge address mode

CU_TR_ADDRESS_MODE_MIRROR = 2

Mirror address mode

CU_TR_ADDRESS_MODE_BORDER = 3

Border address mode

enum CUarray_cubemap_face

Array indices for cube faces

Values

CU_CUBEMAP_FACE_POSITIVE_X = 0x00

Positive X face of cubemap

CU_CUBEMAP_FACE_NEGATIVE_X = 0x01

Negative X face of cubemap

CU_CUBEMAP_FACE_POSITIVE_Y = 0x02

Positive Y face of cubemap

CU_CUBEMAP_FACE_NEGATIVE_Y = 0x03

Negative Y face of cubemap

CU_CUBEMAP_FACE_POSITIVE_Z = 0x04

Positive Z face of cubemap

CU_CUBEMAP_FACE_NEGATIVE_Z = 0x05

Negative Z face of cubemap

enum CUarray_format

Array formats

Values

CU_AD_FORMAT_UNSIGNED_INT8 = 0x01

Unsigned 8-bit integers

CU_AD_FORMAT_UNSIGNED_INT16 = 0x02

Unsigned 16-bit integers

CU_AD_FORMAT_UNSIGNED_INT32 = 0x03

Unsigned 32-bit integers

CU_AD_FORMAT_SIGNED_INT8 = 0x08

Signed 8-bit integers

CU_AD_FORMAT_SIGNED_INT16 = 0x09

Signed 16-bit integers

CU_AD_FORMAT_SIGNED_INT32 = 0x0a

Signed 32-bit integers

CU_AD_FORMAT_HALF = 0x10

16-bit floating point
CU_AD_FORMAT_FLOAT = 0x20
 32-bit floating point

enum CUcomputemode

Compute Modes

Values

CU_COMPUTEMODE_DEFAULT = 0
 Default compute mode (Multiple contexts allowed per device)
CU_COMPUTEMODE_PROHIBITED = 2
 Compute-prohibited mode (No contexts can be created on this device at this time)
CU_COMPUTEMODE_EXCLUSIVE_PROCESS = 3
 Compute-exclusive-process mode (Only one context used by a single process can be present on this device at a time)

enum CUctx_flags

Context creation flags

Values

CU_CTX_SCHED_AUTO = 0x00
 Automatic scheduling
CU_CTX_SCHED_SPIN = 0x01
 Set spin as default scheduling
CU_CTX_SCHED_YIELD = 0x02
 Set yield as default scheduling
CU_CTX_SCHED_BLOCKING_SYNC = 0x04
 Set blocking synchronization as default scheduling
CU_CTX_BLOCKING_SYNC = 0x04
 Set blocking synchronization as default scheduling **Deprecated**
 This flag was deprecated as of CUDA 4.0 and was replaced with **CU_CTX_SCHED_BLOCKING_SYNC**.
CU_CTX_SCHED_MASK = 0x07
CU_CTX_MAP_HOST = 0x08
 Support mapped pinned allocations
CU_CTX_LMEM_RESIZE_TO_MAX = 0x10
 Keep local memory allocation after launch
CU_CTX_FLAGS_MASK = 0x1f

enum CUdevice_attribute

Device properties

Values

CU_DEVICE_ATTRIBUTE_MAX_THREADS_PER_BLOCK = 1

Maximum number of threads per block

CU_DEVICE_ATTRIBUTE_MAX_BLOCK_DIM_X = 2

Maximum block dimension X

CU_DEVICE_ATTRIBUTE_MAX_BLOCK_DIM_Y = 3

Maximum block dimension Y

CU_DEVICE_ATTRIBUTE_MAX_BLOCK_DIM_Z = 4

Maximum block dimension Z

CU_DEVICE_ATTRIBUTE_MAX_GRID_DIM_X = 5

Maximum grid dimension X

CU_DEVICE_ATTRIBUTE_MAX_GRID_DIM_Y = 6

Maximum grid dimension Y

CU_DEVICE_ATTRIBUTE_MAX_GRID_DIM_Z = 7

Maximum grid dimension Z

CU_DEVICE_ATTRIBUTE_MAX_SHARED_MEMORY_PER_BLOCK = 8

Maximum shared memory available per block in bytes

CU_DEVICE_ATTRIBUTE_SHARED_MEMORY_PER_BLOCK = 8

Deprecated, use

CU_DEVICE_ATTRIBUTE_MAX_SHARED_MEMORY_PER_BLOCK

CU_DEVICE_ATTRIBUTE_TOTAL_CONSTANT_MEMORY = 9

Memory available on device for `__constant__` variables in a CUDA C kernel in bytes

CU_DEVICE_ATTRIBUTE_WARP_SIZE = 10

Warp size in threads

CU_DEVICE_ATTRIBUTE_MAX_PITCH = 11

Maximum pitch in bytes allowed by memory copies

CU_DEVICE_ATTRIBUTE_MAX_REGISTERS_PER_BLOCK = 12

Maximum number of 32-bit registers available per block

CU_DEVICE_ATTRIBUTE_REGISTERS_PER_BLOCK = 12

Deprecated, use **CU_DEVICE_ATTRIBUTE_MAX_REGISTERS_PER_BLOCK**

CU_DEVICE_ATTRIBUTE_CLOCK_RATE = 13

Typical clock frequency in kilohertz

CU_DEVICE_ATTRIBUTE_TEXTURE_ALIGNMENT = 14

Alignment requirement for textures

CU_DEVICE_ATTRIBUTE_GPU_OVERLAP = 15

Device can possibly copy memory and execute a kernel concurrently. Deprecated. Use instead **CU_DEVICE_ATTRIBUTE_ASYNC_ENGINE_COUNT**.

CU_DEVICE_ATTRIBUTE_MULTIPROCESSOR_COUNT = 16

Number of multiprocessors on device

CU_DEVICE_ATTRIBUTE_KERNEL_EXEC_TIMEOUT = 17
Specifies whether there is a run time limit on kernels

CU_DEVICE_ATTRIBUTE_INTEGRATED = 18
Device is integrated with host memory

CU_DEVICE_ATTRIBUTE_CAN_MAP_HOST_MEMORY = 19
Device can map host memory into CUDA address space

CU_DEVICE_ATTRIBUTE_COMPUTE_MODE = 20
Compute mode (See [CUcomputemode](#) for details)

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_WIDTH = 21
Maximum 1D texture width

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_WIDTH = 22
Maximum 2D texture width

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_HEIGHT = 23
Maximum 2D texture height

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_WIDTH = 24
Maximum 3D texture width

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_HEIGHT = 25
Maximum 3D texture height

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_DEPTH = 26
Maximum 3D texture depth

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_WIDTH = 27
Maximum 2D layered texture width

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_HEIGHT = 28
Maximum 2D layered texture height

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_LAYERS = 29
Maximum layers in a 2D layered texture

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_ARRAY_WIDTH = 27
Deprecated, use
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_WIDTH

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_ARRAY_HEIGHT = 28
Deprecated, use
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_HEIGHT

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_ARRAY_NUMSLICES = 29
Deprecated, use
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_LAYERS

CU_DEVICE_ATTRIBUTE_SURFACE_ALIGNMENT = 30
Alignment requirement for surfaces

CU_DEVICE_ATTRIBUTE_CONCURRENT_KERNELS = 31
Device can possibly execute multiple kernels concurrently

CU_DEVICE_ATTRIBUTE_ECC_ENABLED = 32
Device has ECC support enabled

CU_DEVICE_ATTRIBUTE_PCI_BUS_ID = 33

PCI bus ID of the device
CU_DEVICE_ATTRIBUTE_PCI_DEVICE_ID = 34
 PCI device ID of the device
CU_DEVICE_ATTRIBUTE_TCC_DRIVER = 35
 Device is using TCC driver model
CU_DEVICE_ATTRIBUTE_MEMORY_CLOCK_RATE = 36
 Peak memory clock frequency in kilohertz
CU_DEVICE_ATTRIBUTE_GLOBAL_MEMORY_BUS_WIDTH = 37
 Global memory bus width in bits
CU_DEVICE_ATTRIBUTE_L2_CACHE_SIZE = 38
 Size of L2 cache in bytes
CU_DEVICE_ATTRIBUTE_MAX_THREADS_PER_MULTIPROCESSOR = 39
 Maximum resident threads per multiprocessor
CU_DEVICE_ATTRIBUTE_ASYNC_ENGINE_COUNT = 40
 Number of asynchronous engines
CU_DEVICE_ATTRIBUTE_UNIFIED_ADDRESSING = 41
 Device shares a unified address space with the host
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_LAYERED_WIDTH = 42
 Maximum 1D layered texture width
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_LAYERED_LAYERS = 43
 Maximum layers in a 1D layered texture
CU_DEVICE_ATTRIBUTE_CAN_TEX2D_GATHER = 44
 Deprecated, do not use.
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_GATHER_WIDTH = 45
 Maximum 2D texture width if CUDA_ARRAY3D_TEXTURE_GATHER is set
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_GATHER_HEIGHT = 46
 Maximum 2D texture height if CUDA_ARRAY3D_TEXTURE_GATHER is set
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_WIDTH_ALTERNATE = 47
 Alternate maximum 3D texture width
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_HEIGHT_ALTERNATE = 48
 Alternate maximum 3D texture height
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_DEPTH_ALTERNATE = 49
 Alternate maximum 3D texture depth
CU_DEVICE_ATTRIBUTE_PCI_DOMAIN_ID = 50
 PCI domain ID of the device
CU_DEVICE_ATTRIBUTE_TEXTURE_PITCH_ALIGNMENT = 51
 Pitch alignment requirement for textures
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURECUBEMAP_WIDTH = 52
 Maximum cubemap texture width/height
CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURECUBEMAP_LAYERED_WIDTH = 53
 Maximum cubemap layered texture width/height

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURECUBEMAP_LAYERED_LAYERS = 54

Maximum layers in a cubemap layered texture

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE1D_WIDTH = 55

Maximum 1D surface width

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_WIDTH = 56

Maximum 2D surface width

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_HEIGHT = 57

Maximum 2D surface height

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE3D_WIDTH = 58

Maximum 3D surface width

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE3D_HEIGHT = 59

Maximum 3D surface height

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE3D_DEPTH = 60

Maximum 3D surface depth

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE1D_LAYERED_WIDTH = 61

Maximum 1D layered surface width

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE1D_LAYERED_LAYERS = 62

Maximum layers in a 1D layered surface

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_LAYERED_WIDTH = 63

Maximum 2D layered surface width

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_LAYERED_HEIGHT = 64

Maximum 2D layered surface height

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_LAYERED_LAYERS = 65

Maximum layers in a 2D layered surface

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACECUBEMAP_WIDTH = 66

Maximum cubemap surface width

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACECUBEMAP_LAYERED_WIDTH = 67

Maximum cubemap layered surface width

CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACECUBEMAP_LAYERED_LAYERS = 68

Maximum layers in a cubemap layered surface

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_LINEAR_WIDTH = 69

Maximum 1D linear texture width

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_WIDTH = 70

Maximum 2D linear texture width

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_HEIGHT = 71

Maximum 2D linear texture height

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_PITCH = 72

Maximum 2D linear texture pitch in bytes

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_MIPMAPPED_WIDTH = 73

Maximum mipmapped 2D texture width

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_MIPMAPPED_HEIGHT = 74

Maximum mipmapped 2D texture height

CU_DEVICE_ATTRIBUTE_COMPUTE_CAPABILITY_MAJOR = 75

Major compute capability version number

CU_DEVICE_ATTRIBUTE_COMPUTE_CAPABILITY_MINOR = 76

Minor compute capability version number

CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_MIPMAPPED_WIDTH = 77

Maximum mipmapped 1D texture width

CU_DEVICE_ATTRIBUTE_STREAM_PRIORITIES_SUPPORTED = 78

Device supports stream priorities

CU_DEVICE_ATTRIBUTE_GLOBAL_L1_CACHE_SUPPORTED = 79

Device supports caching globals in L1

CU_DEVICE_ATTRIBUTE_LOCAL_L1_CACHE_SUPPORTED = 80

Device supports caching locals in L1

CU_DEVICE_ATTRIBUTE_MAX_SHARED_MEMORY_PER_MULTIPROCESSOR = 81

Maximum shared memory available per multiprocessor in bytes

CU_DEVICE_ATTRIBUTE_MAX_REGISTERS_PER_MULTIPROCESSOR = 82

Maximum number of 32-bit registers available per multiprocessor

CU_DEVICE_ATTRIBUTE_MANAGED_MEMORY = 83

Device can allocate managed memory on this system

CU_DEVICE_ATTRIBUTE_MULTI_GPU_BOARD = 84

Device is on a multi-GPU board

CU_DEVICE_ATTRIBUTE_MULTI_GPU_BOARD_GROUP_ID = 85

Unique id for a group of devices on the same multi-GPU board

CU_DEVICE_ATTRIBUTE_HOST_NATIVE_ATOMIC_SUPPORTED = 86

Link between the device and the host supports native atomic operations (this is a placeholder attribute, and is not supported on any current hardware)

CU_DEVICE_ATTRIBUTE_SINGLE_TO_DOUBLE_PRECISION_PERF_RATIO = 87

Ratio of single precision performance (in floating-point operations per second) to double precision performance

CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS = 88

Device supports coherently accessing pageable memory without calling `cudaHostRegister` on it

CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS = 89

Device can coherently access managed memory concurrently with the CPU

CU_DEVICE_ATTRIBUTE_COMPUTE_PREEMPTION_SUPPORTED = 90

Device supports compute preemption.

CU_DEVICE_ATTRIBUTE_CAN_USE_HOST_POINTER_FOR_REGISTERED_MEM = 91

Device can access host registered memory at the same virtual address as the CPU

CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_MEM_OPS = 92

`cuStreamBatchMemOp` and related APIs are supported.

CU_DEVICE_ATTRIBUTE_CAN_USE_64_BIT_STREAM_MEM_OPS = 93

64-bit operations are supported in [cuStreamBatchMemOp](#) and related APIs.

CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_WAIT_VALUE_NOR = 94

[CU_STREAM_WAIT_VALUE_NOR](#) is supported.

CU_DEVICE_ATTRIBUTE_COOPERATIVE_LAUNCH = 95

Device supports launching cooperative kernels via [cuLaunchCooperativeKernel](#)

CU_DEVICE_ATTRIBUTE_COOPERATIVE_MULTI_DEVICE_LAUNCH = 96

Device can participate in cooperative kernels launched via

[cuLaunchCooperativeKernelMultiDevice](#)

CU_DEVICE_ATTRIBUTE_MAX_SHARED_MEMORY_PER_BLOCK_OPTIN = 97

Maximum optin shared memory per block

CU_DEVICE_ATTRIBUTE_CAN_FLUSH_REMOTE_WRITES = 98

Both the [CU_STREAM_WAIT_VALUE_FLUSH](#) flag and the

[CU_STREAM_MEM_OP_FLUSH_REMOTE_WRITES](#) MemOp are supported on the

device. See [Stream memory operations](#) for additional details.

CU_DEVICE_ATTRIBUTE_HOST_REGISTER_SUPPORTED = 99

Device supports host memory registration via [cudaHostRegister](#).

CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS_USES_HOST_PAGE_TABLES = 100

Device accesses pageable memory via the host's page tables.

CU_DEVICE_ATTRIBUTE_DIRECT_MANAGED_MEM_ACCESS_FROM_HOST = 101

The host can directly access managed memory on the device without migration.

CU_DEVICE_ATTRIBUTE_MAX

enum CUdevice_P2PAttribute

P2P Attributes

Values

CU_DEVICE_P2P_ATTRIBUTE_PERFORMANCE_RANK = 0x01

A relative value indicating the performance of the link between two devices

CU_DEVICE_P2P_ATTRIBUTE_ACCESS_SUPPORTED = 0x02

P2P Access is enable

CU_DEVICE_P2P_ATTRIBUTE_NATIVE_ATOMIC_SUPPORTED = 0x03

Atomic operation over the link supported

CU_DEVICE_P2P_ATTRIBUTE_ARRAY_ACCESS_ACCESS_SUPPORTED = 0x04

Accessing CUDA arrays over the link supported

enum CUeglColorFormat

CUDA EGL Color Format - The different planar and multiplanar formats currently supported for [CUDA_EGL](#) interops.

Values**CU_EGL_COLOR_FORMAT_YUV420_PLANAR = 0x00**

Y, U, V in three surfaces, each in a separate surface, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_YUV420_SEMIPLANAR = 0x01

Y, UV in two surfaces (UV as one surface) with VU byte ordering, width, height ratio same as YUV420Planar.

CU_EGL_COLOR_FORMAT_YUV422_PLANAR = 0x02

Y, U, V each in a separate surface, U/V width = 1/2 Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YUV422_SEMIPLANAR = 0x03

Y, UV in two surfaces with VU byte ordering, width, height ratio same as YUV422Planar.

CU_EGL_COLOR_FORMAT_RGB = 0x04

R/G/B three channels in one surface with BGR byte ordering. Only pitch linear format supported.

CU_EGL_COLOR_FORMAT_BGR = 0x05

R/G/B three channels in one surface with RGB byte ordering. Only pitch linear format supported.

CU_EGL_COLOR_FORMAT_ARGB = 0x06

R/G/B/A four channels in one surface with BGRA byte ordering.

CU_EGL_COLOR_FORMAT_RGBA = 0x07

R/G/B/A four channels in one surface with ABGR byte ordering.

CU_EGL_COLOR_FORMAT_L = 0x08

single luminance channel in one surface.

CU_EGL_COLOR_FORMAT_R = 0x09

single color channel in one surface.

CU_EGL_COLOR_FORMAT_YUV444_PLANAR = 0x0A

Y, U, V in three surfaces, each in a separate surface, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YUV444_SEMIPLANAR = 0x0B

Y, UV in two surfaces (UV as one surface) with VU byte ordering, width, height ratio same as YUV444Planar.

CU_EGL_COLOR_FORMAT_YUYV_422 = 0x0C

Y, U, V in one surface, interleaved as UYVY.

CU_EGL_COLOR_FORMAT_UYVY_422 = 0x0D

Y, U, V in one surface, interleaved as YUYV.

CU_EGL_COLOR_FORMAT_ABGR = 0x0E

R/G/B/A four channels in one surface with RGBA byte ordering.

CU_EGL_COLOR_FORMAT_BGRA = 0x0F

R/G/B/A four channels in one surface with ARGB byte ordering.

CU_EGL_COLOR_FORMAT_A = 0x10

Alpha color format - one channel in one surface.

CU_EGL_COLOR_FORMAT_RG = 0x11

R/G color format - two channels in one surface with GR byte ordering

CU_EGL_COLOR_FORMAT_AYUV = 0x12

Y, U, V, A four channels in one surface, interleaved as VUYA.

CU_EGL_COLOR_FORMAT_YVU444_SEMIPLANAR = 0x13

Y, VU in two surfaces (VU as one surface) with UV byte ordering, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YVU422_SEMIPLANAR = 0x14

Y, VU in two surfaces (VU as one surface) with UV byte ordering, U/V width = 1/2 Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YVU420_SEMIPLANAR = 0x15

Y, VU in two surfaces (VU as one surface) with UV byte ordering, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_Y10V10U10_444_SEMIPLANAR = 0x16

Y10, V10U10 in two surfaces (VU as one surface) with UV byte ordering, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_Y10V10U10_420_SEMIPLANAR = 0x17

Y10, V10U10 in two surfaces (VU as one surface) with UV byte ordering, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_Y12V12U12_444_SEMIPLANAR = 0x18

Y12, V12U12 in two surfaces (VU as one surface) with UV byte ordering, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_Y12V12U12_420_SEMIPLANAR = 0x19

Y12, V12U12 in two surfaces (VU as one surface) with UV byte ordering, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_VYUY_ER = 0x1A

Extended Range Y, U, V in one surface, interleaved as YVYU.

CU_EGL_COLOR_FORMAT_UYVY_ER = 0x1B

Extended Range Y, U, V in one surface, interleaved as YUYV.

CU_EGL_COLOR_FORMAT_YUYV_ER = 0x1C

Extended Range Y, U, V in one surface, interleaved as UYVY.

CU_EGL_COLOR_FORMAT_YVYU_ER = 0x1D

Extended Range Y, U, V in one surface, interleaved as VYUY.

CU_EGL_COLOR_FORMAT_YUV_ER = 0x1E

Extended Range Y, U, V three channels in one surface, interleaved as VUY. Only pitch linear format supported.

CU_EGL_COLOR_FORMAT_YUVA_ER = 0x1F

Extended Range Y, U, V, A four channels in one surface, interleaved as AVUY.

CU_EGL_COLOR_FORMAT_AYUV_ER = 0x20

Extended Range Y, U, V, A four channels in one surface, interleaved as VUYA.

CU_EGL_COLOR_FORMAT_YUV444_PLANAR_ER = 0x21

Extended Range Y, U, V in three surfaces, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YUV422_PLANAR_ER = 0x22

Extended Range Y, U, V in three surfaces, U/V width = 1/2 Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YUV420_PLANAR_ER = 0x23

Extended Range Y, U, V in three surfaces, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_YUV444_SEMIPLANAR_ER = 0x24

Extended Range Y, UV in two surfaces (UV as one surface) with VU byte ordering, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YUV422_SEMIPLANAR_ER = 0x25

Extended Range Y, UV in two surfaces (UV as one surface) with VU byte ordering, U/V width = 1/2 Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YUV420_SEMIPLANAR_ER = 0x26

Extended Range Y, UV in two surfaces (UV as one surface) with VU byte ordering, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_YVU444_PLANAR_ER = 0x27

Extended Range Y, V, U in three surfaces, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YVU422_PLANAR_ER = 0x28

Extended Range Y, V, U in three surfaces, U/V width = 1/2 Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YVU420_PLANAR_ER = 0x29

Extended Range Y, V, U in three surfaces, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_YVU444_SEMIPLANAR_ER = 0x2A

Extended Range Y, VU in two surfaces (VU as one surface) with UV byte ordering, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YVU422_SEMIPLANAR_ER = 0x2B

Extended Range Y, VU in two surfaces (VU as one surface) with UV byte ordering, U/V width = 1/2 Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YVU420_SEMIPLANAR_ER = 0x2C

Extended Range Y, VU in two surfaces (VU as one surface) with UV byte ordering, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_BAYER_RGGB = 0x2D

Bayer format - one channel in one surface with interleaved RGGB ordering.

CU_EGL_COLOR_FORMAT_BAYER_BGGR = 0x2E

Bayer format - one channel in one surface with interleaved BGGR ordering.

CU_EGL_COLOR_FORMAT_BAYER_GRBG = 0x2F

Bayer format - one channel in one surface with interleaved GRBG ordering.

CU_EGL_COLOR_FORMAT_BAYER_GBRG = 0x30

Bayer format - one channel in one surface with interleaved GBRG ordering.

CU_EGL_COLOR_FORMAT_BAYER10_RGGB = 0x31

Bayer10 format - one channel in one surface with interleaved RGGB ordering. Out of 16 bits, 10 bits used 6 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER10_BGGR = 0x32

Bayer10 format - one channel in one surface with interleaved BGGR ordering. Out of 16 bits, 10 bits used 6 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER10_GRBG = 0x33

Bayer10 format - one channel in one surface with interleaved GRBG ordering. Out of 16 bits, 10 bits used 6 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER10_GBRG = 0x34

Bayer10 format - one channel in one surface with interleaved GBRG ordering. Out of 16 bits, 10 bits used 6 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER12_RGGB = 0x35

Bayer12 format - one channel in one surface with interleaved RGGB ordering. Out of 16 bits, 12 bits used 4 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER12_BGGR = 0x36

Bayer12 format - one channel in one surface with interleaved BGGR ordering. Out of 16 bits, 12 bits used 4 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER12_GRBG = 0x37

Bayer12 format - one channel in one surface with interleaved GRBG ordering. Out of 16 bits, 12 bits used 4 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER12_GBRG = 0x38

Bayer12 format - one channel in one surface with interleaved GBRG ordering. Out of 16 bits, 12 bits used 4 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER14_RGGB = 0x39

Bayer14 format - one channel in one surface with interleaved RGGB ordering. Out of 16 bits, 14 bits used 2 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER14_BGGR = 0x3A

Bayer14 format - one channel in one surface with interleaved BGGR ordering. Out of 16 bits, 14 bits used 2 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER14_GRBG = 0x3B

Bayer14 format - one channel in one surface with interleaved GRBG ordering. Out of 16 bits, 14 bits used 2 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER14_GBRG = 0x3C

Bayer14 format - one channel in one surface with interleaved GBRG ordering. Out of 16 bits, 14 bits used 2 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER20_RGGB = 0x3D

Bayer20 format - one channel in one surface with interleaved RGGB ordering. Out of 32 bits, 20 bits used 12 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER20_BGGR = 0x3E

Bayer20 format - one channel in one surface with interleaved BGGR ordering. Out of 32 bits, 20 bits used 12 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER20_GRBG = 0x3F

Bayer20 format - one channel in one surface with interleaved GRBG ordering. Out of 32 bits, 20 bits used 12 bits No-op.

CU_EGL_COLOR_FORMAT_BAYER20_GBRG = 0x40

Bayer20 format - one channel in one surface with interleaved GBRG ordering. Out of 32 bits, 20 bits used 12 bits No-op.

CU_EGL_COLOR_FORMAT_YVU444_PLANAR = 0x41

Y, V, U in three surfaces, each in a separate surface, U/V width = Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YVU422_PLANAR = 0x42

Y, V, U in three surfaces, each in a separate surface, U/V width = 1/2 Y width, U/V height = Y height.

CU_EGL_COLOR_FORMAT_YVU420_PLANAR = 0x43

Y, V, U in three surfaces, each in a separate surface, U/V width = 1/2 Y width, U/V height = 1/2 Y height.

CU_EGL_COLOR_FORMAT_BAYER_ISP_RGGB = 0x44

Nvidia proprietary Bayer ISP format - one channel in one surface with interleaved RGGB ordering and mapped to opaque integer datatype.

CU_EGL_COLOR_FORMAT_BAYER_ISP_BGGR = 0x45

Nvidia proprietary Bayer ISP format - one channel in one surface with interleaved BGGR ordering and mapped to opaque integer datatype.

CU_EGL_COLOR_FORMAT_BAYER_ISP_GRBG = 0x46

Nvidia proprietary Bayer ISP format - one channel in one surface with interleaved GRBG ordering and mapped to opaque integer datatype.

CU_EGL_COLOR_FORMAT_BAYER_ISP_GBRG = 0x47

Nvidia proprietary Bayer ISP format - one channel in one surface with interleaved GBRG ordering and mapped to opaque integer datatype.

CU_EGL_COLOR_FORMAT_MAX

enum CUeglFrameType

CUDA EglFrame type - array or pointer

Values

CU_EGL_FRAME_TYPE_ARRAY = 0

Frame type CUDA array

CU_EGL_FRAME_TYPE_PITCH = 1

Frame type pointer

enum CUeglResourceLocationFlags

Resource location flags- system or vidmem

For CUDA context on iGPU, since video and system memory are equivalent - these flags will not have an effect on the execution.

For CUDA context on dGPU, applications can use the flag [CUeglResourceLocationFlags](#) to give a hint about the desired location.

CU_EGL_RESOURCE_LOCATION_SYSTEMEM - the frame data is made resident on the system memory to be accessed by CUDA.

CU_EGL_RESOURCE_LOCATION_VIDMEM - the frame data is made resident on the dedicated video memory to be accessed by CUDA.

There may be an additional latency due to new allocation and data migration, if the frame is produced on a different memory.

Values

CU_EGL_RESOURCE_LOCATION_SYSTEMEM = 0x00

Resource location systemem

CU_EGL_RESOURCE_LOCATION_VIDMEM = 0x01

Resource location vidmem

enum CUevent_flags

Event creation flags

Values

CU_EVENT_DEFAULT = 0x0

Default event flag

CU_EVENT_BLOCKING_SYNC = 0x1

Event uses blocking synchronization

CU_EVENT_DISABLE_TIMING = 0x2

Event will not record timing data

CU_EVENT_INTERPROCESS = 0x4

Event is suitable for interprocess use. **CU_EVENT_DISABLE_TIMING** must be set

enum CUfilter_mode

Texture reference filtering modes

Values

CU_TR_FILTER_MODE_POINT = 0

Point filter mode

CU_TR_FILTER_MODE_LINEAR = 1

Linear filter mode

enum CUfunc_cache

Function cache configurations

Values**CU_FUNC_CACHE_PREFER_NONE = 0x00**

no preference for shared memory or L1 (default)

CU_FUNC_CACHE_PREFER_SHARED = 0x01

prefer larger shared memory and smaller L1 cache

CU_FUNC_CACHE_PREFER_L1 = 0x02

prefer larger L1 cache and smaller shared memory

CU_FUNC_CACHE_PREFER_EQUAL = 0x03

prefer equal sized L1 cache and shared memory

enum CUfunction_attribute

Function properties

Values**CU_FUNC_ATTRIBUTE_MAX_THREADS_PER_BLOCK = 0**

The maximum number of threads per block, beyond which a launch of the function would fail. This number depends on both the function and the device on which the function is currently loaded.

CU_FUNC_ATTRIBUTE_SHARED_SIZE_BYTES = 1

The size in bytes of statically-allocated shared memory required by this function.

This does not include dynamically-allocated shared memory requested by the user at runtime.

CU_FUNC_ATTRIBUTE_CONST_SIZE_BYTES = 2

The size in bytes of user-allocated constant memory required by this function.

CU_FUNC_ATTRIBUTE_LOCAL_SIZE_BYTES = 3

The size in bytes of local memory used by each thread of this function.

CU_FUNC_ATTRIBUTE_NUM_REGS = 4

The number of registers used by each thread of this function.

CU_FUNC_ATTRIBUTE_PTX_VERSION = 5

The PTX virtual architecture version for which the function was compiled. This value is the major PTX version * 10 + the minor PTX version, so a PTX version 1.3 function would return the value 13. Note that this may return the undefined value of 0 for cubins compiled prior to CUDA 3.0.

CU_FUNC_ATTRIBUTE_BINARY_VERSION = 6

The binary architecture version for which the function was compiled. This value is the major binary version * 10 + the minor binary version, so a binary version 1.3 function would return the value 13. Note that this will return a value of 10 for legacy cubins that do not have a properly-encoded binary architecture version.

CU_FUNC_ATTRIBUTE_CACHE_MODE_CA = 7

The attribute to indicate whether the function has been compiled with user specified option "-Xptxas --dlcm=ca" set.

CU_FUNC_ATTRIBUTE_MAX_DYNAMIC_SHARED_SIZE_BYTES = 8

The maximum size in bytes of dynamically-allocated shared memory that can be used by this function. If the user-specified dynamic shared memory size is larger than this value, the launch will fail.

CU_FUNC_ATTRIBUTE_PREFERRED_SHARED_MEMORY_CARVEOUT = 9

On devices where the L1 cache and shared memory use the same hardware resources, this sets the shared memory carveout preference, in percent of the total resources.

This is only a hint, and the driver can choose a different ratio if required to execute the function.

CU_FUNC_ATTRIBUTE_MAX

enum CUgraphicsMapResourceFlags

Flags for mapping and unmapping interop resources

Values

CU_GRAPHICS_MAP_RESOURCE_FLAGS_NONE = 0x00

CU_GRAPHICS_MAP_RESOURCE_FLAGS_READ_ONLY = 0x01

CU_GRAPHICS_MAP_RESOURCE_FLAGS_WRITE_DISCARD = 0x02

enum CUgraphicsRegisterFlags

Flags to register a graphics resource

Values

CU_GRAPHICS_REGISTER_FLAGS_NONE = 0x00

CU_GRAPHICS_REGISTER_FLAGS_READ_ONLY = 0x01

CU_GRAPHICS_REGISTER_FLAGS_WRITE_DISCARD = 0x02

CU_GRAPHICS_REGISTER_FLAGS_SURFACE_LDST = 0x04

CU_GRAPHICS_REGISTER_FLAGS_TEXTURE_GATHER = 0x08

enum CUipcMem_flags

CUDA Ipc Mem Flags

Values

CU_IPC_MEM_LAZY_ENABLE_PEER_ACCESS = 0x1

Automatically enable peer access between remote devices as needed

enum CUjit_cacheMode

Caching modes for dlcm

Values**CU_JIT_CACHE_OPTION_NONE = 0**

Compile with no -dlcm flag specified

CU_JIT_CACHE_OPTION_CG

Compile with L1 cache disabled

CU_JIT_CACHE_OPTION_CA

Compile with L1 cache enabled

enum CUjit_fallback

Cubin matching fallback strategies

Values**CU_PREFER_PTX = 0**

Prefer to compile ptx if exact binary match not found

CU_PREFER_BINARY

Prefer to fall back to compatible binary code if exact match not found

enum CUjit_option

Online compiler and linker options

Values**CU_JIT_MAX_REGISTERS = 0**

Max number of registers that a thread may use. Option type: unsigned int Applies to: compiler only

CU_JIT_THREADS_PER_BLOCKIN: Specifies minimum number of threads per block to target compilation for OUT: Returns the number of threads the compiler actually targeted. This restricts the resource utilization fo the compiler (e.g. max registers) such that a block with the given number of threads should be able to launch based on register limitations. Note, this option does not currently take into account any other resource limitations, such as shared memory utilization. Cannot be combined with [CU_JIT_TARGET](#). Option type: unsigned int Applies to: compiler only**CU_JIT_WALL_TIME**

Overwrites the option value with the total wall clock time, in milliseconds, spent in the compiler and linker Option type: float Applies to: compiler and linker

CU_JIT_INFO_LOG_BUFFERPointer to a buffer in which to print any log messages that are informational in nature (the buffer size is specified via option [CU_JIT_INFO_LOG_BUFFER_SIZE_BYTES](#))

Option type: char * Applies to: compiler and linker

CU_JIT_INFO_LOG_BUFFER_SIZE_BYTES

IN: Log buffer size in bytes. Log messages will be capped at this size (including null terminator) OUT: Amount of log buffer filled with messages Option type: unsigned int Applies to: compiler and linker

CU_JIT_ERROR_LOG_BUFFER

Pointer to a buffer in which to print any log messages that reflect errors (the buffer size is specified via option [CU_JIT_ERROR_LOG_BUFFER_SIZE_BYTES](#)) Option type: char * Applies to: compiler and linker

CU_JIT_ERROR_LOG_BUFFER_SIZE_BYTES

IN: Log buffer size in bytes. Log messages will be capped at this size (including null terminator) OUT: Amount of log buffer filled with messages Option type: unsigned int Applies to: compiler and linker

CU_JIT_OPTIMIZATION_LEVEL

Level of optimizations to apply to generated code (0 - 4), with 4 being the default and highest level of optimizations. Option type: unsigned int Applies to: compiler only

CU_JIT_TARGET_FROM_CUCONTEXT

No option value required. Determines the target based on the current attached context (default) Option type: No option value needed Applies to: compiler and linker

CU_JIT_TARGET

Target is chosen based on supplied [CUjit_target](#). Cannot be combined with [CU_JIT_THREADS_PER_BLOCK](#). Option type: unsigned int for enumerated type [CUjit_target](#) Applies to: compiler and linker

CU_JIT_FALLBACK_STRATEGY

Specifies choice of fallback strategy if matching cubin is not found. Choice is based on supplied [CUjit_fallback](#). This option cannot be used with cuLink* APIs as the linker requires exact matches. Option type: unsigned int for enumerated type [CUjit_fallback](#) Applies to: compiler only

CU_JIT_GENERATE_DEBUG_INFO

Specifies whether to create debug information in output (-g) (0: false, default) Option type: int Applies to: compiler and linker

CU_JIT_LOG_VERBOSE

Generate verbose log messages (0: false, default) Option type: int Applies to: compiler and linker

CU_JIT_GENERATE_LINE_INFO

Generate line number information (-lineinfo) (0: false, default) Option type: int Applies to: compiler only

CU_JIT_CACHE_MODE

Specifies whether to enable caching explicitly (-dlcm) Choice is based on supplied [CUjit_cacheMode_enum](#). Option type: unsigned int for enumerated type [CUjit_cacheMode_enum](#) Applies to: compiler only

CU_JIT_NEW_SM3X_OPT

The below jit options are used for internal purposes only, in this version of CUDA

CU_JIT_FAST_COMPILE

CU_JIT_NUM_OPTIONS

enum CUjit_target

Online compilation targets

Values

CU_TARGET_COMPUTE_20 = 20

Compute device class 2.0

CU_TARGET_COMPUTE_21 = 21

Compute device class 2.1

CU_TARGET_COMPUTE_30 = 30

Compute device class 3.0

CU_TARGET_COMPUTE_32 = 32

Compute device class 3.2

CU_TARGET_COMPUTE_35 = 35

Compute device class 3.5

CU_TARGET_COMPUTE_37 = 37

Compute device class 3.7

CU_TARGET_COMPUTE_50 = 50

Compute device class 5.0

CU_TARGET_COMPUTE_52 = 52

Compute device class 5.2

CU_TARGET_COMPUTE_53 = 53

Compute device class 5.3

CU_TARGET_COMPUTE_60 = 60

Compute device class 6.0.

CU_TARGET_COMPUTE_61 = 61

Compute device class 6.1.

CU_TARGET_COMPUTE_62 = 62

Compute device class 6.2.

CU_TARGET_COMPUTE_70 = 70

Compute device class 7.0.

CU_TARGET_COMPUTE_73 = 73

Compute device class 7.3.

CU_TARGET_COMPUTE_75 = 75

Compute device class 7.5.

enum CUjitInputType

Device code formats

Values**CU_JIT_INPUT_CUBIN = 0**

Compiled device-class-specific device code Applicable options: none

CU_JIT_INPUT_PTX

PTX source code Applicable options: PTX compiler options

CU_JIT_INPUT_FATBINARYBundle of multiple cubins and/or PTX of some device code Applicable options: PTX compiler options, [CU_JIT_FALLBACK_STRATEGY](#)**CU_JIT_INPUT_OBJECT**Host object with embedded device code Applicable options: PTX compiler options, [CU_JIT_FALLBACK_STRATEGY](#)**CU_JIT_INPUT_LIBRARY**Archive of host objects with embedded device code Applicable options: PTX compiler options, [CU_JIT_FALLBACK_STRATEGY](#)**CU_JIT_NUM_INPUT_TYPES**

enum CUlimit

Limits

Values**CU_LIMIT_STACK_SIZE = 0x00**

GPU thread stack size

CU_LIMIT_PRINTF_FIFO_SIZE = 0x01

GPU printf FIFO size

CU_LIMIT_MALLOC_HEAP_SIZE = 0x02

GPU malloc heap size

CU_LIMIT_DEV_RUNTIME_SYNC_DEPTH = 0x03

GPU device runtime launch synchronize depth

CU_LIMIT_DEV_RUNTIME_PENDING_LAUNCH_COUNT = 0x04

GPU device runtime pending launch count

CU_LIMIT_MAX

enum CUmem_advise

Memory advise values

Values**CU_MEM_ADVISE_SET_READ_MOSTLY = 1**

Data will mostly be read and only occasionally be written to

CU_MEM_ADVISE_UNSET_READ_MOSTLY = 2Undo the effect of [CU_MEM_ADVISE_SET_READ_MOSTLY](#)**CU_MEM_ADVISE_SET_PREFERRED_LOCATION = 3**

Set the preferred location for the data as the specified device

CU_MEM_ADVISE_UNSET_PREFERRED_LOCATION = 4

Clear the preferred location for the data

CU_MEM_ADVISE_SET_ACCESSED_BY = 5

Data will be accessed by the specified device, so prevent page faults as much as possible

CU_MEM_ADVISE_UNSET_ACCESSED_BY = 6

Let the Unified Memory subsystem decide on the page faulting policy for the specified device

enum CUmemAttach_flags

CUDA Mem Attach Flags

Values

CU_MEM_ATTACH_GLOBAL = 0x1

Memory can be accessed by any stream on any device

CU_MEM_ATTACH_HOST = 0x2

Memory cannot be accessed by any stream on any device

CU_MEM_ATTACH_SINGLE = 0x4

Memory can only be accessed by a single stream on the associated device

enum CUmemorytype

Memory types

Values

CU_MEMORYTYPE_HOST = 0x01

Host memory

CU_MEMORYTYPE_DEVICE = 0x02

Device memory

CU_MEMORYTYPE_ARRAY = 0x03

Array memory

CU_MEMORYTYPE_UNIFIED = 0x04

Unified device or host memory

enum CUoccupancy_flags

Occupancy calculator flag

Values

CU_OCCUPANCY_DEFAULT = 0x0

Default behavior

CU_OCCUPANCY_DISABLE_CACHING_OVERRIDE = 0x1

Assume global caching is enabled and cannot be automatically turned off

enum CUpointer_attribute

Pointer information

Values

CU_POINTER_ATTRIBUTE_CONTEXT = 1

The **CUcontext** on which a pointer was allocated or registered

CU_POINTER_ATTRIBUTE_MEMORY_TYPE = 2

The **CUmemorytype** describing the physical location of a pointer

CU_POINTER_ATTRIBUTE_DEVICE_POINTER = 3

The address at which a pointer's memory may be accessed on the device

CU_POINTER_ATTRIBUTE_HOST_POINTER = 4

The address at which a pointer's memory may be accessed on the host

CU_POINTER_ATTRIBUTE_P2P_TOKENS = 5

A pair of tokens for use with the nv-p2p.h Linux kernel interface

CU_POINTER_ATTRIBUTE_SYNC_MEMOPS = 6

Synchronize every synchronous memory operation initiated on this region

CU_POINTER_ATTRIBUTE_BUFFER_ID = 7

A process-wide unique ID for an allocated memory region

CU_POINTER_ATTRIBUTE_IS_MANAGED = 8

Indicates if the pointer points to managed memory

CU_POINTER_ATTRIBUTE_DEVICE_ORDINAL = 9

A device ordinal of a device on which a pointer was allocated or registered

enum CUresourcetype

Resource types

Values

CU_RESOURCE_TYPE_ARRAY = 0x00

Array resource

CU_RESOURCE_TYPE_MIPMAPPED_ARRAY = 0x01

Mipmapped array resource

CU_RESOURCE_TYPE_LINEAR = 0x02

Linear resource

CU_RESOURCE_TYPE_PITCH2D = 0x03

Pitch 2D resource

enum CUresourceViewFormat

Resource view format

Values

CU_RES_VIEW_FORMAT_NONE = 0x00
 No resource view format (use underlying resource format)

CU_RES_VIEW_FORMAT_UINT_1X8 = 0x01
 1 channel unsigned 8-bit integers

CU_RES_VIEW_FORMAT_UINT_2X8 = 0x02
 2 channel unsigned 8-bit integers

CU_RES_VIEW_FORMAT_UINT_4X8 = 0x03
 4 channel unsigned 8-bit integers

CU_RES_VIEW_FORMAT_SINT_1X8 = 0x04
 1 channel signed 8-bit integers

CU_RES_VIEW_FORMAT_SINT_2X8 = 0x05
 2 channel signed 8-bit integers

CU_RES_VIEW_FORMAT_SINT_4X8 = 0x06
 4 channel signed 8-bit integers

CU_RES_VIEW_FORMAT_UINT_1X16 = 0x07
 1 channel unsigned 16-bit integers

CU_RES_VIEW_FORMAT_UINT_2X16 = 0x08
 2 channel unsigned 16-bit integers

CU_RES_VIEW_FORMAT_UINT_4X16 = 0x09
 4 channel unsigned 16-bit integers

CU_RES_VIEW_FORMAT_SINT_1X16 = 0x0a
 1 channel signed 16-bit integers

CU_RES_VIEW_FORMAT_SINT_2X16 = 0x0b
 2 channel signed 16-bit integers

CU_RES_VIEW_FORMAT_SINT_4X16 = 0x0c
 4 channel signed 16-bit integers

CU_RES_VIEW_FORMAT_UINT_1X32 = 0x0d
 1 channel unsigned 32-bit integers

CU_RES_VIEW_FORMAT_UINT_2X32 = 0x0e
 2 channel unsigned 32-bit integers

CU_RES_VIEW_FORMAT_UINT_4X32 = 0x0f
 4 channel unsigned 32-bit integers

CU_RES_VIEW_FORMAT_SINT_1X32 = 0x10
 1 channel signed 32-bit integers

CU_RES_VIEW_FORMAT_SINT_2X32 = 0x11
 2 channel signed 32-bit integers

CU_RES_VIEW_FORMAT_SINT_4X32 = 0x12

4 channel signed 32-bit integers
CU_RES_VIEW_FORMAT_FLOAT_1X16 = 0x13
 1 channel 16-bit floating point
CU_RES_VIEW_FORMAT_FLOAT_2X16 = 0x14
 2 channel 16-bit floating point
CU_RES_VIEW_FORMAT_FLOAT_4X16 = 0x15
 4 channel 16-bit floating point
CU_RES_VIEW_FORMAT_FLOAT_1X32 = 0x16
 1 channel 32-bit floating point
CU_RES_VIEW_FORMAT_FLOAT_2X32 = 0x17
 2 channel 32-bit floating point
CU_RES_VIEW_FORMAT_FLOAT_4X32 = 0x18
 4 channel 32-bit floating point
CU_RES_VIEW_FORMAT_UNSIGNED_BC1 = 0x19
 Block compressed 1
CU_RES_VIEW_FORMAT_UNSIGNED_BC2 = 0x1a
 Block compressed 2
CU_RES_VIEW_FORMAT_UNSIGNED_BC3 = 0x1b
 Block compressed 3
CU_RES_VIEW_FORMAT_UNSIGNED_BC4 = 0x1c
 Block compressed 4 unsigned
CU_RES_VIEW_FORMAT_SIGNED_BC4 = 0x1d
 Block compressed 4 signed
CU_RES_VIEW_FORMAT_UNSIGNED_BC5 = 0x1e
 Block compressed 5 unsigned
CU_RES_VIEW_FORMAT_SIGNED_BC5 = 0x1f
 Block compressed 5 signed
CU_RES_VIEW_FORMAT_UNSIGNED_BC6H = 0x20
 Block compressed 6 unsigned half-float
CU_RES_VIEW_FORMAT_SIGNED_BC6H = 0x21
 Block compressed 6 signed half-float
CU_RES_VIEW_FORMAT_UNSIGNED_BC7 = 0x22
 Block compressed 7

enum CUresult

Error codes

Values

CUDA_SUCCESS = 0

The API call returned with no errors. In the case of query calls, this can also mean that the operation being queried is complete (see [cuEventQuery\(\)](#) and [cuStreamQuery\(\)](#)).

CUDA_ERROR_INVALID_VALUE = 1

This indicates that one or more of the parameters passed to the API call is not within an acceptable range of values.

CUDA_ERROR_OUT_OF_MEMORY = 2

The API call failed because it was unable to allocate enough memory to perform the requested operation.

CUDA_ERROR_NOT_INITIALIZED = 3

This indicates that the CUDA driver has not been initialized with `cuInit()` or that initialization has failed.

CUDA_ERROR_DEINITIALIZED = 4

This indicates that the CUDA driver is in the process of shutting down.

CUDA_ERROR_PROFILER_DISABLED = 5

This indicates profiler is not initialized for this run. This can happen when the application is running with external profiling tools like visual profiler.

CUDA_ERROR_PROFILER_NOT_INITIALIZED = 6

Deprecated This error return is deprecated as of CUDA 5.0. It is no longer an error to attempt to enable/disable the profiling via `cuProfilerStart` or `cuProfilerStop` without initialization.

CUDA_ERROR_PROFILER_ALREADY_STARTED = 7

Deprecated This error return is deprecated as of CUDA 5.0. It is no longer an error to call `cuProfilerStart()` when profiling is already enabled.

CUDA_ERROR_PROFILER_ALREADY_STOPPED = 8

Deprecated This error return is deprecated as of CUDA 5.0. It is no longer an error to call `cuProfilerStop()` when profiling is already disabled.

CUDA_ERROR_NO_DEVICE = 100

This indicates that no CUDA-capable devices were detected by the installed CUDA driver.

CUDA_ERROR_INVALID_DEVICE = 101

This indicates that the device ordinal supplied by the user does not correspond to a valid CUDA device.

CUDA_ERROR_INVALID_IMAGE = 200

This indicates that the device kernel image is invalid. This can also indicate an invalid CUDA module.

CUDA_ERROR_INVALID_CONTEXT = 201

This most frequently indicates that there is no context bound to the current thread. This can also be returned if the context passed to an API call is not a valid handle (such as a context that has had `cuCtxDestroy()` invoked on it). This can also be returned if a user mixes different API versions (i.e. 3010 context with 3020 API calls). See `cuCtxGetApiVersion()` for more details.

CUDA_ERROR_CONTEXT_ALREADY_CURRENT = 202

This indicated that the context being supplied as a parameter to the API call was already the active context. **Deprecated** This error return is deprecated as of CUDA 3.2. It is no longer an error to attempt to push the active context via `cuCtxPushCurrent()`.

CUDA_ERROR_MAP_FAILED = 205

This indicates that a map or register operation has failed.

CUDA_ERROR_UNMAP_FAILED = 206

This indicates that an unmap or unregister operation has failed.

CUDA_ERROR_ARRAY_IS_MAPPED = 207

This indicates that the specified array is currently mapped and thus cannot be destroyed.

CUDA_ERROR_ALREADY_MAPPED = 208

This indicates that the resource is already mapped.

CUDA_ERROR_NO_BINARY_FOR_GPU = 209

This indicates that there is no kernel image available that is suitable for the device.

This can occur when a user specifies code generation options for a particular CUDA source file that do not include the corresponding device configuration.

CUDA_ERROR_ALREADY_ACQUIRED = 210

This indicates that a resource has already been acquired.

CUDA_ERROR_NOT_MAPPED = 211

This indicates that a resource is not mapped.

CUDA_ERROR_NOT_MAPPED_AS_ARRAY = 212

This indicates that a mapped resource is not available for access as an array.

CUDA_ERROR_NOT_MAPPED_AS_POINTER = 213

This indicates that a mapped resource is not available for access as a pointer.

CUDA_ERROR_ECC_UNCORRECTABLE = 214

This indicates that an uncorrectable ECC error was detected during execution.

CUDA_ERROR_UNSUPPORTED_LIMIT = 215

This indicates that the `CUlimit` passed to the API call is not supported by the active device.

CUDA_ERROR_CONTEXT_ALREADY_IN_USE = 216

This indicates that the `CUcontext` passed to the API call can only be bound to a single CPU thread at a time but is already bound to a CPU thread.

CUDA_ERROR_PEER_ACCESS_UNSUPPORTED = 217

This indicates that peer access is not supported across the given devices.

CUDA_ERROR_INVALID_PTX = 218

This indicates that a PTX JIT compilation failed.

CUDA_ERROR_INVALID_GRAPHICS_CONTEXT = 219

This indicates an error with OpenGL or DirectX context.

CUDA_ERROR_NVLINK_UNCORRECTABLE = 220

This indicates that an uncorrectable NVLink error was detected during the execution.

CUDA_ERROR_JIT_COMPILER_NOT_FOUND = 221

This indicates that the PTX JIT compiler library was not found.

CUDA_ERROR_INVALID_SOURCE = 300

This indicates that the device kernel source is invalid.

CUDA_ERROR_FILE_NOT_FOUND = 301

This indicates that the file specified was not found.

CUDA_ERROR_SHARED_OBJECT_SYMBOL_NOT_FOUND = 302

This indicates that a link to a shared object failed to resolve.

CUDA_ERROR_SHARED_OBJECT_INIT_FAILED = 303

This indicates that initialization of a shared object failed.

CUDA_ERROR_OPERATING_SYSTEM = 304

This indicates that an OS call failed.

CUDA_ERROR_INVALID_HANDLE = 400

This indicates that a resource handle passed to the API call was not valid. Resource handles are opaque types like [CUstream](#) and [CUevent](#).

CUDA_ERROR_NOT_FOUND = 500

This indicates that a named symbol was not found. Examples of symbols are global/constant variable names, texture names, and surface names.

CUDA_ERROR_NOT_READY = 600

This indicates that asynchronous operations issued previously have not completed yet. This result is not actually an error, but must be indicated differently than [CUDA_SUCCESS](#) (which indicates completion). Calls that may return this value include [cuEventQuery\(\)](#) and [cuStreamQuery\(\)](#).

CUDA_ERROR_ILLEGAL_ADDRESS = 700

While executing a kernel, the device encountered a load or store instruction on an invalid memory address. This leaves the process in an inconsistent state and any further CUDA work will return the same error. To continue using CUDA, the process must be terminated and relaunched.

CUDA_ERROR_LAUNCH_OUT_OF_RESOURCES = 701

This indicates that a launch did not occur because it did not have appropriate resources. This error usually indicates that the user has attempted to pass too many arguments to the device kernel, or the kernel launch specifies too many threads for the kernel's register count. Passing arguments of the wrong size (i.e. a 64-bit pointer when a 32-bit int is expected) is equivalent to passing too many arguments and can also result in this error.

CUDA_ERROR_LAUNCH_TIMEOUT = 702

This indicates that the device kernel took too long to execute. This can only occur if timeouts are enabled - see the device attribute [CU_DEVICE_ATTRIBUTE_KERNEL_EXEC_TIMEOUT](#) for more information. This leaves the process in an inconsistent state and any further CUDA work will return the same error. To continue using CUDA, the process must be terminated and relaunched.

CUDA_ERROR_LAUNCH_INCOMPATIBLE_TEXTURING = 703

This error indicates a kernel launch that uses an incompatible texturing mode.

CUDA_ERROR_PEER_ACCESS_ALREADY_ENABLED = 704

This error indicates that a call to [cuCtxEnablePeerAccess\(\)](#) is trying to re-enable peer access to a context which has already had peer access to it enabled.

CUDA_ERROR_PEER_ACCESS_NOT_ENABLED = 705

This error indicates that `cuCtxDisablePeerAccess()` is trying to disable peer access which has not been enabled yet via `cuCtxEnablePeerAccess()`.

CUDA_ERROR_PRIMARY_CONTEXT_ACTIVE = 708

This error indicates that the primary context for the specified device has already been initialized.

CUDA_ERROR_CONTEXT_IS_DESTROYED = 709

This error indicates that the context current to the calling thread has been destroyed using `cuCtxDestroy`, or is a primary context which has not yet been initialized.

CUDA_ERROR_ASSERT = 710

A device-side assert triggered during kernel execution. The context cannot be used anymore, and must be destroyed. All existing device memory allocations from this context are invalid and must be reconstructed if the program is to continue using CUDA.

CUDA_ERROR_TOO_MANY_PEERS = 711

This error indicates that the hardware resources required to enable peer access have been exhausted for one or more of the devices passed to `cuCtxEnablePeerAccess()`.

CUDA_ERROR_HOST_MEMORY_ALREADY_REGISTERED = 712

This error indicates that the memory range passed to `cuMemHostRegister()` has already been registered.

CUDA_ERROR_HOST_MEMORY_NOT_REGISTERED = 713

This error indicates that the pointer passed to `cuMemHostUnregister()` does not correspond to any currently registered memory region.

CUDA_ERROR_HARDWARE_STACK_ERROR = 714

While executing a kernel, the device encountered a stack error. This can be due to stack corruption or exceeding the stack size limit. This leaves the process in an inconsistent state and any further CUDA work will return the same error. To continue using CUDA, the process must be terminated and relaunched.

CUDA_ERROR_ILLEGAL_INSTRUCTION = 715

While executing a kernel, the device encountered an illegal instruction. This leaves the process in an inconsistent state and any further CUDA work will return the same error. To continue using CUDA, the process must be terminated and relaunched.

CUDA_ERROR_MISALIGNED_ADDRESS = 716

While executing a kernel, the device encountered a load or store instruction on a memory address which is not aligned. This leaves the process in an inconsistent state and any further CUDA work will return the same error. To continue using CUDA, the process must be terminated and relaunched.

CUDA_ERROR_INVALID_ADDRESS_SPACE = 717

While executing a kernel, the device encountered an instruction which can only operate on memory locations in certain address spaces (global, shared, or local), but was supplied a memory address not belonging to an allowed address space. This leaves the process in an inconsistent state and any further CUDA work will return the same error. To continue using CUDA, the process must be terminated and relaunched.

CUDA_ERROR_INVALID_PC = 718

While executing a kernel, the device program counter wrapped its address space. This leaves the process in an inconsistent state and any further CUDA work will return the same error. To continue using CUDA, the process must be terminated and relaunched.

CUDA_ERROR_LAUNCH_FAILED = 719

An exception occurred on the device while executing a kernel. Common causes include dereferencing an invalid device pointer and accessing out of bounds shared memory. This leaves the process in an inconsistent state and any further CUDA work will return the same error. To continue using CUDA, the process must be terminated and relaunched.

CUDA_ERROR_COOPERATIVE_LAUNCH_TOO_LARGE = 720

This error indicates that the number of blocks launched per grid for a kernel that was launched via either `cuLaunchCooperativeKernel` or `cuLaunchCooperativeKernelMultiDevice` exceeds the maximum number of blocks as allowed by `cuOccupancyMaxActiveBlocksPerMultiprocessor` or `cuOccupancyMaxActiveBlocksPerMultiprocessorWithFlags` times the number of multiprocessors as specified by the device attribute `CU_DEVICE_ATTRIBUTE_MULTIPROCESSOR_COUNT`.

CUDA_ERROR_NOT_PERMITTED = 800

This error indicates that the attempted operation is not permitted.

CUDA_ERROR_NOT_SUPPORTED = 801

This error indicates that the attempted operation is not supported on the current system or device.

CUDA_ERROR_UNKNOWN = 999

This indicates that an unknown internal error has occurred.

enum CUshared_carveout

Shared memory carveout configurations

Values

CU_SHAREDMEM_CARVEOUT_DEFAULT = -1

CU_SHAREDMEM_CARVEOUT_MAX_SHARED = 100

< no preference for shared memory or L1 (default)

CU_SHAREDMEM_CARVEOUT_MAX_L1 = 0

< prefer maximum available shared memory, minimum L1 cache

enum CUsharedconfig

Shared memory configurations

Values

CU_SHARED_MEM_CONFIG_DEFAULT_BANK_SIZE = 0x00

set default shared memory bank size

CU_SHARED_MEM_CONFIG_FOUR_BYTE_BANK_SIZE = 0x01

set shared memory bank width to four bytes

CU_SHARED_MEM_CONFIG_EIGHT_BYTE_BANK_SIZE = 0x02

set shared memory bank width to eight bytes

enum CUstream_flags

Stream creation flags

Values

CU_STREAM_DEFAULT = 0x0

Default stream flag

CU_STREAM_NON_BLOCKING = 0x1

Stream does not synchronize with stream 0 (the NULL stream)

enum CUstreamBatchMemOpType

Operations for `cuStreamBatchMemOp`

Values

CU_STREAM_MEM_OP_WAIT_VALUE_32 = 1

Represents a `cuStreamWaitValue32` operation

CU_STREAM_MEM_OP_WRITE_VALUE_32 = 2

Represents a `cuStreamWriteValue32` operation

CU_STREAM_MEM_OP_WAIT_VALUE_64 = 4

Represents a `cuStreamWaitValue64` operation

CU_STREAM_MEM_OP_WRITE_VALUE_64 = 5

Represents a `cuStreamWriteValue64` operation

CU_STREAM_MEM_OP_FLUSH_REMOTE_WRITES = 3

This has the same effect as `CU_STREAM_WAIT_VALUE_FLUSH`, but as a standalone operation.

enum CUstreamWaitValue_flags

Flags for `cuStreamWaitValue32` and `cuStreamWaitValue64`

Values

CU_STREAM_WAIT_VALUE_GEQ = 0x0

Wait until $(\text{int32_t})(*\text{addr} - \text{value}) \geq 0$ (or int64_t for 64 bit values). Note this is a cyclic comparison which ignores wraparound. (Default behavior.)

CU_STREAM_WAIT_VALUE_EQ = 0x1

Wait until `*addr == value`.

CU_STREAM_WAIT_VALUE_AND = 0x2

Wait until `(*addr & value) != 0`.

CU_STREAM_WAIT_VALUE_NOR = 0x3

Wait until `~(*addr | value) != 0`. Support for this operation can be queried with `cuDeviceGetAttribute()` and

`CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_WAIT_VALUE_NOR`.

CU_STREAM_WAIT_VALUE_FLUSH = 1<<30

Follow the wait operation with a flush of outstanding remote writes. This means that, if a remote write operation is guaranteed to have reached the device before the wait can be satisfied, that write is guaranteed to be visible to downstream device work. The device is permitted to reorder remote writes internally. For example, this flag would be required if two remote writes arrive in a defined order, the wait is satisfied by the second write, and downstream work needs to observe the first write. Support for this operation is restricted to selected platforms and can be queried with `CU_DEVICE_ATTRIBUTE_CAN_USE_WAIT_VALUE_FLUSH`.

enum CUstreamWriteValue_flags

Flags for `cuStreamWriteValue32`

Values

CU_STREAM_WRITE_VALUE_DEFAULT = 0x0

Default behavior

CU_STREAM_WRITE_VALUE_NO_MEMORY_BARRIER = 0x1

Permits the write to be reordered with writes which were issued before it, as a performance optimization. Normally, `cuStreamWriteValue32` will provide a memory fence before the write, which has similar semantics to `__threadfence_system()` but is scoped to the stream rather than a CUDA thread.

typedef struct CUarray_st *CUarray

CUDA array

typedef struct CUctx_st *CUcontext

CUDA context

typedef int CUdevice

CUDA device

`typedef unsigned int CUdeviceptr`

CUDA device pointer `CUdeviceptr` is defined as an unsigned integer type whose size matches the size of a pointer on the target platform.

`typedef struct CUeglStreamConnection_st *CUeglStreamConnection`

CUDA EGLStream Connection

`typedef struct CUevent_st *CUevent`

CUDA event

`typedef struct CUfunc_st *CUfunction`

CUDA function

`typedef struct CUgraphicsResource_st *CUgraphicsResource`

CUDA graphics interop resource

`typedef struct CUmipmappedArray_st *CUmipmappedArray`

CUDA mipmapped array

`typedef struct CUmod_st *CUmodule`

CUDA module

`typedef size_t (CUDA_CB *CUoccupancyB2DSize) (int blockSize)`

Block size to per-block dynamic shared memory mapping for a certain kernel

`typedef struct CUstream_st *CUstream`

CUDA stream

```
typedef void (CUDA_CB *CUstreamCallback) (CUstream
hStream, CUresult status, void* userData)
```

CUDA stream callback

```
typedef unsigned long long CUsurfObject
```

An opaque value that represents a CUDA surface object

```
typedef struct CUsurfref_st *CUsurfref
```

CUDA surface reference

```
typedef unsigned long long CUtexObject
```

An opaque value that represents a CUDA texture object

```
typedef struct CUtexref_st *CUtexref
```

CUDA texture reference

```
#define CU_DEVICE_CPU ((CUdevice)-1)
```

Device that represents the CPU

```
#define CU_DEVICE_INVALID ((CUdevice)-2)
```

Device that represents an invalid device

```
#define CU_IPC_HANDLE_SIZE 64
```

CUDA IPC handle size

```
#define CU_LAUNCH_PARAM_BUFFER_POINTER
((void*)0x01)
```

Indicator that the next value in the `extra` parameter to `cuLaunchKernel` will be a pointer to a buffer containing all kernel parameters used for launching kernel `f`. This buffer needs to honor all alignment/padding requirements of the individual parameters. If `CU_LAUNCH_PARAM_BUFFER_SIZE` is not also specified in the `extra` array, then `CU_LAUNCH_PARAM_BUFFER_POINTER` will have no effect.

#define CU_LAUNCH_PARAM_BUFFER_SIZE ((void*)0x02)

Indicator that the next value in the `extra` parameter to `cuLaunchKernel` will be a pointer to a `size_t` which contains the size of the buffer specified with `CU_LAUNCH_PARAM_BUFFER_POINTER`. It is required that `CU_LAUNCH_PARAM_BUFFER_POINTER` also be specified in the `extra` array if the value associated with `CU_LAUNCH_PARAM_BUFFER_SIZE` is not zero.

#define CU_LAUNCH_PARAM_END ((void*)0x00)

End of array terminator for the `extra` parameter to `cuLaunchKernel`

#define CU_MEMHOSTALLOC_DEVICEMAP 0x02

If set, host memory is mapped into CUDA address space and `cuMemHostGetDevicePointer()` may be called on the host pointer. Flag for `cuMemHostAlloc()`

#define CU_MEMHOSTALLOC_PORTABLE 0x01

If set, host memory is portable between CUDA contexts. Flag for `cuMemHostAlloc()`

#define CU_MEMHOSTALLOC_WRITECOMBINED 0x04

If set, host memory is allocated as write-combined - fast to write, faster to DMA, slow to read except via SSE4 streaming load instruction (MOVNTDQA). Flag for `cuMemHostAlloc()`

#define CU_MEMHOSTREGISTER_DEVICEMAP 0x02

If set, host memory is mapped into CUDA address space and `cuMemHostGetDevicePointer()` may be called on the host pointer. Flag for `cuMemHostRegister()`

#define CU_MEMHOSTREGISTER_IOMEMORY 0x04

If set, the passed memory pointer is treated as pointing to some memory-mapped I/O space, e.g. belonging to a third-party PCIe device. On Windows the flag is a no-op. On Linux that memory is marked as non cache-coherent for the GPU and is expected to be physically contiguous. It may return `CUDA_ERROR_NOT_PERMITTED` if run as an unprivileged user, `CUDA_ERROR_NOT_SUPPORTED` on older Linux kernel versions. On all other platforms, it is not supported and `CUDA_ERROR_NOT_SUPPORTED` is returned. Flag for `cuMemHostRegister()`

#define CU_MEMHOSTREGISTER_PORTABLE 0x01

If set, host memory is portable between CUDA contexts. Flag for `cuMemHostRegister()`

#define CU_PARAM_TR_DEFAULT -1

For texture references loaded into the module, use default texunit from texture reference.

#define CU_STREAM_LEGACY ((CUstream)0x1)

Legacy stream handle

Stream handle that can be passed as a `CUstream` to use an implicit stream with legacy synchronization behavior.

See details of the [synchronization behavior](#).

#define CU_STREAM_PER_THREAD ((CUstream)0x2)

Per-thread stream handle

Stream handle that can be passed as a `CUstream` to use an implicit stream with per-thread synchronization behavior.

See details of the [synchronization behavior](#).

#define CU_TRSA_OVERRIDE_FORMAT 0x01

Override the texref format with a format inferred from the array. Flag for `cuTexRefSetArray()`

#define CU_TRSF_NORMALIZED_COORDINATES 0x02

Use normalized texture coordinates in the range [0,1) instead of [0,dim). Flag for `cuTexRefSetFlags()`

#define CU_TRSF_READ_AS_INTEGER 0x01

Read the texture as integers rather than promoting the values to floats in the range [0,1]. Flag for `cuTexRefSetFlags()`

#define CU_TRSF_SRGB 0x10

Perform sRGB->linear conversion during texture read. Flag for `cuTexRefSetFlags()`

`#define CUDA_ARRAY3D_2DARRAY 0x01`

Deprecated, use `CUDA_ARRAY3D_LAYERED`

`#define CUDA_ARRAY3D_CUBEMAP 0x04`

If set, the CUDA array is a collection of six 2D arrays, representing faces of a cube. The width of such a CUDA array must be equal to its height, and Depth must be six. If `CUDA_ARRAY3D_LAYERED` flag is also set, then the CUDA array is a collection of cubemaps and Depth must be a multiple of six.

`#define CUDA_ARRAY3D_DEPTH_TEXTURE 0x10`

This flag if set indicates that the CUDA array is a `DEPTH_TEXTURE`.

`#define CUDA_ARRAY3D_LAYERED 0x01`

If set, the CUDA array is a collection of layers, where each layer is either a 1D or a 2D array and the Depth member of `CUDA_ARRAY3D_DESCRIPTOR` specifies the number of layers, not the depth of a 3D array.

`#define CUDA_ARRAY3D_SURFACE_LDST 0x02`

This flag must be set in order to bind a surface reference to the CUDA array

`#define CUDA_ARRAY3D_TEXTURE_GATHER 0x08`

This flag must be set in order to perform texture gather operations on a CUDA array.

`#define CUDA_COOPERATIVE_LAUNCH_MULTI_DEVICE_NO_POST_LAUNCH_S 0x02`

If set, any subsequent work pushed in a stream that participated in a call to `cuLaunchCooperativeKernelMultiDevice` will only wait for the kernel launched on the GPU corresponding to that stream to complete before it begins execution.

```
#define
CUDA_COOPERATIVE_LAUNCH_MULTI_DEVICE_NO_PRE_LAUNCH_SYNC
0x01
```

If set, each kernel launched as part of `cuLaunchCooperativeKernelMultiDevice` only waits for prior work in the stream corresponding to that GPU to complete before the kernel begins execution.

```
#define CUDA_VERSION 9020
```

CUDA API version number

```
#define MAX_PLANES 3
```

Maximum number of planes per frame

4.2. Error Handling

This section describes the error handling functions of the low-level CUDA driver application programming interface.

```
CUresult cuGetErrorName (CUresult error, const char
**pStr)
```

Gets the string representation of an error code enum name.

Parameters

error

- Error code to convert to string

pStr

- Address of the string pointer.

Returns

```
CUDA_SUCCESS, CUDA_ERROR_INVALID_VALUE
```

Description

Sets `*pStr` to the address of a NULL-terminated string representation of the name of the enum error code `error`. If the error code is not recognized, `CUDA_ERROR_INVALID_VALUE` will be returned and `*pStr` will be set to the NULL address.

See also:

[CUresult](#), [cudaGetErrorName](#)

CUresult cuGetErrorString (CUresult error, const char **pStr)

Gets the string description of an error code.

Parameters**error**

- Error code to convert to string

pStr

- Address of the string pointer.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_VALUE](#)

Description

Sets *pStr to the address of a NULL-terminated string description of the error code `error`. If the error code is not recognized, [CUDA_ERROR_INVALID_VALUE](#) will be returned and *pStr will be set to the NULL address.

See also:

[CUresult](#), [cudaGetErrorString](#)

4.3. Initialization

This section describes the initialization functions of the low-level CUDA driver application programming interface.

CUresult culnit (unsigned int Flags)

Initialize the CUDA driver API.

Parameters**Flags**

- Initialization flag for CUDA.

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_INVALID_VALUE`,
`CUDA_ERROR_INVALID_DEVICE`

Description

Initializes the driver API and must be called before any other function from the driver API. Currently, the `Flags` parameter must be 0. If `cuInit()` has not been called, any function from the driver API will return `CUDA_ERROR_NOT_INITIALIZED`.



Note that this function may also return error codes from previous, asynchronous launches.

4.4. Version Management

This section describes the version management functions of the low-level CUDA driver application programming interface.

CUresult cuDriverGetVersion (int *driverVersion)

Returns the CUDA driver version.

Parameters**driverVersion**

- Returns the CUDA driver version

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_INVALID_VALUE`

Description

Returns in `*driverVersion` the version number of the installed CUDA driver. This function automatically returns `CUDA_ERROR_INVALID_VALUE` if the `driverVersion` argument is `NULL`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cudaDriverGetVersion`, `cudaRuntimeGetVersion`

4.5. Device Management

This section describes the device management functions of the low-level CUDA driver application programming interface.

CUresult cuDeviceGet (CUdevice *device, int ordinal)

Returns a handle to a compute device.

Parameters

device

- Returned device handle

ordinal

- Device number to get handle for

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_DEVICE`

Description

Returns in `*device` a device handle given an ordinal in the range `[0, cuDeviceGetCount()-1]`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDeviceGetAttribute`, `cuDeviceGetCount`, `cuDeviceGetName`, `cuDeviceGetUuid`,
`cuDeviceTotalMem`

CUresult cuDeviceGetAttribute (int *pi, CUdevice_attribute attrib, CUdevice dev)

Returns information about the device.

Parameters

pi

- Returned device attribute value

attrib

- Device attribute to query

dev

- Device handle

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_DEVICE

Description

Returns in *pi the integer value of the attribute attrib on device dev. The supported attributes are:

- ▶ **CU_DEVICE_ATTRIBUTE_MAX_THREADS_PER_BLOCK**: Maximum number of threads per block;
- ▶ **CU_DEVICE_ATTRIBUTE_MAX_BLOCK_DIM_X**: Maximum x-dimension of a block;
- ▶ **CU_DEVICE_ATTRIBUTE_MAX_BLOCK_DIM_Y**: Maximum y-dimension of a block;
- ▶ **CU_DEVICE_ATTRIBUTE_MAX_BLOCK_DIM_Z**: Maximum z-dimension of a block;
- ▶ **CU_DEVICE_ATTRIBUTE_MAX_GRID_DIM_X**: Maximum x-dimension of a grid;
- ▶ **CU_DEVICE_ATTRIBUTE_MAX_GRID_DIM_Y**: Maximum y-dimension of a grid;
- ▶ **CU_DEVICE_ATTRIBUTE_MAX_GRID_DIM_Z**: Maximum z-dimension of a grid;
- ▶ **CU_DEVICE_ATTRIBUTE_MAX_SHARED_MEMORY_PER_BLOCK**: Maximum amount of shared memory available to a thread block in bytes;
- ▶ **CU_DEVICE_ATTRIBUTE_TOTAL_CONSTANT_MEMORY**: Memory available on device for __constant__ variables in a CUDA C kernel in bytes;
- ▶ **CU_DEVICE_ATTRIBUTE_WARP_SIZE**: Warp size in threads;
- ▶ **CU_DEVICE_ATTRIBUTE_MAX_PITCH**: Maximum pitch in bytes allowed by the memory copy functions that involve memory regions allocated through cuMemAllocPitch();

- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_WIDTH`: Maximum 1D texture width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_LINEAR_WIDTH`: Maximum width for a 1D texture bound to linear memory;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_MIPMAPPED_WIDTH`: Maximum mipmapped 1D texture width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_WIDTH`: Maximum 2D texture width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_HEIGHT`: Maximum 2D texture height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_WIDTH`: Maximum width for a 2D texture bound to linear memory;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_HEIGHT`: Maximum height for a 2D texture bound to linear memory;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_PITCH`: Maximum pitch in bytes for a 2D texture bound to linear memory;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_MIPMAPPED_WIDTH`: Maximum mipmapped 2D texture width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_MIPMAPPED_HEIGHT`: Maximum mipmapped 2D texture height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_WIDTH`: Maximum 3D texture width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_HEIGHT`: Maximum 3D texture height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_DEPTH`: Maximum 3D texture depth;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_WIDTH_ALTERNATE`: Alternate maximum 3D texture width, 0 if no alternate maximum 3D texture size is supported;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_HEIGHT_ALTERNATE`: Alternate maximum 3D texture height, 0 if no alternate maximum 3D texture size is supported;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE3D_DEPTH_ALTERNATE`: Alternate maximum 3D texture depth, 0 if no alternate maximum 3D texture size is supported;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURECUBEMAP_WIDTH`: Maximum cubemap texture width or height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_LAYERED_WIDTH`: Maximum 1D layered texture width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_LAYERED_LAYERS`: Maximum layers in a 1D layered texture;

- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_WIDTH`: Maximum 2D layered texture width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_HEIGHT`: Maximum 2D layered texture height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LAYERED_LAYERS`: Maximum layers in a 2D layered texture;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURECUBEMAP_LAYERED_WIDTH`: Maximum cubemap layered texture width or height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURECUBEMAP_LAYERED_LAYERS`: Maximum layers in a cubemap layered texture;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE1D_WIDTH`: Maximum 1D surface width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_WIDTH`: Maximum 2D surface width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_HEIGHT`: Maximum 2D surface height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE3D_WIDTH`: Maximum 3D surface width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE3D_HEIGHT`: Maximum 3D surface height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE3D_DEPTH`: Maximum 3D surface depth;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE1D_LAYERED_WIDTH`: Maximum 1D layered surface width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE1D_LAYERED_LAYERS`: Maximum layers in a 1D layered surface;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_LAYERED_WIDTH`: Maximum 2D layered surface width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_LAYERED_HEIGHT`: Maximum 2D layered surface height;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACE2D_LAYERED_LAYERS`: Maximum layers in a 2D layered surface;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACECUBEMAP_WIDTH`: Maximum cubemap surface width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACECUBEMAP_LAYERED_WIDTH`: Maximum cubemap layered surface width;
- ▶ `CU_DEVICE_ATTRIBUTE_MAXIMUM_SURFACECUBEMAP_LAYERED_LAYERS`: Maximum layers in a cubemap layered surface;
- ▶ `CU_DEVICE_ATTRIBUTE_MAX_REGISTERS_PER_BLOCK`: Maximum number of 32-bit registers available to a thread block;
- ▶ `CU_DEVICE_ATTRIBUTE_CLOCK_RATE`: The typical clock frequency in kilohertz;

- ▶ **CU_DEVICE_ATTRIBUTE_TEXTURE_ALIGNMENT**: Alignment requirement; texture base addresses aligned to textureAlign bytes do not need an offset applied to texture fetches;
- ▶ **CU_DEVICE_ATTRIBUTE_TEXTURE_PITCH_ALIGNMENT**: Pitch alignment requirement for 2D texture references bound to pitched memory;
- ▶ **CU_DEVICE_ATTRIBUTE_GPU_OVERLAP**: 1 if the device can concurrently copy memory between host and device while executing a kernel, or 0 if not;
- ▶ **CU_DEVICE_ATTRIBUTE_MULTIPROCESSOR_COUNT**: Number of multiprocessors on the device;
- ▶ **CU_DEVICE_ATTRIBUTE_KERNEL_EXEC_TIMEOUT**: 1 if there is a run time limit for kernels executed on the device, or 0 if not;
- ▶ **CU_DEVICE_ATTRIBUTE_INTEGRATED**: 1 if the device is integrated with the memory subsystem, or 0 if not;
- ▶ **CU_DEVICE_ATTRIBUTE_CAN_MAP_HOST_MEMORY**: 1 if the device can map host memory into the CUDA address space, or 0 if not;
- ▶ **CU_DEVICE_ATTRIBUTE_COMPUTE_MODE**: Compute mode that device is currently in. Available modes are as follows:
 - ▶ **CU_COMPUTEMODE_DEFAULT**: Default mode - Device is not restricted and can have multiple CUDA contexts present at a single time.
 - ▶ **CU_COMPUTEMODE_PROHIBITED**: Compute-prohibited mode - Device is prohibited from creating new CUDA contexts.
 - ▶ **CU_COMPUTEMODE_EXCLUSIVE_PROCESS**: Compute-exclusive-process mode - Device can have only one context used by a single process at a time.
- ▶ **CU_DEVICE_ATTRIBUTE_CONCURRENT_KERNELS**: 1 if the device supports executing multiple kernels within the same context simultaneously, or 0 if not. It is not guaranteed that multiple kernels will be resident on the device concurrently so this feature should not be relied upon for correctness;
- ▶ **CU_DEVICE_ATTRIBUTE_ECC_ENABLED**: 1 if error correction is enabled on the device, 0 if error correction is disabled or not supported by the device;
- ▶ **CU_DEVICE_ATTRIBUTE_PCI_BUS_ID**: PCI bus identifier of the device;
- ▶ **CU_DEVICE_ATTRIBUTE_PCI_DEVICE_ID**: PCI device (also known as slot) identifier of the device;
- ▶ **CU_DEVICE_ATTRIBUTE_PCI_DOMAIN_ID**: PCI domain identifier of the device
- ▶ **CU_DEVICE_ATTRIBUTE_TCC_DRIVER**: 1 if the device is using a TCC driver. TCC is only available on Tesla hardware running Windows Vista or later;
- ▶ **CU_DEVICE_ATTRIBUTE_MEMORY_CLOCK_RATE**: Peak memory clock frequency in kilohertz;
- ▶ **CU_DEVICE_ATTRIBUTE_GLOBAL_MEMORY_BUS_WIDTH**: Global memory bus width in bits;
- ▶ **CU_DEVICE_ATTRIBUTE_L2_CACHE_SIZE**: Size of L2 cache in bytes. 0 if the device doesn't have L2 cache;

- ▶ `CU_DEVICE_ATTRIBUTE_MAX_THREADS_PER_MULTIPROCESSOR`: Maximum resident threads per multiprocessor;
- ▶ `CU_DEVICE_ATTRIBUTE_UNIFIED_ADDRESSING`: 1 if the device shares a unified address space with the host, or 0 if not;
- ▶ `CU_DEVICE_ATTRIBUTE_COMPUTE_CAPABILITY_MAJOR`: Major compute capability version number;
- ▶ `CU_DEVICE_ATTRIBUTE_COMPUTE_CAPABILITY_MINOR`: Minor compute capability version number;
- ▶ `CU_DEVICE_ATTRIBUTE_GLOBAL_L1_CACHE_SUPPORTED`: 1 if device supports caching globals in L1 cache, 0 if caching globals in L1 cache is not supported by the device;
- ▶ `CU_DEVICE_ATTRIBUTE_LOCAL_L1_CACHE_SUPPORTED`: 1 if device supports caching locals in L1 cache, 0 if caching locals in L1 cache is not supported by the device;
- ▶ `CU_DEVICE_ATTRIBUTE_MAX_SHARED_MEMORY_PER_MULTIPROCESSOR`: Maximum amount of shared memory available to a multiprocessor in bytes; this amount is shared by all thread blocks simultaneously resident on a multiprocessor;
- ▶ `CU_DEVICE_ATTRIBUTE_MAX_REGISTERS_PER_MULTIPROCESSOR`: Maximum number of 32-bit registers available to a multiprocessor; this number is shared by all thread blocks simultaneously resident on a multiprocessor;
- ▶ `CU_DEVICE_ATTRIBUTE_MANAGED_MEMORY`: 1 if device supports allocating managed memory on this system, 0 if allocating managed memory is not supported by the device on this system.
- ▶ `CU_DEVICE_ATTRIBUTE_MULTI_GPU_BOARD`: 1 if device is on a multi-GPU board, 0 if not.
- ▶ `CU_DEVICE_ATTRIBUTE_MULTI_GPU_BOARD_GROUP_ID`: Unique identifier for a group of devices associated with the same board. Devices on the same multi-GPU board will share the same identifier.
- ▶ `CU_DEVICE_ATTRIBUTE_HOST_NATIVE_ATOMIC_SUPPORTED`: 1 if Link between the device and the host supports native atomic operations.
- ▶ `CU_DEVICE_ATTRIBUTE_SINGLE_TO_DOUBLE_PRECISION_PERF_RATIO`: Ratio of single precision performance (in floating-point operations per second) to double precision performance.
- ▶ `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS`: Device supports coherently accessing pageable memory without calling `cudaHostRegister` on it.
- ▶ `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`: Device can coherently access managed memory concurrently with the CPU.
- ▶ `CU_DEVICE_ATTRIBUTE_COMPUTE_PREEMPTION_SUPPORTED`: Device supports Compute Preemption.
- ▶ `CU_DEVICE_ATTRIBUTE_CAN_USE_HOST_POINTER_FOR_REGISTERED_MEM`: Device can access host registered memory at the same virtual address as the CPU.

- ▶ **CU_DEVICE_ATTRIBUTE_MAX_SHARED_MEMORY_PER_BLOCK_OPTIN:**
The maximum per block shared memory size supported on this device. This is the maximum value that can be opted into when using the `cuFuncSetAttribute()` call. For more details see **CU_FUNC_ATTRIBUTE_MAX_DYNAMIC_SHARED_SIZE_BYTES**
- ▶ **CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS_USES_HOST_PAGE_TABLES:**
Device accesses pageable memory via the host's page tables.
- ▶ **CU_DEVICE_ATTRIBUTE_DIRECT_MANAGED_MEM_ACCESS_FROM_HOST:**
The host can directly access managed memory on the device without migration.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDeviceGetCount`, `cuDeviceGetName`, `cuDeviceGetUuid`, `cuDeviceGet`, `cuDeviceTotalMem`, `cudaDeviceGetAttribute`, `cudaGetDeviceProperties`

CUresult cuDeviceGetCount (int *count)

Returns the number of compute-capable devices.

Parameters

count

- Returned number of compute-capable devices

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Returns in `*count` the number of devices with compute capability greater than or equal to 2.0 that are available for execution. If there is no such device, `cuDeviceGetCount()` returns 0.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDeviceGetAttribute`, `cuDeviceGetName`, `cuDeviceGetUuid`, `cuDeviceGet`,
`cuDeviceTotalMem`, `cudaGetDeviceCount`

CUresult cuDeviceGetName (char *name, int len, CUdevice dev)

Returns an identifier string for the device.

Parameters

name

- Returned identifier string for the device

len

- Maximum length of string to store in name

dev

- Device to get identifier string for

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_DEVICE`

Description

Returns an ASCII string identifying the device `dev` in the NULL-terminated string pointed to by `name`. `len` specifies the maximum length of the string that may be returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDeviceGetAttribute`, `cuDeviceGetCount`, `cuDeviceGet`, `cuDeviceTotalMem`,
`cudaGetDeviceProperties`

CUresult cuDeviceGetUuid (CUuuid *uuid, CUdevice dev)

Return an UUID for the device.

Parameters

uuid

- Returned UUID

dev

- Device to get identifier string for

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_INVALID_DEVICE

Description

Returns 16-octets identifying the device `dev` in the structure pointed by the `uuid`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuDeviceGetAttribute](#), [cuDeviceGetCount](#), [cuDeviceGetName](#), [cuDeviceGet](#),
[cuDeviceTotalMem](#), [cudaGetDeviceProperties](#)

CUresult cuDeviceTotalMem (size_t *bytes, CUdevice dev)

Returns the total amount of memory on the device.

Parameters**bytes**

- Returned memory available on device in bytes

dev

- Device handle

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_DEVICE

Description

Returns in `*bytes` the total amount of memory available on the device `dev` in bytes.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDeviceGetAttribute`, `cuDeviceGetCount`, `cuDeviceGetName`, `cuDeviceGetUuid`, `cuDeviceGet`, `cudaMemGetInfo`

4.6. Device Management [DEPRECATED]

This section describes the device management functions of the low-level CUDA driver application programming interface.

`CUresult cuDeviceComputeCapability (int *major, int *minor, CUdevice dev)`

Returns the compute capability of the device.

Parameters

major

- Major revision number

minor

- Minor revision number

dev

- Device handle

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_DEVICE`

Description

Deprecated

This function was deprecated as of CUDA 5.0 and its functionality superseded by `cuDeviceGetAttribute()`.

Returns in `*major` and `*minor` the major and minor revision numbers that define the compute capability of the device `dev`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuDeviceGetAttribute](#), [cuDeviceGetCount](#), [cuDeviceGetName](#), [cuDeviceGetUuid](#), [cuDeviceGet](#), [cuDeviceTotalMem](#)

CUresult cuDeviceGetProperties (CUdevprop *prop, CUdevice dev)

Returns properties for a selected device.

Parameters

prop

- Returned properties of device

dev

- Device to get properties for

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_DEVICE

Description

Deprecated

This function was deprecated as of CUDA 5.0 and replaced by [cuDeviceGetAttribute\(\)](#).

Returns in *prop the properties of device dev. The [CUdevprop](#) structure is defined as:

```
↑ typedef struct CUdevprop_st {
    int maxThreadsPerBlock;
    int maxThreadsDim[3];
    int maxGridSize[3];
    int sharedMemPerBlock;
    int totalConstantMemory;
    int SIMDWidth;
    int memPitch;
    int regsPerBlock;
    int clockRate;
    int textureAlign
} CUdevprop;
```

where:

- ▶ maxThreadsPerBlock is the maximum number of threads per block;
- ▶ maxThreadsDim[3] is the maximum sizes of each dimension of a block;

- ▶ `maxGridSize[3]` is the maximum sizes of each dimension of a grid;
- ▶ `sharedMemPerBlock` is the total amount of shared memory available per block in bytes;
- ▶ `totalConstantMemory` is the total amount of constant memory available on the device in bytes;
- ▶ `SIMDWidth` is the warp size;
- ▶ `memPitch` is the maximum pitch allowed by the memory copy functions that involve memory regions allocated through `cuMemAllocPitch()`;
- ▶ `regsPerBlock` is the total number of registers available per block;
- ▶ `clockRate` is the clock frequency in kilohertz;
- ▶ `textureAlign` is the alignment requirement; texture base addresses that are aligned to `textureAlign` bytes do not need an offset applied to texture fetches.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuDeviceGetAttribute](#), [cuDeviceGetCount](#), [cuDeviceGetName](#), [cuDeviceGetUuid](#), [cuDeviceGet](#), [cuDeviceTotalMem](#)

4.7. Primary Context Management

This section describes the primary context management functions of the low-level CUDA driver application programming interface.

The primary context is unique per device and shared with the CUDA runtime API. These functions allow integration with other libraries using CUDA.

CUresult cuDevicePrimaryCtxGetState (CUdevice dev, unsigned int *flags, int *active)

Get the state of the primary context.

Parameters

dev

- Device to get primary context flags for

flags

- Pointer to store flags

active

- Pointer to store context state; 0 = inactive, 1 = active

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_DEVICE,
 CUDA_ERROR_INVALID_VALUE,

Description

Returns in `*flags` the flags for the primary context of `dev`, and in `*active` whether it is active. See `cuDevicePrimaryCtxSetFlags` for flag values.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDevicePrimaryCtxSetFlags`, `cuCtxGetFlags`, `cudaGetDeviceFlags`

CUresult cuDevicePrimaryCtxRelease (CUdevice dev)

Release the primary context on the GPU.

Parameters**dev**

- Device which primary context is released

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_DEVICE

Description

Releases the primary context interop on the device by decreasing the usage count by 1. If the usage drops to 0 the primary context of device `dev` will be destroyed regardless of how many threads it is current to.

Please note that unlike `cuCtxDestroy()` this method does not pop the context from stack in any circumstances.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDevicePrimaryCtxRetain`, `cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`, `cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPopCurrent`, `cuCtxPushCurrent`, `cuCtxSetCacheConfig`, `cuCtxSetLimit`, `cuCtxSynchronize`

CUresult cuDevicePrimaryCtxReset (CUdevice dev)

Destroy all allocations and reset all state on the primary context.

Parameters

dev

- Device for which primary context is destroyed

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_DEVICE`,
`CUDA_ERROR_PRIMARY_CONTEXT_ACTIVE`

Description

Explicitly destroys and cleans up all resources associated with the current device in the current process.

Note that it is responsibility of the calling function to ensure that no other module in the process is using the device any more. For that reason it is recommended to use `cuDevicePrimaryCtxRelease()` in most cases. However it is safe for other modules to call `cuDevicePrimaryCtxRelease()` even after resetting the device.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDevicePrimaryCtxRetain`, `cuDevicePrimaryCtxRelease`, `cuCtxGetApiVersion`,
`cuCtxGetCacheConfig`, `cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`,
`cuCtxPopCurrent`, `cuCtxPushCurrent`, `cuCtxSetCacheConfig`, `cuCtxSetLimit`,
`cuCtxSynchronize`, `cudaDeviceReset`

CUresult cuDevicePrimaryCtxRetain (CUcontext *pctx, CUdevice dev)

Retain the primary context on the GPU.

Parameters

pctx

- Returned context handle of the new context

dev

- Device for which primary context is requested

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_DEVICE, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Retains the primary context on the device, creating it if necessary, increasing its usage count. The caller must call [cuDevicePrimaryCtxRelease\(\)](#) when done using the context. Unlike [cuCtxCreate\(\)](#) the newly created context is not pushed onto the stack.

Context creation will fail with [CUDA_ERROR_UNKNOWN](#) if the compute mode of the device is [CU_COMPUTEMODE_PROHIBITED](#). The function [cuDeviceGetAttribute\(\)](#) can be used with [CU_DEVICE_ATTRIBUTE_COMPUTE_MODE](#) to determine the compute mode of the device. The nvidia-smi tool can be used to set the compute mode for devices. Documentation for nvidia-smi can be obtained by passing a -h option to it.

Please note that the primary context always supports pinned allocations. Other flags can be specified by [cuDevicePrimaryCtxSetFlags\(\)](#).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuDevicePrimaryCtxRelease](#), [cuDevicePrimaryCtxSetFlags](#), [cuCtxCreate](#),
[cuCtxGetApiVersion](#), [cuCtxGetCacheConfig](#), [cuCtxGetDevice](#), [cuCtxGetFlags](#),
[cuCtxGetLimit](#), [cuCtxPopCurrent](#), [cuCtxPushCurrent](#), [cuCtxSetCacheConfig](#),
[cuCtxSetLimit](#), [cuCtxSynchronize](#)

CUresult cuDevicePrimaryCtxSetFlags (CUdevice dev, unsigned int flags)

Set flags for the primary context.

Parameters

dev

- Device for which the primary context flags are set

flags

- New flags for the device

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_DEVICE,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_PRIMARY_CONTEXT_ACTIVE

Description

Sets the flags for the primary context on the device overwriting perviously set ones. If the primary context is already created `CUDA_ERROR_PRIMARY_CONTEXT_ACTIVE` is returned.

The three LSBs of the `flags` parameter can be used to control how the OS thread, which owns the CUDA context at the time of an API call, interacts with the OS scheduler when waiting for results from the GPU. Only one of the scheduling flags can be set when creating a context.

- ▶ `CU_CTX_SCHED_SPIN`: Instruct CUDA to actively spin when waiting for results from the GPU. This can decrease latency when waiting for the GPU, but may lower the performance of CPU threads if they are performing work in parallel with the CUDA thread.
- ▶ `CU_CTX_SCHED_YIELD`: Instruct CUDA to yield its thread when waiting for results from the GPU. This can increase latency when waiting for the GPU, but can increase the performance of CPU threads performing work in parallel with the GPU.
- ▶ `CU_CTX_SCHED_BLOCKING_SYNC`: Instruct CUDA to block the CPU thread on a synchronization primitive when waiting for the GPU to finish work.
- ▶ `CU_CTX_BLOCKING_SYNC`: Instruct CUDA to block the CPU thread on a synchronization primitive when waiting for the GPU to finish work.

Deprecated: This flag was deprecated as of CUDA 4.0 and was replaced with `CU_CTX_SCHED_BLOCKING_SYNC`.

- ▶ **CU_CTX_SCHED_AUTO**: The default value if the `flags` parameter is zero, uses a heuristic based on the number of active CUDA contexts in the process `C` and the number of logical processors in the system `P`. If $C > P$, then CUDA will yield to other OS threads when waiting for the GPU (**CU_CTX_SCHED_YIELD**), otherwise CUDA will not yield while waiting for results and actively spin on the processor (**CU_CTX_SCHED_SPIN**). However, on low power devices like Tegra, it always defaults to **CU_CTX_SCHED_BLOCKING_SYNC**.
- ▶ **CU_CTX_LMEM_RESIZE_TO_MAX**: Instruct CUDA to not reduce local memory after resizing local memory for a kernel. This can prevent thrashing by local memory allocations when launching many kernels with high local memory usage at the cost of potentially increased memory usage.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuDevicePrimaryCtxRetain](#), [cuDevicePrimaryCtxGetState](#), [cuCtxCreate](#), [cuCtxGetFlags](#), [cudaSetDeviceFlags](#)

4.8. Context Management

This section describes the context management functions of the low-level CUDA driver application programming interface.

CUresult cuCtxCreate (CUcontext *pctx, unsigned int flags, CUdevice dev)

Create a CUDA context.

Parameters

pctx

- Returned context handle of the new context

flags

- Context creation flags

dev

- Device to create context on

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_DEVICE, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Creates a new CUDA context and associates it with the calling thread. The `flags` parameter is described below. The context is created with a usage count of 1 and the caller of `cuCtxCreate()` must call `cuCtxDestroy()` or when done using the context. If a context is already current to the thread, it is supplanted by the newly created context and may be restored by a subsequent call to `cuCtxPopCurrent()`.

The three LSBs of the `flags` parameter can be used to control how the OS thread, which owns the CUDA context at the time of an API call, interacts with the OS scheduler when waiting for results from the GPU. Only one of the scheduling flags can be set when creating a context.

- ▶ **CU_CTX_SCHED_SPIN**: Instruct CUDA to actively spin when waiting for results from the GPU. This can decrease latency when waiting for the GPU, but may lower the performance of CPU threads if they are performing work in parallel with the CUDA thread.
- ▶ **CU_CTX_SCHED_YIELD**: Instruct CUDA to yield its thread when waiting for results from the GPU. This can increase latency when waiting for the GPU, but can increase the performance of CPU threads performing work in parallel with the GPU.
- ▶ **CU_CTX_SCHED_BLOCKING_SYNC**: Instruct CUDA to block the CPU thread on a synchronization primitive when waiting for the GPU to finish work.
- ▶ **CU_CTX_BLOCKING_SYNC**: Instruct CUDA to block the CPU thread on a synchronization primitive when waiting for the GPU to finish work.

Deprecated: This flag was deprecated as of CUDA 4.0 and was replaced with **CU_CTX_SCHED_BLOCKING_SYNC**.

- ▶ **CU_CTX_SCHED_AUTO**: The default value if the `flags` parameter is zero, uses a heuristic based on the number of active CUDA contexts in the process C and the number of logical processors in the system P . If $C > P$, then CUDA will yield to other OS threads when waiting for the GPU (**CU_CTX_SCHED_YIELD**), otherwise CUDA will not yield while waiting for results and actively spin on the processor (**CU_CTX_SCHED_SPIN**). However, on low power devices like Tegra, it always defaults to **CU_CTX_SCHED_BLOCKING_SYNC**.
- ▶ **CU_CTX_MAP_HOST**: Instruct CUDA to support mapped pinned allocations. This flag must be set in order to allocate pinned host memory that is accessible to the GPU.

- ▶ `CU_CTX_LMEM_RESIZE_TO_MAX`: Instruct CUDA to not reduce local memory after resizing local memory for a kernel. This can prevent thrashing by local memory allocations when launching many kernels with high local memory usage at the cost of potentially increased memory usage.

Context creation will fail with `CUDA_ERROR_UNKNOWN` if the compute mode of the device is `CU_COMPUTEMODE_PROHIBITED`. The function `cuDeviceGetAttribute()` can be used with `CU_DEVICE_ATTRIBUTE_COMPUTE_MODE` to determine the compute mode of the device. The `nvidia-smi` tool can be used to set the compute mode for * devices. Documentation for `nvidia-smi` can be obtained by passing a `-h` option to it.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`, `cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPopCurrent`, `cuCtxPushCurrent`, `cuCtxSetCacheConfig`, `cuCtxSetLimit`, `cuCtxSynchronize`

CUresult cuCtxDestroy (CUcontext ctx)

Destroy a CUDA context.

Parameters

`ctx`
- Context to destroy

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Destroys the CUDA context specified by `ctx`. The context `ctx` will be destroyed regardless of how many threads it is current to. It is the responsibility of the calling function to ensure that no API call issues using `ctx` while `cuCtxDestroy()` is executing.

If `ctx` is current to the calling thread then `ctx` will also be popped from the current thread's context stack (as though `cuCtxPopCurrent()` were called). If `ctx` is current to other threads, then `ctx` will remain current to those threads, and attempting to access `ctx` from those threads will result in the error `CUDA_ERROR_CONTEXT_IS_DESTROYED`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuCtxGetApiVersion](#), [cuCtxGetCacheConfig](#), [cuCtxGetDevice](#), [cuCtxGetFlags](#), [cuCtxGetLimit](#), [cuCtxPopCurrent](#), [cuCtxPushCurrent](#), [cuCtxSetCacheConfig](#), [cuCtxSetLimit](#), [cuCtxSynchronize](#)

CUresult cuCtxGetApiVersion (CUcontext ctx, unsigned int *version)

Gets the context's API version.

Parameters

ctx

- Context to check

version

- Pointer to version

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_UNKNOWN](#)

Description

Returns a version number in `version` corresponding to the capabilities of the context (e.g. 3010 or 3020), which library developers can use to direct callers to a specific API version. If `ctx` is NULL, returns the API version used to create the currently bound context.

Note that new API versions are only introduced when context capabilities are changed that break binary compatibility, so the API version and driver version may be different. For example, it is valid for the API version to be 3020 while the driver version is 4020.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`,
`cuCtxPopCurrent`, `cuCtxPushCurrent`, `cuCtxSetCacheConfig`, `cuCtxSetLimit`,
`cuCtxSynchronize`

CUresult cuCtxGetCacheConfig (CUfunc_cache *pconfig)

Returns the preferred cache configuration for the current context.

Parameters

pconfig

- Returned cache configuration

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

On devices where the L1 cache and shared memory use the same hardware resources, this function returns through `pconfig` the preferred cache configuration for the current context. This is only a preference. The driver will use the requested configuration if possible, but it is free to choose a different configuration if required to execute functions.

This will return a `pconfig` of `CU_FUNC_CACHE_PREFER_NONE` on devices where the size of the L1 cache and shared memory are fixed.

The supported cache configurations are:

- ▶ `CU_FUNC_CACHE_PREFER_NONE`: no preference for shared memory or L1 (default)
- ▶ `CU_FUNC_CACHE_PREFER_SHARED`: prefer larger shared memory and smaller L1 cache
- ▶ `CU_FUNC_CACHE_PREFER_L1`: prefer larger L1 cache and smaller shared memory
- ▶ `CU_FUNC_CACHE_PREFER_EQUAL`: prefer equal sized L1 cache and shared memory



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuCtxDestroy](#), [cuCtxGetApiVersion](#), [cuCtxGetDevice](#), [cuCtxGetFlags](#), [cuCtxGetLimit](#), [cuCtxPopCurrent](#), [cuCtxPushCurrent](#), [cuCtxSetCacheConfig](#), [cuCtxSetLimit](#), [cuCtxSynchronize](#), [cuFuncSetCacheConfig](#), [cudaDeviceGetCacheConfig](#)

CUresult cuCtxGetCurrent (CUcontext *pctx)

Returns the CUDA context bound to the calling CPU thread.

Parameters

pctx

- Returned context handle

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#),

Description

Returns in *pctx the CUDA context bound to the calling CPU thread. If no context is bound to the calling CPU thread then *pctx is set to NULL and [CUDA_SUCCESS](#) is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxSetCurrent](#), [cuCtxCreate](#), [cuCtxDestroy](#), [cudaGetDevice](#)

CUresult cuCtxGetDevice (CUdevice *device)

Returns the device ID for the current context.

Parameters

device

- Returned device ID for the current context

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#),

Description

Returns in `*device` the ordinal of the current context's device.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuCtxDestroy](#), [cuCtxGetApiVersion](#), [cuCtxGetCacheConfig](#), [cuCtxGetFlags](#), [cuCtxGetLimit](#), [cuCtxPopCurrent](#), [cuCtxPushCurrent](#), [cuCtxSetCacheConfig](#), [cuCtxSetLimit](#), [cuCtxSynchronize](#), [cudaGetDevice](#)

CUresult cuCtxGetFlags (unsigned int *flags)

Returns the flags for the current context.

Parameters

flags

- Pointer to store flags of current context

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#),

Description

Returns in `*flags` the flags of the current context. See [cuCtxCreate](#) for flag values.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuCtxGetApiVersion](#), [cuCtxGetCacheConfig](#), [cuCtxGetCurrent](#),
[cuCtxGetDevice](#) [cuCtxGetLimit](#), [cuCtxGetSharedMemConfig](#),
[cuCtxGetStreamPriorityRange](#), [cudaGetDeviceFlags](#)

CUresult cuCtxGetLimit (size_t *pvalue, CUlimit limit)

Returns resource limits.

Parameters

pvalue

- Returned size of limit

limit

- Limit to query

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_VALUE,
CUDA_ERROR_UNSUPPORTED_LIMIT

Description

Returns in *pvalue the current size of limit. The supported CUlimit values are:

- ▶ **CU_LIMIT_STACK_SIZE**: stack size in bytes of each GPU thread.
- ▶ **CU_LIMIT_PRINTF_FIFO_SIZE**: size in bytes of the FIFO used by the printf() device system call.
- ▶ **CU_LIMIT_MALLOC_HEAP_SIZE**: size in bytes of the heap used by the malloc() and free() device system calls.
- ▶ **CU_LIMIT_DEV_RUNTIME_SYNC_DEPTH**: maximum grid depth at which a thread can issue the device runtime call `cudaDeviceSynchronize()` to wait on child grid launches to complete.
- ▶ **CU_LIMIT_DEV_RUNTIME_PENDING_LAUNCH_COUNT**: maximum number of outstanding device runtime launches that can be made from this context.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`,
`cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxPopCurrent`, `cuCtxPushCurrent`,
`cuCtxSetCacheConfig`, `cuCtxSetLimit`, `cuCtxSynchronize`, `cudaDeviceGetLimit`

CUresult cuCtxGetSharedMemConfig (CUsharedconfig *pConfig)

Returns the current shared memory configuration for the current context.

Parameters

pConfig

- returned shared memory configuration

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

This function will return in `pConfig` the current size of shared memory banks in the current context. On devices with configurable shared memory banks, `cuCtxSetSharedMemConfig` can be used to change this setting, so that all subsequent kernel launches will by default use the new bank size. When `cuCtxGetSharedMemConfig` is called on devices without configurable shared memory, it will return the fixed bank size of the hardware.

The returned bank configurations can be either:

- ▶ `CU_SHARED_MEM_CONFIG_FOUR_BYTE_BANK_SIZE`: shared memory bank width is four bytes.
- ▶ `CU_SHARED_MEM_CONFIG_EIGHT_BYTE_BANK_SIZE`: shared memory bank width will eight bytes.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`,
`cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPopCurrent`, `cuCtxPushCurrent`,
`cuCtxSetLimit`, `cuCtxSynchronize`, `cuCtxGetSharedMemConfig`, `cuFuncSetCacheConfig`,
`cudaDeviceGetSharedMemConfig`

CUresult cuCtxGetStreamPriorityRange (int *leastPriority, int *greatestPriority)

Returns numerical values that correspond to the least and greatest stream priorities.

Parameters

leastPriority

- Pointer to an int in which the numerical value for least stream priority is returned

greatestPriority

- Pointer to an int in which the numerical value for greatest stream priority is returned

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_VALUE,

Description

Returns in `*leastPriority` and `*greatestPriority` the numerical values that correspond to the least and greatest stream priorities respectively. Stream priorities follow a convention where lower numbers imply greater priorities. The range of meaningful stream priorities is given by [`*greatestPriority`, `*leastPriority`]. If the user attempts to create a stream with a priority value that is outside the meaningful range as specified by this API, the priority is automatically clamped down or up to either `*leastPriority` or `*greatestPriority` respectively. See [cuStreamCreateWithPriority](#) for details on creating a priority stream. A NULL may be passed in for `*leastPriority` or `*greatestPriority` if the value is not desired.

This function will return '0' in both `*leastPriority` and `*greatestPriority` if the current context's device does not support stream priorities (see [cuDeviceGetAttribute](#)).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamCreateWithPriority](#), [cuStreamGetPriority](#), [cuCtxGetDevice](#), [cuCtxGetFlags](#), [cuCtxSetLimit](#), [cuCtxSynchronize](#), [cudaDeviceGetStreamPriorityRange](#)

CUresult cuCtxPopCurrent (CUcontext *pctx)

Pops the current CUDA context from the current CPU thread.

Parameters

pctx

- Returned new context handle

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT

Description

Pops the current CUDA context from the CPU thread and passes back the old context handle in *pctx. That context may then be made current to a different CPU thread by calling `cuCtxPushCurrent()`.

If a context was current to the CPU thread before `cuCtxCreate()` or `cuCtxPushCurrent()` was called, this function makes that context current to the CPU thread again.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`,
`cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPushCurrent`,
`cuCtxSetCacheConfig`, `cuCtxSetLimit`, `cuCtxSynchronize`

CUresult cuCtxPushCurrent (CUcontext ctx)

Pushes a context on the current CPU thread.

Parameters

ctx

- Context to push

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Pushes the given context `ctx` onto the CPU thread's stack of current contexts. The specified context becomes the CPU thread's current context, so all CUDA functions that operate on the current context are affected.

The previous current context may be made current again by calling `cuCtxDestroy()` or `cuCtxPopCurrent()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`, `cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPopCurrent`, `cuCtxSetCacheConfig`, `cuCtxSetLimit`, `cuCtxSynchronize`

CUresult cuCtxSetCacheConfig (CUfunc_cache config)

Sets the preferred cache configuration for the current context.

Parameters

`config`

- Requested cache configuration

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

On devices where the L1 cache and shared memory use the same hardware resources, this sets through `config` the preferred cache configuration for the current context. This is only a preference. The driver will use the requested configuration if possible, but it is free to choose a different configuration if required to execute the function. Any function preference set via `cuFuncSetCacheConfig()` will be preferred over this context-wide setting. Setting the context-wide cache configuration to `CU_FUNC_CACHE_PREFER_NONE` will cause subsequent kernel launches to prefer to not change the cache configuration unless required to launch the kernel.

This setting does nothing on devices where the size of the L1 cache and shared memory are fixed.

Launching a kernel with a different preference than the most recent preference setting may insert a device-side synchronization point.

The supported cache configurations are:

- ▶ `CU_FUNC_CACHE_PREFER_NONE`: no preference for shared memory or L1 (default)
- ▶ `CU_FUNC_CACHE_PREFER_SHARED`: prefer larger shared memory and smaller L1 cache
- ▶ `CU_FUNC_CACHE_PREFER_L1`: prefer larger L1 cache and smaller shared memory
- ▶ `CU_FUNC_CACHE_PREFER_EQUAL`: prefer equal sized L1 cache and shared memory



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`, `cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPopCurrent`, `cuCtxPushCurrent`, `cuCtxSetLimit`, `cuCtxSynchronize`, `cuFuncSetCacheConfig`, `cudaDeviceSetCacheConfig`

CUresult cuCtxSetCurrent (CUcontext ctx)

Binds the specified CUDA context to the calling CPU thread.

Parameters

`ctx`

- Context to bind to the calling CPU thread

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`

Description

Binds the specified CUDA context to the calling CPU thread. If `ctx` is `NULL` then the CUDA context previously bound to the calling CPU thread is unbound and `CUDA_SUCCESS` is returned.

If there exists a CUDA context stack on the calling CPU thread, this will replace the top of that stack with `ctx`. If `ctx` is `NULL` then this will be equivalent to popping the top

of the calling CPU thread's CUDA context stack (or a no-op if the calling CPU thread's CUDA context stack is empty).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxGetCurrent](#), [cuCtxCreate](#), [cuCtxDestroy](#), [cudaSetDevice](#)

CUresult cuCtxSetLimit (CUlimit limit, size_t value)

Set resource limits.

Parameters

limit

- Limit to set

value

- Size of limit

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_VALUE](#),
[CUDA_ERROR_UNSUPPORTED_LIMIT](#), [CUDA_ERROR_OUT_OF_MEMORY](#)

Description

Setting `limit` to `value` is a request by the application to update the current limit maintained by the context. The driver is free to modify the requested value to meet h/w requirements (this could be clamping to minimum or maximum values, rounding up to nearest element size, etc). The application can use [cuCtxGetLimit\(\)](#) to find out exactly what the limit has been set to.

Setting each [CUlimit](#) has its own specific restrictions, so each is discussed here.

- ▶ [CU_LIMIT_STACK_SIZE](#) controls the stack size in bytes of each GPU thread.
- ▶ [CU_LIMIT_PRINTF_FIFO_SIZE](#) controls the size in bytes of the FIFO used by the `printf()` device system call. Setting [CU_LIMIT_PRINTF_FIFO_SIZE](#) must be performed before launching any kernel that uses the `printf()` device system call, otherwise [CUDA_ERROR_INVALID_VALUE](#) will be returned.
- ▶ [CU_LIMIT_MALLOC_HEAP_SIZE](#) controls the size in bytes of the heap used by the `malloc()` and `free()` device system calls. Setting [CU_LIMIT_MALLOC_HEAP_SIZE](#) must be performed before launching any kernel that uses the `malloc()` or `free()` device system calls, otherwise [CUDA_ERROR_INVALID_VALUE](#) will be returned.

- ▶ `CU_LIMIT_DEV_RUNTIME_SYNC_DEPTH` controls the maximum nesting depth of a grid at which a thread can safely call `cudaDeviceSynchronize()`. Setting this limit must be performed before any launch of a kernel that uses the device runtime and calls `cudaDeviceSynchronize()` above the default sync depth, two levels of grids. Calls to `cudaDeviceSynchronize()` will fail with error code `cudaErrorSyncDepthExceeded` if the limitation is violated. This limit can be set smaller than the default or up the maximum launch depth of 24. When setting this limit, keep in mind that additional levels of sync depth require the driver to reserve large amounts of device memory which can no longer be used for user allocations. If these reservations of device memory fail, `cuCtxSetLimit` will return `CUDA_ERROR_OUT_OF_MEMORY`, and the limit can be reset to a lower value. This limit is only applicable to devices of compute capability 3.5 and higher. Attempting to set this limit on devices of compute capability less than 3.5 will result in the error `CUDA_ERROR_UNSUPPORTED_LIMIT` being returned.
- ▶ `CU_LIMIT_DEV_RUNTIME_PENDING_LAUNCH_COUNT` controls the maximum number of outstanding device runtime launches that can be made from the current context. A grid is outstanding from the point of launch up until the grid is known to have been completed. Device runtime launches which violate this limitation fail and return `cudaErrorLaunchPendingCountExceeded` when `cudaGetLastError()` is called after launch. If more pending launches than the default (2048 launches) are needed for a module using the device runtime, this limit can be increased. Keep in mind that being able to sustain additional pending launches will require the driver to reserve larger amounts of device memory upfront which can no longer be used for allocations. If these reservations fail, `cuCtxSetLimit` will return `CUDA_ERROR_OUT_OF_MEMORY`, and the limit can be reset to a lower value. This limit is only applicable to devices of compute capability 3.5 and higher. Attempting to set this limit on devices of compute capability less than 3.5 will result in the error `CUDA_ERROR_UNSUPPORTED_LIMIT` being returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`, `cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPopCurrent`, `cuCtxPushCurrent`, `cuCtxSetCacheConfig`, `cuCtxSynchronize`, `cudaDeviceSetLimit`

CUresult cuCtxSetSharedMemConfig (CUsharedconfig config)

Sets the shared memory configuration for the current context.

Parameters

config

- requested shared memory configuration

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

On devices with configurable shared memory banks, this function will set the context's shared memory bank size which is used for subsequent kernel launches.

Changing the shared memory configuration between launches may insert a device side synchronization point between those launches.

Changing the shared memory bank size will not increase shared memory usage or affect occupancy of kernels, but may have major effects on performance. Larger bank sizes will allow for greater potential bandwidth to shared memory, but will change what kinds of accesses to shared memory will result in bank conflicts.

This function will do nothing on devices with fixed shared memory bank size.

The supported bank configurations are:

- ▶ **CU_SHARED_MEM_CONFIG_DEFAULT_BANK_SIZE**: set bank width to the default initial setting (currently, four bytes).
- ▶ **CU_SHARED_MEM_CONFIG_FOUR_BYTE_BANK_SIZE**: set shared memory bank width to be natively four bytes.
- ▶ **CU_SHARED_MEM_CONFIG_EIGHT_BYTE_BANK_SIZE**: set shared memory bank width to be natively eight bytes.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuCtxDestroy](#), [cuCtxGetApiVersion](#), [cuCtxGetCacheConfig](#),
[cuCtxGetDevice](#), [cuCtxGetFlags](#), [cuCtxGetLimit](#), [cuCtxPopCurrent](#), [cuCtxPushCurrent](#),
[cuCtxSetLimit](#), [cuCtxSynchronize](#), [cuCtxGetSharedMemConfig](#), [cuFuncSetCacheConfig](#),
[cudaDeviceSetSharedMemConfig](#)

CUresult cuCtxSynchronize (void)

Block for a context's tasks to complete.

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT

Description

Blocks until the device has completed all preceding requested tasks. [cuCtxSynchronize\(\)](#) returns an error if one of the preceding tasks failed. If the context was created with the [CU_CTX_SCHED_BLOCKING_SYNC](#) flag, the CPU thread will block until the GPU context has finished its work.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuCtxDestroy](#), [cuCtxGetApiVersion](#), [cuCtxGetCacheConfig](#),
[cuCtxGetDevice](#), [cuCtxGetFlags](#), [cuCtxGetLimit](#), [cuCtxPopCurrent](#), [cuCtxPushCurrent](#),
[cuCtxSetCacheConfig](#), [cuCtxSetLimit](#), [cudaDeviceSynchronize](#)

4.9. Context Management [DEPRECATED]

This section describes the deprecated context management functions of the low-level CUDA driver application programming interface.

CUresult cuCtxAttach (CUcontext *pctx, unsigned int flags)

Increment a context's usage-count.

Parameters

pctx

- Returned context handle of the current context

flags

- Context attach flags (must be 0)

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Deprecated

Note that this function is deprecated and should not be used.

Increments the usage count of the context and passes back a context handle in `*pctx` that must be passed to `cuCtxDetach()` when the application is done with the context. `cuCtxAttach()` fails if there is no context current to the thread.

Currently, the `flags` parameter must be 0.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxDetach`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`,
`cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPopCurrent`, `cuCtxPushCurrent`,
`cuCtxSetCacheConfig`, `cuCtxSetLimit`, `cuCtxSynchronize`

CUresult cuCtxDetach (CUcontext ctx)

Decrement a context's usage-count.

Parameters

`ctx`

- Context to destroy

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT

Description

Deprecated

Note that this function is deprecated and should not be used.

Decrements the usage count of the context `ctx`, and destroys the context if the usage count goes to 0. The context must be a handle that was passed back by `cuCtxCreate()` or `cuCtxAttach()`, and must be current to the calling thread.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuCtxDestroy`, `cuCtxGetApiVersion`, `cuCtxGetCacheConfig`, `cuCtxGetDevice`, `cuCtxGetFlags`, `cuCtxGetLimit`, `cuCtxPopCurrent`, `cuCtxPushCurrent`, `cuCtxSetCacheConfig`, `cuCtxSetLimit`, `cuCtxSynchronize`

4.10. Module Management

This section describes the module management functions of the low-level CUDA driver application programming interface.

CUresult cuLinkAddData (CUlinkState state, CUjitInputType type, void *data, size_t size, const char *name, unsigned int numOptions, CUjit_option *options, void **optionValues)

Add an input to a pending linker invocation.

Parameters

state

A pending linker action.

type

The type of the input data.

data

The input data. PTX must be NULL-terminated.

size

The length of the input data.

name

An optional name for this input in log messages.

numOptions

Size of options.

options

Options to be applied only for this input (overrides options from [cuLinkCreate](#)).

optionValues

Array of option values, each cast to void *.

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_IMAGE,
 CUDA_ERROR_INVALID_PTX, CUDA_ERROR_OUT_OF_MEMORY,
 CUDA_ERROR_NO_BINARY_FOR_GPU

Description

Ownership of data is retained by the caller. No reference is retained to any inputs after this call returns.

This method accepts only compiler options, which are used if the data must be compiled from PTX, and does not accept any of CU_JIT_WALL_TIME, CU_JIT_INFO_LOG_BUFFER, CU_JIT_ERROR_LOG_BUFFER, CU_JIT_TARGET_FROM_CUCONTEXT, or CU_JIT_TARGET.

See also:

[cuLinkCreate](#), [cuLinkAddFile](#), [cuLinkComplete](#), [cuLinkDestroy](#)

CUresult cuLinkAddFile (CUlinkState state, CUjitInputType type, const char *path, unsigned int numOptions, CUjit_option *options, void **optionValues)

Add a file input to a pending linker invocation.

Parameters**state**

A pending linker action

type

The type of the input data

path

Path to the input file

numOptions

Size of options

options

Options to be applied only for this input (overrides options from [cuLinkCreate](#))

optionValues

Array of option values, each cast to void *

Returns

CUDA_SUCCESS, CUDA_ERROR_FILE_NOT_FOUND
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_INVALID_IMAGE, CUDA_ERROR_INVALID_PTX,
 CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_NO_BINARY_FOR_GPU

Description

No reference is retained to any inputs after this call returns.

This method accepts only compiler options, which are used if the input must be compiled from PTX, and does not accept any of `CU_JIT_WALL_TIME`, `CU_JIT_INFO_LOG_BUFFER`, `CU_JIT_ERROR_LOG_BUFFER`, `CU_JIT_TARGET_FROM_CUCONTEXT`, or `CU_JIT_TARGET`.

This method is equivalent to invoking `cuLinkAddData` on the contents of the file.

See also:

`cuLinkCreate`, `cuLinkAddData`, `cuLinkComplete`, `cuLinkDestroy`

CUresult cuLinkComplete (CUlinkState state, void **cubinOut, size_t *sizeOut)

Complete a pending linker invocation.

Parameters**state**

A pending linker invocation

cubinOut

On success, this will point to the output image

sizeOut

Optional parameter to receive the size of the generated image

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
 CUDA_ERROR_OUT_OF_MEMORY

Description

Completes the pending linker action and returns the cubin image for the linked device code, which can be used with `cuModuleLoadData`. The cubin is owned by `state`, so it should be loaded before `state` is destroyed via `cuLinkDestroy`. This call does not destroy `state`.

See also:

`cuLinkCreate`, `cuLinkAddData`, `cuLinkAddFile`, `cuLinkDestroy`, `cuModuleLoadData`

CUresult cuLinkCreate (unsigned int numOptions, CUjit_option *options, void **optionValues, CUlinkState *stateOut)

Creates a pending JIT linker invocation.

Parameters**numOptions**

Size of options arrays

options

Array of linker and compiler options

optionValues

Array of option values, each cast to void *

stateOut

On success, this will contain a CUlinkState to specify and complete this action

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_OUT_OF_MEMORY`,
`CUDA_ERROR_JIT_COMPILER_NOT_FOUND`

Description

If the call is successful, the caller owns the returned CUlinkState, which should eventually be destroyed with `cuLinkDestroy`. The device code machine size (32 or 64 bit) will match the calling application.

Both linker and compiler options may be specified. Compiler options will be applied to inputs to this linker action which must be compiled from PTX. The options `CU_JIT_WALL_TIME`, `CU_JIT_INFO_LOG_BUFFER_SIZE_BYTES`, and `CU_JIT_ERROR_LOG_BUFFER_SIZE_BYTES` will accumulate data until the CUlinkState is destroyed.

`optionValues` must remain valid for the life of the CUlinkState if output options are used. No other references to inputs are maintained after this call returns.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuLinkAddData](#), [cuLinkAddFile](#), [cuLinkComplete](#), [cuLinkDestroy](#)

CUresult cuLinkDestroy (CUlinkState state)

Destroys state for a JIT linker invocation.

Parameters**state**

State object for the linker invocation

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_HANDLE](#)

Description**See also:**

[cuLinkCreate](#)

CUresult cuModuleGetFunction (CUfunction *hfunc, CUmodule hmod, const char *name)

Returns a function handle.

Parameters**hfunc**

- Returned function handle

hmod

- Module to retrieve function from

name

- Name of function to retrieve

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#), [CUDA_ERROR_NOT_FOUND](#)

Description

Returns in **hfunc* the handle of the function of name *name* located in module *hmod*. If no function of that name exists, [cuModuleGetFunction\(\)](#) returns [CUDA_ERROR_NOT_FOUND](#).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuModuleGetGlobal](#), [cuModuleGetTexRef](#), [cuModuleLoad](#), [cuModuleLoadData](#), [cuModuleLoadDataEx](#), [cuModuleLoadFatBinary](#), [cuModuleUnload](#)

CUresult cuModuleGetGlobal (CUdeviceptr *dptr, size_t *bytes, CUmodule hmod, const char *name)

Returns a global pointer from a module.

Parameters

dptr

- Returned global device pointer

bytes

- Returned global size in bytes

hmod

- Module to retrieve global from

name

- Name of global to retrieve

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#), [CUDA_ERROR_NOT_FOUND](#)

Description

Returns in `*dptr` and `*bytes` the base pointer and size of the global of name `name` located in module `hmod`. If no variable of that name exists, `cuModuleGetGlobal()` returns [CUDA_ERROR_NOT_FOUND](#). Both parameters `dptr` and `bytes` are optional. If one of them is `NULL`, it is ignored.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuModuleGetFunction`, `cuModuleGetTexRef`, `cuModuleLoad`, `cuModuleLoadData`,
`cuModuleLoadDataEx`, `cuModuleLoadFatBinary`, `cuModuleUnload`,
`cudaGetSymbolAddress`, `cudaGetSymbolSize`

CUresult cuModuleGetSurfRef (CUsurfref *pSurfRef, CUmodule hmod, const char *name)

Returns a handle to a surface reference.

Parameters

pSurfRef

- Returned surface reference

hmod

- Module to retrieve surface reference from

name

- Name of surface reference to retrieve

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_NOT_FOUND`

Description

Returns in `*pSurfRef` the handle of the surface reference of name `name` in the module `hmod`. If no surface reference of that name exists, `cuModuleGetSurfRef()` returns `CUDA_ERROR_NOT_FOUND`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuModuleGetFunction`, `cuModuleGetGlobal`, `cuModuleGetTexRef`, `cuModuleLoad`,
`cuModuleLoadData`, `cuModuleLoadDataEx`, `cuModuleLoadFatBinary`,
`cuModuleUnload`, `cudaGetSurfaceReference`

CUresult cuModuleGetTexRef (CUtexref *pTexRef, CUmodule hmod, const char *name)

Returns a handle to a texture reference.

Parameters

pTexRef

- Returned texture reference

hmod

- Module to retrieve texture reference from

name

- Name of texture reference to retrieve

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_NOT_FOUND

Description

Returns in *pTexRef the handle of the texture reference of name name in the module hmod. If no texture reference of that name exists, cuModuleGetTexRef() returns CUDA_ERROR_NOT_FOUND. This texture reference handle should not be destroyed, since it will be destroyed when the module is unloaded.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuModuleGetFunction, cuModuleGetGlobal, cuModuleGetSurfRef, cuModuleLoad, cuModuleLoadData, cuModuleLoadDataEx, cuModuleLoadFatBinary, cuModuleUnload, cudaGetTextureReference

CUresult cuModuleLoad (CUmodule *module, const char *fname)

Loads a compute module.

Parameters

module

- Returned module

fname

- Filename of module to load

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_PTX,
 CUDA_ERROR_NOT_FOUND, CUDA_ERROR_OUT_OF_MEMORY,
 CUDA_ERROR_FILE_NOT_FOUND, CUDA_ERROR_NO_BINARY_FOR_GPU,
 CUDA_ERROR_SHARED_OBJECT_SYMBOL_NOT_FOUND,
 CUDA_ERROR_SHARED_OBJECT_INIT_FAILED,
 CUDA_ERROR_JIT_COMPILER_NOT_FOUND

Description

Takes a filename `fname` and loads the corresponding module `module` into the current context. The CUDA driver API does not attempt to lazily allocate the resources needed by a module; if the memory for functions and data (constant and global) needed by the module cannot be allocated, `cuModuleLoad()` fails. The file should be a cubin file as output by `nvcc`, or a PTX file either as output by `nvcc` or handwritten, or a fatbin file as output by `nvcc` from toolchain 4.0 or later.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuModuleGetFunction`, `cuModuleGetGlobal`, `cuModuleGetTexRef`, `cuModuleLoadData`, `cuModuleLoadDataEx`, `cuModuleLoadFatBinary`, `cuModuleUnload`

CUresult cuModuleLoadData (CUmodule *module, const void *image)

Load a module's data.

Parameters

module

- Returned module

image

- Module data to load

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_PTX,
CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_NO_BINARY_FOR_GPU,
CUDA_ERROR_SHARED_OBJECT_SYMBOL_NOT_FOUND,
CUDA_ERROR_SHARED_OBJECT_INIT_FAILED,
CUDA_ERROR_JIT_COMPILER_NOT_FOUND

Description

Takes a pointer `image` and loads the corresponding module `module` into the current context. The pointer may be obtained by mapping a cubin or PTX or fatbin file, passing a cubin or PTX or fatbin file as a NULL-terminated text string, or incorporating a cubin or fatbin object into the executable resources and using operating system calls such as Windows `FindResource()` to obtain the pointer.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuModuleGetFunction](#), [cuModuleGetGlobal](#), [cuModuleGetTexRef](#), [cuModuleLoad](#), [cuModuleLoadDataEx](#), [cuModuleLoadFatBinary](#), [cuModuleUnload](#)

CUresult cuModuleLoadDataEx (CUmodule *module, const void *image, unsigned int numOptions, CUjit_option *options, void **optionValues)

Load a module's data with options.

Parameters

module

- Returned module

image

- Module data to load

numOptions

- Number of options

options

- Options for JIT

optionValues

- Option values for JIT

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_PTX,
 CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_NO_BINARY_FOR_GPU,
 CUDA_ERROR_SHARED_OBJECT_SYMBOL_NOT_FOUND,
 CUDA_ERROR_SHARED_OBJECT_INIT_FAILED,
 CUDA_ERROR_JIT_COMPILER_NOT_FOUND

Description

Takes a pointer `image` and loads the corresponding module `module` into the current context. The pointer may be obtained by mapping a cubin or PTX or fatbin file, passing a cubin or PTX or fatbin file as a NULL-terminated text string, or incorporating a cubin or fatbin object into the executable resources and using operating system calls such as Windows `FindResource()` to obtain the pointer. Options are passed as an array via `options` and any corresponding parameters are passed in `optionValues`. The number of total options is supplied via `numOptions`. Any outputs will be returned via `optionValues`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuModuleGetFunction`, `cuModuleGetGlobal`, `cuModuleGetTexRef`, `cuModuleLoad`,
`cuModuleLoadData`, `cuModuleLoadFatBinary`, `cuModuleUnload`

CUresult cuModuleLoadFatBinary (CUmodule *module, const void *fatCubin)

Load a module's data.

Parameters**module**

- Returned module

fatCubin

- Fat binary to load

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_PTX`,
`CUDA_ERROR_NOT_FOUND`, `CUDA_ERROR_OUT_OF_MEMORY`,
`CUDA_ERROR_NO_BINARY_FOR_GPU`,
`CUDA_ERROR_SHARED_OBJECT_SYMBOL_NOT_FOUND`,
`CUDA_ERROR_SHARED_OBJECT_INIT_FAILED`,
`CUDA_ERROR_JIT_COMPILER_NOT_FOUND`

Description

Takes a pointer `fatCubin` and loads the corresponding module `module` into the current context. The pointer represents a fat binary object, which is a collection of different cubin and/or PTX files, all representing the same device code, but compiled and optimized for different architectures.

Prior to CUDA 4.0, there was no documented API for constructing and using fat binary objects by programmers. Starting with CUDA 4.0, fat binary objects can be constructed by providing the `-fatbin` option to `nvcc`. More information can be found in the `nvcc` document.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuModuleGetFunction](#), [cuModuleGetGlobal](#), [cuModuleGetTexRef](#), [cuModuleLoad](#), [cuModuleLoadData](#), [cuModuleLoadDataEx](#), [cuModuleUnload](#)

CUresult cuModuleUnload (CUmodule hmod)

Unloads a module.

Parameters

hmod

- Module to unload

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Unloads a module `hmod` from the current context.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuModuleGetFunction](#), [cuModuleGetGlobal](#), [cuModuleGetTexRef](#), [cuModuleLoad](#), [cuModuleLoadData](#), [cuModuleLoadDataEx](#), [cuModuleLoadFatBinary](#)

4.11. Memory Management

This section describes the memory management functions of the low-level CUDA driver application programming interface.

CUresult cuArray3DCreate (CUarray *pHandle, const CUDA_ARRAY3D_DESCRIPTOR *pAllocateArray)

Creates a 3D CUDA array.

Parameters

pHandle

- Returned array

pAllocateArray

- 3D array descriptor

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY,
 CUDA_ERROR_UNKNOWN

Description

Creates a CUDA array according to the `CUDA_ARRAY3D_DESCRIPTOR` structure `pAllocateArray` and returns a handle to the new CUDA array in `*pHandle`. The `CUDA_ARRAY3D_DESCRIPTOR` is defined as:

```
↑ typedef struct {
    unsigned int Width;
    unsigned int Height;
    unsigned int Depth;
    CUarray_format Format;
    unsigned int NumChannels;
    unsigned int Flags;
} CUDA_ARRAY3D_DESCRIPTOR;
```

where:

- ▶ Width, Height, and Depth are the width, height, and depth of the CUDA array (in elements); the following types of CUDA arrays can be allocated:
 - ▶ A 1D array is allocated if Height and Depth extents are both zero.
 - ▶ A 2D array is allocated if only Depth extent is zero.
 - ▶ A 3D array is allocated if all three extents are non-zero.
 - ▶ A 1D layered CUDA array is allocated if only Height is zero and the `CUDA_ARRAY3D_LAYERED` flag is set. Each layer is a 1D array. The number of layers is determined by the depth extent.
 - ▶ A 2D layered CUDA array is allocated if all three extents are non-zero and the `CUDA_ARRAY3D_LAYERED` flag is set. Each layer is a 2D array. The number of layers is determined by the depth extent.
 - ▶ A cubemap CUDA array is allocated if all three extents are non-zero and the `CUDA_ARRAY3D_CUBEMAP` flag is set. Width must be equal to Height, and Depth must be six. A cubemap is a special type of 2D layered CUDA array, where the six layers represent the six faces of a cube. The order of the six layers in memory is the same as that listed in `CUarray_cubemap_face`.
 - ▶ A cubemap layered CUDA array is allocated if all three extents are non-zero, and both, `CUDA_ARRAY3D_CUBEMAP` and `CUDA_ARRAY3D_LAYERED` flags are set. Width must be equal to Height, and Depth must be a multiple of six. A cubemap layered CUDA array is a special type of 2D layered CUDA array

that consists of a collection of cubemaps. The first six layers represent the first cubemap, the next six layers form the second cubemap, and so on.

- ▶ Format specifies the format of the elements; `CUarray_format` is defined as:

```
typedef enum CUarray_format_enum {
    CU_AD_FORMAT_UNSIGNED_INT8 = 0x01,
    CU_AD_FORMAT_UNSIGNED_INT16 = 0x02,
    CU_AD_FORMAT_UNSIGNED_INT32 = 0x03,
    CU_AD_FORMAT_SIGNED_INT8 = 0x08,
    CU_AD_FORMAT_SIGNED_INT16 = 0x09,
    CU_AD_FORMAT_SIGNED_INT32 = 0x0a,
    CU_AD_FORMAT_HALF = 0x10,
    CU_AD_FORMAT_FLOAT = 0x20
} CUarray_format;
```

- ▶ NumChannels specifies the number of packed components per CUDA array element; it may be 1, 2, or 4;
- ▶ Flags may be set to
 - ▶ `CUDA_ARRAY3D_LAYERED` to enable creation of layered CUDA arrays. If this flag is set, Depth specifies the number of layers, not the depth of a 3D array.
 - ▶ `CUDA_ARRAY3D_SURFACE_LDST` to enable surface references to be bound to the CUDA array. If this flag is not set, `cuSurfRefSetArray` will fail when attempting to bind the CUDA array to a surface reference.
 - ▶ `CUDA_ARRAY3D_CUBEMAP` to enable creation of cubemaps. If this flag is set, Width must be equal to Height, and Depth must be six. If the `CUDA_ARRAY3D_LAYERED` flag is also set, then Depth must be a multiple of six.
 - ▶ `CUDA_ARRAY3D_TEXTURE_GATHER` to indicate that the CUDA array will be used for texture gather. Texture gather can only be performed on 2D CUDA arrays.

Width, Height and Depth must meet certain size requirements as listed in the following table. All values are specified in elements. Note that for brevity's sake, the full name of the device attribute is not specified. For ex., `TEXTURE1D_WIDTH` refers to the device attribute `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_WIDTH`.

Note that 2D CUDA arrays have different size requirements if the `CUDA_ARRAY3D_TEXTURE_GATHER` flag is set. Width and Height must not be greater than `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_GATHER_WIDTH` and `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_GATHER_HEIGHT` respectively, in that case.

CUDA array type	Valid extents that must always be met {(width range in elements), (height range), (depth range)}	Valid extents with <code>CUDA_ARRAY3D_SURFACE_LDST</code> set {(width range in elements), (height range), (depth range)}
1D	{ (1, <code>TEXTURE1D_WIDTH</code>), 0, 0 }	{ (1, <code>SURFACE1D_WIDTH</code>), 0, 0 }

2D	{ (1,TEXTURE2D_WIDTH), (1,TEXTURE2D_HEIGHT), 0 }	{ (1,SURFACE2D_WIDTH), (1,SURFACE2D_HEIGHT), 0 }
3D	{ (1,TEXTURE3D_WIDTH), (1,TEXTURE3D_HEIGHT), (1,TEXTURE3D_DEPTH) } OR { (1,TEXTURE3D_WIDTH_ALTERNATE (1,TEXTURE3D_HEIGHT_ALTERNATE (1,TEXTURE3D_DEPTH_ALTERNATE)	{ (1,SURFACE3D_WIDTH), (1,SURFACE3D_HEIGHT), (1,SURFACE3D_DEPTH) }
1D Layered	{ (1,TEXTURE1D_LAYERED_WIDTH), 0, (1,TEXTURE1D_LAYERED_LAYERS) }	{ (1,SURFACE1D_LAYERED_WIDTH), 0, (1,SURFACE1D_LAYERED_LAYERS) }
2D Layered	{ (1,TEXTURE2D_LAYERED_WIDTH), (1,TEXTURE2D_LAYERED_HEIGHT), (1,TEXTURE2D_LAYERED_LAYERS) }	{ (1,SURFACE2D_LAYERED_WIDTH), (1,SURFACE2D_LAYERED_HEIGHT), (1,SURFACE2D_LAYERED_LAYERS) }
Cubemap	{ (1,TEXTURECUBEMAP_WIDTH), (1,TEXTURECUBEMAP_WIDTH), 6 }	{ (1,SURFACECUBEMAP_WIDTH), (1,SURFACECUBEMAP_WIDTH), 6 }
Cubemap Layered	{ (1,TEXTURECUBEMAP_LAYERED_WIDTH), (1,TEXTURECUBEMAP_LAYERED_WIDTH), (1,TEXTURECUBEMAP_LAYERED_LAYERS) }	{ (1,SURFACECUBEMAP_LAYERED_WIDTH), (1,SURFACECUBEMAP_LAYERED_WIDTH), (1,SURFACECUBEMAP_LAYERED_LAYERS) }

Here are examples of CUDA array descriptions:

Description for a CUDA array of 2048 floats:

```
↑ CUDA_ARRAY3D_DESCRIPTOR desc;
   desc.Format = CU_AD_FORMAT_FLOAT;
   desc.NumChannels = 1;
   desc.Width = 2048;
   desc.Height = 0;
   desc.Depth = 0;
```

Description for a 64 x 64 CUDA array of floats:

```
↑ CUDA_ARRAY3D_DESCRIPTOR desc;
   desc.Format = CU_AD_FORMAT_FLOAT;
   desc.NumChannels = 1;
   desc.Width = 64;
   desc.Height = 64;
   desc.Depth = 0;
```

Description for a width x height x depth CUDA array of 64-bit, 4x16-bit float16's:

```
↑ CUDA_ARRAY3D_DESCRIPTOR desc;
   desc.FormatFlags = CU_AD_FORMAT_HALF;
   desc.NumChannels = 4;
   desc.Width = width;
   desc.Height = height;
   desc.Depth = depth;
```



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaMalloc3DArray`

CUresult cuArray3DGetDescriptor (CUDA_ARRAY3D_DESCRIPTOR *pArrayDescriptor, CUarray hArray)

Get a 3D CUDA array descriptor.

Parameters

`pArrayDescriptor`

- Returned 3D array descriptor

`hArray`

- 3D array to get descriptor of

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`

Description

Returns in `*pArrayDescriptor` a descriptor containing information on the format and dimensions of the CUDA array `hArray`. It is useful for subroutines that have been passed a CUDA array, but need to know the CUDA array parameters for validation or other purposes.

This function may be called on 1D and 2D arrays, in which case the `Height` and/or `Depth` members of the descriptor struct will be set to 0.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuArray3DCreate](#), [cuArrayCreate](#), [cuArrayDestroy](#), [cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#), [cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#), [cuMemcpy3D](#), [cuMemcpy3DAsync](#), [cuMemcpyAtoA](#), [cuMemcpyAtoD](#), [cuMemcpyAtoH](#), [cuMemcpyAtoHAsync](#), [cuMemcpyDtoA](#), [cuMemcpyDtoD](#), [cuMemcpyDtoDAsync](#), [cuMemcpyDtoH](#), [cuMemcpyDtoHAsync](#), [cuMemcpyHtoA](#), [cuMemcpyHtoAAsync](#), [cuMemcpyHtoD](#), [cuMemcpyHtoDAsync](#), [cuMemFree](#), [cuMemFreeHost](#), [cuMemGetAddressRange](#), [cuMemGetInfo](#), [cuMemHostAlloc](#), [cuMemHostGetDevicePointer](#), [cuMemsetD2D8](#), [cuMemsetD2D16](#), [cuMemsetD2D32](#), [cuMemsetD8](#), [cuMemsetD16](#), [cuMemsetD32](#), [cudaArrayGetInfo](#)

CUresult cuArrayCreate (CUarray *pHandle, const CUDA_ARRAY_DESCRIPTOR *pAllocateArray)

Creates a 1D or 2D CUDA array.

Parameters

pHandle

- Returned array

pAllocateArray

- Array descriptor

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY,
CUDA_ERROR_UNKNOWN

Description

Creates a CUDA array according to the [CUDA_ARRAY_DESCRIPTOR](#) structure `pAllocateArray` and returns a handle to the new CUDA array in `*pHandle`. The [CUDA_ARRAY_DESCRIPTOR](#) is defined as:

```
↑ typedef struct {
    unsigned int Width;
    unsigned int Height;
    CUarray_format Format;
    unsigned int NumChannels;
} CUDA_ARRAY_DESCRIPTOR;
```

where:

- ▶ Width, and Height are the width, and height of the CUDA array (in elements); the CUDA array is one-dimensional if height is 0, two-dimensional otherwise;
- ▶ Format specifies the format of the elements; `CUarray_format` is defined as:

```
↑ typedef enum CUarray_format_enum {
    CU_AD_FORMAT_UNSIGNED_INT8 = 0x01,
    CU_AD_FORMAT_UNSIGNED_INT16 = 0x02,
    CU_AD_FORMAT_UNSIGNED_INT32 = 0x03,
    CU_AD_FORMAT_SIGNED_INT8 = 0x08,
    CU_AD_FORMAT_SIGNED_INT16 = 0x09,
    CU_AD_FORMAT_SIGNED_INT32 = 0x0a,
    CU_AD_FORMAT_HALF = 0x10,
    CU_AD_FORMAT_FLOAT = 0x20
} CUarray_format;
```

- ▶ NumChannels specifies the number of packed components per CUDA array element; it may be 1, 2, or 4;

Here are examples of CUDA array descriptions:

Description for a CUDA array of 2048 floats:

```
↑ CUDA_ARRAY_DESCRIPTOR desc;
    desc.Format = CU_AD_FORMAT_FLOAT;
    desc.NumChannels = 1;
    desc.Width = 2048;
    desc.Height = 1;
```

Description for a 64 x 64 CUDA array of floats:

```
↑ CUDA_ARRAY_DESCRIPTOR desc;
    desc.Format = CU_AD_FORMAT_FLOAT;
    desc.NumChannels = 1;
    desc.Width = 64;
    desc.Height = 64;
```

Description for a width x height CUDA array of 64-bit, 4x16-bit float16's:

```
↑ CUDA_ARRAY_DESCRIPTOR desc;
    desc.FormatFlags = CU_AD_FORMAT_HALF;
    desc.NumChannels = 4;
    desc.Width = width;
    desc.Height = height;
```

Description for a width x height CUDA array of 16-bit elements, each of which is two 8-bit unsigned chars:

```
↑ CUDA_ARRAY_DESCRIPTOR arrayDesc;
    desc.FormatFlags = CU_AD_FORMAT_UNSIGNED_INT8;
    desc.NumChannels = 2;
    desc.Width = width;
    desc.Height = height;
```



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaMallocArray

CUresult cuArrayDestroy (CUarray hArray)

Destroys a CUDA array.

Parameters

hArray

- Array to destroy

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_ARRAY_IS_MAPPED

Description

Destroys the CUDA array hArray.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaFreeArray

CUresult cuArrayGetDescriptor (CUDA_ARRAY_DESCRIPTOR *pArrayDescriptor, CUarray hArray)

Get a 1D or 2D CUDA array descriptor.

Parameters

pArrayDescriptor

- Returned array descriptor

hArray

- Array to get descriptor of

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE

Description

Returns in *pArrayDescriptor a descriptor containing information on the format and dimensions of the CUDA array hArray. It is useful for subroutines that have been passed a CUDA array, but need to know the CUDA array parameters for validation or other purposes.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy,
cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync,
cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA,
cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA,
cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync,
cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync,
cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo,
cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16,
cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaArrayGetInfo

CUresult cuDeviceGetByPCIBusId (CUdevice *dev, const char *pciBusId)

Returns a handle to a compute device.

Parameters

dev

- Returned device handle

pciBusId

- String in one of the following forms: [domain]:[bus]:[device].[function] [domain]:[bus]:[device] [bus]:[device].[function] where `domain`, `bus`, `device`, and `function` are all hexadecimal values

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
CUDA_ERROR_INVALID_DEVICE

Description

Returns in `*device` a device handle given a PCI bus ID string.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuDeviceGet](#), [cuDeviceGetAttribute](#), [cuDeviceGetPCIBusId](#), [cudaDeviceGetByPCIBusId](#)

CUresult cuDeviceGetPCIBusId (char *pciBusId, int len, CUdevice dev)

Returns a PCI Bus Id string for the device.

Parameters

pciBusId

- Returned identifier string for the device in the following format [domain]:[bus]:[device].[function] where `domain`, `bus`, `device`, and `function` are all hexadecimal values. `pciBusId` should be large enough to store 13 characters including the NULL-terminator.

len

- Maximum length of string to store in name

dev

- Device to get identifier string for

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_INVALID_DEVICE

Description

Returns an ASCII string identifying the device `dev` in the NULL-terminated string pointed to by `pciBusId`. `len` specifies the maximum length of the string that may be returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDeviceGet`, `cuDeviceGetAttribute`, `cuDeviceGetByPCIBusId`, `cudaDeviceGetPCIBusId`

CUresult `culpcCloseMemHandle` (CUdeviceptr `dptr`)

Close memory mapped with `cuIpcOpenMemHandle`.

Parameters**dptr**

- Device pointer returned by `cuIpcOpenMemHandle`

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_MAP_FAILED, CUDA_ERROR_INVALID_HANDLE,

Description

Unmaps memory returned by `cuIpcOpenMemHandle`. The original allocation in the exporting process as well as imported mappings in other processes will be unaffected.

Any resources used to enable peer access will be freed if this is the last mapping using them.

IPC functionality is restricted to devices with support for unified addressing on Linux and Windows operating systems. IPC functionality on Windows is restricted to GPUs in TCC mode

See also:

[cuMemAlloc](#), [cuMemFree](#), [cuIpcGetEventHandle](#), [cuIpcOpenEventHandle](#), [cuIpcGetMemHandle](#), [cuIpcOpenMemHandle](#), [cudaIpcCloseMemHandle](#)

CUresult cuIpcGetEventHandle (CUipcEventHandle *pHandle, CUevent event)

Gets an interprocess handle for a previously allocated event.

Parameters

pHandle

- Pointer to a user allocated [CUipcEventHandle](#) in which to return the opaque event handle

event

- Event allocated with [CU_EVENT_INTERPROCESS](#) and [CU_EVENT_DISABLE_TIMING](#) flags.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_HANDLE](#), [CUDA_ERROR_OUT_OF_MEMORY](#), [CUDA_ERROR_MAP_FAILED](#)

Description

Takes as input a previously allocated event. This event must have been created with the [CU_EVENT_INTERPROCESS](#) and [CU_EVENT_DISABLE_TIMING](#) flags set. This opaque handle may be copied into other processes and opened with [cuIpcOpenEventHandle](#) to allow efficient hardware synchronization between GPU work in different processes.

After the event has been opened in the importing process, [cuEventRecord](#), [cuEventSynchronize](#), [cuStreamWaitEvent](#) and [cuEventQuery](#) may be used in either process. Performing operations on the imported event after the exported event has been freed with [cuEventDestroy](#) will result in undefined behavior.

IPC functionality is restricted to devices with support for unified addressing on Linux and Windows operating systems. IPC functionality on Windows is restricted to GPUs in TCC mode

See also:

[cuEventCreate](#), [cuEventDestroy](#), [cuEventSynchronize](#), [cuEventQuery](#),
[cuStreamWaitEvent](#), [cuIpcOpenEventHandle](#), [cuIpcGetMemHandle](#),
[cuIpcOpenMemHandle](#), [cuIpcCloseMemHandle](#), [cudaIpcGetEventHandle](#)

CUresult cuIpcGetMemHandle (CUipcMemHandle *pHandle, CUdeviceptr dptr)

Gets an interprocess memory handle for an existing device memory allocation.

Parameters

pHandle

- Pointer to user allocated [CUipcMemHandle](#) to return the handle in.

dptr

- Base pointer to previously allocated device memory

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_HANDLE](#),
[CUDA_ERROR_OUT_OF_MEMORY](#), [CUDA_ERROR_MAP_FAILED](#),

Description

Takes a pointer to the base of an existing device memory allocation created with [cuMemAlloc](#) and exports it for use in another process. This is a lightweight operation and may be called multiple times on an allocation without adverse effects.

If a region of memory is freed with [cuMemFree](#) and a subsequent call to [cuMemAlloc](#) returns memory with the same device address, [cuIpcGetMemHandle](#) will return a unique handle for the new memory.

IPC functionality is restricted to devices with support for unified addressing on Linux and Windows operating systems. IPC functionality on Windows is restricted to GPUs in TCC mode

See also:

[cuMemAlloc](#), [cuMemFree](#), [cuIpcGetEventHandle](#), [cuIpcOpenEventHandle](#),
[cuIpcOpenMemHandle](#), [cuIpcCloseMemHandle](#), [cudaIpcGetMemHandle](#)

CUresult culpcOpenEventHandle (CUevent *phEvent, CUipcEventHandle handle)

Opens an interprocess event handle for use in the current process.

Parameters

phEvent

- Returns the imported event

handle

- Interprocess handle to open

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_MAP_FAILED, CUDA_ERROR_PEER_ACCESS_UNSUPPORTED,
CUDA_ERROR_INVALID_HANDLE

Description

Opens an interprocess event handle exported from another process with [cuIpcGetEventHandle](#). This function returns a [CUevent](#) that behaves like a locally created event with the [CU_EVENT_DISABLE_TIMING](#) flag specified. This event must be freed with [cuEventDestroy](#).

Performing operations on the imported event after the exported event has been freed with [cuEventDestroy](#) will result in undefined behavior.

IPC functionality is restricted to devices with support for unified addressing on Linux and Windows operating systems. IPC functionality on Windows is restricted to GPUs in TCC mode

See also:

[cuEventCreate](#), [cuEventDestroy](#), [cuEventSynchronize](#), [cuEventQuery](#),
[cuStreamWaitEvent](#), [cuIpcGetEventHandle](#), [cuIpcGetMemHandle](#),
[cuIpcOpenMemHandle](#), [cuIpcCloseMemHandle](#), [cudaIpcOpenEventHandle](#)

CUresult culpcOpenMemHandle (CUdeviceptr *pdptr, CUipcMemHandle handle, unsigned int Flags)

Opens an interprocess memory handle exported from another process and returns a device pointer usable in the local process.

Parameters

pdptr

- Returned device pointer

handle

- [CUipcMemHandle](#) to open

Flags

- Flags for this operation. Must be specified as [CU_IPC_MEM_LAZY_ENABLE_PEER_ACCESS](#)

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_MAP_FAILED](#), [CUDA_ERROR_INVALID_HANDLE](#),
[CUDA_ERROR_TOO_MANY_PEERS](#)

Description

Maps memory exported from another process with [cuIpcGetMemHandle](#) into the current device address space. For contexts on different devices [cuIpcOpenMemHandle](#) can attempt to enable peer access between the devices as if the user called [cuCtxEnablePeerAccess](#). This behavior is controlled by the [CU_IPC_MEM_LAZY_ENABLE_PEER_ACCESS](#) flag. [cuDeviceCanAccessPeer](#) can determine if a mapping is possible.

Contexts that may open [CUipcMemHandles](#) are restricted in the following way. [CUipcMemHandles](#) from each [CUdevice](#) in a given process may only be opened by one [CUcontext](#) per [CUdevice](#) per other process.

Memory returned from [cuIpcOpenMemHandle](#) must be freed with [cuIpcCloseMemHandle](#).

Calling [cuMemFree](#) on an exported memory region before calling [cuIpcCloseMemHandle](#) in the importing context will result in undefined behavior.

IPC functionality is restricted to devices with support for unified addressing on Linux and Windows operating systems. IPC functionality on Windows is restricted to GPUs in TCC mode



No guarantees are made about the address returned in `*pdptr`. In particular, multiple processes may not receive the same address for the same `handle`.

See also:

`cuMemAlloc`, `cuMemFree`, `cuIpcGetEventHandle`, `cuIpcOpenEventHandle`,
`cuIpcGetMemHandle`, `cuIpcCloseMemHandle`, `cuCtxEnablePeerAccess`,
`cuDeviceCanAccessPeer`, `cudaIpcOpenMemHandle`

CUresult cuMemAlloc (CUdeviceptr *dptr, size_t bytesize)

Allocates device memory.

Parameters

`dptr`

- Returned device pointer

`bytesize`

- Requested allocation size in bytes

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_OUT_OF_MEMORY`

Description

Allocates `bytesize` bytes of linear memory on the device and returns in `*dptr` a pointer to the allocated memory. The allocated memory is suitably aligned for any kind of variable. The memory is not cleared. If `bytesize` is 0, `cuMemAlloc()` returns `CUDA_ERROR_INVALID_VALUE`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`,
`cuArrayGetDescriptor`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`,
`cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`,
`cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`,

cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaMalloc

CUresult cuMemAllocHost (void **pp, size_t bytesize)

Allocates page-locked host memory.

Parameters

pp

- Returned host pointer to page-locked memory

bytesize

- Requested allocation size in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY

Description

Allocates `bytesize` bytes of host memory that is page-locked and accessible to the device. The driver tracks the virtual memory ranges allocated with this function and automatically accelerates calls to functions such as `cuMemcpy()`. Since the memory can be accessed directly by the device, it can be read or written with much higher bandwidth than pageable memory obtained with functions such as `malloc()`. Allocating excessive amounts of memory with `cuMemAllocHost()` may degrade system performance, since it reduces the amount of memory available to the system for paging. As a result, this function is best used sparingly to allocate staging areas for data exchange between host and device.

Note all host memory allocated using `cuMemHostAlloc()` will automatically be immediately accessible to all contexts on all devices which support unified addressing (as may be queried using `CU_DEVICE_ATTRIBUTE_UNIFIED_ADDRESSING`). The device pointer that may be used to access this host memory from those contexts is always equal to the returned host pointer `*pp`. See [Unified Addressing](#) for additional details.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaMallocHost

CUresult cuMemAllocManaged (CUdeviceptr *dptr, size_t bytesize, unsigned int flags)

Allocates memory that will be automatically managed by the Unified Memory system.

Parameters**dptr**

- Returned device pointer

bytesize

- Requested allocation size in bytes

flags

- Must be one of `CU_MEM_ATTACH_GLOBAL` or `CU_MEM_ATTACH_HOST`

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`, `CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`, `CUDA_ERROR_NOT_SUPPORTED`, `CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_OUT_OF_MEMORY`

Description

Allocates `bytesize` bytes of managed memory on the device and returns in `*dptr` a pointer to the allocated memory. If the device doesn't support allocating managed memory, `CUDA_ERROR_NOT_SUPPORTED` is returned. Support for managed memory can be queried using the device attribute `CU_DEVICE_ATTRIBUTE_MANAGED_MEMORY`. The allocated memory is suitably aligned for any kind of variable. The memory is not cleared. If `bytesize` is 0, `cuMemAllocManaged` returns `CUDA_ERROR_INVALID_VALUE`. The pointer is valid on the CPU and on all GPUs in the system that support managed memory. All accesses to this pointer must obey the Unified Memory programming model.

`flags` specifies the default stream association for this allocation. `flags` must be one of `CU_MEM_ATTACH_GLOBAL` or `CU_MEM_ATTACH_HOST`. If `CU_MEM_ATTACH_GLOBAL` is specified, then this memory is accessible from any stream on any device. If `CU_MEM_ATTACH_HOST` is specified, then the allocation should not be accessed from devices that have a zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`; an explicit call to `cuStreamAttachMemAsync` will be required to enable access on such devices.

If the association is later changed via `cuStreamAttachMemAsync` to a single stream, the default association as specified during `cuMemAllocManaged` is restored when that stream is destroyed. For `__managed__` variables, the default association is always `CU_MEM_ATTACH_GLOBAL`. Note that destroying a stream is an asynchronous operation, and as a result, the change to default association won't happen until all work in the stream has completed.

Memory allocated with `cuMemAllocManaged` should be released with `cuMemFree`.

Device memory oversubscription is possible for GPUs that have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`. Managed memory on such GPUs may be evicted from device memory to host memory at any time by the Unified Memory driver in order to make room for other allocations.

In a multi-GPU system where all GPUs have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`, managed memory may not be populated when this API returns and instead may be populated on access. In such systems, managed memory can migrate to any processor's memory at any time. The Unified Memory driver will employ heuristics to maintain data locality and prevent excessive page faults to the extent possible. The application can also guide the driver about memory usage patterns via `cuMemAdvise`. The application can also explicitly migrate memory to a desired processor's memory via `cuMemPrefetchAsync`.

In a multi-GPU system where all of the GPUs have a zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS` and all the GPUs have peer-to-peer support with each other, the physical storage for managed memory is created on the GPU which is active at the time `cuMemAllocManaged` is called. All other GPUs will reference the data at reduced bandwidth via peer mappings over the PCIe bus. The Unified Memory driver does not migrate memory among such GPUs.

In a multi-GPU system where not all GPUs have peer-to-peer support with each other and where the value of the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS` is zero for at least one of those GPUs, the location chosen for physical storage of managed memory is system-dependent.

- ▶ On Linux, the location chosen will be device memory as long as the current set of active contexts are on devices that either have peer-to-peer support with each other or have a non-zero value for the device attribute

`CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`. If there is an active context on a GPU that does not have a non-zero value for that device attribute and it does not have peer-to-peer support with the other devices that have active contexts on them, then the location for physical storage will be 'zero-copy' or host memory. Note that this means that managed memory that is located in device memory is migrated to host memory if a new context is created on a GPU that doesn't have a non-zero value for the device attribute and does not support peer-to-peer with at least one of the other devices that has an active context. This in turn implies that context creation may fail if there is insufficient host memory to migrate all managed allocations.

- ▶ On Windows, the physical storage is always created in 'zero-copy' or host memory. All GPUs will reference the data at reduced bandwidth over the PCIe bus. In these circumstances, use of the environment variable `CUDA_VISIBLE_DEVICES` is recommended to restrict CUDA to only use those GPUs that have peer-to-peer support. Alternatively, users can also set `CUDA_MANAGED_FORCE_DEVICE_ALLOC` to a non-zero value to force the driver to always use device memory for physical storage. When this environment variable is set to a non-zero value, all contexts created in that process on devices that support managed memory have to be peer-to-peer compatible with each other. Context creation will fail if a context is created on a device that supports managed memory and is not peer-to-peer compatible with any of the other managed memory supporting devices on which contexts were previously created, even if those contexts have been destroyed. These environment variables are described in the CUDA programming guide under the "CUDA environment variables" section.
- ▶ On ARM, managed memory is not available on discrete gpu with Drive PX-2.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cuDeviceGetAttribute`, `cuStreamAttachMemAsync`, `cudaMallocManaged`

CUresult cuMemAllocPitch (CUdeviceptr *dptr, size_t *pPitch, size_t WidthInBytes, size_t Height, unsigned int ElementSizeBytes)

Allocates pitched device memory.

Parameters

dptr

- Returned device pointer

pPitch

- Returned pitch of allocation in bytes

WidthInBytes

- Requested allocation width in bytes

Height

- Requested allocation height in rows

ElementSizeBytes

- Size of largest reads/writes for range

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY

Description

Allocates at least `WidthInBytes * Height` bytes of linear memory on the device and returns in `*dptr` a pointer to the allocated memory. The function may pad the allocation to ensure that corresponding pointers in any given row will continue to meet the alignment requirements for coalescing as the address is updated from row to row. `ElementSizeBytes` specifies the size of the largest reads and writes that will be performed on the memory range. `ElementSizeBytes` may be 4, 8 or 16 (since coalesced memory transactions are not possible on other data sizes). If `ElementSizeBytes` is smaller than the actual read/write size of a kernel, the kernel will run correctly, but possibly at reduced speed. The pitch returned in `*pPitch` by `cuMemAllocPitch()` is the width in bytes of the allocation. The intended usage of pitch is as a separate parameter of the allocation, used to compute addresses within the 2D array. Given the row and column of an array element of type `T`, the address is computed as:

```
↑ T* pElement = (T*)((char*)BaseAddress + Row * Pitch) + Column;
```

The pitch returned by `cuMemAllocPitch()` is guaranteed to work with `cuMemcpy2D()` under all circumstances. For allocations of 2D arrays, it is recommended that programmers consider performing pitch allocations using `cuMemAllocPitch()`. Due to alignment restrictions in the hardware, this is especially true if the application will be

performing 2D memory copies between different regions of device memory (whether linear memory or CUDA arrays).

The byte alignment of the pitch returned by `cuMemAllocPitch()` is guaranteed to match or exceed the alignment requirement for texture binding with `cuTexRefSetAddress2D()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaMallocPitch`

CUresult cuMemcpy (CUdeviceptr dst, CUdeviceptr src, size_t ByteCount)

Copies memory.

Parameters

dst

- Destination unified virtual address space pointer

src

- Source unified virtual address space pointer

ByteCount

- Size of memory copy in bytes

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Copies data between two pointers. `dst` and `src` are base pointers of the destination and source, respectively. `ByteCount` specifies the number of bytes to copy. Note that this function infers the type of the transfer (host to host, host to device, device to device, or device to host) from the pointer values. This function is only allowed in contexts which support unified addressing.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits *synchronous* behavior for most use cases.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaMemcpy`, `cudaMemcpyToSymbol`, `cudaMemcpyFromSymbol`

CUresult cuMemcpy2D (const CUDA_MEMCPY2D *pCopy)

Copies memory for 2D arrays.

Parameters

`pCopy`

- Parameters for the memory copy

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Perform a 2D memory copy according to the parameters specified in `pCopy`. The `CUDA_MEMCPY2D` structure is defined as:

```
↑ typedef struct CUDA_MEMCPY2D_st {
    unsigned int srcXInBytes, srcY;
    CUmemorytype srcMemoryType;
    const void *srcHost;
    CUdeviceptr srcDevice;
    CUarray srcArray;
    unsigned int srcPitch;

    unsigned int dstXInBytes, dstY;
    CUmemorytype dstMemoryType;
    void *dstHost;
    CUdeviceptr dstDevice;
    CUarray dstArray;
    unsigned int dstPitch;

    unsigned int WidthInBytes;
    unsigned int Height;
} CUDA_MEMCPY2D;
```

where:

- ▶ `srcMemoryType` and `dstMemoryType` specify the type of memory of the source and destination, respectively; `CUmemorytype_enum` is defined as:

```
↑ typedef enum CUmemorytype_enum {
    CU_MEMORYTYPE_HOST = 0x01,
    CU_MEMORYTYPE_DEVICE = 0x02,
    CU_MEMORYTYPE_ARRAY = 0x03,
    CU_MEMORYTYPE_UNIFIED = 0x04
} CUmemorytype;
```

If `srcMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `srcDevice` and `srcPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `srcArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `srcMemoryType` is `CU_MEMORYTYPE_HOST`, `srcHost` and `srcPitch` specify the (host) base address of the source data and the bytes per row to apply. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_DEVICE`, `srcDevice` and `srcPitch` specify the (device) base address of the source data and the bytes per row to apply. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_ARRAY`, `srcArray` specifies the handle of the source data. `srcHost`, `srcDevice` and `srcPitch` are ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_HOST`, `dstHost` and `dstPitch` specify the (host) base address of the destination data and the bytes per row to apply. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `dstDevice` and `dstPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `dstArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `dstMemoryType` is `CU_MEMORYTYPE_DEVICE`, `dstDevice` and `dstPitch` specify the (device) base address of the destination data and the bytes per row to apply. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_ARRAY`, `dstArray` specifies the handle of the destination data. `dstHost`, `dstDevice` and `dstPitch` are ignored.

- ▶ `srcXInBytes` and `srcY` specify the base address of the source data for the copy.

For host pointers, the starting address is

```
↑ void* Start = (void*)((char*)srcHost+srcY*srcPitch + srcXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr Start = srcDevice+srcY*srcPitch+srcXInBytes;
```

For CUDA arrays, `srcXInBytes` must be evenly divisible by the array element size.

- ▶ `dstXInBytes` and `dstY` specify the base address of the destination data for the copy.

For host pointers, the base address is

```
↑ void* dstStart = (void*)((char*)dstHost+dstY*dstPitch + dstXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr dstStart = dstDevice+dstY*dstPitch+dstXInBytes;
```

For CUDA arrays, `dstXInBytes` must be evenly divisible by the array element size.

- ▶ `WidthInBytes` and `Height` specify the width (in bytes) and height of the 2D copy being performed.
- ▶ If specified, `srcPitch` must be greater than or equal to `WidthInBytes + srcXInBytes`, and `dstPitch` must be greater than or equal to `WidthInBytes + dstXInBytes`.

`cuMemcpy2D()` returns an error if any pitch is greater than the maximum allowed (`CU_DEVICE_ATTRIBUTE_MAX_PITCH`). `cuMemAllocPitch()` passes back pitches that always work with `cuMemcpy2D()`. On intra-device memory copies (device to device, CUDA array to device, CUDA array to CUDA array), `cuMemcpy2D()` may fail for pitches not computed by `cuMemAllocPitch()`. `cuMemcpy2DUnaligned()` does not have this restriction, but may run significantly slower in the cases where `cuMemcpy2D()` would have returned an error code.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits *synchronous* behavior for most use cases.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaMemcpy2D, cudaMemcpy2DToArray, cudaMemcpy2DFromArray

CUresult cuMemcpy2DAsync (const CUDA_MEMCPY2D *pCopy, CUstream hStream)

Copies memory for 2D arrays.

Parameters**pCopy**

- Parameters for the memory copy

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Perform a 2D memory copy according to the parameters specified in `pCopy`. The `CUDA_MEMCPY2D` structure is defined as:

```
↑ typedef struct CUDA_MEMCPY2D_st {
    unsigned int srcXInBytes, srcY;
    CUmemorytype srcMemoryType;
    const void *srcHost;
    CUdeviceptr srcDevice;
    CUarray srcArray;
    unsigned int srcPitch;
    unsigned int dstXInBytes, dstY;
    CUmemorytype dstMemoryType;
    void *dstHost;
    CUdeviceptr dstDevice;
    CUarray dstArray;
    unsigned int dstPitch;
    unsigned int WidthInBytes;
    unsigned int Height;
} CUDA_MEMCPY2D;
```

where:

- ▶ `srcMemoryType` and `dstMemoryType` specify the type of memory of the source and destination, respectively; `CUmemorytype_enum` is defined as:

```
↑ typedef enum CUmemorytype_enum {
    CU_MEMORYTYPE_HOST = 0x01,
    CU_MEMORYTYPE_DEVICE = 0x02,
    CU_MEMORYTYPE_ARRAY = 0x03,
    CU_MEMORYTYPE_UNIFIED = 0x04
} CUmemorytype;
```

If `srcMemoryType` is `CU_MEMORYTYPE_HOST`, `srcHost` and `srcPitch` specify the (host) base address of the source data and the bytes per row to apply. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `srcDevice` and `srcPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `srcArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `srcMemoryType` is `CU_MEMORYTYPE_DEVICE`, `srcDevice` and `srcPitch` specify the (device) base address of the source data and the bytes per row to apply. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_ARRAY`, `srcArray` specifies the handle of the source data. `srcHost`, `srcDevice` and `srcPitch` are ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `dstDevice` and `dstPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `dstArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `dstMemoryType` is `CU_MEMORYTYPE_HOST`, `dstHost` and `dstPitch` specify the (host) base address of the destination data and the bytes per row to apply. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_DEVICE`, `dstDevice` and `dstPitch` specify the (device) base address of the destination data and the bytes per row to apply. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_ARRAY`, `dstArray` specifies the handle of the destination data. `dstHost`, `dstDevice` and `dstPitch` are ignored.

- ▶ `srcXInBytes` and `srcY` specify the base address of the source data for the copy.

For host pointers, the starting address is

```
↑ void* Start = (void*)((char*)srcHost+srcY*srcPitch + srcXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr Start = srcDevice+srcY*srcPitch+srcXInBytes;
```

For CUDA arrays, `srcXInBytes` must be evenly divisible by the array element size.

- ▶ `dstXInBytes` and `dstY` specify the base address of the destination data for the copy.

For host pointers, the base address is

```
↑ void* dstStart = (void*)((char*)dstHost+dstY*dstPitch + dstXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr dstStart = dstDevice+dstY*dstPitch+dstXInBytes;
```

For CUDA arrays, `dstXInBytes` must be evenly divisible by the array element size.

- ▶ `WidthInBytes` and `Height` specify the width (in bytes) and height of the 2D copy being performed.
- ▶ If specified, `srcPitch` must be greater than or equal to `WidthInBytes + srcXInBytes`, and `dstPitch` must be greater than or equal to `WidthInBytes + dstXInBytes`.
- ▶ If specified, `srcPitch` must be greater than or equal to `WidthInBytes + srcXInBytes`, and `dstPitch` must be greater than or equal to `WidthInBytes + dstXInBytes`.
- ▶ If specified, `srcHeight` must be greater than or equal to `Height + srcY`, and `dstHeight` must be greater than or equal to `Height + dstY`.

`cuMemcpy2DAsync()` returns an error if any pitch is greater than the maximum allowed (`CU_DEVICE_ATTRIBUTE_MAX_PITCH`). `cuMemAllocPitch()` passes back pitches that always work with `cuMemcpy2D()`. On intra-device memory copies (device to device, CUDA array to device, CUDA array to CUDA array), `cuMemcpy2DAsync()` may fail for pitches not computed by `cuMemAllocPitch()`.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async, cuMemsetD8, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cuMemsetD32Async, cudaMemcpy2DAsync, cudaMemcpy2DToArrayAsync, cudaMemcpy2DFromArrayAsync

CUresult cuMemcpy2DUnaligned (const CUDA_MEMCPY2D *pCopy)

Copies memory for 2D arrays.

Parameters**pCopy**

- Parameters for the memory copy

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Perform a 2D memory copy according to the parameters specified in `pCopy`. The `CUDA_MEMCPY2D` structure is defined as:

```
↑ typedef struct CUDA_MEMCPY2D_st {
    unsigned int srcXInBytes, srcY;
    CUmemorytype srcMemoryType;
    const void *srcHost;
    CUdeviceptr srcDevice;
    CUarray srcArray;
    unsigned int srcPitch;
    unsigned int dstXInBytes, dstY;
    CUmemorytype dstMemoryType;
    void *dstHost;
    CUdeviceptr dstDevice;
    CUarray dstArray;
    unsigned int dstPitch;
    unsigned int WidthInBytes;
    unsigned int Height;
} CUDA_MEMCPY2D;
```

where:

- ▶ `srcMemoryType` and `dstMemoryType` specify the type of memory of the source and destination, respectively; `CUmemorytype_enum` is defined as:

```
↑ typedef enum CUmemorytype_enum {
    CU_MEMORYTYPE_HOST = 0x01,
    CU_MEMORYTYPE_DEVICE = 0x02,
    CU_MEMORYTYPE_ARRAY = 0x03,
    CU_MEMORYTYPE_UNIFIED = 0x04
} CUmemorytype;
```

If `srcMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `srcDevice` and `srcPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `srcArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `srcMemoryType` is `CU_MEMORYTYPE_HOST`, `srcHost` and `srcPitch` specify the (host) base address of the source data and the bytes per row to apply. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_DEVICE`, `srcDevice` and `srcPitch` specify the (device) base address of the source data and the bytes per row to apply. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_ARRAY`, `srcArray` specifies the handle of the source data. `srcHost`, `srcDevice` and `srcPitch` are ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `dstDevice` and `dstPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `dstArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `dstMemoryType` is `CU_MEMORYTYPE_HOST`, `dstHost` and `dstPitch` specify the (host) base address of the destination data and the bytes per row to apply. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_DEVICE`, `dstDevice` and `dstPitch` specify the (device) base address of the destination data and the bytes per row to apply. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_ARRAY`, `dstArray` specifies the handle of the destination data. `dstHost`, `dstDevice` and `dstPitch` are ignored.

- ▶ `srcXInBytes` and `srcY` specify the base address of the source data for the copy.

For host pointers, the starting address is

```
↑ void* Start = (void*)((char*)srcHost+srcY*srcPitch + srcXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr Start = srcDevice+srcY*srcPitch+srcXInBytes;
```

For CUDA arrays, `srcXInBytes` must be evenly divisible by the array element size.

- ▶ `dstXInBytes` and `dstY` specify the base address of the destination data for the copy.

For host pointers, the base address is

```
↑ void* dstStart = (void*)((char*)dstHost+dstY*dstPitch + dstXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr dstStart = dstDevice+dstY*dstPitch+dstXInBytes;
```

For CUDA arrays, `dstXInBytes` must be evenly divisible by the array element size.

- ▶ `WidthInBytes` and `Height` specify the width (in bytes) and height of the 2D copy being performed.
- ▶ If specified, `srcPitch` must be greater than or equal to `WidthInBytes + srcXInBytes`, and `dstPitch` must be greater than or equal to `WidthInBytes + dstXInBytes`.

`cuMemcpy2D()` returns an error if any pitch is greater than the maximum allowed (`CU_DEVICE_ATTRIBUTE_MAX_PITCH`). `cuMemAllocPitch()` passes back pitches that always work with `cuMemcpy2D()`. On intra-device memory copies (device to device, CUDA array to device, CUDA array to CUDA array), `cuMemcpy2D()` may fail for pitches not computed by `cuMemAllocPitch()`. `cuMemcpy2DUnaligned()` does not have this restriction, but may run significantly slower in the cases where `cuMemcpy2D()` would have returned an error code.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.

- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaMemcpy2D, cudaMemcpy2DToArray, cudaMemcpy2DFromArray

CUresult cuMemcpy3D (const CUDA_MEMCPY3D *pCopy)

Copies memory for 3D arrays.

Parameters

pCopy

- Parameters for the memory copy

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Perform a 3D memory copy according to the parameters specified in `pCopy`. The `CUDA_MEMCPY3D` structure is defined as:

```
↑ typedef struct CUDA_MEMCPY3D_st {
    unsigned int srcXInBytes, srcY, srcZ;
    unsigned int srcLOD;
    CUmemorytype srcMemoryType;
    const void *srcHost;
    CUdeviceptr srcDevice;
    CUarray srcArray;
    unsigned int srcPitch; // ignored when src is array
    unsigned int srcHeight; // ignored when src is array; may
be 0 if Depth==1

    unsigned int dstXInBytes, dstY, dstZ;
    unsigned int dstLOD;
    CUmemorytype dstMemoryType;
    void *dstHost;
    CUdeviceptr dstDevice;
    CUarray dstArray;
    unsigned int dstPitch; // ignored when dst is array
    unsigned int dstHeight; // ignored when dst is array; may
be 0 if Depth==1

    unsigned int WidthInBytes;
    unsigned int Height;
    unsigned int Depth;
} CUDA_MEMCPY3D;
```

where:

- ▶ `srcMemoryType` and `dstMemoryType` specify the type of memory of the source and destination, respectively; `CUmemorytype_enum` is defined as:

```
↑ typedef enum CUmemorytype_enum {
    CU_MEMORYTYPE_HOST = 0x01,
    CU_MEMORYTYPE_DEVICE = 0x02,
    CU_MEMORYTYPE_ARRAY = 0x03,
    CU_MEMORYTYPE_UNIFIED = 0x04
} CUmemorytype;
```

If `srcMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `srcDevice` and `srcPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `srcArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `srcMemoryType` is `CU_MEMORYTYPE_HOST`, `srcHost`, `srcPitch` and `srcHeight` specify the (host) base address of the source data, the bytes per row, and the height of each 2D slice of the 3D array. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_DEVICE`, `srcDevice`, `srcPitch` and `srcHeight` specify the (device) base address of the source data, the bytes per row, and the height of each 2D slice of the 3D array. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_ARRAY`, `srcArray` specifies the handle of the source data. `srcHost`, `srcDevice`, `srcPitch` and `srcHeight` are ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `dstDevice` and `dstPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `dstArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `dstMemoryType` is `CU_MEMORYTYPE_HOST`, `dstHost` and `dstPitch` specify the (host) base address of the destination data, the bytes per row, and the height of each 2D slice of the 3D array. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_DEVICE`, `dstDevice` and `dstPitch` specify the (device) base address of the destination data, the bytes per row, and the height of each 2D slice of the 3D array. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_ARRAY`, `dstArray` specifies the handle of the destination data. `dstHost`, `dstDevice`, `dstPitch` and `dstHeight` are ignored.

- ▶ `srcXInBytes`, `srcY` and `srcZ` specify the base address of the source data for the copy.

For host pointers, the starting address is

```
↑ void* Start = (void*)((char*)srcHost+(srcZ*srcHeight+srcY)*srcPitch +
srcXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr Start = srcDevice+(srcZ*srcHeight+srcY)*srcPitch+srcXInBytes;
```

For CUDA arrays, `srcXInBytes` must be evenly divisible by the array element size.

- ▶ `dstXInBytes`, `dstY` and `dstZ` specify the base address of the destination data for the copy.

For host pointers, the base address is

```
↑ void* dstStart = (void*)((char*)dstHost+(dstZ*dstHeight+dstY)*dstPitch +
dstXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr dstStart = dstDevice+(dstZ*dstHeight+dstY)*dstPitch+dstXInBytes;
```

For CUDA arrays, `dstXInBytes` must be evenly divisible by the array element size.

- ▶ `WidthInBytes`, `Height` and `Depth` specify the width (in bytes), height and depth of the 3D copy being performed.
- ▶ If specified, `srcPitch` must be greater than or equal to `WidthInBytes + srcXInBytes`, and `dstPitch` must be greater than or equal to `WidthInBytes + dstXInBytes`.
- ▶ If specified, `srcHeight` must be greater than or equal to `Height + srcY`, and `dstHeight` must be greater than or equal to `Height + dstY`.

`cuMemcpy3D()` returns an error if any pitch is greater than the maximum allowed (`CUDA_DEVICE_ATTRIBUTE_MAX_PITCH`).

The `srcLOD` and `dstLOD` members of the `CUDA_MEMCPY3D` structure must be set to 0.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaMemcpy3D`

CUresult cuMemcpy3DAsync (const CUDA_MEMCPY3D *pCopy, CUstream hStream)

Copies memory for 3D arrays.

Parameters

pCopy

- Parameters for the memory copy

hStream

- Stream identifier

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`, `CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`, `CUDA_ERROR_INVALID_VALUE`

Description

Perform a 3D memory copy according to the parameters specified in `pCopy`. The `CUDA_MEMCPY3D` structure is defined as:

```
↑ typedef struct CUDA_MEMCPY3D_st {
    unsigned int srcXInBytes, srcY, srcZ;
    unsigned int srcLOD;
    CUmemorytype srcMemoryType;
    const void *srcHost;
    CUdeviceptr srcDevice;
    CUarray srcArray;
    unsigned int srcPitch; // ignored when src is array
    unsigned int srcHeight; // ignored when src is array; may
be 0 if Depth==1

    unsigned int dstXInBytes, dstY, dstZ;
    unsigned int dstLOD;
    CUmemorytype dstMemoryType;
    void *dstHost;
    CUdeviceptr dstDevice;
    CUarray dstArray;
    unsigned int dstPitch; // ignored when dst is array
    unsigned int dstHeight; // ignored when dst is array; may
be 0 if Depth==1

    unsigned int WidthInBytes;
    unsigned int Height;
    unsigned int Depth;
} CUDA_MEMCPY3D;
```

where:

- ▶ `srcMemoryType` and `dstMemoryType` specify the type of memory of the source and destination, respectively; `CUmemorytype_enum` is defined as:

```
↑ typedef enum CUmemorytype_enum {
    CU_MEMORYTYPE_HOST = 0x01,
    CU_MEMORYTYPE_DEVICE = 0x02,
    CU_MEMORYTYPE_ARRAY = 0x03,
    CU_MEMORYTYPE_UNIFIED = 0x04
} CUmemorytype;
```

If `srcMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `srcDevice` and `srcPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `srcArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `srcMemoryType` is `CU_MEMORYTYPE_HOST`, `srcHost`, `srcPitch` and `srcHeight` specify the (host) base address of the source data, the bytes per row, and the height of each 2D slice of the 3D array. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_DEVICE`, `srcDevice`, `srcPitch` and `srcHeight` specify the (device) base address of the source data, the bytes per row, and the height of each 2D slice of the 3D array. `srcArray` is ignored.

If `srcMemoryType` is `CU_MEMORYTYPE_ARRAY`, `srcArray` specifies the handle of the source data. `srcHost`, `srcDevice`, `srcPitch` and `srcHeight` are ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_UNIFIED`, `dstDevice` and `dstPitch` specify the (unified virtual address space) base address of the source data and the bytes per row to apply. `dstArray` is ignored. This value may be used only if unified addressing is supported in the calling context.

If `dstMemoryType` is `CU_MEMORYTYPE_HOST`, `dstHost` and `dstPitch` specify the (host) base address of the destination data, the bytes per row, and the height of each 2D slice of the 3D array. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_DEVICE`, `dstDevice` and `dstPitch` specify the (device) base address of the destination data, the bytes per row, and the height of each 2D slice of the 3D array. `dstArray` is ignored.

If `dstMemoryType` is `CU_MEMORYTYPE_ARRAY`, `dstArray` specifies the handle of the destination data. `dstHost`, `dstDevice`, `dstPitch` and `dstHeight` are ignored.

- ▶ `srcXInBytes`, `srcY` and `srcZ` specify the base address of the source data for the copy.

For host pointers, the starting address is

```
↑ void* Start = (void*)((char*)srcHost+(srcZ*srcHeight+srcY)*srcPitch +
srcXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr Start = srcDevice+(srcZ*srcHeight+srcY)*srcPitch+srcXInBytes;
```

For CUDA arrays, `srcXInBytes` must be evenly divisible by the array element size.

- ▶ `dstXInBytes`, `dstY` and `dstZ` specify the base address of the destination data for the copy.

For host pointers, the base address is

```
↑ void* dstStart = (void*)((char*)dstHost+(dstZ*dstHeight+dstY)*dstPitch +
dstXInBytes);
```

For device pointers, the starting address is

```
↑ CUdeviceptr dstStart = dstDevice+(dstZ*dstHeight+dstY)*dstPitch+dstXInBytes;
```

For CUDA arrays, `dstXInBytes` must be evenly divisible by the array element size.

- ▶ `WidthInBytes`, `Height` and `Depth` specify the width (in bytes), height and depth of the 3D copy being performed.
- ▶ If specified, `srcPitch` must be greater than or equal to `WidthInBytes + srcXInBytes`, and `dstPitch` must be greater than or equal to `WidthInBytes + dstXInBytes`.
- ▶ If specified, `srcHeight` must be greater than or equal to `Height + srcY`, and `dstHeight` must be greater than or equal to `Height + dstY`.

`cuMemcpy3DAsync()` returns an error if any pitch is greater than the maximum allowed (`CU_DEVICE_ATTRIBUTE_MAX_PITCH`).

The `srcLOD` and `dstLOD` members of the `CUDA_MEMCPY3D` structure must be set to 0.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D8Async`, `cuMemsetD2D16`, `cuMemsetD2D16Async`, `cuMemsetD2D32`, `cuMemsetD2D32Async`, `cuMemsetD8`, `cuMemsetD8Async`, `cuMemsetD16`, `cuMemsetD16Async`, `cuMemsetD32`, `cuMemsetD32Async`, `cudaMemcpy3DAsync`

CUresult cuMemcpy3DPeer (const CUDA_MEMCPY3D_PEER *pCopy)

Copies memory between contexts.

Parameters

`pCopy`

- Parameters for the memory copy

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Perform a 3D memory copy according to the parameters specified in `pCopy`. See the definition of the `CUDA_MEMCPY3D_PEER` structure for documentation of its parameters.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits `synchronous` behavior for most use cases.

See also:

`cuMemcpyDtoD`, `cuMemcpyPeer`, `cuMemcpyDtoDAsync`, `cuMemcpyPeerAsync`, `cuMemcpy3DPeerAsync`, `cudaMemcpy3DPeer`

CUresult cuMemcpy3DPeerAsync (const CUDA_MEMCPY3D_PEER *pCopy, CUstream hStream)

Copies memory between contexts asynchronously.

Parameters

`pCopy`

- Parameters for the memory copy

`hStream`

- Stream identifier

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Perform a 3D memory copy according to the parameters specified in `pCopy`. See the definition of the `CUDA_MEMCPY3D_PEER` structure for documentation of its parameters.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits `asynchronous` behavior for most use cases.
- ▶ This function uses standard `default stream` semantics.

See also:

[cuMemcpyDtoD](#), [cuMemcpyPeer](#), [cuMemcpyDtoDAsync](#), [cuMemcpyPeerAsync](#),
[cuMemcpy3DPeerAsync](#), [cudaMemcpy3DPeerAsync](#)

CUresult cuMemcpyAsync (CUdeviceptr dst, CUdeviceptr src, size_t ByteCount, CUstream hStream)

Copies memory asynchronously.

Parameters**dst**

- Destination unified virtual address space pointer

src

- Source unified virtual address space pointer

ByteCount

- Size of memory copy in bytes

hStream

- Stream identifier

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Copies data between two pointers. `dst` and `src` are base pointers of the destination and source, respectively. `ByteCount` specifies the number of bytes to copy. Note that this function infers the type of the transfer (host to host, host to device, device to device, or device to host) from the pointer values. This function is only allowed in contexts which support unified addressing.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),
[cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#), [cuMemcpy3D](#),

cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async, cuMemsetD8, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cuMemsetD32Async, cudaMemcpyAsync, cudaMemcpyToSymbolAsync, cudaMemcpyFromSymbolAsync

CUresult cuMemcpyAtoA (CUarray dstArray, size_t dstOffset, CUarray srcArray, size_t srcOffset, size_t ByteCount)

Copies memory from Array to Array.

Parameters

dstArray

- Destination array

dstOffset

- Offset in bytes of destination array

srcArray

- Source array

srcOffset

- Offset in bytes of source array

ByteCount

- Size of memory copy in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Copies from one 1D CUDA array to another. `dstArray` and `srcArray` specify the handles of the destination and source CUDA arrays for the copy, respectively. `dstOffset` and `srcOffset` specify the destination and source offsets in bytes into the CUDA arrays. `ByteCount` is the number of bytes to be copied. The size of the elements in the CUDA arrays need not be the same format, but the elements must be the same size; and count must be evenly divisible by that size.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits *synchronous* behavior for most use cases.

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#), [cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#), [cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#), [cuMemcpy3D](#), [cuMemcpy3DAsync](#), [cuMemcpyAtoD](#), [cuMemcpyAtoH](#), [cuMemcpyAtoHAsync](#), [cuMemcpyDtoA](#), [cuMemcpyDtoD](#), [cuMemcpyDtoDAsync](#), [cuMemcpyDtoH](#), [cuMemcpyDtoHAsync](#), [cuMemcpyHtoA](#), [cuMemcpyHtoAAsync](#), [cuMemcpyHtoD](#), [cuMemcpyHtoDAsync](#), [cuMemFree](#), [cuMemFreeHost](#), [cuMemGetAddressRange](#), [cuMemGetInfo](#), [cuMemHostAlloc](#), [cuMemHostGetDevicePointer](#), [cuMemsetD2D8](#), [cuMemsetD2D16](#), [cuMemsetD2D32](#), [cuMemsetD8](#), [cuMemsetD16](#), [cuMemsetD32](#), [cudaMemcpyArrayToArray](#)

CUresult cuMemcpyAtoD (CUdeviceptr dstDevice, CUarray srcArray, size_t srcOffset, size_t ByteCount)

Copies memory from Array to Device.

Parameters**dstDevice**

- Destination device pointer

srcArray

- Source array

srcOffset

- Offset in bytes of source array

ByteCount

- Size of memory copy in bytes

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#), [CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#), [CUDA_ERROR_INVALID_VALUE](#)

Description

Copies from one 1D CUDA array to device memory. `dstDevice` specifies the base pointer of the destination and must be naturally aligned with the CUDA array elements. `srcArray` and `srcOffset` specify the CUDA array handle and the offset in bytes into

the array where the copy is to begin. `ByteCount` specifies the number of bytes to copy and must be evenly divisible by the array element size.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaMemcpyFromArray`

CUresult cuMemcpyAtoH (void *dstHost, CUarray srcArray, size_t srcOffset, size_t ByteCount)

Copies memory from Array to Host.

Parameters

dstHost

- Destination device pointer

srcArray

- Source array

srcOffset

- Offset in bytes of source array

ByteCount

- Size of memory copy in bytes

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Copies from one 1D CUDA array to host memory. `dstHost` specifies the base pointer of the destination. `srcArray` and `srcOffset` specify the CUDA array handle and starting offset in bytes of the source data. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaMemcpyFromArray`

CUresult cuMemcpyAtoHAsync (void *dstHost, CUarray srcArray, size_t srcOffset, size_t ByteCount, CUstream hStream)

Copies memory from Array to Host.

Parameters

dstHost

- Destination pointer

srcArray

- Source array

srcOffset

- Offset in bytes of source array

ByteCount

- Size of memory copy in bytes

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Copies from one 1D CUDA array to host memory. `dstHost` specifies the base pointer of the destination. `srcArray` and `srcOffset` specify the CUDA array handle and starting offset in bytes of the source data. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`,
`cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`,
`cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`,
`cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`,
`cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`,
`cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`,
`cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`,
`cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`,
`cuMemsetD2D8Async`, `cuMemsetD2D16`, `cuMemsetD2D16Async`,
`cuMemsetD2D32`, `cuMemsetD2D32Async`, `cuMemsetD8`, `cuMemsetD8Async`,
`cuMemsetD16`, `cuMemsetD16Async`, `cuMemsetD32`, `cuMemsetD32Async`,
`cudaMemcpyFromArrayAsync`

CUresult cuMemcpyDtoA (CUarray dstArray, size_t dstOffset, CUdeviceptr srcDevice, size_t ByteCount)

Copies memory from Device to Array.

Parameters

dstArray

- Destination array

dstOffset

- Offset in bytes of destination array

srcDevice

- Source device pointer

ByteCount

- Size of memory copy in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Copies from device memory to a 1D CUDA array. `dstArray` and `dstOffset` specify the CUDA array handle and starting index of the destination data. `srcDevice` specifies the base pointer of the source. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`,
`cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`,
`cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`,
`cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`,
`cuMemcpyAtoHAsync`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`,
`cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`,
`cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`,
`cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`,
`cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`,
`cudaMemcpyToArray`

CUresult cuMemcpyDtoD (CUdeviceptr dstDevice, CUdeviceptr srcDevice, size_t ByteCount)

Copies memory from Device to Device.

Parameters**dstDevice**

- Destination device pointer

srcDevice

- Source device pointer

ByteCount

- Size of memory copy in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Copies from device memory to device memory. `dstDevice` and `srcDevice` are the base pointers of the destination and source, respectively. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`,
`cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`,
`cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`,
`cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`,
`cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`,
`cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`,
`cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`,
`cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`,
`cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaMemcpy`,
`cudaMemcpyToSymbol`, `cudaMemcpyFromSymbol`

CUresult cuMemcpyDtoDAsync (CUdeviceptr dstDevice, CUdeviceptr srcDevice, size_t ByteCount, CUstream hStream)

Copies memory from Device to Device.

Parameters**dstDevice**

- Destination device pointer

srcDevice

- Source device pointer

ByteCount

- Size of memory copy in bytes

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Copies from device memory to device memory. `dstDevice` and `srcDevice` are the base pointers of the destination and source, respectively. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`,
`cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`,
`cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`,
`cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`,
`cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoH`,
`cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`,
`cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`,
`cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`,
`cuMemsetD2D8Async`, `cuMemsetD2D16`, `cuMemsetD2D16Async`, `cuMemsetD2D32`,
`cuMemsetD2D32Async`, `cuMemsetD8`, `cuMemsetD8Async`, `cuMemsetD16`,
`cuMemsetD16Async`, `cuMemsetD32`, `cuMemsetD32Async`, `cudaMemcpyAsync`,
`cudaMemcpyToSymbolAsync`, `cudaMemcpyFromSymbolAsync`

CUresult cuMemcpyDtoH (void *dstHost, CUdeviceptr srcDevice, size_t ByteCount)

Copies memory from Device to Host.

Parameters

dstHost

- Destination host pointer

srcDevice

- Source device pointer

ByteCount

- Size of memory copy in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Copies from device to host memory. `dstHost` and `srcDevice` specify the base pointers of the destination and source, respectively. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`,
`cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`,
`cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`,
`cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`,
`cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`,
`cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`,
`cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`,
`cuMemGetInfo`, `cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`,
`cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`,
`cudaMemcpy`, `cudaMemcpyFromSymbol`

CUresult cuMemcpyDtoHAsync (void *dstHost, CUdeviceptr srcDevice, size_t ByteCount, CUstream hStream)

Copies memory from Device to Host.

Parameters

dstHost

- Destination host pointer

srcDevice

- Source device pointer

ByteCount

- Size of memory copy in bytes

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Copies from device to host memory. `dstHost` and `srcDevice` specify the base pointers of the destination and source, respectively. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`,

cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async, cuMemsetD8, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cuMemsetD32Async, cudaMemcpyAsync, cudaMemcpyFromSymbolAsync

CUresult cudaMemcpyHtoA (CUarray dstArray, size_t dstOffset, const void *srcHost, size_t ByteCount)

Copies memory from Host to Array.

Parameters

dstArray

- Destination array

dstOffset

- Offset in bytes of destination array

srcHost

- Source host pointer

ByteCount

- Size of memory copy in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Copies from host memory to a 1D CUDA array. `dstArray` and `dstOffset` specify the CUDA array handle and starting offset in bytes of the destination data. `pSrc` specifies the base address of the source. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits *synchronous* behavior for most use cases.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cudaMemcpy2D, cudaMemcpy2DAsync, cudaMemcpy2DUnaligned, cudaMemcpy3D, cudaMemcpy3DAsync, cudaMemcpyAtoA, cudaMemcpyAtoD, cudaMemcpyAtoH,

cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaMemcpyToArray

CUresult cuMemcpyHtoAAsync (CUarray dstArray, size_t dstOffset, const void *srcHost, size_t ByteCount, CUstream hStream)

Copies memory from Host to Array.

Parameters

dstArray

- Destination array

dstOffset

- Offset in bytes of destination array

srcHost

- Source host pointer

ByteCount

- Size of memory copy in bytes

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Copies from host memory to a 1D CUDA array. `dstArray` and `dstOffset` specify the CUDA array handle and starting offset in bytes of the destination data. `srcHost` specifies the base address of the source. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async, cuMemsetD8, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cuMemsetD32Async, cudaMemcpyToArrayAsync

CUresult cudaMemcpyHtoD (CUdeviceptr dstDevice, const void *srcHost, size_t ByteCount)

Copies memory from Host to Device.

Parameters**dstDevice**

- Destination device pointer

srcHost

- Source host pointer

ByteCount

- Size of memory copy in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Copies from host memory to device memory. `dstDevice` and `srcHost` are the base addresses of the destination and source, respectively. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaMemcpy, cudaMemcpyToSymbol

CUresult cuMemcpyHtoDAsync (CUdeviceptr dstDevice, const void *srcHost, size_t ByteCount, CUstream hStream)

Copies memory from Host to Device.

Parameters**dstDevice**

- Destination device pointer

srcHost

- Source host pointer

ByteCount

- Size of memory copy in bytes

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Copies from host memory to device memory. `dstDevice` and `srcHost` are the base addresses of the destination and source, respectively. `ByteCount` specifies the number of bytes to copy.



► Note that this function may also return error codes from previous, asynchronous launches.

- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async, cuMemsetD8, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cuMemsetD32Async, cudaMemcpyAsync, cudaMemcpyToSymbolAsync

CUresult cuMemcpyPeer (CUdeviceptr dstDevice, CUcontext dstContext, CUdeviceptr srcDevice, CUcontext srcContext, size_t ByteCount)

Copies device memory between two contexts.

Parameters**dstDevice**

- Destination device pointer

dstContext

- Destination context

srcDevice

- Source device pointer

srcContext

- Source context

ByteCount

- Size of memory copy in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Copies from device memory in one context to device memory in another context. `dstDevice` is the base device pointer of the destination memory and `dstContext` is the destination context. `srcDevice` is the base device pointer of the source memory and `srcContext` is the source pointer. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **synchronous** behavior for most use cases.

See also:

[cuMemcpyDtoD](#), [cuMemcpy3DPeer](#), [cuMemcpyDtoDAsync](#), [cuMemcpyPeerAsync](#), [cuMemcpy3DPeerAsync](#), [cudaMemcpyPeer](#)

CUresult cuMemcpyPeerAsync (CUdeviceptr dstDevice, CUcontext dstContext, CUdeviceptr srcDevice, CUcontext srcContext, size_t ByteCount, CUstream hStream)

Copies device memory between two contexts asynchronously.

Parameters

dstDevice

- Destination device pointer

dstContext

- Destination context

srcDevice

- Source device pointer

srcContext

- Source context

ByteCount

- Size of memory copy in bytes

hStream

- Stream identifier

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#), [CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#), [CUDA_ERROR_INVALID_VALUE](#)

Description

Copies from device memory in one context to device memory in another context. `dstDevice` is the base device pointer of the destination memory and `dstContext` is the destination context. `srcDevice` is the base device pointer of the source memory and `srcContext` is the source pointer. `ByteCount` specifies the number of bytes to copy.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits [asynchronous](#) behavior for most use cases.
- ▶ This function uses standard [default stream](#) semantics.

See also:

[cuMemcpyDtoD](#), [cuMemcpyPeer](#), [cuMemcpy3DPeer](#), [cuMemcpyDtoDAsync](#), [cuMemcpy3DPeerAsync](#), [cudaMemcpyPeerAsync](#)

CUresult cuMemFree (CUdeviceptr dptr)

Frees device memory.

Parameters

`dptr`

- Pointer to memory to free

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Frees the memory space pointed to by `dptr`, which must have been returned by a previous call to [cuMemAlloc\(\)](#) or [cuMemAllocPitch\(\)](#).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),

cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaFree

CUresult cuMemFreeHost (void *p)

Frees page-locked host memory.

Parameters

p
- Pointer to memory to free

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Frees the memory space pointed to by *p*, which must have been returned by a previous call to `cuMemAllocHost()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16, cuMemsetD32, cudaFreeHost

CUresult cuMemGetAddressRange (CUdeviceptr *pbase, size_t *psize, CUdeviceptr dptr)

Get information on memory allocations.

Parameters

pbase

- Returned base address

psize

- Returned size of device memory allocation

dptr

- Device pointer to query

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_NOT_FOUND, CUDA_ERROR_INVALID_VALUE

Description

Returns the base address in *pbase* and size in *psize* of the allocation by `cuMemAlloc()` or `cuMemAllocPitch()` that contains the input pointer *dptr*. Both parameters *pbase* and *psize* are optional. If one of them is NULL, it is ignored.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`,
`cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`,
`cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`,
`cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`,
`cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`,
`cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`,
`cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetInfo`,
`cuMemHostAlloc`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`,
`cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`

CUresult cuMemGetInfo (size_t *free, size_t *total)

Gets free and total memory.

Parameters

free

- Returned free memory in bytes

total

- Returned total memory in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns in *free and *total respectively, the free and total amount of memory available for allocation by the CUDA context, in bytes.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy,
 cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch,
 cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D,
 cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH,
 cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync,
 cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync,
 cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost,
 cuMemGetAddressRange, cuMemHostAlloc, cuMemHostGetDevicePointer,
 cuMemsetD2D8, cuMemsetD2D16, cuMemsetD2D32, cuMemsetD8, cuMemsetD16,
 cuMemsetD32, cudaMemGetInfo

CUresult cuMemHostAlloc (void **pp, size_t bytesize, unsigned int Flags)

Allocates page-locked host memory.

Parameters

pp

- Returned host pointer to page-locked memory

bytesize

- Requested allocation size in bytes

Flags

- Flags for allocation request

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY

Description

Allocates `bytesize` bytes of host memory that is page-locked and accessible to the device. The driver tracks the virtual memory ranges allocated with this function and automatically accelerates calls to functions such as `cuMemcpyHtoD()`. Since the memory can be accessed directly by the device, it can be read or written with much higher bandwidth than pageable memory obtained with functions such as `malloc()`. Allocating excessive amounts of pinned memory may degrade system performance, since it reduces the amount of memory available to the system for paging. As a result, this function is best used sparingly to allocate staging areas for data exchange between host and device.

The `Flags` parameter enables different options to be specified that affect the allocation, as follows.

- ▶ **CU_MEMHOSTALLOC_PORTABLE**: The memory returned by this call will be considered as pinned memory by all CUDA contexts, not just the one that performed the allocation.
- ▶ **CU_MEMHOSTALLOC_DEVICEMAP**: Maps the allocation into the CUDA address space. The device pointer to the memory may be obtained by calling `cuMemHostGetDevicePointer()`.
- ▶ **CU_MEMHOSTALLOC_WRITECOMBINED**: Allocates the memory as write-combined (WC). WC memory can be transferred across the PCI Express bus more quickly on some system configurations, but cannot be read efficiently by most CPUs. WC memory is a good option for buffers that will be written by the CPU and read by the GPU via mapped pinned memory or host->device transfers.

All of these flags are orthogonal to one another: a developer may allocate memory that is portable, mapped and/or write-combined with no restrictions.

The CUDA context must have been created with the `CU_CTX_MAP_HOST` flag in order for the `CU_MEMHOSTALLOC_DEVICEMAP` flag to have any effect.

The `CU_MEMHOSTALLOC_DEVICEMAP` flag may be specified on CUDA contexts for devices that do not support mapped pinned memory. The failure is deferred to `cuMemHostGetDevicePointer()` because the memory may be mapped into other CUDA contexts via the `CU_MEMHOSTALLOC_PORTABLE` flag.

The memory allocated by this function must be freed with `cuMemFreeHost()`.

Note all host memory allocated using `cuMemHostAlloc()` will automatically be immediately accessible to all contexts on all devices which support unified addressing (as may be queried using `CU_DEVICE_ATTRIBUTE_UNIFIED_ADDRESSING`). Unless the flag `CU_MEMHOSTALLOC_WRITECOMBINED` is specified, the device pointer that may be used to access this host memory from those contexts is always equal to the returned host pointer `*pp`. If the flag `CU_MEMHOSTALLOC_WRITECOMBINED` is specified, then the function `cuMemHostGetDevicePointer()` must be used to query the device pointer, even if the context supports unified addressing. See [Unified Addressing](#) for additional details.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaHostAlloc`

CUresult cuMemHostGetDevicePointer (CUdeviceptr *pdptr, void *p, unsigned int Flags)

Passes back device pointer of mapped pinned memory.

Parameters

pdptr

- Returned device pointer

p

- Host pointer

Flags

- Options (must be 0)

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Passes back the device pointer `pdptr` corresponding to the mapped, pinned host buffer `p` allocated by `cuMemHostAlloc`.

`cuMemHostGetDevicePointer()` will fail if the `CU_MEMHOSTALLOC_DEVICEMAP` flag was not specified at the time the memory was allocated, or if the function is called on a GPU that does not support mapped pinned memory.

For devices that have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CAN_USE_HOST_POINTER_FOR_REGISTERED_MEM`, the memory can also be accessed from the device using the host pointer `p`. The device pointer returned by `cuMemHostGetDevicePointer()` may or may not match the original host pointer `p` and depends on the devices visible to the application. If all devices visible to the application have a non-zero value for the device attribute, the device pointer returned by `cuMemHostGetDevicePointer()` will match the original pointer `p`. If any device visible to the application has a zero value for the device attribute, the device pointer returned by `cuMemHostGetDevicePointer()` will not match the original host pointer `p`, but it will be suitable for use on all devices provided Unified Virtual Addressing is enabled. In such systems, it is valid to access the memory using either pointer on devices that have a non-zero value for the device attribute. Note however that such devices should access the memory using only one of the two pointers and not both.

`Flags` provides for future releases. For now, it must be set to 0.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`, `cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`, `cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`, `cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`, `cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`, `cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`, `cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`, `cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`, `cuMemsetD2D8`, `cuMemsetD2D16`, `cuMemsetD2D32`, `cuMemsetD8`, `cuMemsetD16`, `cuMemsetD32`, `cudaHostGetDevicePointer`

CUresult cuMemHostGetFlags (unsigned int *pFlags, void *p)

Passes back flags that were used for a pinned allocation.

Parameters

pFlags

- Returned flags word

p

- Host pointer

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Passes back the flags `pFlags` that were specified when allocating the pinned host buffer `p` allocated by `cuMemHostAlloc`.

`cuMemHostGetFlags()` will fail if the pointer does not reside in an allocation performed by `cuMemAllocHost()` or `cuMemHostAlloc()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuMemAllocHost](#), [cuMemHostAlloc](#), [cudaHostGetFlags](#)

CUresult cuMemHostRegister (void *p, size_t bytesize, unsigned int Flags)

Registers an existing host memory range for use by CUDA.

Parameters

p

- Host pointer to memory to page-lock

bytesize

- Size in bytes of the address range to page-lock

Flags

- Flags for allocation request

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY,
 CUDA_ERROR_HOST_MEMORY_ALREADY_REGISTERED,
 CUDA_ERROR_NOT_PERMITTED, CUDA_ERROR_NOT_SUPPORTED

Description

Page-locks the memory range specified by `p` and `bytesize` and maps it for the device(s) as specified by `Flags`. This memory range also is added to the same tracking mechanism as `cuMemHostAlloc` to automatically accelerate calls to functions such as `cuMemcpyHtoD()`. Since the memory can be accessed directly by the device, it can be read or written with much higher bandwidth than pageable memory that has not been registered. Page-locking excessive amounts of memory may degrade system performance, since it reduces the amount of memory available to the system for paging. As a result, this function is best used sparingly to register staging areas for data exchange between host and device.

This function has limited support on Mac OS X. OS 10.7 or higher is required.

The `Flags` parameter enables different options to be specified that affect the allocation, as follows.

- ▶ `CU_MEMHOSTREGISTER_PORTABLE`: The memory returned by this call will be considered as pinned memory by all CUDA contexts, not just the one that performed the allocation.
- ▶ `CU_MEMHOSTREGISTER_DEVICEMAP`: Maps the allocation into the CUDA address space. The device pointer to the memory may be obtained by calling `cuMemHostGetDevicePointer()`.
- ▶ `CU_MEMHOSTREGISTER_IOMEMORY`: The pointer is treated as pointing to some I/O memory space, e.g. the PCI Express resource of a 3rd party device.

All of these flags are orthogonal to one another: a developer may page-lock memory that is portable or mapped with no restrictions.

The CUDA context must have been created with the `CU_CTX_MAP_HOST` flag in order for the `CU_MEMHOSTREGISTER_DEVICEMAP` flag to have any effect.

The `CU_MEMHOSTREGISTER_DEVICEMAP` flag may be specified on CUDA contexts for devices that do not support mapped pinned memory. The failure is deferred to `cuMemHostGetDevicePointer()` because the memory may be mapped into other CUDA contexts via the `CU_MEMHOSTREGISTER_PORTABLE` flag.

For devices that have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CAN_USE_HOST_POINTER_FOR_REGISTERED_MEM`, the memory can also be accessed from the device using the host pointer `p`. The device pointer returned by `cuMemHostGetDevicePointer()` may or may not match the original host pointer `ptr` and depends on the devices visible to the application. If all devices visible to the application have a non-zero value for the device attribute, the device pointer returned by `cuMemHostGetDevicePointer()` will match the original pointer `ptr`. If any device visible to the application has a zero value for the device attribute, the device pointer returned by `cuMemHostGetDevicePointer()` will not match the original host pointer `ptr`, but it will be suitable for use on all devices provided Unified Virtual Addressing is enabled. In such systems, it is valid to access the memory using either pointer on devices that have a non-zero value for the device attribute. Note however that such devices should access the memory using only one of the two pointers and not both.

The memory page-locked by this function must be unregistered with `cuMemHostUnregister()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuMemHostUnregister`, `cuMemHostGetFlags`, `cuMemHostGetDevicePointer`, `cudaHostRegister`

CUresult cuMemHostUnregister (void *p)

Unregisters a memory range that was registered with cuMemHostRegister.

Parameters

p

- Host pointer to memory to unregister

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY,
 CUDA_ERROR_HOST_MEMORY_NOT_REGISTERED,

Description

Unmaps the memory range whose base address is specified by *p*, and makes it pageable again.

The base address must be the same one specified to cuMemHostRegister().



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuMemHostRegister, cudaHostUnregister

CUresult cuMemsetD16 (CUdeviceptr dstDevice, unsigned short us, size_t N)

Initializes device memory.

Parameters

dstDevice

- Destination device pointer

us

- Value to set

N

- Number of elements

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Sets the memory range of N 16-bit values to the specified value us . The `dstDevice` pointer must be two byte aligned.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),
[cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#),
[cuMemcpy3D](#), [cuMemcpy3DAsync](#), [cuMemcpyAtoA](#), [cuMemcpyAtoD](#),
[cuMemcpyAtoH](#), [cuMemcpyAtoHAsync](#), [cuMemcpyDtoA](#), [cuMemcpyDtoD](#),
[cuMemcpyDtoDAsync](#), [cuMemcpyDtoH](#), [cuMemcpyDtoHAsync](#), [cuMemcpyHtoA](#),
[cuMemcpyHtoAAsync](#), [cuMemcpyHtoD](#), [cuMemcpyHtoDAsync](#), [cuMemFree](#),
[cuMemFreeHost](#), [cuMemGetAddressRange](#), [cuMemGetInfo](#), [cuMemHostAlloc](#),
[cuMemHostGetDevicePointer](#), [cuMemsetD2D8](#), [cuMemsetD2D8Async](#),
[cuMemsetD2D16](#), [cuMemsetD2D16Async](#), [cuMemsetD2D32](#), [cuMemsetD2D32Async](#),
[cuMemsetD8](#), [cuMemsetD8Async](#), [cuMemsetD16Async](#), [cuMemsetD32](#),
[cuMemsetD32Async](#), [cudaMemset](#)

CUresult cuMemsetD16Async (CUdeviceptr dstDevice, unsigned short us, size_t N, CUstream hStream)

Sets device memory.

Parameters**dstDevice**

- Destination device pointer

us

- Value to set

N

- Number of elements

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Sets the memory range of N 16-bit values to the specified value us . The `dstDevice` pointer must be two byte aligned.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).
- ▶ This function uses standard [default stream semantics](#).

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),
[cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#),
[cuMemcpy3D](#), [cuMemcpy3DAsync](#), [cuMemcpyAtoA](#), [cuMemcpyAtoD](#),
[cuMemcpyAtoH](#), [cuMemcpyAtoHAsync](#), [cuMemcpyDtoA](#), [cuMemcpyDtoD](#),
[cuMemcpyDtoDAsync](#), [cuMemcpyDtoH](#), [cuMemcpyDtoHAsync](#), [cuMemcpyHtoA](#),
[cuMemcpyHtoAAsync](#), [cuMemcpyHtoD](#), [cuMemcpyHtoDAsync](#), [cuMemFree](#),
[cuMemFreeHost](#), [cuMemGetAddressRange](#), [cuMemGetInfo](#), [cuMemHostAlloc](#),
[cuMemHostGetDevicePointer](#), [cuMemsetD2D8](#), [cuMemsetD2D8Async](#),
[cuMemsetD2D16](#), [cuMemsetD2D16Async](#), [cuMemsetD2D32](#), [cuMemsetD2D32Async](#),
[cuMemsetD8](#), [cuMemsetD8Async](#), [cuMemsetD16](#), [cuMemsetD32](#), [cuMemsetD32Async](#),
[cudaMemsetAsync](#)

CUresult cuMemsetD2D16 (CUdeviceptr dstDevice, size_t dstPitch, unsigned short us, size_t Width, size_t Height)

Initializes device memory.

Parameters**dstDevice**

- Destination device pointer

dstPitch

- Pitch of destination device pointer

us

- Value to set

Width

- Width of row

Height

- Number of rows

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Sets the 2D memory range of `width` 16-bit values to the specified value `us`. `Height` specifies the number of rows to set, and `dstPitch` specifies the number of bytes between each row. The `dstDevice` pointer and `dstPitch` offset must be two byte aligned. This function performs fastest when the pitch is one that has been passed back by `cuMemAllocPitch()`.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),
[cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#),
[cuMemcpy3D](#), [cuMemcpy3DAsync](#), [cuMemcpyAtoA](#), [cuMemcpyAtoD](#),
[cuMemcpyAtoH](#), [cuMemcpyAtoHAsync](#), [cuMemcpyDtoA](#), [cuMemcpyDtoD](#),
[cuMemcpyDtoDAsync](#), [cuMemcpyDtoH](#), [cuMemcpyDtoHAsync](#), [cuMemcpyHtoA](#),
[cuMemcpyHtoAAsync](#), [cuMemcpyHtoD](#), [cuMemcpyHtoDAsync](#), [cuMemFree](#),
[cuMemFreeHost](#), [cuMemGetAddressRange](#), [cuMemGetInfo](#), [cuMemHostAlloc](#),
[cuMemHostGetDevicePointer](#), [cuMemsetD2D8](#), [cuMemsetD2D8Async](#),
[cuMemsetD2D16Async](#), [cuMemsetD2D32](#), [cuMemsetD2D32Async](#), [cuMemsetD8](#),
[cuMemsetD8Async](#), [cuMemsetD16](#), [cuMemsetD16Async](#), [cuMemsetD32](#),
[cuMemsetD32Async](#), [cudaMemset2D](#)

CUresult cuMemsetD2D16Async (CUdeviceptr dstDevice, size_t dstPitch, unsigned short us, size_t Width, size_t Height, CUstream hStream)

Sets device memory.

Parameters

dstDevice

- Destination device pointer

dstPitch

- Pitch of destination device pointer

us

- Value to set

Width

- Width of row

Height

- Number of rows

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Sets the 2D memory range of `Width` 16-bit values to the specified value `us`. `Height` specifies the number of rows to set, and `dstPitch` specifies the number of bytes between each row. The `dstDevice` pointer and `dstPitch` offset must be two byte aligned. This function performs fastest when the pitch is one that has been passed back by `cuMemAllocPitch()`.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).
- ▶ This function uses standard [default stream](#) semantics.

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),
[cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#),
[cuMemcpy3D](#), [cuMemcpy3DAsync](#), [cuMemcpyAtoA](#), [cuMemcpyAtoD](#),
[cuMemcpyAtoH](#), [cuMemcpyAtoHAsync](#), [cuMemcpyDtoA](#), [cuMemcpyDtoD](#),
[cuMemcpyDtoDAsync](#), [cuMemcpyDtoH](#), [cuMemcpyDtoHAsync](#), [cuMemcpyHtoA](#),
[cuMemcpyHtoAAsync](#), [cuMemcpyHtoD](#), [cuMemcpyHtoDAsync](#), [cuMemFree](#),
[cuMemFreeHost](#), [cuMemGetAddressRange](#), [cuMemGetInfo](#), [cuMemHostAlloc](#),
[cuMemHostGetDevicePointer](#), [cuMemsetD2D8](#), [cuMemsetD2D8Async](#),
[cuMemsetD2D16](#), [cuMemsetD2D32](#), [cuMemsetD2D32Async](#), [cuMemsetD8](#),
[cuMemsetD8Async](#), [cuMemsetD16](#), [cuMemsetD16Async](#), [cuMemsetD32](#),
[cuMemsetD32Async](#), [cudaMemset2DAsync](#)

CUresult cuMemsetD2D32 (CUdeviceptr dstDevice, size_t dstPitch, unsigned int ui, size_t Width, size_t Height)

Initializes device memory.

Parameters

dstDevice

- Destination device pointer

dstPitch

- Pitch of destination device pointer

ui

- Value to set

Width

- Width of row

Height

- Number of rows

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Sets the 2D memory range of `Width` 32-bit values to the specified value `ui`. `Height` specifies the number of rows to set, and `dstPitch` specifies the number of bytes between each row. The `dstDevice` pointer and `dstPitch` offset must be four byte aligned. This function performs fastest when the pitch is one that has been passed back by [cuMemAllocPitch\(\)](#).



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32Async, cuMemsetD8, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cuMemsetD32Async, cudaMemset2D

CUresult cuMemsetD2D32Async (CUdeviceptr dstDevice, size_t dstPitch, unsigned int ui, size_t Width, size_t Height, CUstream hStream)

Sets device memory.

Parameters**dstDevice**

- Destination device pointer

dstPitch

- Pitch of destination device pointer

ui

- Value to set

Width

- Width of row

Height

- Number of rows

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Sets the 2D memory range of `Width` 32-bit values to the specified value `ui`. `Height` specifies the number of rows to set, and `dstPitch` specifies the number of bytes between each row. The `dstDevice` pointer and `dstPitch` offset must be four byte aligned. This function performs fastest when the pitch is one that has been passed back by `cuMemAllocPitch()`.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).
- ▶ This function uses standard [default stream semantics](#).

See also:

`cuArray3DCreate`, `cuArray3DGetDescriptor`, `cuArrayCreate`, `cuArrayDestroy`,
`cuArrayGetDescriptor`, `cuMemAlloc`, `cuMemAllocHost`, `cuMemAllocPitch`,
`cuMemcpy2D`, `cuMemcpy2DAsync`, `cuMemcpy2DUnaligned`,
`cuMemcpy3D`, `cuMemcpy3DAsync`, `cuMemcpyAtoA`, `cuMemcpyAtoD`,
`cuMemcpyAtoH`, `cuMemcpyAtoHAsync`, `cuMemcpyDtoA`, `cuMemcpyDtoD`,
`cuMemcpyDtoDAsync`, `cuMemcpyDtoH`, `cuMemcpyDtoHAsync`, `cuMemcpyHtoA`,
`cuMemcpyHtoAAsync`, `cuMemcpyHtoD`, `cuMemcpyHtoDAsync`, `cuMemFree`,
`cuMemFreeHost`, `cuMemGetAddressRange`, `cuMemGetInfo`, `cuMemHostAlloc`,
`cuMemHostGetDevicePointer`, `cuMemsetD2D8`, `cuMemsetD2D8Async`,
`cuMemsetD2D16`, `cuMemsetD2D16Async`, `cuMemsetD2D32`, `cuMemsetD8`,
`cuMemsetD8Async`, `cuMemsetD16`, `cuMemsetD16Async`, `cuMemsetD32`,
`cuMemsetD32Async`, `cudaMemset2DAsync`

CUresult cuMemsetD2D8 (CUdeviceptr dstDevice, size_t dstPitch, unsigned char uc, size_t Width, size_t Height)

Initializes device memory.

Parameters

`dstDevice`

- Destination device pointer

dstPitch

- Pitch of destination device pointer

uc

- Value to set

Width

- Width of row

Height

- Number of rows

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Sets the 2D memory range of width 8-bit values to the specified value `uc`. `Height` specifies the number of rows to set, and `dstPitch` specifies the number of bytes between each row. This function performs fastest when the pitch is one that has been passed back by `cuMemAllocPitch()`.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),
[cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#),
[cuMemcpy3D](#), [cuMemcpy3DAsync](#), [cuMemcpyAtoA](#), [cuMemcpyAtoD](#),
[cuMemcpyAtoH](#), [cuMemcpyAtoHAsync](#), [cuMemcpyDtoA](#), [cuMemcpyDtoD](#),
[cuMemcpyDtoDAsync](#), [cuMemcpyDtoH](#), [cuMemcpyDtoHAsync](#), [cuMemcpyHtoA](#),
[cuMemcpyHtoAAsync](#), [cuMemcpyHtoD](#), [cuMemcpyHtoDAsync](#), [cuMemFree](#),
[cuMemFreeHost](#), [cuMemGetAddressRange](#), [cuMemGetInfo](#), [cuMemHostAlloc](#),
[cuMemHostGetDevicePointer](#), [cuMemsetD2D8Async](#), [cuMemsetD2D16](#),
[cuMemsetD2D16Async](#), [cuMemsetD2D32](#), [cuMemsetD2D32Async](#), [cuMemsetD8](#),
[cuMemsetD8Async](#), [cuMemsetD16](#), [cuMemsetD16Async](#), [cuMemsetD32](#),
[cuMemsetD32Async](#), [cudaMemset2D](#)

CUresult cuMemsetD2D8Async (CUdeviceptr dstDevice, size_t dstPitch, unsigned char uc, size_t Width, size_t Height, CUstream hStream)

Sets device memory.

Parameters

dstDevice

- Destination device pointer

dstPitch

- Pitch of destination device pointer

uc

- Value to set

Width

- Width of row

Height

- Number of rows

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Sets the 2D memory range of `Width` 8-bit values to the specified value `uc`. `Height` specifies the number of rows to set, and `dstPitch` specifies the number of bytes between each row. This function performs fastest when the pitch is one that has been passed back by `cuMemAllocPitch()`.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).
- ▶ This function uses standard [default stream semantics](#).

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),

[cuMemcpy2D](#), [cuMemcpy2DAsync](#), [cuMemcpy2DUnaligned](#),
[cuMemcpy3D](#), [cuMemcpy3DAsync](#), [cuMemcpyAtoA](#), [cuMemcpyAtoD](#),
[cuMemcpyAtoH](#), [cuMemcpyAtoHAsync](#), [cuMemcpyDtoA](#), [cuMemcpyDtoD](#),
[cuMemcpyDtoDAsync](#), [cuMemcpyDtoH](#), [cuMemcpyDtoHAsync](#), [cuMemcpyHtoA](#),
[cuMemcpyHtoAAsync](#), [cuMemcpyHtoD](#), [cuMemcpyHtoDAsync](#), [cuMemFree](#),
[cuMemFreeHost](#), [cuMemGetAddressRange](#), [cuMemGetInfo](#), [cuMemHostAlloc](#),
[cuMemHostGetDevicePointer](#), [cuMemsetD2D8](#), [cuMemsetD2D16](#),
[cuMemsetD2D16Async](#), [cuMemsetD2D32](#), [cuMemsetD2D32Async](#), [cuMemsetD8](#),
[cuMemsetD8Async](#), [cuMemsetD16](#), [cuMemsetD16Async](#), [cuMemsetD32](#),
[cuMemsetD32Async](#), [cudaMemset2DAsync](#)

CUresult cuMemsetD32 (CUdeviceptr dstDevice, unsigned int ui, size_t N)

Initializes device memory.

Parameters

dstDevice

- Destination device pointer

ui

- Value to set

N

- Number of elements

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Sets the memory range of N 32-bit values to the specified value ui. The dstDevice pointer must be four byte aligned.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).

See also:

[cuArray3DCreate](#), [cuArray3DGetDescriptor](#), [cuArrayCreate](#), [cuArrayDestroy](#),
[cuArrayGetDescriptor](#), [cuMemAlloc](#), [cuMemAllocHost](#), [cuMemAllocPitch](#),

cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned,
 cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD,
 cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD,
 cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA,
 cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree,
 cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc,
 cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async,
 cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async,
 cuMemsetD8, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async,
 cuMemsetD32Async, cudaMemset

CUresult cuMemsetD32Async (CUdeviceptr dstDevice, unsigned int ui, size_t N, CUstream hStream)

Sets device memory.

Parameters

dstDevice

- Destination device pointer

ui

- Value to set

N

- Number of elements

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Sets the memory range of N 32-bit values to the specified value ui. The dstDevice pointer must be four byte aligned.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).
- ▶ This function uses standard [default stream](#) semantics.

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async, cuMemsetD8, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cudaMemsetAsync

CUresult cuMemsetD8 (CUdeviceptr dstDevice, unsigned char uc, size_t N)

Initializes device memory.

Parameters

dstDevice

- Destination device pointer

uc

- Value to set

N

- Number of elements

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Sets the memory range of N 8-bit values to the specified value uc.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async, cuMemsetD8Async, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cuMemsetD32Async, cudaMemset

CUresult cuMemsetD8Async (CUdeviceptr dstDevice, unsigned char uc, size_t N, CUstream hStream)

Sets device memory.

Parameters

dstDevice

- Destination device pointer

uc

- Value to set

N

- Number of elements

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE

Description

Sets the memory range of N 8-bit values to the specified value uc.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ See also [memset synchronization details](#).
- ▶ This function uses standard [default stream semantics](#).

See also:

cuArray3DCreate, cuArray3DGetDescriptor, cuArrayCreate, cuArrayDestroy, cuArrayGetDescriptor, cuMemAlloc, cuMemAllocHost, cuMemAllocPitch, cuMemcpy2D, cuMemcpy2DAsync, cuMemcpy2DUnaligned, cuMemcpy3D, cuMemcpy3DAsync, cuMemcpyAtoA, cuMemcpyAtoD, cuMemcpyAtoH, cuMemcpyAtoHAsync, cuMemcpyDtoA, cuMemcpyDtoD, cuMemcpyDtoDAsync, cuMemcpyDtoH, cuMemcpyDtoHAsync, cuMemcpyHtoA, cuMemcpyHtoAAsync, cuMemcpyHtoD, cuMemcpyHtoDAsync, cuMemFree, cuMemFreeHost, cuMemGetAddressRange, cuMemGetInfo, cuMemHostAlloc, cuMemHostGetDevicePointer, cuMemsetD2D8, cuMemsetD2D8Async, cuMemsetD2D16, cuMemsetD2D16Async, cuMemsetD2D32, cuMemsetD2D32Async, cuMemsetD8, cuMemsetD16, cuMemsetD16Async, cuMemsetD32, cuMemsetD32Async, cudaMemsetAsync

CUresult cuMipmappedArrayCreate (CUmipmappedArray *pHandle, const CUDA_ARRAY3D_DESCRIPTOR *pMipmappedArrayDesc, unsigned int numMipmapLevels)

Creates a CUDA mipmapped array.

Parameters**pHandle**

- Returned mipmapped array

pMipmappedArrayDesc

- mipmapped array descriptor

numMipmapLevels

- Number of mipmap levels

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Creates a CUDA mipmapped array according to the `CUDA_ARRAY3D_DESCRIPTOR` structure `pMipmappedArrayDesc` and returns a handle to the new CUDA mipmapped array in `*pHandle`. `numMipmapLevels` specifies the number of mipmap levels to be allocated. This value is clamped to the range $[1, 1 + \text{floor}(\log_2(\max(\text{width}, \text{height}, \text{depth})))]$.

The `CUDA_ARRAY3D_DESCRIPTOR` is defined as:

```
typedef struct {
    unsigned int Width;
    unsigned int Height;
    unsigned int Depth;
    CUarray_format Format;
    unsigned int NumChannels;
    unsigned int Flags;
} CUDA_ARRAY3D_DESCRIPTOR;
```

where:

- ▶ Width, Height, and Depth are the width, height, and depth of the CUDA array (in elements); the following types of CUDA arrays can be allocated:
 - ▶ A 1D mipmapped array is allocated if Height and Depth extents are both zero.
 - ▶ A 2D mipmapped array is allocated if only Depth extent is zero.
 - ▶ A 3D mipmapped array is allocated if all three extents are non-zero.
 - ▶ A 1D layered CUDA mipmapped array is allocated if only Height is zero and the `CUDA_ARRAY3D_LAYERED` flag is set. Each layer is a 1D array. The number of layers is determined by the depth extent.
 - ▶ A 2D layered CUDA mipmapped array is allocated if all three extents are non-zero and the `CUDA_ARRAY3D_LAYERED` flag is set. Each layer is a 2D array. The number of layers is determined by the depth extent.
 - ▶ A cubemap CUDA mipmapped array is allocated if all three extents are non-zero and the `CUDA_ARRAY3D_CUBEMAP` flag is set. Width must be equal to Height, and Depth must be six. A cubemap is a special type of 2D layered CUDA array, where the six layers represent the six faces of a cube. The order of the six layers in memory is the same as that listed in `CUarray_cubemap_face`.
 - ▶ A cubemap layered CUDA mipmapped array is allocated if all three extents are non-zero, and both, `CUDA_ARRAY3D_CUBEMAP` and `CUDA_ARRAY3D_LAYERED` flags are set. Width must be equal to Height, and Depth must be a multiple of six. A cubemap layered CUDA array is a special type of 2D layered CUDA array that consists of a collection of cubemaps. The first six layers represent the first cubemap, the next six layers form the second cubemap, and so on.
- ▶ Format specifies the format of the elements; `CUarray_format` is defined as:

```
typedef enum CUarray_format_enum {
    CU_AD_FORMAT_UNSIGNED_INT8 = 0x01,
    CU_AD_FORMAT_UNSIGNED_INT16 = 0x02,
    CU_AD_FORMAT_UNSIGNED_INT32 = 0x03,
    CU_AD_FORMAT_SIGNED_INT8 = 0x08,
    CU_AD_FORMAT_SIGNED_INT16 = 0x09,
    CU_AD_FORMAT_SIGNED_INT32 = 0x0a,
    CU_AD_FORMAT_HALF = 0x10,
    CU_AD_FORMAT_FLOAT = 0x20
} CUarray_format;
```

- ▶ NumChannels specifies the number of packed components per CUDA array element; it may be 1, 2, or 4;

- ▶ Flags may be set to
 - ▶ `CUDA_ARRAY3D_LAYERED` to enable creation of layered CUDA mipmapped arrays. If this flag is set, `Depth` specifies the number of layers, not the depth of a 3D array.
 - ▶ `CUDA_ARRAY3D_SURFACE_LDST` to enable surface references to be bound to individual mipmap levels of the CUDA mipmapped array. If this flag is not set, `cuSurfRefSetArray` will fail when attempting to bind a mipmap level of the CUDA mipmapped array to a surface reference.
 - ▶ `CUDA_ARRAY3D_CUBEMAP` to enable creation of mipmapped cubemaps. If this flag is set, `Width` must be equal to `Height`, and `Depth` must be six. If the `CUDA_ARRAY3D_LAYERED` flag is also set, then `Depth` must be a multiple of six.
 - ▶ `CUDA_ARRAY3D_TEXTURE_GATHER` to indicate that the CUDA mipmapped array will be used for texture gather. Texture gather can only be performed on 2D CUDA mipmapped arrays.

`Width`, `Height` and `Depth` must meet certain size requirements as listed in the following table. All values are specified in elements. Note that for brevity's sake, the full name of the device attribute is not specified. For ex., `TEXTURE1D_MIPMAPPED_WIDTH` refers to the device attribute `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_MIPMAPPED_WIDTH`.

CUDA array type	Valid extents that must always be met {(width range in elements), (height range), (depth range)}	Valid extents with <code>CUDA_ARRAY3D_SURFACE_LDST</code> set {(width range in elements), (height range), (depth range)}
1D	{ (1, <code>TEXTURE1D_MIPMAPPED_WIDTH</code>), 0, 0 }	{ (1, <code>SURFACE1D_WIDTH</code>), 0, 0 }
2D	{ (1, <code>TEXTURE2D_MIPMAPPED_WIDTH</code>), (1, <code>TEXTURE2D_MIPMAPPED_HEIGHT</code>), 0 }	{ (1, <code>SURFACE2D_WIDTH</code>), (1, <code>SURFACE2D_HEIGHT</code>), 0 }
3D	{ (1, <code>TEXTURE3D_WIDTH</code>), (1, <code>TEXTURE3D_HEIGHT</code>), (1, <code>TEXTURE3D_DEPTH</code>) } OR { (1, <code>TEXTURE3D_WIDTH_ALTERNATE</code>), (1, <code>TEXTURE3D_HEIGHT_ALTERNATE</code>), (1, <code>TEXTURE3D_DEPTH_ALTERNATE</code>) }	{ (1, <code>SURFACE3D_WIDTH</code>), (1, <code>SURFACE3D_HEIGHT</code>), (1, <code>SURFACE3D_DEPTH</code>) }
1D Layered	{ (1, <code>TEXTURE1D_LAYERED_WIDTH</code>), 0, (1, <code>TEXTURE1D_LAYERED_LAYERS</code>) }	{ (1, <code>SURFACE1D_LAYERED_WIDTH</code>), 0, (1, <code>SURFACE1D_LAYERED_LAYERS</code>) }

2D Layered	{ (1,TEXTURE2D_LAYERED_WIDTH), { (1,SURFACE2D_LAYERED_WIDTH), (1,TEXTURE2D_LAYERED_HEIGHT), (1,SURFACE2D_LAYERED_HEIGHT), (1,TEXTURE2D_LAYERED_LAYERS) } (1,SURFACE2D_LAYERED_LAYERS) }
Cubemap	{ (1,TEXTURECUBEMAP_WIDTH), { (1,SURFACECUBEMAP_WIDTH), (1,TEXTURECUBEMAP_WIDTH), 6 } (1,SURFACECUBEMAP_WIDTH), 6 }
Cubemap Layered	{ (1,TEXTURECUBEMAP_LAYERED_WIDTH), { (1,SURFACECUBEMAP_LAYERED_WIDTH), (1,TEXTURECUBEMAP_LAYERED_WIDTH), (1,SURFACECUBEMAP_LAYERED_WIDTH), (1,TEXTURECUBEMAP_LAYERED_LAYERS) } (1,SURFACECUBEMAP_LAYERED_LAYERS) }



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuMipmappedArrayDestroy](#), [cuMipmappedArrayGetLevel](#), [cuArrayCreate](#),
[cudaMallocMipmappedArray](#)

CUresult cuMipmappedArrayDestroy (CUmipmappedArray hMipmappedArray)

Destroys a CUDA mipmapped array.

Parameters

hMipmappedArray

- Mipmapped array to destroy

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_ARRAY_IS_MAPPED

Description

Destroys the CUDA mipmapped array hMipmappedArray.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuMipmappedArrayCreate`, `cuMipmappedArrayGetLevel`, `cuArrayCreate`,
`cudaFreeMipmappedArray`

CUresult cuMipmappedArrayGetLevel (CUarray *pLevelArray, CUmipmappedArray hMipmappedArray, unsigned int level)

Gets a mipmap level of a CUDA mipmapped array.

Parameters

pLevelArray

- Returned mipmap level CUDA array

hMipmappedArray

- CUDA mipmapped array

level

- Mipmap level

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`

Description

Returns in `*pLevelArray` a CUDA array that represents a single mipmap level of the CUDA mipmapped array `hMipmappedArray`.

If `level` is greater than the maximum number of levels in this mipmapped array, `CUDA_ERROR_INVALID_VALUE` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuMipmappedArrayCreate`, `cuMipmappedArrayDestroy`, `cuArrayCreate`,
`cudaGetMipmappedArrayLevel`

4.12. Unified Addressing

This section describes the unified addressing functions of the low-level CUDA driver application programming interface.

Overview

CUDA devices can share a unified address space with the host. For these devices there is no distinction between a device pointer and a host pointer -- the same pointer value may be used to access memory from the host program and from a kernel running on the device (with exceptions enumerated below).

Supported Platforms

Whether or not a device supports unified addressing may be queried by calling `cuDeviceGetAttribute()` with the device attribute `CU_DEVICE_ATTRIBUTE_UNIFIED_ADDRESSING`.

Unified addressing is automatically enabled in 64-bit processes

Looking Up Information from Pointer Values

It is possible to look up information about the memory which backs a pointer value. For instance, one may want to know if a pointer points to host or device memory. As another example, in the case of device memory, one may want to know on which CUDA device the memory resides. These properties may be queried using the function `cuPointerGetAttribute()`

Since pointers are unique, it is not necessary to specify information about the pointers specified to the various copy functions in the CUDA API. The function `cuMemcpy()` may be used to perform a copy between two pointers, ignoring whether they point to host or device memory (making `cuMemcpyHtoD()`, `cuMemcpyDtoD()`, and `cuMemcpyDtoH()` unnecessary for devices supporting unified addressing). For multidimensional copies, the memory type `CU_MEMORYTYPE_UNIFIED` may be used to specify that the CUDA driver should infer the location of the pointer from its value.

Automatic Mapping of Host Allocated Host Memory

All host memory allocated in all contexts using `cuMemAllocHost()` and `cuMemHostAlloc()` is always directly accessible from all contexts on all devices that support unified addressing. This is the case regardless of whether or not the flags `CU_MEMHOSTALLOC_PORTABLE` and `CU_MEMHOSTALLOC_DEVICEMAP` are specified.

The pointer value through which allocated host memory may be accessed in kernels on all devices that support unified addressing is the same as the pointer value through which that memory is accessed on the host, so it is not necessary to call `cuMemHostGetDevicePointer()` to get the device pointer for these allocations.

Note that this is not the case for memory allocated using the flag `CU_MEMHOSTALLOC_WRITECOMBINED`, as discussed below.

Automatic Registration of Peer Memory

Upon enabling direct access from a context that supports unified addressing to another peer context that supports unified addressing using `cuCtxEnablePeerAccess()` all memory allocated in the peer context using `cuMemAlloc()` and `cuMemAllocPitch()` will immediately be accessible by the current context. The device pointer value through which any peer memory may be accessed in the current context is the same pointer value through which that memory may be accessed in the peer context.

Exceptions, Disjoint Addressing

Not all memory may be accessed on devices through the same pointer value through which they are accessed on the host. These exceptions are host memory registered using `cuMemHostRegister()` and host memory allocated using the flag `CU_MEMHOSTALLOC_WRITECOMBINED`. For these exceptions, there exists a distinct host and device address for the memory. The device address is guaranteed to not overlap any valid host pointer range and is guaranteed to have the same value across all contexts that support unified addressing.

This device address may be queried using `cuMemHostGetDevicePointer()` when a context using unified addressing is current. Either the host or the unified device pointer value may be used to refer to this memory through `cuMemcpy()` and similar functions using the `CU_MEMORYTYPE_UNIFIED` memory type.

CUresult cuMemAdvise (CUdeviceptr devPtr, size_t count, CUmem_advise advice, CUdevice device)

Advise about the usage of a given memory range.

Parameters

devPtr

- Pointer to memory to set the advice for

count

- Size in bytes of the memory range

advice

- Advice to be applied for the specified memory range

device

- Device to apply the advice for

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_INVALID_VALUE`,
`CUDA_ERROR_INVALID_DEVICE`

Description

Advise the Unified Memory subsystem about the usage pattern for the memory range starting at `devPtr` with a size of `count` bytes. The start address and end address of the memory range will be rounded down and rounded up respectively to be aligned to CPU page size before the advice is applied. The memory range must refer to managed memory allocated via `cuMemAllocManaged` or declared via `__managed__` variables. The memory range could also refer to system-allocated pageable memory provided it represents a valid, host-accessible region of memory and all additional constraints imposed by `advice` as outlined below are also satisfied. Specifying an invalid system-allocated pageable memory range results in an error being returned.

The `advice` parameter can take the following values:

- ▶ **CU_MEM_ADVICE_SET_READ_MOSTLY**: This implies that the data is mostly going to be read from and only occasionally written to. Any read accesses from any processor to this region will create a read-only copy of at least the accessed pages in that processor's memory. Additionally, if `cuMemPrefetchAsync` is called on this region, it will create a read-only copy of the data on the destination processor. If any processor writes to this region, all copies of the corresponding page will be invalidated except for the one where the write occurred. The `device` argument is ignored for this advice. Note that for a page to be read-duplicated, the accessing processor must either be the CPU or a GPU that has a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`. Also, if a context is created on a device that does not have the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS` set, then read-duplication will not occur until all such contexts are destroyed. If the memory region refers to valid system-allocated pageable memory, then the accessing device must have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS` for a read-only copy to be created on that device. Note however that if the accessing device also has a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS_USES_HOST_PAGE_TABLES`, then setting this advice will not create a read-only copy when that device accesses this memory region.
- ▶ **CU_MEM_ADVICE_UNSET_READ_MOSTLY**: Undoes the effect of `CU_MEM_ADVICE_SET_READ_MOSTLY` and also prevents the Unified Memory driver from attempting heuristic read-duplication on the memory range. Any read-duplicated copies of the data will be collapsed into a single copy. The location for the collapsed copy will be the preferred location if the page has a preferred location and one of the read-duplicated copies was resident at that location. Otherwise, the location chosen is arbitrary.
- ▶ **CU_MEM_ADVICE_SET_PREFERRED_LOCATION**: This advice sets the preferred location for the data to be the memory belonging to `device`. Passing

in `CU_DEVICE_CPU` for `device` sets the preferred location as host memory. If `device` is a GPU, then it must have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`. Setting the preferred location does not cause data to migrate to that location immediately. Instead, it guides the migration policy when a fault occurs on that memory region. If the data is already in its preferred location and the faulting processor can establish a mapping without requiring the data to be migrated, then data migration will be avoided. On the other hand, if the data is not in its preferred location or if a direct mapping cannot be established, then it will be migrated to the processor accessing it. It is important to note that setting the preferred location does not prevent data prefetching done using `cuMemPrefetchAsync`. Having a preferred location can override the page thrash detection and resolution logic in the Unified Memory driver. Normally, if a page is detected to be constantly thrashing between for example host and device memory, the page may eventually be pinned to host memory by the Unified Memory driver. But if the preferred location is set as device memory, then the page will continue to thrash indefinitely. If `CU_MEM_ADVISE_SET_READ_MOSTLY` is also set on this memory region or any subset of it, then the policies associated with that advice will override the policies of this advice, unless read accesses from `device` will not result in a read-only copy being created on that device as outlined in description for the advice `CU_MEM_ADVISE_SET_READ_MOSTLY`. If the memory region refers to valid system-allocated pageable memory, then `device` must have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS`. Additionally, if `device` has a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS_USES_HOST_PAGE_TABLES`, then this call has no effect. Note however that this behavior may change in the future.

- ▶ `CU_MEM_ADVISE_UNSET_PREFERRED_LOCATION`: Undoes the effect of `CU_MEM_ADVISE_SET_PREFERRED_LOCATION` and changes the preferred location to none.
- ▶ `CU_MEM_ADVISE_SET_ACCESSED_BY`: This advice implies that the data will be accessed by `device`. Passing in `CU_DEVICE_CPU` for `device` will set the advice for the CPU. If `device` is a GPU, then the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS` must be non-zero. This advice does not cause data migration and has no impact on the location of the data per se. Instead, it causes the data to always be mapped in the specified processor's page tables, as long as the location of the data permits a mapping to be established. If the data gets migrated for any reason, the mappings are updated accordingly. This advice is recommended in scenarios where data locality is not important, but avoiding faults is. Consider for example a system containing multiple GPUs with peer-to-peer access enabled, where the data located on one GPU is occasionally accessed by peer GPUs. In such

scenarios, migrating data over to the other GPUs is not as important because the accesses are infrequent and the overhead of migration may be too high. But preventing faults can still help improve performance, and so having a mapping set up in advance is useful. Note that on CPU access of this data, the data may be migrated to host memory because the CPU typically cannot access device memory directly. Any GPU that had the `CU_MEM_ADVISE_SET_ACCESSED_BY` flag set for this data will now have its mapping updated to point to the page in host memory. If `CU_MEM_ADVISE_SET_READ_MOSTLY` is also set on this memory region or any subset of it, then the policies associated with that advice will override the policies of this advice. Additionally, if the preferred location of this memory region or any subset of it is also `device`, then the policies associated with `CU_MEM_ADVISE_SET_PREFERRED_LOCATION` will override the policies of this advice. If the memory region refers to valid system-allocated pageable memory, then `device` must have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS`. Additionally, if `device` has a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS_USES_HOST_PAGE_TABLES`, then this call has no effect.

- ▶ `CU_MEM_ADVISE_UNSET_ACCESSED_BY`: Undoes the effect of `CU_MEM_ADVISE_SET_ACCESSED_BY`. Any mappings to the data from `device` may be removed at any time causing accesses to result in non-fatal page faults. If the memory region refers to valid system-allocated pageable memory, then `device` must have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS`. Additionally, if `device` has a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS_USES_HOST_PAGE_TABLES`, then this call has no effect.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits *asynchronous* behavior for most use cases.
- ▶ This function uses standard *default stream* semantics.

See also:

`cuMemcpy`, `cuMemcpyPeer`, `cuMemcpyAsync`, `cuMemcpy3DPeerAsync`, `cuMemPrefetchAsync`, `cudaMemAdvise`

CUresult cuMemPrefetchAsync (CUdeviceptr devPtr, size_t count, CUdevice dstDevice, CUstream hStream)

Prefetches memory to the specified destination device.

Parameters

devPtr

- Pointer to be prefetched

count

- Size in bytes

dstDevice

- Destination device to prefetch to

hStream

- Stream to enqueue prefetch operation

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_VALUE,
CUDA_ERROR_INVALID_DEVICE

Description

Prefetches memory to the specified destination device. `devPtr` is the base device pointer of the memory to be prefetched and `dstDevice` is the destination device. `count` specifies the number of bytes to copy. `hStream` is the stream in which the operation is enqueued. The memory range must refer to managed memory allocated via [cuMemAllocManaged](#) or declared via `__managed__` variables.

Passing in `CU_DEVICE_CPU` for `dstDevice` will prefetch the data to host memory. If `dstDevice` is a GPU, then the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS` must be non-zero. Additionally, `hStream` must be associated with a device that has a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`.

The start address and end address of the memory range will be rounded down and rounded up respectively to be aligned to CPU page size before the prefetch operation is enqueued in the stream.

If no physical memory has been allocated for this region, then this memory region will be populated and mapped on the destination device. If there's insufficient memory to prefetch the desired region, the Unified Memory driver may evict pages from other [cuMemAllocManaged](#) allocations to host memory in order to make room. Device memory allocated using [cuMemAlloc](#) or [cuArrayCreate](#) will not be evicted.

By default, any mappings to the previous location of the migrated pages are removed and mappings for the new location are only setup on `dstDevice`. The exact behavior

however also depends on the settings applied to this memory range via `cuMemAdvise` as described below:

If `CU_MEM_ADVISE_SET_READ_MOSTLY` was set on any subset of this memory range, then that subset will create a read-only copy of the pages on `dstDevice`.

If `CU_MEM_ADVISE_SET_PREFERRED_LOCATION` was called on any subset of this memory range, then the pages will be migrated to `dstDevice` even if `dstDevice` is not the preferred location of any pages in the memory range.

If `CU_MEM_ADVISE_SET_ACCESSED_BY` was called on any subset of this memory range, then mappings to those pages from all the appropriate processors are updated to refer to the new location if establishing such a mapping is possible. Otherwise, those mappings are cleared.

Note that this API is not required for functionality and only serves to improve performance by allowing the application to migrate data to a suitable location before it is accessed. Memory accesses to this range are always coherent and are allowed even when the data is actively being migrated.

Note that this function is asynchronous with respect to the host and all work on other devices.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits `asynchronous` behavior for most use cases.
- ▶ This function uses standard `default stream` semantics.

See also:

`cuMemcpy`, `cuMemcpyPeer`, `cuMemcpyAsync`, `cuMemcpy3DPeerAsync`, `cuMemAdvise`, `cudaMemPrefetchAsync`

`CUresult cuMemRangeGetAttribute (void *data, size_t dataSize, CUmem_range_attribute attribute, CUdeviceptr devPtr, size_t count)`

Query an attribute of a given memory range.

Parameters

`data`

- A pointers to a memory location where the result of each attribute query will be written to.

dataSize

- Array containing the size of data

attribute

- The attribute to query

devPtr

- Start of the range to query

count

- Size of the range to query

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_VALUE,
CUDA_ERROR_INVALID_DEVICE

Description

Query an attribute about the memory range starting at `devPtr` with a size of `count` bytes. The memory range must refer to managed memory allocated via [cuMemAllocManaged](#) or declared via `__managed__` variables.

The `attribute` parameter can take the following values:

- ▶ **CU_MEM_RANGE_ATTRIBUTE_READ_MOSTLY**: If this attribute is specified, `data` will be interpreted as a 32-bit integer, and `dataSize` must be 4. The result returned will be 1 if all pages in the given memory range have read-duplication enabled, or 0 otherwise.
- ▶ **CU_MEM_RANGE_ATTRIBUTE_PREFERRED_LOCATION**: If this attribute is specified, `data` will be interpreted as a 32-bit integer, and `dataSize` must be 4. The result returned will be a GPU device id if all pages in the memory range have that GPU as their preferred location, or it will be `CU_DEVICE_CPU` if all pages in the memory range have the CPU as their preferred location, or it will be `CU_DEVICE_INVALID` if either all the pages don't have the same preferred location or some of the pages don't have a preferred location at all. Note that the actual location of the pages in the memory range at the time of the query may be different from the preferred location.
- ▶ **CU_MEM_RANGE_ATTRIBUTE_ACCESSED_BY**: If this attribute is specified, `data` will be interpreted as an array of 32-bit integers, and `dataSize` must be a non-zero multiple of 4. The result returned will be a list of device ids that had **CU_MEM_ADVISE_SET_ACCESSED_BY** set for that entire memory range. If any device does not have that advice set for the entire memory range, that device will not be included. If `data` is larger than the number of devices that have that advice set for that memory range, `CU_DEVICE_INVALID` will be returned in all the extra space provided. For ex., if `dataSize` is 12 (i.e. `data` has 3 elements) and only device 0 has the advice set, then the result returned will be `{ 0, CU_DEVICE_INVALID, CU_DEVICE_INVALID }`. If `data` is smaller than the number of devices that have that advice set, then only as many devices will be

returned as can fit in the array. There is no guarantee on which specific devices will be returned, however.

- ▶ **CU_MEM_RANGE_ATTRIBUTE_LAST_PREFETCH_LOCATION**: If this attribute is specified, `data` will be interpreted as a 32-bit integer, and `dataSize` must be 4. The result returned will be the last location to which all pages in the memory range were prefetched explicitly via `cuMemPrefetchAsync`. This will either be a GPU id or `CU_DEVICE_CPU` depending on whether the last location for prefetch was a GPU or the CPU respectively. If any page in the memory range was never explicitly prefetched or if all pages were not prefetched to the same location, `CU_DEVICE_INVALID` will be returned. Note that this simply returns the last location that the application requested to prefetch the memory range to. It gives no indication as to whether the prefetch operation to that location has completed or even begun.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ This function exhibits **asynchronous** behavior for most use cases.
- ▶ This function uses standard **default stream** semantics.

See also:

`cuMemRangeGetAttributes`, `cuMemPrefetchAsync`, `cuMemAdvise`, `cudaMemRangeGetAttribute`

CUresult cuMemRangeGetAttributes (void **data, size_t *dataSizes, CUmem_range_attribute *attributes, size_t numAttributes, CUdeviceptr devPtr, size_t count)

Query attributes of a given memory range.

Parameters

data

- A two-dimensional array containing pointers to memory locations where the result of each attribute query will be written to.

dataSizes

- Array containing the sizes of each result

attributes

- An array of attributes to query (`numAttributes` and the number of attributes in this array should match)

numAttributes

- Number of attributes to query

devPtr

- Start of the range to query

count

- Size of the range to query

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_INVALID_DEVICE

Description

Query attributes of the memory range starting at `devPtr` with a size of `count` bytes. The memory range must refer to managed memory allocated via `cuMemAllocManaged` or declared via `__managed__` variables. The `attributes` array will be interpreted to have `numAttributes` entries. The `dataSizes` array will also be interpreted to have `numAttributes` entries. The results of the query will be stored in `data`.

The list of supported attributes are given below. Please refer to `cuMemRangeGetAttribute` for attribute descriptions and restrictions.

- ▶ CU_MEM_RANGE_ATTRIBUTE_READ_MOSTLY
- ▶ CU_MEM_RANGE_ATTRIBUTE_PREFERRED_LOCATION
- ▶ CU_MEM_RANGE_ATTRIBUTE_ACCESSED_BY
- ▶ CU_MEM_RANGE_ATTRIBUTE_LAST_PREFETCH_LOCATION



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuMemRangeGetAttribute`, `cuMemAdvise` `cuMemPrefetchAsync`,
`cudaMemRangeGetAttributes`

CUresult cuPointerGetAttribute (void *data, CUpointer_attribute attribute, CUdeviceptr ptr)

Returns information about a pointer.

Parameters**data**

- Returned pointer attribute value

attribute

- Pointer attribute to query

ptr

- Pointer

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_DEVICE

Description

The supported attributes are:

► **CU_POINTER_ATTRIBUTE_CONTEXT:**

Returns in `*data` the `CUcontext` in which `ptr` was allocated or registered. The type of `data` must be `CUcontext *`.

If `ptr` was not allocated by, mapped by, or registered with a `CUcontext` which uses unified virtual addressing then `CUDA_ERROR_INVALID_VALUE` is returned.

► **CU_POINTER_ATTRIBUTE_MEMORY_TYPE:**

Returns in `*data` the physical memory type of the memory that `ptr` addresses as a `CUmemorytype` enumerated value. The type of `data` must be unsigned int.

If `ptr` addresses device memory then `*data` is set to `CU_MEMORYTYPE_DEVICE`. The particular `CUdevice` on which the memory resides is the `CUdevice` of the `CUcontext` returned by the `CU_POINTER_ATTRIBUTE_CONTEXT` attribute of `ptr`.

If `ptr` addresses host memory then `*data` is set to `CU_MEMORYTYPE_HOST`.

If `ptr` was not allocated by, mapped by, or registered with a `CUcontext` which uses unified virtual addressing then `CUDA_ERROR_INVALID_VALUE` is returned.

If the current `CUcontext` does not support unified virtual addressing then `CUDA_ERROR_INVALID_CONTEXT` is returned.

► **CU_POINTER_ATTRIBUTE_DEVICE_POINTER:**

Returns in `*data` the device pointer value through which `ptr` may be accessed by kernels running in the current `CUcontext`. The type of `data` must be `CUdeviceptr *`.

If there exists no device pointer value through which kernels running in the current `CUcontext` may access `ptr` then `CUDA_ERROR_INVALID_VALUE` is returned.

If there is no current `CUcontext` then `CUDA_ERROR_INVALID_CONTEXT` is returned.

Except in the exceptional disjoint addressing cases discussed below, the value returned in `*data` will equal the input value `ptr`.

▶ **CU_POINTER_ATTRIBUTE_HOST_POINTER:**

Returns in `*data` the host pointer value through which `ptr` may be accessed by the host program. The type of `data` must be `void**`. If there exists no host pointer value through which the host program may directly access `ptr` then `CUDA_ERROR_INVALID_VALUE` is returned.

Except in the exceptional disjoint addressing cases discussed below, the value returned in `*data` will equal the input value `ptr`.

▶ **CU_POINTER_ATTRIBUTE_P2P_TOKENS:**

Returns in `*data` two tokens for use with the `nv-p2p.h` Linux kernel interface. `data` must be a struct of type `CUDA_POINTER_ATTRIBUTE_P2P_TOKENS`.

`ptr` must be a pointer to memory obtained from `:cuMemAlloc()`. Note that `p2pToken` and `vaSpaceToken` are only valid for the lifetime of the source allocation. A subsequent allocation at the same address may return completely different tokens. Querying this attribute has a side effect of setting the attribute `CU_POINTER_ATTRIBUTE_SYNC_MEMOPS` for the region of memory that `ptr` points to.

▶ **CU_POINTER_ATTRIBUTE_SYNC_MEMOPS:**

A boolean attribute which when set, ensures that synchronous memory operations initiated on the region of memory that `ptr` points to will always synchronize. See further documentation in the section titled "API synchronization behavior" to learn more about cases when synchronous memory operations can exhibit asynchronous behavior.

▶ **CU_POINTER_ATTRIBUTE_BUFFER_ID:**

Returns in `*data` a buffer ID which is guaranteed to be unique within the process. `data` must point to an unsigned long long.

`ptr` must be a pointer to memory obtained from a CUDA memory allocation API. Every memory allocation from any of the CUDA memory allocation APIs will have a unique ID over a process lifetime. Subsequent allocations do not reuse IDs from previous freed allocations. IDs are only unique within a single process.

▶ **CU_POINTER_ATTRIBUTE_IS_MANAGED:**

Returns in `*data` a boolean that indicates whether the pointer points to managed memory or not.

▶ **CU_POINTER_ATTRIBUTE_DEVICE_ORDINAL:**

Returns in `*data` an integer representing a device ordinal of a device against which the memory was allocated or registered.

Note that for most allocations in the unified virtual address space the host and device pointer for accessing the allocation will be the same. The exceptions to this are

- ▶ user memory registered using `cuMemHostRegister`
- ▶ host memory allocated using `cuMemHostAlloc` with the `CU_MEMHOSTALLOC_WRITECOMBINED` flag For these types of allocation there will exist separate, disjoint host and device addresses for accessing the allocation. In particular
 - ▶ The host address will correspond to an invalid unmapped device address (which will result in an exception if accessed from the device)
 - ▶ The device address will correspond to an invalid unmapped host address (which will result in an exception if accessed from the host). For these types of allocations, querying `CU_POINTER_ATTRIBUTE_HOST_POINTER` and `CU_POINTER_ATTRIBUTE_DEVICE_POINTER` may be used to retrieve the host and device addresses from either address.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuPointerSetAttribute`, `cuMemAlloc`, `cuMemFree`, `cuMemAllocHost`, `cuMemFreeHost`, `cuMemHostAlloc`, `cuMemHostRegister`, `cuMemHostUnregister`, `cudaPointerGetAttributes`

`CUresult cuPointerGetAttributes (unsigned int numAttributes, CUpointer_attribute *attributes, void **data, CUdeviceptr ptr)`

Returns information about a pointer.

Parameters

numAttributes

- Number of attributes to query

attributes

- An array of attributes to query (numAttributes and the number of attributes in this array should match)

data

- A two-dimensional array containing pointers to memory locations where the result of each attribute query will be written to.

ptr

- Pointer to query

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_INVALID_DEVICE

Description

The supported attributes are (refer to `cuPointerGetAttribute` for attribute descriptions and restrictions):

- ▶ CU_POINTER_ATTRIBUTE_CONTEXT
- ▶ CU_POINTER_ATTRIBUTE_MEMORY_TYPE
- ▶ CU_POINTER_ATTRIBUTE_DEVICE_POINTER
- ▶ CU_POINTER_ATTRIBUTE_HOST_POINTER
- ▶ CU_POINTER_ATTRIBUTE_SYNC_MEMOPS
- ▶ CU_POINTER_ATTRIBUTE_BUFFER_ID
- ▶ CU_POINTER_ATTRIBUTE_IS_MANAGED
- ▶ CU_POINTER_ATTRIBUTE_DEVICE_ORDINAL

Unlike `cuPointerGetAttribute`, this function will not return an error when the `ptr` encountered is not a valid CUDA pointer. Instead, the attributes are assigned default NULL values and CUDA_SUCCESS is returned.

If `ptr` was not allocated by, mapped by, or registered with a `CUcontext` which uses UVA (Unified Virtual Addressing), `CUDA_ERROR_INVALID_CONTEXT` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuPointerGetAttribute`, `cuPointerSetAttribute`, `cudaPointerGetAttributes`

CUresult cuPointerSetAttribute (const void *value, CUpointer_attribute attribute, CUdeviceptr ptr)

Set attributes on a previously allocated memory region.

Parameters

value

- Pointer to memory containing the value to be set

attribute

- Pointer attribute to set

ptr

- Pointer to a memory region allocated using CUDA memory allocation APIs

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_DEVICE

Description

The supported attributes are:

► **CU_POINTER_ATTRIBUTE_SYNC_MEMOPS:**

A boolean attribute that can either be set (1) or unset (0). When set, the region of memory that `ptr` points to is guaranteed to always synchronize memory operations that are synchronous. If there are some previously initiated synchronous memory operations that are pending when this attribute is set, the function does not return until those memory operations are complete. See further documentation in the section titled "API synchronization behavior" to learn more about cases when synchronous memory operations can exhibit asynchronous behavior. `value` will be considered as a pointer to an unsigned integer to which this attribute is to be set.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuPointerGetAttribute](#), [cuPointerGetAttributes](#), [cuMemAlloc](#), [cuMemFree](#),
[cuMemAllocHost](#), [cuMemFreeHost](#), [cuMemHostAlloc](#), [cuMemHostRegister](#),
[cuMemHostUnregister](#)

4.13. Stream Management

This section describes the stream management functions of the low-level CUDA driver application programming interface.

CUresult cuStreamAddCallback (CUstream hStream, CUstreamCallback callback, void *userData, unsigned int flags)

Add a callback to a compute stream.

Parameters

hStream

- Stream to add callback to

callback

- The function to call once preceding stream operations are complete

userData

- User specified data to be passed to the callback function

flags

- Reserved for future use, must be 0

Returns

CU_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_NOT_SUPPORTED

Description

Adds a callback to be called on the host after all currently enqueued items in the stream have completed. For each cuStreamAddCallback call, the callback will be executed exactly once. The callback will block later work in the stream until it is finished.

The callback may be passed [CUDA_SUCCESS](#) or an error code. In the event of a device error, all subsequently executed callbacks will receive an appropriate [CUresult](#).

Callbacks must not make any CUDA API calls. Attempting to use a CUDA API will result in [CUDA_ERROR_NOT_PERMITTED](#). Callbacks must not perform any synchronization that may depend on outstanding device work or other callbacks that are not mandated to run earlier. Callbacks without a mandated order (in independent streams) execute in undefined order and may be serialized.

For the purposes of Unified Memory, callback execution makes a number of guarantees:

- ▶ The callback stream is considered idle for the duration of the callback. Thus, for example, a callback may always use memory attached to the callback stream.
- ▶ The start of execution of a callback has the same effect as synchronizing an event recorded in the same stream immediately prior to the callback. It thus synchronizes streams which have been "joined" prior to the callback.

- ▶ Adding device work to any stream does not have the effect of making the stream active until all preceding callbacks have executed. Thus, for example, a callback might use global attached memory even if work has been added to another stream, if it has been properly ordered with an event.
- ▶ Completion of a callback does not cause a stream to become active except as described above. The callback stream will remain idle if no device work follows the callback, and will remain idle across consecutive callbacks without device work in between. Thus, for example, stream synchronization can be done by signaling from a callback at the end of the stream.



- ▶ This function uses standard **default stream** semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamCreate](#), [cuStreamQuery](#), [cuStreamSynchronize](#), [cuStreamWaitEvent](#), [cuStreamDestroy](#), [cuMemAllocManaged](#), [cuStreamAttachMemAsync](#), [cudaStreamAddCallback](#)

CUresult cuStreamAttachMemAsync (CUstream hStream, CUdeviceptr dptr, size_t length, unsigned int flags)

Attach memory to a stream asynchronously.

Parameters

hStream

- Stream in which to enqueue the attach operation

dptr

- Pointer to memory (must be a pointer to managed memory or to a valid host-accessible region of system-allocated pageable memory)

length

- Length of memory

flags

- Must be one of [CUmemAttach_flags](#)

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#), [CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#), [CUDA_ERROR_INVALID_HANDLE](#), [CUDA_ERROR_NOT_SUPPORTED](#)

Description

Enqueues an operation in `hStream` to specify stream association of `length` bytes of memory starting from `dptr`. This function is a stream-ordered operation, meaning that it is dependent on, and will only take effect when, previous work in stream has completed. Any previous association is automatically replaced.

`dptr` must point to one of the following types of memories:

- ▶ managed memory declared using the `__managed__` keyword or allocated with `cuMemAllocManaged`.
- ▶ a valid host-accessible region of system-allocated pageable memory. This type of memory may only be specified if the device associated with the stream reports a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_PAGEABLE_MEMORY_ACCESS`.

For managed allocations, `length` must be either zero or the entire allocation's size. Both indicate that the entire allocation's stream association is being changed. Currently, it is not possible to change stream association for a portion of a managed allocation.

For pageable host allocations, `length` must be non-zero.

The stream association is specified using `flags` which must be one of `CUmemAttach_flags`. If the `CU_MEM_ATTACH_GLOBAL` flag is specified, the memory can be accessed by any stream on any device. If the `CU_MEM_ATTACH_HOST` flag is specified, the program makes a guarantee that it won't access the memory on the device from any stream on a device that has a zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`.

If the `CU_MEM_ATTACH_SINGLE` flag is specified and `hStream` is associated with a device that has a zero value for the device attribute `CU_DEVICE_ATTRIBUTE_CONCURRENT_MANAGED_ACCESS`, the program makes a guarantee that it will only access the memory on the device from `hStream`. It is illegal to attach singly to the NULL stream, because the NULL stream is a virtual global stream and not a specific stream. An error will be returned in this case.

When memory is associated with a single stream, the Unified Memory system will allow CPU access to this memory region so long as all operations in `hStream` have completed, regardless of whether other streams are active. In effect, this constrains exclusive ownership of the managed memory region by an active GPU to per-stream activity instead of whole-GPU activity.

Accessing memory on the device from streams that are not associated with it will produce undefined results. No error checking is performed by the Unified Memory system to ensure that kernels launched into other streams do not access this region.

It is a program's responsibility to order calls to `cuStreamAttachMemAsync` via events, synchronization or other means to ensure legal access to memory at all times. Data

visibility and coherency will be changed appropriately for all kernels which follow a stream-association change.

If `hStream` is destroyed while data is associated with it, the association is removed and the association reverts to the default visibility of the allocation as specified at `cuMemAllocManaged`. For `__managed__` variables, the default association is always `CU_MEM_ATTACH_GLOBAL`. Note that destroying a stream is an asynchronous operation, and as a result, the change to default association won't happen until all work in the stream has completed.



- ▶ This function uses standard `default stream` semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuStreamCreate`, `cuStreamQuery`, `cuStreamSynchronize`, `cuStreamWaitEvent`, `cuStreamDestroy`, `cuMemAllocManaged`, `cudaStreamAttachMemAsync`

CUresult cuStreamCreate (CUstream *phStream, unsigned int Flags)

Create a stream.

Parameters

phStream

- Returned newly created stream

Flags

- Parameters for stream creation

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`, `CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`, `CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_OUT_OF_MEMORY`

Description

Creates a stream and returns a handle in `phStream`. The `Flags` argument determines behaviors of the stream. Valid values for `Flags` are:

- ▶ `CU_STREAM_DEFAULT`: Default stream creation flag.

- ▶ **CU_STREAM_NON_BLOCKING**: Specifies that work running in the created stream may run concurrently with work in stream 0 (the NULL stream), and that the created stream should perform no implicit synchronization with stream 0.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamDestroy](#), [cuStreamCreateWithPriority](#), [cuStreamGetPriority](#), [cuStreamGetFlags](#), [cuStreamWaitEvent](#), [cuStreamQuery](#), [cuStreamSynchronize](#), [cuStreamAddCallback](#), [cudaStreamCreate](#), [cudaStreamCreateWithFlags](#)

CUresult cuStreamCreateWithPriority (CUstream *phStream, unsigned int flags, int priority)

Create a stream with the given priority.

Parameters

phStream

- Returned newly created stream

flags

- Flags for stream creation. See [cuStreamCreate](#) for a list of valid flags

priority

- Stream priority. Lower numbers represent higher priorities. See [cuCtxGetStreamPriorityRange](#) for more information about meaningful stream priorities that can be passed.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#), [CUDA_ERROR_OUT_OF_MEMORY](#)

Description

Creates a stream with the specified priority and returns a handle in `phStream`. This API alters the scheduler priority of work in the stream. Work in a higher priority stream may preempt work already executing in a low priority stream.

`priority` follows a convention where lower numbers represent higher priorities. '0' represents default priority. The range of meaningful numerical priorities can be queried using [cuCtxGetStreamPriorityRange](#). If the specified priority is outside the numerical

range returned by `cuCtxGetStreamPriorityRange`, it will automatically be clamped to the lowest or the highest number in the range.



- ▶ Note that this function may also return error codes from previous, asynchronous launches.
- ▶ Stream priorities are supported only on GPUs with compute capability 3.5 or higher.
- ▶ In the current implementation, only compute kernels launched in priority streams are affected by the stream's priority. Stream priorities have no effect on host-to-device and device-to-host memory operations.

See also:

`cuStreamDestroy`, `cuStreamCreate`, `cuStreamGetPriority`, `cuCtxGetStreamPriorityRange`, `cuStreamGetFlags`, `cuStreamWaitEvent`, `cuStreamQuery`, `cuStreamSynchronize`, `cuStreamAddCallback`, `cudaStreamCreateWithPriority`

CUresult cuStreamDestroy (CUstream hStream)

Destroys a stream.

Parameters

hStream

- Stream to destroy

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Destroys the stream specified by `hStream`.

In case the device is still doing work in the stream `hStream` when `cuStreamDestroy()` is called, the function will return immediately and the resources associated with `hStream` will be released automatically once the device has completed all work in `hStream`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuStreamCreate`, `cuStreamWaitEvent`, `cuStreamQuery`, `cuStreamSynchronize`,
`cuStreamAddCallback`, `cudaStreamDestroy`

CUresult cuStreamGetCtx (CUstream hStream, CUcontext *pctx)

Query the context associated with a stream.

Parameters

hStream

- Handle to the stream to be queried

pctx

- Returned context associated with the stream

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_HANDLE`,

Description

Returns the CUDA context that the stream is associated with.

The stream handle `hStream` can refer to any of the following:

- ▶ a stream created via any of the CUDA driver APIs such as `cuStreamCreate` and `cuStreamCreateWithPriority`, or their runtime API equivalents such as `cudaStreamCreate`, `cudaStreamCreateWithFlags` and `cudaStreamCreateWithPriority`. The returned context is the context that was active in the calling thread when the stream was created. Passing an invalid handle will result in undefined behavior.
- ▶ any of the special streams such as the NULL stream, `CU_STREAM_LEGACY` and `CU_STREAM_PER_THREAD`. The runtime API equivalents of these are also accepted, which are `NULL`, `cudaStreamLegacy` and `cudaStreamPerThread` respectively. Specifying any of the special handles will return the context current to the calling thread. If no context is current to the calling thread, `CUDA_ERROR_INVALID_CONTEXT` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamDestroy](#), [cuStreamCreateWithPriority](#), [cuStreamGetPriority](#), [cuStreamGetFlags](#), [cuStreamWaitEvent](#), [cuStreamQuery](#), [cuStreamSynchronize](#), [cuStreamAddCallback](#), [cudaStreamCreate](#), [cudaStreamCreateWithFlags](#)

CUresult cuStreamGetFlags (CUstream hStream, unsigned int *flags)

Query the flags of a given stream.

Parameters

hStream

- Handle to the stream to be queried

flags

- Pointer to an unsigned integer in which the stream's flags are returned. The value returned in `flags` is a logical 'OR' of all flags that were used while creating this stream. See [cuStreamCreate](#) for the list of valid flags.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#), [CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#), [CUDA_ERROR_INVALID_VALUE](#), [CUDA_ERROR_INVALID_HANDLE](#), [CUDA_ERROR_OUT_OF_MEMORY](#)

Description

Query the flags of a stream created using [cuStreamCreate](#) or [cuStreamCreateWithPriority](#) and return the flags in `flags`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamDestroy](#), [cuStreamCreate](#), [cuStreamGetPriority](#), [cudaStreamGetFlags](#)

CUresult cuStreamGetPriority (CUstream hStream, int *priority)

Query the priority of a given stream.

Parameters

hStream

- Handle to the stream to be queried

priority

- Pointer to a signed integer in which the stream's priority is returned

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE,
 CUDA_ERROR_OUT_OF_MEMORY

Description

Query the priority of a stream created using [cuStreamCreate](#) or [cuStreamCreateWithPriority](#) and return the priority in `priority`. Note that if the stream was created with a priority outside the numerical range returned by [cuCtxGetStreamPriorityRange](#), this function returns the clamped priority. See [cuStreamCreateWithPriority](#) for details about priority clamping.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamDestroy](#), [cuStreamCreate](#), [cuStreamCreateWithPriority](#),
[cuCtxGetStreamPriorityRange](#), [cuStreamGetFlags](#), [cudaStreamGetPriority](#)

CUresult cuStreamQuery (CUstream hStream)

Determine status of a compute stream.

Parameters

hStream

- Stream to query status of

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_NOT_READY

Description

Returns `CUDA_SUCCESS` if all operations in the stream specified by `hStream` have completed, or `CUDA_ERROR_NOT_READY` if not.

For the purposes of Unified Memory, a return value of `CUDA_SUCCESS` is equivalent to having called `cuStreamSynchronize()`.



- ▶ This function uses standard `default stream` semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuStreamCreate`, `cuStreamWaitEvent`, `cuStreamDestroy`, `cuStreamSynchronize`,
`cuStreamAddCallback`, `cudaStreamQuery`

CUresult cuStreamSynchronize (CUstream hStream)

Wait until a stream's tasks are completed.

Parameters

hStream

- Stream to wait for

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE

Description

Waits until the device has completed all operations in the stream specified by `hStream`. If the context was created with the `CU_CTX_SCHED_BLOCKING_SYNC` flag, the CPU thread will block until the stream is finished with all of its tasks.



- ▶ This function uses standard `default stream` semantics.

- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamCreate](#), [cuStreamDestroy](#), [cuStreamWaitEvent](#), [cuStreamQuery](#),
[cuStreamAddCallback](#), [cudaStreamSynchronize](#)

CUresult cuStreamWaitEvent (CUstream hStream, CUevent hEvent, unsigned int Flags)

Make a compute stream wait on an event.

Parameters

hStream

- Stream to wait

hEvent

- Event to wait on (may not be NULL)

Flags

- Parameters for the operation (must be 0)

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_HANDLE](#),

Description

Makes all future work submitted to `hStream` wait for all work captured in `hEvent`. See [cuEventRecord\(\)](#) for details on what is captured by an event. The synchronization will be performed efficiently on the device when applicable. `hEvent` may be from a different context or device than `hStream`.



- ▶ This function uses standard [default stream](#) semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamCreate](#), [cuEventRecord](#), [cuStreamQuery](#), [cuStreamSynchronize](#),
[cuStreamAddCallback](#), [cuStreamDestroy](#), [cudaStreamWaitEvent](#)

4.14. Event Management

This section describes the event management functions of the low-level CUDA driver application programming interface.

CUresult cuEventCreate (CUevent *phEvent, unsigned int Flags)

Creates an event.

Parameters

phEvent

- Returns newly created event

Flags

- Event creation flags

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY

Description

Creates an event *phEvent for the current context with the flags specified via Flags. Valid flags include:

- ▶ **CU_EVENT_DEFAULT**: Default event creation flag.
- ▶ **CU_EVENT_BLOCKING_SYNC**: Specifies that the created event should use blocking synchronization. A CPU thread that uses `cuEventSynchronize()` to wait on an event created with this flag will block until the event has actually been recorded.
- ▶ **CU_EVENT_DISABLE_TIMING**: Specifies that the created event does not need to record timing data. Events created with this flag specified and the **CU_EVENT_BLOCKING_SYNC** flag not specified will provide the best performance when used with `cuStreamWaitEvent()` and `cuEventQuery()`.
- ▶ **CU_EVENT_INTERPROCESS**: Specifies that the created event may be used as an interprocess event by `cuIpcGetEventHandle()`. **CU_EVENT_INTERPROCESS** must be specified along with **CU_EVENT_DISABLE_TIMING**.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuEventRecord](#), [cuEventQuery](#), [cuEventSynchronize](#), [cuEventDestroy](#),
[cuEventElapsedTime](#), [cudaEventCreate](#), [cudaEventCreateWithFlags](#)

CUresult cuEventDestroy (CUevent hEvent)

Destroys an event.

Parameters**hEvent**

- Event to destroy

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_HANDLE](#)

Description

Destroys the event specified by `hEvent`.

An event may be destroyed before it is complete (i.e., while `cuEventQuery()` would return [CUDA_ERROR_NOT_READY](#)). In this case, the call does not block on completion of the event, and any associated resources will automatically be released asynchronously at completion.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuEventCreate](#), [cuEventRecord](#), [cuEventQuery](#), [cuEventSynchronize](#),
[cuEventElapsedTime](#), [cudaEventDestroy](#)

CUresult cuEventElapsedTime (float *pMilliseconds, CUevent hStart, CUevent hEnd)

Computes the elapsed time between two events.

Parameters**pMilliseconds**

- Time between `hStart` and `hEnd` in ms

hStart

- Starting event

hEnd

- Ending event

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_NOT_READY

Description

Computes the elapsed time between two events (in milliseconds with a resolution of around 0.5 microseconds).

If either event was last recorded in a non-NULL stream, the resulting time may be greater than expected (even if both used the same stream handle). This happens because the `cuEventRecord()` operation takes place asynchronously and there is no guarantee that the measured latency is actually just between the two events. Any number of other different stream operations could execute in between the two measured events, thus altering the timing in a significant way.

If `cuEventRecord()` has not been called on either event then `CUDA_ERROR_INVALID_HANDLE` is returned. If `cuEventRecord()` has been called on both events but one or both of them has not yet been completed (that is, `cuEventQuery()` would return `CUDA_ERROR_NOT_READY` on at least one of the events), `CUDA_ERROR_NOT_READY` is returned. If either event was created with the `CU_EVENT_DISABLE_TIMING` flag, then this function will return `CUDA_ERROR_INVALID_HANDLE`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuEventCreate`, `cuEventRecord`, `cuEventQuery`, `cuEventSynchronize`, `cuEventDestroy`,
`cudaEventElapsedTime`

CUresult cuEventQuery (CUevent hEvent)

Queries an event's status.

Parameters

hEvent

- Event to query

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_NOT_READY

Description

Queries the status of all work currently captured by hEvent. See cuEventRecord() for details on what is captured by an event.

Returns CUDA_SUCCESS if all captured work has been completed, or CUDA_ERROR_NOT_READY if any captured work is incomplete.

For the purposes of Unified Memory, a return value of CUDA_SUCCESS is equivalent to having called cuEventSynchronize().



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuEventCreate, cuEventRecord, cuEventSynchronize, cuEventDestroy,
cuEventElapsedTime, cudaEventQuery

CUresult cuEventRecord (CUevent hEvent, CUstream hStream)

Records an event.

Parameters

hEvent

- Event to record

hStream

- Stream to record event for

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_INVALID_VALUE

Description

Captures in `hEvent` the contents of `hStream` at the time of this call. `hEvent` and `hStream` must be from the same context. Calls such as `cuEventQuery()` or `cuStreamWaitEvent()` will then examine or wait for completion of the work that was captured. Uses of `hStream` after this call do not modify `hEvent`. See note on default stream behavior for what is captured in the default case.

`cuEventRecord()` can be called multiple times on the same event and will overwrite the previously captured state. Other APIs such as `cuStreamWaitEvent()` use the most recently captured state at the time of the API call, and are not affected by later calls to `cuEventRecord()`. Before the first call to `cuEventRecord()`, an event represents an empty set of work, so for example `cuEventQuery()` would return `CUDA_SUCCESS`.



- ▶ This function uses standard [default stream](#) semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuEventCreate`, `cuEventQuery`, `cuEventSynchronize`, `cuStreamWaitEvent`,
`cuEventDestroy`, `cuEventElapsedTime`, `cudaEventRecord`

CUresult cuEventSynchronize (CUevent hEvent)

Waits for an event to complete.

Parameters**hEvent**

- Event to wait for

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE

Description

Waits until the completion of all work currently captured in `hEvent`. See [cuEventRecord\(\)](#) for details on what is captured by an event.

Waiting for an event that was created with the `CU_EVENT_BLOCKING_SYNC` flag will cause the calling CPU thread to block until the event has been completed by the device. If the `CU_EVENT_BLOCKING_SYNC` flag has not been set, then the CPU thread will busy-wait until the event has been completed by the device.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuEventCreate](#), [cuEventRecord](#), [cuEventQuery](#), [cuEventDestroy](#), [cuEventElapsedTime](#), [cudaEventSynchronize](#)

4.15. Stream memory operations

This section describes the stream memory operations of the low-level CUDA driver application programming interface.

The whole set of operations is disabled by default. Users are required to explicitly enable them, e.g. on Linux by passing the kernel module parameter shown below: `modprobe nvidia NVreg_EnableStreamMemOps=1` There is currently no way to enable these operations on other operating systems.

Users can programmatically query whether the device supports these operations with [cuDeviceGetAttribute\(\)](#) and `CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_MEM_OPS`.

Support for the `CU_STREAM_WAIT_VALUE_NOR` flag can be queried with `CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_WAIT_VALUE_NOR`.

Support for the [cuStreamWriteValue64\(\)](#) and [cuStreamWaitValue64\(\)](#) functions, as well as for the `CU_STREAM_MEM_OP_WAIT_VALUE_64` and `CU_STREAM_MEM_OP_WRITE_VALUE_64` flags, can be queried with `CU_DEVICE_ATTRIBUTE_CAN_USE_64_BIT_STREAM_MEM_OPS`.

Support for both `CU_STREAM_WAIT_VALUE_FLUSH` and `CU_STREAM_MEM_OP_FLUSH_REMOTE_WRITES` requires dedicated platform hardware features and can be queried with [cuDeviceGetAttribute\(\)](#) and `CU_DEVICE_ATTRIBUTE_CAN_FLUSH_REMOTE_WRITES`.

Note that all memory pointers passed as parameters to these operations are device pointers. Where necessary a device pointer should be obtained, for example with `cuMemHostGetDevicePointer()`.

None of the operations accepts pointers to managed memory buffers (`cuMemAllocManaged`).

CUresult cuStreamBatchMemOp (CUstream stream, unsigned int count, CUstreamBatchMemOpParams *paramArray, unsigned int flags)

Batch operations to synchronize the stream via memory operations.

Parameters

stream

The stream to enqueue the operations in.

count

The number of operations in the array. Must be less than 256.

paramArray

The types and parameters of the individual operations.

flags

Reserved for future expansion; must be 0.

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_INVALID_VALUE`,
`CUDA_ERROR_NOT_SUPPORTED`

Description

This is a batch version of `cuStreamWaitValue32()` and `cuStreamWriteValue32()`. Batching operations may avoid some performance overhead in both the API call and the device execution versus adding them to the stream in separate API calls. The operations are enqueued in the order they appear in the array.

See `CUstreamBatchMemOpType` for the full set of supported operations, and `cuStreamWaitValue32()`, `cuStreamWaitValue64()`, `cuStreamWriteValue32()`, and `cuStreamWriteValue64()` for details of specific operations.

Basic support for this can be queried with `cuDeviceGetAttribute()` and `CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_MEM_OPS`. See related APIs for details on querying support for specific operations.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamWaitValue32](#), [cuStreamWaitValue64](#), [cuStreamWriteValue32](#), [cuStreamWriteValue64](#), [cuMemHostRegister](#)

CUresult cuStreamWaitValue32 (CUstream stream, CUdeviceptr addr, cuuint32_t value, unsigned int flags)

Wait on a memory location.

Parameters

stream

The stream to synchronize on the memory location.

addr

The memory location to wait on.

value

The value to compare with the memory location.

flags

See [CUstreamWaitValue_flags](#).

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_VALUE](#), [CUDA_ERROR_NOT_SUPPORTED](#)

Description

Enqueues a synchronization of the stream on the given memory location. Work ordered after the operation will block until the given condition on the memory is satisfied. By default, the condition is to wait for $(\text{int32_t})(*\text{addr} - \text{value}) \geq 0$, a cyclic greater-or-equal. Other condition types can be specified via `flags`.

If the memory was registered via [cuMemHostRegister\(\)](#), the device pointer should be obtained with [cuMemHostGetDevicePointer\(\)](#). This function cannot be used with managed memory ([cuMemAllocManaged](#)).

Support for this can be queried with [cuDeviceGetAttribute\(\)](#) and [CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_MEM_OPS](#).

Support for [CU_STREAM_WAIT_VALUE_NOR](#) can be queried with [cuDeviceGetAttribute\(\)](#) and [CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_WAIT_VALUE_NOR](#).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamWaitValue64](#), [cuStreamWriteValue32](#), [cuStreamWriteValue64](#)
[cuStreamBatchMemOp](#), [cuMemHostRegister](#), [cuStreamWaitEvent](#)

CUresult cuStreamWaitValue64 (CUstream stream, CUdeviceptr addr, cuuint64_t value, unsigned int flags)

Wait on a memory location.

Parameters

stream

The stream to synchronize on the memory location.

addr

The memory location to wait on.

value

The value to compare with the memory location.

flags

See [CUstreamWaitValue_flags](#).

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_VALUE](#),
[CUDA_ERROR_NOT_SUPPORTED](#)

Description

Enqueues a synchronization of the stream on the given memory location. Work ordered after the operation will block until the given condition on the memory is satisfied. By default, the condition is to wait for $(\text{int64}_t)(*addr - \text{value}) \geq 0$, a cyclic greater-or-equal. Other condition types can be specified via `flags`.

If the memory was registered via [cuMemHostRegister\(\)](#), the device pointer should be obtained with [cuMemHostGetDevicePointer\(\)](#).

Support for this can be queried with [cuDeviceGetAttribute\(\)](#) and [CU_DEVICE_ATTRIBUTE_CAN_USE_64_BIT_STREAM_MEM_OPS](#).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamWaitValue32](#), [cuStreamWriteValue32](#), [cuStreamWriteValue64](#),
[cuStreamBatchMemOp](#), [cuMemHostRegister](#), [cuStreamWaitEvent](#)

CUresult cuStreamWriteValue32 (CUstream stream, CUdeviceptr addr, cuuint32_t value, unsigned int flags)

Write a value to memory.

Parameters

stream

The stream to do the write in.

addr

The device address to write to.

value

The value to write.

flags

See [CUstreamWriteValue_flags](#).

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_VALUE](#),
[CUDA_ERROR_NOT_SUPPORTED](#)

Description

Write a value to memory. Unless the [CU_STREAM_WRITE_VALUE_NO_MEMORY_BARRIER](#) flag is passed, the write is preceded by a system-wide memory fence, equivalent to a `__threadfence_system()` but scoped to the stream rather than a CUDA thread.

If the memory was registered via [cuMemHostRegister\(\)](#), the device pointer should be obtained with [cuMemHostGetDevicePointer\(\)](#). This function cannot be used with managed memory ([cuMemAllocManaged](#)).

Support for this can be queried with [cuDeviceGetAttribute\(\)](#) and [CU_DEVICE_ATTRIBUTE_CAN_USE_STREAM_MEM_OPS](#).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuStreamWriteValue64](#), [cuStreamWaitValue32](#), [cuStreamWaitValue64](#), [cuStreamBatchMemOp](#), [cuMemHostRegister](#), [cuEventRecord](#)

CUresult cuStreamWriteValue64 (CUstream stream, CUdeviceptr addr, cuuint64_t value, unsigned int flags)

Write a value to memory.

Parameters

stream

The stream to do the write in.

addr

The device address to write to.

value

The value to write.

flags

See [CUstreamWriteValue_flags](#).

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_VALUE](#), [CUDA_ERROR_NOT_SUPPORTED](#)

Description

Write a value to memory. Unless the [CU_STREAM_WRITE_VALUE_NO_MEMORY_BARRIER](#) flag is passed, the write is preceded by a system-wide memory fence, equivalent to a `__threadfence_system()` but scoped to the stream rather than a CUDA thread.

If the memory was registered via [cuMemHostRegister\(\)](#), the device pointer should be obtained with [cuMemHostGetDevicePointer\(\)](#).

Support for this can be queried with [cuDeviceGetAttribute\(\)](#) and [CU_DEVICE_ATTRIBUTE_CAN_USE_64_BIT_STREAM_MEM_OPS](#).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuStreamWriteValue32`, `cuStreamWaitValue32`, `cuStreamWaitValue64`,
`cuStreamBatchMemOp`, `cuMemHostRegister`, `cuEventRecord`

4.16. Execution Control

This section describes the execution control functions of the low-level CUDA driver application programming interface.

CUresult cuFuncGetAttribute (int *pi, CUfunction_attribute attrib, CUfunction hfunc)

Returns information about a function.

Parameters**pi**

- Returned attribute value

attrib

- Attribute requested

hfunc

- Function to query attribute of

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_HANDLE`, `CUDA_ERROR_INVALID_VALUE`

Description

Returns in `*pi` the integer value of the attribute `attrib` on the kernel given by `hfunc`. The supported attributes are:

- ▶ `CU_FUNC_ATTRIBUTE_MAX_THREADS_PER_BLOCK`: The maximum number of threads per block, beyond which a launch of the function would fail. This number depends on both the function and the device on which the function is currently loaded.
- ▶ `CU_FUNC_ATTRIBUTE_SHARED_SIZE_BYTES`: The size in bytes of statically-allocated shared memory per block required by this function. This does not include dynamically-allocated shared memory requested by the user at runtime.
- ▶ `CU_FUNC_ATTRIBUTE_CONST_SIZE_BYTES`: The size in bytes of user-allocated constant memory required by this function.

- ▶ `CU_FUNC_ATTRIBUTE_LOCAL_SIZE_BYTES`: The size in bytes of local memory used by each thread of this function.
- ▶ `CU_FUNC_ATTRIBUTE_NUM_REGS`: The number of registers used by each thread of this function.
- ▶ `CU_FUNC_ATTRIBUTE_PTX_VERSION`: The PTX virtual architecture version for which the function was compiled. This value is the major PTX version * 10 + the minor PTX version, so a PTX version 1.3 function would return the value 13. Note that this may return the undefined value of 0 for cubins compiled prior to CUDA 3.0.
- ▶ `CU_FUNC_ATTRIBUTE_BINARY_VERSION`: The binary architecture version for which the function was compiled. This value is the major binary version * 10 + the minor binary version, so a binary version 1.3 function would return the value 13. Note that this will return a value of 10 for legacy cubins that do not have a properly-encoded binary architecture version.
- ▶ `CU_FUNC_CACHE_MODE_CA`: The attribute to indicate whether the function has been compiled with user specified option "-Xptxas --dlcm=ca" set .
- ▶ `CU_FUNC_ATTRIBUTE_MAX_DYNAMIC_SHARED_SIZE_BYTES`: The maximum size in bytes of dynamically-allocated shared memory.
- ▶ `CU_FUNC_ATTRIBUTE_PREFERRED_SHARED_MEMORY_CARVEOUT`: Preferred shared memory-L1 cache split ratio in percent of shared memory.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxGetCacheConfig`, `cuCtxSetCacheConfig`, `cuFuncSetCacheConfig`, `cuLaunchKernel`, `cudaFuncGetAttributes` `cudaFuncSetAttribute`

CUresult cuFuncSetAttribute (CUfunction hfunc, CUfunction_attribute attrib, int value)

Sets information about a function.

Parameters

hfunc

- Function to query attribute of

attrib

- Attribute requested

value

- The value to set

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_HANDLE`, `CUDA_ERROR_INVALID_VALUE`

Description

This call sets the value of a specified attribute `attrib` on the kernel given by `hfunc` to an integer value specified by `val`. This function returns `CUDA_SUCCESS` if the new value of the attribute could be successfully set. If the set fails, this call will return an error. Not all attributes can have values set. Attempting to set a value on a read-only attribute will result in an error (`CUDA_ERROR_INVALID_VALUE`).

Supported attributes for the `cuFuncSetAttribute` call are:

- ▶ `CU_FUNC_ATTRIBUTE_MAX_DYNAMIC_SHARED_SIZE_BYTES`:
 This maximum size in bytes of dynamically-allocated shared memory. The value should contain the requested maximum size of dynamically-allocated shared memory. The sum of this value and the function attribute `CU_FUNC_ATTRIBUTE_SHARED_SIZE_BYTES` cannot exceed the device attribute `CU_DEVICE_ATTRIBUTE_MAX_SHARED_MEMORY_PER_BLOCK_OPTIN`. The maximal size of requestable dynamic shared memory may differ by GPU architecture.
- ▶ `CU_FUNC_ATTRIBUTE_PREFERRED_SHARED_MEMORY_CARVEOUT`: On devices where the L1 cache and shared memory use the same hardware resources, this sets the shared memory carveout preference, in percent of the total resources. This is only a hint, and the driver can choose a different ratio if required to execute the function.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxGetCacheConfig`, `cuCtxSetCacheConfig`, `cuFuncSetCacheConfig`, `cuLaunchKernel`,
`cudaFuncGetAttributes`, `cudaFuncSetAttribute`

CUresult cuFuncSetCacheConfig (CUfunction hfunc, CUfunc_cache config)

Sets the preferred cache configuration for a device function.

Parameters

hfunc

- Kernel to configure cache for

config

- Requested cache configuration

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_VALUE,
CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED,
CUDA_ERROR_INVALID_CONTEXT

Description

On devices where the L1 cache and shared memory use the same hardware resources, this sets through `config` the preferred cache configuration for the device function `hfunc`. This is only a preference. The driver will use the requested configuration if possible, but it is free to choose a different configuration if required to execute `hfunc`. Any context-wide preference set via `cuCtxSetCacheConfig()` will be overridden by this per-function setting unless the per-function setting is `CU_FUNC_CACHE_PREFER_NONE`. In that case, the current context-wide setting will be used.

This setting does nothing on devices where the size of the L1 cache and shared memory are fixed.

Launching a kernel with a different preference than the most recent preference setting may insert a device-side synchronization point.

The supported cache configurations are:

- ▶ `CU_FUNC_CACHE_PREFER_NONE`: no preference for shared memory or L1 (default)
- ▶ `CU_FUNC_CACHE_PREFER_SHARED`: prefer larger shared memory and smaller L1 cache
- ▶ `CU_FUNC_CACHE_PREFER_L1`: prefer larger L1 cache and smaller shared memory
- ▶ `CU_FUNC_CACHE_PREFER_EQUAL`: prefer equal sized L1 cache and shared memory



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxGetCacheConfig](#), [cuCtxSetCacheConfig](#), [cuFuncGetAttribute](#), [cuLaunchKernel](#), [cudaFuncSetCacheConfig](#)

CUresult cuFuncSetSharedMemConfig (CUfunction hfunc, CUsharedconfig config)

Sets the shared memory configuration for a device function.

Parameters

hfunc

- kernel to be given a shared memory config

config

- requested shared memory configuration

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_VALUE,
CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED,
CUDA_ERROR_INVALID_CONTEXT

Description

On devices with configurable shared memory banks, this function will force all subsequent launches of the specified device function to have the given shared memory bank size configuration. On any given launch of the function, the shared memory configuration of the device will be temporarily changed if needed to suit the function's preferred configuration. Changes in shared memory configuration between subsequent launches of functions, may introduce a device side synchronization point.

Any per-function setting of shared memory bank size set via [cuFuncSetSharedMemConfig](#) will override the context wide setting set with [cuCtxSetSharedMemConfig](#).

Changing the shared memory bank size will not increase shared memory usage or affect occupancy of kernels, but may have major effects on performance. Larger bank sizes will allow for greater potential bandwidth to shared memory, but will change what kinds of accesses to shared memory will result in bank conflicts.

This function will do nothing on devices with fixed shared memory bank size.

The supported bank configurations are:

- ▶ `CU_SHARED_MEM_CONFIG_DEFAULT_BANK_SIZE`: use the context's shared memory configuration when launching this function.
- ▶ `CU_SHARED_MEM_CONFIG_FOUR_BYTE_BANK_SIZE`: set shared memory bank width to be natively four bytes when launching this function.
- ▶ `CU_SHARED_MEM_CONFIG_EIGHT_BYTE_BANK_SIZE`: set shared memory bank width to be natively eight bytes when launching this function.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxGetCacheConfig`, `cuCtxSetCacheConfig`, `cuCtxGetSharedMemConfig`, `cuCtxSetSharedMemConfig`, `cuFuncGetAttribute`, `cuLaunchKernel`, `cudaFuncSetSharedMemConfig`

CUresult cuLaunchCooperativeKernel (CUfunction f, unsigned int gridDimX, unsigned int gridDimY, unsigned int gridDimZ, unsigned int blockDimX, unsigned int blockDimY, unsigned int blockDimZ, unsigned int sharedMemBytes, CUstream hStream, void **kernelParams)

Launches a CUDA function where thread blocks can cooperate and synchronize as they execute.

Parameters

f

- Kernel to launch

gridDimX

- Width of grid in blocks

gridDimY

- Height of grid in blocks

gridDimZ

- Depth of grid in blocks

blockDimX

- X dimension of each thread block

blockDimY

- Y dimension of each thread block

blockDimZ

- Z dimension of each thread block

sharedMemBytes

- Dynamic shared-memory size per thread block in bytes

hStream

- Stream identifier

kernelParams

- Array of pointers to kernel parameters

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_INVALID_IMAGE,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_LAUNCH_FAILED,
 CUDA_ERROR_LAUNCH_OUT_OF_RESOURCES,
 CUDA_ERROR_LAUNCH_TIMEOUT,
 CUDA_ERROR_LAUNCH_INCOMPATIBLE_TEXTURING,
 CUDA_ERROR_COOPERATIVE_LAUNCH_TOO_LARGE,
 CUDA_ERROR_SHARED_OBJECT_INIT_FAILED

Description

Invokes the kernel `f` on a `gridDimX x gridDimY x gridDimZ` grid of blocks. Each block contains `blockDimX x blockDimY x blockDimZ` threads.

`sharedMemBytes` sets the amount of dynamic shared memory that will be available to each thread block.

The device on which this kernel is invoked must have a non-zero value for the device attribute `CU_DEVICE_ATTRIBUTE_COOPERATIVE_LAUNCH`.

The total number of blocks launched cannot exceed the maximum number of blocks per multiprocessor as returned by `cuOccupancyMaxActiveBlocksPerMultiprocessor` (or `cuOccupancyMaxActiveBlocksPerMultiprocessorWithFlags`) times the number of multiprocessors as specified by the device attribute `CU_DEVICE_ATTRIBUTE_MULTIPROCESSOR_COUNT`.

The kernel cannot make use of CUDA dynamic parallelism.

Kernel parameters must be specified via `kernelParams`. If `f` has `N` parameters, then `kernelParams` needs to be an array of `N` pointers. Each of `kernelParams[0]` through `kernelParams[N-1]` must point to a region of memory from which the actual kernel parameter will be copied. The number of kernel parameters and their offsets and sizes do not need to be specified as that information is retrieved directly from the kernel's image.

Calling `cuLaunchCooperativeKernel()` sets persistent function state that is the same as function state set through `cuLaunchKernel` API

When the kernel `f` is launched via `cuLaunchCooperativeKernel()`, the previous block shape, shared size and parameter info associated with `f` is overwritten.

Note that to use `cuLaunchCooperativeKernel()`, the kernel `f` must either have been compiled with toolchain version 3.2 or later so that it will contain kernel parameter information, or have no kernel parameters. If either of these conditions is not met, then `cuLaunchCooperativeKernel()` will return `CUDA_ERROR_INVALID_IMAGE`.



- ▶ This function uses standard **default stream** semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxGetCacheConfig`, `cuCtxSetCacheConfig`, `cuFuncSetCacheConfig`,
`cuFuncGetAttribute`, `cuLaunchCooperativeKernelMultiDevice`,
`cudaLaunchCooperativeKernel`

CUresult cuLaunchCooperativeKernelMultiDevice (CUDA_LAUNCH_PARAMS *launchParamsList, unsigned int numDevices, unsigned int flags)

Launches CUDA functions on multiple devices where thread blocks can cooperate and synchronize as they execute.

Parameters

launchParamsList

- List of launch parameters, one per device

numDevices

- Size of the `launchParamsList` array

flags

- Flags to control launch behavior

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_HANDLE`, `CUDA_ERROR_INVALID_IMAGE`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_LAUNCH_FAILED`,
`CUDA_ERROR_LAUNCH_OUT_OF_RESOURCES`,

CUDA_ERROR_LAUNCH_TIMEOUT,
 CUDA_ERROR_LAUNCH_INCOMPATIBLE_TEXTURING,
 CUDA_ERROR_COOPERATIVE_LAUNCH_TOO_LARGE,
 CUDA_ERROR_SHARED_OBJECT_INIT_FAILED

Description

Invokes kernels as specified in the `launchParamsList` array where each element of the array specifies all the parameters required to perform a single kernel launch. These kernels can cooperate and synchronize as they execute. The size of the array is specified by `numDevices`.

No two kernels can be launched on the same device. All the devices targeted by this multi-device launch must be identical.

All devices must have a non-zero value for the device attribute

`CU_DEVICE_ATTRIBUTE_COOPERATIVE_MULTI_DEVICE_LAUNCH`.

All kernels launched must be identical with respect to the compiled code. Note that any `__device__`, `__constant__` or `__managed__` variables present in the module that owns the kernel launched on each device, are independently instantiated on every device. It is the application's responsibility to ensure these variables are initialized and used appropriately.

The size of the grids as specified in blocks, the size of the blocks themselves and the amount of shared memory used by each thread block must also match across all launched kernels.

The streams used to launch these kernels must have been created via either `cuStreamCreate` or `cuStreamCreateWithPriority`. The NULL stream or `CU_STREAM_LEGACY` or `CU_STREAM_PER_THREAD` cannot be used.

The total number of blocks launched per kernel cannot exceed the maximum number of blocks per multiprocessor as returned by `cuOccupancyMaxActiveBlocksPerMultiprocessor` (or `cuOccupancyMaxActiveBlocksPerMultiprocessorWithFlags`) times the number of multiprocessors as specified by the device attribute `CU_DEVICE_ATTRIBUTE_MULTIPROCESSOR_COUNT`. Since the total number of blocks launched per device has to match across all devices, the maximum number of blocks that can be launched per device will be limited by the device with the least number of multiprocessors.

The kernels cannot make use of CUDA dynamic parallelism.

The `CUDA_LAUNCH_PARAMS` structure is defined as:

```
↑
typedef struct CUDA_LAUNCH_PARAMS_st
{
    CUfunction function;
    unsigned int gridDimX;
    unsigned int gridDimY;
    unsigned int gridDimZ;
    unsigned int blockDimX;
    unsigned int blockDimY;
    unsigned int blockDimZ;
    unsigned int sharedMemBytes;
    CUstream hStream;
    void **kernelParams;
} CUDA_LAUNCH_PARAMS;
```

where:

- ▶ `CUDA_LAUNCH_PARAMS::function` specifies the kernel to be launched. All functions must be identical with respect to the compiled code.
- ▶ `CUDA_LAUNCH_PARAMS::gridDimX` is the width of the grid in blocks. This must match across all kernels launched.
- ▶ `CUDA_LAUNCH_PARAMS::gridDimY` is the height of the grid in blocks. This must match across all kernels launched.
- ▶ `CUDA_LAUNCH_PARAMS::gridDimZ` is the depth of the grid in blocks. This must match across all kernels launched.
- ▶ `CUDA_LAUNCH_PARAMS::blockDimX` is the X dimension of each thread block. This must match across all kernels launched.
- ▶ `CUDA_LAUNCH_PARAMS::blockDimY` is the Y dimension of each thread block. This must match across all kernels launched.
- ▶ `CUDA_LAUNCH_PARAMS::blockDimZ` is the Z dimension of each thread block. This must match across all kernels launched.
- ▶ `CUDA_LAUNCH_PARAMS::sharedMemBytes` is the dynamic shared-memory size per thread block in bytes. This must match across all kernels launched.
- ▶ `CUDA_LAUNCH_PARAMS::hStream` is the handle to the stream to perform the launch in. This cannot be the NULL stream or `CU_STREAM_LEGACY` or `CU_STREAM_PER_THREAD`. The CUDA context associated with this stream must match that associated with `CUDA_LAUNCH_PARAMS::function`.
- ▶ `CUDA_LAUNCH_PARAMS::kernelParams` is an array of pointers to kernel parameters. If `CUDA_LAUNCH_PARAMS::function` has N parameters, then `CUDA_LAUNCH_PARAMS::kernelParams` needs to be an array of N pointers. Each of `CUDA_LAUNCH_PARAMS::kernelParams[0]` through `CUDA_LAUNCH_PARAMS::kernelParams[N-1]` must point to a region of memory from which the actual kernel parameter will be copied. The number of kernel parameters and their offsets and sizes do not need to be specified as that information is retrieved directly from the kernel's image.

By default, the kernel won't begin execution on any GPU until all prior work in all the specified streams has completed. This behavior can be overridden by specifying the

flag `CUDA_COOPERATIVE_LAUNCH_MULTI_DEVICE_NO_PRE_LAUNCH_SYNC`. When this flag is specified, each kernel will only wait for prior work in the stream corresponding to that GPU to complete before it begins execution.

Similarly, by default, any subsequent work pushed in any of the specified streams will not begin execution until the kernels on all GPUs have completed. This behavior can be overridden by specifying the flag `CUDA_COOPERATIVE_LAUNCH_MULTI_DEVICE_NO_POST_LAUNCH_SYNC`. When this flag is specified, any subsequent work pushed in any of the specified streams will only wait for the kernel launched on the GPU corresponding to that stream to complete before it begins execution.

Calling `cuLaunchCooperativeKernelMultiDevice()` sets persistent function state that is the same as function state set through `cuLaunchKernel` API when called individually for each element in `launchParamsList`.

When kernels are launched via `cuLaunchCooperativeKernelMultiDevice()`, the previous block shape, shared size and parameter info associated with each `CUDA_LAUNCH_PARAMS::function` in `launchParamsList` is overwritten.

Note that to use `cuLaunchCooperativeKernelMultiDevice()`, the kernels must either have been compiled with toolchain version 3.2 or later so that it will contain kernel parameter information, or have no kernel parameters. If either of these conditions is not met, then `cuLaunchCooperativeKernelMultiDevice()` will return `CUDA_ERROR_INVALID_IMAGE`.



- ▶ This function uses standard `default stream` semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxGetCacheConfig`, `cuCtxSetCacheConfig`, `cuFuncSetCacheConfig`,
`cuFuncGetAttribute`, `cuLaunchCooperativeKernel`,
`cudaLaunchCooperativeKernelMultiDevice`

`CUresult cuLaunchKernel (CUfunction f, unsigned int gridDimX, unsigned int gridDimY, unsigned int gridDimZ, unsigned int blockDimX, unsigned int blockDimY,`

unsigned int blockDimZ, unsigned int sharedMemBytes, CUstream hStream, void **kernelParams, void **extra)

Launches a CUDA function.

Parameters

f

- Kernel to launch

gridDimX

- Width of grid in blocks

gridDimY

- Height of grid in blocks

gridDimZ

- Depth of grid in blocks

blockDimX

- X dimension of each thread block

blockDimY

- Y dimension of each thread block

blockDimZ

- Z dimension of each thread block

sharedMemBytes

- Dynamic shared-memory size per thread block in bytes

hStream

- Stream identifier

kernelParams

- Array of pointers to kernel parameters

extra

- Extra options

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_INVALID_IMAGE,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_LAUNCH_FAILED,
 CUDA_ERROR_LAUNCH_OUT_OF_RESOURCES,
 CUDA_ERROR_LAUNCH_TIMEOUT,
 CUDA_ERROR_LAUNCH_INCOMPATIBLE_TEXTURING,
 CUDA_ERROR_SHARED_OBJECT_INIT_FAILED

Description

Invokes the kernel `f` on a `gridDimX` x `gridDimY` x `gridDimZ` grid of blocks. Each block contains `blockDimX` x `blockDimY` x `blockDimZ` threads.

`sharedMemBytes` sets the amount of dynamic shared memory that will be available to each thread block.

Kernel parameters to `f` can be specified in one of two ways:

1) Kernel parameters can be specified via `kernelParams`. If `f` has `N` parameters, then `kernelParams` needs to be an array of `N` pointers. Each of `kernelParams[0]` through `kernelParams[N-1]` must point to a region of memory from which the actual kernel parameter will be copied. The number of kernel parameters and their offsets and sizes do not need to be specified as that information is retrieved directly from the kernel's image.

2) Kernel parameters can also be packaged by the application into a single buffer that is passed in via the `extra` parameter. This places the burden on the application of knowing each kernel parameter's size and alignment/padding within the buffer. Here is an example of using the `extra` parameter in this manner:

```
↑
    size_t argBufferSize;
    char argBuffer[256];

    // populate argBuffer and argBufferSize

    void *config[] = {
        CU_LAUNCH_PARAM_BUFFER_POINTER, argBuffer,
        CU_LAUNCH_PARAM_BUFFER_SIZE,   &argBufferSize,
        CU_LAUNCH_PARAM_END
    };
    status = cuLaunchKernel(f, gx, gy, gz, bx, by, bz, sh, s, NULL,
config);
```

The `extra` parameter exists to allow `cuLaunchKernel` to take additional less commonly used arguments. `extra` specifies a list of names of extra settings and their corresponding values. Each extra setting name is immediately followed by the corresponding value. The list must be terminated with either `NULL` or `CU_LAUNCH_PARAM_END`.

- ▶ `CU_LAUNCH_PARAM_END`, which indicates the end of the `extra` array;
- ▶ `CU_LAUNCH_PARAM_BUFFER_POINTER`, which specifies that the next value in `extra` will be a pointer to a buffer containing all the kernel parameters for launching kernel `f`;
- ▶ `CU_LAUNCH_PARAM_BUFFER_SIZE`, which specifies that the next value in `extra` will be a pointer to a `size_t` containing the size of the buffer specified with `CU_LAUNCH_PARAM_BUFFER_POINTER`;

The error `CUDA_ERROR_INVALID_VALUE` will be returned if kernel parameters are specified with both `kernelParams` and `extra` (i.e. both `kernelParams` and `extra` are non-NULL).

Calling `cuLaunchKernel()` sets persistent function state that is the same as function state set through the following deprecated APIs: `cuFuncSetBlockShape()`, `cuFuncSetSharedSize()`, `cuParamSetSize()`, `cuParamSeti()`, `cuParamSetf()`, `cuParamSetv()`.

When the kernel `f` is launched via `cuLaunchKernel()`, the previous block shape, shared size and parameter info associated with `f` is overwritten.

Note that to use `cuLaunchKernel()`, the kernel `f` must either have been compiled with toolchain version 3.2 or later so that it will contain kernel parameter information, or have no kernel parameters. If either of these conditions is not met, then `cuLaunchKernel()` will return `CUDA_ERROR_INVALID_IMAGE`.



- ▶ This function uses standard **default stream** semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxGetCacheConfig`, `cuCtxSetCacheConfig`, `cuFuncSetCacheConfig`,
`cuFuncGetAttribute`, `cudaLaunchKernel`

4.17. Execution Control [DEPRECATED]

This section describes the deprecated execution control functions of the low-level CUDA driver application programming interface.

CUresult cuFuncSetBlockShape (CUfunction hfunc, int x, int y, int z)

Sets the block-dimensions for the function.

Parameters

hfunc

- Kernel to specify dimensions of

x

- X dimension

y

- Y dimension

z

- Z dimension

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_HANDLE`, `CUDA_ERROR_INVALID_VALUE`

Description

Deprecated

Specifies the x , y , and z dimensions of the thread blocks that are created when the kernel given by `hfunc` is launched.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuFuncSetSharedSize](#), [cuFuncSetCacheConfig](#), [cuFuncGetAttribute](#), [cuParamSetSize](#), [cuParamSeti](#), [cuParamSetf](#), [cuParamSetv](#), [cuLaunch](#), [cuLaunchGrid](#), [cuLaunchGridAsync](#), [cuLaunchKernel](#)

CUresult cuFuncSetSharedSize (CUfunction hfunc, unsigned int bytes)

Sets the dynamic shared-memory size for the function.

Parameters

hfunc

- Kernel to specify dynamic shared-memory size for

bytes

- Dynamic shared-memory size per thread in bytes

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_HANDLE](#), [CUDA_ERROR_INVALID_VALUE](#)

Description

Deprecated

Sets through `bytes` the amount of dynamic shared memory that will be available to each thread block when the kernel given by `hfunc` is launched.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuFuncSetBlockShape`, `cuFuncSetCacheConfig`, `cuFuncGetAttribute`,
`cuParamSetSize`, `cuParamSeti`, `cuParamSetf`, `cuParamSetv`, `cuLaunch`, `cuLaunchGrid`,
`cuLaunchGridAsync`, `cuLaunchKernel`

CUresult cuLaunch (CUfunction f)

Launches a CUDA function.

Parameters

f

- Kernel to launch

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_LAUNCH_FAILED`,
`CUDA_ERROR_LAUNCH_OUT_OF_RESOURCES`,
`CUDA_ERROR_LAUNCH_TIMEOUT`,
`CUDA_ERROR_LAUNCH_INCOMPATIBLE_TEXTURING`,
`CUDA_ERROR_SHARED_OBJECT_INIT_FAILED`

Description

Deprecated

Invokes the kernel `f` on a 1 x 1 x 1 grid of blocks. The block contains the number of threads specified by a previous call to `cuFuncSetBlockShape()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuFuncSetBlockShape`, `cuFuncSetSharedSize`, `cuFuncGetAttribute`, `cuParamSetSize`,
`cuParamSetf`, `cuParamSeti`, `cuParamSetv`, `cuLaunchGrid`, `cuLaunchGridAsync`,
`cuLaunchKernel`

CUresult cuLaunchGrid (CUfunction f, int grid_width, int grid_height)

Launches a CUDA function.

Parameters

f

- Kernel to launch

grid_width

- Width of grid in blocks

grid_height

- Height of grid in blocks

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_LAUNCH_FAILED,
CUDA_ERROR_LAUNCH_OUT_OF_RESOURCES,
CUDA_ERROR_LAUNCH_TIMEOUT,
CUDA_ERROR_LAUNCH_INCOMPATIBLE_TEXTURING,
CUDA_ERROR_SHARED_OBJECT_INIT_FAILED

Description

Deprecated

Invokes the kernel `f` on a `grid_width` x `grid_height` grid of blocks. Each block contains the number of threads specified by a previous call to `cuFuncSetBlockShape()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuFuncSetBlockShape`, `cuFuncSetSharedSize`, `cuFuncGetAttribute`, `cuParamSetSize`,
`cuParamSetf`, `cuParamSeti`, `cuParamSetv`, `cuLaunch`, `cuLaunchGridAsync`,
`cuLaunchKernel`

CUresult cuLaunchGridAsync (CUfunction f, int grid_width, int grid_height, CUstream hStream)

Launches a CUDA function.

Parameters

f

- Kernel to launch

grid_width

- Width of grid in blocks

grid_height

- Height of grid in blocks

hStream

- Stream identifier

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_LAUNCH_FAILED,
 CUDA_ERROR_LAUNCH_OUT_OF_RESOURCES,
 CUDA_ERROR_LAUNCH_TIMEOUT,
 CUDA_ERROR_LAUNCH_INCOMPATIBLE_TEXTURING,
 CUDA_ERROR_SHARED_OBJECT_INIT_FAILED

Description

Deprecated

Invokes the kernel `f` on a `grid_width` x `grid_height` grid of blocks. Each block contains the number of threads specified by a previous call to `cuFuncSetBlockShape()`.



- ▶ In certain cases where cubins are created with no ABI (i.e., using `ptxas --abi-compile no`), this function may serialize kernel launches. In order to force the CUDA driver to retain asynchronous behavior, set the `CU_CTX_LMEM_RESIZE_TO_MAX` flag during context creation (see `cuCtxCreate`).
- ▶ This function uses standard `default stream` semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuFuncSetBlockShape](#), [cuFuncSetSharedSize](#), [cuFuncGetAttribute](#), [cuParamSetSize](#), [cuParamSetf](#), [cuParamSeti](#), [cuParamSetv](#), [cuLaunch](#), [cuLaunchGrid](#), [cuLaunchKernel](#)

CUresult cuParamSetf (CUfunction hfunc, int offset, float value)

Adds a floating-point parameter to the function's argument list.

Parameters

hfunc

- Kernel to add parameter to

offset

- Offset to add parameter to argument list

value

- Value of parameter

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Deprecated

Sets a floating-point parameter that will be specified the next time the kernel corresponding to `hfunc` will be invoked. `offset` is a byte offset.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuFuncSetBlockShape](#), [cuFuncSetSharedSize](#), [cuFuncGetAttribute](#), [cuParamSetSize](#), [cuParamSeti](#), [cuParamSetv](#), [cuLaunch](#), [cuLaunchGrid](#), [cuLaunchGridAsync](#), [cuLaunchKernel](#)

CUresult cuParamSeti (CUfunction hfunc, int offset, unsigned int value)

Adds an integer parameter to the function's argument list.

Parameters

hfunc

- Kernel to add parameter to

offset

- Offset to add parameter to argument list

value

- Value of parameter

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Deprecated

Sets an integer parameter that will be specified the next time the kernel corresponding to `hfunc` will be invoked. `offset` is a byte offset.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuFuncSetBlockShape](#), [cuFuncSetSharedSize](#), [cuFuncGetAttribute](#), [cuParamSetSize](#),
[cuParamSetf](#), [cuParamSetv](#), [cuLaunch](#), [cuLaunchGrid](#), [cuLaunchGridAsync](#),
[cuLaunchKernel](#)

CUresult cuParamSetSize (CUfunction hfunc, unsigned int numbytes)

Sets the parameter size for the function.

Parameters

hfunc

- Kernel to set parameter size for

numbytes

- Size of parameter list in bytes

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Deprecated

Sets through `numbytes` the total size in bytes needed by the function parameters of the kernel corresponding to `hfunc`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuFuncSetBlockShape`, `cuFuncSetSharedSize`, `cuFuncGetAttribute`, `cuParamSetf`,
`cuParamSeti`, `cuParamSetv`, `cuLaunch`, `cuLaunchGrid`, `cuLaunchGridAsync`,
`cuLaunchKernel`

CUresult cuParamSetTexRef (CUfunction hfunc, int texunit, CUtexref hTexRef)

Adds a texture-reference to the function's argument list.

Parameters**hfunc**

- Kernel to add texture-reference to

texunit

- Texture unit (must be `CU_PARAM_TR_DEFAULT`)

hTexRef

- Texture-reference to add to argument list

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Deprecated

Makes the CUDA array or linear memory bound to the texture reference `hTexRef` available to a device program as a texture. In this version of CUDA, the texture-reference must be obtained via `cuModuleGetTexRef()` and the `texunit` parameter must be set to `CU_PARAM_TR_DEFAULT`.



Note that this function may also return error codes from previous, asynchronous launches.

CUresult cuParamSetv (CUfunction hfunc, int offset, void *ptr, unsigned int numbytes)

Adds arbitrary data to the function's argument list.

Parameters

hfunc

- Kernel to add data to

offset

- Offset to add data to argument list

ptr

- Pointer to arbitrary data

numbytes

- Size of data to copy in bytes

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Deprecated

Copies an arbitrary amount of data (specified in `numbytes`) from `ptr` into the parameter space of the kernel corresponding to `hfunc`. `offset` is a byte offset.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuFuncSetBlockShape](#), [cuFuncSetSharedSize](#), [cuFuncGetAttribute](#), [cuParamSetSize](#), [cuParamSetf](#), [cuParamSeti](#), [cuLaunch](#), [cuLaunchGrid](#), [cuLaunchGridAsync](#), [cuLaunchKernel](#)

4.18. Occupancy

This section describes the occupancy calculation functions of the low-level CUDA driver application programming interface.

CUresult cuOccupancyMaxActiveBlocksPerMultiprocessor (int *numBlocks, CUfunction func, int blockSize, size_t dynamicSMemSize)

Returns occupancy of a function.

Parameters**numBlocks**

- Returned occupancy

func

- Kernel for which occupancy is calculated

blockSize

- Block size the kernel is intended to be launched with

dynamicSMemSize

- Per-block dynamic shared memory usage intended, in bytes

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#), [CUDA_ERROR_UNKNOWN](#)

Description

Returns in *numBlocks the number of the maximum active blocks per streaming multiprocessor.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cudaOccupancyMaxActiveBlocksPerMultiprocessor`

CUresult

`cuOccupancyMaxActiveBlocksPerMultiprocessorWithFlags`
 (int *numBlocks, CUfunction func, int blockSize, size_t
 dynamicSMemSize, unsigned int flags)

Returns occupancy of a function.

Parameters

numBlocks

- Returned occupancy

func

- Kernel for which occupancy is calculated

blockSize

- Block size the kernel is intended to be launched with

dynamicSMemSize

- Per-block dynamic shared memory usage intended, in bytes

flags

- Requested behavior for the occupancy calculator

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_UNKNOWN`

Description

Returns in *numBlocks the number of the maximum active blocks per streaming multiprocessor.

The Flags parameter controls how special cases are handled. The valid flags are:

- ▶ `CU_OCCUPANCY_DEFAULT`, which maintains the default behavior as `cuOccupancyMaxActiveBlocksPerMultiprocessor`;
- ▶ `CU_OCCUPANCY_DISABLE_CACHING_OVERRIDE`, which suppresses the default behavior on platform where global caching affects occupancy. On such platforms, if caching is enabled, but per-block SM resource usage would result in zero occupancy, the occupancy calculator will calculate the occupancy as if caching is disabled. Setting `CU_OCCUPANCY_DISABLE_CACHING_OVERRIDE` makes the occupancy calculator to return 0 in such cases. More information can be found about this feature in the "Unified L1/Texture Cache" section of the Maxwell tuning guide.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cudaOccupancyMaxActiveBlocksPerMultiprocessorWithFlags`

CUresult cuOccupancyMaxPotentialBlockSize (int *minGridSize, int *blockSize, CUfunction func, CUoccupancyB2DSize blockSizeToDynamicSMemSize, size_t dynamicSMemSize, int blockSizeLimit)

Suggest a launch configuration with reasonable occupancy.

Parameters

minGridSize

- Returned minimum grid size needed to achieve the maximum occupancy

blockSize

- Returned maximum block size that can achieve the maximum occupancy

func

- Kernel for which launch configuration is calculated

blockSizeToDynamicSMemSize

- A function that calculates how much per-block dynamic shared memory `func` uses based on the block size

dynamicSMemSize

- Dynamic shared memory usage intended, in bytes

blockSizeLimit

- The maximum block size `func` is designed to handle

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_UNKNOWN`

Description

Returns in `*blockSize` a reasonable block size that can achieve the maximum occupancy (or, the maximum number of active warps with the fewest blocks per multiprocessor), and in `*minGridSize` the minimum grid size to achieve the maximum occupancy.

If `blockSizeLimit` is 0, the configurator will use the maximum block size permitted by the device / function instead.

If per-block dynamic shared memory allocation is not needed, the user should leave both `blockSizeToDynamicSMemSize` and `dynamicSMemSize` as 0.

If per-block dynamic shared memory allocation is needed, then if the dynamic shared memory size is constant regardless of block size, the size should be passed through `dynamicSMemSize`, and `blockSizeToDynamicSMemSize` should be NULL.

Otherwise, if the per-block dynamic shared memory size varies with different block sizes, the user needs to provide a unary function through `blockSizeToDynamicSMemSize` that computes the dynamic shared memory needed by `func` for any given block size. `dynamicSMemSize` is ignored. An example signature is:

```
↑ // Take block size, returns dynamic shared memory needed
   size_t blockToSmem(int blockSize);
```



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cudaOccupancyMaxPotentialBlockSize](#)

CUresult cuOccupancyMaxPotentialBlockSizeWithFlags (int *minGridSize, int *blockSize, CUfunction func, CUoccupancyB2DSize blockSizeToDynamicSMemSize, size_t dynamicSMemSize, int blockSizeLimit, unsigned int flags)

Suggest a launch configuration with reasonable occupancy.

Parameters

minGridSize

- Returned minimum grid size needed to achieve the maximum occupancy

blockSize

- Returned maximum block size that can achieve the maximum occupancy

func

- Kernel for which launch configuration is calculated

blockSizeToDynamicSMemSize

- A function that calculates how much per-block dynamic shared memory `func` uses based on the block size

dynamicSMemSize

- Dynamic shared memory usage intended, in bytes

blockSizeLimit

- The maximum block size `func` is designed to handle

flags

- Options

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_UNKNOWN

Description

An extended version of `cuOccupancyMaxPotentialBlockSize`. In addition to arguments passed to `cuOccupancyMaxPotentialBlockSize`, `cuOccupancyMaxPotentialBlockSizeWithFlags` also takes a `Flags` parameter.

The `Flags` parameter controls how special cases are handled. The valid flags are:

- ▶ `CU_OCCUPANCY_DEFAULT`, which maintains the default behavior as `cuOccupancyMaxPotentialBlockSize`;
- ▶ `CU_OCCUPANCY_DISABLE_CACHING_OVERRIDE`, which suppresses the default behavior on platform where global caching affects occupancy. On such platforms, the launch configurations that produces maximal occupancy might not support global caching. Setting `CU_OCCUPANCY_DISABLE_CACHING_OVERRIDE` guarantees that the the produced launch configuration is global caching compatible at a potential cost of occupancy. More information can be found about this feature in the "Unified L1/Texture Cache" section of the Maxwell tuning guide.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cudaOccupancyMaxPotentialBlockSizeWithFlags`

4.19. Texture Reference Management

This section describes the texture reference management functions of the low-level CUDA driver application programming interface.

CUresult cuTexRefGetAddress (CUdeviceptr *pdptr, CUtexref hTexRef)

Gets the address associated with a texture reference.

Parameters

pdptr

- Returned device address

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Returns in *pdptr the base address bound to the texture reference hTexRef, or returns CUDA_ERROR_INVALID_VALUE if the texture reference is not bound to any device memory range.

See also:

cuTexRefSetAddress, cuTexRefSetAddress2D, cuTexRefSetAddressMode,
cuTexRefSetArray, cuTexRefSetFilterMode, cuTexRefSetFlags, cuTexRefSetFormat,
cuTexRefGetAddressMode, cuTexRefGetArray, cuTexRefGetFilterMode,
cuTexRefGetFlags, cuTexRefGetFormat

CUresult cuTexRefGetAddressMode (CUaddress_mode *pam, CUtexref hTexRef, int dim)

Gets the addressing mode used by a texture reference.

Parameters

pam

- Returned addressing mode

hTexRef

- Texture reference

dim

- Dimension

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns in *pam the addressing mode corresponding to the dimension dim of the texture reference hTexRef. Currently, the only valid value for dim are 0 and 1.

See also:

cuTexRefSetAddress, cuTexRefSetAddress2D, cuTexRefSetAddressMode,
 cuTexRefSetArray, cuTexRefSetFilterMode, cuTexRefSetFlags, cuTexRefSetFormat,
 cuTexRefGetAddress, cuTexRefGetArray, cuTexRefGetFilterMode, cuTexRefGetFlags,
 cuTexRefGetFormat

CUresult cuTexRefGetArray (CUarray *phArray, CUtexref hTexRef)

Gets the array bound to a texture reference.

Parameters**phArray**

- Returned array

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns in *phArray the CUDA array bound to the texture reference hTexRef, or returns CUDA_ERROR_INVALID_VALUE if the texture reference is not bound to any CUDA array.

See also:

cuTexRefSetAddress, cuTexRefSetAddress2D, cuTexRefSetAddressMode,
 cuTexRefSetArray, cuTexRefSetFilterMode, cuTexRefSetFlags, cuTexRefSetFormat,

[cuTexRefGetAddress](#), [cuTexRefGetAddressMode](#), [cuTexRefGetFilterMode](#),
[cuTexRefGetFlags](#), [cuTexRefGetFormat](#)

CUresult cuTexRefGetBorderColor (float *pBorderColor, CUtexref hTexRef)

Gets the border color used by a texture reference.

Parameters

pBorderColor

- Returned Type and Value of RGBA color

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns in `pBorderColor`, values of the RGBA color used by the texture reference `hTexRef`. The color value is of type float and holds color components in the following sequence: `pBorderColor[0]` holds 'R' component `pBorderColor[1]` holds 'G' component `pBorderColor[2]` holds 'B' component `pBorderColor[3]` holds 'A' component

See also:

[cuTexRefSetAddressMode](#), [cuTexRefSetAddressMode](#), [cuTexRefSetBorderColor](#)

CUresult cuTexRefGetFilterMode (CUfilter_mode *pfm, CUtexref hTexRef)

Gets the filter-mode used by a texture reference.

Parameters

pfm

- Returned filtering mode

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns in `*pfm` the filtering mode of the texture reference `hTexRef`.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFilterMode`, `cuTexRefSetFlags`, `cuTexRefSetFormat`,
`cuTexRefGetAddress`, `cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFlags`,
`cuTexRefGetFormat`

CUresult cuTexRefGetFlags (unsigned int *pFlags, CUtexref hTexRef)

Gets the flags used by a texture reference.

Parameters**pFlags**

- Returned flags

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns in `*pFlags` the flags of the texture reference `hTexRef`.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFilterMode`, `cuTexRefSetFlags`, `cuTexRefSetFormat`,
`cuTexRefGetAddress`, `cuTexRefGetAddressMode`, `cuTexRefGetArray`,
`cuTexRefGetFilterMode`, `cuTexRefGetFormat`

CUresult cuTexRefGetFormat (CUarray_format *pFormat, int *pNumChannels, CUtexref hTexRef)

Gets the format used by a texture reference.

Parameters

pFormat

- Returned format

pNumChannels

- Returned number of components

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Returns in `*pFormat` and `*pNumChannels` the format and number of components of the CUDA array bound to the texture reference `hTexRef`. If `pFormat` or `pNumChannels` is NULL, it will be ignored.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFilterMode`, `cuTexRefSetFlags`, `cuTexRefSetFormat`,
`cuTexRefGetAddress`, `cuTexRefGetAddressMode`, `cuTexRefGetArray`,
`cuTexRefGetFilterMode`, `cuTexRefGetFlags`

CUresult cuTexRefGetMaxAnisotropy (int *pmaxAniso, CUtexref hTexRef)

Gets the maximum anisotropy for a texture reference.

Parameters

pmaxAniso

- Returned maximum anisotropy

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns the maximum anisotropy in `pmaxAniso` that's used when reading memory through the texture reference `hTexRef`.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`

CUresult cuTexRefGetMipmapFilterMode (CUfilter_mode *pfm, CUtexref hTexRef)

Gets the mipmap filtering mode for a texture reference.

Parameters**pfm**

- Returned mipmap filtering mode

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns the mipmap filtering mode in `pfm` that's used when reading memory through the texture reference `hTexRef`.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`

CUresult cuTexRefGetMipmapLevelBias (float *pbias, CUtexref hTexRef)

Gets the mipmap level bias for a texture reference.

Parameters

pbias

- Returned mipmap level bias

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Returns the mipmap level bias in `pbias` that's added to the specified mipmap level when reading memory through the texture reference `hTexRef`.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`

CUresult cuTexRefGetMipmapLevelClamp (float *pminMipmapLevelClamp, float *pmaxMipmapLevelClamp, CUtexref hTexRef)

Gets the min/max mipmap level clamps for a texture reference.

Parameters

pminMipmapLevelClamp

- Returned mipmap min level clamp

pmaxMipmapLevelClamp

- Returned mipmap max level clamp

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns the min/max mipmap level clamps in `pminMipmapLevelClamp` and `pmaxMipmapLevelClamp` that's used when reading memory through the texture reference `hTexRef`.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`

CUresult cuTexRefGetMipmappedArray (CUmipmappedArray *phMipmappedArray, CUtexref hTexRef)

Gets the mipmapped array bound to a texture reference.

Parameters**phMipmappedArray**

- Returned mipmapped array

hTexRef

- Texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns in `*phMipmappedArray` the CUDA mipmapped array bound to the texture reference `hTexRef`, or returns `CUDA_ERROR_INVALID_VALUE` if the texture reference is not bound to any CUDA mipmapped array.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFilterMode`, `cuTexRefSetFlags`, `cuTexRefSetFormat`,
`cuTexRefGetAddress`, `cuTexRefGetAddressMode`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`

CUresult cuTexRefSetAddress (size_t *ByteOffset, CUtexref hTexRef, CUdeviceptr dptr, size_t bytes)

Binds an address as a texture reference.

Parameters

ByteOffset

- Returned byte offset

hTexRef

- Texture reference to bind

dptr

- Device pointer to bind

bytes

- Size of memory to bind in bytes

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Binds a linear address range to the texture reference `hTexRef`. Any previous address or CUDA array state associated with the texture reference is superseded by this function. Any memory previously bound to `hTexRef` is unbound.

Since the hardware enforces an alignment requirement on texture base addresses, `cuTexRefSetAddress()` passes back a byte offset in `*ByteOffset` that must be applied to texture fetches in order to read from the desired memory. This offset must be divided by the texel size and passed to kernels that read from the texture so they can be applied to the `tex1Dfetch()` function.

If the device memory pointer was returned from `cuMemAlloc()`, the offset is guaranteed to be 0 and NULL may be passed as the `ByteOffset` parameter.

The total number of elements (or texels) in the linear address range cannot exceed `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_LINEAR_WIDTH`. The number of elements is computed as $(\text{bytes} / \text{bytesPerElement})$, where `bytesPerElement` is determined from the data format and number of components set using `cuTexRefSetFormat()`.

See also:

[cuTexRefSetAddress2D](#), [cuTexRefSetAddressMode](#), [cuTexRefSetArray](#),
[cuTexRefSetFilterMode](#), [cuTexRefSetFlags](#), [cuTexRefSetFormat](#), [cuTexRefGetAddress](#),
[cuTexRefGetAddressMode](#), [cuTexRefGetArray](#), [cuTexRefGetFilterMode](#),
[cuTexRefGetFlags](#), [cuTexRefGetFormat](#), [cudaBindTexture](#)

CUresult cuTexRefSetAddress2D (CUtexref hTexRef, const CUDA_ARRAY_DESCRIPTOR *desc, CUdeviceptr dptr, size_t Pitch)

Binds an address as a 2D texture reference.

Parameters**hTexRef**

- Texture reference to bind

desc

- Descriptor of CUDA array

dptr

- Device pointer to bind

Pitch

- Line pitch in bytes

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Binds a linear address range to the texture reference `hTexRef`. Any previous address or CUDA array state associated with the texture reference is superseded by this function. Any memory previously bound to `hTexRef` is unbound.

Using a `tex2D()` function inside a kernel requires a call to either [cuTexRefSetArray\(\)](#) to bind the corresponding texture reference to an array, or [cuTexRefSetAddress2D\(\)](#) to bind the texture reference to linear memory.

Function calls to [cuTexRefSetFormat\(\)](#) cannot follow calls to [cuTexRefSetAddress2D\(\)](#) for the same texture reference.

It is required that `dptr` be aligned to the appropriate hardware-specific texture alignment. You can query this value using the device attribute [CU_DEVICE_ATTRIBUTE_TEXTURE_ALIGNMENT](#). If an unaligned `dptr` is supplied, [CUDA_ERROR_INVALID_VALUE](#) is returned.

`Pitch` has to be aligned to the hardware-specific texture pitch alignment. This value can be queried using the device attribute `CU_DEVICE_ATTRIBUTE_TEXTURE_PITCH_ALIGNMENT`. If an unaligned `Pitch` is supplied, `CUDA_ERROR_INVALID_VALUE` is returned.

`Width` and `Height`, which are specified in elements (or texels), cannot exceed `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_WIDTH` and `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_HEIGHT` respectively. `Pitch`, which is specified in bytes, cannot exceed `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_PITCH`.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddressMode`, `cuTexRefSetArray`, `cuTexRefSetFilterMode`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`, `cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`, `cuTexRefGetFlags`, `cuTexRefGetFormat`, `cudaBindTexture2D`

CUresult cuTexRefSetAddressMode (CUtexref hTexRef, int dim, CUaddress_mode am)

Sets the addressing mode for a texture reference.

Parameters

hTexRef

- Texture reference

dim

- Dimension

am

- Addressing mode to set

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`, `CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`, `CUDA_ERROR_INVALID_VALUE`

Description

Specifies the addressing mode `am` for the given dimension `dim` of the texture reference `hTexRef`. If `dim` is zero, the addressing mode is applied to the first parameter

of the functions used to fetch from the texture; if `dim` is 1, the second, and so on.

`CUaddress_mode` is defined as:

```
↑ typedef enum CUaddress_mode_enum {
    CU_TR_ADDRESS_MODE_WRAP = 0,
    CU_TR_ADDRESS_MODE_CLAMP = 1,
    CU_TR_ADDRESS_MODE_MIRROR = 2,
    CU_TR_ADDRESS_MODE_BORDER = 3
} CUaddress_mode;
```

Note that this call has no effect if `hTexRef` is bound to linear memory. Also, if the flag, `CU_TRSF_NORMALIZED_COORDINATES`, is not set, the only supported address mode is `CU_TR_ADDRESS_MODE_CLAMP`.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetArray`,
`cuTexRefSetFilterMode`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`, `cudaBindTexture`, `cudaBindTexture2D`,
`cudaBindTextureToArray`, `cudaBindTextureToMipmappedArray`

CUresult cuTexRefSetArray (CUtexref hTexRef, CUarray hArray, unsigned int Flags)

Binds an array as a texture reference.

Parameters

hTexRef

- Texture reference to bind

hArray

- Array to bind

Flags

- Options (must be `CU_TRSA_OVERRIDE_FORMAT`)

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Binds the CUDA array `hArray` to the texture reference `hTexRef`. Any previous address or CUDA array state associated with the texture reference is superseded by this function. `Flags` must be set to `CU_TRSA_OVERRIDE_FORMAT`. Any CUDA array previously bound to `hTexRef` is unbound.

See also:

[cuTexRefSetAddress](#), [cuTexRefSetAddress2D](#), [cuTexRefSetAddressMode](#),
[cuTexRefSetFilterMode](#), [cuTexRefSetFlags](#), [cuTexRefSetFormat](#), [cuTexRefGetAddress](#),
[cuTexRefGetAddressMode](#), [cuTexRefGetArray](#), [cuTexRefGetFilterMode](#),
[cuTexRefGetFlags](#), [cuTexRefGetFormat](#), [cudaBindTextureToArray](#)

CUresult cuTexRefSetBorderColor (CUtexref hTexRef, float *pBorderColor)

Sets the border color for a texture reference.

Parameters**hTexRef**

- Texture reference

pBorderColor

- RGBA color

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Specifies the value of the RGBA color via the `pBorderColor` to the texture reference `hTexRef`. The color value supports only float type and holds color components in the following sequence: `pBorderColor[0]` holds 'R' component `pBorderColor[1]` holds 'G' component `pBorderColor[2]` holds 'B' component `pBorderColor[3]` holds 'A' component

Note that the color values can be set only when the Address mode is set to `CU_TR_ADDRESS_MODE_BORDER` using [cuTexRefSetAddressMode](#). Applications using integer border color values have to "reinterpret_cast" their values to float.

See also:

[cuTexRefSetAddressMode](#), [cuTexRefGetAddressMode](#), [cuTexRefGetBorderColor](#),
[cudaBindTexture](#), [cudaBindTexture2D](#), [cudaBindTextureToArray](#),
[cudaBindTextureToMipmappedArray](#)

CUresult cuTexRefSetFilterMode (CUtexref hTexRef, CUfilter_mode fm)

Sets the filtering mode for a texture reference.

Parameters

hTexRef

- Texture reference

fm

- Filtering mode to set

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Specifies the filtering mode `fm` to be used when reading memory through the texture reference `hTexRef`. `CUfilter_mode_enum` is defined as:

```
↑ typedef enum CUfilter_mode_enum {
    CU_TR_FILTER_MODE_POINT = 0,
    CU_TR_FILTER_MODE_LINEAR = 1
} CUfilter_mode;
```

Note that this call has no effect if `hTexRef` is bound to linear memory.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`, `cudaBindTextureToArray`

CUresult cuTexRefSetFlags (CUtexref hTexRef, unsigned int Flags)

Sets the flags for a texture reference.

Parameters

hTexRef

- Texture reference

Flags

- Optional flags to set

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Specifies optional flags via `Flags` to specify the behavior of data returned through the texture reference `hTexRef`. The valid flags are:

- ▶ `CU_TRSF_READ_AS_INTEGER`, which suppresses the default behavior of having the texture promote integer data to floating point data in the range [0, 1]. Note that texture with 32-bit integer format would not be promoted, regardless of whether or not this flag is specified;
- ▶ `CU_TRSF_NORMALIZED_COORDINATES`, which suppresses the default behavior of having the texture coordinates range from [0, Dim) where Dim is the width or height of the CUDA array. Instead, the texture coordinates [0, 1.0) reference the entire breadth of the array dimension;

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFilterMode`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`, `cudaBindTexture`, `cudaBindTexture2D`,
`cudaBindTextureToArray`, `cudaBindTextureToMipmappedArray`

CUresult cuTexRefSetFormat (CUtexref hTexRef, CUarray_format fmt, int NumPackedComponents)

Sets the format for a texture reference.

Parameters**hTexRef**

- Texture reference

fmt

- Format to set

NumPackedComponents

- Number of components per array element

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Specifies the format of the data to be read by the texture reference `hTexRef`. `fmt` and `NumPackedComponents` are exactly analogous to the `Format` and `NumChannels` members of the `CUDA_ARRAY_DESCRIPTOR` structure: They specify the format of each component and the number of components per array element.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`, `cuTexRefSetArray`, `cuTexRefSetFilterMode`, `cuTexRefSetFlags`, `cuTexRefGetAddress`, `cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`, `cuTexRefGetFlags`, `cuTexRefGetFormat`, `cudaCreateChannelDesc`, `cudaBindTexture`, `cudaBindTexture2D`, `cudaBindTextureToArray`, `cudaBindTextureToMipmappedArray`

CUresult cuTexRefSetMaxAnisotropy (CUtexref hTexRef, unsigned int maxAniso)

Sets the maximum anisotropy for a texture reference.

Parameters

`hTexRef`

- Texture reference

`maxAniso`

- Maximum anisotropy

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`, `CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`, `CUDA_ERROR_INVALID_VALUE`

Description

Specifies the maximum anisotropy `maxAniso` to be used when reading memory through the texture reference `hTexRef`.

Note that this call has no effect if `hTexRef` is bound to linear memory.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`, `cuTexRefSetArray`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`, `cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`, `cuTexRefGetFlags`, `cuTexRefGetFormat`, `cudaBindTextureToArray`, `cudaBindTextureToMipmappedArray`

CUresult cuTexRefSetMipmapFilterMode (CUtexref hTexRef, CUfilter_mode fm)

Sets the mipmap filtering mode for a texture reference.

Parameters

hTexRef

- Texture reference

fm

- Filtering mode to set

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Specifies the mipmap filtering mode `fm` to be used when reading memory through the texture reference `hTexRef`. `CUfilter_mode_enum` is defined as:

```
↑ typedef enum CUfilter_mode_enum {
    CU_TR_FILTER_MODE_POINT = 0,
    CU_TR_FILTER_MODE_LINEAR = 1
} CUfilter_mode;
```

Note that this call has no effect if `hTexRef` is not bound to a mipmapped array.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`, `cudaBindTextureToMipmappedArray`

CUresult cuTexRefSetMipmapLevelBias (CUtexref hTexRef, float bias)

Sets the mipmap level bias for a texture reference.

Parameters

hTexRef

- Texture reference

bias

- Mipmap level bias

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Specifies the mipmap level bias `bias` to be added to the specified mipmap level when reading memory through the texture reference `hTexRef`.

Note that this call has no effect if `hTexRef` is not bound to a mipmapped array.

See also:

`cuTexRefSetAddress`, `cuTexRefSetAddress2D`, `cuTexRefSetAddressMode`,
`cuTexRefSetArray`, `cuTexRefSetFlags`, `cuTexRefSetFormat`, `cuTexRefGetAddress`,
`cuTexRefGetAddressMode`, `cuTexRefGetArray`, `cuTexRefGetFilterMode`,
`cuTexRefGetFlags`, `cuTexRefGetFormat`, `cudaBindTextureToMipmappedArray`

CUresult cuTexRefSetMipmapLevelClamp (CUtexref hTexRef, float minMipmapLevelClamp, float maxMipmapLevelClamp)

Sets the mipmap min/max mipmap level clamps for a texture reference.

Parameters**hTexRef**

- Texture reference

minMipmapLevelClamp

- Mipmap min level clamp

maxMipmapLevelClamp

- Mipmap max level clamp

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Specifies the min/max mipmap level clamps, `minMipmapLevelClamp` and `maxMipmapLevelClamp` respectively, to be used when reading memory through the texture reference `hTexRef`.

Note that this call has no effect if `hTexRef` is not bound to a mipmapped array.

See also:

[cuTexRefSetAddress](#), [cuTexRefSetAddress2D](#), [cuTexRefSetAddressMode](#),
[cuTexRefSetArray](#), [cuTexRefSetFlags](#), [cuTexRefSetFormat](#), [cuTexRefGetAddress](#),
[cuTexRefGetAddressMode](#), [cuTexRefGetArray](#), [cuTexRefGetFilterMode](#),
[cuTexRefGetFlags](#), [cuTexRefGetFormat](#), [cudaBindTextureToMipmappedArray](#)

CUresult cuTexRefSetMipmappedArray (CUtexref hTexRef, CUmipmappedArray hMipmappedArray, unsigned int Flags)

Binds a mipmapped array to a texture reference.

Parameters**hTexRef**

- Texture reference to bind

hMipmappedArray

- Mipmapped array to bind

Flags

- Options (must be [CU_TRSA_OVERRIDE_FORMAT](#))

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Binds the CUDA mipmapped array `hMipmappedArray` to the texture reference `hTexRef`. Any previous address or CUDA array state associated with the texture reference is superseded by this function. `Flags` must be set to [CU_TRSA_OVERRIDE_FORMAT](#). Any CUDA array previously bound to `hTexRef` is unbound.

See also:

[cuTexRefSetAddress](#), [cuTexRefSetAddress2D](#), [cuTexRefSetAddressMode](#),
[cuTexRefSetFilterMode](#), [cuTexRefSetFlags](#), [cuTexRefSetFormat](#), [cuTexRefGetAddress](#),
[cuTexRefGetAddressMode](#), [cuTexRefGetArray](#), [cuTexRefGetFilterMode](#),
[cuTexRefGetFlags](#), [cuTexRefGetFormat](#), [cudaBindTextureToMipmappedArray](#)

4.20. Texture Reference Management [DEPRECATED]

This section describes the deprecated texture reference management functions of the low-level CUDA driver application programming interface.

CUresult cuTexRefCreate (CUtexref *pTexRef)

Creates a texture reference.

Parameters

pTexRef

- Returned texture reference

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Deprecated

Creates a texture reference and returns its handle in *pTexRef. Once created, the application must call `cuTexRefSetArray()` or `cuTexRefSetAddress()` to associate the reference with allocated memory. Other texture reference functions are used to specify the format and interpretation (addressing, filtering, etc.) to be used when the memory is read through this texture reference.

See also:

[cuTexRefDestroy](#)

CUresult cuTexRefDestroy (CUtexref hTexRef)

Destroys a texture reference.

Parameters

hTexRef

- Texture reference to destroy

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Deprecated

Destroys the texture reference specified by `hTexRef`.

See also:

`cuTexRefCreate`

4.21. Surface Reference Management

This section describes the surface reference management functions of the low-level CUDA driver application programming interface.

CUresult cuSurfRefGetArray (CUarray *phArray, CUsurfref hSurfRef)

Passes back the CUDA array bound to a surface reference.

Parameters**phArray**

- Surface reference handle

hSurfRef

- Surface reference handle

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns in `*phArray` the CUDA array bound to the surface reference `hSurfRef`, or returns `CUDA_ERROR_INVALID_VALUE` if the surface reference is not bound to any CUDA array.

See also:

[cuModuleGetSurfRef](#), [cuSurfRefSetArray](#)

CUresult cuSurfRefSetArray (CUsurfref hSurfRef, CUarray hArray, unsigned int Flags)

Sets the CUDA array for a surface reference.

Parameters

hSurfRef

- Surface reference handle

hArray

- CUDA array handle

Flags

- set to 0

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Sets the CUDA array `hArray` to be read and written by the surface reference `hSurfRef`. Any previous CUDA array state associated with the surface reference is superseded by this function. `Flags` must be set to 0. The [CUDA_ARRAY3D_SURFACE_LDST](#) flag must have been set for the CUDA array. Any CUDA array previously bound to `hSurfRef` is unbound.

See also:

[cuModuleGetSurfRef](#), [cuSurfRefGetArray](#), [cudaBindSurfaceToArray](#)

4.22. Texture Object Management

This section describes the texture object management functions of the low-level CUDA driver application programming interface. The texture object API is only supported on devices of compute capability 3.0 or higher.

CUresult cuTexObjectCreate (CUtexObject *pTexObject, const CUDA_RESOURCE_DESC *pResDesc,

```
const CUDA_TEXTURE_DESC *pTexDesc, const
CUDA_RESOURCE_VIEW_DESC *pResViewDesc)
```

Creates a texture object.

Parameters

pTexObject

- Texture object to create

pResDesc

- Resource descriptor

pTexDesc

- Texture descriptor

pResViewDesc

- Resource view descriptor

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Creates a texture object and returns it in `pTexObject`. `pResDesc` describes the data to texture from. `pTexDesc` describes how the data should be sampled. `pResViewDesc` is an optional argument that specifies an alternate format for the data described by `pResDesc`, and also describes the subresource region to restrict access to when texturing. `pResViewDesc` can only be specified if the type of resource is a CUDA array or a CUDA mipmapped array.

Texture objects are only supported on devices of compute capability 3.0 or higher. Additionally, a texture object is an opaque value, and, as such, should only be accessed through CUDA API calls.

The `CUDA_RESOURCE_DESC` structure is defined as:

```

↑ typedef struct CUDA_RESOURCE_DESC_st
  {
    CUresourcetype resType;

    union {
      struct {
        CUarray hArray;
      } array;
      struct {
        CUmipmappedArray hMipmappedArray;
      } mipmap;
      struct {
        CUdeviceptr devPtr;
        CUarray_format format;
        unsigned int numChannels;
        size_t sizeInBytes;
      } linear;
      struct {
        CUdeviceptr devPtr;
        CUarray_format format;
        unsigned int numChannels;
        size_t width;
        size_t height;
        size_t pitchInBytes;
      } pitch2D;
    } res;

    unsigned int flags;
  } CUDA_RESOURCE_DESC;

```

where:

- ▶ `CUDA_RESOURCE_DESC::resType` specifies the type of resource to texture from. `CUresourcetype` is defined as:

```

↑ typedef enum CUresourcetype_enum {
    CU_RESOURCE_TYPE_ARRAY = 0x00,
    CU_RESOURCE_TYPE_MIPMAPPED_ARRAY = 0x01,
    CU_RESOURCE_TYPE_LINEAR = 0x02,
    CU_RESOURCE_TYPE_PITCH2D = 0x03
} CUresourcetype;

```

If `CUDA_RESOURCE_DESC::resType` is set to `CU_RESOURCE_TYPE_ARRAY`, `CUDA_RESOURCE_DESC::res::array::hArray` must be set to a valid CUDA array handle.

If `CUDA_RESOURCE_DESC::resType` is set to `CU_RESOURCE_TYPE_MIPMAPPED_ARRAY`, `CUDA_RESOURCE_DESC::res::mipmap::hMipmappedArray` must be set to a valid CUDA mipmapped array handle.

If `CUDA_RESOURCE_DESC::resType` is set to `CU_RESOURCE_TYPE_LINEAR`, `CUDA_RESOURCE_DESC::res::linear::devPtr` must be set to a valid device pointer, that is aligned to `CU_DEVICE_ATTRIBUTE_TEXTURE_ALIGNMENT`. `CUDA_RESOURCE_DESC::res::linear::format` and `CUDA_RESOURCE_DESC::res::linear::numChannels` describe the format of each component and the number of components per array element.

`CUDA_RESOURCE_DESC::res::linear::sizeInBytes` specifies the size of the array in bytes. The total number of elements in the linear address range cannot exceed `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE1D_LINEAR_WIDTH`. The number of elements is computed as $(\text{sizeInBytes} / (\text{sizeof}(\text{format}) * \text{numChannels}))$.

If `CUDA_RESOURCE_DESC::resType` is set to `CU_RESOURCE_TYPE_PITCH2D`, `CUDA_RESOURCE_DESC::res::pitch2D::devPtr` must be set to a valid device pointer, that is aligned to `CU_DEVICE_ATTRIBUTE_TEXTURE_ALIGNMENT`. `CUDA_RESOURCE_DESC::res::pitch2D::format` and `CUDA_RESOURCE_DESC::res::pitch2D::numChannels` describe the format of each component and the number of components per array element. `CUDA_RESOURCE_DESC::res::pitch2D::width` and `CUDA_RESOURCE_DESC::res::pitch2D::height` specify the width and height of the array in elements, and cannot exceed `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_WIDTH` and `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_HEIGHT` respectively. `CUDA_RESOURCE_DESC::res::pitch2D::pitchInBytes` specifies the pitch between two rows in bytes and has to be aligned to `CU_DEVICE_ATTRIBUTE_TEXTURE_PITCH_ALIGNMENT`. Pitch cannot exceed `CU_DEVICE_ATTRIBUTE_MAXIMUM_TEXTURE2D_LINEAR_PITCH`.

- ▶ flags must be set to zero.

The `CUDA_TEXTURE_DESC` struct is defined as

```
typedef struct CUDA_TEXTURE_DESC_st {
    CUaddress_mode addressMode[3];
    CUfilter_mode filterMode;
    unsigned int flags;
    unsigned int maxAnisotropy;
    CUfilter_mode mipmapFilterMode;
    float mipmapLevelBias;
    float minMipmapLevelClamp;
    float maxMipmapLevelClamp;
} CUDA_TEXTURE_DESC;
```

where

- ▶ `CUDA_TEXTURE_DESC::addressMode` specifies the addressing mode for each dimension of the texture data. `CUaddress_mode` is defined as:

```
typedef enum CUaddress_mode_enum {
    CU_TR_ADDRESS_MODE_WRAP = 0,
    CU_TR_ADDRESS_MODE_CLAMP = 1,
    CU_TR_ADDRESS_MODE_MIRROR = 2,
    CU_TR_ADDRESS_MODE_BORDER = 3
} CUaddress_mode;
```

This is ignored if `CUDA_RESOURCE_DESC::resType` is `CU_RESOURCE_TYPE_LINEAR`. Also, if the flag, `CU_TRSF_NORMALIZED_COORDINATES` is not set, the only supported address mode is `CU_TR_ADDRESS_MODE_CLAMP`.

- ▶ `CUDA_TEXTURE_DESC::filterMode` specifies the filtering mode to be used when fetching from the texture. `CUfilter_mode` is defined as:

```
↑
    typedef enum CUfilter_mode_enum {
        CU_TR_FILTER_MODE_POINT = 0,
        CU_TR_FILTER_MODE_LINEAR = 1
    } CUfilter_mode;
```

This is ignored if `CUDA_RESOURCE_DESC::resType` is `CU_RESOURCE_TYPE_LINEAR`.

- ▶ `CUDA_TEXTURE_DESC::flags` can be any combination of the following:
 - ▶ `CU_TRSF_READ_AS_INTEGER`, which suppresses the default behavior of having the texture promote integer data to floating point data in the range [0, 1]. Note that texture with 32-bit integer format would not be promoted, regardless of whether or not this flag is specified.
 - ▶ `CU_TRSF_NORMALIZED_COORDINATES`, which suppresses the default behavior of having the texture coordinates range from [0, Dim) where Dim is the width or height of the CUDA array. Instead, the texture coordinates [0, 1.0) reference the entire breadth of the array dimension; Note that for CUDA mipmapped arrays, this flag has to be set.
- ▶ `CUDA_TEXTURE_DESC::maxAnisotropy` specifies the maximum anisotropy ratio to be used when doing anisotropic filtering. This value will be clamped to the range [1,16].
- ▶ `CUDA_TEXTURE_DESC::mipmapFilterMode` specifies the filter mode when the calculated mipmap level lies between two defined mipmap levels.
- ▶ `CUDA_TEXTURE_DESC::mipmapLevelBias` specifies the offset to be applied to the calculated mipmap level.
- ▶ `CUDA_TEXTURE_DESC::minMipmapLevelClamp` specifies the lower end of the mipmap level range to clamp access to.
- ▶ `CUDA_TEXTURE_DESC::maxMipmapLevelClamp` specifies the upper end of the mipmap level range to clamp access to.

The `CUDA_RESOURCE_VIEW_DESC` struct is defined as

```
↑
    typedef struct CUDA_RESOURCE_VIEW_DESC_st
    {
        CUresourceViewFormat format;
        size_t width;
        size_t height;
        size_t depth;
        unsigned int firstMipmapLevel;
        unsigned int lastMipmapLevel;
        unsigned int firstLayer;
        unsigned int lastLayer;
    } CUDA_RESOURCE_VIEW_DESC;
```

where:

- ▶ `CUDA_RESOURCE_VIEW_DESC::format` specifies how the data contained in the CUDA array or CUDA mipmapped array should be interpreted. Note that this can incur a change in size of the texture data. If the resource view format is a block compressed format, then the underlying CUDA array or CUDA mipmapped array has to have a base of format `CU_AD_FORMAT_UNSIGNED_INT32` with 2 or 4 channels, depending on the block compressed format. For ex., BC1 and BC4 require the underlying CUDA array to have a format of `CU_AD_FORMAT_UNSIGNED_INT32` with 2 channels. The other BC formats require the underlying resource to have the same base format but with 4 channels.
- ▶ `CUDA_RESOURCE_VIEW_DESC::width` specifies the new width of the texture data. If the resource view format is a block compressed format, this value has to be 4 times the original width of the resource. For non block compressed formats, this value has to be equal to that of the original resource.
- ▶ `CUDA_RESOURCE_VIEW_DESC::height` specifies the new height of the texture data. If the resource view format is a block compressed format, this value has to be 4 times the original height of the resource. For non block compressed formats, this value has to be equal to that of the original resource.
- ▶ `CUDA_RESOURCE_VIEW_DESC::depth` specifies the new depth of the texture data. This value has to be equal to that of the original resource.
- ▶ `CUDA_RESOURCE_VIEW_DESC::firstMipmapLevel` specifies the most detailed mipmap level. This will be the new mipmap level zero. For non-mipmapped resources, this value has to be zero. `CUDA_TEXTURE_DESC::minMipmapLevelClamp` and `CUDA_TEXTURE_DESC::maxMipmapLevelClamp` will be relative to this value. For ex., if the `firstMipmapLevel` is set to 2, and a `minMipmapLevelClamp` of 1.2 is specified, then the actual minimum mipmap level clamp will be 3.2.
- ▶ `CUDA_RESOURCE_VIEW_DESC::lastMipmapLevel` specifies the least detailed mipmap level. For non-mipmapped resources, this value has to be zero.
- ▶ `CUDA_RESOURCE_VIEW_DESC::firstLayer` specifies the first layer index for layered textures. This will be the new layer zero. For non-layered resources, this value has to be zero.
- ▶ `CUDA_RESOURCE_VIEW_DESC::lastLayer` specifies the last layer index for layered textures. For non-layered resources, this value has to be zero.

See also:

`cuTexObjectDestroy`, `cudaCreateTextureObject`

CUresult cuTexObjectDestroy (CUtexObject texObject)

Destroys a texture object.

Parameters

texObject

- Texture object to destroy

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Destroys the texture object specified by `texObject`.

See also:

`cuTexObjectCreate`, `cudaDestroyTextureObject`

CUresult cuTexObjectGetResourceDesc (CUDA_RESOURCE_DESC *pResDesc, CUtexObject texObject)

Returns a texture object's resource descriptor.

Parameters

pResDesc

- Resource descriptor

texObject

- Texture object

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Returns the resource descriptor for the texture object specified by `texObject`.

See also:

[cuTexObjectCreate](#), [cudaGetTextureObjectResourceDesc](#),

CUresult cuTexObjectGetResourceViewDesc (CUDA_RESOURCE_VIEW_DESC *pResViewDesc, CUtexObject texObject)

Returns a texture object's resource view descriptor.

Parameters**pResViewDesc**

- Resource view descriptor

texObject

- Texture object

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Returns the resource view descriptor for the texture object specified by `texObject`. If no resource view was set for `texObject`, the `CUDA_ERROR_INVALID_VALUE` is returned.

See also:

[cuTexObjectCreate](#), [cudaGetTextureObjectResourceViewDesc](#)

CUresult cuTexObjectGetTextureDesc (CUDA_TEXTURE_DESC *pTexDesc, CUtexObject texObject)

Returns a texture object's texture descriptor.

Parameters**pTexDesc**

- Texture descriptor

texObject

- Texture object

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns the texture descriptor for the texture object specified by `texObject`.

See also:

`cuTexObjectCreate`, `cudaGetTextureObjectTextureDesc`

4.23. Surface Object Management

This section describes the surface object management functions of the low-level CUDA driver application programming interface. The surface object API is only supported on devices of compute capability 3.0 or higher.

**CUresult cuSurfObjectCreate (CUSurfObject
 *pSurfObject, const CUDA_RESOURCE_DESC *pResDesc)**

Creates a surface object.

Parameters**pSurfObject**

- Surface object to create

pResDesc

- Resource descriptor

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Creates a surface object and returns it in `pSurfObject`. `pResDesc` describes the data to perform surface load/stores on. `CUDA_RESOURCE_DESC::resType` must be `CU_RESOURCE_TYPE_ARRAY` and `CUDA_RESOURCE_DESC::res::array::hArray` must be set to a valid CUDA array handle. `CUDA_RESOURCE_DESC::flags` must be set to zero.

Surface objects are only supported on devices of compute capability 3.0 or higher. Additionally, a surface object is an opaque value, and, as such, should only be accessed through CUDA API calls.

See also:

[cuSurfObjectDestroy](#), [cudaCreateSurfaceObject](#)

CUresult cuSurfObjectDestroy (CUsurfObject surfObject)

Destroys a surface object.

Parameters

surfObject

- Surface object to destroy

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Destroys the surface object specified by `surfObject`.

See also:

[cuSurfObjectCreate](#), [cudaDestroySurfaceObject](#)

CUresult cuSurfObjectGetResourceDesc (CUDA_RESOURCE_DESC *pResDesc, CUsurfObject surfObject)

Returns a surface object's resource descriptor.

Parameters

pResDesc

- Resource descriptor

surfObject

- Surface object

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Returns the resource descriptor for the surface object specified by `surfObject`.

See also:

`cuSurfObjectCreate`, `cudaGetSurfaceObjectResourceDesc`

4.24. Peer Context Memory Access

This section describes the direct peer context memory access functions of the low-level CUDA driver application programming interface.

CUresult cuCtxDisablePeerAccess (CUcontext peerContext)

Disables direct access to memory allocations in a peer context and unregisters any registered allocations.

Parameters**peerContext**

- Peer context to disable direct access to

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED,
 CUDA_ERROR_PEER_ACCESS_NOT_ENABLED,
 CUDA_ERROR_INVALID_CONTEXT,

Description

Returns `CUDA_ERROR_PEER_ACCESS_NOT_ENABLED` if direct peer access has not yet been enabled from `peerContext` to the current context.

Returns `CUDA_ERROR_INVALID_CONTEXT` if there is no current context, or if `peerContext` is not a valid context.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuDeviceCanAccessPeer](#), [cuCtxEnablePeerAccess](#), [cudaDeviceDisablePeerAccess](#)

CUresult cuCtxEnablePeerAccess (CUcontext peerContext, unsigned int Flags)

Enables direct access to memory allocations in a peer context.

Parameters

peerContext

- Peer context to enable direct access to from the current context

Flags

- Reserved for future use and must be set to 0

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED,
 CUDA_ERROR_PEER_ACCESS_ALREADY_ENABLED,
 CUDA_ERROR_TOO_MANY_PEERS, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_PEER_ACCESS_UNSUPPORTED, CUDA_ERROR_INVALID_VALUE

Description

If both the current context and `peerContext` are on devices which support unified addressing (as may be queried using [CU_DEVICE_ATTRIBUTE_UNIFIED_ADDRESSING](#)) and same major compute capability, then on success all allocations from `peerContext` will immediately be accessible by the current context. See [Unified Addressing](#) for additional details.

Note that access granted by this call is unidirectional and that in order to access memory from the current context in `peerContext`, a separate symmetric call to [cuCtxEnablePeerAccess\(\)](#) is required.

There is a system-wide maximum of eight peer connections per device.

Returns [CUDA_ERROR_PEER_ACCESS_UNSUPPORTED](#) if [cuDeviceCanAccessPeer\(\)](#) indicates that the [CUdevice](#) of the current context cannot directly access memory from the [CUdevice](#) of `peerContext`.

Returns `CUDA_ERROR_PEER_ACCESS_ALREADY_ENABLED` if direct access of `peerContext` from the current context has already been enabled.

Returns `CUDA_ERROR_TOO_MANY_PEERS` if direct peer access is not possible because hardware resources required for peer access have been exhausted.

Returns `CUDA_ERROR_INVALID_CONTEXT` if there is no current context, `peerContext` is not a valid context, or if the current context is `peerContext`.

Returns `CUDA_ERROR_INVALID_VALUE` if `Flags` is not 0.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuDeviceCanAccessPeer`, `cuCtxDisablePeerAccess`, `cudaDeviceEnablePeerAccess`

CUresult cuDeviceCanAccessPeer (int *canAccessPeer, CUdevice dev, CUdevice peerDev)

Queries if a device may directly access a peer device's memory.

Parameters

`canAccessPeer`

- Returned access capability

`dev`

- Device from which allocations on `peerDev` are to be directly accessed.

`peerDev`

- Device on which the allocations to be directly accessed by `dev` reside.

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_DEVICE`

Description

Returns in `*canAccessPeer` a value of 1 if contexts on `dev` are capable of directly accessing memory from contexts on `peerDev` and 0 otherwise. If direct access of `peerDev` from `dev` is possible, then access may be enabled on two specific contexts by calling `cuCtxEnablePeerAccess()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxEnablePeerAccess](#), [cuCtxDisablePeerAccess](#), [cudaDeviceCanAccessPeer](#)

CUresult cuDeviceGetP2PAttribute (int *value, CUdevice_P2PAttribute attrib, CUdevice srcDevice, CUdevice dstDevice)

Queries attributes of the link between two devices.

Parameters

value

- Returned value of the requested attribute

attrib

- The requested attribute of the link between `srcDevice` and `dstDevice`.

srcDevice

- The source device of the target link.

dstDevice

- The destination device of the target link.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_DEVICE](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Returns in `*value` the value of the requested attribute `attrib` of the link between `srcDevice` and `dstDevice`. The supported attributes are:

- ▶ [CU_DEVICE_P2P_ATTRIBUTE_PERFORMANCE_RANK](#): A relative value indicating the performance of the link between two devices.
- ▶ [CU_DEVICE_P2P_ATTRIBUTE_ACCESS_SUPPORTED](#) P2P: 1 if P2P Access is enable.
- ▶ [CU_DEVICE_P2P_ATTRIBUTE_NATIVE_ATOMIC_SUPPORTED](#): 1 if Atomic operations over the link are supported.
- ▶ [CU_DEVICE_P2P_ATTRIBUTE_CUDA_ARRAY_ACCESS_SUPPORTED](#): 1 if `cudaArray` can be accessed over the link.

Returns `CUDA_ERROR_INVALID_DEVICE` if `srcDevice` or `dstDevice` are not valid or if they represent the same device.

Returns `CUDA_ERROR_INVALID_VALUE` if `attrib` is not valid or if `value` is a null pointer.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxEnablePeerAccess`, `cuCtxDisablePeerAccess`, `cuDeviceCanAccessPeer`, `cudaDeviceGetP2PAttribute`

4.25. Graphics Interoperability

This section describes the graphics interoperability functions of the low-level CUDA driver application programming interface.

`CUresult cuGraphicsMapResources (unsigned int count, CUgraphicsResource *resources, CUstream hStream)`

Map graphics resources for access by CUDA.

Parameters

count

- Number of resources to map

resources

- Resources to map for CUDA usage

hStream

- Stream with which to synchronize

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_HANDLE`, `CUDA_ERROR_ALREADY_MAPPED`,
`CUDA_ERROR_UNKNOWN`

Description

Maps the `count` graphics resources in `resources` for access by CUDA.

The resources in `resources` may be accessed by CUDA until they are unmapped. The graphics API from which `resources` were registered should not access any resources while they are mapped by CUDA. If an application does so, the results are undefined.

This function provides the synchronization guarantee that any graphics calls issued before `cuGraphicsMapResources()` will complete before any subsequent CUDA work issued in `stream` begins.

If `resources` includes any duplicate entries then `CUDA_ERROR_INVALID_HANDLE` is returned. If any of `resources` are presently mapped for access by CUDA then `CUDA_ERROR_ALREADY_MAPPED` is returned.



- ▶ This function uses standard `default stream` semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsResourceGetMappedPointer`, `cuGraphicsSubResourceGetMappedArray`,
`cuGraphicsUnmapResources`, `cudaGraphicsMapResources`

CUresult

`cuGraphicsResourceGetMappedMipmappedArray` (`CUmipmappedArray *pMipmappedArray`, `CUgraphicsResource resource`)

Get a mipmapped array through which to access a mapped graphics resource.

Parameters

`pMipmappedArray`

- Returned mipmapped array through which `resource` may be accessed

`resource`

- Mapped resource to access

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_NOT_MAPPED`, `CUDA_ERROR_NOT_MAPPED_AS_ARRAY`

Description

Returns in `*pMipmappedArray` a mipmapped array through which the mapped graphics resource `resource`. The value set in `*pMipmappedArray` may change every time that `resource` is mapped.

If `resource` is not a texture then it cannot be accessed via a mipmapped array and `CUDA_ERROR_NOT_MAPPED_AS_ARRAY` is returned. If `resource` is not mapped then `CUDA_ERROR_NOT_MAPPED` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsResourceGetMappedPointer`,
`cudaGraphicsResourceGetMappedMipmappedArray`

CUresult cuGraphicsResourceGetMappedPointer (CUdeviceptr *pDevPtr, size_t *pSize, CUgraphicsResource resource)

Get a device pointer through which to access a mapped graphics resource.

Parameters

`pDevPtr`

- Returned pointer through which `resource` may be accessed

`pSize`

- Returned size of the buffer accessible starting at `*pPointer`

`resource`

- Mapped resource to access

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_NOT_MAPPED`, `CUDA_ERROR_NOT_MAPPED_AS_POINTER`

Description

Returns in `*pDevPtr` a pointer through which the mapped graphics resource `resource` may be accessed. Returns in `pSize` the size of the memory in bytes which

may be accessed from that pointer. The value set in `pPointer` may change every time that resource is mapped.

If `resource` is not a buffer then it cannot be accessed via a pointer and `CUDA_ERROR_NOT_MAPPED_AS_POINTER` is returned. If `resource` is not mapped then `CUDA_ERROR_NOT_MAPPED` is returned. *



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsMapResources`, `cuGraphicsSubResourceGetMappedArray`,
`cudaGraphicsResourceGetMappedPointer`

CUresult cuGraphicsResourceSetMapFlags (CUgraphicsResource resource, unsigned int flags)

Set usage flags for mapping a graphics resource.

Parameters

resource

- Registered resource to set flags for

flags

- Parameters for resource mapping

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_ALREADY_MAPPED`

Description

Set flags for mapping the graphics resource `resource`.

Changes to `flags` will take effect the next time `resource` is mapped. The `flags` argument may be any of the following:

- ▶ `CU_GRAPHICS_MAP_RESOURCE_FLAGS_NONE`: Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA kernels. This is the default value.
- ▶ `CU_GRAPHICS_MAP_RESOURCE_FLAGS_READONLY`: Specifies that CUDA kernels which access this resource will not write to this resource.

- ▶ `CUDA_GRAPHICS_MAP_RESOURCE_FLAGS_WRITEDISCARD`: Specifies that CUDA kernels which access this resource will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.

If `resource` is presently mapped for access by CUDA then `CUDA_ERROR_ALREADY_MAPPED` is returned. If `flags` is not one of the above values then `CUDA_ERROR_INVALID_VALUE` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsMapResources`, `cudaGraphicsResourceSetMapFlags`

CUresult cuGraphicsSubResourceGetMappedArray (CUarray *pArray, CUgraphicsResource resource, unsigned int arrayIndex, unsigned int mipLevel)

Get an array through which to access a subresource of a mapped graphics resource.

Parameters

pArray

- Returned array through which a subresource of `resource` may be accessed

resource

- Mapped resource to access

arrayIndex

- Array index for array textures or cubemap face index as defined by `CUarray_cubemap_face` for cubemap textures for the subresource to access

mipLevel

- Mipmap level for the subresource to access

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_NOT_MAPPED`, `CUDA_ERROR_NOT_MAPPED_AS_ARRAY`

Description

Returns in `*pArray` an array through which the subresource of the mapped graphics resource `resource` which corresponds to array index `arrayIndex` and mipmap level

`mipLevel` may be accessed. The value set in `*pArray` may change every time that resource is mapped.

If `resource` is not a texture then it cannot be accessed via an array and `CUDA_ERROR_NOT_MAPPED_AS_ARRAY` is returned. If `arrayIndex` is not a valid array index for `resource` then `CUDA_ERROR_INVALID_VALUE` is returned. If `mipLevel` is not a valid mipmap level for `resource` then `CUDA_ERROR_INVALID_VALUE` is returned. If `resource` is not mapped then `CUDA_ERROR_NOT_MAPPED` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsResourceGetMappedPointer`, `cudaGraphicsSubResourceGetMappedArray`

CUresult cuGraphicsUnmapResources (unsigned int count, CUgraphicsResource *resources, CUstream hStream)

Unmap graphics resources.

Parameters

count

- Number of resources to unmap

resources

- Resources to unmap

hStream

- Stream with which to synchronize

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_HANDLE`, `CUDA_ERROR_NOT_MAPPED`,
`CUDA_ERROR_UNKNOWN`

Description

Unmaps the `count` graphics resources in `resources`.

Once unmapped, the resources in `resources` may not be accessed by CUDA until they are mapped again.

This function provides the synchronization guarantee that any CUDA work issued in `stream` before `cuGraphicsUnmapResources()` will complete before any subsequently issued graphics work begins.

If `resources` includes any duplicate entries then `CUDA_ERROR_INVALID_HANDLE` is returned. If any of `resources` are not presently mapped for access by CUDA then `CUDA_ERROR_NOT_MAPPED` is returned.



- ▶ This function uses standard `default stream` semantics.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsMapResources`, `cudaGraphicsUnmapResources`

CUresult cuGraphicsUnregisterResource (CUgraphicsResource resource)

Unregisters a graphics resource for access by CUDA.

Parameters

resource

- Resource to unregister

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_HANDLE`, `CUDA_ERROR_UNKNOWN`

Description

Unregisters the graphics resource `resource` so it is not accessible by CUDA unless registered again.

If `resource` is invalid then `CUDA_ERROR_INVALID_HANDLE` is returned.



- Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsD3D9RegisterResource`, `cuGraphicsD3D10RegisterResource`,
`cuGraphicsD3D11RegisterResource`, `cuGraphicsGLRegisterBuffer`,
`cuGraphicsGLRegisterImage`, `cudaGraphicsUnregisterResource`

4.26. Profiler Control

This section describes the profiler control functions of the low-level CUDA driver application programming interface.

`CUresult cuProfilerInitialize (const char *configFile, const char *outputFile, CUoutput_mode outputMode)`

Initialize the profiling.

Parameters

`configFile`

- Name of the config file that lists the counters/options for profiling.

`outputFile`

- Name of the outputFile where the profiling results will be stored.

`outputMode`

- `outputMode`, can be `CU_OUT_KEY_VALUE_PAIR` or `CU_OUT_CSV`.

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_PROFILER_DISABLED`

Description

Using this API user can initialize the CUDA profiler by specifying the configuration file, output file and output file format. This API is generally used to profile different set of counters by looping the kernel launch. The `configFile` parameter can be used to select profiling options including profiler counters. Refer to the "Compute Command Line Profiler User Guide" for supported profiler options and counters.

Limitation: The CUDA profiler cannot be initialized with this API if another profiling tool is already active, as indicated by the `CUDA_ERROR_PROFILER_DISABLED` return code.

Typical usage of the profiling APIs is as follows:

```
for each set of counters/options { cuProfilerInitialize(); //Initialize profiling, set
the counters or options in the config file ... cuProfilerStart(); // code to be profiled
cuProfilerStop(); ... cuProfilerStart(); // code to be profiled cuProfilerStop(); ... }
```



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuProfilerStart](#), [cuProfilerStop](#), [cudaProfilerInitialize](#)

CUresult cuProfilerStart (void)

Enable profiling.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_CONTEXT](#)

Description

Enables profile collection by the active profiling tool for the current context. If profiling is already enabled, then [cuProfilerStart\(\)](#) has no effect.

[cuProfilerStart](#) and [cuProfilerStop](#) APIs are used to programmatically control the profiling granularity by allowing profiling to be done only on selective pieces of code.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuProfilerInitialize](#), [cuProfilerStop](#), [cudaProfilerStart](#)

CUresult cuProfilerStop (void)

Disable profiling.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_CONTEXT](#)

Description

Disables profile collection by the active profiling tool for the current context. If profiling is already disabled, then [cuProfilerStop\(\)](#) has no effect.

[cuProfilerStart](#) and [cuProfilerStop](#) APIs are used to programmatically control the profiling granularity by allowing profiling to be done only on selective pieces of code.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuProfilerInitialize](#), [cuProfilerStart](#), [cudaProfilerStop](#)

4.27. OpenGL Interoperability

This section describes the OpenGL interoperability functions of the low-level CUDA driver application programming interface. Note that mapping of OpenGL resources is performed with the graphics API agnostic, resource mapping interface described in [Graphics Interoperability](#).

OpenGL Interoperability [DEPRECATED]

enum CUGLDeviceList

CUDA devices corresponding to an OpenGL device

Values

CU_GL_DEVICE_LIST_ALL = 0x01

The CUDA devices for all GPUs used by the current OpenGL context

CU_GL_DEVICE_LIST_CURRENT_FRAME = 0x02

The CUDA devices for the GPUs used by the current OpenGL context in its currently rendering frame

CU_GL_DEVICE_LIST_NEXT_FRAME = 0x03

The CUDA devices for the GPUs to be used by the current OpenGL context in the next frame

CUresult cuGLGetDevices (unsigned int *pCudaDeviceCount, CUdevice *pCudaDevices, unsigned int cudaDeviceCount, CUGLDeviceList deviceList)

Gets the CUDA devices associated with the current OpenGL context.

Parameters

pCudaDeviceCount

- Returned number of CUDA devices.

pCudaDevices

- Returned CUDA devices.

cudaDeviceCount

- The size of the output device array pCudaDevices.

deviceList

- The set of devices to return.

Returns

CUDA_SUCCESS, CUDA_ERROR_NO_DEVICE,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_GRAPHICS_CONTEXT

Description

Returns in *pCudaDeviceCount the number of CUDA-compatible devices corresponding to the current OpenGL context. Also returns in *pCudaDevices at most cudaDeviceCount of the CUDA-compatible devices corresponding to the current OpenGL context. If any of the GPUs being used by the current OpenGL context are not CUDA capable then the call will return CUDA_ERROR_NO_DEVICE.

The deviceList argument may be any of the following:

- ▶ **CU_GL_DEVICE_LIST_ALL**: Query all devices used by the current OpenGL context.
- ▶ **CU_GL_DEVICE_LIST_CURRENT_FRAME**: Query the devices used by the current OpenGL context to render the current frame (in SLI).
- ▶ **CU_GL_DEVICE_LIST_NEXT_FRAME**: Query the devices used by the current OpenGL context to render the next frame (in SLI). Note that this is a prediction, it can't be guaranteed that this is correct in all cases.



- ▶ This function is not supported on Mac OS X.
- ▶ Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuWGLGetDevice](#), [cudaGLGetDevices](#)

CUresult cuGraphicsGLRegisterBuffer (CUgraphicsResource *pCudaResource, GLuint buffer, unsigned int Flags)

Registers an OpenGL buffer object.

Parameters

pCudaResource

- Pointer to the returned object handle

buffer

- name of buffer object to be registered

Flags

- Register flags

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_ALREADY_MAPPED, CUDA_ERROR_INVALID_CONTEXT,

Description

Registers the buffer object specified by `buffer` for access by CUDA. A handle to the registered object is returned as `pCudaResource`. The register flags `Flags` specify the intended usage, as follows:

- ▶ `CU_GRAPHICS_REGISTER_FLAGS_NONE`: Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA. This is the default value.
- ▶ `CU_GRAPHICS_REGISTER_FLAGS_READ_ONLY`: Specifies that CUDA will not write to this resource.
- ▶ `CU_GRAPHICS_REGISTER_FLAGS_WRITE_DISCARD`: Specifies that CUDA will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsUnregisterResource`, `cuGraphicsMapResources`,
`cuGraphicsResourceGetMappedPointer`, `cudaGraphicsGLRegisterBuffer`

CUresult cuGraphicsGLRegisterImage (CUgraphicsResource *pCudaResource, GLuint image, GLenum target, unsigned int Flags)

Register an OpenGL texture or renderbuffer object.

Parameters

pCudaResource

- Pointer to the returned object handle

image

- name of texture or renderbuffer object to be registered

target

- Identifies the type of object specified by image

Flags

- Register flags

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_ALREADY_MAPPED, CUDA_ERROR_INVALID_CONTEXT,

Description

Registers the texture or renderbuffer object specified by image for access by CUDA. A handle to the registered object is returned as pCudaResource.

target must match the type of the object, and must be one of GL_TEXTURE_2D, GL_TEXTURE_RECTANGLE, GL_TEXTURE_CUBE_MAP, GL_TEXTURE_3D, GL_TEXTURE_2D_ARRAY, or GL_RENDERBUFFER.

The register flags Flags specify the intended usage, as follows:

- ▶ CU_GRAPHICS_REGISTER_FLAGS_NONE: Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA. This is the default value.
- ▶ CU_GRAPHICS_REGISTER_FLAGS_READ_ONLY: Specifies that CUDA will not write to this resource.
- ▶ CU_GRAPHICS_REGISTER_FLAGS_WRITE_DISCARD: Specifies that CUDA will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.
- ▶ CU_GRAPHICS_REGISTER_FLAGS_SURFACE_LDST: Specifies that CUDA will bind this resource to a surface reference.
- ▶ CU_GRAPHICS_REGISTER_FLAGS_TEXTURE_GATHER: Specifies that CUDA will perform texture gather operations on this resource.

The following image formats are supported. For brevity's sake, the list is abbreviated. For ex., {GL_R, GL_RG} X {8, 16} would expand to the following 4 formats {GL_R8, GL_R16, GL_RG8, GL_RG16} :

- ▶ GL_RED, GL_RG, GL_RGBA, GL_LUMINANCE, GL_ALPHA, GL_LUMINANCE_ALPHA, GL_INTENSITY
- ▶ {GL_R, GL_RG, GL_RGBA} X {8, 16, 16F, 32F, 8UI, 16UI, 32UI, 8I, 16I, 32I}
- ▶ {GL_LUMINANCE, GL_ALPHA, GL_LUMINANCE_ALPHA, GL_INTENSITY} X {8, 16, 16F_ARB, 32F_ARB, 8UI_EXT, 16UI_EXT, 32UI_EXT, 8I_EXT, 16I_EXT, 32I_EXT}

The following image classes are currently disallowed:

- ▶ Textures with borders
- ▶ Multisampled renderbuffers



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsUnregisterResource](#), [cuGraphicsMapResources](#),
[cuGraphicsSubResourceGetMappedArray](#), [cudaGraphicsGLRegisterImage](#)

CUresult cuWGLGetDevice (CUdevice *pDevice, HGPUNV hGpu)

Gets the CUDA device associated with hGpu.

Parameters

pDevice

- Device associated with hGpu

hGpu

- Handle to a GPU, as queried via WGL_NV_gpu_affinity()

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE

Description

Returns in *pDevice the CUDA device associated with a hGpu, if applicable.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGLMapBufferObject`, `cuGLRegisterBufferObject`, `cuGLUnmapBufferObject`,
`cuGLUnregisterBufferObject`, `cuGLUnmapBufferObjectAsync`,
`cuGLSetBufferObjectMapFlags`, `cudaWGLGetDevice`

4.27.1. OpenGL Interoperability [DEPRECATED]

OpenGL Interoperability

This section describes deprecated OpenGL interoperability functionality.

enum `CUGLmap_flags`

Flags to map or unmap a resource

Values

`CU_GL_MAP_RESOURCE_FLAGS_NONE = 0x00`
`CU_GL_MAP_RESOURCE_FLAGS_READ_ONLY = 0x01`
`CU_GL_MAP_RESOURCE_FLAGS_WRITE_DISCARD = 0x02`

`CUresult cuGLCtxCreate (CUcontext *pCtx, unsigned int Flags, CUdevice device)`

Create a CUDA context for interoperability with OpenGL.

Parameters

`pCtx`

- Returned CUDA context

Flags

- Options for CUDA context creation

`device`

- Device on which to create the context

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_OUT_OF_MEMORY`

Description

Deprecated This function is deprecated as of Cuda 5.0.

This function is deprecated and should no longer be used. It is no longer necessary to associate a CUDA context with an OpenGL context in order to achieve maximum interoperability performance.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuGLInit](#), [cuGLMapBufferObject](#), [cuGLRegisterBufferObject](#), [cuGLUnmapBufferObject](#), [cuGLUnregisterBufferObject](#), [cuGLMapBufferObjectAsync](#), [cuGLUnmapBufferObjectAsync](#), [cuGLSetBufferObjectMapFlags](#), [cuWGLGetDevice](#)

CUresult cuGLInit (void)

Initializes OpenGL interoperability.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_UNKNOWN](#)

Description

Deprecated This function is deprecated as of Cuda 3.0.

Initializes OpenGL interoperability. This function is deprecated and calling it is no longer required. It may fail if the needed OpenGL driver facilities are not available.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGLMapBufferObject](#), [cuGLRegisterBufferObject](#), [cuGLUnmapBufferObject](#),
[cuGLUnregisterBufferObject](#), [cuGLMapBufferObjectAsync](#),
[cuGLUnmapBufferObjectAsync](#), [cuGLSetBufferObjectMapFlags](#), [cuWGLGetDevice](#)

CUresult cuGLMapBufferObject (CUdeviceptr *dptr, size_t *size, GLuint buffer)

Maps an OpenGL buffer object.

Parameters

dptr

- Returned mapped base pointer

size

- Returned size of mapping

buffer

- The name of the buffer object to map

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_MAP_FAILED

Description

Deprecated This function is deprecated as of Cuda 3.0.

Maps the buffer object specified by `buffer` into the address space of the current CUDA context and returns in `*dptr` and `*size` the base pointer and size of the resulting mapping.

There must be a valid OpenGL context bound to the current thread when this function is called. This must be the same context, or a member of the same shareGroup, as the context that was bound when the buffer was registered.

All streams in the current CUDA context are synchronized with the current GL context.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsMapResources](#)

CUresult cuGLMapBufferObjectAsync (CUdeviceptr *dptr, size_t *size, GLuint buffer, CUstream hStream)

Maps an OpenGL buffer object.

Parameters

dptr

- Returned mapped base pointer

size

- Returned size of mapping

buffer

- The name of the buffer object to map

hStream

- Stream to synchronize

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_MAP_FAILED

Description

Deprecated This function is deprecated as of Cuda 3.0.

Maps the buffer object specified by `buffer` into the address space of the current CUDA context and returns in `*dptr` and `*size` the base pointer and size of the resulting mapping.

There must be a valid OpenGL context bound to the current thread when this function is called. This must be the same context, or a member of the same `shareGroup`, as the context that was bound when the buffer was registered.

Stream `hStream` in the current CUDA context is synchronized with the current GL context.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsMapResources](#)

CUresult cuGLRegisterBufferObject (GLuint buffer)

Registers an OpenGL buffer object.

Parameters

buffer

- The name of the buffer object to register.

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_ALREADY_MAPPED

Description

Deprecated This function is deprecated as of Cuda 3.0.

Registers the buffer object specified by `buffer` for access by CUDA. This function must be called before CUDA can map the buffer object. There must be a valid OpenGL context bound to the current thread when this function is called, and the buffer name is resolved by that context.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsGLRegisterBuffer](#)

CUresult cuGLSetBufferObjectMapFlags (GLuint buffer, unsigned int Flags)

Set the map flags for an OpenGL buffer object.

Parameters

buffer

- Buffer object to unmap

Flags

- Map flags

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_NOT_INITIALIZED`,
`CUDA_ERROR_INVALID_HANDLE`, `CUDA_ERROR_ALREADY_MAPPED`,
`CUDA_ERROR_INVALID_CONTEXT`,

Description

Deprecated This function is deprecated as of Cuda 3.0.

Sets the map flags for the buffer object specified by `buffer`.

Changes to `Flags` will take effect the next time `buffer` is mapped. The `Flags` argument may be any of the following:

- ▶ `CU_GL_MAP_RESOURCE_FLAGS_NONE`: Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA kernels. This is the default value.
- ▶ `CU_GL_MAP_RESOURCE_FLAGS_READ_ONLY`: Specifies that CUDA kernels which access this resource will not write to this resource.
- ▶ `CU_GL_MAP_RESOURCE_FLAGS_WRITE_DISCARD`: Specifies that CUDA kernels which access this resource will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.

If `buffer` has not been registered for use with CUDA, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `buffer` is presently mapped for access by CUDA, then `CUDA_ERROR_ALREADY_MAPPED` is returned.

There must be a valid OpenGL context bound to the current thread when this function is called. This must be the same context, or a member of the same `shareGroup`, as the context that was bound when the buffer was registered.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsResourceSetMapFlags](#)

CUresult cuGLUnmapBufferObject (GLuint buffer)

Unmaps an OpenGL buffer object.

Parameters

buffer

- Buffer object to unmap

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Deprecated This function is deprecated as of Cuda 3.0.

Unmaps the buffer object specified by `buffer` for access by CUDA.

There must be a valid OpenGL context bound to the current thread when this function is called. This must be the same context, or a member of the same shareGroup, as the context that was bound when the buffer was registered.

All streams in the current CUDA context are synchronized with the current GL context.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsUnmapResources](#)

CUresult cuGLUnmapBufferObjectAsync (GLuint buffer, CUstream hStream)

Unmaps an OpenGL buffer object.

Parameters**buffer**

- Name of the buffer object to unmap

hStream

- Stream to synchronize

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE

Description

Deprecated This function is deprecated as of Cuda 3.0.

Unmaps the buffer object specified by `buffer` for access by CUDA.

There must be a valid OpenGL context bound to the current thread when this function is called. This must be the same context, or a member of the same shareGroup, as the context that was bound when the buffer was registered.

Stream `hStream` in the current CUDA context is synchronized with the current GL context.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsUnmapResources](#)

CUresult cuGLUnregisterBufferObject (GLuint buffer)

Unregister an OpenGL buffer object.

Parameters

buffer

- Name of the buffer object to unregister

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#),
[CUDA_ERROR_INVALID_VALUE](#)

Description

Deprecated This function is deprecated as of Cuda 3.0.

Unregisters the buffer object specified by `buffer`. This releases any resources associated with the registered buffer. After this call, the buffer may no longer be mapped for access by CUDA.

There must be a valid OpenGL context bound to the current thread when this function is called. This must be the same context, or a member of the same shareGroup, as the context that was bound when the buffer was registered.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsUnregisterResource](#)

4.28. Direct3D 9 Interoperability

This section describes the Direct3D 9 interoperability functions of the low-level CUDA driver application programming interface. Note that mapping of Direct3D 9 resources is performed with the graphics API agnostic, resource mapping interface described in [Graphics Interoperability](#).

Direct3D 9 Interoperability [DEPRECATED]

enum CUd3d9DeviceList

CUDA devices corresponding to a D3D9 device

Values

CU_D3D9_DEVICE_LIST_ALL = 0x01

The CUDA devices for all GPUs used by a D3D9 device

CU_D3D9_DEVICE_LIST_CURRENT_FRAME = 0x02

The CUDA devices for the GPUs used by a D3D9 device in its currently rendering frame

CU_D3D9_DEVICE_LIST_NEXT_FRAME = 0x03

The CUDA devices for the GPUs to be used by a D3D9 device in the next frame

CUresult cuD3D9CtxCreate (CUcontext *pCtx, CUdevice *pCudaDevice, unsigned int Flags, IDirect3DDevice9 *pD3DDevice)

Create a CUDA context for interoperability with Direct3D 9.

Parameters

pCtx

- Returned newly created CUDA context

pCudaDevice

- Returned pointer to the device on which the context was created

Flags

- Context creation flags (see [cuCtxCreate\(\)](#) for details)

pD3DDevice

- Direct3D device to create interoperability context with

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Creates a new CUDA context, enables interoperability for that context with the Direct3D device `pD3DDevice`, and associates the created CUDA context with the calling thread. The created **CUcontext** will be returned in `*pCtx`. Direct3D resources from this device may be registered and mapped through the lifetime of this CUDA context. If `pCudaDevice` is non-NULL then the **CUdevice** on which this CUDA context was created will be returned in `*pCudaDevice`.

On success, this call will increase the internal reference count on `pD3DDevice`. This reference count will be decremented upon destruction of this context through `cuCtxDestroy()`. This context will cease to function if `pD3DDevice` is destroyed or encounters an error.

Note that this function is never required for correct functionality. Use of this function will result in accelerated interoperability only when the operating system is Windows Vista or Windows 7, and the device `pD3DDevice` is not an `IDirect3DDevice9Ex`. In all other circumstances, this function is not necessary.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuD3D9GetDevice`, `cuGraphicsD3D9RegisterResource`

CUresult cuD3D9CtxCreateOnDevice (CUcontext *pCtx, unsigned int flags, IDirect3DDevice9 *pD3DDevice, CUdevice cudaDevice)

Create a CUDA context for interoperability with Direct3D 9.

Parameters

`pCtx`

- Returned newly created CUDA context

`flags`

- Context creation flags (see `cuCtxCreate()` for details)

pD3DDevice

- Direct3D device to create interoperability context with

cudaDevice

- The CUDA device on which to create the context. This device must be among the devices returned when querying CU_D3D9_DEVICES_ALL from [cuD3D9GetDevices](#).

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Creates a new CUDA context, enables interoperability for that context with the Direct3D device `pD3DDevice`, and associates the created CUDA context with the calling thread. The created [CUcontext](#) will be returned in `*pCtx`. Direct3D resources from this device may be registered and mapped through the lifetime of this CUDA context.

On success, this call will increase the internal reference count on `pD3DDevice`. This reference count will be decremented upon destruction of this context through [cuCtxDestroy\(\)](#). This context will cease to function if `pD3DDevice` is destroyed or encounters an error.

Note that this function is never required for correct functionality. Use of this function will result in accelerated interoperability only when the operating system is Windows Vista or Windows 7, and the device `pD3DDevice` is not an `IDirect3DDevice9Ex`. In all other circumstances, this function is not necessary.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D9GetDevices](#), [cuGraphicsD3D9RegisterResource](#)

CUresult cuD3D9GetDevice (CUdevice *pCudaDevice, const char *pszAdapterName)

Gets the CUDA device corresponding to a display adapter.

Parameters**pCudaDevice**

- Returned CUDA device corresponding to `pszAdapterName`

pszAdapterName

- Adapter name to query for device

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_NOT_FOUND, CUDA_ERROR_UNKNOWN

Description

Returns in *pCudaDevice the CUDA-compatible device corresponding to the adapter name pszAdapterName obtained from EnumDisplayDevices() or IDirect3D9::GetAdapterIdentifier().

If no device on the adapter with name pszAdapterName is CUDA-compatible, then the call will fail.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuD3D9CtxCreate, cudaD3D9GetDevice

CUresult cuD3D9GetDevices (unsigned int *pCudaDeviceCount, CUdevice *pCudaDevices, unsigned int cudaDeviceCount, IDirect3DDevice9 *pD3D9Device, CUd3d9DeviceList deviceList)

Gets the CUDA devices corresponding to a Direct3D 9 device.

Parameters**pCudaDeviceCount**

- Returned number of CUDA devices corresponding to pD3D9Device

pCudaDevices

- Returned CUDA devices corresponding to pD3D9Device

cudaDeviceCount

- The size of the output device array pCudaDevices

pD3D9Device

- Direct3D 9 device to query for CUDA devices

deviceList

- The set of devices to return. This set may be `CU_D3D9_DEVICE_LIST_ALL` for all devices, `CU_D3D9_DEVICE_LIST_CURRENT_FRAME` for the devices used to render the current frame (in SLI), or `CU_D3D9_DEVICE_LIST_NEXT_FRAME` for the devices used to render the next frame (in SLI).

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_NO_DEVICE`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_NOT_FOUND`,
`CUDA_ERROR_UNKNOWN`

Description

Returns in `*pCudaDeviceCount` the number of CUDA-compatible device corresponding to the Direct3D 9 device `pD3D9Device`. Also returns in `*pCudaDevices` at most `cudaDeviceCount` of the CUDA-compatible devices corresponding to the Direct3D 9 device `pD3D9Device`.

If any of the GPUs being used to render `pDevice` are not CUDA capable then the call will return `CUDA_ERROR_NO_DEVICE`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuD3D9CtxCreate`, `cudaD3D9GetDevices`

CUresult cuD3D9GetDirect3DDevice (IDirect3DDevice9 ****ppD3DDevice)**

Get the Direct3D 9 device against which the current CUDA context was created.

Parameters**ppD3DDevice**

- Returned Direct3D device corresponding to CUDA context

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`
`CUDA_ERROR_INVALID_GRAPHICS_CONTEXT`

Description

Returns in `*ppD3DDevice` the Direct3D device against which this CUDA context was created in `cuD3D9CtxCreate()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuD3D9GetDevice`, `cudaD3D9GetDirect3DDevice`

CUresult cuGraphicsD3D9RegisterResource (CUgraphicsResource *pCudaResource, IDirect3DResource9 *pD3DResource, unsigned int Flags)

Register a Direct3D 9 resource for access by CUDA.

Parameters

`pCudaResource`

- Returned graphics resource handle

`pD3DResource`

- Direct3D resource to register

Flags

- Parameters for resource registration

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_OUT_OF_MEMORY`, `CUDA_ERROR_UNKNOWN`

Description

Registers the Direct3D 9 resource `pD3DResource` for access by CUDA and returns a CUDA handle to `pD3DResource` in `pCudaResource`. The handle returned in `pCudaResource` may be used to map and unmap this resource until it is unregistered. On success this call will increase the internal reference count on `pD3DResource`. This reference count will be decremented when this resource is unregistered through `cuGraphicsUnregisterResource()`.

This call is potentially high-overhead and should not be called every frame in interactive applications.

The type of `pD3DResource` must be one of the following.

- ▶ `IDirect3DVertexBuffer9`: may be accessed through a device pointer
- ▶ `IDirect3DIndexBuffer9`: may be accessed through a device pointer
- ▶ `IDirect3DSurface9`: may be accessed through an array. Only stand-alone objects of type `IDirect3DSurface9` may be explicitly shared. In particular, individual mipmap levels and faces of cube maps may not be registered directly. To access individual surfaces associated with a texture, one must register the base texture object.
- ▶ `IDirect3DBaseTexture9`: individual surfaces on this texture may be accessed through an array.

The `Flags` argument may be used to specify additional parameters at register time. The valid values for this parameter are

- ▶ `CU_GRAPHICS_REGISTER_FLAGS_NONE`: Specifies no hints about how this resource will be used.
- ▶ `CU_GRAPHICS_REGISTER_FLAGS_SURFACE_LDST`: Specifies that CUDA will bind this resource to a surface reference.
- ▶ `CU_GRAPHICS_REGISTER_FLAGS_TEXTURE_GATHER`: Specifies that CUDA will perform texture gather operations on this resource.

Not all Direct3D resources of the above types may be used for interoperability with CUDA. The following are some limitations.

- ▶ The primary rendertarget may not be registered with CUDA.
- ▶ Resources allocated as shared may not be registered with CUDA.
- ▶ Textures which are not of a format which is 1, 2, or 4 channels of 8, 16, or 32-bit integer or floating-point data cannot be shared.
- ▶ Surfaces of depth or stencil formats cannot be shared.

A complete list of supported formats is as follows:

- ▶ `D3DFMT_L8`
- ▶ `D3DFMT_L16`
- ▶ `D3DFMT_A8R8G8B8`
- ▶ `D3DFMT_X8R8G8B8`
- ▶ `D3DFMT_G16R16`
- ▶ `D3DFMT_A8B8G8R8`
- ▶ `D3DFMT_A8`
- ▶ `D3DFMT_A8L8`
- ▶ `D3DFMT_Q8W8V8U8`
- ▶ `D3DFMT_V16U16`
- ▶ `D3DFMT_A16B16G16R16F`
- ▶ `D3DFMT_A16B16G16R16`
- ▶ `D3DFMT_R32F`

- ▶ D3DFMT_G16R16F
- ▶ D3DFMT_A32B32G32R32F
- ▶ D3DFMT_G32R32F
- ▶ D3DFMT_R16F

If Direct3D interoperability is not initialized for this context using `cuD3D9CtxCreate` then `CUDA_ERROR_INVALID_CONTEXT` is returned. If `pD3DResource` is of incorrect type or is already registered then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pD3DResource` cannot be registered then `CUDA_ERROR_UNKNOWN` is returned. If `Flags` is not one of the above specified value then `CUDA_ERROR_INVALID_VALUE` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuD3D9CtxCreate`, `cuGraphicsUnregisterResource`, `cuGraphicsMapResources`, `cuGraphicsSubResourceGetMappedArray`, `cuGraphicsResourceGetMappedPointer`, `cudaGraphicsD3D9RegisterResource`

4.28.1. Direct3D 9 Interoperability [DEPRECATED]

Direct3D 9 Interoperability

This section describes deprecated Direct3D 9 interoperability functionality.

enum CUd3d9map_flags

Flags to map or unmap a resource

Values

```
CU_D3D9_MAPRESOURCE_FLAGS_NONE = 0x00
CU_D3D9_MAPRESOURCE_FLAGS_READONLY = 0x01
CU_D3D9_MAPRESOURCE_FLAGS_WRITEDISCARD = 0x02
```

enum CUd3d9register_flags

Flags to register a resource

Values

```
CU_D3D9_REGISTER_FLAGS_NONE = 0x00
CU_D3D9_REGISTER_FLAGS_ARRAY = 0x01
```

CUresult cuD3D9MapResources (unsigned int count, IDirect3DResource9 **ppResource)

Map Direct3D resources for access by CUDA.

Parameters

count

- Number of resources in ppResource

ppResource

- Resources to map for CUDA usage

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_ALREADY_MAPPED,
 CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 3.0.

Maps the `count` Direct3D resources in `ppResource` for access by CUDA.

The resources in `ppResource` may be accessed in CUDA kernels until they are unmapped. Direct3D should not access any resources while they are mapped by CUDA. If an application does so the results are undefined.

This function provides the synchronization guarantee that any Direct3D calls issued before `cuD3D9MapResources()` will complete before any CUDA kernels issued after `cuD3D9MapResources()` begin.

If any of `ppResource` have not been registered for use with CUDA or if `ppResource` contains any duplicate entries, then `CUDA_ERROR_INVALID_HANDLE` is returned. If any of `ppResource` are presently mapped for access by CUDA, then `CUDA_ERROR_ALREADY_MAPPED` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsMapResources](#)

CUresult cuD3D9RegisterResource (IDirect3DResource9 *pResource, unsigned int Flags)

Register a Direct3D resource for access by CUDA.

Parameters

pResource

- Resource to register for CUDA access

Flags

- Flags for resource registration

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 3.0.

Registers the Direct3D resource pResource for access by CUDA.

If this call is successful, then the application will be able to map and unmap this resource until it is unregistered through [cuD3D9UnregisterResource\(\)](#). Also on success, this call will increase the internal reference count on pResource. This reference count will be decremented when this resource is unregistered through [cuD3D9UnregisterResource\(\)](#).

This call is potentially high-overhead and should not be called every frame in interactive applications.

The type of pResource must be one of the following.

- ▶ IDirect3DVertexBuffer9: Cannot be used with Flags set to CU_D3D9_REGISTER_FLAGS_ARRAY.
- ▶ IDirect3DIndexBuffer9: Cannot be used with Flags set to CU_D3D9_REGISTER_FLAGS_ARRAY.
- ▶ IDirect3DSurface9: Only stand-alone objects of type IDirect3DSurface9 may be explicitly shared. In particular, individual mipmap levels and faces of cube maps may not be registered directly. To access individual surfaces associated with a texture, one must register the base texture object. For restrictions on the Flags parameter, see type IDirect3DBaseTexture9.
- ▶ IDirect3DBaseTexture9: When a texture is registered, all surfaces associated with the all mipmap levels of all faces of the texture will be accessible to CUDA.

The `Flags` argument specifies the mechanism through which CUDA will access the Direct3D resource. The following values are allowed.

- ▶ `CUDA_D3D9_REGISTER_FLAGS_NONE`: Specifies that CUDA will access this resource through a `CUDA_DEVICEPTR`. The pointer, size, and (for textures), pitch for each subresource of this allocation may be queried through `cuD3D9ResourceGetMappedPointer()`, `cuD3D9ResourceGetMappedSize()`, and `cuD3D9ResourceGetMappedPitch()` respectively. This option is valid for all resource types.
- ▶ `CUDA_D3D9_REGISTER_FLAGS_ARRAY`: Specifies that CUDA will access this resource through a `CUDA_ARRAY` queried on a sub-resource basis through `cuD3D9ResourceGetMappedArray()`. This option is only valid for resources of type `IDirect3DSurface9` and subtypes of `IDirect3DBaseTexture9`.

Not all Direct3D resources of the above types may be used for interoperability with CUDA. The following are some limitations.

- ▶ The primary rendertarget may not be registered with CUDA.
- ▶ Resources allocated as shared may not be registered with CUDA.
- ▶ Any resources allocated in `D3DPOOL_SYSTEMMEM` or `D3DPOOL_MANAGED` may not be registered with CUDA.
- ▶ Textures which are not of a format which is 1, 2, or 4 channels of 8, 16, or 32-bit integer or floating-point data cannot be shared.
- ▶ Surfaces of depth or stencil formats cannot be shared.

If Direct3D interoperability is not initialized on this context, then `CUDA_ERROR_INVALID_CONTEXT` is returned. If `pResource` is of incorrect type (e.g. is a non-stand-alone `IDirect3DSurface9`) or is already registered, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` cannot be registered then `CUDA_ERROR_UNKNOWN` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsD3D9RegisterResource](#)

CUresult cuD3D9ResourceGetMappedArray (CUarray *pArray, IDirect3DResource9 *pResource, unsigned int Face, unsigned int Level)

Get an array through which to access a subresource of a Direct3D resource which has been mapped for access by CUDA.

Parameters

pArray

- Returned array corresponding to subresource

pResource

- Mapped resource to access

Face

- Face of resource to access

Level

- Level of resource to access

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_NOT_MAPPED

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in *pArray an array through which the subresource of the mapped Direct3D resource pResource which corresponds to Face and Level may be accessed. The value set in pArray may change every time that pResource is mapped.

If pResource is not registered then CUDA_ERROR_INVALID_HANDLE is returned. If pResource was not registered with usage flags CU_D3D9_REGISTER_FLAGS_ARRAY then CUDA_ERROR_INVALID_HANDLE is returned. If pResource is not mapped then CUDA_ERROR_NOT_MAPPED is returned.

For usage requirements of Face and Level parameters, see cuD3D9ResourceGetMappedPointer().



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsSubResourceGetMappedArray`

CUresult cuD3D9ResourceGetMappedPitch (`size_t *pPitch`, `size_t *pPitchSlice`, `IDirect3DResource9 *pResource`, `unsigned int Face`, `unsigned int Level`)

Get the pitch of a subresource of a Direct3D resource which has been mapped for access by CUDA.

Parameters

pPitch

- Returned pitch of subresource

pPitchSlice

- Returned Z-slice pitch of subresource

pResource

- Mapped resource to access

Face

- Face of resource to access

Level

- Level of resource to access

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_NOT_MAPPED`

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in `*pPitch` and `*pPitchSlice` the pitch and Z-slice pitch of the subresource of the mapped Direct3D resource `pResource`, which corresponds to `Face` and `Level`. The values set in `pPitch` and `pPitchSlice` may change every time that `pResource` is mapped.

The pitch and Z-slice pitch values may be used to compute the location of a sample on a surface as follows.

For a 2D surface, the byte offset of the sample at position `x`, `y` from the base pointer of the surface is:

$$y * \text{pitch} + (\text{bytes per pixel}) * x$$

For a 3D surface, the byte offset of the sample at position `x`, `y`, `z` from the base pointer of the surface is:

$z * \text{slicePitch} + y * \text{pitch} + (\text{bytes per pixel}) * x$

Both parameters `pPitch` and `pPitchSlice` are optional and may be set to `NULL`.

If `pResource` is not of type `IDirect3DBaseTexture9` or one of its subtypes or if `pResource` has not been registered for use with CUDA, then `cudaErrorInvalidResourceHandle` is returned. If `pResource` was not registered with usage flags `CU_D3D9_REGISTER_FLAGS_NONE`, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` is not mapped for access by CUDA then `CUDA_ERROR_NOT_MAPPED` is returned.

For usage requirements of `Face` and `Level` parameters, see `cuD3D9ResourceGetMappedPointer()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsSubResourceGetMappedArray`

CUresult cuD3D9ResourceGetMappedPointer (CUdeviceptr *pDevPtr, IDirect3DResource9 *pResource, unsigned int Face, unsigned int Level)

Get the pointer through which to access a subresource of a Direct3D resource which has been mapped for access by CUDA.

Parameters

pDevPtr

- Returned pointer corresponding to subresource

pResource

- Mapped resource to access

Face

- Face of resource to access

Level

- Level of resource to access

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_NOT_MAPPED`

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in `*pDevPtr` the base pointer of the subresource of the mapped Direct3D resource `pResource`, which corresponds to `Face` and `Level`. The value set in `pDevPtr` may change every time that `pResource` is mapped.

If `pResource` is not registered, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` was not registered with usage flags `CU_D3D9_REGISTER_FLAGS_NONE`, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` is not mapped, then `CUDA_ERROR_NOT_MAPPED` is returned.

If `pResource` is of type `IDirect3DCubeTexture9`, then `Face` must one of the values enumerated by type `D3DCUBEMAP_FACES`. For all other types `Face` must be 0. If `Face` is invalid, then `CUDA_ERROR_INVALID_VALUE` is returned.

If `pResource` is of type `IDirect3DBaseTexture9`, then `Level` must correspond to a valid mipmap level. At present only mipmap level 0 is supported. For all other types `Level` must be 0. If `Level` is invalid, then `CUDA_ERROR_INVALID_VALUE` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsResourceGetMappedPointer](#)

CUresult cuD3D9ResourceGetMappedSize (size_t *pSize, IDirect3DResource9 *pResource, unsigned int Face, unsigned int Level)

Get the size of a subresource of a Direct3D resource which has been mapped for access by CUDA.

Parameters

pSize

- Returned size of subresource

pResource

- Mapped resource to access

Face

- Face of resource to access

Level

- Level of resource to access

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE,
 CUDA_ERROR_NOT_MAPPED

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in `*pSize` the size of the subresource of the mapped Direct3D resource `pResource`, which corresponds to `Face` and `Level`. The value set in `pSize` may change every time that `pResource` is mapped.

If `pResource` has not been registered for use with CUDA, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` was not registered with usage flags `CU_D3D9_REGISTER_FLAGS_NONE`, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` is not mapped for access by CUDA, then `CUDA_ERROR_NOT_MAPPED` is returned.

For usage requirements of `Face` and `Level` parameters, see `cuD3D9ResourceGetMappedPointer`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsResourceGetMappedPointer`

CUresult cuD3D9ResourceGetSurfaceDimensions (`size_t *pWidth`, `size_t *pHeight`, `size_t *pDepth`, `IDirect3DResource9 *pResource`, `unsigned int Face`, `unsigned int Level`)

Get the dimensions of a registered surface.

Parameters**pWidth**

- Returned width of surface

pHeight

- Returned height of surface

pDepth

- Returned depth of surface

pResource

- Registered resource to access

Face

- Face of resource to access

Level

- Level of resource to access

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in `*pWidth`, `*pHeight`, and `*pDepth` the dimensions of the subresource of the mapped Direct3D resource `pResource`, which corresponds to `Face` and `Level`.

Because anti-aliased surfaces may have multiple samples per pixel, it is possible that the dimensions of a resource will be an integer factor larger than the dimensions reported by the Direct3D runtime.

The parameters `pWidth`, `pHeight`, and `pDepth` are optional. For 2D surfaces, the value returned in `*pDepth` will be 0.

If `pResource` is not of type `IDirect3DBaseTexture9` or `IDirect3DSurface9` or if `pResource` has not been registered for use with CUDA, then `CUDA_ERROR_INVALID_HANDLE` is returned.

For usage requirements of `Face` and `Level` parameters, see `cuD3D9ResourceGetMappedPointer()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsSubResourceGetMappedArray`

CUresult cuD3D9ResourceSetMapFlags (IDirect3DResource9 *pResource, unsigned int Flags)

Set usage flags for mapping a Direct3D resource.

Parameters

pResource

- Registered resource to set flags for

Flags

- Parameters for resource mapping

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_ALREADY_MAPPED

Description

Deprecated This function is deprecated as of Cuda 3.0.

Set `Flags` for mapping the Direct3D resource `pResource`.

Changes to `Flags` will take effect the next time `pResource` is mapped. The `Flags` argument may be any of the following:

- ▶ `CU_D3D9_MAPRESOURCE_FLAGS_NONE`: Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA kernels. This is the default value.
- ▶ `CU_D3D9_MAPRESOURCE_FLAGS_READONLY`: Specifies that CUDA kernels which access this resource will not write to this resource.
- ▶ `CU_D3D9_MAPRESOURCE_FLAGS_WRITEDISCARD`: Specifies that CUDA kernels which access this resource will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.

If `pResource` has not been registered for use with CUDA, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` is presently mapped for access by CUDA, then `CUDA_ERROR_ALREADY_MAPPED` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsResourceSetMapFlags`

CUresult cuD3D9UnmapResources (unsigned int count, IDirect3DResource9 **ppResource)

Unmaps Direct3D resources.

Parameters

count

- Number of resources to unmap for CUDA

ppResource

- Resources to unmap for CUDA

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_NOT_MAPPED, CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 3.0.

Unmaps the `count` Direct3D resources in `ppResource`.

This function provides the synchronization guarantee that any CUDA kernels issued before `cuD3D9UnmapResources()` will complete before any Direct3D calls issued after `cuD3D9UnmapResources()` begin.

If any of `ppResource` have not been registered for use with CUDA or if `ppResource` contains any duplicate entries, then `CUDA_ERROR_INVALID_HANDLE` is returned. If any of `ppResource` are not presently mapped for access by CUDA, then `CUDA_ERROR_NOT_MAPPED` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsUnmapResources`

CUresult cuD3D9UnregisterResource (IDirect3DResource9 *pResource)

Unregister a Direct3D resource.

Parameters

pResource

- Resource to unregister

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 3.0.

Unregisters the Direct3D resource pResource so it is not accessible by CUDA unless registered again.

If pResource is not registered, then CUDA_ERROR_INVALID_HANDLE is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsUnregisterResource](#)

4.29. Direct3D 10 Interoperability

This section describes the Direct3D 10 interoperability functions of the low-level CUDA driver application programming interface. Note that mapping of Direct3D 10 resources is performed with the graphics API agnostic, resource mapping interface described in [Graphics Interoperability](#).

Direct3D 10 Interoperability [DEPRECATED]

enum CUd3d10DeviceList

CUDA devices corresponding to a D3D10 device

Values**CU_D3D10_DEVICE_LIST_ALL = 0x01**

The CUDA devices for all GPUs used by a D3D10 device

CU_D3D10_DEVICE_LIST_CURRENT_FRAME = 0x02

The CUDA devices for the GPUs used by a D3D10 device in its currently rendering frame

CU_D3D10_DEVICE_LIST_NEXT_FRAME = 0x03

The CUDA devices for the GPUs to be used by a D3D10 device in the next frame

CUresult cuD3D10GetDevice (CUdevice *pCudaDevice, IDXGIAdapter *pAdapter)

Gets the CUDA device corresponding to a display adapter.

Parameters**pCudaDevice**

- Returned CUDA device corresponding to pAdapter

pAdapter

- Adapter to query for CUDA device

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_NOT_FOUND, CUDA_ERROR_UNKNOWN

Description

Returns in *pCudaDevice the CUDA-compatible device corresponding to the adapter pAdapter obtained from IDXGIFactory::EnumAdapters.

If no device on pAdapter is CUDA-compatible then the call will fail.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

cuD3D10GetDevices, cudaD3D10GetDevice

CUresult cuD3D10GetDevices (unsigned int *pCudaDeviceCount, CUdevice *pCudaDevices, unsigned

int cudaDeviceCount, ID3D10Device *pD3D10Device, CUd3d10DeviceList deviceList)

Gets the CUDA devices corresponding to a Direct3D 10 device.

Parameters

pCudaDeviceCount

- Returned number of CUDA devices corresponding to pD3D10Device

pCudaDevices

- Returned CUDA devices corresponding to pD3D10Device

cudaDeviceCount

- The size of the output device array pCudaDevices

pD3D10Device

- Direct3D 10 device to query for CUDA devices

deviceList

- The set of devices to return. This set may be `CU_D3D10_DEVICE_LIST_ALL` for all devices, `CU_D3D10_DEVICE_LIST_CURRENT_FRAME` for the devices used to render the current frame (in SLI), or `CU_D3D10_DEVICE_LIST_NEXT_FRAME` for the devices used to render the next frame (in SLI).

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_NO_DEVICE`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_NOT_FOUND`,
`CUDA_ERROR_UNKNOWN`

Description

Returns in *pCudaDeviceCount the number of CUDA-compatible device corresponding to the Direct3D 10 device pD3D10Device. Also returns in *pCudaDevices at most cudaDeviceCount of the CUDA-compatible devices corresponding to the Direct3D 10 device pD3D10Device.

If any of the GPUs being used to render pDevice are not CUDA capable then the call will return `CUDA_ERROR_NO_DEVICE`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D10GetDevice](#), [cudaD3D10GetDevices](#)

CUresult cuGraphicsD3D10RegisterResource (CUgraphicsResource *pCudaResource, ID3D10Resource *pD3DResource, unsigned int Flags)

Register a Direct3D 10 resource for access by CUDA.

Parameters

pCudaResource

- Returned graphics resource handle

pD3DResource

- Direct3D resource to register

Flags

- Parameters for resource registration

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Registers the Direct3D 10 resource `pD3DResource` for access by CUDA and returns a CUDA handle to `pD3DResource` in `pCudaResource`. The handle returned in `pCudaResource` may be used to map and unmap this resource until it is unregistered. On success this call will increase the internal reference count on `pD3DResource`. This reference count will be decremented when this resource is unregistered through `cuGraphicsUnregisterResource()`.

This call is potentially high-overhead and should not be called every frame in interactive applications.

The type of `pD3DResource` must be one of the following.

- ▶ ID3D10Buffer: may be accessed through a device pointer.
- ▶ ID3D10Texture1D: individual subresources of the texture may be accessed via arrays
- ▶ ID3D10Texture2D: individual subresources of the texture may be accessed via arrays
- ▶ ID3D10Texture3D: individual subresources of the texture may be accessed via arrays

The `Flags` argument may be used to specify additional parameters at register time. The valid values for this parameter are

- ▶ CU_GRAPHICS_REGISTER_FLAGS_NONE: Specifies no hints about how this resource will be used.

- ▶ `CU_GRAPHICS_REGISTER_FLAGS_SURFACE_LDST`: Specifies that CUDA will bind this resource to a surface reference.
- ▶ `CU_GRAPHICS_REGISTER_FLAGS_TEXTURE_GATHER`: Specifies that CUDA will perform texture gather operations on this resource.

Not all Direct3D resources of the above types may be used for interoperability with CUDA. The following are some limitations.

- ▶ The primary rendertarget may not be registered with CUDA.
- ▶ Textures which are not of a format which is 1, 2, or 4 channels of 8, 16, or 32-bit integer or floating-point data cannot be shared.
- ▶ Surfaces of depth or stencil formats cannot be shared.

A complete list of supported DXGI formats is as follows. For compactness the notation `A_{B,C,D}` represents `A_B`, `A_C`, and `A_D`.

- ▶ `DXGI_FORMAT_A8_UNORM`
- ▶ `DXGI_FORMAT_B8G8R8A8_UNORM`
- ▶ `DXGI_FORMAT_B8G8R8X8_UNORM`
- ▶ `DXGI_FORMAT_R16_FLOAT`
- ▶ `DXGI_FORMAT_R16G16B16A16_{FLOAT,SINT,SNORM,UINT,UNORM}`
- ▶ `DXGI_FORMAT_R16G16_{FLOAT,SINT,SNORM,UINT,UNORM}`
- ▶ `DXGI_FORMAT_R16_{SINT,SNORM,UINT,UNORM}`
- ▶ `DXGI_FORMAT_R32_FLOAT`
- ▶ `DXGI_FORMAT_R32G32B32A32_{FLOAT,SINT,UINT}`
- ▶ `DXGI_FORMAT_R32G32_{FLOAT,SINT,UINT}`
- ▶ `DXGI_FORMAT_R32_{SINT,UINT}`
- ▶ `DXGI_FORMAT_R8G8B8A8_{SINT,SNORM,UINT,UNORM,UNORM_SRGB}`
- ▶ `DXGI_FORMAT_R8G8_{SINT,SNORM,UINT,UNORM}`
- ▶ `DXGI_FORMAT_R8_{SINT,SNORM,UINT,UNORM}`

If `pD3DResource` is of incorrect type or is already registered then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pD3DResource` cannot be registered then `CUDA_ERROR_UNKNOWN` is returned. If `Flags` is not one of the above specified value then `CUDA_ERROR_INVALID_VALUE` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsUnregisterResource`, `cuGraphicsMapResources`,
`cuGraphicsSubResourceGetMappedArray`, `cuGraphicsResourceGetMappedPointer`,
`cudaGraphicsD3D10RegisterResource`

4.29.1. Direct3D 10 Interoperability [DEPRECATED]

Direct3D 10 Interoperability

This section describes deprecated Direct3D 10 interoperability functionality.

enum `CUD3D10map_flags`

Flags to map or unmap a resource

Values

`CU_D3D10_MAPRESOURCE_FLAGS_NONE = 0x00`

`CU_D3D10_MAPRESOURCE_FLAGS_READONLY = 0x01`

`CU_D3D10_MAPRESOURCE_FLAGS_WRITEDISCARD = 0x02`

enum `CUD3D10register_flags`

Flags to register a resource

Values

`CU_D3D10_REGISTER_FLAGS_NONE = 0x00`

`CU_D3D10_REGISTER_FLAGS_ARRAY = 0x01`

`CUresult cuD3D10CtxCreate (CUcontext *pCtx, CUdevice *pCudaDevice, unsigned int Flags, ID3D10Device *pD3DDevice)`

Create a CUDA context for interoperability with Direct3D 10.

Parameters

`pCtx`

- Returned newly created CUDA context

`pCudaDevice`

- Returned pointer to the device on which the context was created

Flags

- Context creation flags (see `cuCtxCreate()` for details)

`pD3DDevice`

- Direct3D device to create interoperability context with

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 5.0.

This function is deprecated and should no longer be used. It is no longer necessary to associate a CUDA context with a D3D10 device in order to achieve maximum interoperability performance.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D10GetDevice](#), [cuGraphicsD3D10RegisterResource](#)

CUresult cuD3D10CtxCreateOnDevice (CUcontext *pCtx, unsigned int flags, ID3D10Device *pD3DDevice, CUdevice cudaDevice)

Create a CUDA context for interoperability with Direct3D 10.

Parameters**pCtx**

- Returned newly created CUDA context

flags

- Context creation flags (see [cuCtxCreate\(\)](#) for details)

pD3DDevice

- Direct3D device to create interoperability context with

cudaDevice

- The CUDA device on which to create the context. This device must be among the devices returned when querying CU_D3D10_DEVICES_ALL from [cuD3D10GetDevices](#).

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
 CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 5.0.

This function is deprecated and should no longer be used. It is no longer necessary to associate a CUDA context with a D3D10 device in order to achieve maximum interoperability performance.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D10GetDevices](#), [cuGraphicsD3D10RegisterResource](#)

CUresult cuD3D10GetDirect3DDevice (ID3D10Device **ppD3DDevice)

Get the Direct3D 10 device against which the current CUDA context was created.

Parameters

ppD3DDevice

- Returned Direct3D device corresponding to CUDA context

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT

Description

Deprecated This function is deprecated as of CUDA 5.0.

This function is deprecated and should no longer be used. It is no longer necessary to associate a CUDA context with a D3D10 device in order to achieve maximum interoperability performance.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D10GetDevice](#)

CUresult cuD3D10MapResources (unsigned int count, ID3D10Resource **ppResources)

Map Direct3D resources for access by CUDA.

Parameters

count

- Number of resources to map for CUDA

ppResources

- Resources to map for CUDA

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_ALREADY_MAPPED,
 CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 3.0.

Maps the `count` Direct3D resources in `ppResources` for access by CUDA.

The resources in `ppResources` may be accessed in CUDA kernels until they are unmapped. Direct3D should not access any resources while they are mapped by CUDA. If an application does so, the results are undefined.

This function provides the synchronization guarantee that any Direct3D calls issued before `cuD3D10MapResources()` will complete before any CUDA kernels issued after `cuD3D10MapResources()` begin.

If any of `ppResources` have not been registered for use with CUDA or if `ppResources` contains any duplicate entries, then `CUDA_ERROR_INVALID_HANDLE` is returned. If any of `ppResources` are presently mapped for access by CUDA, then `CUDA_ERROR_ALREADY_MAPPED` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsMapResources](#)

CUresult cuD3D10RegisterResource (ID3D10Resource *pResource, unsigned int Flags)

Register a Direct3D resource for access by CUDA.

Parameters

pResource

- Resource to register

Flags

- Parameters for resource registration

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 3.0.

Registers the Direct3D resource pResource for access by CUDA.

If this call is successful, then the application will be able to map and unmap this resource until it is unregistered through [cuD3D10UnregisterResource\(\)](#). Also on success, this call will increase the internal reference count on pResource. This reference count will be decremented when this resource is unregistered through [cuD3D10UnregisterResource\(\)](#).

This call is potentially high-overhead and should not be called every frame in interactive applications.

The type of pResource must be one of the following.

- ▶ ID3D10Buffer: Cannot be used with Flags set to CU_D3D10_REGISTER_FLAGS_ARRAY.
- ▶ ID3D10Texture1D: No restrictions.
- ▶ ID3D10Texture2D: No restrictions.
- ▶ ID3D10Texture3D: No restrictions.

The Flags argument specifies the mechanism through which CUDA will access the Direct3D resource. The following values are allowed.

- ▶ CU_D3D10_REGISTER_FLAGS_NONE: Specifies that CUDA will access this resource through a CUdeviceptr. The pointer, size, and (for textures), pitch for each subresource of this allocation may be queried through [cuD3D10ResourceGetMappedPointer\(\)](#), [cuD3D10ResourceGetMappedSize\(\)](#),

and `cuD3D10ResourceGetMappedPitch()` respectively. This option is valid for all resource types.

- ▶ `CU_D3D10_REGISTER_FLAGS_ARRAY`: Specifies that CUDA will access this resource through a `CUarray` queried on a sub-resource basis through `cuD3D10ResourceGetMappedArray()`. This option is only valid for resources of type `ID3D10Texture1D`, `ID3D10Texture2D`, and `ID3D10Texture3D`.

Not all Direct3D resources of the above types may be used for interoperability with CUDA. The following are some limitations.

- ▶ The primary rendertarget may not be registered with CUDA.
- ▶ Resources allocated as shared may not be registered with CUDA.
- ▶ Textures which are not of a format which is 1, 2, or 4 channels of 8, 16, or 32-bit integer or floating-point data cannot be shared.
- ▶ Surfaces of depth or stencil formats cannot be shared.

If Direct3D interoperability is not initialized on this context then `CUDA_ERROR_INVALID_CONTEXT` is returned. If `pResource` is of incorrect type or is already registered, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` cannot be registered, then `CUDA_ERROR_UNKNOWN` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsD3D10RegisterResource](#)

`CUresult cuD3D10ResourceGetMappedArray (CUarray *pArray, ID3D10Resource *pResource, unsigned int SubResource)`

Get an array through which to access a subresource of a Direct3D resource which has been mapped for access by CUDA.

Parameters

pArray

- Returned array corresponding to subresource

pResource

- Mapped resource to access

SubResource

- Subresource of pResource to access

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE,
 CUDA_ERROR_NOT_MAPPED

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in `*pArray` an array through which the subresource of the mapped Direct3D resource `pResource`, which corresponds to `SubResource` may be accessed. The value set in `pArray` may change every time that `pResource` is mapped.

If `pResource` is not registered, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` was not registered with usage flags `CU_D3D10_REGISTER_FLAGS_ARRAY`, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` is not mapped, then `CUDA_ERROR_NOT_MAPPED` is returned.

For usage requirements of the `SubResource` parameter, see `cuD3D10ResourceGetMappedPointer()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsSubResourceGetMappedArray`

CUresult cuD3D10ResourceGetMappedPitch (`size_t *pPitch`, `size_t *pPitchSlice`, `ID3D10Resource *pResource`, `unsigned int SubResource`)

Get the pitch of a subresource of a Direct3D resource which has been mapped for access by CUDA.

Parameters**pPitch**

- Returned pitch of subresource

pPitchSlice

- Returned Z-slice pitch of subresource

pResource

- Mapped resource to access

SubResource

- Subresource of `pResource` to access

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE,
 CUDA_ERROR_NOT_MAPPED

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in `*pPitch` and `*pPitchSlice` the pitch and Z-slice pitch of the subresource of the mapped Direct3D resource `pResource`, which corresponds to `SubResource`. The values set in `pPitch` and `pPitchSlice` may change every time that `pResource` is mapped.

The pitch and Z-slice pitch values may be used to compute the location of a sample on a surface as follows.

For a 2D surface, the byte offset of the sample at position `x`, `y` from the base pointer of the surface is:

$$y * \text{pitch} + (\text{bytes per pixel}) * x$$

For a 3D surface, the byte offset of the sample at position `x`, `y`, `z` from the base pointer of the surface is:

$$z * \text{slicePitch} + y * \text{pitch} + (\text{bytes per pixel}) * x$$

Both parameters `pPitch` and `pPitchSlice` are optional and may be set to NULL.

If `pResource` is not of type `IDirect3DBaseTexture10` or one of its subtypes or if `pResource` has not been registered for use with CUDA, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` was not registered with usage flags `CU_D3D10_REGISTER_FLAGS_NONE`, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` is not mapped for access by CUDA, then `CUDA_ERROR_NOT_MAPPED` is returned.

For usage requirements of the `SubResource` parameter, see `cuD3D10ResourceGetMappedPointer()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuGraphicsSubResourceGetMappedArray`

CUresult cuD3D10ResourceGetMappedPointer (CUdeviceptr *pDevPtr, ID3D10Resource *pResource, unsigned int SubResource)

Get a pointer through which to access a subresource of a Direct3D resource which has been mapped for access by CUDA.

Parameters

pDevPtr

- Returned pointer corresponding to subresource

pResource

- Mapped resource to access

SubResource

- Subresource of pResource to access

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE,
 CUDA_ERROR_NOT_MAPPED

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in *pDevPtr the base pointer of the subresource of the mapped Direct3D resource pResource, which corresponds to SubResource. The value set in pDevPtr may change every time that pResource is mapped.

If pResource is not registered, then [CUDA_ERROR_INVALID_HANDLE](#) is returned. If pResource was not registered with usage flags [CU_D3D10_REGISTER_FLAGS_NONE](#), then [CUDA_ERROR_INVALID_HANDLE](#) is returned. If pResource is not mapped, then [CUDA_ERROR_NOT_MAPPED](#) is returned.

If pResource is of type ID3D10Buffer, then SubResource must be 0. If pResource is of any other type, then the value of SubResource must come from the subresource calculation in [D3D10CalcSubResource\(\)](#).



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsResourceGetMappedPointer](#)

CUresult cuD3D10ResourceGetMappedSize (size_t *pSize, ID3D10Resource *pResource, unsigned int SubResource)

Get the size of a subresource of a Direct3D resource which has been mapped for access by CUDA.

Parameters

pSize

- Returned size of subresource

pResource

- Mapped resource to access

SubResource

- Subresource of pResource to access

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_NOT_MAPPED

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in *pSize the size of the subresource of the mapped Direct3D resource pResource, which corresponds to SubResource. The value set in pSize may change every time that pResource is mapped.

If pResource has not been registered for use with CUDA, then CUDA_ERROR_INVALID_HANDLE is returned. If pResource was not registered with usage flags CU_D3D10_REGISTER_FLAGS_NONE, then CUDA_ERROR_INVALID_HANDLE is returned. If pResource is not mapped for access by CUDA, then CUDA_ERROR_NOT_MAPPED is returned.

For usage requirements of the SubResource parameter, see cuD3D10ResourceGetMappedPointer().



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsResourceGetMappedPointer](#)

CUresult cuD3D10ResourceGetSurfaceDimensions (size_t *pWidth, size_t *pHeight, size_t *pDepth, ID3D10Resource *pResource, unsigned int SubResource)

Get the dimensions of a registered surface.

Parameters

pWidth

- Returned width of surface

pHeight

- Returned height of surface

pDepth

- Returned depth of surface

pResource

- Registered resource to access

SubResource

- Subresource of pResource to access

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE

Description

Deprecated This function is deprecated as of CUDA 3.0.

Returns in *pWidth, *pHeight, and *pDepth the dimensions of the subresource of the mapped Direct3D resource pResource, which corresponds to SubResource.

Because anti-aliased surfaces may have multiple samples per pixel, it is possible that the dimensions of a resource will be an integer factor larger than the dimensions reported by the Direct3D runtime.

The parameters pWidth, pHeight, and pDepth are optional. For 2D surfaces, the value returned in *pDepth will be 0.

If pResource is not of type IDirect3DBaseTexture10 or IDirect3DSurface10 or if pResource has not been registered for use with CUDA, then CUDA_ERROR_INVALID_HANDLE is returned.

For usage requirements of the SubResource parameter, see cuD3D10ResourceGetMappedPointer().



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsSubResourceGetMappedArray](#)

CUresult cuD3D10ResourceSetMapFlags (ID3D10Resource *pResource, unsigned int Flags)

Set usage flags for mapping a Direct3D resource.

Parameters

pResource

- Registered resource to set flags for

Flags

- Parameters for resource mapping

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED, CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT, CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_ALREADY_MAPPED

Description

Deprecated This function is deprecated as of CUDA 3.0.

Set flags for mapping the Direct3D resource pResource.

Changes to flags will take effect the next time pResource is mapped. The Flags argument may be any of the following.

- ▶ CU_D3D10_MAPRESOURCE_FLAGS_NONE: Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA kernels. This is the default value.
- ▶ CU_D3D10_MAPRESOURCE_FLAGS_READONLY: Specifies that CUDA kernels which access this resource will not write to this resource.
- ▶ CU_D3D10_MAPRESOURCE_FLAGS_WRITEDISCARD: Specifies that CUDA kernels which access this resource will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.

If `pResource` has not been registered for use with CUDA, then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pResource` is presently mapped for access by CUDA then `CUDA_ERROR_ALREADY_MAPPED` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsResourceSetMapFlags](#)

CUresult cuD3D10UnmapResources (unsigned int count, ID3D10Resource **ppResources)

Unmap Direct3D resources.

Parameters

count

- Number of resources to unmap for CUDA

ppResources

- Resources to unmap for CUDA

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_NOT_MAPPED`, `CUDA_ERROR_UNKNOWN`

Description

Deprecated This function is deprecated as of CUDA 3.0.

Unmaps the `count` Direct3D resources in `ppResources`.

This function provides the synchronization guarantee that any CUDA kernels issued before `cuD3D10UnmapResources()` will complete before any Direct3D calls issued after `cuD3D10UnmapResources()` begin.

If any of `ppResources` have not been registered for use with CUDA or if `ppResources` contains any duplicate entries, then `CUDA_ERROR_INVALID_HANDLE` is returned. If any of `ppResources` are not presently mapped for access by CUDA, then `CUDA_ERROR_NOT_MAPPED` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsUnmapResources](#)

CUresult cuD3D10UnregisterResource (ID3D10Resource *pResource)

Unregister a Direct3D resource.

Parameters

pResource

- Resources to unregister

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_HANDLE, CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 3.0.

Unregisters the Direct3D resource `pResource` so it is not accessible by CUDA unless registered again.

If `pResource` is not registered, then `CUDA_ERROR_INVALID_HANDLE` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsUnregisterResource](#)

4.30. Direct3D 11 Interoperability

This section describes the Direct3D 11 interoperability functions of the low-level CUDA driver application programming interface. Note that mapping of Direct3D 11 resources is performed with the graphics API agnostic, resource mapping interface described in [Graphics Interoperability](#).

Direct3D 11 Interoperability [DEPRECATED]

enum CUd3d11DeviceList

CUDA devices corresponding to a D3D11 device

Values

CU_D3D11_DEVICE_LIST_ALL = 0x01

The CUDA devices for all GPUs used by a D3D11 device

CU_D3D11_DEVICE_LIST_CURRENT_FRAME = 0x02

The CUDA devices for the GPUs used by a D3D11 device in its currently rendering frame

CU_D3D11_DEVICE_LIST_NEXT_FRAME = 0x03

The CUDA devices for the GPUs to be used by a D3D11 device in the next frame

CUresult cuD3D11GetDevice (CUdevice *pCudaDevice, IDXGIAdapter *pAdapter)

Gets the CUDA device corresponding to a display adapter.

Parameters

pCudaDevice

- Returned CUDA device corresponding to pAdapter

pAdapter

- Adapter to query for CUDA device

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_NO_DEVICE,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_NOT_FOUND,
 CUDA_ERROR_UNKNOWN

Description

Returns in *pCudaDevice the CUDA-compatible device corresponding to the adapter pAdapter obtained from IDXGIFactory::EnumAdapters.

If no device on pAdapter is CUDA-compatible the call will return CUDA_ERROR_NO_DEVICE.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D11GetDevices](#), [cudaD3D11GetDevice](#)

CUresult cuD3D11GetDevices (unsigned int *pCudaDeviceCount, CUdevice *pCudaDevices, unsigned int cudaDeviceCount, ID3D11Device *pD3D11Device, CUd3d11DeviceList deviceList)

Gets the CUDA devices corresponding to a Direct3D 11 device.

Parameters

pCudaDeviceCount

- Returned number of CUDA devices corresponding to pD3D11Device

pCudaDevices

- Returned CUDA devices corresponding to pD3D11Device

cudaDeviceCount

- The size of the output device array pCudaDevices

pD3D11Device

- Direct3D 11 device to query for CUDA devices

deviceList

- The set of devices to return. This set may be [CU_D3D11_DEVICE_LIST_ALL](#) for all devices, [CU_D3D11_DEVICE_LIST_CURRENT_FRAME](#) for the devices used to render the current frame (in SLI), or [CU_D3D11_DEVICE_LIST_NEXT_FRAME](#) for the devices used to render the next frame (in SLI).

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#), [CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_NO_DEVICE](#), [CUDA_ERROR_INVALID_VALUE](#), [CUDA_ERROR_NOT_FOUND](#), [CUDA_ERROR_UNKNOWN](#)

Description

Returns in *pCudaDeviceCount the number of CUDA-compatible device corresponding to the Direct3D 11 device pD3D11Device. Also returns in *pCudaDevices at most cudaDeviceCount of the CUDA-compatible devices corresponding to the Direct3D 11 device pD3D11Device.

If any of the GPUs being used to render `pDevice` are not CUDA capable then the call will return `CUDA_ERROR_NO_DEVICE`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuD3D11GetDevice`, `cudaD3D11GetDevices`

CUresult cuGraphicsD3D11RegisterResource (CUgraphicsResource *pCudaResource, ID3D11Resource *pD3DResource, unsigned int Flags)

Register a Direct3D 11 resource for access by CUDA.

Parameters

pCudaResource

- Returned graphics resource handle

pD3DResource

- Direct3D resource to register

Flags

- Parameters for resource registration

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_OUT_OF_MEMORY`, `CUDA_ERROR_UNKNOWN`

Description

Registers the Direct3D 11 resource `pD3DResource` for access by CUDA and returns a CUDA handle to `pD3DResource` in `pCudaResource`. The handle returned in `pCudaResource` may be used to map and unmap this resource until it is unregistered. On success this call will increase the internal reference count on `pD3DResource`. This reference count will be decremented when this resource is unregistered through `cuGraphicsUnregisterResource()`.

This call is potentially high-overhead and should not be called every frame in interactive applications.

The type of `pD3DResource` must be one of the following.

- ▶ ID3D11Buffer: may be accessed through a device pointer.
- ▶ ID3D11Texture1D: individual subresources of the texture may be accessed via arrays
- ▶ ID3D11Texture2D: individual subresources of the texture may be accessed via arrays
- ▶ ID3D11Texture3D: individual subresources of the texture may be accessed via arrays

The `Flags` argument may be used to specify additional parameters at register time. The valid values for this parameter are

- ▶ `CU_GRAPHICS_REGISTER_FLAGS_NONE`: Specifies no hints about how this resource will be used.
- ▶ `CU_GRAPHICS_REGISTER_FLAGS_SURFACE_LDST`: Specifies that CUDA will bind this resource to a surface reference.
- ▶ `CU_GRAPHICS_REGISTER_FLAGS_TEXTURE_GATHER`: Specifies that CUDA will perform texture gather operations on this resource.

Not all Direct3D resources of the above types may be used for interoperability with CUDA. The following are some limitations.

- ▶ The primary rendertarget may not be registered with CUDA.
- ▶ Textures which are not of a format which is 1, 2, or 4 channels of 8, 16, or 32-bit integer or floating-point data cannot be shared.
- ▶ Surfaces of depth or stencil formats cannot be shared.

A complete list of supported DXGI formats is as follows. For compactness the notation `A_{B,C,D}` represents `A_B`, `A_C`, and `A_D`.

- ▶ `DXGI_FORMAT_A8_UNORM`
- ▶ `DXGI_FORMAT_B8G8R8A8_UNORM`
- ▶ `DXGI_FORMAT_B8G8R8X8_UNORM`
- ▶ `DXGI_FORMAT_R16_FLOAT`
- ▶ `DXGI_FORMAT_R16G16B16A16_{FLOAT,SINT,SNORM,UINT,UNORM}`
- ▶ `DXGI_FORMAT_R16G16_{FLOAT,SINT,SNORM,UINT,UNORM}`
- ▶ `DXGI_FORMAT_R16_{SINT,SNORM,UINT,UNORM}`
- ▶ `DXGI_FORMAT_R32_FLOAT`
- ▶ `DXGI_FORMAT_R32G32B32A32_{FLOAT,SINT,UINT}`
- ▶ `DXGI_FORMAT_R32G32_{FLOAT,SINT,UINT}`
- ▶ `DXGI_FORMAT_R32_{SINT,UINT}`
- ▶ `DXGI_FORMAT_R8G8B8A8_{SINT,SNORM,UINT,UNORM,UNORM_SRGB}`
- ▶ `DXGI_FORMAT_R8G8_{SINT,SNORM,UINT,UNORM}`
- ▶ `DXGI_FORMAT_R8_{SINT,SNORM,UINT,UNORM}`

If `pD3DResource` is of incorrect type or is already registered then `CUDA_ERROR_INVALID_HANDLE` is returned. If `pD3DResource` cannot be registered then `CUDA_ERROR_UNKNOWN` is returned. If `Flags` is not one of the above specified value then `CUDA_ERROR_INVALID_VALUE` is returned.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuGraphicsUnregisterResource](#), [cuGraphicsMapResources](#),
[cuGraphicsSubResourceGetMappedArray](#), [cuGraphicsResourceGetMappedPointer](#),
[cudaGraphicsD3D11RegisterResource](#)

4.30.1. Direct3D 11 Interoperability [DEPRECATED]

Direct3D 11 Interoperability

This section describes deprecated Direct3D 11 interoperability functionality.

CUresult cuD3D11CtxCreate (CUcontext *pCtx, CUdevice *pCudaDevice, unsigned int Flags, ID3D11Device *pD3DDevice)

Create a CUDA context for interoperability with Direct3D 11.

Parameters

pCtx

- Returned newly created CUDA context

pCudaDevice

- Returned pointer to the device on which the context was created

Flags

- Context creation flags (see [cuCtxCreate\(\)](#) for details)

pD3DDevice

- Direct3D device to create interoperability context with

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_VALUE](#),
[CUDA_ERROR_OUT_OF_MEMORY](#), [CUDA_ERROR_UNKNOWN](#)

Description

Deprecated This function is deprecated as of CUDA 5.0.

This function is deprecated and should no longer be used. It is no longer necessary to associate a CUDA context with a D3D11 device in order to achieve maximum interoperability performance.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D11GetDevice](#), [cuGraphicsD3D11RegisterResource](#)

CUresult cuD3D11CtxCreateOnDevice (CUcontext *pCtx, unsigned int flags, ID3D11Device *pD3DDevice, CUdevice cudaDevice)

Create a CUDA context for interoperability with Direct3D 11.

Parameters

pCtx

- Returned newly created CUDA context

flags

- Context creation flags (see [cuCtxCreate\(\)](#) for details)

pD3DDevice

- Direct3D device to create interoperability context with

cudaDevice

- The CUDA device on which to create the context. This device must be among the devices returned when querying CU_D3D11_DEVICES_ALL from [cuD3D11GetDevices](#).

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_VALUE,
CUDA_ERROR_OUT_OF_MEMORY, CUDA_ERROR_UNKNOWN

Description

Deprecated This function is deprecated as of CUDA 5.0.

This function is deprecated and should no longer be used. It is no longer necessary to associate a CUDA context with a D3D11 device in order to achieve maximum interoperability performance.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D11GetDevices](#), [cuGraphicsD3D11RegisterResource](#)

CUresult cuD3D11GetDirect3DDevice (ID3D11Device **ppD3DDevice)

Get the Direct3D 11 device against which the current CUDA context was created.

Parameters

ppD3DDevice

- Returned Direct3D device corresponding to CUDA context

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_DEINITIALIZED](#),
[CUDA_ERROR_NOT_INITIALIZED](#), [CUDA_ERROR_INVALID_CONTEXT](#)

Description

Deprecated This function is deprecated as of CUDA 5.0.

This function is deprecated and should no longer be used. It is no longer necessary to associate a CUDA context with a D3D11 device in order to achieve maximum interoperability performance.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuD3D11GetDevice](#)

4.31. VDPAU Interoperability

This section describes the VDPAU interoperability functions of the low-level CUDA driver application programming interface.

CUresult cuGraphicsVDPAURegisterOutputSurface (CUgraphicsResource *pCudaResource, VdpOutputSurface vdpSurface, unsigned int flags)

Registers a VDPAU VdpOutputSurface object.

Parameters

pCudaResource

- Pointer to the returned object handle

vdpSurface

- The VdpOutputSurface to be registered

flags

- Map flags

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_ALREADY_MAPPED, CUDA_ERROR_INVALID_CONTEXT,

Description

Registers the VdpOutputSurface specified by vdpSurface for access by CUDA. A handle to the registered object is returned as pCudaResource. The surface's intended usage is specified using flags, as follows:

- ▶ CU_GRAPHICS_MAP_RESOURCE_FLAGS_NONE: Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA. This is the default value.
- ▶ CU_GRAPHICS_MAP_RESOURCE_FLAGS_READ_ONLY: Specifies that CUDA will not write to this resource.
- ▶ CU_GRAPHICS_MAP_RESOURCE_FLAGS_WRITE_DISCARD: Specifies that CUDA will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.

The VdpOutputSurface is presented as an array of subresources that may be accessed using pointers returned by cuGraphicsSubResourceGetMappedArray. The exact number of valid arrayIndex values depends on the VDPAU surface format. The mapping is shown in the table below. mipLevel must be 0.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuVDPAUCtxCreate](#), [cuGraphicsVDPAURegisterVideoSurface](#),
[cuGraphicsUnregisterResource](#), [cuGraphicsResourceSetMapFlags](#),
[cuGraphicsMapResources](#), [cuGraphicsUnmapResources](#),
[cuGraphicsSubResourceGetMappedArray](#), [cuVDPAUGetDevice](#),
[cudaGraphicsVDPAURegisterOutputSurface](#)

CUresult cuGraphicsVDPAURegisterVideoSurface (CUgraphicsResource *pCudaResource, VdpVideoSurface vdpSurface, unsigned int flags)

Registers a VDPAU VdpVideoSurface object.

Parameters**pCudaResource**

- Pointer to the returned object handle

vdpSurface

- The VdpVideoSurface to be registered

flags

- Map flags

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_HANDLE](#),
[CUDA_ERROR_ALREADY_MAPPED](#), [CUDA_ERROR_INVALID_CONTEXT](#),

Description

Registers the VdpVideoSurface specified by `vdpSurface` for access by CUDA. A handle to the registered object is returned as `pCudaResource`. The surface's intended usage is specified using `flags`, as follows:

- ▶ [CU_GRAPHICS_MAP_RESOURCE_FLAGS_NONE](#): Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA. This is the default value.
- ▶ [CU_GRAPHICS_MAP_RESOURCE_FLAGS_READ_ONLY](#): Specifies that CUDA will not write to this resource.
- ▶ [CU_GRAPHICS_MAP_RESOURCE_FLAGS_WRITE_DISCARD](#): Specifies that CUDA will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.

The VdpVideoSurface is presented as an array of subresources that may be accessed using pointers returned by [cuGraphicsSubResourceGetMappedArray](#). The exact number

of valid `arrayIndex` values depends on the VDPAU surface format. The mapping is shown in the table below. `mipLevel` must be 0.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuVDPAUCtxCreate`, `cuGraphicsVDPAURegisterOutputSurface`,
`cuGraphicsUnregisterResource`, `cuGraphicsResourceSetMapFlags`,
`cuGraphicsMapResources`, `cuGraphicsUnmapResources`,
`cuGraphicsSubResourceGetMappedArray`, `cuVDPAUGetDevice`,
`cudaGraphicsVDPAURegisterVideoSurface`

CUresult cuVDPAUCtxCreate (CUcontext *pCtx, unsigned int flags, CUdevice device, VdpDevice vdpDevice, VdpGetProcAddress *vdpGetProcAddress)

Create a CUDA context for interoperability with VDPAU.

Parameters

pCtx

- Returned CUDA context

flags

- Options for CUDA context creation

device

- Device on which to create the context

vdpDevice

- The VdpDevice to interop with

vdpGetProcAddress

- VDPAU's VdpGetProcAddress function pointer

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`, `CUDA_ERROR_OUT_OF_MEMORY`

Description

Creates a new CUDA context, initializes VDPAU interoperability, and associates the CUDA context with the calling thread. It must be called before performing any other

VDPAU interoperability operations. It may fail if the needed VDPAU driver facilities are not available. For usage of the `flags` parameter, see `cuCtxCreate()`.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

`cuCtxCreate`, `cuGraphicsVDPAURegisterVideoSurface`,
`cuGraphicsVDPAURegisterOutputSurface`, `cuGraphicsUnregisterResource`,
`cuGraphicsResourceSetMapFlags`, `cuGraphicsMapResources`,
`cuGraphicsUnmapResources`, `cuGraphicsSubResourceGetMappedArray`,
`cuVDPAUGetDevice`

CUresult cuVDPAUGetDevice (CUdevice *pDevice, VdpDevice vdpDevice, VdpGetProcAddress *vdpGetProcAddress)

Gets the CUDA device associated with a VDPAU device.

Parameters

pDevice

- Device associated with `vdpDevice`

vdpDevice

- A `VdpDevice` handle

vdpGetProcAddress

- VDPAU's `VdpGetProcAddress` function pointer

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_DEINITIALIZED`,
`CUDA_ERROR_NOT_INITIALIZED`, `CUDA_ERROR_INVALID_CONTEXT`,
`CUDA_ERROR_INVALID_VALUE`

Description

Returns in `*pDevice` the CUDA device associated with a `vdpDevice`, if applicable.



Note that this function may also return error codes from previous, asynchronous launches.

See also:

[cuCtxCreate](#), [cuVDPAUCtxCreate](#), [cuGraphicsVDPAURegisterVideoSurface](#), [cuGraphicsVDPAURegisterOutputSurface](#), [cuGraphicsUnregisterResource](#), [cuGraphicsResourceSetMapFlags](#), [cuGraphicsMapResources](#), [cuGraphicsUnmapResources](#), [cuGraphicsSubResourceGetMappedArray](#), [cudaVDPAUGetDevice](#)

4.32. EGL Interoperability

This section describes the EGL interoperability functions of the low-level CUDA driver application programming interface.

CUresult cuEGLStreamConsumerAcquireFrame (**CUeglStreamConnection *conn**, **CUgraphicsResource *pCudaResource**, **CUstream *pStream**, **unsigned int timeout**)

Acquire an image frame from the EGLStream with CUDA as a consumer.

Parameters**conn**

- Connection on which to acquire

pCudaResource

- CUDA resource on which the stream frame will be mapped for use.

pStream

- CUDA stream for synchronization and any data migrations implied by [CUeglResourceLocationFlags](#).

timeout

- Desired timeout in usec.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_HANDLE](#),

Description

Acquire an image frame from EGLStreamKHR.

[cuGraphicsResourceGetMappedEglFrame](#) can be called on `pCudaResource` to get [CUeglFrame](#).

See also:

`cuEGLStreamConsumerConnect`, `cuEGLStreamConsumerDisconnect`,
`cuEGLStreamConsumerAcquireFrame`, `cuEGLStreamConsumerReleaseFrame`,
`cudaEGLStreamConsumerAcquireFrame`

CUresult cuEGLStreamConsumerConnect (CUeglStreamConnection *conn, EGLStreamKHR stream)

Connect CUDA to EGLStream as a consumer.

Parameters

conn

- Pointer to the returned connection handle

stream

- EGLStreamKHR handle

Returns

`CUDA_SUCCESS`, `CUDA_ERROR_INVALID_HANDLE`,
`CUDA_ERROR_INVALID_CONTEXT`,

Description

Connect CUDA as a consumer to EGLStreamKHR specified by `stream`.

The EGLStreamKHR is an EGL object that transfers a sequence of image frames from one API to another.

See also:

`cuEGLStreamConsumerConnect`, `cuEGLStreamConsumerDisconnect`,
`cuEGLStreamConsumerAcquireFrame`, `cuEGLStreamConsumerReleaseFrame`,
`cudaEGLStreamConsumerConnect`

CUresult cuEGLStreamConsumerConnectWithFlags (CUeglStreamConnection *conn, EGLStreamKHR stream, unsigned int flags)

Connect CUDA to EGLStream as a consumer with given flags.

Parameters

conn

- Pointer to the returned connection handle

stream

- EGLStreamKHR handle

flags

- Flags denote intended location - system or video.

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_INVALID_CONTEXT,

Description

Connect CUDA as a consumer to EGLStreamKHR specified by `stream` with specified `flags` defined by `CUeglResourceLocationFlags`.

The flags specify whether the consumer wants to access frames from system memory or video memory. Default is `CU_EGL_RESOURCE_LOCATION_VIDMEM`.

See also:

`cuEGLStreamConsumerConnect`, `cuEGLStreamConsumerDisconnect`,
`cuEGLStreamConsumerAcquireFrame`, `cuEGLStreamConsumerReleaseFrame`,
`cudaEGLStreamConsumerConnectWithFlags`

CUresult cuEGLStreamConsumerDisconnect (CUeglStreamConnection *conn)

Disconnect CUDA as a consumer to EGLStream .

Parameters**conn**

- Connection to disconnect.

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_INVALID_CONTEXT,

Description

Disconnect CUDA as a consumer to EGLStreamKHR.

The EGLStreamKHR is an EGL object that transfers a sequence of image frames from one API to another.

See also:

cuEGLStreamConsumerConnect, cuEGLStreamConsumerDisconnect,
 cuEGLStreamConsumerAcquireFrame, cuEGLStreamConsumerReleaseFrame,
 cudaEGLStreamConsumerDisconnect

CUresult cuEGLStreamConsumerReleaseFrame (CUeglStreamConnection *conn, CUgraphicsResource pCudaResource, CUstream *pStream)

Releases the last frame acquired from the EGLStream.

Parameters

conn

- Connection on which to release

pCudaResource

- CUDA resource whose corresponding frame is to be released

pStream

- CUDA stream on which release will be done.

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,

Description

Release the acquired image frame specified by pCudaResource to EGLStreamKHR.

The EGLStreamKHR is an EGL object that transfers a sequence of image frames from one API to another.

See also:

cuEGLStreamConsumerConnect, cuEGLStreamConsumerDisconnect,
 cuEGLStreamConsumerAcquireFrame, cuEGLStreamConsumerReleaseFrame,
 cudaEGLStreamConsumerReleaseFrame

CUresult cuEGLStreamProducerConnect (CUeglStreamConnection *conn, EGLStreamKHR stream, EGLint width, EGLint height)

Connect CUDA to EGLStream as a producer.

Parameters

conn

- Pointer to the returned connection handle

stream

- EGLStreamKHR handle

width

- width of the image to be submitted to the stream

height

- height of the image to be submitted to the stream

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_INVALID_CONTEXT,

Description

Connect CUDA as a producer to EGLStreamKHR specified by `stream`.

The EGLStreamKHR is an EGL object that transfers a sequence of image frames from one API to another.

See also:

cuEGLStreamProducerConnect, cuEGLStreamProducerDisconnect,
cuEGLStreamProducerPresentFrame, cudaEGLStreamProducerConnect

CUresult cuEGLStreamProducerDisconnect (CUeglStreamConnection *conn)

Disconnect CUDA as a producer to EGLStream .

Parameters**conn**

- Connection to disconnect.

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_INVALID_CONTEXT,

Description

Disconnect CUDA as a producer to EGLStreamKHR.

The EGLStreamKHR is an EGL object that transfers a sequence of image frames from one API to another.

See also:

[cuEGLStreamProducerConnect](#), [cuEGLStreamProducerDisconnect](#),
[cuEGLStreamProducerPresentFrame](#), [cudaEGLStreamProducerDisconnect](#)

CUresult cuEGLStreamProducerPresentFrame (CUeglStreamConnection *conn, CUeglFrame eglframe, CUstream *pStream)

Present a CUDA eglFrame to the EGLStream with CUDA as a producer.

Parameters

conn

- Connection on which to present the CUDA array

eglframe

- CUDA Eglstream Proucer Frame handle to be sent to the consumer over EglStream.

pStream

- CUDA stream on which to present the frame.

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_HANDLE](#),

Description

The EGLStreamKHR is an EGL object that transfers a sequence of image frames from one API to another.

The [CUeglFrame](#) is defined as:

```
typedef struct CUeglFrame_st {
    union {
        CUarray pArray[MAX_PLANES];
        void* pPitch[MAX_PLANES];
    } frame;
    unsigned int width;
    unsigned int height;
    unsigned int depth;
    unsigned int pitch;
    unsigned int planeCount;
    unsigned int numChannels;
    CUeglFrameType frameType;
    CUeglColorFormat eglColorFormat;
    CUarray_format cuFormat;
} CUeglFrame;
```

See also:

[cuEGLStreamProducerConnect](#), [cuEGLStreamProducerDisconnect](#),
[cuEGLStreamProducerReturnFrame](#), [cudaEGLStreamProducerPresentFrame](#)

CUresult cuEGLStreamProducerReturnFrame (CUeglStreamConnection *conn, CUeglFrame *eglframe, CUstream *pStream)

Return the CUDA eglFrame to the EGLStream released by the consumer.

Parameters

conn

- Connection on which to return

eglframe

- CUDA Eglstream Proucer Frame handle returned from the consumer over EglStream.

pStream

- CUDA stream on which to return the frame.

Returns

CUDA_SUCCESS, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_LAUNCH_TIMEOUT

Description

The EGLStreamKHR is an EGL object that transfers a sequence of image frames from one API to another.

This API can potentially return CUDA_ERROR_LAUNCH_TIMEOUT if the consumer has not returned a frame to EGL stream. If timeout is returned the application can retry.

See also:

cuEGLStreamProducerConnect, cuEGLStreamProducerDisconnect,
cuEGLStreamProducerPresentFrame, cudaEGLStreamProducerReturnFrame

CUresult cuEventCreateFromEGLSync (CUevent *phEvent, EGLSyncKHR eglSync, unsigned int flags)

Creates an event from EGLSync object.

Parameters

phEvent

- Returns newly created event

eglSync

- Opaque handle to EGLSync object

flags

- Event creation flags

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
 CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
 CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_OUT_OF_MEMORY

Description

Creates an event *phEvent from an EGLSyncKHR eglSync with the flags specified via flags. Valid flags include:

- ▶ **CU_EVENT_DEFAULT**: Default event creation flag.
- ▶ **CU_EVENT_BLOCKING_SYNC**: Specifies that the created event should use blocking synchronization. A CPU thread that uses `cuEventSynchronize()` to wait on an event created with this flag will block until the event has actually been completed.

`cuEventRecord` and `TimingData` are not supported for events created from `EGLSync`.

The `EGLSyncKHR` is an opaque handle to an EGL sync object. `typedef void* EGLSyncKHR`

See also:

`cuEventQuery`, `cuEventSynchronize`, `cuEventDestroy`

CUresult cuGraphicsEGLRegisterImage (CUgraphicsResource *pCudaResource, EGLImageKHR image, unsigned int flags)

Registers an EGL image.

Parameters**pCudaResource**

- Pointer to the returned object handle

image

- An `EGLImageKHR` image which can be used to create target resource.

flags

- Map flags

Returns

[CUDA_SUCCESS](#), [CUDA_ERROR_INVALID_HANDLE](#),
[CUDA_ERROR_ALREADY_MAPPED](#), [CUDA_ERROR_INVALID_CONTEXT](#),

Description

Registers the EGLImageKHR specified by `image` for access by CUDA. A handle to the registered object is returned as `pCudaResource`. Additional Mapping/Unmapping is not required for the registered resource and [cuGraphicsResourceGetMappedEglFrame](#) can be directly called on the `pCudaResource`.

The application will be responsible for synchronizing access to shared objects. The application must ensure that any pending operation which access the objects have completed before passing control to CUDA. This may be accomplished by issuing and waiting for `glFinish` command on all GLcontexts (for OpenGL and likewise for other APIs). The application will be also responsible for ensuring that any pending operation on the registered CUDA resource has completed prior to executing subsequent commands in other APIs accessing the same memory objects. This can be accomplished by calling `cuCtxSynchronize` or `cuEventSynchronize` (preferably).

The surface's intended usage is specified using `flags`, as follows:

- ▶ `CU_GRAPHICS_MAP_RESOURCE_FLAGS_NONE`: Specifies no hints about how this resource will be used. It is therefore assumed that this resource will be read from and written to by CUDA. This is the default value.
- ▶ `CU_GRAPHICS_MAP_RESOURCE_FLAGS_READ_ONLY`: Specifies that CUDA will not write to this resource.
- ▶ `CU_GRAPHICS_MAP_RESOURCE_FLAGS_WRITE_DISCARD`: Specifies that CUDA will not read from this resource and will write over the entire contents of the resource, so none of the data previously stored in the resource will be preserved.

The EGLImageKHR is an object which can be used to create EGLImage target resource. It is defined as a void pointer. `typedef void* EGLImageKHR`

See also:

[cuGraphicsEGLRegisterImage](#), [cuGraphicsUnregisterResource](#),
[cuGraphicsResourceSetMapFlags](#), [cuGraphicsMapResources](#),
[cuGraphicsUnmapResources](#), [cudaGraphicsEGLRegisterImage](#)

CUresult cuGraphicsResourceGetMappedEglFrame (CUeglFrame *eglFrame, CUgraphicsResource resource, unsigned int index, unsigned int mipLevel)

Get an eglFrame through which to access a registered EGL graphics resource.

Parameters

eglFrame

- Returned eglFrame.

resource

- Registered resource to access.

index

- Index for cubemap surfaces.

mipLevel

- Mipmap level for the subresource to access.

Returns

CUDA_SUCCESS, CUDA_ERROR_DEINITIALIZED,
CUDA_ERROR_NOT_INITIALIZED, CUDA_ERROR_INVALID_CONTEXT,
CUDA_ERROR_INVALID_VALUE, CUDA_ERROR_INVALID_HANDLE,
CUDA_ERROR_NOT_MAPPED

Description

Returns in *eglFrame an eglFrame pointer through which the registered graphics resource resource may be accessed. This API can only be called for EGL graphics resources.

The CUeglFrame is defined as:

```
↑ typedef struct CUeglFrame_st {
    union {
        CUarray pArray[MAX_PLANES];
        void* pPitch[MAX_PLANES];
    } frame;
    unsigned int width;
    unsigned int height;
    unsigned int depth;
    unsigned int pitch;
    unsigned int planeCount;
    unsigned int numChannels;
    CUeglFrameType frameType;
    CUeglColorFormat eglColorFormat;
    CUarray_format cuFormat;
} CUeglFrame;
```

If resource is not registered then CUDA_ERROR_NOT_MAPPED is returned. *

See also:

`cuGraphicsMapResources`, `cuGraphicsSubResourceGetMappedArray`,
`cuGraphicsResourceGetMappedPointer`, `cudaGraphicsResourceGetMappedEglFrame`

Chapter 5.

DATA STRUCTURES

Here are the data structures with brief descriptions:

`CUDA_ARRAY3D_DESCRIPTOR`
`CUDA_ARRAY_DESCRIPTOR`
`CUDA_LAUNCH_PARAMS`
`CUDA_MEMCPY2D`
`CUDA_MEMCPY3D`
`CUDA_MEMCPY3D_PEER`
`CUDA_POINTER_ATTRIBUTE_P2P_TOKENS`
`CUDA_RESOURCE_DESC`
`CUDA_RESOURCE_VIEW_DESC`
`CUDA_TEXTURE_DESC`
`CUdevprop`
`CUeglFrame`
`CUipcEventHandle`
`CUipcMemHandle`
`CUstreamBatchMemOpParams`

5.1. `CUDA_ARRAY3D_DESCRIPTOR` Struct Reference

3D array descriptor

`size_t` `CUDA_ARRAY3D_DESCRIPTOR::Depth`

Depth of 3D array

`unsigned int` `CUDA_ARRAY3D_DESCRIPTOR::Flags`

Flags

CUarray_format CUDA_ARRAY3D_DESCRIPTOR::Format

Array format

size_t CUDA_ARRAY3D_DESCRIPTOR::Height

Height of 3D array

unsigned int CUDA_ARRAY3D_DESCRIPTOR::NumChannels

Channels per array element

size_t CUDA_ARRAY3D_DESCRIPTOR::Width

Width of 3D array

5.2. CUDA_ARRAY_DESCRIPTOR Struct Reference

Array descriptor

CUarray_format CUDA_ARRAY_DESCRIPTOR::Format

Array format

size_t CUDA_ARRAY_DESCRIPTOR::Height

Height of array

unsigned int CUDA_ARRAY_DESCRIPTOR::NumChannels

Channels per array element

size_t CUDA_ARRAY_DESCRIPTOR::Width

Width of array

5.3. CUDA_LAUNCH_PARAMS Struct Reference

Kernel launch parameters

unsigned int CUDA_LAUNCH_PARAMS::blockDimX

X dimension of each thread block

unsigned int CUDA_LAUNCH_PARAMS::blockDimY

Y dimension of each thread block

unsigned int CUDA_LAUNCH_PARAMS::blockDimZ

Z dimension of each thread block

CUfunction CUDA_LAUNCH_PARAMS::function

Kernel to launch

unsigned int CUDA_LAUNCH_PARAMS::gridDimX

Width of grid in blocks

unsigned int CUDA_LAUNCH_PARAMS::gridDimY

Height of grid in blocks

unsigned int CUDA_LAUNCH_PARAMS::gridDimZ

Depth of grid in blocks

CUstream CUDA_LAUNCH_PARAMS::hStream

Stream identifier

****CUDA_LAUNCH_PARAMS::kernelParams**

Array of pointers to kernel parameters

unsigned int CUDA_LAUNCH_PARAMS::sharedMemBytes

Dynamic shared-memory size per thread block in bytes

5.4. CUDA_MEMCPY2D Struct Reference

2D memory copy parameters

CUarray CUDA_MEMCPY2D::dstArray

Destination array reference

CUdeviceptr CUDA_MEMCPY2D::dstDevice

Destination device pointer

void *CUDA_MEMCPY2D::dstHost

Destination host pointer

CUmemorytype CUDA_MEMCPY2D::dstMemoryType

Destination memory type (host, device, array)

size_t CUDA_MEMCPY2D::dstPitch

Destination pitch (ignored when dst is array)

size_t CUDA_MEMCPY2D::dstXInBytes

Destination X in bytes

size_t CUDA_MEMCPY2D::dstY

Destination Y

size_t CUDA_MEMCPY2D::Height

Height of 2D memory copy

CUarray CUDA_MEMCPY2D::srcArray

Source array reference

CUdeviceptr CUDA_MEMCPY2D::srcDevice

Source device pointer

const void *CUDA_MEMCPY2D::srcHost

Source host pointer

CUmemorytype CUDA_MEMCPY2D::srcMemoryType

Source memory type (host, device, array)

size_t CUDA_MEMCPY2D::srcPitch

Source pitch (ignored when src is array)

size_t CUDA_MEMCPY2D::srcXInBytes

Source X in bytes

size_t CUDA_MEMCPY2D::srcY

Source Y

size_t CUDA_MEMCPY2D::WidthInBytes

Width of 2D memory copy in bytes

5.5. CUDA_MEMCPY3D Struct Reference

3D memory copy parameters

size_t CUDA_MEMCPY3D::Depth

Depth of 3D memory copy

CUarray CUDA_MEMCPY3D::dstArray

Destination array reference

CUdeviceptr CUDA_MEMCPY3D::dstDevice

Destination device pointer

size_t CUDA_MEMCPY3D::dstHeight

Destination height (ignored when dst is array; may be 0 if Depth==1)

void *CUDA_MEMCPY3D::dstHost

Destination host pointer

size_t CUDA_MEMCPY3D::dstLOD

Destination LOD

CUmemorytype CUDA_MEMCPY3D::dstMemoryType

Destination memory type (host, device, array)

size_t CUDA_MEMCPY3D::dstPitch

Destination pitch (ignored when dst is array)

size_t CUDA_MEMCPY3D::dstXInBytes

Destination X in bytes

size_t CUDA_MEMCPY3D::dstY

Destination Y

size_t CUDA_MEMCPY3D::dstZ

Destination Z

size_t CUDA_MEMCPY3D::Height

Height of 3D memory copy

void *CUDA_MEMCPY3D::reserved0

Must be NULL

void *CUDA_MEMCPY3D::reserved1

Must be NULL

CUarray CUDA_MEMCPY3D::srcArray

Source array reference

CUdeviceptr CUDA_MEMCPY3D::srcDevice

Source device pointer

size_t CUDA_MEMCPY3D::srcHeight

Source height (ignored when src is array; may be 0 if Depth==1)

const void *CUDA_MEMCPY3D::srcHost

Source host pointer

size_t CUDA_MEMCPY3D::srcLOD

Source LOD

CUmemorytype CUDA_MEMCPY3D::srcMemoryType

Source memory type (host, device, array)

size_t CUDA_MEMCPY3D::srcPitch

Source pitch (ignored when src is array)

size_t CUDA_MEMCPY3D::srcXInBytes

Source X in bytes

size_t CUDA_MEMCPY3D::srcY

Source Y

size_t CUDA_MEMCPY3D::srcZ

Source Z

size_t CUDA_MEMCPY3D::WidthInBytes

Width of 3D memory copy in bytes

5.6. CUDA_MEMCPY3D_PEER Struct Reference

3D memory cross-context copy parameters

size_t CUDA_MEMCPY3D_PEER::Depth

Depth of 3D memory copy

CUarray CUDA_MEMCPY3D_PEER::dstArray

Destination array reference

CUcontext CUDA_MEMCPY3D_PEER::dstContext

Destination context (ignored with dstMemoryType is `CU_MEMORYTYPE_ARRAY`)

CUdeviceptr CUDA_MEMCPY3D_PEER::dstDevice

Destination device pointer

size_t CUDA_MEMCPY3D_PEER::dstHeight

Destination height (ignored when dst is array; may be 0 if Depth==1)

void *CUDA_MEMCPY3D_PEER::dstHost

Destination host pointer

size_t CUDA_MEMCPY3D_PEER::dstLOD

Destination LOD

CUmemorytype CUDA_MEMCPY3D_PEER::dstMemoryType

Destination memory type (host, device, array)

size_t CUDA_MEMCPY3D_PEER::dstPitch

Destination pitch (ignored when dst is array)

size_t CUDA_MEMCPY3D_PEER::dstXInBytes

Destination X in bytes

size_t CUDA_MEMCPY3D_PEER::dstY

Destination Y

size_t CUDA_MEMCPY3D_PEER::dstZ

Destination Z

size_t CUDA_MEMCPY3D_PEER::Height

Height of 3D memory copy

CUarray CUDA_MEMCPY3D_PEER::srcArray

Source array reference

CUcontext CUDA_MEMCPY3D_PEER::srcContext

Source context (ignored with srcMemoryType is CU_MEMORYTYPE_ARRAY)

CUdeviceptr CUDA_MEMCPY3D_PEER::srcDevice

Source device pointer

size_t CUDA_MEMCPY3D_PEER::srcHeight

Source height (ignored when src is array; may be 0 if Depth==1)

const void *CUDA_MEMCPY3D_PEER::srcHost

Source host pointer

size_t CUDA_MEMCPY3D_PEER::srcLOD

Source LOD

CUmemorytype CUDA_MEMCPY3D_PEER::srcMemoryType

Source memory type (host, device, array)

size_t CUDA_MEMCPY3D_PEER::srcPitch

Source pitch (ignored when src is array)

size_t CUDA_MEMCPY3D_PEER::srcXInBytes

Source X in bytes

size_t CUDA_MEMCPY3D_PEER::srcY

Source Y

`size_t CUDA_MEMCPY3D_PEER::srcZ`

Source Z

`size_t CUDA_MEMCPY3D_PEER::WidthInBytes`

Width of 3D memory copy in bytes

5.7. CUDA_POINTER_ATTRIBUTE_P2P_TOKENS Struct Reference

GPU Direct v3 tokens

5.8. CUDA_RESOURCE_DESC Struct Reference

CUDA Resource descriptor

`CUdeviceptr CUDA_RESOURCE_DESC::devPtr`

Device pointer

`unsigned int CUDA_RESOURCE_DESC::flags`

Flags (must be zero)

`CUarray_format CUDA_RESOURCE_DESC::format`

Array format

`CUarray CUDA_RESOURCE_DESC::hArray`

CUDA array

`size_t CUDA_RESOURCE_DESC::height`

Height of the array in elements

`CUmipmappedArray`

`CUDA_RESOURCE_DESC::hMipmappedArray`

CUDA mipmapped array

`unsigned int CUDA_RESOURCE_DESC::numChannels`

Channels per array element

`size_t CUDA_RESOURCE_DESC::pitchInBytes`

Pitch between two rows in bytes

`CUresourcetype CUDA_RESOURCE_DESC::resType`

Resource type

`size_t CUDA_RESOURCE_DESC::sizeInBytes`

Size in bytes

`size_t CUDA_RESOURCE_DESC::width`

Width of the array in elements

5.9. CUDA_RESOURCE_VIEW_DESC Struct Reference

Resource view descriptor

`size_t CUDA_RESOURCE_VIEW_DESC::depth`

Depth of the resource view

`unsigned int CUDA_RESOURCE_VIEW_DESC::firstLayer`

First layer index

`unsigned int
CUDA_RESOURCE_VIEW_DESC::firstMipmapLevel`

First defined mipmap level

`CUresourceViewFormat
CUDA_RESOURCE_VIEW_DESC::format`

Resource view format

`size_t CUDA_RESOURCE_VIEW_DESC::height`

Height of the resource view

`unsigned int CUDA_RESOURCE_VIEW_DESC::lastLayer`

Last layer index

`unsigned int
CUDA_RESOURCE_VIEW_DESC::lastMipmapLevel`

Last defined mipmap level

`size_t CUDA_RESOURCE_VIEW_DESC::width`

Width of the resource view

5.10. CUDA_TEXTURE_DESC Struct Reference

Texture descriptor

`CUaddress_mode CUDA_TEXTURE_DESC::addressMode`

Address modes

`float CUDA_TEXTURE_DESC::borderColor`

Border Color

`CUfilter_mode CUDA_TEXTURE_DESC::filterMode`

Filter mode

`unsigned int CUDA_TEXTURE_DESC::flags`

Flags

`unsigned int CUDA_TEXTURE_DESC::maxAnisotropy`

Maximum anisotropy ratio

float CUDA_TEXTURE_DESC::maxMipmapLevelClamp

Mipmap maximum level clamp

float CUDA_TEXTURE_DESC::minMipmapLevelClamp

Mipmap minimum level clamp

CUfilter_mode

CUDA_TEXTURE_DESC::mipmapFilterMode

Mipmap filter mode

float CUDA_TEXTURE_DESC::mipmapLevelBias

Mipmap level bias

5.11. CUdevprop Struct Reference

Legacy device properties

int CUdevprop::clockRate

Clock frequency in kilohertz

int CUdevprop::maxGridSize

Maximum size of each dimension of a grid

int CUdevprop::maxThreadsDim

Maximum size of each dimension of a block

int CUdevprop::maxThreadsPerBlock

Maximum number of threads per block

int CUdevprop::memPitch

Maximum pitch in bytes allowed by memory copies

int CUdevprop::regsPerBlock

32-bit registers available per block

int CUdevprop::sharedMemPerBlock

Shared memory available per block in bytes

int CUdevprop::SIMDWidth

Warp size in threads

int CUdevprop::textureAlign

Alignment requirement for textures

int CUdevprop::totalConstantMemory

Constant memory available on device in bytes

5.13. CUipcEventHandle Struct Reference

CUDA IPC event handle

5.14. CUipcMemHandle Struct Reference

CUDA IPC mem handle

5.15. CUstreamBatchMemOpParams Union Reference

Per-operation parameters for `cuStreamBatchMemOp`

5.12. Difference between the driver and runtime APIs

The driver and runtime APIs are very similar and can for the most part be used interchangeably. However, there are some key differences worth noting between the two.

Complexity vs. control

The runtime API eases device code management by providing implicit initialization, context management, and module management. This leads to simpler code, but it also lacks the level of control that the driver API has.

In comparison, the driver API offers more fine-grained control, especially over contexts and module loading. Kernel launches are much more complex to implement, as the execution configuration and kernel parameters must be specified with explicit function calls. However, unlike the runtime, where all the kernels are automatically loaded during initialization and stay loaded for as long as the program runs, with the driver API it is possible to only keep the modules that are currently needed loaded, or even dynamically reload modules. The driver API is also language-independent as it only deals with cubin objects.

Context management

Context management can be done through the driver API, but is not exposed in the runtime API. Instead, the runtime API decides itself which context to use for a thread: if a context has been made current to the calling thread through the driver API, the runtime will use that, but if there is no such context, it uses a "primary context." Primary contexts are created as needed, one per device per process, are reference-counted, and are then destroyed when there are no more references to them. Within one process, all users of the runtime API will share the primary context, unless a context has been made current to each thread. The context that the runtime uses, i.e, either the current context or primary context, can be synchronized with `cudaDeviceSynchronize()`, and destroyed with `cudaDeviceReset()`.

Using the runtime API with primary contexts has its tradeoffs, however. It can cause trouble for users writing plug-ins for larger software packages, for example, because if all plug-ins run in the same process, they will all share a context but will likely have no way to communicate with each other. So, if one of them calls `cudaDeviceReset()` after finishing all its CUDA work, the other plug-ins will fail because the context they were using was destroyed without their knowledge. To avoid this issue, CUDA clients can use the driver API to create and set the current context, and then use the runtime API to work with it. However, contexts may consume significant resources, such as device memory, extra host threads, and performance costs of context switching on the device. This runtime-driver context sharing is important when using the driver API in conjunction with libraries built on the runtime API, such as cuBLAS or cuFFT.

Chapter 6.

DATA FIELDS

Here is a list of all documented struct and union fields with links to the struct/union documentation for each field:

A

addressMode

[CUDA_TEXTURE_DESC](#)

B

blockDimX

[CUDA_LAUNCH_PARAMS](#)

blockDimY

[CUDA_LAUNCH_PARAMS](#)

blockDimZ

[CUDA_LAUNCH_PARAMS](#)

borderColor

[CUDA_TEXTURE_DESC](#)

C

clockRate

[CUdevprop](#)

cuFormat

[CUeglFrame](#)

D

depth

[CUDA_RESOURCE_VIEW_DESC](#)

[CUeglFrame](#)

Depth

[CUDA_MEMCPY3D_PEER](#)

[CUDA_ARRAY3D_DESCRIPTOR](#)

CUDA_MEMCPY3D

devPtr

CUDA_RESOURCE_DESC

dstArray

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

CUDA_MEMCPY2D

dstContext

CUDA_MEMCPY3D_PEER

dstDevice

CUDA_MEMCPY2D

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

dstHeight

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

dstHost

CUDA_MEMCPY2D

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

dstLOD

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

dstMemoryType

CUDA_MEMCPY2D

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

dstPitch

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

CUDA_MEMCPY2D

dstXInBytes

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

CUDA_MEMCPY2D

dstY

CUDA_MEMCPY3D

CUDA_MEMCPY2D

CUDA_MEMCPY3D_PEER

dstZ

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

E**eglColorFormat**

CUeglFrame

F**filterMode**

CUDA_TEXTURE_DESC

firstLayer

CUDA_RESOURCE_VIEW_DESC

firstMipmapLevel

CUDA_RESOURCE_VIEW_DESC

flags

CUDA_RESOURCE_DESC

CUDA_TEXTURE_DESC

Flags

CUDA_ARRAY3D_DESCRIPTOR

Format

CUDA_ARRAY_DESCRIPTOR

format

CUDA_RESOURCE_VIEW_DESC

Format

CUDA_ARRAY3D_DESCRIPTOR

format

CUDA_RESOURCE_DESC

frameType

CUeglFrame

function

CUDA_LAUNCH_PARAMS

G**gridDimX**

CUDA_LAUNCH_PARAMS

gridDimY

CUDA_LAUNCH_PARAMS

gridDimZ

CUDA_LAUNCH_PARAMS

H**hArray**

CUDA_RESOURCE_DESC

height

CUDA_RESOURCE_DESC

CUeglFrame

Height

CUDA_ARRAY_DESCRIPTOR
 CUDA_ARRAY3D_DESCRIPTOR
 CUDA_MEMCPY3D_PEER
 CUDA_MEMCPY3D
 CUDA_MEMCPY2D

height

CUDA_RESOURCE_VIEW_DESC

hMipmappedArray

CUDA_RESOURCE_DESC

hStream

CUDA_LAUNCH_PARAMS

K**kernelParams**

CUDA_LAUNCH_PARAMS

L**lastLayer**

CUDA_RESOURCE_VIEW_DESC

lastMipmapLevel

CUDA_RESOURCE_VIEW_DESC

M**maxAnisotropy**

CUDA_TEXTURE_DESC

maxGridSize

CUdevprop

maxMipmapLevelClamp

CUDA_TEXTURE_DESC

maxThreadsDim

CUdevprop

maxThreadsPerBlock

CUdevprop

memPitch

CUdevprop

minMipmapLevelClamp

CUDA_TEXTURE_DESC

mipmapFilterMode

CUDA_TEXTURE_DESC

mipmapLevelBias

CUDA_TEXTURE_DESC

N**NumChannels**

CUDA_ARRAY_DESCRIPTOR

numChannels

CUeglFrame

CUDA_RESOURCE_DESC

NumChannels

CUDA_ARRAY3D_DESCRIPTOR

P**pArray**

CUeglFrame

pitch

CUeglFrame

pitchInBytes

CUDA_RESOURCE_DESC

planeCount

CUeglFrame

pPitch

CUeglFrame

R**regsPerBlock**

CUdevprop

reserved0

CUDA_MEMCPY3D

reserved1

CUDA_MEMCPY3D

resType

CUDA_RESOURCE_DESC

S**sharedMemBytes**

CUDA_LAUNCH_PARAMS

sharedMemPerBlock

CUdevprop

SIMDWidth

CUdevprop

sizeInBytes

CUDA_RESOURCE_DESC

srcArray

CUDA_MEMCPY2D

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

srcContext

CUDA_MEMCPY3D_PEER

srcDevice

CUDA_MEMCPY2D

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

srcHeight

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

srcHost

CUDA_MEMCPY3D_PEER

CUDA_MEMCPY2D

CUDA_MEMCPY3D

srcLOD

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

srcMemoryType

CUDA_MEMCPY2D

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

srcPitch

CUDA_MEMCPY3D

CUDA_MEMCPY2D

CUDA_MEMCPY3D_PEER

srcXInBytes

CUDA_MEMCPY3D_PEER

CUDA_MEMCPY3D

CUDA_MEMCPY2D

srcY

CUDA_MEMCPY2D

CUDA_MEMCPY3D

CUDA_MEMCPY3D_PEER

srcZ

CUDA_MEMCPY3D_PEER

CUDA_MEMCPY3D

T

textureAlign

CUdevprop

totalConstantMemory

CUdevprop

W**Width**

CUDA_ARRAY_DESCRIPTOR
CUDA_ARRAY3D_DESCRIPTOR

width

CUDA_RESOURCE_VIEW_DESC
CUeglFrame
CUDA_RESOURCE_DESC

WidthInBytes

CUDA_MEMCPY3D_PEER
CUDA_MEMCPY3D
CUDA_MEMCPY2D

Chapter 7.

DEPRECATED LIST

Global CU_CTX_BLOCKING_SYNC

This flag was deprecated as of CUDA 4.0 and was replaced with CU_CTX_SCHED_BLOCKING_SYNC.

Global CUDA_ERROR_PROFILER_NOT_INITIALIZED

This error return is deprecated as of CUDA 5.0. It is no longer an error to attempt to enable/disable the profiling via cuProfilerStart or cuProfilerStop without initialization.

Global CUDA_ERROR_PROFILER_ALREADY_STARTED

This error return is deprecated as of CUDA 5.0. It is no longer an error to call cuProfilerStart() when profiling is already enabled.

Global CUDA_ERROR_PROFILER_ALREADY_STOPPED

This error return is deprecated as of CUDA 5.0. It is no longer an error to call cuProfilerStop() when profiling is already disabled.

Global CUDA_ERROR_CONTEXT_ALREADY_CURRENT

This error return is deprecated as of CUDA 3.2. It is no longer an error to attempt to push the active context via cuCtxPushCurrent().

Global cuDeviceComputeCapability

Global cuDeviceGetProperties

Global cuCtxAttach

Global cuCtxDetach

Global cuFuncSetBlockShape

Global cuFuncSetSharedSize

Global cuLaunch

Global cuLaunchGrid

Global cuLaunchGridAsync

Global cuParamSetf

Global cuParamSeti

Global cuParamSetSize

Global cuParamSetTexRef

Global cuParamSetv

Global cuTexRefCreate**Global cuTexRefDestroy****Global cuGLCtxCreate**

This function is deprecated as of Cuda 5.0.

Global cuGLInit

This function is deprecated as of Cuda 3.0.

Global cuGLMapBufferObject

This function is deprecated as of Cuda 3.0.

Global cuGLMapBufferObjectAsync

This function is deprecated as of Cuda 3.0.

Global cuGLRegisterBufferObject

This function is deprecated as of Cuda 3.0.

Global cuGLSetBufferObjectMapFlags

This function is deprecated as of Cuda 3.0.

Global cuGLUnmapBufferObject

This function is deprecated as of Cuda 3.0.

Global cuGLUnmapBufferObjectAsync

This function is deprecated as of Cuda 3.0.

Global cuGLUnregisterBufferObject

This function is deprecated as of Cuda 3.0.

Global cuD3D9MapResources

This function is deprecated as of CUDA 3.0.

Global cuD3D9RegisterResource

This function is deprecated as of CUDA 3.0.

Global cuD3D9ResourceGetMappedArray

This function is deprecated as of CUDA 3.0.

Global cuD3D9ResourceGetMappedPitch

This function is deprecated as of CUDA 3.0.

Global cuD3D9ResourceGetMappedPointer

This function is deprecated as of CUDA 3.0.

Global cuD3D9ResourceGetMappedSize

This function is deprecated as of CUDA 3.0.

Global cuD3D9ResourceGetSurfaceDimensions

This function is deprecated as of CUDA 3.0.

Global cuD3D9ResourceSetMapFlags

This function is deprecated as of Cuda 3.0.

Global cuD3D9UnmapResources

This function is deprecated as of CUDA 3.0.

Global cuD3D9UnregisterResource

This function is deprecated as of CUDA 3.0.

Global cuD3D10CtxCreate

This function is deprecated as of CUDA 5.0.

Global cuD3D10CtxCreateOnDevice

This function is deprecated as of CUDA 5.0.

Global cuD3D10GetDirect3DDevice

This function is deprecated as of CUDA 5.0.

Global cuD3D10MapResources

This function is deprecated as of CUDA 3.0.

Global cuD3D10RegisterResource

This function is deprecated as of CUDA 3.0.

Global cuD3D10ResourceGetMappedArray

This function is deprecated as of CUDA 3.0.

Global cuD3D10ResourceGetMappedPitch

This function is deprecated as of CUDA 3.0.

Global cuD3D10ResourceGetMappedPointer

This function is deprecated as of CUDA 3.0.

Global cuD3D10ResourceGetMappedSize

This function is deprecated as of CUDA 3.0.

Global cuD3D10ResourceGetSurfaceDimensions

This function is deprecated as of CUDA 3.0.

Global cuD3D10ResourceSetMapFlags

This function is deprecated as of CUDA 3.0.

Global cuD3D10UnmapResources

This function is deprecated as of CUDA 3.0.

Global cuD3D10UnregisterResource

This function is deprecated as of CUDA 3.0.

Global cuD3D11CtxCreate

This function is deprecated as of CUDA 5.0.

Global cuD3D11CtxCreateOnDevice

This function is deprecated as of CUDA 5.0.

Global cuD3D11GetDirect3DDevice

This function is deprecated as of CUDA 5.0.

Notice

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

Trademarks

NVIDIA and the NVIDIA logo are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2007-2017 NVIDIA Corporation. All rights reserved.