<table>
<thead>
<tr>
<th>2.9</th>
<th>Textures .......................................................................................................................... 38</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.9.1</td>
<td>Texture Buffers .............................................................................................................. 40</td>
</tr>
<tr>
<td>2.10</td>
<td>Samplers .......................................................................................................................... 41</td>
</tr>
<tr>
<td>2.11</td>
<td>Imageblocks ..................................................................................................................... 43</td>
</tr>
<tr>
<td>2.12</td>
<td>Aggregate Types .............................................................................................................. 45</td>
</tr>
<tr>
<td>2.12.1</td>
<td>Arrays of Textures, Texture Buffers, and Samplers .......................................................... 45</td>
</tr>
<tr>
<td>2.12.1.1</td>
<td>Array Element Access with its Operator .......................................................................... 46</td>
</tr>
<tr>
<td>2.12.1.2</td>
<td>Array Capacity ............................................................................................................... 47</td>
</tr>
<tr>
<td>2.12.1.3</td>
<td>Constructors for Templated Arrays ................................................................................ 47</td>
</tr>
<tr>
<td>2.12.2</td>
<td>Structures of Buffers, Textures, and Samplers ................................................................ 48</td>
</tr>
<tr>
<td>2.13</td>
<td>Argument Buffers ............................................................................................................ 49</td>
</tr>
<tr>
<td>2.13.1</td>
<td>Tier 2 Hardware Support for Argument Buffers ............................................................... 51</td>
</tr>
<tr>
<td>2.14</td>
<td>Uniform Type .................................................................................................................... 52</td>
</tr>
<tr>
<td>2.14.1</td>
<td>The Need for a Uniform Type .......................................................................................... 52</td>
</tr>
<tr>
<td>2.14.2</td>
<td>Behavior of the Uniform Type ......................................................................................... 53</td>
</tr>
<tr>
<td>2.14.3</td>
<td>Uniform Control Flow ....................................................................................................... 54</td>
</tr>
<tr>
<td>2.15</td>
<td>Visible Function Table ................................................................................................... 55</td>
</tr>
<tr>
<td>2.16</td>
<td>Function Group Attribute .............................................................................................. 56</td>
</tr>
<tr>
<td>2.17</td>
<td>Ray Tracing Types .......................................................................................................... 56</td>
</tr>
<tr>
<td>2.17.1</td>
<td>Ray Tracing Intersection Tags ......................................................................................... 56</td>
</tr>
<tr>
<td>2.17.2</td>
<td>Ray Type .......................................................................................................................... 60</td>
</tr>
<tr>
<td>2.17.3</td>
<td>Intersection Function Table ............................................................................................ 60</td>
</tr>
<tr>
<td>2.17.4</td>
<td>Intersection Result Type ................................................................................................ 62</td>
</tr>
<tr>
<td>2.17.5</td>
<td>Intersector Type ............................................................................................................. 63</td>
</tr>
<tr>
<td>2.17.6</td>
<td>Acceleration Structure Type ............................................................................................ 63</td>
</tr>
<tr>
<td>2.17.7</td>
<td>Intersection Query Type ................................................................................................ 66</td>
</tr>
<tr>
<td>2.18</td>
<td>Interpolant Type ............................................................................................................. 66</td>
</tr>
<tr>
<td>2.19</td>
<td>Mesh Shader Types .......................................................................................................... 67</td>
</tr>
<tr>
<td>2.19.1</td>
<td>Mesh Grid Property Type ................................................................................................ 67</td>
</tr>
<tr>
<td>2.19.2</td>
<td>Mesh Type ........................................................................................................................ 67</td>
</tr>
<tr>
<td>2.20</td>
<td>Type Conversions and Re-interpretating Data .................................................................. 71</td>
</tr>
<tr>
<td>2.21</td>
<td>Implicit Type Conversions ............................................................................................. 72</td>
</tr>
<tr>
<td>3</td>
<td>Operators .......................................................................................................................... 73</td>
</tr>
<tr>
<td>3.1</td>
<td>Scalar and Vector Operators .......................................................................................... 73</td>
</tr>
<tr>
<td>3.2</td>
<td>Matrix Operators ............................................................................................................. 76</td>
</tr>
<tr>
<td>4</td>
<td>Address Spaces ............................................................................................................... 79</td>
</tr>
<tr>
<td>4.1</td>
<td>device Address Space ...................................................................................................... 79</td>
</tr>
<tr>
<td>4.2</td>
<td>constant Address Space .................................................................................................. 80</td>
</tr>
<tr>
<td>4.3</td>
<td>thread Address Space ..................................................................................................... 80</td>
</tr>
<tr>
<td>4.4</td>
<td>threadgroup Address Space ............................................................................................ 81</td>
</tr>
<tr>
<td>4.4.1</td>
<td>SIMD-groups and Quad-groups ...................................................................................... 82</td>
</tr>
</tbody>
</table>
6.16.3 Copying Commands of an Indirect Command Buffer ........................................ 252
6.17 Variable Rasterization Rate ................................................................................. 253
6.18 Ray Tracing Functions ......................................................................................... 254
6.18.1 Acceleration Structure Functions ...................................................................... 254
6.18.2 Intersector Intersect Functions ......................................................................... 255
6.18.3 Intersector functions to control traversal behavior ........................................... 259
6.18.4 Intersection Query Functions ............................................................................ 263
6.18.5 Indirect Instance Descriptors ............................................................................ 270
6.18.6 Curve Utility Functions ...................................................................................... 271

7 Numerical Compliance .......................................................................................... 273
7.1 INF, NaN, and Denormalized Numbers ................................................................. 273
7.2 Rounding Mode ...................................................................................................... 273
7.3 Floating-Point Exceptions ...................................................................................... 273
7.4 Relative Error as ULPs ......................................................................................... 273
7.5 Edge Case Behavior in Flush to Zero Mode .......................................................... 280
7.6 Conversion Rules for Floating-Point and Integer Types ........................................ 281
7.7 Texture Addressing and Conversion Rules ........................................................... 281
7.7.1 Conversion Rules for Normalized Integer Pixel Data Types ............................... 281
7.7.1.1 Converting Normalized Integer Pixel Data Types to Floating-Point Values .... 281
7.7.1.2 Converting Floating-Point Values to Normalized Integer Pixel Data Types ....... 282
7.7.2 Conversion Rules for Half-Precision Floating-Point Pixel Data Type ................. 283
7.7.3 Conversion Rules for Single-Precision Floating-Point Pixel Data Type ............... 284
7.7.4 Conversion Rules for 10- and 11-bit Floating-Point Pixel Data Type .................. 284
7.7.5 Conversion Rules for 9-bit Floating-Point Pixel Data Type with a 5-bit Exponent ... 284
7.7.6 Conversion Rules for Signed and Unsigned Integer Pixel Data Types ............... 285
7.7.7 Conversion Rules for sRGBA and sBGRA Textures ............................................. 285

Tables and Figures

Table 1.1. Rounding mode ........................................................................................... 15
Figure 1. Normalized device coordinate system ......................................................... 20
Figure 2. Viewport coordinate system ........................................................................ 20
Figure 3. Normalized 2D texture coordinate system ................................................... 21
Table 2.1. Metal scalar data types ............................................................................... 22
Table 2.2. Size and alignment of scalar data types ....................................................... 23
Table 2.3. Size and alignment of vector data types ....................................................... 25
Table 2.4. Size and alignment of packed vector data types ......................................... 31
Table 2.5. Size and alignment of matrix data types ...................................................... 33
Table 2.6. Metal pixel data types .................................................................................. 36
Table 2.7. Sampler state enumeration values ............................................................... 41
Table 2.8. Imageblock slices and compatible target texture formats ......................... 44
Table 2.9. Intersection Tags ........................................................................................ 57
Table 2.10. Mesh Template Parameter ........................................................................ 67
Table 2.11. Mesh Vertex Attributes ............................................................................. 68

2023-06-02 | Copyright © 2023 Apple Inc. | All Rights Reserved.
Page 7 of 287
Table 7.1. Accuracy of single-precision floating-point operations and functions.......................... 273
Table 7.2. Accuracy of single-precision operations and functions with fast math enabled........ 275
Table 7.3. Accuracy of half-precision floating-point operations and functions......................... 278
Table 7.4. Accuracy of brain floating-point operations and functions .................................... 279
Table 7.5. Accuracy of brain floating-point operations and functions with fast math enabled 280
Table 7.6. Conversion to a normalized float value ................................................................. 282
Table 7.7. Conversion from floating-point to a normalized integer value ................................. 283
Table 7.8. Conversion between integer pixel data types ......................................................... 285
1 Introduction

1.1 Purpose of This Document

Metal enables you to develop apps that take advantage of the graphics and compute processing power of the GPU. This document describes the Metal Shading Language (MSL), which you will use to write a shader program, which is graphics and data-parallel compute code that runs on the GPU. Shader programs run on different programmable units of the GPU. MSL is a single, unified language that allows tighter integration between the graphics and compute programs. Since MSL is C++-based, you will find it familiar and easy to use.

MSL works with the Metal framework, which manages the execution and optionally the compilation of the Metal programs. Metal uses clang and LLVM so you get a compiler that delivers optimized performance on the GPU.

1.2 Organization of This Specification

This document is organized into the following chapters:

- This chapter, “Introduction,” is an introduction to this document that covers the similarities and differences between Metal and C++14. It also details the options for the Metal compiler, including preprocessor directives, options for math intrinsics, and options for controlling optimization.
- “Data Types” lists the Metal data types, including types that represent vectors, matrices, buffers, textures, and samplers. It also discusses type alignment and type conversion.
- “Operators” lists the Metal operators.
- “Address Spaces” describes disjoint address spaces for allocating memory objects with access restrictions.
- “Function and Variable Declarations” details how to declare functions and variables, with optional attributes that specify restrictions.
- “Metal Standard Library” defines a collection of built-in Metal functions.
- “Numerical Compliance” describes requirements for representing floating-point numbers, including accuracy in mathematical operations.

iOS and macOS support for features (functions, enumerations, types, attributes, or operators) described in this document is available since Metal 1, unless otherwise indicated.

For the rest of this document, the abbreviation X.Y stands for “Metal version X.Y”; for example, 2.1 indicates Metal 2.1. Please note that though a feature is supported in MSL shading language, it may not be supported on all GPUs. Please refer to the Metal Feature Set Tables at developer.apple.com.

1.3 References

C++14

1.4  Metal and C++14

The Metal programming language is a C++14-based Specification with extensions and restrictions. Refer to the C++14 Specification (also known as the ISO/IEC JTC1/SC22/WG21 N4431 Language Specification) for a detailed description of the language grammar.

This section and its subsections describe the modifications and restrictions to the C++14 language supported in Metal.

For more about Metal preprocessing directives and compiler options, see section 1.5 of this document.

1.4.1  Overloading

Metal supports overloading, as defined by section 13 of the C++14 Specification. Metal extends the function overloading rules to include the address space attribute of an argument. You cannot overload Metal graphics and kernel functions. (For a definition of graphics and kernel functions, see section 5.1 of this document.)

1.4.2  Templates

Metal supports templates, as defined by section 14 of the C++14 Specification.

1.4.3  Preprocessing Directives

Metal supports the preprocessing directives, as defined by section 16 of the C++14 Specification.

1.4.4  Restrictions

The following C++14 features are not available in Metal (section numbers in this list refer to the C++14 Specification):

- lambda expressions (section 5.1.2)
- dynamic_cast operator (section 5.2.7)
- type identification (section 5.2.8)
- new and delete operators (sections 5.3.4 and 5.3.5)
- noexcept operator (section 5.3.7)
- goto statement (section 6.6)
- register, thread_local storage attributes (section 7.1.1)
- virtual function attribute (section 7.1.2)
- derived classes (section 10, section 11)
• exception handling (section 15)

Do not use the C++ standard library in Metal code. Instead, Metal has its own standard library, as discussed in section 5 of this document.

Metal restricts the use of pointers:

• You must declare arguments to Metal graphics and kernel functions that are pointers with the `Metal device, constant, threadgroup, or threadgroup_imageblock address space` attribute. (For more about Metal address space attributes, see section 4 of this document.)

• Function pointers are supported in Metal 2.3. Prior to Metal 2.3, function pointers are not supported.

Metal supports recursive function calls (C++ section 5.2.2, item 9) in compute (kernel) context starting with Metal 2.4.

A Metal function cannot be called `main`.

1.5 Compiler and Preprocessor

You can use the Metal compiler online (with the appropriate APIs to compile Metal sources) or offline. You can load Metal sources that are compiled offline as binaries, using the appropriate Metal APIs.

This section explains the compiler options supported by the Metal compiler and categorizes them as preprocessor options, options for math intrinsics, options that control optimization, miscellaneous compilation options, and linking.

1.5.1 Preprocessor Compiler Options

The following options control the Metal preprocessor that is run on each program source before actual compilation:

-D name

Predefine `name` as a macro, with definition 1.

-D name=definition

The contents of `definition` are tokenized and processed as if they appeared in a `#define` directive. This option allows you to compile Metal code to enable or disable features. You may use this option multiple times, and the preprocessor processes the definitions in the order in which they appear.

-I dir

Add the directory `dir` to the search path of directories for header files. This option is only available for the offline compiler.
1.5.2 Preprocessor Definitions

The Metal compiler sets a number of preprocessor definitions by default, including:

```c
__METAL_VERSION__   // Set to the Metal language revision
__METAL_MACOS__     // Set if compiled with the macOS Metal language
__METAL_IOS__       // Set if compiled with the iOS Metal language
__METAL__          // Set if compiled with the unified Metal language
                   // Set with -std=metal3.0 or above.
```

You can use definitions to conditionally apply shading language features that are only available on later language version (see Compiler Options Controlling the Language Version).

The version number is MajorMinorPatch. For example, for Metal 1.2, patch 0, __METAL_VERSION__ is 120; for Metal 2.1, patch 1, __METAL_VERSION__ is 211.

To conditionally include code that uses features introduced in Metal 2, you can use the preprocessor definition in code, as follows:

```c
#if __METAL_VERSION__ >= 200
// Code that requires features introduced in Metal 2.
#endif
```

1.5.3 Math Intrinsics Compiler Options

These options control compiler behavior regarding floating-point arithmetic, trading off between speed and correctness.

- `ffast-math` (default)
- `fno-fast-math`

For more about math functions, see section 6.5. For more about the relative errors of ordinary and fast math functions, see section 7.4.

These options enable (the default) or disable the optimizations for floating-point arithmetic that may violate the IEEE 754 standard. They also enable or disable the high precision variant of math functions for single precision floating-point scalar and vector types.

The fast math optimizations for floating-point arithmetic include:

- No NaNs: Allow optimizations to assume the arguments and result are not NaN ("Not a Number").
- No INFs: Allow optimizations to assume the arguments and result are not positive or negative infinity.
- No Signed Zeros: Allow optimizations to treat the sign of a zero argument or result as insignificant.
- Allow Reciprocal: Allow optimizations to use the reciprocal of an argument rather than perform a division.
- Allow Reassociation: Allow algebraically equivalent transformations, such as reassociating floating-point operations that may dramatically change the floating-point results.
• Allow Contract: Allow floating-point contraction across statements. For example, allow fusing a multiple followed by an additional into a single fused-multiply-add.

When utilizing fast math in your program, it is important to understand that the compiler can assume certain properties and make optimizations accordingly. For example, the use of fast math asserts that the shader will never generate Inf or NaN. If the program has an expression X/Y, the compiler can assume Y is never zero as this could potentially result in positive/negative infinite or NaN, depending on the value of X. If Y can be zero, the developer would have an undefined program if compiled with fast math.

By default, the compiler allows floating point contractions. For example, a*b+c may be converted to a single fused-multiply-add. These contractions could lead to computation differences if other expressions are not contracted. To disable allowing the compiler to contractions, pass the following option:

-ffp-contract=off

The compiler also supports using a program to disable contractions:

#pragma STDC FP_CONTRACT OFF

1.5.4 Invariance Compiler Options

If you are building with an SDK that supports iOS 14 or macOS 11, you need to pass the following option to support vertex invariance.

-fpreserve-invariance

Preserve invariant for computations marked with [[invariant]] in vertex shaders. If not set, [[invariant]] is ignored.

In previous versions of Metal, [[invariant]] was a best-effort analysis to mark which operations need to be invariant and may fail in certain cases. This is replaced with a conservative invariant model where the compiler marks operations that doesn’t go into an invariant calculation. This will guarantee anything that is invariant calculation remains invariant. This option may reduce performance as it may prevent certain optimizations to preserve invariance.

1.5.5 Optimization Compiler Options

These options control the optimization level of the compiler.

-02
Optimize for performance (default).

-0s
Like -02 with extra optimizations to reduce code size.
1.5.6  Max Total Threadgroup Size Option

All OS: since Metal 3.0

This option specifies the number of threads (value) in a threadgroup for every function in the translation unit.

    -fmax-total-threads-per-threadgroup=<value>

The attribute `[[max_total_threads_per_threadgroup]] function attribute described in section 5.1.3, section 5.1.7, and section 5.1.8 takes precedence over the compile option.

This option is useful for setting the option to enable functions compiled for a dynamic library to be compatible with a PSO.

1.5.7  Texture Write Rounding Mode

Configure the rounding mode for texture writes to floating-point pixel types by setting the –ftexture-write-rounding-mode compiler flag to one of the options in Table 1.1.

<table>
<thead>
<tr>
<th>Rounding mode</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>native (default)</td>
<td>Texture writes use the hardware’s native rounding strategy.</td>
</tr>
<tr>
<td>rte</td>
<td>Texture writes round to the nearest even number.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.3.</td>
<td></td>
</tr>
<tr>
<td>rtz</td>
<td>Texture writes round toward zero.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.3.</td>
<td></td>
</tr>
</tbody>
</table>

The –ftexture-write-rounding-mode flag is available for these SDKs:
- macOS 11 or later
- iOS 14 or later

For more information about which GPU families support rounding modes other than native, see the Metal Feature Set Tables.
1.5.8 Compiler Options to Enable Modules

The compiler supports multiple options to control the use of modules. These options are only available for the offline compiler:

- **fmodules**
  Enable the modules feature.

- **fimplicit-module-maps**
  Enable the implicit search for module map files named `module.modulemap` or a similar name. By default, `-fmodules` enables this option. (The compiler option `-fno-implicit-module-maps` disables this option.)

- **fno-implicit-module-maps**
  Disable the implicit search for module map files named `module.modulemap`. Module map files are only loaded if they are explicitly specified with `-fmodule-map-file` or transitively used by another module map file.

- **fmodules-cache-path=<directory>**
  Specify the path to the modules cache. If not provided, the compiler selects a system-appropriate default.

- **fmodule-map-file=<file>**
  Load the specified module map file, if a header from its directory or one of its subdirectories is loaded.

If you are building with an SDK that supports iOS 16 or macOS 13, `-fmodules` has the following additional options.

- **fmodules=[mode]**
  Supported values for modes are:

  - stdlib: Enable the modules feature but restrict the search for module maps to the Metal standard library. Enabled by default with an SDK that supports iOS 16 or macOS 13.
  - all: Enable the modules feature (equivalent to `-fmodules`).
  - none: Disable the modules feature.

1.5.9 Compiler Options Controlling the Language Version

The following option controls the version of the unified graphics and computing language accepted by the compiler:

- **std=**
  Determine the language revision to use. A value for this option must be provided, which must be one of:

  - **ios-metal1.0**: Support the unified graphics and computing language revision 1.0 programs for iOS 8. [[ deprecated and will be removed in future OS ]]
• **ios-metal1.1**: Support the unified graphics and computing language revision 1.1 programs for iOS 9.

• **ios-metal1.2**: Support the unified graphics and computing language revision 1.2 programs for iOS 10.

• **ios-metal2.0**: Support the unified graphics and computing language revision 2.0 programs for iOS 11.

• **ios-metal2.1**: Support the unified graphics and computing language revision 2.1 programs for iOS 12.

• **ios-metal2.2**: Support the unified graphics and computing language revision 2.2 programs for iOS 13.

• **ios-metal2.3**: Support the unified graphics and computing language revision 2.3 programs for iOS 14.

• **ios-metal2.4**: Support the unified graphics and computing language revision 2.4 programs for iOS 15.

• **macos-metal1.1 or osx-metal1.1**: Support the unified graphics and computing language revision 1.1 programs for macOS 10.11.

• **macos-metal1.2 or osx-metal1.2**: Support the unified graphics and computing language revision 1.2 programs for macOS 10.12.

• **macos-metal2.0 or osx-metal2.0**: Support the unified graphics and computing language revision 2.0 programs for macOS 10.13.

• **macos-metal2.1**: Support the unified graphics and computing language revision 2.1 programs for macOS 10.14.

• **macos-metal2.2**: Support the unified graphics and computing language revision 2.2 programs for macOS 10.15.

• **macos-metal2.3**: Support the unified graphics and computing language revision 2.3 programs for macOS 11.0.

• **macos-metal2.4**: Support the unified graphics and computing language revision 2.4 programs for macOS 12.0.

Note that macos-* is available in macOS10.13 SDK and later.

As of macOS13.0, iOS 16 and tvOS 16, Metal has a unified the shading language between the platforms.

• **metal3.0**: Support the unified graphics and computing language revision 3.0 programs for macOS 13.0, iOS 16, and tvOS 16.

• **metal3.1**: Support the unified graphics and computing language revision 3.1 programs for macOS 14.0, iOS 17, and tvOS 17.
1.5.10  Compiler Options to Request or Suppress Warnings

The following options are available:

- **-Werror**
  Make all warnings into errors.

- **-w**
  Inhibit all warning messages.

1.5.11  Target Conditionals

Metal defines several macros which one can use to determine what platform the shader is running on. The following macros are defined in `<TargetConditionals.h>`:

- **TARGET_OS_MAC**
  Generated code will run under Mac OS X variant

- **TARGET_OS_OSX**
  Generated code will run under OS X devices

- **TARGET_OS_IPHONE**
  Generated code for firmware, devices or simulator

- **TARGET_OS_IOS**
  Generated code will run under iOS

- **TARGET_OS_TV**
  Generated code will run under Apple TV OS

- **TARGET_OS_MACCATALYST**
  Generated code will run under MacOS

- **TARGET_OS_SIMULATOR**
  Generated code will run under a simulator

Note that this header is not part of `<metal_stdlib>`.

1.5.12  Dynamic Library Linker options

The Metal compiler driver can pass options to the linker. Here is a brief description of some of these options. See the Metal linker for more information.

- **-dynamiclib**
  Specify that the output is a dynamic library.

- **-install_name**
  Used with `-dynamiclib` to specify the location of where the dynamic library is expected be installed and found by the loader. Use with `@executable_path` and `@loader_path`.

1.5.13  Options for Compiling to GPU binaries

The following options are available for compiling to a GPU binary if you are building with an SDK that supports iOS 16 or macOS 13.

- **-arch [architecture]**
  Specify the architecture to build for.
-gpu-family [gpu family name]
Specify the architectures associated with the MTLGPUFamily to build for. See MTLGPUFamily in Metal API for the list of available families.

-N [descriptor.mtlp-json]
Specify the pipeline descriptors in Metal script format. The descriptor files must end in .mtlp-json.

### 1.5.14 Options for Generating Metal Library Symbol Files
If you are building with an SDK that supports iOS 15 or macOS 12, the following option is available to generate a Metal library symbol file.

-record-sources
Enable the compiler to store source information into the AIR or Metal library file (.metallib).

-record-sources=flat
Enable the compiler to store source information if generating an AIR file. Enable the compiler to store the source information in a symbol companion file (.metallibsym) if generating a Metal Library file.

See Generating and Loading a Metal Symbol File at developer.apple.com for more information.

### 1.6 Metal Coordinate Systems
Metal defines several standard coordinate systems to represent transformed graphics data at different stages along the rendering pipeline.

A four-dimensional homogenous vector \((x, y, z, w)\) specifies a three-dimensional point in clip-space coordinates. A vertex shader generates positions in clip-space coordinates. Metal divides the \(x\), \(y\), and \(z\) values by \(w\) to convert clip-space coordinates into normalized device coordinates.

Normalized device coordinates use a left-handed coordinate system (see Figure 1) and map to positions in the viewport. These coordinates are independent of viewport size. The lower-left corner of the viewport is at an \((x, y)\) coordinate of \((-1.0, -1.0)\) and the upper corner is at \((1.0, 1.0)\). Positive-\(z\) values point away from the camera (“into the screen”). The visible portion of the \(z\) coordinate is between 0.0 and 1.0. The Metal rendering pipeline clips primitives to this box.
Figure 1. Normalized device coordinate system

The rasterizer stage transforms normalized-device coordinates (NDC) into *viewport coordinates* (see Figure 2). The \((x, y)\) coordinates in this space are measured in pixels, with the origin in the top-left corner of the viewport and positive values going to the right and down. You specify viewports in this coordinate space, and the Metal maps NDC coordinates to the extents of the viewport.

If you are using variable rasterization rate (see Section 6.15), then the viewport coordinate system is a logical coordinate system independent of the render target’s physical layout. A rate map determines the relationship between coordinates in this logical coordinate system (sometimes called screen space) and pixels in the render targets (physical coordinates).

Figure 2. Viewport coordinate system

*Texture coordinates* use a similar coordinate system to viewport coordinates. Texture coordinates can also be specified using *normalized texture coordinates*. For 2D textures, normalized texture coordinates are values from \(0.0\) to \(1.0\) in both \(x\) and \(y\) directions, as seen in Figure 3. A value of \((0.0, 0.0)\) specifies the pixel at the first byte of the image data (the top-left corner of the image). A value of \((1.0, 1.0)\) specifies the pixel at the last byte of the image data (the bottom-right corner of the image).
Figure 3. Normalized 2D texture coordinate system
2 Data Types

This chapter details the Metal data types, including types that represent vectors and matrices. The chapter also discusses atomic data types, buffers, textures, samplers, arrays, user-defined structures, type alignment, and type conversion.

2.1 Scalar Data Types

Metal supports the scalar types listed in Table 2.1. Metal does not support the double, long long, unsigned long long, and long double data types.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bool</td>
<td>A conditional data type that has the value of either true or false. The value true expands to the integer constant 1, and the value false expands to the integer constant 0.</td>
</tr>
<tr>
<td>char</td>
<td>A signed two’s complement 8-bit integer.</td>
</tr>
<tr>
<td>int8_t</td>
<td></td>
</tr>
<tr>
<td>unsigned char uchar uint8_t</td>
<td>An unsigned 8-bit integer.</td>
</tr>
<tr>
<td>short</td>
<td>A signed two’s complement 16-bit integer.</td>
</tr>
<tr>
<td>int16_t</td>
<td></td>
</tr>
<tr>
<td>unsigned short ushort uint16_t</td>
<td>An unsigned 16-bit integer.</td>
</tr>
<tr>
<td>int</td>
<td>A signed two’s complement 32-bit integer.</td>
</tr>
<tr>
<td>int32_t</td>
<td></td>
</tr>
<tr>
<td>unsigned int uint int32_t</td>
<td>An unsigned 32-bit integer.</td>
</tr>
<tr>
<td>long</td>
<td>A signed two’s complement 64-bit integer.</td>
</tr>
<tr>
<td>int64_t</td>
<td>All OS: Since Metal 2.2.</td>
</tr>
<tr>
<td>unsigned long uint64_t</td>
<td>An unsigned 64-bit integer.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.2.</td>
<td></td>
</tr>
</tbody>
</table>
Type | Description
--- | ---
**half** | A 16-bit floating-point. The half data type must conform to the IEEE 754 binary16 storage format.

**bfloat**

Since Metal 3.1. | A 16-bit brain floating point. The bfloat data type is a truncated version of float for machine learning applications, using an 8-bit (7 explicitly stored) rather than 24-bit mantissa.

**float** | A 32-bit floating-point. The float data type must conform to the IEEE 754 single precision storage format.

**size_t** | An unsigned integer type of the result of the sizeof operator. This is a 64-bit unsigned integer.

**ptrdiff_t** | A signed integer type that is the result of subtracting two pointers. This is a 64-bit signed integer.

**void** | The void type comprises an empty set of values; it is an incomplete type that cannot be completed.

Metal supports:

- the f or F suffix to specify a single precision floating-point literal value (such as 0.5f or 0.5F).
- the h or H suffix to specify a half precision floating-point literal value (such as 0.5h or 0.5H).
- the bf or suffix to specify a brain precision floating-point literal value (such as 0.5bf or 0.5BF).
- the u or U suffix for unsigned integer literals.
- the l or L suffix for signed long integer literals.

Table 2.2 lists the size and alignment of most of the scalar data types.

**Table 2.2. Size and alignment of scalar data types**

<table>
<thead>
<tr>
<th>Type</th>
<th>Size (in bytes)</th>
<th>Alignment (in bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bool</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>char int8_t</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>unsigned char uchar uint8_t</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Type</td>
<td>Size (in bytes)</td>
<td>Alignment (in bytes)</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>short</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>int16_t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unsigned short</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ushort</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uint16_t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>int</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>int32_t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unsigned int</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uint</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uint32_t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>long</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>int64_t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unsigned long</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uint64_t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>size_t</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>half</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>bfloat</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>float</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

### 2.2 Vector Data Types

Metal supports a subset of the vector data types implemented by the system vector math library. Metal supported these vector type names, where \( n \) is 2, 3, or 4, representing a 2-, 3-, or 4- component vector type, respectively:

- booln
- charn
- shortn
- intrn
- longn
- ucharn
- ushortn
- uintn
- ulongtn
- halfn
- bfloatn (supported since Metal 3.1)
- floatn
Metal also supports \( \text{vec<T,n>} \) where \( T \) is a valid scalar type and \( n \) is 2, 3, or 4, representing a 2-, 3-, or 4- component vector type.

Table 2.3 lists the size and alignment of the vector data types.

### Table 2.3. Size and alignment of vector data types

<table>
<thead>
<tr>
<th>Type</th>
<th>Size (in bytes)</th>
<th>Alignment (in bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bool2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>bool3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>bool4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>char2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>uchar2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>char3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>uchar3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>char4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>uchar4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>short2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>ushort2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>short3</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>ushort3</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>short4</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>ushort4</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>int2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>uint2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>int3</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>uint3</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>int4</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>uint4</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>long2</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>ulong2</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>long3</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>ulong3</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>long4</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>ulong4</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>half2</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
### 2.2.1 Accessing Vector Components

You can use an array index to access vector components. Array index 0 refers to the first component of the vector, index 1 to the second component, and so on. The following examples show various ways to access array components:

```cpp
float4 pos = float4(1.0f, 2.0f, 3.0f, 4.0f);

float x = pos[0];    // x = 1.0
float z = pos[2];    // z = 3.0

float4 vA = float4(1.0f, 2.0f, 3.0f, 4.0f);
float4 vB;

for (int i=0; i<4; i++)
    vB[i] = vA[i] * 2.0f    // vB = (2.0, 4.0, 6.0, 8.0);
```

Metal supports using a period (.) as a selection operator to access vector components, using letters that may indicate coordinate or color data:

```cpp
<vector_data_type>.xyzw
<vector_data_type>.rgba
```

The following code initializes a vector test and then uses the .xyzw or .rgba selection syntax to access individual components:
int4 test = int4(0, 1, 2, 3);
int a = test.x; // a = 0
int b = test.y; // b = 1
int c = test.z; // c = 2
int d = test.w; // d = 3
int e = test.r; // e = 0
int f = test.g; // f = 1
int g = test.b; // g = 2
int h = test.a; // h = 3

The component selection syntax allows the selection of multiple components:

```
float4 c;
c.xyzw = float4(1.0f, 2.0f, 3.0f, 4.0f);
c.z = 1.0f;
c.xy = float2(3.0f, 4.0f);
c.xyz = float3(3.0f, 4.0f, 5.0f);
```

The component selection syntax also allows the permutation or replication of components:

```
float4 pos = float4(1.0f, 2.0f, 3.0f, 4.0f);
float4 swiz = pos.wzyx; // swiz = (4.0f, 3.0f, 2.0f, 1.0f)
float4 dup = pos.xxyy; // dup = (1.0f, 1.0f, 2.0f, 2.0f)
```

The component group notation can occur on the left-hand side (lvalue) of an expression. To form the lvalue, you may apply swizzling. The resulting lvalue may be either the scalar or vector type, depending on number of components specified. Each component must be a supported scalar or vector type. The resulting lvalue of vector type must not contain duplicate components.

```
float4 pos = float4(1.0f, 2.0f, 3.0f, 4.0f);
// pos = (5.0, 2.0, 3.0, 6.0)
pos.xw = float2(5.0f, 6.0f);
// pos = (8.0, 2.0, 3.0, 7.0)
pos.wx = float2(7.0f, 8.0f);
// pos = (3.0, 5.0, 9.0, 7.0)
```
pos.xyz = float3(3.0f, 5.0f, 9.0f);

When assigning a swizzled value to a variable, the GPU may need to read the existing value, modify it, and write the result back. The assignment to pos.xw in the example above, causes the GPU to load the float4 value, shuffle values 5.0f and 6.0f into it, and the write back the result back into pos. If two threads write to different components of the vector at the same time, the result is undefined.

The following methods of vector component access are not permitted and result in a compile-time error:

- Accessing components beyond those declared for the vector type is an error.
  2-component vector data types can only access .xy or .rg elements. 3-component vector data types can only access .xyz or .rgb elements.
  
```
float2 pos;    // This is a 2-component vector.
pos.x = 1.0f;  // x is legal and so is y.
pos.z = 1.0f;  // z is illegal and so is w. z is the 3rd
              // component.
float3 pos;    // This is a 3-component vector.
pos.z = 1.0f;  // z is legal for a 3-component vector.
pos.w = 1.0f;  // This is illegal. w is the 4th component.
```

- Accessing the same component twice on the left-hand side is ambiguous and is an error:
  
```
// This is illegal because 'x' is used twice.
pos.xx = float2(3.0f, 4.0f);
```

- Accessing a different number of components is an error:
  
```
// This is illegal due to a mismatch between float2 and float4.
pos.xy = float4(1.0f, 2.0f, 3.0f, 4.0f);
```

- Intermixing the .rgba and .xyzw syntax in a single access is an error:
  
```
float4 pos = float4(1.0f, 2.0f, 3.0f, 4.0f);
pos.x = 1.0f;  // OK
pos.g = 2.0f;  // OK

// These are illegal due to mixing rgba and xyzw attributes.
pos.xg = float2(3.0f, 4.0f);
float3 coord = pos.ryz;
```

- A pointer or reference to a vector with swizzles is an error:
  
```
float4 pos = float4(1.0f, 2.0f, 3.0f, 4.0f);
my_func(&pos.xy);  // This is an illegal pointer to a swizzle.
The `sizeof` operator on a vector type returns the size of the vector, which is given as the number of components * size of each component. For example, `sizeof(float4)` returns 16 and `sizeof(half4)` returns 8.

### 2.2.2 Vector Constructors

You can use constructors to create vectors from a set of scalars or vectors. The parameter signature determines how to construct and initialize a vector. For instance, if the vector is initialized with only a single scalar parameter, all components of the constructed vector are set to that scalar value.

If you construct a vector from multiple scalars, one or more vectors, or a mixture of scalars and vectors, the vector's components are constructed in order from the components of the arguments. The arguments are consumed from left to right. Each argument has all its components consumed, in order, before any components from the next argument are consumed.

This is a list of constructors for `float4`:

- `float4(float x);`
- `float4(float x, float y, float z, float w);`
- `float4(float2 a, float2 b);`
- `float4(float2 a, float b, float c);`
- `float4(float a, float b, float2 c);`
- `float4(float a, float2 b, float c);`
- `float4(float3 a, float b);`
- `float4(float a, float3 b);`
- `float4(float4 x);`

This is a list of constructors for `float3`:

- `float3(float x);`
- `float3(float x, float y, float z);`
- `float3(float a, float2 b);`
- `float3(float2 a, float b);`
- `float3(float3 x);`

This is a list of constructors for `float2`:

- `float2(float x);`
- `float2(float x, float y);`
float2(float2 x);

The following examples illustrate uses of the aforementioned constructors:
float x = 1.0f, y = 2.0f, z = 3.0f, w = 4.0f;
float4 a = float4(0.0f);
float4 b = float4(x, y, z, w);
float2 c = float2(5.0f, 6.0f);

float2 a = float2(x, y);
float2 b = float2(z, w);
float4 x = float4(a.xy, b.xy);

Under-initializing a vector constructor results in a compile-time error.

2.2.3 Packed Vector Types
You must align the vector data types described in section 2.2 to the size of the vector. You can also require their vector data to be tightly packed; for example, a vertex structure that may contain position, normal, tangent vectors and texture coordinates tightly packed and passed as a buffer to a vertex function.

The supported packed vector type names are:
- packed_charn
- packed_shortn
- packed_intn
- packed_ucharn
- packed_ushortn
- packed_uintn
- packed_halfn
- packed_bfloatn (supported since Metal 3.1)
- packed_floatn
- packed_longn (supported since Metal 2.3)

Where n is 2, 3, or 4 representing a 2-, 3-, or 4- component vector type, respectively. (The packed_booln vector type names are reserved.)

Metal also supports packed_vec<T,n> where T is a valid scalar type and n is 2, 3, or 4, representing a 2-, 3-, or 4- component packed vector type.

Table 2.4 lists the size and alignment of the packed vector data types.
Table 2.4. Size and alignment of packed vector data types

<table>
<thead>
<tr>
<th>Type</th>
<th>Size (in bytes)</th>
<th>Alignment (in bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>packed_char2,</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>packed_uchar2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_char3,</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>packed_uchar3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_char4,</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>packed_uchar4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_short2,</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>packed_ushort2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_short3,</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>packed_ushort3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_short4,</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>packed_ushort4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_int2,</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>packed_uint2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_int3,</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>packed_uint3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_int4,</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>packed_uint4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packed_half2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>packed_half3</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>packed_half4</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>packed_bfloat2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>packed_bfloat3</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>packed_bfloat4</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>packed_float2</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>packed_float3</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>packed_float4</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>packed_long2</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>packed_long3</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>packed_long4</td>
<td>32</td>
<td>8</td>
</tr>
</tbody>
</table>
Packed vector data types are typically used as a data storage format. Metal supports the assignment, arithmetic, logical, relational, and copy constructor operators for packed vector data types. Metal also supports loads and stores from a packed vector data type to an aligned vector data type and vice-versa.

Examples:

```c
device float4 *buffer;
device packed_float4 *packed_buffer;
int i;
packed_float4 f ( buffer[i] );
pack_buffer[i] = buffer[i];

// An operator used to convert from packed_float4 to float4.
buffer[i] = float4( packed_buffer[i] );
```

You can use an array index to access components of a packed vector data type. Since Metal 2.1, you can use .xyzw or .rgba selection syntax to access components of a packed vector data type. The semantics and restrictions when swizzling for packed vector data type are the same as for vector types.

Example:

```c
packed_float4 f;
f[0] = 1.0f;  // OK
f.x = 1.0f;  // OK since Metal 2.1
```

### 2.3 Matrix Data Types

Metal supports a subset of the matrix data types implemented by the system math library. The supported matrix type names are:

- halfnxm
- floatnxm

Where \(n\) and \(m\) are numbers of columns and rows. \(n\) and \(m\) must be 2, 3, or 4. A matrix of type floatnxm is composed of \(n\) floatm vectors. Similarly, a matrix of type halfnxm is composed of \(n\) halfm vectors.

Metal also supports matrix\(<T, c, r>\) where \(T\) is a valid floating point type, \(c\) is 2, 3, or 4, and \(r\) is 2, 3, or 4.

Table 2.5 lists the size and alignment of the matrix data types.
### Table 2.5. Size and alignment of matrix data types

<table>
<thead>
<tr>
<th>Type</th>
<th>Size (in bytes)</th>
<th>Alignment (in bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>half2x2</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>half2x3</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>half2x4</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>half3x2</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>half3x3</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>half3x4</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>half4x2</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>half4x3</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>half4x4</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>float2x2</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>float2x3</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>float2x4</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>float3x2</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>float3x3</td>
<td>48</td>
<td>16</td>
</tr>
<tr>
<td>float3x4</td>
<td>48</td>
<td>16</td>
</tr>
<tr>
<td>float4x2</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>float4x3</td>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td>float4x4</td>
<td>64</td>
<td>16</td>
</tr>
</tbody>
</table>

#### 2.3.1 Accessing Matrix Components

You can use the array subscripting syntax to access the components of a matrix. Applying a single subscript to a matrix treats the matrix as an array of column vectors. Two subscripts select a column and then a row. The top column is column 0. A second subscript then operates on the resulting vector, as defined earlier for vectors.

```cpp
float4x4 m;

// This sets the 2nd column to all 2.0.
m[1] = float4(2.0f);
```
// This sets the 1st element of the 1st column to 1.0.
m[0][0] = 1.0f;
// This sets the 4th element of the 3rd column to 3.0.
m[2][3] = 3.0f;

You can access floatnxm and halfnxm matrices as an array of n floatm or n halfm entries.

Accessing a component outside the bounds of a matrix with a non-constant expression results in undefined behavior. Accessing a matrix component that is outside the bounds of the matrix with a constant expression generates a compile-time error.

### 2.3.2 Matrix Constructors

You can use constructors to create matrices from a set of scalars, vectors, or matrices. The parameter signature determines how to construct and initialize a matrix. For example, if you initialize a matrix with only a single scalar parameter, the result is a matrix that contains that scalar for all components of the matrix’s diagonal, with the remaining components initialized to 0.0. For example, a call to:

```cpp
float4x4(fval);
```

Where `fval` is a scalar floating-point value constructs a matrix with these initial contents:

```
fval 0.0 0.0 0.0
0.0 fval 0.0 0.0
0.0 0.0 fval 0.0
0.0 0.0 0.0 fval
```

You can also construct a matrix from another matrix that has the same number of rows and columns. For example:

```cpp
float3x4(float3x4);
float3x4(half3x4);
```

Matrix components are constructed and consumed in column-major order. The matrix constructor must have just enough values specified in its arguments to initialize every component in the constructed matrix object. Providing more arguments than necessary results in an error. Under-initializing a matrix constructor results in a compile-time error.

A matrix of type `T` with `n` columns and `m` rows can also be constructed from `n` vectors of type `T` with `m` components. The following examples are legal constructors:

```cpp
float2x2(float2, float2);
float3x3(float3, float3, float3);
```
float3x2(float2, float2, float2);

Since Metal 2, a matrix of type T with n columns and m rows can also be constructed from n * m scalars of type T. The following examples are legal constructors:
float2x2(float, float, float, float);
float3x2(float, float, float, float, float, float);

The following are examples of matrix constructors that are not supported. A matrix cannot be constructed from combinations of vectors and scalars.
// Not supported.
float2x3(float2 a, float b, float2 c, float d);

### 2.4 SIMD-group Matrix Data Types

All OS: SIMD-group Matrix Types are supported since Metal 2.3.

Metal supports a matrix type `simdgroup_matrix<T, Cols, Rows>` defined in `<metal_simdgroup_matrix>`. Operations on SIMD-group matrices are executed cooperatively by threads in the SIMD-group. Therefore, all operations must be executed only under uniform control-flow within the SIMD-group or the behavior is undefined.

Metal supports the following SIMD-group matrix type names, where T is half, bfloat (since Metal 3.1) or float and Cols and Rows are 8:
- `simdgroup_half8x8`
- `simdgroup_bfloat8x8` (supported since Metal 3.1)
- `simdgroup_float8x8`

The mapping of matrix elements to threads in the SIMD-group is unspecified. For a description of what functions are supported on SIMD-group matrices, see section 6.7

### 2.5 Alignment of Data Types

You can use the `alignas` alignment specifier to specify the alignment requirement of a type or an object. You may also apply the `alignas` specifier to the declaration of a variable or a data member of a structure or class. You may also apply it to the declaration of a structure, class, or enumeration type.

The Metal compiler is responsible for aligning data items to the appropriate alignment as required by the data type. For arguments to a graphics or kernel function declared to be a pointer to a data type, the Metal compiler assumes that the object referenced by the pointer is always appropriately aligned as required by the data type.
2.6 Atomic Data Types

Objects of atomic types are free from data races. If one thread writes to an atomic object while another thread reads from it, the behavior is well-defined.

Metal supports `atomic<T>`, where `T` can be `int`, `uint`, `bool`, or `ulong` for all OSes that support Metal 2.4 and later, or `T` can be `float` for all OSes that support Metal 3 and later.

Metal provides these type aliases for atomic types:

- `atomic_int` A type of alias of `atomic<int>` for OSes that support Metal 1 and later.
- `atomic_uint` A type of alias of `atomic<GLuint>` for OSes that support Metal 1 and later.
- `atomic_bool` A type of alias of `atomic<bool>` for OSes that support Metal 2.4 and later.
- `atomic_ulong` A type of alias of `atomic<GLuint>` for OSes that support Metal 2.4 and later.
- `atomic_float` A type of alias of `atomic<float>` for OSes that support Metal 3 and later.

Metal atomic functions (as described in section 6.15) can only use Metal atomic data types. These atomic functions are a subset of the C++14 atomic and synchronization functions.

2.7 Pixel Data Types

iOS: Pixel data types supported since Metal 2.
macOS: Pixel data types supported since Metal 2.3.

The Metal pixel data type is a templated type that describes the pixel format type and its corresponding ALU type. The ALU type represents the type returned by a load operation and the input type specified for a store operation. Pixel data types are generally available in all address spaces. (For more about address spaces, see section 4.)

Table 2.6 lists supported pixel data types in MSL, as well as their size and alignment.

<table>
<thead>
<tr>
<th>Pixel Data Type</th>
<th>Supported values of T</th>
<th>Size (in bytes)</th>
<th>Alignment (in bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r8unorm&lt;T&gt;</td>
<td>half or float</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>r8snorm&lt;T&gt;</td>
<td>half or float</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>r16unorm&lt;T&gt;</td>
<td>float</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>r16snorm&lt;T&gt;</td>
<td>float</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>rg8unorm&lt;T&gt;</td>
<td>half2 or float2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Pixel Data Type</td>
<td>Supported values of T</td>
<td>Size (in bytes)</td>
<td>Alignment (in bytes)</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------</td>
<td>-----------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>rg8snorm&lt;T&gt;</td>
<td>half2 or float2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>rg16unorm&lt;T&gt;</td>
<td>float2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>rg16snorm&lt;T&gt;</td>
<td>float2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>rgba8unorm&lt;T&gt;</td>
<td>half4 or float4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>srgba8unorm&lt;T&gt;</td>
<td>half4 or float4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>rgba8snorm&lt;T&gt;</td>
<td>half4 or float4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>rgba16unorm&lt;T&gt;</td>
<td>float4</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>rgba16snorm&lt;T&gt;</td>
<td>float4</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>rgb10a2&lt;T&gt;</td>
<td>half4 or float4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>rg11b10f&lt;T&gt;</td>
<td>half3 or float3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>rgb9e5&lt;T&gt;</td>
<td>half3 or float3</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Only assignments and equality/inequality comparisons between the pixel data types and their corresponding ALU types are allowed. (The following examples show the `buffer(n)` attribute, which is explained in section 5.2.1.)

Example:

```c
kernel void
my_kernel(device rgba8unorm<half4> *p [[buffer(0)]],
           uint gid [[thread_position_in_grid]], …)
{
    rgba8unorm<half4> x = p[index]; half4 val = p[gid];
    …
    p[gid] = val;
    p[index] = x;
}
```

Example:

```c
struct Foo {
    rgba8unorm<half4> a;
};

kernel void
my_kernel(device Foo *p [[buffer(0)]],
           uint gid [[thread_position_in_grid]], …)
{
    half4 a = p[gid].a;
}
```
2.8 Buffers

MSL implements a buffer as a pointer to a built-in or user defined data type described in the device, constant, or threadgroup address space. (For more about these address space attributes, see sections 4.1, 4.2, and 4.4, respectively.)

Ordinary Metal buffers may contain:

- Basic types such as float and int
- Vector and matrix types
- Arrays of buffer types
- Structures of buffer types
- Unions of buffer types

Note: As of Metal 2.3, Metal supports buffers that contain long or ulong data types.

The example below shows buffers as arguments to a function that performs the Phong interpolation model. The first two arguments are buffers in the device address space. The third argument is a buffer in the constant address space.

```cpp
vertex ColorInOut
phong_vertex(const device packed_float3* vertices [[buffer(0)]],
             const device packed_float3* normals [[buffer(1)]],
             constant AAPL::uniforms_t& uniforms [[buffer(2)]],
             unsigned int vid [[vertex_id]])
{
  ...
}
```

For more about the buffer(n) attribute used in the example, see section 5.2.1.

For details about argument buffers, see section 2.13.

2.9 Textures

The texture data type is a handle to one-, two-, or three-dimensional texture data that corresponds to all or a portion of a single mipmap level of a texture. The following templates define specific texture data types:

```cpp
enum class access { sample, read, write, read_write };
texture1d<T, access a = access::sample>
```
texture1d_array<T, access a = access::sample>
texture2d<T, access a = access::sample>
texture2d_array<T, access a = access::sample>
texture3d<T, access a = access::sample>
texturecube<T, access a = access::sample>
texturecube_array<T, access a = access::sample>
texture2d_ms<T, access a = access::read>
texture2d_ms_array<T, access a = access::read>

To use sample_compare with a depth format, you need to declare one of the following texture types:

depth2d<T, access a = access::sample>
depth2d_array<T, access a = access::sample>
depthcube<T, access a = access::sample>
depthcube_array<T, access a = access::sample>
depth2d_ms<T, access a = access::read>
depth2d_ms_array<T, access a = access::read>

macOS supports texture2d_ms_array and depth2d_ms_array since Metal 2. All other types supported since Metal 1.

iOS supports all types except texture2d_ms_array and depth2d_ms_array since Metal 1.

T specifies the color type of one of the components returned when reading from a texture or the color type of one of the components specified when writing to the texture. For texture types (except depth texture types), T can be half, float, short, ushort, int, or uint. For depth texture types, T must be float.

If T is int or short, the data associated with the texture must use a signed integer format. If T is uint or ushort, the data associated with the texture must use an unsigned integer format. If T is half, the data associated with the texture must either be a normalized (signed or unsigned integer) or half-precision format. If T is float, the data associated with the texture must either be a normalized (signed or unsigned integer), half or single-precision format.

These access attributes describe support for accessing a texture:

- sample — A graphics or kernel function can sample the texture object. sample implies the ability to read from a texture with and without a sampler.
- read — Without a sampler, a graphics or kernel function can only read the texture object.
- write — A graphics or kernel function can write to the texture object.
- read_write — A graphics or kernel function can read and write to the texture object.

All OS: read_write access supported since Metal 1.2. Other access qualifiers since Metal 1.
Multisampled textures only support the `read` attribute. Depth textures only support the `sample` and `read` attributes.

The following example uses access qualifiers with texture object arguments:

```cpp
void foo (texture2d<float> imgA [[texture(0)]],
          texture2d<float, access::read> imgB [[texture(1)]],
          texture2d<float, access::write> imgC [[texture(2)]])
{
    ...
}
```

(For a description of the `texture` attribute, see section 5.2.1.)

You can use a texture type as the variable type for any variables declared inside a function. The `access` attribute for variables of texture type declared inside a function must be `access::read` or `access::sample`. Declaring variables inside a function to be a texture type without using `access::read` or `access::sample` qualifiers causes a compilation error.

Examples:

```cpp
void foo (texture2d<float> imgA [[texture(0)]],
          texture2d<float, access::read> imgB [[texture(1)]],
          texture2d<float, access::write> imgC [[texture(2)]])
{
    texture2d<float> x = imgA; // OK
    texture2d<float, access::read> y = imgB; // OK
    texture2d<float, access::write> z; // This is illegal.
    ...
}
```

### 2.9.1 Texture Buffers

All OS: Texture buffers supported since Metal 2.1.

A texture buffer is a texture type that can access a large 1D array of pixel data and perform dynamic type conversion between pixel formats on that data with optimized performance. Texture buffers handle type conversion more efficiently than other techniques, allowing access to a larger element count, and handling out-of-bounds read access. Similar type conversion can be achieved without texture buffers by either:

- Reading the pixel data (just like any other array) from a texture object and performing the pixel transformation to the desired format.

- Wrapping a texture object around the data of a buffer object, and then accessing the shared buffer data via the texture. This wrapping technique provides the pixel conversion, but requires an extra processing step, and the size of the texture is limited.

The following template defines the opaque type `texture_buffer`, which you can use like any texture:

```cpp
texture_buffer<T, access a = access::read>
```

`access` can be one of `read`, `write`, or `read_write`. 
T specifies the type of a component returned when reading from a texture buffer or the type of component specified when writing to a texture buffer. For a texture buffer, T can be one of half, float, short, ushort, int, or uint.

For a format without an alpha channel (such as R, RG, or RGB), an out-of-bounds read returns (0, 0, 0, 1). For a format with alpha (such as RGBA), an out-of-bounds read returns (0, 0, 0, 0). For some devices, an out-of-bounds read might have a performance penalty.

An out-of-bounds write is ignored.

A texture buffer can support more texture data than a generic 1D texture, which has is a maximum width of 16384. However, you cannot sample a texture buffer.

A texture buffer also converts data, delivering it in the requested texture format, regardless of the source’s format. When creating a texture buffer, you can specify the format of the data in the buffer (for example, RGBA8Unorm), and later the shader function can read it as a converted type (such as float4). As a result, a single pipeline state object can access data stored in different pixel formats without recompilation.

A texture buffer, like a texture type, can be declared as the type of a local variable to a shader function. For information about arrays of texture buffers, see section 2.12.1. For more about texture buffer, see section 6.12.16.

### 2.10 Samplers

The sampler type identifies how to sample a texture. The Metal API allows you to create a sampler object and pass it in an argument to a graphics or kernel function. You can describe a sampler object in the program source instead of in the API. For these cases, you can only specify a subset of the sampler state: the addressing mode, filter mode, normalized coordinates, and comparison function.

Table 2.7 lists the supported sampler state enumerations and their associated values (and defaults). You can specify these states when a sampler is initialized in Metal program source.

<table>
<thead>
<tr>
<th>Enumeration</th>
<th>Valid Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>coord</td>
<td>normalized (default) pixel</td>
<td>When sampling from a texture, specifies whether the texture coordinates are normalized values.</td>
</tr>
<tr>
<td>address</td>
<td>repeat mirrored_repeat clamp_to_edge (default) clamp_to_zero clamp_to_border</td>
<td>Sets the addressing mode for all texture coordinates.</td>
</tr>
<tr>
<td>Enumeration</td>
<td>Valid Values</td>
<td>Description</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>s_address</td>
<td>repeat, mirrored_repeat, clamp_to_edge (default), clamp_to_zero, clamp_to_border</td>
<td>Sets the addressing mode for individual texture coordinates.</td>
</tr>
<tr>
<td>t_address</td>
<td>repeat, mirrored_repeat, clamp_to_edge (default), clamp_to_zero, clamp_to_border</td>
<td></td>
</tr>
<tr>
<td>r_address</td>
<td>repeat, mirrored_repeat, clamp_to_edge (default), clamp_to_zero, clamp_to_border</td>
<td></td>
</tr>
<tr>
<td>border_color</td>
<td>transparent_black (default), opaque_black, opaque_white</td>
<td>Specifies the border color to use with the clamp_to_border addressing mode.</td>
</tr>
<tr>
<td>filter</td>
<td>nearest (default), linear</td>
<td>Sets the magnification and minification filtering modes for texture sampling.</td>
</tr>
<tr>
<td>mag_filter</td>
<td>nearest (default), linear</td>
<td>Sets the magnification filtering mode for texture sampling.</td>
</tr>
<tr>
<td>min_filter</td>
<td>nearest (default), linear</td>
<td>Sets the minification filtering mode for texture sampling.</td>
</tr>
<tr>
<td>mip_filter</td>
<td>none (default), nearest, linear</td>
<td>Sets the mipmap filtering mode for texture sampling. If none, the texture is sampled as if it has a single mip level. All samples are read from level 0.</td>
</tr>
<tr>
<td>compare_func</td>
<td>never (default), less, less_equal, greater, greater_equal, equal, not_equal, always</td>
<td>Sets the comparison test used by the sample_compare and gather_compare texture functions.</td>
</tr>
</tbody>
</table>

macOS: Supports clamp_to_border address mode and border_color since Metal 1.2.

iOS: Supports clamp_to_border address mode or border_color since Metal 2.3.

With clamp_to_border, sampling outside a texture only uses the border color for the texture coordinate (and does not use any colors at the edge of the texture). If the address mode is clamp_to_border, then border_color is valid.

clamp_to_zero is equivalent to clamp_to_border with a border color of transparent_black (0.0, 0.0, 0.0) with the alpha component value from the texture. If clamp_to_zero is the address mode for one or more texture coordinates, the other texture coordinates can use an address mode of clamp_to_border if the border color is transparent_black. Otherwise, the behavior is undefined.
If `coord` is set to `pixel`, the `min_filter` and `mag_filter` values must be the same, the `mip_filter` value must be `none`, and the address modes must be either `clamp_to_zero`, `clamp_to_border`, or `clamp_to_edge`.

In addition to the enumeration types, you can also specify the maximum anisotropic filtering and an LOD (level-of-detail) range for a sampler:

```metal
max_anisotropy(int value)
lod_clamp(float min, float max)
```

The following Metal program source illustrates several ways to declare samplers. (The `sampler(n)` attribute that appears in the code below is explained in section 5.2.1.) Note that samplers or constant buffers declared in program source do not need these attribute qualifiers. You must use `constexpr` to declare samplers that are initialized in MSL source.

```metal
constexpr sampler s(coord::pixel,  
    address::clamp_to_zero,  
    filter::linear);

constexpr sampler a(coord::normalized);

costexpr sampler b(address::repeat);

costexpr sampler s(address::clamp_to_zero,  
    filter::linear,  
    compare_func::less);

costexpr sampler s(address::clamp_to_zero,  
    filter::linear,  
    compare_func::less,  
    max_anisotropy(10),  
    lod_clamp(0.0f, MAXFLOAT));
```

```metal
kernel void  
my_kernel(device float4 *p [[buffer(0)]],  
    texture2d<float> img [[texture(0)]],  
    sampler smp [[sampler(3)]],  
    ...  
)  
{  
    ...  
}
```

### 2.11 Imageblocks

iOS: Supports imageblocks since Metal 2.

macOS: Supports imageblocks since Metal 2.3.
An imageblock is a 2D data structure (represented by width, height, and number of samples) allocated in threadgroup memory that is an efficient mechanism for processing 2D image data. Each element of the structure can be a scalar or vector integer or floating-point data type, pixel data types (specified in Table 2.6 in section 2.7), an array of these types, or structures built using these types. The data layout of the imageblock is opaque. You can use an (x, y) coordinate and optionally the sample index to access the elements in the imageblock. The elements in the imageblock associated with a specific (x, y) are the per-thread imageblock data or just the imageblock data.

Section 5.6 details imageblock attributes, including the [[imageblock_data(type)]] attribute. Section 6.13 lists the built-in functions for imageblocks.

Imageblocks are only used with fragment and kernel functions. Sections 5.6.3 and 5.6.4 describe how to access an imageblock in a fragment or kernel function, respectively.

For fragment functions, you can access only the fragment’s imageblock data (identified by the fragment’s pixel position in the tile). Use the tile size to derive the imageblock dimensions.

For kernel functions, all threads in the threadgroup can access the imageblock. You typically derive the imageblock dimensions from the threadgroup size, before you specify the imageblock dimensions.

An imageblock slice refers to a region in the imageblock that describes the values of a given element in the imageblock data structure for all pixel locations or threads in the imageblock. The storage type of the imageblock slice must be compatible with the texture format of the target texture, as listed in Table 2.8.

### Table 2.8. Imageblock slices and compatible target texture formats

<table>
<thead>
<tr>
<th>Pixel Storage Type</th>
<th>Compatible Texture Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>float, half</td>
<td>R32Float, R16Float, R8Unorm, R8Snorm, R16Unorm, R16Snorm</td>
</tr>
<tr>
<td>float2, half2</td>
<td>RG32Float, RG16Float, RG8Unorm, RG8Snorm, RG16Unorm, RG16Snorm</td>
</tr>
<tr>
<td>float4, half4</td>
<td>RGBA32Float, RGBA16Float, RGBA8Unorm, RGBA8Snorm, RGBA16Unorm, RGBA16Snorm, RGBA10A2Unorm, RGBA11B10Float, RGBA9E5Float</td>
</tr>
<tr>
<td>int, short</td>
<td>R32Sint, R16Sint, R8Sint</td>
</tr>
<tr>
<td>int2, short2</td>
<td>RG32Sint, RG16Sint, RG8Sint</td>
</tr>
<tr>
<td>int4, short4</td>
<td>RGBA32Sint, RGBA16Sint, RGBA8Sint</td>
</tr>
<tr>
<td>uint, ushort</td>
<td>R32Uint, R16Uint, R8Uint</td>
</tr>
<tr>
<td>uint2, ushort2</td>
<td>RG32Uint, RG16Uint, RG8Uint</td>
</tr>
<tr>
<td>uint4, ushort4</td>
<td>RGBA32Uint, RGBA16Uint, RGBA8Uint</td>
</tr>
</tbody>
</table>
### 2.12 Aggregate Types

Metal supports several aggregate types: arrays, structures, classes, and unions.

Do not specify a structure member with an address space attribute, unless the member is a pointer type. All members of an aggregate type must belong to the same address space. (For more about address spaces, see section 4.)

#### 2.12.1 Arrays of Textures, Texture Buffers, and Samplers

iOS: Arrays of textures supported since Metal 1.2; arrays of samplers since Metal 2; arrays of texture buffers since Metal 2.1.

macOS: Arrays of textures supported since Metal 2; arrays of samplers since Metal 2; arrays of texture buffers since Metal 2.1.

Declare an array of textures as either:

<table>
<thead>
<tr>
<th>Pixel Storage Type</th>
<th>Compatible Texture Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>r8unorm&lt;T&gt;</td>
<td>A8Unorm, R8Unorm</td>
</tr>
<tr>
<td>r8snorm&lt;T&gt;</td>
<td>R8Snorm</td>
</tr>
<tr>
<td>r16unorm&lt;T&gt;</td>
<td>R16Unorm</td>
</tr>
<tr>
<td>r16snorm&lt;T&gt;</td>
<td>R16Snorm</td>
</tr>
<tr>
<td>rg8unorm&lt;T&gt;</td>
<td>RG8Unorm</td>
</tr>
<tr>
<td>rg8snorm&lt;T&gt;</td>
<td>RG8Snorm</td>
</tr>
<tr>
<td>rg16unorm&lt;T&gt;</td>
<td>RG16Unorm</td>
</tr>
<tr>
<td>rg16snorm&lt;T&gt;</td>
<td>RG16Snorm</td>
</tr>
<tr>
<td>rgba8unorm&lt;T&gt;</td>
<td>RGBA8Unorm, BGRA8Unorm</td>
</tr>
<tr>
<td>srgba8unorm&lt;T&gt;</td>
<td>RGBA8Unorm_sRGB, BGRA8Unorm_sRGB</td>
</tr>
<tr>
<td>rgba8snorm&lt;T&gt;</td>
<td>RGBA8Snorm, BGRA8Unorm</td>
</tr>
<tr>
<td>rgba16unorm&lt;T&gt;</td>
<td>RGBA16Unorm</td>
</tr>
<tr>
<td>rgba16snorm&lt;T&gt;</td>
<td>RGBA16Snorm</td>
</tr>
<tr>
<td>rgb10a2&lt;T&gt;</td>
<td>RGB10A2Unorm</td>
</tr>
<tr>
<td>rg11b10f&lt;T&gt;</td>
<td>RG11B10Float</td>
</tr>
<tr>
<td>rgb9e5&lt;T&gt;</td>
<td>RGB9E5Float</td>
</tr>
</tbody>
</table>
array<typename T, size_t N>
const array<typename T, size_t N>

typename shall be a texture type declared with the access::read or access::sample attribute. On macOS since Metal 2 and on iOS with at least A12 hardware since Metal 2.2, support for an array of writeable textures (access::write) is also available. (For more about texture types, see section 2.9.)

Construct an array of texture buffers (see section 2.9.1) with the access::read qualifier using:
array<texture_buffer<T>, size_t N>

Declare an array of samplers as either:
array<sampler, size_t N>
const array<sampler, size_t N>

You can pass an array of textures or an array of samplers as an argument to a function (graphics, kernel, or user function) or declare an array of textures or samples as a local variable inside a function. You can also declare an array of samplers in program scope. Unless used in an argument buffer (see section 2.13), you cannot declare an array<T, N> type (an array of textures, texture buffers, or samplers) in a structure.

MSL also adds support for array_ref<T>. An array_ref<T> represents an immutable array of size() elements of type T. T must be a sampler type or a supported texture type, including texture buffers. The storage for the array is not owned by the array_ref<T> object. Implicit conversions are provided from types with contiguous iterators like metal::array. A common use for array_ref<T> is to pass an array of textures as an argument to functions so they can accept a variety of array types.

The array_ref<T> type cannot be passed as an argument to graphics and kernel functions. However, the array_ref<T> type can be passed as an argument to user functions. The array_ref<T> type cannot be declared as local variables inside functions.

The member functions listed in sections 2.12.1.1 to 2.12.1.3 are available for the array of textures, array of samplers, and the array_ref<T> types.

2.12.1.1 Array Element Access with its Operator

Elements of an array of textures, texture buffers, or samplers can be accessed using the [] operator:
reference operator[] (size_t pos);

Elements of an array of textures, texture buffers, or samplers, or a templated type array_ref<T> can be accessed using the following variant of the [] operator:
constexpr const_reference operator[] (size_t pos) const;
2.12.1.2 Array Capacity

`size()` returns the number of elements in an array of textures, texture buffers, or samplers.

```cpp
constexpr size_t size();
constexpr size_t size() const;
```

**Example:**

```cpp
kernel void
my_kernel(const array<texture2d<float>, 10> src [[texture(0)]],
            texture2d<float, access::write> dst [[texture(10)]],
            ...)
{
    for (int i = 0; i < src.size(); i++)
    {
        if (is_null_texture(src[i]))
            break;
        process_image(src[i], dst);
    }
}
```

2.12.1.3 Constructors for Templated Arrays

```cpp
constexpr array_ref();
constexpr array_ref(const array_ref &);
array_ref & operator=(const array_ref &);
constexpr array_ref(const T * array, size_t length);

template<size_t N>
constexpr array_ref(const T(&a)[N]);

template<typename T>
constexpr array_ref<T> make_array_ref(const T * array, size_t length)

template<typename T, size_t N>
constexpr array_ref<T> make_array_ref(const T(&a)[N])
```

**Examples of constructing arrays:**

```cpp
float4 foo(array_ref<texture2d<float>> src)
{
    float4 clr(0.0f);
    for (int i = 0; i < src.size; i++)
    {
        clr += process_texture(src[i]);
    }
    return clr;
}
```

```cpp
kernel void
my_kernel_A(const array<texture2d<float>, 10> src [[texture(0)]],
             texture2d<float, access::write> dst [[texture(10)]],
             ...)
```
texture2d<float, access::write> dst [[texture(10)]],
...
{
  float4 clr = foo(src);
  ...
}

kernel void
my_kernel_B(const array<texture2d<float>, 20> src [[texture(0)]],
texture2d<float, access::write> dst [[texture(10)]],
...
{
  float4 clr = foo(src);
  ...
}

Below is an example of an array of samplers declared in program scope:
constexpr array<sampler, 2> = {
  sampler(address::clamp_to_zero),
sampler(coord::pixel) 
};

2.12.2 Structures of Buffers, Textures, and Samplers
Arguments to a graphics, kernel, visible, or user function can be a structure or a nested structure with members that are buffers, textures, or samplers only. You must pass such a structure by value. Each member of such a structure passed as the argument type to a graphics or kernel function can have an attribute to specify its location (as described in section 5.2.1).
Example of a structure passed as an argument:
struct Foo {
  texture2d<float> a [[texture(0)]];
  depth2d<float> b [[texture(1)]];
};

[[kernel]] void
my_kernel(Foo f)
{...

You can also nest structures, as shown in the following example:
struct Foo {
  texture2d<float> a [[texture(0)]];
  depth2d<float> b [[texture(1)]];
};

struct Bar {
  Foo f;
}
sampler s [[sampler(0)]];

[[kernel]] void
my_kernel(Bar b)
{...}

Below are examples of invalid use-cases that shall result in a compilation error:

[[kernel]] void
my_kernel(device Foo& f) // This is an illegal use.
{...}

struct MyResources {
    texture2d<float> a [[texture(0)]];
    depth2d<float>   b [[texture(1)]];
    int c;
};

[[kernel]] void
my_kernel(MyResources r) // This is an illegal use.
{...}

### 2.13 Argument Buffers

All OS: Argument buffers supported since Metal 2.

Argument buffers extend the basic buffer types to include pointers (buffers), textures, texture buffers, and samplers. However, argument buffers cannot contain unions. The following example specifies an argument buffer structure called Foo for a function:

```
struct Foo {
    texture2d<float, access::write> a;
    depth2d<float> b;
    sampler c;
    texture2d<float> d;
    device float4* e;
    texture2d<float> f;
    texture_buffer<float> g;
    int h;
};
```

```
kerneld void
my_kernel(constant Foo & f [[buffer(0)]]))
{...}
```
Arrays of textures and samplers can be declared using the existing `array<T, N>` templated type. Arrays of all other legal buffer types can also be declared using C-style array syntax.

Members of argument buffers can be assigned a generic `[[id(n)]]` attribute, where `n` is a 32-bit unsigned integer that can be used to identify the buffer element from the Metal API. Argument buffers can be distinguished from regular buffers if they contain buffers, textures, samplers, or any element with the `[[id]]` attribute.

The same index may not be assigned to more than one member of an argument buffer. Manually assigned indices do not need to be contiguous, but they must be monotonically increasing. In the following example, index 0 is automatically assigned to `foo1`. The `[[id(n)]]` attribute specifies the index offsets for the `t1` and `t2` structure members. Since `foo2` has no specified index, it is automatically assigned the next index, 4, which is determined by adding 1 to the maximum ID used by the previous structure member.

```cpp
struct Foo {
    texture2d<float> t1 [[id(1)]];  // Assigned index 1
    texture2d<float> t2 [[id(3)]];  // Assigned index 3
};
struct Bar {
    Foo foo1;  // foo1 assigned idx 0, t1 and t2 assigned idx 1 and 3
    Foo foo2;  // foo2 assigned idx 4, t1 and t2 assigned idx 5 and 7
};
```

If you omit the `[[id]]` attribute, an ID is automatically assigned according to the following rules:

1. IDs are assigned to structure members in order, by adding 1 to the maximum ID used by the previous structure member. In the example below, the indices are not provided, so indices 0 and 1 are automatically assigned.

```cpp
struct MaterialTexture {
    texture2d<float> tex;  // Assigned index 0
    float4 uvScaleOffset;  // Assigned index 1
};
```

2. IDs are assigned to array elements in order, by adding 1 to the maximum ID used by the previous array element. In the example below, indices 1-3 are automatically assigned to the three array elements of `texs1`. Indices 4-5 are automatically assigned to the fields in `materials[0]`, indices 6-7 to `materials[1]`, and indices 8-9 to `materials[2]`. The `[[id(20)]]` attribute starts by assigning index 20 to constants.

```cpp
struct Material {
    float4 diffuse;  // Assigned index 0
    array<texture2d<float>, 3> texs1;  // Assigned indices 1-3
    MaterialTexture materials[3];  // Assigned indices 4-9
    int constants [[id(20)]] [4];  // Assigned indices 20-23
};
```
3. If a structure member or array element $E$ is itself a structure or array, its structure members or array elements are assigned indices according to rules 1 and 2 recursively, starting from the ID assigned to $E$. In the following example, index 4 is explicitly provided for the nested structure called `normal`, so its elements (previously defined as `tex` and `uvScaleOffset`) are assigned IDs 4 and 5, respectively. The elements of the nested structure called `specular` are assigned IDs 6 and 7 by adding one to the maximum ID (5) used by the previous member.

```c
struct Material {
    MaterialTexture diffuse; // Assigned indices 0, 1
    MaterialTexture normal [[id(4)]]; // Assigned indices 4, 5
    MaterialTexture specular; // Assigned indices 6, 7
}
```

4. Top-level argument buffer arguments are assigned IDs starting from 0, according to the previous three rules.

### 2.13.1 Tier 2 Hardware Support for Argument Buffers

With Tier 2 hardware, argument buffers have the following additional capabilities that are not available with Tier 1 hardware.

You can access argument buffers through pointer indexing. This syntax shown below refers to an array of consecutive, independently encoded argument buffers:

```c
kernel void
kern(constant Resources *resArray [[buffer(0)]])
{
    constant Resources &resources = resArray[3];
}
```

```c
struct TStruct {
    texture2d<float> tex;
};
kernel void
kern(constant TStruct *textures [[buffer(0)]]);
```

To support GPU driven pipelines and indirect draw calls and dispatches, you can copy resources between structures and arrays within a function, as shown below:

```c
kernel void
copy(constant Foo & src [[buffer(0)]],
    device Foo & dst [[buffer(1)]])
{
    dst.a = src.d;
    ...
}
```
Samplers cannot be copied from the thread address space to the device address space. As a result, samplers can only be copied into an argument buffer directly from another argument buffer. The example below shows both legal and illegal copying:

```cpp
struct Resources {
    sampler sam;
};
kernell void
copy(device Resources *src,
    device Resources *dst,
    sampler sam1)
{
    constexpr sampler sam2;
    dst->sam = src->sam; // Legal: device -> device
    dst->sam = sam1;     // Illegal: thread -> device
    dst->sam = sam2;     // Illegal: thread -> device
}
```

Argument buffers can contain pointers to other argument buffers:

```cpp
struct Textures {
    texture2d<float> diffuse;
    texture2d<float> specular;
};
struct Material {
    device Textures *textures;
};
fragment float4
fragFunc(device Material & material);
```

### 2.14 Uniform Type

All OS: uniform types supported since Metal 2.

#### 2.14.1 The Need for a Uniform Type

In the following function example, the variable `i` is used to index into an array of textures given by `texInput`. The variable `i` is non-uniform; that is, it can have a different value for threads executing the graphics or kernel function for a draw or dispatch call, as shown in the example below. Therefore, the texture sampling hardware has to handle a sample request that can refer to different textures for threads executing the graphics or kernel function for a draw or dispatch call.

```cpp
kernel void
my_kernel(array<texture2d<float>, 10> texInput,
    array<texture2d<float>, 10> texOutput,
    sampler s,
    ...,
```
uint2 gid [[thread_position_in_grid]]
{
    int i = ...;
    float4 color = texInput[i].sample(s, float2(gid));
    ...
    texOutput[i].write(color, float2(gid));
}

If the variable i has the same value for all threads (is uniform) executing the graphics or kernel function of a draw or dispatch call and if this information was communicated to the hardware, then the texture sampling hardware can apply appropriate optimizations. A similar argument can be made for texture writes, where a variable computed at runtime is used as an index into an array of textures or to index into one or more buffers.

To indicate that this variable is uniform for all threads executing the graphics or kernel function of a draw or dispatch call, MSL adds a new template class called uniform (available in the header metal_uniform) that can be for declaring variables inside a graphics or kernel function. This template class can only be instantiated with arithmetic types (such as Boolean, integer, and floating point) and vector types.

The code below is a modified version of the previous example, where the variable i is declared as a uniform type:

```mlsl
kernel void
my_kernel(array<texture2d<float>, 10> texInput,
            array<texture2d<float>, 10> texOutput,
            sampler s,
            ...
            uint2 gid [[thread_position_in_grid]])
{
    uniform<int> i = ...;
    float4 color = texInput[i].sample(s, float2(gid));
    ...
    texOutput[i].write(color, float2(gid));
}
```

### 2.14.2 Behavior of the Uniform Type

If a variable is of the uniform type, and the variable does not have the same value for all threads executing the kernel or graphics function, then the behavior is undefined.

Uniform variables implicitly type convert to non-uniform types. Assigning the result of an expression computed using uniform variables to a uniform variable is legal, but assigning a non-uniform variable to a uniform variable results in a compile-time error. In the following example, the multiplication legally converts the uniform variable x into non-uniform product z. However, assigning the non-uniform variable z to the uniform variable b results in a compile-time error.

```mlsl
uniform<int> x = ...;
int y = ...;
```
int z = x*y;       // x is converted to a non-uniform for a multiply
uniform<int> b = z; // illegal; compile-time error

To declare an array of uniform elements:
uniform<float> bar[10]; // elements stored in bar array are uniform

The uniform type is legal for both parameters and the return type of a function. For example:
uniform<int> foo(...); // foo returns a uniform integer value
int bar(uniform<int> a, ...);

It is legal to declare a pointer to a uniform type, but not legal to declare a uniform pointer. For example:
device uniform<int> *ptr; // values pointed to by ptr are uniform
uniform<device int *> ptr; // illegal; compile-time error

The results of expressions that combine uniform with non-uniform variables are non-uniform. If the non-uniform result is assigned to a uniform variable, as in the example below, the behaviors is undefined. (The front-end might generate a compile-time error, but it is not guaranteed to do so.)
uniform<int> i = ...;
int j = ...;
if (i < j) {  // non-uniform result for expression (i < j)
    ...  
    i++; // compile-time error, undefined behavior
}

The following example is similar:
bool p = ... // non-uniform condition.
uniform<int> a = ..., b = ...;
uniform<int> c = p ? a : b; // compile-time error, undefined behavior

### 2.14.3 Uniform Control Flow

When a control flow conditional test is based on a uniform quantity, all program instances follow the same path at that conditional test in a function. Code for control flow based on uniform quantities should be more efficient than code for control flow based on non-uniform quantities.
2.15 Visible Function Table

All OS: type is supported since Metal 2.3.

Defined in the header `<metal_visible_function_table>`, the `visible_function_table` type is used to represent a table of function pointers to visible functions (see section 5.1.4) stored in device memory. In Metal 2.3, it can be used in a compute (kernel) function. As of Metal 2.4, it can be used in fragment, vertex, and tile functions. It is an opaque type, and the content of the table cannot be modified from the GPU. A `visible_function_table` type can be used in an argument buffer or directly passed to qualified function using a buffer binding point.

To declare a `visible_function_table` type with a template parameter `T` where `T` is the signature of the function stored in the table, use the following template function.

```cpp
visible_function_table<typename T> functions;
```

The following example shows how to declare a table that is compatible with a function whose definition is "[[visible]] int func(float f)".

```cpp
visible_function_table<int(float)> functions;
```

To get a visible function pointer from the table, use the [] operator.

```cpp
using fnptr = T (*)(...) [[visible]]
fnptr operator[](uint index) const;
```

`size()` returns the number of function pointer entries in the table.

```cpp
uint size() const
```

`empty()` returns true if the table is empty.

```cpp
bool empty() const
```

The following function can be used to determine if a table is a null `visible_function_table`. A null `visible_function_table` is a table that is not pointing to anything.

```cpp
bool is_null_visible_function_table(visible_function_table<T>);
```

The following example shows how the table can be passed in a buffer.

```cpp
using TFuncSig = void(float, int);
kernell void F(uint tid [[thread_position_in_grid]],
   device float* buf [[buffer(0)]],
   visible_function_table<TFuncSig> table [[buffer(1)]])
{
   uint tsize = table.size();
   table[tid % tsize](buf[tid], tid);
}
```
2.16 Function Group Attribute

All OS: `[[function_group]]` is supported since Metal 2.3.

The optional `[[function_group]]` attribute can be used to indicate the possible groups of functions being called from an indirect call through a function pointer or `visible_function_table`. This is a compiler hint to enable the compiler to optimize the call site. The groups of functions are specified as string literal arguments of the attribute. This attribute can be applied in three different contexts:

- variable declarations with an initializer expression -- It affects all indirect call expressions in the initializer expressions.
- expression statements -- It affects all the indirect call expressions of the given expression.
- return statements -- It affects all the indirect call expressions of the return value expression.

The following examples show how `[[function_group]]` can be used.

```cpp
float h(visible_function_table<float(float)> table, 
        float (*fnptr[3])(float))
{
    // indirect call to table[0] is restricted to "group1"
    [[function_groups("group1")]] float x = table[0](1.0f);

    // indirect call to `fnptr[0]` can call any function
    x += fnptr[0](2.0f);

    // indirect call to `fnptr[1]` is restricted to "group2"+"group3"
    [[function_groups("group2", "group3")]] return x + fnptr[1](3.0f);
}
```

2.17 Ray Tracing Types

All OS: Ray tracing types are supported since Metal 2.3. These types are defined in the header `<metal_raytracing>` in the namespace `metal::raytracing`. In Metal 2.3, these types are only supported in a compute function (kernel functions) except where noted below. As of Metal 2.4, they are also supported in vertex, fragment, and tile functions. In Metal 3.1, raytracing supports curves and multi-level instancing.

2.17.1 Ray Tracing Intersection Tags

All OS: Ray tracing intersection tags are supported since Metal 2.3.
The intersection_tags are defined in the header <metal_raytracing> in the namespace metal::raytracing. They are listed in Table 2.9 and are used in ray tracing when defining

- intersection functions ([[[intersection]]] section 5.1.6)
- intersection function tables (intersection_function_table section 2.17.3)
- intersection results (intersection_result section 2.17.4)
- intersector types and associated functions (intersector section 6.18.2)
- acceleration structure types (acceleration_structure section 2.17.6 and 6.18.1)
- intersection queries (intersection_query section 6.18.4).

The tags are used to configure the ray tracing process and control the behavior and semantics of the different types and tables. The tags identify the type of accelerator structure being intersected, the built-in parameters available for intersection functions, the type of intersection function in an intersection function table, the methods available on intersector type or intersection query object, and the data returned in an intersection result type.

The intersection_tags must match in tag type and order between related uses of intersection_function_table, intersection_result, intersector, and intersection_query, or the compiler will generate an error. The acceleration structure type being intersected has to match the ordering of instancing, primitive_motion, and instance_motion tags if they are present on the other raytracing types used to intersect the acceleration structure. When calling intersection functions in an intersection function table, the developer must ensure they use the same ordered set of tags, or else the result is undefined.

### Table 2.9. Intersection Tags

<table>
<thead>
<tr>
<th>Intersection Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>instancing</td>
<td>This tag indicates intersection functions can read the built-in instance_id and/or user_instance_id as described in section 5.2.3.7, and the acceleration structure is an instance acceleration structure. The intersector&lt;intersection_tags...&gt;::intersect() function and intersection_query&lt;intersection_tags...&gt; assume that the acceleration structure needs to be an instance_acceleration_structure and it returns the instance_id value.</td>
</tr>
<tr>
<td>Intersection Tag</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>triangle_data</td>
<td>This tag indicates triangle intersection functions can read input parameters with barycentric_coord or front_facing attribute as described in section 5.2.3.7. This tag cannot be used in defining an acceleration structure. The intersector&lt;intersection_tags...&gt;::intersect() function and intersection_query&lt; intersection_tags...&gt; returns the triangle_barycentric_coord and triangle_front_facing values.</td>
</tr>
<tr>
<td>world_space_data</td>
<td>This tag indicates intersection functions declared with this tag can query world_space_origin, world_space_direction, object_to_world_transform, and world_to_object_transform as described in section 5.2.3.7. This tag cannot be used in defining an acceleration structure or intersection_query. It enables support for world space data in intersector and intersection_function_table.</td>
</tr>
<tr>
<td>primitive_motion</td>
<td>This tag enables support for primitive level motion in intersector, intersection_query, intersection_function_table, and acceleration structures.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4</td>
<td></td>
</tr>
<tr>
<td>instance_motion</td>
<td>This tag enables support for instance level motion in intersector, intersection_query, intersection_function_table, and acceleration structure.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4</td>
<td></td>
</tr>
<tr>
<td>extended_limits</td>
<td>This tag indicates acceleration structures passed to intersection functions are built with extended limits for the number of primitives, number of geometries, number of instances, and increases the number of bits used for visibility masks. This tag cannot be used in defining an acceleration structure.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4</td>
<td></td>
</tr>
<tr>
<td>curve_data</td>
<td>This tag makes the curve_parameter of the curve intersection point available as a field of intersection_result object from methods of the intersection_query objects, and as input parameter to intersection functions as described in section 5.2.3.7.</td>
</tr>
<tr>
<td>Since Metal 3.1</td>
<td></td>
</tr>
<tr>
<td>Intersection Tag</td>
<td>Description</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>max_levels&lt;Count&gt; Since Metal 3.1</td>
<td>This tag enables support for multi-level instancing in intersector, intersection_query and intersection_function_table. It cannot be used in acceleration structures. Count is a template parameter that determines the maximum number of acceleration structure levels that can be traversed. It must be between [2, 16] for intersection_query. It must be [2,32] for intersector. For intersection_function_table, it needs to match it use with intersection_query or intersector.</td>
</tr>
</tbody>
</table>

In Metal 2.3, the following are valid combinations of intersection tags:

- no tags
- triangle_data
- instancing
- instancing, triangle_data
- instancing, world_space_data
- instancing, triangle_data, world_space_data

Metal 2.4 adds the following are additional valid combinations:

- primitive_motion
- triangle_data, primitive_motion
- instancing, primitive_motion
- instancing, triangle_data, primitive_motion
- instancing, world_space_data, primitive_motion
- instancing, triangle_data, world_space_data, primitive_motion
- instance_motion
- instancing, instance_motion
- instancing, triangle_data, instance_motion
- instancing, world_space_data, instance_motion
- instancing, triangle_data, world_space_data, instance_motion
- instancing, primitive_motion, instance_motion
- instancing, triangle_data, primitive_motion, instance_motion
- instancing, world_space_data, primitive_motion, instance_motion
- instancing, triangle_data, world_space_data, primitive_motion, instance_motion
The extended_limits tag may be added to all combinations listed above. In Metal 3.1, curve_data may be added to all combinations listed above. The intersection tag max_levels<Count> may be added to any combination above containing instancing.

### 2.17.2 Ray Type

The ray struct is a container for the properties of the ray required for an intersection.

```cpp
struct ray {
    ray(float3 origin = 0.0f, float3 direction = 0.0f, 
        float min_distance = 0.0f, float max_distance = INFINITY);
    float3 origin;
    float3 direction;
    float min_distance;
    float max_distance;
};
```

The ray's origin and direction field are in world space. When a ray object is passed into a custom intersection or triangle intersection function, the min_distance and max_distance fields will be based on the current search interval: As candidate intersections are discovered, max_distance will decrease to match the newly narrowed search interval. Within intersection functions, the origin and direction will be in object space.

A ray can be invalid. Examples of invalid rays include:

- INF's or NaN's in origin or direction
- min_distance == NaN or max_distance == NaN
- min_distance == INF (Note that max_distance may be positive INF).
- length(ray.direction) == 0.0
- min_distance > max_distance
- min_distance < 0.0 or max_distance < 0.0

The ray direction does not need to be normalized, although it does need to be non-zero.

### 2.17.3 Intersection Function Table

The `intersection_function_table<intersection_tags...>` struct type describes a table of custom intersection functions passed into the shader as defined from section 5.1.6. The intersection tags are defined from Table 2.9. The intersection tags on `intersection_function_table` type and the intersection functions must match. An example of such a declaration is

```cpp
intersection_function_table<triangle_data, instancing> intersectionFuncs;
```

Call the following function to check if the `intersection_function_table` is null.

2023-06-02 | Copyright © 2023 Apple Inc. | All Rights Reserved.

Page 60 of 287
bool
is_null_intersection_function_table(
    intersection_function_table< intersection_tags...>)

Call the following member function to check if the intersection_function_table is empty.
bool empty() const

Call the following member function to return the number of entries in intersection_function_table.
uint size() const

Metal 3 supports the following function: get_buffer and get_visible_function_table.

Call the following member function to return the buffer at index from the intersection_function_table, where T is a pointer or reference in the device or constant address space.

template<typename T>
    T get_buffer(uint index) const

Call the following member function to return the visible_function_table<T> at index from the intersection_function_table. T is the signature of the function stored in the table.

template <typename T> visible_function_table<T>
    get_visible_function_table(uint index) const;

Metal 3.1 supports the following functions: set_buffer and set_visible_function_table.

Call the following member functions to set the device or constant buffer object at the index position in the intersection_function_table entry.
void set_buffer(const device void *buf, uint index)
void set_buffer(constant void *buf, uint index)

Call the following member function to set the visible_function_table at the index position in the intersection_function_table, where T is the signature of the function stored in the table.

template<typename T>
    void set_visible_function_table(visible_function_table<T> vft, uint index)
2.17.4 Intersection Result Type

The results of an intersection are returned in an
intersection_result<intersection_tags...> structure where
intersection_tags are defined in Table 2.9. The return struct is defined as
class intersection_type {
    none,
    triangle,
    bounding_box,
    curve // Available as of Metal 3.1
};

template <typename... intersection_tags>
struct intersection_result {
    intersection_type type;
    float distance;
    uint primitive_id;
    uint geometry_id;

    const device void *primitive_data; // Available as of Metal 3

    // Available only if intersection_tags include instancing without
    // max_levels<Count>
    uint instance_id;
    uint user_instance_id; // Available as of Metal 2.4

    // As of Metal 3.1, replaces instance_id and user_instance_id with
    // an array if intersection_tags include instancing and
    // max_levels<Count>.
    uint instance_count; // The number of instances
    // intersected by the ray.

    uint instance_id[Count - 1]; // The instance IDs of instances
    // intersected by the ray.

    uint user_instance_id[Count - 1]; // The user instance IDs of
    // instances intersected by
    // the ray.

    // Available only if intersection_tags include triangle_data.
    float2 triangle_barycentric_coord;
    bool triangle_front_facing;

    // As of Metal 2.4, the following is available only if
    // intersection_tags include world_space_data and instancing
    float4x3 world_to_object_transform;
float4x3 object_to_world_transform;

// As of Metal 3.1, the following is available only if
// intersection_tags include curve_data.
float curve_parameter;
};

If a ray is invalid, an intersection::none is returned.
The distance returned is in world space.
For vertex attributes v0, v1, and v2, the attribute value at the specified triangle barycentric point is:
    v1 * triangle_barycentric_coord.x +
    v2 * triangle_barycentric_coord.y +
    v0 * (1.0f - (triangle_barycentric_coord.x +
               triangle_barycentric_coord.y))

2.17.5 Intersector Type

The intersector<intersection_tags...> struct type defines an object that controls
the acceleration structure traversal and defines functions to intersect rays like intersect().
Use the intersection_tags (described in Table 2.9) when creating the intersector to
specialize on which types of acceleration structure it operates on and which functions are
available (see section 6.18.2). Intersection tags on the intersector type must match their
associated intersection function (section 5.1.6), or the behavior is undefined.

// Create a default intersector
intersector<> primitiveIntersector;

// Create an intersector that is specialized to support triangle and
// world space data.
intersector<triangle_data, instancing, world_space_data> instanceInter;

The intersector<intersection_tags...> struct type provides a convenience type for
the intersection result type defined in section 2.17.5:
    intersector<intersection_tags...>::result

2.17.6 Acceleration Structure Type

All OS: Acceleration Structure Type has been supported since Metal 2.3.
All OS: Acceleration Structure Template Type has been supported since Metal 2.4.

Metal 2.3 supports two types of acceleration structure:
• primitive_acceleration_structure
• instance_acceleration_structure.

These are opaque objects that can be bound directly using buffer binding points or via argument buffers:

```plaintext
struct AccelerationStructs {
    primitive_acceleration_structure prim_accel;
    instance_acceleration_structure inst_accel;
    array<primitive_acceleration_structure, 2> prim_accel_array;
    array<instance_acceleration_structure, 2> inst_accel_array;
};

[[kernel]]
void intersectInstancesKernel(
    primitive_acceleration_structure prim_accel [[buffer(0)]],
    instance_acceleration_structure inst_accel [[buffer(1)]],
    device AccelerationStructs *accels [[buffer(3)]]) {...}
```

It is possible to create default initialized variables of such types, and the default value is the null value for the acceleration structures.

In Metal 2.4, the acceleration structure is replaced with a templatized type acceleration_structure<intersection_tags...>. The template parameter intersection_tags can be empty or a combination of instancing, primitive_motion, or instance_motion as defined in Table 2.9. Intersection Tags. For example, the following defines an instance acceleration structure that supports primitive motion.

```plaintext
acceleration_structure<instancing, primitive_motion> accel_struct;
```

The following combinations of tags can be used to declare a primitive acceleration structure

- no tags
- primitive_motion

The following combinations of tags can be used to declare an instance acceleration structure

- instancing
- instancing, primitive_motion
- instancing, instance_motion
- instancing, primitive_motion, instance_motion
To maintain backward compatibility, `primitive_acceleration_structure` is aliased to `acceleration_structure<>` and `instance_acceleration_structure` is aliased to `acceleration_structure<instancing>`.

As before, these are opaque objects that can be bound directly using buffer binding points or via argument buffers:

```c
struct AccelerationMotionStructs {
    acceleration_structure<primitive_motion> prim_motion_accel;
    acceleration_structure<instancing, instance_motion> inst_motion_accel;
    array<acceleration_structure<>[], 2> prim_accel_array;
    array<acceleration_structure<instancing>[], 2> inst_accel_array;
};
[[kernel]]
void intersectMotionKernel(
    acceleration_structure<primitive_motion> prim [[buffer(15)]],
    acceleration_structure<instancing, instance_motion> inst [[buffer(16)]],
    device AccelerationMotionStructs *accels [[buffer(17)]]) {
    ...
}
```

When binding these acceleration structures from the Metal API to the compute or graphic functions, the acceleration structures’ type must match what is defined in the shader. For instance acceleration structures, one can bind instance acceleration structures without support for `primitive_motion` to a shader that expects instance acceleration structures with `primitive_motion`. For example, a Metal buffer with an instance acceleration structure that can be passed to a shader with `acceleration_structure<instancing>` can also be given to a shader with `acceleration_structure<instancing, primitive_motion>`. This capability allows developers to write one shader function that can handle either an acceleration structure with or without `primitive_motion` at the cost of the ray tracing runtime checking for `primitive_motion`. To avoid this cost, developers can write two functions where one uses an acceleration structure with `primitive_motion` and one without.

See section 6.18.1 for the functions to call if the acceleration structure is `null`. 
2.17.7  Intersection Query Type

All OS: Intersection Query is supported since Metal 2.4

The `intersection_query<interation_tags...>` type defines an object that enables users to fully control the raytracing process and when to call custom intersection code. The intersection query object provides a set of functions to advance the query through an acceleration structure and query traversal information. Use the `intersection_tags` (defined in Table 2.9) when creating the `intersection_query<interation_tags...>` type to specialize the type of acceleration structure and what functions are available (see section 6.18.4). It supports the following combinations of intersection tags:

- no tags
- triangle_data
- instancing
- instancing, triangle_data

Metal 3.1 supports the following additional combinations:

- instancing, max_levels<Count>
- instancing, triangle_data, max_levels<Count>

In Metal 3.1, `curve_data` may be added to all combinations listed above.

The `intersection_query<interation_tags...>` type has the following restrictions

- it cannot be used for members of a structure/union
- it cannot be returned from a function
- it cannot be assigned to

These restrictions prevent the intersection query object from being copied.

2.18  Interpolant Type

All OS: Interpolant Type is supported since Metal 2.3.

The interpolant type `interpolant<T,P>` defined in `<metal_interpolate>` is a templatized type that encapsulates a fragment shader input for pull-model interpolation (section 6.11). Type parameters T and P represent the input’s data type and perspective-correctness, respectively. Supported values for T are the scalar and vector floating-point types. Supported values of P are the types `interpolation::perspective` and `interpolation::no_perspective`.

You can declare a variable with the `interpolant<T,P>` type only in the following contexts:

- As a fragment shader input argument with `[[stage_in]]`. Such a declaration must match a corresponding vertex shader output argument of type T with the same name or `[[user(name)]]` attribute. The declaration can’t have a sampling-and-interpolation attribute (section 5.4).
As a local or temporary variable, which needs to be initialized as a copy of the above.

An `interpolant<T,P>` variable is not automatically convertible to a value of type `T`. Instead, retrieve a value by calling one of several interpolation methods (see section 6.11). The interpolation shall be perspective-correct if the value of `P` is `interpolation::perspective`.

### 2.19 Mesh Shader Types

All OS: Mesh shader types have been supported since Metal 3. These types are used in the mesh pipeline to render geometry and are defined in the header `<metal_mesh>`.

#### 2.19.1 Mesh Grid Property Type

All OS: Mesh Grid PropertyType has been supported since Metal 3.

An object function (see section 5.1.7) can use the `mesh_grid_properties` type to specify the size of the mesh grid to dispatch for a given threadgroup from the object stage.

Call the following member function to control the number of threadgroups of the mesh grid that will be dispatched.

```cpp
void set_threadgroups_per_grid(uint3)
```

If the member function `set_threadgroups_per_grid` for a given threadgroup of the object grid is never called, then no mesh grid will be dispatched for the given object grid threadgroup. Calls to `set_threadgroups_per_grid` behave as a write to `threadgroup` memory performed by each thread.

#### 2.19.2 Mesh Type

All OS: Mesh Type has been supported since Metal 3.

A mesh function (see section 5.1.8) can use an argument of type `mesh<V, P, NV, NP, t>` struct type to represent the exported mesh data. Table 2.10 describes the `mesh` template parameters.

<table>
<thead>
<tr>
<th>Template Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>V is the vertex type.</td>
</tr>
<tr>
<td>P</td>
<td>P is the primitive type.</td>
</tr>
<tr>
<td>NV</td>
<td>NV is the maximum number of vertices.</td>
</tr>
<tr>
<td>Template Parameter</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>NP</td>
<td>NP is the maximum number of primitives.</td>
</tr>
</tbody>
</table>
| t                  | t specifies the topology of the mesh. It is one of the following enum values:  
```c
enum topology {
    point,
    line,
    triangle
}
```

A valid vertex type $V$ follows the same rules as the vertex function return type defined in section 5.2.3.3 with the following restrictions. The vertex type can be either

- A float4 represents the vertex position
or a user defined structure:
- Includes a field with the `[[position]]` attribute.
- May include other fields of scalar or vector of integer or floating-point type.
- Supports the following attributes from Table 2.11. Each attribute can be used once within the vertex type.

### Table 2.11. Mesh Vertex Attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
</table>
| clip_distance| float or float[n]  
$\ n$ needs to be known at compile time | Distance from vertex to clipping plane |
<p>| invariant    | Not applicable; needs to be used with <code>[[position]]</code> | Marks the output position such that if the sequence of operations used to compute the output position in multiple vertex shaders is identical, there is a high likelihood that the resulting output position computed by these vertex shaders are the same value. Requires users to pass <code>-fpreserve-invariance</code>. See the description below for more information. |</p>
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>point_size</td>
<td>float</td>
<td>Size of a point primitive</td>
</tr>
<tr>
<td>position</td>
<td>float4</td>
<td>The transformed vertex position</td>
</tr>
<tr>
<td>shared</td>
<td>Not applicable</td>
<td>If present, then for every amplification_id, the output shall have the same value.</td>
</tr>
</tbody>
</table>

A valid primitive type follows the same rules as fragment input section 5.2.3.4. A valid primitive type is either

- **void** indicating no per-primitive type

or a user-defined structure

- Includes fields of scalar or vector of integer or floating-point type

- Supports only the following attributes from Table 2.12. Each attribute can be used once within the primitive type.

**Table 2.12. Mesh Primitive Attributes**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>primitive_culled</td>
<td>bool</td>
<td>If set to true, the primitive is not rendered.</td>
</tr>
<tr>
<td>primitive_id</td>
<td>uint</td>
<td>The per-primitive identifier used with barycentric coordinates.</td>
</tr>
<tr>
<td>render_target_array_index</td>
<td>uchar, ushort, or uint</td>
<td>The render target array index, which refers to the face of a cubemap, data at a specified depth of a 3D texture, an array slice of a texture array, an array slice, or face of a cubemap array. For a cubemap, the render target array index is the face index, which is a value from 0 to 5. For a cubemap array the render target array index is computed as: array slice index * 6 + face index.</td>
</tr>
<tr>
<td>viewport_array_index</td>
<td>uchar, ushort, or uint</td>
<td>The viewport (and scissor rectangle) index value of the primitive.</td>
</tr>
</tbody>
</table>
If the mesh\(<V, P, NV, NP, t>\) does not specify a field with [[primitive_culled]], the behavior is the primitive is rendered. If the fragment shader reads the field, the value read is false because that fragment invocation belongs to a non-culled primitive.

Interpolation and sampling qualifiers are accepted on the vertex and primitive type members. The behavior is specified in section 5.2.3.4.

To minimize the possible user errors in mesh-fragment linking, the names of fields for user-defined vertex and primitive type needs to be unique between the vertex and primitive type.

An example of mesh\(<V, P, NV, NP, t>\) is

```cpp
struct VertexOut {
    float4 position [[position]];
};
struct PrimitiveOut {
    float color [[flat]];
};

using custom_mesh_t = metal::mesh<VertexOut, PrimitiveOut, 64, 64, metal::topology::triangle>;
```

The mesh types contain the following static data member below.

**Table 2.13. Mesh Static Members**

<table>
<thead>
<tr>
<th>Member variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>uint max_vertices</td>
<td>The maximum number of vertices in the mesh ((NV)).</td>
</tr>
<tr>
<td>uint max_primitive</td>
<td>The maximum number of primitives in the mesh ((NP)).</td>
</tr>
<tr>
<td>uint indices_per_primitive</td>
<td>The number of indices per primitive based on topology (t).</td>
</tr>
<tr>
<td>uint max_indices</td>
<td>The maximum number of indices ((max_primitive * indices_per_primitive)).</td>
</tr>
</tbody>
</table>

Call the following member function to set the vertex at index \(I\) in the range \([0, max\_vertices)\).

```cpp
void set_vertex(uint I, V v)
```

If \(P\) is not void, call the following member function to set the primitive at index \(I\) in the range \([0, max\_primitive)\).
void set_primitive(uint I, P p)

Call the following member to set the primitive count where \( c \) is in the range \([0, \text{max}_{\text{primitive}}]\).

void set_primitive_count(uint c)

Call the following member to set the index where \( I \) is in the range \([0, \text{max}_{\text{indices}}]\).

void set_index(uint I, uchar v)

It is legal to call the following \texttt{set_indices} functions to set the indices if the position in the index buffer is valid and if the position in the index buffer is a multiple of 2 (uchar2 overload) or 2 (uchar4 overload). The index \( I \) needs to be in the range \([0, \text{max}_{\text{indices}}]\).

void set_indices(uint I, uchar2 v)
void set_indices(uint I, uchar4 v)

\section*{2.20 Type Conversions and Re-interpreting Data}

The \texttt{static\_cast} operator converts from a scalar or vector type to another scalar or vector type using the default rounding mode with no saturation (when converting to floating-point, round ties to even; when converting to an integer, round toward zero). If the source type is a scalar or vector Boolean, the value \texttt{false} is converted to zero, and the value \texttt{true} is converted to one.

Metal adds an \texttt{as\_type\langle\texttt{type\_id}\rangle} operator to allow any scalar or vector data type (that is not a pointer) to be reinterpreted as another scalar or vector data type of the same size. The bits in the operand are returned directly without modification as the new type. The usual type promotion for function arguments is not performed.

For example, \texttt{as\_type\langle\texttt{float}\rangle(0x3f800000)} returns 1.0f, which is the value of the bit pattern 0x3f800000 if viewed as an IEEE-754 single precision value.

Using the \texttt{as\_type\langle\texttt{type\_id}\rangle} operator to reinterpret data to a type with a different number of bytes results in an error.

Examples of legal and illegal type conversions:

float f = 1.0f;
// Legal. Contains: 0x3f800000
uint u = as\_type\langle\texttt{uint}\rangle(f);

// Legal. Contains:
// (\texttt{int4})(0x3f800000, 0x40000000, 0x40400000, 0x40800000)
float4 f = float4(1.0f, 2.0f, 3.0f, 4.0f);
int4 i = as\_type\langle\texttt{int4}\rangle(f);

int i;
// Legal.
short2 j = as_type<short2>(i);

half4 f;
// Error. Result and operand have different sizes
float4 g = as_type<float4>(f);

float4 f;
// Legal. g.xyz has same values as f.xyz.
float3 g = as_type<float3>(f);

2.21 Implicit Type Conversions

Implicit conversions between scalar built-in types (except void) are supported. When an implicit conversion is done, it is not just a re-interpretation of the expression’s value but a conversion of that value to an equivalent value in the new type. For example, the integer value 5 is converted to the floating-point value 5.0. Bfloat is an extended floating point type that only allows implicit conversion to a type of greater floating point rank. While bfloat can be implicitly converted to float, it cannot be implicitly converted to half, and neither float nor half can be implicitly converted to bfloat.

All vector types are considered to have a higher conversion rank than scalar types. Implicit conversions from a vector type to another vector or scalar type are not permitted and a compilation error results. For example, the following attempt to convert from a 4-component integer vector to a 4-component floating-point vector fails.

int4 i;
float4 f = i;  // compile error.

Implicit conversions from scalar-to-vector types are supported. The scalar value is replicated in each element of the vector. The scalar may also be subject to the usual arithmetic conversion to the element type used by the vector.

For example:
float4 f = 2.0f;  // f = (2.0f, 2.0f, 2.0f, 2.0f)

Implicit conversions from scalar-to-matrix types and vector-to-matrix types are not supported and a compilation error results. Implicit conversions from a matrix type to another matrix, vector or scalar type are not permitted and a compilation error results.

Implicit conversions for pointer types follow the rules described in the C++14 Specification.
3 Operators

All OS: Scalar, vector, and matrix operators supported since Metal 1.

For indirect command buffers, the assignment operator (\(\equiv\)) does not copy the contents of a command. For more about copying commands in indirect command buffers, see section 6.16.3.

3.1 Scalar and Vector Operators

This section lists both binary and unary operators and describes their actions on scalar and vector operands.

1. The arithmetic binary operators, add (\(+\)), subtract (\(-\)), multiply (\(\times\)) and divide (\(/\)), act upon scalar and vector, integer, and floating-point data type operands. Following the usual arithmetic conversions, all arithmetic operators return a result of the same built-in type (integer or floating point) as the type of the operands. After conversion, the following cases are valid:
   - If the two operands of the arithmetic binary operator are scalars, the result of the operation is a scalar.
   - If one operand is a scalar, and the other operand is a vector,
     - The scalar is converted to the element type used by the vector operand.
     - The scalar type is then widened to a vector that has the same number of components as the vector operand.
     - The operation is performed component-wise, which results in a same size vector.
   - If the two operands are vectors of the same size, the operation is performed component-wise, which results in a same size vector.

Division on integer types that result in a value that lies outside of the range bounded by the maximum and minimum representable values of the integer type, such as TYPE_MIN/\(-1\) for signed integer types or division by zero, does not cause an exception but results in an unspecified value. Division by zero for floating-point types results in \(\pm\infty\) or NaN, as prescribed by IEEE-754. (For more about the numerical accuracy of floating-point operations, see section 7.)

Because bfloat and half are not implicitly convertible to each other, the operators do not support mixing bfloat and half.

2. The modulus operator (\(\%\)) acts upon scalar and vector integer data type operands. The modulus operator returns a result of the same built-in type as the type of the operands, after the usual arithmetic conversions. The following cases are valid:
   - If the two operands of the modulus operator are scalars, the result of the operation is a scalar.
   - If one operand is a scalar, and the other is a vector:
     - The scalar is converted to the element type of the vector operand.
• The scalar type is then widened to a vector that has the same number of components as the vector operand.
• The operation is performed component-wise, which results in a same-size vector.
• If the two operands are vectors of the same size, the operation is performed component-wise, which results in a same-size vector.

For any component computed with a second operand that is zero, the modulus operator result is undefined. If one or both operands are negative, the results are undefined. Results for other components with non-zero operands remain defined.

If both operands are non-negative, the remainder is non-negative.

3. The arithmetic unary operators (+ and −) act upon scalar and vector, integer, and floating-point type operands.

4. The arithmetic post- and pre-increment and decrement operators (−− and ++) have scalar and vector integer type operands. All unary operators work component-wise on their operands. The result is the same type as the operand. For post- and pre-increment and decrement, the expression needs to be assignable to an lvalue. Pre-increment and post-decrement add or subtract 1 to the contents of the expression on which they operate, and the value of the pre-increment or pre-decrement expression is the resulting value of that modification. Post-increment and post-decrement expressions add or subtract 1 to the contents of the expression on which they operate, but the resulting expression has the expression’s value before execution of the post-increment or post-decrement.

5. The relational operators [greater-than (>), less-than (<), greater-than or equal to (≥), and less-than or equal to (≤)] act upon scalar and vector, integer, and floating-point type operands. The result is a Boolean (bool type) scalar or vector. After converting the operand type, the following cases are valid:

• If the two operands of the relational operator are scalars, the result of the operation is a Boolean.
• If one operand is a scalar, and the other is a vector:
  • The scalar is converted to the element type of the vector operand.
  • The scalar type is then widened to a vector that has the same number of components as the vector operand.
  • The operation is performed component-wise, which results in a Boolean vector.
• If the two operands are vectors of the same size, the operation is performed component-wise, which results in a same-size Boolean vector.

If either argument is a NaN, the relational operator returns false. To test a relational operation on any or all elements of a vector, use the any and all built-in functions in the context of an if(...) statement. (For more about any and all functions, see section 6.4.)

6. The equality operators, equal (==) and not equal (!=), act upon scalar and vector, integer and floating-point type operands. All equality operators result in a Boolean scalar or vector. After converting the operand type, the following cases are valid:
• If the two operands of the equality operator are scalars, the result of the operation is a Boolean.
• If one operand is a scalar, and the other is a vector:
  • The scalar is converted to the element type of the vector operand.
  • The scalar type is then widened to a vector that has the same number of components as the vector operand.
  • The operation is performed component-wise, which results in a Boolean vector.
• If the two operands are vectors of the same size, the operation is performed component-wise, which results in a same-size Boolean vector.

All other cases of implicit conversions are illegal. If one or both arguments is NaN, the equality operator equal (==) returns false. If one or both arguments is NaN, the equality operator not equal (!=) returns true.

7. The bitwise operators [and (&), or (|), exclusive or (^), not (~)] can act upon all scalar and vector built-in type operands, except the built-in scalar and vector floating-point types.
   • For built-in vector types, the bitwise operators are applied component-wise.
   • If one operand is a scalar and the other is a vector,
     • The scalar is converted to the element type used by the vector operand.
     • The scalar type is then widened to a vector that has the same number of components as the vector operand.
     • The bitwise operation is performed component-wise resulting in a same-size vector.

8. The logical operators [and (&&), or (||)] act upon two operands that are Boolean expressions. The result is a scalar or vector Boolean.
9. The logical unary operator not (!) acts upon one operand that is a Boolean expression. The result is a scalar or vector Boolean.
10. The ternary selection operator (?:) acts upon three operands that are expressions (exp1?exp2:exp3). This operator evaluates the first expression exp1, which must result in a scalar Boolean. If the result is true, the second expression is evaluated; if false, the third expression is evaluated. Only one of the second and third expressions is evaluated. The second and third expressions can be of any type if:
   • the types of the second and third expressions match,
   • or there is a type conversion for one of the expressions that can make their types match (for more about type conversions, see section 2.12),
   • or one expression is a vector and the other is a scalar, and the scalar can be widened to the same type as the vector type. The resulting matching type is the type of the entire expression.

11. The ones’ complement operator (~) acts upon one operand that needs to be of a scalar or vector integer type. The result is the ones’ complement of its operand.

The right-shift (>>) and left-shift (<<) operators act upon all scalar and vector integer type operands. For built-in vector types, the operators are applied component-wise. For the right-shift (>>) and left-shift (<<) operators, if the first operand is a scalar, the
rightmost operand needs to be a scalar. If the first operand is a vector, the rightmost operand can be a vector or scalar.

The result of `E1 << E2` is `E1` left-shifted by the `log2(N)` least significant bits in `E2` viewed as an unsigned integer value:

- If `E1` is a scalar, `N` is the number of bits used to represent the data type of `E1`.
- Or if `E1` is a vector, `N` is the number of bits used to represent the type of `E1` elements.

For the left-shift operator, the vacated bits are filled with zeros.

The result of `E1 >> E2` is `E1` right-shifted by the `log2(N)` least significant bits in `E2` viewed as an unsigned integer value:

- If `E1` is a scalar, `N` is the number of bits used to represent the data type of `E1`.
- Or if `E1` is a vector, `N` is the number of bits used to represent the type of `E1` elements.

For the right-shift operator, if `E1` has an unsigned type or if `E1` has a signed type and a nonnegative value, the vacated bits are filled with zeros. If `E1` has a signed type and a negative value, the vacated bits are filled with ones.

12. The assignment operator behaves as described by the C++14 Specification. For the `lvalue = expression` assignment operation, if `expression` is a scalar type and `lvalue` is a vector type, the scalar is converted to the element type used by the vector operand. The scalar type is then widened to a vector that has the same number of components as the vector operand. The operation is performed component-wise, which results in a same size vector.

Other C++14 operators that are not detailed above (such as `sizeof(T)`, unary `&` operator, and comma `(,)` operator) behave as described in the C++14 Specification.

Unsigned integers shall obey the laws of arithmetic modulo `2n`, where `n` is the number of bits in the value representation of that particular size of integer. The result of signed integer overflow is undefined.

For integral operands the divide (`/`) operator yields the algebraic quotient with any fractional part discarded. (This is often called truncation towards zero.) If the quotient `a/b` is representable in the type of the result, `(a/b)*b + a%b` is equal to `a`.

### 3.2 Matrix Operators

The arithmetic operators add `(+),` subtract `(-)` operate on matrices. Both matrices must have the same numbers of rows and columns. The operation is done component-wise resulting in the same size matrix. The arithmetic operator multiply `(*)` acts upon:

- a scalar and a matrix
- a matrix and a scalar
- a vector and a matrix
- a matrix and a vector
- a matrix and a matrix
If one operand is a scalar, the scalar value is multiplied to each component of the matrix resulting in the same-size matrix. A right vector operand is treated as a column vector and a left vector operand as a row vector. For vector-to-matrix, matrix-to-vector, and matrix-to-matrix multiplication, the number of columns of the left operand needs to be equal to the number of rows of the right operand. The multiply operation does a linear algebraic multiply, yielding a vector or a matrix that has the same number of rows as the left operand and the same number of columns as the right operand.

The following examples presume these vector, matrix, and scalar variables are initialized. The order of partial sums for the vector-to-matrix, matrix-to-vector, and matrix-to-matrix multiplication operations described below is undefined.

```c
float3 v;
float3x3 m, n;
float a = 3.0f;
```

The matrix-to-scalar multiplication:

```c
float3x3 m1 = m * a;
```

is equivalent to:

```c
m1[0][0] = m[0][0] * a;
m1[0][1] = m[0][1] * a;
m1[0][2] = m[0][2] * a;
m1[1][0] = m[1][0] * a;
m1[1][1] = m[1][1] * a;
m1[1][2] = m[1][2] * a;
m1[2][0] = m[2][0] * a;
m1[2][1] = m[2][1] * a;
m1[2][2] = m[2][2] * a;
```

The vector-to-matrix multiplication:

```c
float3 u = v * m;
```

is equivalent to:

```c
u.x = dot(v, m[0]);
u.y = dot(v, m[1]);
u.z = dot(v, m[2]);
```

The matrix-to-vector multiplication:

```c
float3 u = m * v;
```
is equivalent to:

\[ u.x = m[0].x \times v.x + m[1].x \times v.y + m[2].x \times v.z; \]
\[ u.y = m[0].y \times v.x + m[1].y \times v.y + m[2].y \times v.z; \]
\[ u.z = m[0].z \times v.x + m[1].z \times v.y + m[2].z \times v.z; \]

The matrix-to-matrix multiplication:

```c
float3x3 r = m * n; // m, n are float3x3
```

is equivalent to:

```c
r[0].x = m[0].x \times n[0].x + m[1].x \times n[0].y + m[2].x \times n[0].z;
```
```c
r[0].y = m[0].y \times n[0].x + m[1].y \times n[0].y + m[2].y \times n[0].z;
```
```c
r[0].z = m[0].z \times n[0].x + m[1].z \times n[0].y + m[2].z \times n[0].z;
```
```c
r[1].x = m[0].x \times n[1].x + m[1].x \times n[1].y + m[2].x \times n[1].z;
```
```c
r[1].y = m[0].y \times n[1].x + m[1].y \times n[1].y + m[2].y \times n[1].z;
```
```c
r[1].z = m[0].z \times n[1].x + m[1].z \times n[1].y + m[2].z \times n[1].z;
```
```c
r[2].x = m[0].x \times n[2].x + m[1].x \times n[2].y + m[2].x \times n[2].z;
```
```c
r[2].y = m[0].y \times n[2].x + m[1].y \times n[2].y + m[2].y \times n[2].z;
```
```c
r[2].z = m[0].z \times n[2].x + m[1].z \times n[2].y + m[2].z \times n[2].z;
```
4 Address Spaces

The Metal memory model describes the behavior and structure of memory objects in MSL programs. An address space attribute specifies the region of memory from where buffer memory objects are allocated. These attributes describe disjoint address spaces that can also specify access restrictions:

- device (see section 4.1)
- constant (see section 4.2)
- thread (see section 4.3)
- threadgroup (see section 4.4)
- threadgroup_imageblock (see section 4.5)
- ray_data (see section 4.6)
- object_data (see section 4.7)

All OS: device, threadgroup, constant, and thread attribute support since Metal 1 and ray_data attribute support since Metal 2.3.

iOS: Supports the threadgroup_imageblock attribute since Metal 2.

macOS: Supports the threadgroup_imageblock attribute since Metal 2.3.

All arguments to a graphics or kernel function that are a pointer or reference to a type needs to be declared with an address space attribute. For graphics functions, an argument that is a pointer or reference to a type needs to be declared in the device or constant address space. For kernel functions, an argument that is a pointer or reference to a type needs to be declared in the device, threadgroup, threadgroup_imageblock, or constant address space. The following example introduces the use of several address space attributes. (The threadgroup attribute is supported here for the pointer l_data only if foo is called by a kernel function, as detailed in section 4.4.)

```c
void foo(device int *g_data,
         threadgroup int *l_data,
         constant float *c_data)
{
}
```

The address space for a variable at program scope needs to be constant.

Any variable that is a pointer or reference needs to be declared with one of the address space attributes discussed in this section. If an address space attribute is missing on a pointer or reference type declaration, a compilation error occurs.

4.1 device Address Space

The device address space name refers to buffer memory objects allocated from the device memory pool that are both readable and writeable.

A buffer memory object can be declared as a pointer or reference to a scalar, vector or user-defined structure. In an app, Metal API calls allocate the memory for the buffer object, which determines the actual size of the buffer memory.
Some examples are:

```cpp
// An array of a float vector with four components.
device float4 *color;

struct Foo {
    float a[3];
    int b[2];
};

// An array of Foo elements.
device Foo *my_info;
```

Since you always allocate texture objects from the device address space, you do not need the `device` `address` attribute for texture types. You cannot directly access the elements of a texture object, so use the built-in functions to read from and write to a texture object (see section 6.12).

### 4.2 constant Address Space

The `constant address` space name refers to buffer memory objects allocated from the device memory pool that are read-only. You must declare variables in program scope in the `constant address` space and initialize them during the declaration statement. The initializer(s) expression must be a core constant expression. (Refer to section 5.20 of the C++14 specification.) The compiler may evaluate a core constant expression at compile time. Variables in program scope have the same lifetime as the program, and their values persist between calls to any of the compute or graphics functions in the program.

```cpp
constant float samples[] = { 1.0f, 2.0f, 3.0f, 4.0f };;
```

Pointers or references to the `constant address` space are allowed as arguments to functions. Writing to variables declared in the `constant address` space is a compile-time error. Declaring such a variable without initialization is also a compile-time error.

Buffers in the `constant address` space passed to kernel, vertex, and fragment functions have minimum alignment requirements based on the GPU. See “Minimum constant buffer offset alignment” in the Metal Feature Set Tables for more information.

### 4.3 thread Address Space

The `thread address` space refers to the per-thread memory address space. Variables allocated in this address space are not visible to other threads. Variables declared inside a graphics or kernel function are allocated in the `thread address` space.
[[kernel]] void
my_kernel(…)
{
    // A float allocated in the per-thread address space
    float x;

    // A pointer to variable x in per-thread address space
    thread float * p = &x;
    ...
}

4.4 threadgroup Address Space

A GPU compute unit can execute multiple threads concurrently in a threadgroup, and a GPU can execute a separate threadgroup for each of its compute units.

Threads in a threadgroup can work together by sharing data in threadgroup memory, which is faster on most devices than sharing data in device memory. Use the threadgroup address space to:

- Allocate a threadgroup variable in a kernel, mesh, or object function.
- Define a kernel, fragment, or object function parameter that’s a pointer to a threadgroup address.

See the Metal Feature Set Tables to learn which GPUs support threadgroup space arguments for fragment shaders.

Threadgroup variables in a kernel, mesh, or object function only exist for the lifetime of the threadgroup that executes the kernel. Threadgroup variables in a mid-render kernel function are persistent across mid-render and fragment kernel functions over a tile.

This example kernel demonstrates how to declare both variables and arguments in the threadgroup address space. (The [[threadgroup]] attribute in the code below is explained in section 5.2.1.)

kernel void
my_kernel(threadgroup float *sharedParameter [[threadgroup(0)]],
           ...)
{
    // Allocate a float in the threadgroup address space.
    threadgroup float sharedFloat;

    // Allocate an array of 10 floats in the threadgroup address space.
    threadgroup float sharedFloatArray[10];
    ...
}

For more information about the [[threadgroup(0)]] attribute, see section 5.2.1.
4.4.1 SIMD-groups and Quad-groups

macOS: SIMD-group support since Metal 2. Quad-group support since Metal 2.1.

iOS: Some SIMD-group function support since Metal 2.2. Quad-group support since Metal 2.

Within a threadgroup, you can divide threads into SIMD-groups, which are collections of threads that execute concurrently. The mapping to SIMD-groups is invariant for the duration of a kernel's execution, across dispatches of a given kernel with the same launch parameters, and from one threadgroup to another within the dispatch (excluding the trailing edge threadgroups in the presence of nonuniform threadgroup sizes). In addition, all SIMD-groups within a threadgroup needs to be the same size, apart from the SIMD-group with the maximum index, which may be smaller, if the size of the threadgroup is not evenly divisible by the size of the SIMD-groups.

A quad-group is a SIMD-group with the thread execution width of 4.

For more about kernel function attributes for SIMD-groups and quad-groups, see section 5.2.3.6. For more about threads and thread synchronization, see section 6.9 and its subsections:

- For more about thread synchronization functions, including a SIMD-group barrier, see section 6.9.1.
- For more about SIMD-group functions, see section 6.9.2.
- For more about quad-group functions, see section 6.9.3.

4.5 threadgroup_imageblock Address Space

The threadgroup_imageblock address space refers to objects allocated in threadgroup memory that are only accessible using an imageblock<T, L> object (see section 2.11). A pointer to a user-defined type allocated in the threadgroup_imageblock address space can be an argument to a tile shading function (see section 5.1.9). There is exactly one threadgroup per tile, and each threadgroup can access the threadgroup memory and the imageblock associated with its tile.

- Variables allocated in the threadgroup_imageblock address space in a kernel function are allocated for each threadgroup executing the kernel, are shared by all threads in a threadgroup, and exist only for the lifetime of the threadgroup that executes the kernel. Each thread in the threadgroup uses explicit 2D coordinates to access imageblocks. Do not assume any particular spatial relationship between the threads and the imageblock. The threadgroup dimensions may be smaller than the tile size.

4.6 ray_data Address Space

All OS: ray_data address supported since Metal 2.3.

The ray_data address space refers to objects allocated in a memory that is only accessible in an intersection function (see section 5.1.6) for Ray Tracing. Intersection functions can read and write to a custom payload using [[payload]] attribute (see Table 5.10) in the ray_data address space. When a shader calls intersect() (see section 6.18.2) with a payload, the
system copies the payload to the `ray_data` address space, calls the intersection function, and when the intersection function returns, it copies the payload back out.

### 4.7 object_data Address Space

All OS: object_data address supported since Metal 3.

Object functions use the object_data address space to pass a payload to a mesh function (see section 5.2.3.9). The object_data address space behaves like the threadgroup address space in that the programming model is explicitly cooperative within the threadgroup. Developers should use the threads in the threadgroup to efficiently compute the payload and value `mesh_grid_properties::set_threadgroups_per_grid`. The payload in the object_data address space is not explicitly bound or initialized, and the implementation manages its lifetime.
5 Function and Variable Declarations

This chapter describes how you declare functions, arguments, and variables. It also details how you often use attributes to specify restrictions to functions, arguments, and variables.

5.1 Functions

Metal 1 or later supports the kernel, vertex, and fragment attributes for every OS. Metal 2.3 or later supports the C++ attributes:

- [[vertex]] or vertex (See section 5.1.1)
- [[fragment]] or fragment (See section 5.1.2)
- [[kernel]] or kernel (See section 5.1.3)
- [[visible]] (See section 5.1.4)
- [[stitchable]] (See section 5.1.5)
- [[intersection(...)]] (See section 5.1.6)
- [[object]] (See section 5.1.7)
- [[mesh]] (See section 5.1.8)

Make a function accessible to the Metal API by adding one of these function attributes at the start of a function, which makes it a qualified function. Kernel, vertex, and fragment functions can't call one another without triggering a compilation error, but they may call other functions that use the [[visible]] attribute. They can also call functions with the [[intersection(...)]] attribute by calling intersect() (see section 6.18.2).

Prior to Metal 2.2, the Metal compiler ignores namespace identifiers for kernel, vertex, and fragment functions. In Metal 2.2 and later, if you declare a qualified function within a namespace, you must include the namespace identifier with the function’s name each time you refer it to a Metal Framework API. This example declares two kernel functions in different namespaces.

namespace outer {
    [[kernel]] void functionA() {...
    namespace inner {
        [[kernel]] void functionB() {...
    }
}...}

Refer to a function in a namespace by prepending the function’s name with the namespace’s identifier followed by two colons.

Outer::functionA

Similarly, refer to a function in a nested namespace by prepending the function’s name with all namespaces in order and separating each with two colons.

Outer::inner::functionB
5.1.1 Vertex Functions

You can declare the vertex or since Metal 2.3 [[vertex]] attribute only for a graphics function. Metal executes a vertex function for each vertex in the vertex stream and generates per-vertex output. The following example shows the syntax for declaring a vertex.

```plaintext
vertex void
my_vertex_func(...) 
{...}
[[vertex]] void
vertex_func2(...) 
{...}
```

For a vertex function, the return type identifies the output generated by the function. If the vertex function does not generate output, it shall return void and can only be used in a render pipeline with rasterization disabled.

5.1.1.1 Post-Tessellation Vertex Functions

All OS: post-tessellation vertex function (patch attribute) supported since Metal 1.2.

The post-tessellation vertex function calculates the vertex data for each surface sample on the patch produced by the fixed-function tessellator. The inputs to the post-tessellation vertex function are:

- Per-patch data.
- Patch control point data.
- The tessellator stage output (the normalized vertex location on the patch).

The post-tessellation vertex function generates the final vertex data for the tessellated triangles. For example, to add additional detail (such as displacement mapping values) to the rendered geometry, the post-tessellation vertex function can sample a texture to modify the vertex position by a displacement value.

After the post-tessellation vertex function has executed, the tessellated primitives are rasterized.

The post-tessellation vertex function is a vertex function identified using the ordinary vertex function attribute.

5.1.1.2 Patch Type and Number of Control Points Per-Patch

The [[patch]] attribute is required for the post-tessellation vertex function.

For macOS, the [[patch(patch-type, N)]] attribute must specify both the patch type (patch-type is either quad or triangle) and the number of control points in the patch (N needs to be a value from 0 to 32). For iOS, specifying the patch-type is required, but the number of control points is optional.
If the number of control points are specified in the post-tessellation vertex function, this number must match the number of control points provided to the `drawPatches` or `drawIndexedPatches` API.

Example:
```
[[patch(quad)]]
[[vertex]] vertex_output
my_post_tessellation_vertex(...)
{...}

[[patch(quad, 16)]]
[[vertex]] vertex_output
my_bezier_vertex(...)
{...}
```

### 5.1.2 Fragment Functions

You can declare the `fragment` or since Metal 2.3 `[[fragment]]` attribute only for a graphics function. Metal executes a fragment function for each fragment in the fragment stream and their associated data and generates per-fragment output. The following example shows the syntax for declaring a fragment function with the `fragment` attribute.

```
[[fragment]] void
my_fragment_func(...)
{...}
[[fragment]] void
fragment_func2(...){...}
```

For graphics functions, the return type identifies whether the output generated by the function is either per-vertex or per-fragment. If the fragment function does not generate output, it shall return `void`.

To request performing fragment tests before the fragment function executes, use the `[[early_fragment_tests]]` function attribute with a fragment function, as shown in the example below.

```
[[early_fragment_tests]]
fragment float4
my_fragment( ... )
{...}
```
It is an error if the return type of the fragment function declared with the `[[early_fragment_tests]]` attribute includes a depth or stencil value; that is, if the return type of this fragment function includes an element declared with the `[[depth(depth_attribute)]]` or `[[stencil]]` attribute.

It is an error to use the `[[early_fragment_tests]]` attribute with any function that is not a fragment function; that is, not declared with the `fragment` attribute.

### 5.1.3 Compute Functions (Kernels)

A compute function (also called a "kernel") is a data-parallel function that is executed over a 1-, 2-, or 3D grid. The following example shows the syntax for declaring a compute function with the `kernel` or since Metal 2.3 `[[kernel]]` attribute.

```objc
kernel
void my_kernel(…) {...}
```

```objc
[[kernel]]
void my_kernel2(…) {...}
```

Functions declared with the `kernel` or `[[kernel]]` attribute must return `void`.

You can use the `[[max_total_threads_per_threadgroup]]` function attribute with a kernel function to specify the maximum threads per threadgroup.

Below is an example of a kernel function that uses this attribute:

```objc
[[max_total_threads_per_threadgroup(x)]]
kernlel void
my_kernel(…)
{…}
```

If the `[[max_total_threads_per_threadgroup]]` value is greater than the `MTLDevice maxThreadsPerThreadgroup` property, then compute pipeline state creation shall fail.

### 5.1.4 Visible Functions

All OS `[[visible]]` functions are supported since Metal 2.3.

A function with a `[[visible]]` attribute is a function that is visible from the Metal Framework API, i.e., one may get a `MTLFunction` object of this function. It is legal to take the address of a visible function and get a visible function pointer. The visible function pointers may be used with the `visible_function_table` type (section 2.15). It is legal for other functions to directly call a visible function. Note that visible function, like other qualified functions, is split into their own translation unit. When a function directly calls a visible function, the developer needs to pass it in the pipeline descriptor.
The following example with the [[visible]] attribute.

```cpp
[[visible]] float my_visible(device int *data, int data_offset) {...}
```

### 5.1.5 Stitchable Functions

All OS: [[stitchable]] functions are supported since Metal 2.4.

A function with a [[stitchable]] attribute is a function that can be used in the Metal Framework Function Stitching API. The [[stitchable]] attribute implies [[visible]], which means that stitchable functions can be used in all contexts where a visible function can be used as described in Sec 5.1.4. The compiler will generate additional metadata for stitchable functions to enable these functions to be used with the Metal Function Stitching API. Developers should use this attribute only if they need this functionality as the metadata will increase the code size of the function.

```cpp
[[stitchable]] float my_func(device float *data, texture2d<float> tex) {...}
```

### 5.1.6 Intersection Functions

All OS: [[intersection(primitive_type, intersection_tags...)]] functions are supported since Metal 2.3

One can declare a custom intersection function to be used with raytracing by using the [[intersection(primitive_type, intersection_tags...)]] attribute. Intersection functions are called when the shader calls `intersect()` (see section 6.18) to determine if a potential ray intersection is valid or if traversal should continue. Metal supports the following types of intersection functions:

#### Table 5.1. Intersection Function Primitive Types

<table>
<thead>
<tr>
<th>Primitive Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>triangle</td>
<td>Indicates that this is an intersection function that extends the default triangle intersection test.</td>
</tr>
<tr>
<td>bounding_box</td>
<td>Indicates that this is an intersection function which is run when a ray intersects the bounding box.</td>
</tr>
<tr>
<td>curve</td>
<td>Indicates that this is an intersection function that extends the default curve intersection test.</td>
</tr>
</tbody>
</table>

You may pass zero or more intersection tags as described in Table 2.9 from section 2.17. Some examples are:
[[intersection(triangle, triangle_data, instancing, world_space_data)]]
bool triangleIntersectionFunction(...) {...}

[[intersection(bounding_box, triangle_data, instancing, world_space_data)]]
UserResult boundingBoxIntersectionFunction(...) {...}

The intersection function `primitive_type` and `intersection_tags` control the allowable input and output attributes (see Section 5.2.3.7).
Intersection functions support passing buffer arguments from device and constant address space.
Intersection functions don’t support passing texture arguments to an intersection function. However, you can pass a texture using an argument buffer.
Intersection functions don’t support threadgroup memory.
Intersection functions don’t support threadgroup_barrier or simdgroup_barrier. If they are used, the result is undefined.
Intersection functions may or may not be run in the same SIMD-group as the thread which launched the intersection operation: The implementation is permitted to regroup or repack candidate intersections to improve efficiency before launching SIMD-groups to do intersection testing.
If the acceleration structure traversal finds a procedural box primitive, and the intersection function is a triangle tester (or vice versa), this is an application error and behavior is undefined.

5.1.7 Object Functions

All OS: `[[object]]` functions are supported since Metal 3.
A function with an `[[object]]` attribute is an object function in the mesh pipeline. An object function is a data-parallel function executed over a 1-, 2-, or 3D compute grid that can launch compute grids to a second mesh stage and with a data payload. Object functions must return `void`.
Input built-in variables to object functions are described in section 5.2.3.9. The `[[payload]]` attribute tags a buffer that the object function exports to the mesh shader as a read-only buffer. It may be specified once per function.
You can use the `[[max_total_threads_per_threadgroup]]` function attribute with an object function to specify the maximum threads per threadgroup.

You can use the `[[max_total_threadgroups_per_mesh_grid(size)]]` on an object function to specify the maximum threadgroups per mesh grid. The following is an example using the `[[object]]` attribute.
#define kMeshThreadgroups 32

struct ObjectOutput {
  // User-defined payload; one entry for each mesh threadgroup. This
  // is an array because the data will be shared by the mesh grid.
  float value[kMeshThreadgroups];
};

[[object, max_total_threadgroups_per_mesh_grid(kMeshThreadgroups)]]
void objectShader(uint threadgroup_size [[threads_per_threadgroup]],
  uint lane [[thread_index_in_threadgroup]],
  object_data ObjectOutput& output [[payload]],
  mesh_grid_properties mgp) {...}

## 5.1.8 Mesh Functions

All OS: [[[mesh]]] functions are supported since Metal 3.

A function with a [[[mesh]]] attribute is a mesh function in the mesh pipeline. A mesh function is a data-parallel function that can optionally export a mesh object representing a chunk of geometry to the rasterization pipeline. The mesh object is a parameter of the mesh function. If no mesh object is exported, rasterization is disabled. Input built-in variables to mesh functions are described in section 5.2.3.10. Mesh functions must return `void`.

You can use the `[[max_total_threads_per_threadgroup]]` function attribute with a mesh function to specify the maximum threads per threadgroup. The following is an example using the `[[mesh]]` attribute.

```cpp
struct vertex_t {
  float4 clip_pos [[position]];  
  float3 world_pos;  
  float3 color;  
  // other user-defined properties
};

struct primitive_t {
  float3 normal;
};

// A mesh declaration that can export one cube.
using cube_mesh_t = metal::mesh<vertex_t, primitive_t, 
  8 /*corners*/,
  6*2 /*faces*/,
  metal::topology::triangle>;

struct view_info_t {
  float4x4 view_proj;
};
```
struct cube_info_t {
    float4x3 world;
    float3 color;
};

[[mesh, max_total_threads_per_threadgroup(12)]]
void cube_stage(cube_mesh_t output,
    const object_data cube_info_t &cube [[payload]],
    constant view_info_t &view [[buffer(0)]],
    uint gid [[threadgroup_position_in_grid]],
    uint lane [[thread_index_in_threadgroup]]) {...}

5.1.9  Tile Functions

iOS: Tile function support since Metal 2.

macOS: Tile function support since Metal 2.3.

A tile shading function is a special type of compute kernel or fragment function that can execute inline with graphics operations and take advantage of the Tile-Based Deferred Rendering (TBDR) architecture. With TBDR, commands are buffered until a large list of commands accumulates. The hardware divides the framebuffer into tiles and then renders only the primitives that are visible within each tile. Tile shading functions support performing compute operations in the middle of rendering, which can access memory more efficiently by reducing round trips to memory and utilizing high-bandwidth local memory.

A tile function launches a set of threads called a dispatch, which is organized into threadgroups and grids. You may launch threads at any point in a render pass and as often as needed. Tile functions barrier against previous and subsequent draws, so a tile function does not execute until all earlier draws have completed. Likewise, later draws do not execute until the tile function completes.

GPUs always process each tile and each dispatch to completion. Before processing the next tile, all draws and dispatches for a tile launch in submission.

Tile functions have access to 32 KB of threadgroup memory that may be divided between imageblock storage and threadgroup storage. (For more about the threadgroup memory size, see section 4.4.) The imageblock size is dependent on the tile width, tile height, and the bit depth of each sample. Either the render pass attachments (which use implicit imageblock layout; see section 5.6.3.1) or function-declared structures (which use explicit imageblock layout; see section 5.6.3.2) determines the bit depth of the sample. For more about how kernel functions utilize the threadgroup_imageblock address space, see section 4.5.

5.1.10  Host Name Attribute

Starting from Metal 2.2, you can override the default name that the Metal Framework API uses to refer to a qualified function. Add the [[host_name(name)]] attribute to the function declaration, where name is the string literal that the Metal Framework API will use to reference the function name. The compiler raises a compile time error if you give different functions the same name. For example,
5.1.11  Templated Qualified Functions

Starting from Metal 2.2, you can use templates for qualified functions (e.g. vertex, fragment, visible, and kernel functions) declarations. You must explicitly instantiate the template to force the compiler to emit code for a given specialization. For example,

```metal
template<typename T>
kernel void bar(device T *x) { ... }
// Explicit specialization of `bar<T>` with [T = int]
template kernel void bar(device int *);
```

The compiler gives all specializations the same name unless one uses the `[[host_name(name)]]` attribute to provide a different name for each specialization.

```metal
// Explicit specialization of `bar<T>` with [T = int] and host_name
// "bar_int"
template [[host_name("bar_int")]] kernel void bar(device int *);

// Explicit specialization of `bar<T>` with [T = float] and host_name
// "bar_float"
template [[host_name("bar_float")]] kernel void bar(device float *);
```

5.2  Function Arguments and Variables

Most inputs and outputs to graphics (vertex or fragment) and kernel functions are passed as arguments. (Initialized variables in the constant address space and samplers declared in program scope are inputs and outputs that do not have to be passed as arguments.) Starting in Metal 3.1, built-in input variables for kernel, mesh, and object shaders can now be declared in program scope, avoiding the need for passing them as arguments. This applies if they are not used in a dynamic library or separately compiled binary function.

Arguments to graphics and kernel functions can be any of the following:

- Device buffer — A pointer or reference to any data type in the **device** address space (see section 2.8).
- Constant buffer — A pointer or reference to any data type in the **constant** address space (see section 2.8).
- A **texture** object (see section 2.9) or an array of textures.
- A **texture_buffer** object (see section 2.9.1) or an array of texture buffers.
- A **sampler** object (see section 2.10) or an array of samplers.

- A buffer shared between threads in a threadgroup — a pointer to a type in the threadgroup address space that can only be used as arguments for kernel functions.
- An imageblock (see section 2.11).
- An argument buffer (see section 2.13).
- A visible function table (see section 2.15) for kernel functions. As of Metal 2.4, visible function table can also be used in graphic functions.
- An intersection function table (see section 2.17.3) for kernel functions.
- An acceleration structure (see section 6.18.1) for intersection functions.
- A structure with elements that are buffers, textures, or texture buffers.

Buffers (device) specified as argument values to a graphics or kernel function cannot alias; that is, a buffer passed as an argument value cannot overlap another buffer passed to a separate argument of the same graphics or kernel function.

You cannot declare arguments to graphics and kernel functions to be of type size_t, ptdiff_t, or a structure and/or union that contains members declared to be one of these built-in scalar types.

The arguments to these functions are often specified with attributes to provide further guidance on their use. Attributes are used to specify:

- The resource location for the argument (see section 5.2.1).
- Built-in variables that support communicating data between fixed-function and programmable pipeline stages (see section 5.2.3).
- Which data is sent down the pipeline from vertex function to fragment function (see section 5.2.4).

### 5.2.1 Locating Buffer, Texture, and Sampler Arguments

For each argument, an attribute can be optionally specified to identify the location of a buffer, texture, or sampler to use for this argument type. The Metal framework API uses this attribute to identify the location for these argument types.

- Device and constant buffers: [[buffer(index)]]
- Textures (including texture buffers): [[texture(index)]]
- Samplers: [[sampler(index)]]
- Threadgroup buffers: [[threadgroup(index)]]

The `index` value is an unsigned integer that identifies the location of an assigned buffer, texture or sampler argument. (A texture buffer is a specific type of texture.) The proper syntax is for the attribute to follow the argument or variable name.

The example below is a simple kernel function, `add_vectors`, that adds an array of two buffers in the device address space, `inA` and `inB`, and returns the result in the buffer out. The attributes `([buffer(index)])` specify the buffer locations for the function arguments.

```c
[[kernel]] void
add_vectors(const device float4 *inA [[buffer(0)]],
             const device float4 *inB [[buffer(1)]],
             device float4 *out [[buffer(2)]],
             uint id [[thread_position_in_grid]])
```
out[id] = inA[id] + inB[id];
}

The example below shows attributes used for function arguments of several different types (a buffer, a texture, and a sampler):
[[kernel]] void
my_kernel(device float4 *p [[buffer(0)]],
        texture2d<float> img [[texture(0)]],
        sampler sam [[sampler(1)]])
{...

If the location indices are not specified, the Metal compiler assigns them using the first available location index. In the following example, src is assigned texture index 0, dst texture index 1, s sampler index 0, and u buffer index 0:
kernel void
my_kernel(texture2d<half> src,
        texture2d<half, access::write> dst,
        sampler s,
        device myUserInfo *u)
{...

In the following example, some kernel arguments have explicitly assigned location indices and some do not. src is explicitly assigned texture index 0, and f is explicitly assigned buffer index 10. If you assign location indices using function constants (section 5.8), the compiler does not consider those entries when assigning indices. The other arguments are assigned the first available location index: dst texture index 1, s sampler index 0, and u buffer index 0.
kernel void
my_kernel(texture2d<half> src [[texture(0)]],
        texture2d<half, access::write> dst,
        sampler s,
        device myUserInfo *u,
        device float *f [[buffer(10)]])
{...

Each attribute (buffer, threadgroup, texture, and sampler) represents a group of resources. The index values specified on the arguments shall be unique within each resource group. Multiple buffer, texture or sampler arguments with the same index value generate a compilation error unless they are declared with a function constant attribute (see section 5.8.1).

5.2.1.1 Vertex Function Example with Resources and Outputs to Device Memory
The following example is a vertex function, render_vertex, which outputs to device memory in the array xform_output, which is a function argument specified with the device attribute
(introduced in section 4.1). All the render_vertex function arguments are specified with the buffer(0), buffer(1), buffer(2), and buffer(3) attributes (introduced in section 5.2.1). For more about the position attribute shown in this example, see section 5.2.3.3.

```cpp
struct VertexOutput {
    float4 position [[position]];
    float4 color;
    float2 texcoord;
};

struct VertexInput {
    float4 position;
    float3 normal;
    float2 texcoord;
};

constexpr constant uint MAX_LIGHTS = 4;

struct LightDesc {
    uint num_lights;
    float4 light_position[MAX_LIGHTS];
    float4 light_color[MAX_LIGHTS];
    float4 light_attenuation_factors[MAX_LIGHTS];
};

vertex void render_vertex(const device VertexInput* v_in [[buffer(0)]],
                          constant float4x4& mvp_matrix [[buffer(1)]],
                          constant LightDesc& light_desc [[buffer(2)]],
                          device VertexOutput* xform_output [[buffer(3)]],
                          uint v_id [[vertex_id]])
{
    VertexOutput v_out;
    v_out.position = v_in[v_id].position * mvp_matrix;
    v_out.color = do_lighting(v_in[v_id].position, v_in[v_id].normal, light_desc);
    v_out.texcoord = v_in[v_id].texcoord;

    // Output the position to a buffer.
    xform_output[v_id] = v_out;
}
```

### 5.2.1.2 Raster Order Groups

All OS: raster order group attribute supported since Metal 2.

Loads and stores to buffers (in device memory) and textures in a fragment function are unordered. The [[raster_order_group(index)]] attribute used for a buffer or texture
guarantees the order of accesses for any overlapping fragments from different primitives that map to the same \((x, y)\) pixel coordinate and sample, if per-sample shading is active.

The \([[\text{raster-order-group(index)}]]\) attribute can be specified on a texture (which is always in device memory) or a buffer that is declared in device memory, but not in either the threadgroup or constant address space. The \([[\text{raster-order-group(index)}]]\) attribute cannot be used with a structure or class.

Fragment function invocations that mark overlapping accesses to a buffer or texture with the \([[\text{raster-order-group(index)}]]\) attribute are executed in the same order as the geometry is submitted. For overlapping fragment function invocations, writes performed by a fragment function invocation to a buffer or texture marked with the \([[\text{raster-order-group(index)}]]\) attribute needs to be available to be read by a subsequent invocation and must not affect reads by a previous invocation. Similarly, reads performed by a fragment function invocation must reflect writes by a previous invocation and must not reflect writes by a subsequent invocation.

The \(\text{index in } [[\text{raster-order-group(index)}]]\) is an integer value that specifies a rasterizer order ID, which provides finer grained control over the ordering of loads and stores. For example, if two buffers A and B are marked with different rasterizer order ID values, then loads and stores to buffers A and B for overlapping fragments can be synchronized independently.

Example:

```metal
fragment void
my_fragment(texture2d<float, access::read_write> texA
        [[raster-order-group(0), texture(0)]],
...
{
    ushort2 coord;
    float4 clr = texA.read(coord);
    // do operations on clr
    clr = ...;
    texA.write(clr, coord);
}
```

For an argument buffer, you can use the \([[\text{raster-order-group(index)}]]\) attribute on a buffer or texture member in a structure.

### 5.2.2 Attributes to Locate Per-Vertex Inputs

A vertex function can read per-vertex inputs by indexing into a buffer(s) passed as arguments to the vertex function using the vertex and instance IDs. In addition, you can also declare per-vertex input with the \([[\text{stage_in}}]]\) attribute and pass that input as an argument. For per-vertex input passed as an argument declared with the \([[\text{stage_in}}]]\) attribute, each element of the per-vertex input must specify the vertex attribute location as \([[\text{attribute(index)}}]]\). For more about the \([[\text{stage_in}}]]\) attribute, see section 5.2.4.

The \(\text{index value}\) is an unsigned integer that identifies the assigned vertex input location. The proper syntax is for the attribute to follow the argument or variable name. The Metal API uses

---

2023-06-02 | Copyright © 2023 Apple Inc. | All Rights Reserved.
this attribute to identify the location of the vertex buffer and describe the vertex data such as the buffer to fetch the per-vertex data from, its data format, and its stride.

The following example shows how to assign vertex attributes to elements of a vertex input structure that is passed to a vertex function using the `stage_in` attribute:

```cpp
struct VertexInput {
    float4 position [[attribute(0)]];
    float3 normal   [[attribute(1)]];
    half4 color     [[attribute(2)]];
    half2 texcoord  [[attribute(3)]];
};

constexpr constant uint MAX_LIGHTS = 4;

struct LightDesc {
    uint num_lights;
    float4 light_position[MAX_LIGHTS];
    float4 light_color[MAX_LIGHTS];
    float4 light_attenuation_factors[MAX_LIGHTS];
};

castexpr sampler s = sampler(coord::normalized,
address::clamp_to_zero,
    filter::linear);

derived vertex VertexOutput render_vertex(VertexInput v_in [[stage_in]],
    constant float4x4& mvp_matrix [[buffer(1)]],
    constant LightDesc& lights [[buffer(2)]],
    uint v_id [[vertex_id]])
{
    VertexOutput v_out;
    ... return v_out;
}
```

The example below shows how both buffers and the `stage_in` attribute can be used to fetch per-vertex inputs in a vertex function.

```cpp
struct VertexInput {
    float4 position [[attribute(0)]];
    float3 normal   [[attribute(1)]];
};

struct VertexInput2 {
    half4 color;
    half2 texcoord[4];
};
```
constexpr constant uint MAX_LIGHTS = 4;

struct LightDesc {
    uint num_lights;
    float4 light_position[MAX_LIGHTS];
    float4 light_color[MAX_LIGHTS];
    float4 light_attenuation_factors[MAX_LIGHTS];
};

castexpr sampler s = sampler(coord::normalized,
address::clamp_to_zero,
    filter::linear);

vertex VertexOutput
render_vertex(VertexInput v_in [[stage_in]],
    VertexInput2 v_in2 [[buffer(0)]],
    constant float4x4& mvp_matrix [[buffer(1)]],
    constant LightDesc& lights [[buffer(2)]],
    uint v_id [[vertex_id]])
{
    VertexOutput vOut;
    ...
    return vOut;
}

A post-tessellation vertex function can read the per-patch and patch control-point data. The post-tessellation vertex function specifies the patch control-point data as the following templated type:

patch_control_point<T>

Where T is a user defined structure. Each element of T must specify an attribute location using [[attribute(index)]].

All OS: patch control-point templated type supported since Metal 1.2.

The patch_control_point<T> type supports these member functions:
• constexpr size_t size() const;
which returns the number of control-points in the patch.
• constexpr const_reference operator[](size_t pos) const;
which returns the data for a specific patch control point that pos identifies.

Example:

struct ControlPoint {
    int3 patchParam [[attribute(0)]];
    float3 P [[attribute(1)]];
    float3 P1 [[attribute(2)]];
    float3 P2 [[attribute(3)]];
    float2 vSegments [[attribute(4)]];
}
};

struct PerPatchData {
    float4 patchConstant    [[attribute(5)]];
    float4 someOtherPatchConstant [[attribute(6)]];
};

struct PatchData {
    patch_control_point<ControlPoint> cp;   // Control-point data
    PerPatchData patchData;                // Per-patch data
};

[[patch(quad)]]
vertex VertexOutput
post_tess_vertex_func(PatchData input [[stage_in ]], ...)
{...

5.2.3 Attributes for Built-in Variables

Some graphics operations occur in the fixed-function pipeline stages and need to provide values to or receive values from graphics functions. *Built-in* input and output variables are used to communicate values between the graphics (vertex and fragment) functions and the fixed-function graphics pipeline stages. Attributes are used with arguments and the return type of graphics functions to identify these built-in variables.

5.2.3.1 Vertex Function Input Attributes

Table 5.2 lists the built-in attributes that can be specified for arguments to a vertex function and the corresponding data types with which they can be used.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>amplification_count</td>
<td>ushort or uint</td>
<td>The number of output vertices produced for each vertex instance.</td>
</tr>
<tr>
<td>amplification_id</td>
<td></td>
<td>The array index offset mappings for viewport and render target array indices, which enables routing an amplified vertex to a different viewport and render target.</td>
</tr>
<tr>
<td>base_instance</td>
<td></td>
<td>The base instance value added to each instance identifier before reading per-instance data.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>base_vertex</td>
<td>ushort or uint</td>
<td>The base vertex value added to each vertex identifier before reading per-vertex data.</td>
</tr>
<tr>
<td>instance_id</td>
<td>ushort or uint</td>
<td>The per-instance identifier, which includes the base instance value if one is specified.</td>
</tr>
<tr>
<td>vertex_id</td>
<td>ushort or uint</td>
<td>The per-vertex identifier, which includes the base vertex value if one is specified.</td>
</tr>
</tbody>
</table>

The default value for [[amplification_count]] is 1, which indicates that vertex amplification is disabled.

The value for [[amplification_id]] shall be in the range [0, amplification_count).

Notes on vertex function input attribute types:

- If the type for declaring [[vertex_id]] is uint, the type for declaring [[base_vertex]] needs to be uint or ushort.
- If the type for declaring [[vertex_id]] is ushort, the type for declaring [[base_vertex]] needs to be ushort.
- If the type for declaring [[instance_id]] is uint, the type for declaring [[base_instance]] needs to be uint or ushort.
- If the type for declaring [[instance_id]] is ushort, the type for declaring [[base_instance]] needs to be ushort.

5.2.3.2 Post-Tessellation Vertex Function Input Attributes

Table 5.3 lists the built-in attributes that can be specified for arguments to a post-tessellation vertex function and the corresponding data types with which they can be used.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>base_instance</td>
<td>ushort or uint</td>
<td>The base instance value added to each instance identifier before reading per-instance data.</td>
</tr>
<tr>
<td>instance_id</td>
<td>ushort or uint</td>
<td>The per-instance identifier, which includes the base instance value if one is specified.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>patch_id</td>
<td>ushort or uint</td>
<td>The patch identifier.</td>
</tr>
<tr>
<td>position_in_patch</td>
<td>float2 or float3</td>
<td>Defines the location on the patch being evaluated. For quad patches, must be float2. For triangle patches, must be float3.</td>
</tr>
</tbody>
</table>

All OS: All attributes in Table 5.3 are supported since Metal 1.2.

Notes on vertex function input attributes:

- If the type for declaring [[instance_id]] is uint, the type for declaring [[base_instance]] needs to be uint or ushort.
- If the type for declaring [[instance_id]] is ushort, the type for declaring [[base_instance]] needs to be ushort.

5.2.3.3 Vertex Function Output Attributes

Table 5.4 lists the built-in attributes that can be specified for a return type of a vertex function or the members of a structure that a vertex function returns (and their corresponding data types).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>clip_distance</td>
<td>float or float[n]</td>
<td>Distance from vertex to clipping plane</td>
</tr>
<tr>
<td>invariant</td>
<td>Not applicable; needs to be used with [[position]]</td>
<td>Marks the output position such that if the sequence of operations used to compute the output position in multiple vertex shaders is identical, there is a high likelihood that the resulting output position computed by these vertex shaders are the same value. Requires users to pass -fpreserve-invariance. Please see description below for more information.</td>
</tr>
<tr>
<td>point_size</td>
<td>float</td>
<td>Size of a point primitive</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>position</td>
<td>float4</td>
<td>The transformed vertex position</td>
</tr>
<tr>
<td>render_target_array_index</td>
<td>uchar, ushort, or uint</td>
<td>The array index that refers to one of:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1) an array slice of a texture array,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) data at a specified depth of a 3D texture,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) the face of a cubemap, or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4) a specified face of a specified array slice of a cubemap array.</td>
</tr>
<tr>
<td>shared</td>
<td>Not applicable</td>
<td>If present, then for every amplification_id, the output shall have the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>same value.</td>
</tr>
<tr>
<td>viewport_array_index</td>
<td>uchar, ushort, or uint</td>
<td>The viewport (and scissor rectangle) index value of the primitive.</td>
</tr>
</tbody>
</table>

All OS: Attributes in Table 5.4 are supported since Metal 1, unless otherwise indicated.

A cubemap is represented as a render target array with six layers, one for each face, and [[render_target_array_index]] is the face index, which is a value from 0 to 5. For a cubemap array, the [[render_target_array_index]] is computed as: array_slice_index * 6 + face_index.

You must return the same value of [[render_target_array_index]] for every vertex in a primitive. If values differ, the behavior and value passed to the fragment function are undefined. The same behavior applies to primitives generated by tessellation. If [[render_target_array_index]] is out-of-bounds (that is, greater than or equal to renderTargetArrayLength), the hardware interprets this value as 0. For more about [[render_target_array_index]] as fragment function input, see section 5.2.3.4.

[[viewport_array_index]] enables specifying one viewport and scissor rectangle from multiple active viewports and scissor rectangles. If the vertex function does not specify [[viewport_array_index]], the output viewport array index value is 0. For more about [[viewport_array_index]], see section 5.10.

[[invariant]] indicates that the floating-point math used in multiple function passes must generate a vertex position that matches exactly for every pass. [[invariant]] may only be used for a position in a vertex function (fields with the [[position]] attribute) to indicate the result of the calculation for the output is invariant. Compilers prior to iOS 14.0 and macOS 11.0, the calculation is likely (although not guaranteed) to be invariant. This calculation is now guaranteed to be invariant when passing -fpreserve-invariance option or setting the preserveInvariance on the MTLCompilerOptions from the Metal API for runtime compilation. Note that [[invariant]] is ignored if the options are not passed. This position invariance is essential for techniques such as shadow volumes or a z-prepass.
If the return type of a vertex function is not `void`, it must include the vertex position. If the vertex return type is `float4`, then it always refers to the vertex position, and the `[[position]]` attribute must not be specified. If the vertex return type is a structure, it must include an element declared with the `[[position]]` attribute.

The following example describes a vertex function called `process_vertex`. The function returns a user-defined structure called `VertexOutput`, which contains a built-in variable that represents the vertex position, so it requires the `[[position]]` attribute.

```cpp
struct VertexOutput {
    float4 position [[position]];
    float4 color;
    float2 texcoord;
}

vertex VertexOutput
process_vertex(...) {
    VertexOutput v_out;
    // compute per-vertex output
    ...
    return v_out;
}
```

Post-tessellation vertex function outputs are the same as a regular vertex function.

If vertex amplification is enabled, and if a vertex output variable has the same value for every `[[amplification_id]]` attribute, the vertex output is considered *shared*. A vertex output that is shared may use a single varying output slot, which is a limited resource. Vertex outputs that are not shared consume more than one varying output slot. (The Metal framework call `MTLRenderPipelineDescriptor maxVertexAmplificationCount` returns the number of varying slots that may be used to pass the amplified data to fragment function invocations, which impacts the number of total available varying slots.)

By default, all built-in vertex outputs are shared, except for those with the `[[position]]` attribute. By default, all other vertex outputs are not shared. To explicitly specify that the output is shared, use the `[[shared]]` attribute with a vertex output variable.

If the shader compiler can deduce that a vertex output variable has the same value for every `amplification_id`, the compiler may mark that vertex output as shared. The compiler may not mark vertex outputs as shared in any of these cases:

- The output value depends on the `[[amplification_id]]`.
- An atomic read-modify-write operation returns the output value.
- The shader loads the output value from volatile memory.

### 5.2.3.4 Fragment Function Input Attributes

Table 5.5 lists the built-in attributes that can be specified for arguments of a fragment function (and their corresponding data types).
If the return type of a vertex function is not `void`, it must include the vertex position. If the vertex return type is `float4`, this always refers to the vertex position (and the `[[position]]` attribute need not be specified). If the vertex return type is a structure, it must include an element declared with the `[[position]]` attribute.

### Table 5.5. Attributes for fragment function input arguments

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>amplification_count</td>
<td>ushort or uint</td>
<td>The number of output vertices produced for each vertex instance.</td>
</tr>
<tr>
<td>amplification_id</td>
<td>ushort or uint</td>
<td>The array index offset mappings for viewport and render target array indices, which enables routing an amplified vertex to a different viewport and render target.</td>
</tr>
<tr>
<td>barycentric_coord</td>
<td>float, float2, or float3</td>
<td>The barycentric coordinates.</td>
</tr>
<tr>
<td>color(m)</td>
<td>floatn, halfn, intn, uintn, shortn, or ushortn</td>
<td>The input value read from a color attachment. The index m indicates which color attachment to read from.</td>
</tr>
<tr>
<td>front-facing</td>
<td>bool</td>
<td>This value is <code>true</code> if the fragment belongs to a front-facing primitive.</td>
</tr>
<tr>
<td>point_coord</td>
<td>float2</td>
<td>Two-dimensional coordinates, which range from 0.0 to 1.0 across a point primitive, specifying the location of the current fragment within the point primitive.</td>
</tr>
<tr>
<td>position</td>
<td>float4</td>
<td>Describes the window-relative coordinate (x, y, z, 1/w) values for the fragment.</td>
</tr>
<tr>
<td>primitive_id</td>
<td>uint</td>
<td>The per-primitive identifier used with barycentric coordinates.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>render_target_array_index</td>
<td>uchar, ushort, or uint</td>
<td>The render target array index, which refers to the face of a cubemap, data at a specified depth of a 3D texture, an array slice of a texture array, an array slice, or face of a cubemap array. For a cubemap, the render target array index is the face index, which is a value from 0 to 5. For a cubemap array the render target array index is computed as: array slice index * 6 + face index.</td>
</tr>
<tr>
<td>sample_id</td>
<td>uint</td>
<td>The sample number of the sample currently being processed.</td>
</tr>
<tr>
<td>sample_mask</td>
<td>uint</td>
<td>The set of samples covered by the primitive generating the fragment during multisample rasterization.</td>
</tr>
<tr>
<td>sample_mask, post_depth_coverage</td>
<td>uint</td>
<td>The set of samples covered by the primitive generating the fragment after application of the early depth and stencil tests during multisample rasterization. The early_fragment_tests attribute needs to be used on the fragment function; otherwise the compilation fails.</td>
</tr>
<tr>
<td>thread_index_in_quadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a quad-group.</td>
</tr>
<tr>
<td>thread_index_in_simdgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a SIMD-group.</td>
</tr>
<tr>
<td>threads_per_simdgroup</td>
<td>ushort or uint</td>
<td>The thread execution width of a SIMD-group.</td>
</tr>
<tr>
<td>viewport_array_index</td>
<td>uint</td>
<td>The viewport (and scissor rectangle) index value of the primitive.</td>
</tr>
</tbody>
</table>

A variable declared with the `[[position]]` attribute as input to a fragment function can only be declared with the `center_no_perspective` sampling and interpolation attribute. (See section 5.4.)
For [[color(m)], m is used to specify the color attachment index when accessing (reading or writing) multiple color attachments in a fragment function.

The [[sample_mask]] attribute can only be declared once for a fragment function input.

The value of [[render_target_array_index]] in the fragment function is the same value written from the vertex function, even if the specified value is out of range.

For more about [[viewport_array_index]], see section 5.10.

The default value for [[amplification_count]] is 1, which indicates that vertex amplification is disabled.

The value for [[amplification_id]] shall be in the range [0, amplification_count).

For a specified [[amplification_id]] attribute value, the [[viewport_array_index]] and [[render_target_array_index]] built-in fragment input values are added to (offset by) the values that the corresponding MTLVertexAmplificationViewMapping structure provides.

The following example describes the structure MyVertexOut that is both a vertex function return type and a fragment function input type. MyVertexOut uses the [[amplification_id]] attribute for the input argument amp_id to amplify the position and ampData members. Use of the [[shared]] attribute explicitly ensures the texcoord member as having the same value for all varyings under vertex amplification, as described in section 5.2.3.3.

In the vertex function myVertex, the [[amplification_id]] and [[amplification_count]] attributes specify the vertex function input variables for vertex amplification, as detailed in section 5.2.3.1. The shader compiler deduces that the normal member has the same value for every [[amplification_id]], so the compiler marks it as shared in vertex output.

In the fragment function myFragment, the same [[amplification_id]] and [[amplification_count]] attributes specify fragment function input variables. If vertex amplification is enabled, then amp_id determines the mapping (MTLVertexAmplificationViewMapping structure) from which to select the viewport array index (viewportArrayIndexOffset member).

```cpp
struct MyVertexIn {
    float4 position [[attribute(0)]];
    float3 normal   [[attribute(1)]];
    float3 tangent  [[attribute(2)]];
    float2 texcoord [[attribute(3)]];
};

struct MyVertexOut {
    float4 position [[position]];
    float3 normal;
    float3 tangent;
    float3 bitangent;
    float2 texcoord [[shared]]; // explicitly shared.
    float  ampData;
```
ushort viewport [[viewport_array_index]]; // implicitly shared
}
constexpr ushort MAX_AMP = 2;

vertex MyVertexOut myVertex(MyVertexIn in [[stage_in]],
    constant float4x4 view_proj[MAX_AMP],
    constant float data[MAX_AMP],
    ushort amp_id [[amplification_id]],
    ushort amp_count [[amplification_count]], ...)
{
    MyVertexOut vert;
    vert.position  = view_proj[amp_id] * in.position; // deduced amplified
    vert.normal    = in.normal; // deduced shared
    vert.tangent   = ...;
    vert.bitangent = ...;
    vert.texcoord  = ...;
    vert.ampData   = data[amp_id]; // not shared
    vert.viewport  = 1;
    return vert;
}

fragment float4 myFragment(MyVertexOut in [[ stage_in ]],
    ushort amp_id [[amplification_id]],
    ushort amp_count [[amplification_count]], ...
) {
    // For MTLVertexAmplificationViewMapping = {{1,3},{2,4}}
    // when amp_id == 0, in.viewport == 2
    // when amp_id == 1, in.viewport == 3
    ushort viewport = in.viewport;
    ...
}

A fragment function input declared with the [[barycentric_coord]] attribute can only be declared with either the center_perspective (default) or center_no_perspective sampling and interpolation attributes. The barycentric coordinates and per-pixel primitive ID can be passed as fragment function input in structures organized as shown in these examples:

struct FragmentInput0 {
    uint primitive_id [[primitive_id]];
    // [[center_perspective]] is the default, so it can be omitted.
    float3 barycentric_coord [[barycentric_coord, center_perspective]];
};

struct FragmentInput1 {
    uint primitive_id [[primitive_id]];
    float2 linear_barycentric_coord [[barycentric_coord, center_no_perspective]];
};

By storing the barycentric coordinates and per-pixel primitive ID, your shader can manually read and interpolate the vertices of a drawn primitive within the fragment phase or defer this interpolation to a separate pass. In the deferred interpolation scenario, you can use a thin buffer

2023-06-02 | Copyright © 2023 Apple Inc. | All Rights Reserved.

Page 107 of 287
during the geometry pass to store a minimal set of surface data, including pre-clipped barycentric coordinates. At a later stage, you must have enough data to reconstruct the original vertex indices from the primitive ID data and to correlate the barycentric coordinates to those vertex indices.

When applying the `barycentric_coord` attribute to an input argument (or to a field of an argument) with more components than the dimension of the primitive, the remaining elements are initialized with 0.0f. For example, for

```c
fragment float4
frag (float3 coord [[barycentric_coord]]) { ... }
```

- When drawing a point, `coord.yz` is `float2(0.0f).`
- When drawing a line, `coord.z` is `0.0f`.

When applying the `barycentric_coord` attribute to an input argument (or to a field of an argument) with fewer components than the dimension of the primitive, the remaining elements are ignored.

Table 5.6 lists attributes that can be specified for tile arguments that are input to a fragment function. The data types for declaring `[[pixel_position_in_tile]]` and `[[pixels_per_tile]]` must match.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>pixel_position_in_tile</code></td>
<td>ushort2 or uint2</td>
<td>(x, y) position of the fragment in the tile.</td>
</tr>
<tr>
<td><code>pixels_per_tile</code></td>
<td>ushort2 or uint2</td>
<td>(width, height) of the tile in pixels.</td>
</tr>
<tr>
<td><code>tile_index</code></td>
<td>ushort or uint</td>
<td>1D tile index.</td>
</tr>
<tr>
<td><code>render_target_array_index</code></td>
<td>uchar, ushort,</td>
<td>The render target array index, which refers to the face of a cubemap, data at a specified depth of a 3D texture, an array slice of a texture array, an array slice, or face of a cubemap array. For a cubemap, the render target array index is the face index, which is a value from 0 to 5. For a cubemap array the render target array index is computed as: array slice index * 6 + face index.</td>
</tr>
</tbody>
</table>
macOS: attributes in Table 5.6 are supported since Metal 2.3.
iOS: attributes in Table 5.6 are supported since Metal 2.
[[tile_index]] is a value from [0, n), where n is the number of tiles in the render target.

5.2.3.5 Fragment Function Output Attributes
The return type of a fragment function describes the per-fragment output. You must use the attributes listed in Table 5.7 to specify that a fragment function can output one or more render-target color values, a depth value, a sampling coverage mask, or a stencil reference value. If the depth value is not output by the fragment function, the depth value generated by the rasterizer is output to the depth attachment.

**Table 5.7. Attributes for fragment function return types**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>color(m)</td>
<td>float, halfn, intn, uintn, shortn, or ushortn</td>
<td>Color value output for a color attachment. m is the color attachment index and needs to be known at compile time. The index i can be used to specify one or more colors output by a fragment function for a given color attachment and is an input to the blend equation.</td>
</tr>
<tr>
<td>color(m), index(i)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>depth(depth_argument)</td>
<td>float</td>
<td>Depth value output using the function specified by depth_argument.</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample_mask</td>
<td>uint</td>
<td>Coverage mask.</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stencil</td>
<td>uint</td>
<td>Stencil reference value to be used in a stencil test.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.1.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The color attachment index m for fragment output is specified in the same way as it is for [[color(m)]] for fragment input (see discussion for Table 5.5). Multiple elements in the fragment function return type that use the same color attachment index for blending needs to be declared with the same data type.

If there is only a single-color attachment in a fragment function, then [[color(m)]] is optional. If [[color(m)]] is not specified, the attachment index is 0. If multiple color attachments are specified, [[color(m)]] needs to be specified for all color values. See examples of specifying the color attachment in sections 5.5 and 5.8.1.5.

If index(i) is not specified in the attribute, the default is an index of 0. If index(i) is specified, the value of i needs to be known at compile time.

If a fragment function writes a depth value, the depth_argument needs to be specified with one of the following values:
any
greater
less
You cannot use the [[stencil]] attribute in fragment-based tile shading functions. The [[stencil]] attribute is not compatible with the [[early_fragment_tests]] function attribute.

If the fragment function does not output the stencil value, the setStencilReferenceValue: or setStencilFrontReferenceValue:backReferenceValue: method of MTLRenderCommandEncoder can set the stencil reference value.

The following example shows how color attachment indices can be specified. Color values written in clr_f write to color attachment index 0, clr_i to color attachment index 1, and clr_ui to color attachment index 2.

```swift
struct MyFragmentOutput {
    // color attachment 0
    float4 clr_f [[color(0)]]; // color attachment 1
    int4 clr_i [[color(1)]]; // color attachment 2
    uint4 clr_ui [[color(2)]];}

fragment MyFragmentOutput
my_fragment(...) {
    MyFragmentOutput f;
    ... f.clr_f = ...;
    ... return f;
}
```

If a color attachment index is used as both an input to and an output of a fragment function, the data types associated with the input argument and output declared with this color attachment index must match.

5.2.3.6 Kernel Function Input Attributes

When a kernel function is submitted for execution, it executes over an N-dimensional grid of threads, where N is one, two or three. A thread is an instance of the kernel function that
executes for each point in this grid, and thread_position_in_grid identifies its position in the grid.

Within a compute unit, a threadgroup is partitioned into multiple smaller groups for execution. The execution width of the compute unit, referred to as the threads_per_simdgroup, determines the recommended size of this smaller group. For best performance, make the total number of threads in the threadgroup a multiple of the threads_per_simdgroup.

Threadgroups are assigned a unique position within the grid (referred to as threadgroup_position_in_grid). Threads are assigned a unique position within a threadgroup (referred to as thread_position_in_threadgroup). The unique scalar index of a thread within a threadgroup is given by thread_index_in_threadgroup.

Each thread’s position in the grid and position in the threadgroup are N-dimensional tuples. Threadgroups are assigned a position using a similar approach to that used for threads. Threads are assigned to a threadgroup and given a position in the threadgroup with components in the range from zero to the size of the threadgroup size in that dimension minus one.

When a kernel function is submitted for execution, the number of threadgroups and the threadgroup size are specified, or the number of threads in the grid and the threadgroup size are specified. For example, consider a kernel function submitted for execution that uses a 2D grid where the number of threadgroups specified are \((Wx, Wy)\) and the threadgroup size is \((Sx, Sy)\). Let \((wx, wy)\) be the position of each threadgroup in the grid \((threadgroup_position_in_grid)\) and \((lx, ly)\) be the position of each thread in the threadgroup \((thread_position_in_threadgroup)\).

The thread position in the grid \((thread_position_in_grid)\) is:
\[
(gx, gy) = (wx * Sx + lx, wy * Sy + ly)
\]
The grid size \((threads_per_grid)\) is:
\[
(Gx, Gy) = (Wx * Sx, Wy * Sy)
\]
In cases other than a tile function, the thread index in the threadgroup \((thread_index_in_threadgroup)\) is determined by: \(ly * Sx + lx\)

For a tile function, the thread index is not a linear mapping from the \(lx\) and \(ly\) values. Each thread in a tile function is guaranteed to get a unique index in the range \([0, Sx * Sy)\).

Within a threadgroup, threads are divided into SIMD-groups in an implementation-defined fashion. Any given thread in a SIMD-group can query its SIMD lane ID and which SIMD-group it is a member of.

Table 5.8 lists the built-in attributes that can be specified for arguments to a kernel function and the corresponding data types with which they can be used. Starting in Metal 3.1, the built-in attributes can be specified on global (program scope) variables to be used in a kernel context.
Table 5.8. Attributes for kernel function input arguments

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dispatch_quadgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The quad-group execution width of a threadgroup specified at dispatch.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dispatch_simdgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The SIMD-group execution width of a threadgroup specified at dispatch.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dispatch_threads_per_threadgroup</td>
<td>ushort, ushort2,</td>
<td>The thread execution width of a threadgroup for threads specified</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td>ushort3, uint, uint2, or</td>
<td>at dispatch.</td>
</tr>
<tr>
<td></td>
<td>uint3</td>
<td></td>
</tr>
<tr>
<td>grid_origin</td>
<td>ushort, ushort2,</td>
<td>The origin (offset) of the grid over which compute threads that read</td>
</tr>
<tr>
<td>All OS: Since Metal 1.2.</td>
<td>ushort3, uint, uint2, or</td>
<td>per-thread stage-in data are launched.</td>
</tr>
<tr>
<td></td>
<td>uint3</td>
<td></td>
</tr>
<tr>
<td>grid_size</td>
<td>ushort, ushort2,</td>
<td>The maximum size of the grid over which compute threads that read</td>
</tr>
<tr>
<td>All OS: Since Metal 1.2.</td>
<td>ushort3, uint, uint2, or</td>
<td>per-thread stage-in data are launched.</td>
</tr>
<tr>
<td></td>
<td>uint3</td>
<td></td>
</tr>
<tr>
<td>quadgroup_index_in_threadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a quad-group within a threadgroup.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>quadgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The quad-group execution width of a threadgroup.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>simdgroup_index_in_threadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a SIMD-group within a threadgroup.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>simdgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The SIMD-group execution width of a threadgroup.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td><code>thread_execution_width</code></td>
<td>ushort or uint</td>
<td>The thread execution width of a SIMD-group (compute unit).</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td></td>
<td>[[ Deprecated as of Metal 3 – use <code>threads_per_simdgroup</code> ]]</td>
</tr>
<tr>
<td><code>thread_index_in_quadgroup</code></td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
<td>iOS: Since Metal 2.</td>
</tr>
<tr>
<td><code>thread_index_in_simdgroup</code></td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a SIMD-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.</td>
<td></td>
<td>iOS: Since Metal 2.2.</td>
</tr>
<tr>
<td><code>thread_index_in_threadgroup</code></td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a threadgroup.</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>thread_position_in_grid</code></td>
<td>ushort, ushort2,</td>
<td>The thread’s position in an N-dimensional grid of threads.</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td>ushort3, uint, uint2, or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uint3</td>
<td></td>
</tr>
<tr>
<td><code>thread_position_in_threadgroup</code></td>
<td>ushort, ushort2,</td>
<td>The thread’s unique position within a threadgroup</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td>ushort3, uint, uint2, or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uint3</td>
<td></td>
</tr>
<tr>
<td><code>threadgroup_position_in_grid</code></td>
<td>ushort, ushort2,</td>
<td>The threadgroup’s unique position within a grid.</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td>ushort3, uint, uint2, or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uint3</td>
<td></td>
</tr>
<tr>
<td><code>threadgroups_per_grid</code></td>
<td>ushort, ushort2,</td>
<td>The number of threadgroups in a grid.</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td>ushort3, uint, uint2, or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uint3</td>
<td></td>
</tr>
<tr>
<td><code>threads_per_grid</code></td>
<td>ushort, ushort2,</td>
<td>The grid size.</td>
</tr>
<tr>
<td>All OS: Since Metal 1.</td>
<td>ushort3, uint, uint2, or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uint3</td>
<td></td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>threads_per_simdgroup</td>
<td>ushort or uint</td>
<td>The thread execution width of a SIMD-group (compute unit).</td>
</tr>
<tr>
<td>threads_per_threadgroup</td>
<td>ushort, ushort2,</td>
<td>The thread execution width of a threadgroup.</td>
</tr>
<tr>
<td></td>
<td>ushort3, uint, uint2,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>or uint3</td>
<td></td>
</tr>
</tbody>
</table>

All OS: Since Metal 1.2, `grid_origin` and `grid_size` supported.

macOS: Since Metal 2, SIMD-group attributes supported. Since Metal 2.1, quad-group attributes supported. Other attributes supported since Metal 1.

iOS: Since Metal 2, quad-group attributes supported. Since Metal 2, SIMD-group attributes supported.

Other attributes supported since Metal 1.

All OS: Since Metal 3.1, global (program scope) variables can now specify these attributes except when they are used in a dynamic library or separately compiled binary function.

For standard Metal compute functions (other than tile functions), SIMD-groups are linear and one-dimensional. (Threadgroups may be multi-dimensional.) The number of SIMD-groups in a threadgroup ([[simdgroups_per_threadgroup]]) is the total number threads in the threadgroup ([[threads_per_threadgroup]]) divided by the SIMD-group size ([[threads_per_simdgroup]]):

\[
\text{simdgroups\_per\_threadgroup} = \text{ceil}(\text{threads\_per\_threadgroup}/\text{threads\_per\_simdgroup})
\]

Similarly, the number of quad-groups in a threadgroup (quadgroups_per_threadgroup) is the total number of threads in threadgroup divided by 4, which is the thread execution width of a quad-group:

\[
\text{quadgroups\_per\_threadgroup} = \text{ceil}(\text{threads\_per\_threadgroup}/4)
\]

For tile functions, threads are arranged as 2 x 2 quads. For a 2D grid where the number of threadgroups specified are \((W_x, W_y)\), simdgroups_per_threadgroup is computed by:

\[
\text{simdgroups\_per\_threadgroup} = \text{ceil}(W_x/2) \ast 2 \ast \text{ceil}(W_y/2) \ast 2 / \text{threads\_per\_simdgroup}
\]

\[
\text{simdgroups\_per\_threadgroup} = \text{ceil}(W_x/2) \ast \text{ceil}(W_y/2) \ast 4 / \text{threads\_per\_simdgroup}
\]

For tile functions, quadgroups_per_threadgroup is computed by:

\[
\text{quadgroups\_per\_threadgroup} = \text{ceil}(W_x/2) \ast 2 \ast \text{ceil}(W_y/2) \ast 2 / 4
\]

\[
\text{quadgroups\_per\_threadgroup} = \text{ceil}(W_x/2) \ast \text{ceil}(W_y/2)
\]
[[dispatch_simdgroups_per_threadgroup]] and
[[dispatch_quadgroups_per_threadgroup]] are similarly computed for threads
specified at dispatch.

SIMD-groups execute concurrently within a given threadgroup and make independent forward
progress with respect to each other, in the absence of threadgroup barrier operations. The
thread index in a SIMD-group (given by [[thread_index_in_simdgroup]]) is a value
between 0 and SIMD-group size – 1, inclusive. Similarly, the thread index in a quad-group
(given by [[thread_index_in_quadgroup]]) is a value between 0 and 3, inclusive.

In Metal 2, the number of threads in the grid does not have to be a multiple of the number of
threads in a threadgroup. It is therefore possible that the actual threadgroup size of a specific
threadgroup may be smaller than the threadgroup size specified in the dispatch. The
[[threads_per_threadgroup]] attribute specifies the actual threadgroup size for a given
threadgroup executing the kernel. The [[dispatch_threads_per_threadgroup]]
attribute is the threadgroup size specified at dispatch.

Notes on kernel function attributes:

- The type for declaring [[thread_position_in_grid]], [[threads_per_grid]],
  [[thread_position_in_threadgroup]], [[threads_per_threadgroup]],
  [[threadgroup_position_in_grid]],
  [[dispatch_threads_per_threadgroup]], and [[threadgroups_per_threadgroup]]
  needs to be a scalar type or a vector type. If it is a vector type, the number of components
  for the vector types for declaring these arguments need to match.
- The data types for declaring [[thread_position_in_grid]] and
  [[threads_per_grid]] need to match.
- The data types for declaring [[thread_position_in_threadgroup]],
  [[threads_per_threadgroup]], and
  [[dispatch_threads_per_threadgroup]] need to match.
- If [[thread_position_in_threadgroup]] is type uint, uint2 or uint3,
  [[thread_index_in_threadgroup]] needs to be type uint.
- The types for declaring [[thread_index_in_simdgroup]],
  [[threads_per_simdgroup]], [[simdgroup_index_in_threadgroup]],
  [[simdgroups_per_threadgroup]],
  [[dispatch_simdgroups_per_threadgroup]],
  [[quadgroup_index_in_threadgroup]], [[quadgroups_per_threadgroup]],
  and [[dispatch_quadgroups_per_threadgroup]] need to be ushort or uint.
  The types for declaring these built-in variables need to match.
- [[threads_per_simdgroup]] and [[thread_execution_width]] are aliases of
  one another that reference the same concept.

**Table 5.9. Attributes for kernel function tile input arguments**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>render_target_array_index</td>
<td>uchar, ushort, or uint</td>
<td>The render target array index, which refers to the face of a cubemap, data at a specified</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>depth of a 3D texture, an array slice of a texture array, an array slice, or face of a cubemap array. For a cubemap, the render target array index is the face index, which is a value from 0 to 5. For a cubemap array the render target array index is computed as: array slice index * 6 + face index.</td>
</tr>
</tbody>
</table>

macOS: attributes in Table 5.9 are supported since Metal 2.3.

iOS: attributes in Table 5.9 are supported since Metal 2.

### 5.2.3.7 Intersection Function Input Attributes

Table 5.10 lists the built-in attributes that can be specified for arguments to a custom intersection function (see section 5.1.6). Some built-in attributes can be used when specific values of `primitive_type` and `intersection_tags` are specified on the intersection function.

For example, `instance_id` is available if `intersection_tags` contains `instancing`.

```cpp
[[intersection(triangle, triangle_data, instancing, world_space_data)]]
bool triangleIntersectionFunction(..., uint id [[instance_id]], ...) {
...}
```

Any such restriction is listed in the description of the attribute.

### Table 5.10. Attributes for intersection function input arguments

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>origin</td>
<td>float3</td>
<td>Ray origin in object space.</td>
</tr>
<tr>
<td>direction</td>
<td>float3</td>
<td>Ray direction in object space.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>min_distance</td>
<td>float</td>
<td>Ray min distance.</td>
</tr>
<tr>
<td>max_distance</td>
<td>float</td>
<td>Passed by reference. Returns the current closest intersection max distance. The intersector initializes the initial value with the ray’s maximum distance and the value decreases as the intersector finds intersections.</td>
</tr>
<tr>
<td>payload</td>
<td>User type. Passed by reference.</td>
<td>User defined payload passed by the calling thread. Needs to be specified to allow matching payload table by intersect() (section 6.18.2).</td>
</tr>
<tr>
<td>geometry_id</td>
<td>ushort or uint</td>
<td>The per-geometry id.</td>
</tr>
<tr>
<td>primitive_id</td>
<td>ushort or uint</td>
<td>The per-primitive identifier. For curves, this is a curve segment index.</td>
</tr>
<tr>
<td>instance_id</td>
<td>ushort, uint or array_ref&lt;uint&gt;</td>
<td>The per-instance identifier. Available if intersection_tags include instancing. Since Metal 3.1, if intersection_tags include max_levels&lt;Count&gt;, the type must be array_ref&lt;uint&gt;. Otherwise, it is ushort or uint.</td>
</tr>
<tr>
<td>world_space_origin</td>
<td>float3</td>
<td>Origin in world space. Available if intersection_tags include world_space_data.</td>
</tr>
<tr>
<td>world_space_direction</td>
<td>float3</td>
<td>Direction in world space. Available if intersection_tags include world_space_data.</td>
</tr>
<tr>
<td>barycentric_coord</td>
<td>float2</td>
<td>The barycentric coordinates. Available if the primitive_type is triangle and intersection_tag include triangle_data.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>front_facing</td>
<td>bool</td>
<td>This value is true if the triangle front face is visible from the ray origin. Available if <code>intersection_tags</code> include <code>triangle_data</code>.</td>
</tr>
<tr>
<td>distance</td>
<td>float</td>
<td>Distance along the ray at the triangle intersection. Available if the <code>primitive_type</code> is <code>triangle</code>.</td>
</tr>
<tr>
<td>opaque</td>
<td>bool</td>
<td>If this primitive should be considered opaque or not. Available if the <code>primitive_type</code> is a <code>bounding_box</code>.</td>
</tr>
<tr>
<td>instance_intersection_function_table_offset</td>
<td>ushort or uint</td>
<td>Offset into the intersection function table used to select the intersection instance.</td>
</tr>
<tr>
<td>geometry_intersection_function_table_offset</td>
<td>ushort or uint</td>
<td>Offset into the geometry object used to select to select the intersection instance.</td>
</tr>
<tr>
<td>time</td>
<td>float</td>
<td>Ray intersection time. Available if <code>intersection_tags</code> include <code>primitive_motion</code>. All OS: Since Metal 2.4.</td>
</tr>
<tr>
<td>motion_start_time</td>
<td>float</td>
<td>Motion start time for this geometry. Available if <code>intersection_tags</code> include <code>primitive_motion</code>. All OS: Since Metal 2.4.</td>
</tr>
<tr>
<td>motion_end_time</td>
<td>float</td>
<td>Motion end time for this geometry. Available if <code>intersection_tags</code> include <code>primitive_motion</code>. All OS: Since Metal 2.4.</td>
</tr>
<tr>
<td>key_frame_count</td>
<td>ushort or uint</td>
<td>Number of key frames. Available if <code>intersection_tags</code> include <code>primitive_motion</code>. All OS: Since Metal 2.4.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------</td>
<td>--------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>object_to_world_transform</td>
<td>float4x3</td>
<td>Object space to world space transformation matrix. Available if intersection_tags include instancing and world_space_data. If intersection_tags include instance_motion, the matrix is interpolated based on the time.</td>
</tr>
<tr>
<td>world_to_object_transform</td>
<td>float4x3</td>
<td>World space to object space transformation matrix. Available if intersection_tags include instancing and world_space_data. If intersection_tags include instance_motion, the matrix is interpolated based on the time.</td>
</tr>
<tr>
<td>user_instance_id</td>
<td>ushort, uint or array_ref&lt;uint&gt;</td>
<td>User defined instance id. Available if intersection_tags include instancing. Since Metal 3.1, if intersection_tags include max_levels&lt;Count&gt;, the type must be array_ref&lt;uint&gt;. Otherwise, it is ushort or uint.</td>
</tr>
<tr>
<td>primitive_data</td>
<td>const device T* or const device T&amp;</td>
<td>Per-primitive data. The data is read-only and passed in the device address space.</td>
</tr>
<tr>
<td>curve_parameter</td>
<td>float</td>
<td>The value which would need to be passed to the curve basis functions to reconstruct the position corresponding to the intersection along the curve segment. This will be exactly 0.0F or 1.0F if, and only if, the ray intersects a curve end cap or elbow. Available if intersection_tags include curve_data. See section 6.18.6 for a set of curve utility functions.</td>
</tr>
</tbody>
</table>
For vertex attributes \( v_0, v_1, \) and \( v_2, \) the attribute value at the specified barycentric point is:
\[
v_1 \times barycentric\_coord.x + \\
v_2 \times barycentric\_coord.y + \\
v_0 \times (1.0f - (barycentric\_coord.x + barycentric\_coord.y))
\]

The type for a parameter with the \([\text{payload}]\) attribute is of the form \( \text{ray\_data\ T \&}. \) It is passed by reference to the intersection functions, and it is allocated in the \( \text{ray\_data\ address\ space} \). The type \( \text{T} \) of the payload can be or contain the following types:
- device or constant pointers or references
- integer types
- enum types
- floating point types
- vector types
- arrays of such types
- struct and union (except for \( \text{atomic<T'> and imageblock<T'>} \)).

### 5.2.3.8 Intersection Function Output Attributes

Table 5.11 lists the built-in attributes that can be specified for a return type of a \( [[\text{intersection}(\text{primitive\_type,\ intersection\_tags...})]] \) function (and their corresponding data types).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>accept_intersection</td>
<td>bool</td>
<td>If true, this primitive becomes the next committed hit: if it is the nearest, it will be returned from ( \text{intersect()} ).</td>
</tr>
<tr>
<td>continue_search</td>
<td>bool</td>
<td>If the hit is accepted (([\text{accept_intersection]} == \text{true} )), ( \text{continue_search\ indicates\ if\ the\ search\ should\ continue.\ If\ \text{continue_search\ is\ true,\ \text{intersect()}\ will\ continue\ to\ search\ for\ a\ closer\ hit.}\ If\ \text{false,\ no\ further\ searching\ is\ done.\ The\ current\ nearest\ hit\ is\ returned\ from\ \text{intersect()}).\ Defaults\ to\ \text{true.}\ Even\ if\ \text{true}\ is\ returned,\ a committed hit will immediately halt searching if ( \text{accept_any_intersection()}\ is\ \text{true}).}</td>
</tr>
<tr>
<td>distance</td>
<td>float</td>
<td>This returns the distance along the ray of a hit found within the bounding box. If the hit is rejected</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>([[accept_intersection]] == false), this return value is ignored. Available if the primitive_type is a bounding_box.</td>
</tr>
</tbody>
</table>

For triangle intersection functions, [[accept_intersection]] is the only required return value. If the function returns a bool without an attribute, then it is assumed to be [[accept_intersection]].

The value of [[distance]] needs to be greater than or equal to the value of [[min_distance]] and it needs to be less than or equal to the value of [[max_distance]] and within the custom primitive’s bounding box (inclusive), or the results are undefined. If the value of [[distance]] is the same as the value of [[max_distance]], then accepting this hit takes precedence over the previous hit at the same distance.

Any changes made to the ray payload take effect regardless of how the intersection function returns: Rejected primitives can have side effects to memory that are observed by future intersection shader threads.

Writes to device memory also occur even for rejected primitives. Those writes are visible to other threads via the usual memory consistency and coherency rules (at present, only atomics will be coherent, and only relaxed consistency is supported). Intersection functions may be invoked even if the ray does not intersect the primitive’s bounding box. For example, implementations may group multiple primitives into one acceleration structure leaf node.

Below is an example of an intersection function of a bounding box.

```cpp
struct IntersectionResult {
    bool continueSearch [[continue_search]];  // Renew a search if a hit isn’t found.
    bool accept [[accept_intersection]];     // TRUE = Accept this hit, accept previous hit if equal.
    float distance [[distance]];             // Distance from the origin to the hit.
};

[[intersection(bounding_box)]]
IntersectionResult sphereIntersectionFunction(
    float3 origin [[origin]],                  // Origin of ray.
    float3 direction [[direction]],           // Direction of ray.
    uint primitiveIndex [[primitive_id]],     // Primitive index.
    ray_data float2& resources [[payload]],   // Ray data.
    float min_distance [[min_distance]],      // Minimum distance from origin.
    float max_distance [[max_distance]])     // Maximum distance from origin.
{
    ...}
```

### 5.2.3.9 Object Function Input Attributes

All OS: Since Metal 3
All OS: Since Metal 3.1, global variables can now specify these attributes except when used in a dynamic library or separately compiled binary function.

Object functions use the same execution model as a kernel function (see section 5.2.3.6), where it executes over an N-dimensional grid of threads. Object functions arguments can be samplers, textures, arguments of type mesh_grid_properties, and buffers in the device, constant, and threadgroup address space.

Object functions support a subset of the built-in attributes of a kernel function and [[amplification_count]] and [[payload]]. The semantics of [[amplification_count]] is the same as in section 5.2.3.1 Vertex Function Input Attributes. Table 5.12 lists the built-in attributes that can be specified for arguments to an object function and the corresponding data types with which they can be used. Starting in Metal 3.1, the built-in attributes in Table 5.12 can be specified on program scope variables except for amplification_count and payload.

### Table 5.12. Attributes for Object Function

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>amplification_count</td>
<td>ushort or uint</td>
<td>The number of output vertices produced for each vertex instance.</td>
</tr>
<tr>
<td>dispatch_quadgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The quad-group execution width of a threadgroup specified at dispatch.</td>
</tr>
<tr>
<td>dispatch_simdgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The SIMD-group execution width of a threadgroup specified at dispatch.</td>
</tr>
<tr>
<td>dispatch_threads_per_threadgroup</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The thread execution width of a threadgroup for threads specified at dispatch.</td>
</tr>
<tr>
<td>payload</td>
<td>Pointer or I-value reference to User Defined T in object_data address space</td>
<td>The payload is data passed to the mesh shader from the object shader. The payload pointer or reference is the same for all threads in the threadgroup. The payload memory is assumed uninitialized at the entry of the object function.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>quadgroup_index_in_threadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a quad-group within a threadgroup.</td>
</tr>
<tr>
<td>quadgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The quad-group execution width of a threadgroup.</td>
</tr>
<tr>
<td>simdgroup_index_in_threadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a SIMD-group within a threadgroup.</td>
</tr>
<tr>
<td>simdgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The SIMD-group execution width of a threadgroup.</td>
</tr>
<tr>
<td>thread_index_in_quadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a quad-group.</td>
</tr>
<tr>
<td>thread_index_in_simdgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a SIMD-group.</td>
</tr>
<tr>
<td>thread_index_in_threadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a threadgroup.</td>
</tr>
<tr>
<td>thread_position_in_grid</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The thread’s position in an N-dimensional grid of threads.</td>
</tr>
<tr>
<td>thread_position_in_threadgroup</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The thread’s unique position within a threadgroup</td>
</tr>
<tr>
<td>threadgroup_position_in_grid</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The threadgroup’s unique position within a grid.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>threadgroups_per_grid</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The number of threadgroups in a grid.</td>
</tr>
<tr>
<td>threads_per_grid</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The grid size.</td>
</tr>
<tr>
<td>threads_per_simdgroup</td>
<td>ushort or uint</td>
<td>The thread execution width of a SIMD-group.</td>
</tr>
<tr>
<td>threads_per_threadgroup</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The thread execution width of a threadgroup.</td>
</tr>
</tbody>
</table>

Object function attributes have the same restrictions as kernel function attributes:

- **The type for declaring** `[[thread_position_in_grid]], [[threads_per_grid]], [[thread_position_in_threadgroup]], [[threads_per_threadgroup]], [[threadgroup_position_in_grid]], [[dispatch_threads_per_threadgroup]], and `[[threadgroups_per_grid]]` needs to be a scalar type or a vector type. If it's a vector type, the number of components for the vector types for declaring these arguments need to match.

- **The data types for declaring** `[[thread_position_in_grid]]` and `[[threads_per_grid]]` need to match.

- **The data types for declaring** `[[thread_position_in_threadgroup]], [[threads_per_threadgroup]], and [[dispatch_threads_per_threadgroup]]` need to match.

- **If** `[[thread_position_in_threadgroup]]` **is type** `uint, uint2` or `uint3`, `[[thread_index_in_threadgroup]]` **needs to be** `type` `uint`.

- **The types for declaring** `[[thread_index_in_simdgroup]], [[threads_per_simdgroup]], [[simdgroupposition_in_threadgroup]], [[simdgroups_per_threadgroup]], [[dispatch_simdgroups_per_threadgroup]], [[quadgroup_index_in_threadgroup]], [[quadgroups_per_threadgroup]], and [[dispatch_quadgroups_per_threadgroup]]` **need to be** `ushort` or `uint`. **The types for declaring** these built-in variables need to match.
5.2.3.10 Mesh Function Input Attributes

All OS: Since Metal 3

All OS: Since Metal 3.1, global variables can now specify these attributes except when they are used in a dynamic library or separately compiled binary function.

Mesh functions use the same execution model as a kernel function (see section 5.2.3.6), where it executes over an N-dimensional grid of threads. Mesh functions arguments can be from samplers, textures, arguments of type mesh<V, P, NV, NP, t>, and buffers of device and constant. If the mesh function has a mesh<V, P, NV, NP, t> argument, it points to an opaque handle for memory representing the mesh to export. The underlying memory referenced by the mesh<V, P, NV, NP, t> argument is shared among threads of a given threadgroup.

Mesh functions support a subset of the built-in attributes of a kernel function and also [[amplification_count]], [[amplification_id]], and [[payload]] attributes. The semantics of [[amplification_count]] and [[amplification_id]] is the same as in section 5.2.3.1 Vertex Function Input Attributes. Table 5.13 lists the built-in attributes that can be specified for arguments to a mesh function and the corresponding data types with which they can be used. Starting in Metal 3.1, the built-in attributes in Table 5.13 can be specified on program scope variables except for amplification_count, amplification_id, and payload.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Corresponding Data Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>amplification_count</td>
<td>ushort or uint</td>
<td>The number of output vertices produced for each primitive instance.</td>
</tr>
<tr>
<td>amplification_id</td>
<td>ushort or uint</td>
<td>The array index offset mappings for viewport and render target array indices, which enables routing an amplified vertex to a different viewport and render target.</td>
</tr>
<tr>
<td>dispatch_quadgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The quad-group execution width of a threadgroup specified at dispatch.</td>
</tr>
<tr>
<td>dispatch_simdgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The SIMD-group execution width of a threadgroup specified at dispatch.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>dispatch_threads_per_threadgroup</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The thread execution width of a threadgroup for threads specified at dispatch.</td>
</tr>
<tr>
<td>payload</td>
<td>Pointer or l-value reference to User Defined T in object_data address space. Needs to be const qualified.</td>
<td>The payload is data passed to the mesh shader from the object shader. The payload pointer or reference is the same for all threads in the mesh grid. The payload memory is read-only in the mesh function.</td>
</tr>
<tr>
<td>quadgroup_index_in_threadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a quad-group within a threadgroup.</td>
</tr>
<tr>
<td>quadgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The quad-group execution width of a threadgroup.</td>
</tr>
<tr>
<td>simdgroup_index_in_threadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a SIMD-group within a threadgroup.</td>
</tr>
<tr>
<td>simdgroups_per_threadgroup</td>
<td>ushort or uint</td>
<td>The SIMD-group execution width of a threadgroup.</td>
</tr>
<tr>
<td>thread_index_in_quadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a quad-group.</td>
</tr>
<tr>
<td>thread_index_in_simdgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a SIMD-group.</td>
</tr>
<tr>
<td>thread_index_in_threadgroup</td>
<td>ushort or uint</td>
<td>The scalar index of a thread within a threadgroup.</td>
</tr>
<tr>
<td>thread_position_in_grid</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The thread’s position in an N-dimensional grid of threads.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Corresponding Data Types</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>thread_position_in_threadgroup</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The thread’s unique position within a threadgroup</td>
</tr>
<tr>
<td>threadgroup_position_in_grid</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The threadgroup’s unique position within a grid.</td>
</tr>
<tr>
<td>threadgroups_per_grid</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The number of threadgroups in a grid.</td>
</tr>
<tr>
<td>threads_per_grid</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The grid size.</td>
</tr>
<tr>
<td>threads_per_simdgroup</td>
<td>ushort or uint</td>
<td>The thread execution width of a SIMD-group.</td>
</tr>
<tr>
<td>threads_per_threadgroup</td>
<td>ushort, ushort2, ushort3, uint, uint2, or uint3</td>
<td>The thread execution width of a threadgroup.</td>
</tr>
</tbody>
</table>

Mesh function attributes have the same restrictions as kernel function attributes:

- **The type for declaring** `[[thread_position_in_grid]], [[threads_per_grid]], [[thread_position_in_threadgroup]], [[threads_per_threadgroup]], [[threadgroup_position_in_grid]], [[dispatch_threads_per_threadgroup]], and `[[threadgroups_per_grid]] needs to be a scalar type or a vector type. If it’s a vector type, the number of components for the vector types for declaring these arguments need to match.
- **The data types for declaring** `[[thread_position_in_grid]]` and `[[threads_per_grid]]` need to match.
- **The data types for declaring** `[[thread_position_in_threadgroup]], [[threads_per_threadgroup]], and` `[[dispatch_threads_per_threadgroup]] need to match.
5.2.4 Input Assembly Attribute

Vertex function output and the rasterizer-generated fragments become the per-fragment inputs to a fragment function. The [[stage_in]] attribute can assemble the per-fragment inputs.

A vertex function can read per-vertex inputs by indexing into buffer(s) passed as arguments to the vertex function using the vertex and instance IDs. To assemble per-vertex inputs and pass them as arguments to a vertex function, declare the inputs with the [[stage_in]] attribute.

A kernel function reads per-thread inputs by indexing into buffer(s) or texture(s) passed as arguments to the kernel function using the thread position in grid or thread position in threadgroup IDs. In addition, to pass per-thread inputs as arguments to a kernel function, declaring the inputs with the [[stage_in]] attribute.

You can declare only one argument of the vertex, fragment, or kernel function with the [[stage_in]] attribute. For a user-defined structure declared with the [[stage_in]] attribute, the members of the structure can be:

- A scalar integer or floating-point value.
- A vector of integer or floating-point values.
- An interpolant<T,P> value for fragment function input.

You cannot use the stage_in attribute to declare members of the structure that are packed vectors, matrices, structures, bitfields, references or pointers to a type, or arrays of scalars, vectors, or matrices.

5.2.4.1 Vertex Function Output Example

The following example shows how to pass per-vertex inputs using the stage_in attribute:

```c
struct VertexOutput {
    float4 position [[position]];
    float4 color;
    float2 texcoord;
};

struct VertexInput {
    float4 position [[attribute(0)]];
    float3 normal   [[attribute(1)]];
    half4 color     [[attribute(2)]];
};
```
half2 texcoord [[attribute(3)]];
};

constexpr constant uint MAX_LIGHTS = 4;

struct LightDesc {
    uint num_lights;
    float4 light_position[MAX_LIGHTS];
    float4 light_color[MAX_LIGHTS];
    float4 light_attenuation_factors[MAX_LIGHTS];
};

constexpr sampler s = sampler(coord::normalized,
    address::clamp_to_zero,
    filter::linear);

vertex VertexOutput
render_vertex(VertexInput v_in [[stage_in]],
    constant float4x4& mvp_matrix [[buffer(1)]],
    constant LightDesc& lights [[buffer(2)]],
    uint v_id [[vertex_id]])
{
    VertexOutput v_out;
    v_out.position = v_in.position * mvp_matrix;
    v_out.color = do_lighting(v_in.position, v_in.normal, lights);
    ... return v_out;
}

5.2.4.2 Fragment Function Input Example

An example in section 5.2.3.3 previously introduces the process_vertex vertex function, which returns a VertexOutput structure per vertex. In the following example, the output from process_vertex is pipelined to become input for a fragment function called render_pixel, so the first argument of the fragment function uses the [[stage_in]] attribute and uses the incoming VertexOutput type. (In render_pixel, the imgA and imgB 2D textures call the built-in function sample, which is introduced in section 6.12.3).

struct VertexOutput2 {
    float4 position [[position]];
    float4 color;
    float2 texcoord;
};

struct VertexInputData {
    float4 position;
    float3 normal;
    float2 texcoord;
};
constexpr constant uint MAX_LIGHTS = 4;

struct LightDesc {
    uint num_lights;
    float4 light_position[MAX_LIGHTS];
    float4 light_color[MAX_LIGHTS];
    float4 light_attenuation_factors[MAX_LIGHTS];
};

constexpr sampler s = sampler(coord::normalized,
    address::clamp_to_edge,
    filter::linear);

vertex VertexOutput2
render_vertex(const device VertexInputData *v_in [[buffer(0)]],
    constant float4x4& mvp_matrix [[buffer(1)]],
    constant LightDesc& lights [[buffer(2)]],
    uint v_id [[vertex_id]])
{
    VertexOutput v_out;
    v_out.position = v_in[v_id].position * mvp_matrix;
    v_out.color = do_lighting(v_in[v_id].position,
        v_in[v_id].normal, lights);
    v_out.texcoord = v_in[v_id].texcoord;
    return v_out;
}

fragment float4
render_pixel(VertexOutput2 input [[stage_in]],
    texture2d<float> imgA [[texture(0)]],
    texture2d<float> imgB [[texture(1)]])
{
    float4 tex_clr0 = imgA.sample(s, input.texcoord);
    float4 tex_clr1 = imgB.sample(s, input.texcoord);

    // Compute color.
    float4 clr = compute_color(tex_clr0, tex_clr1, ...);
    return clr;
}

5.2.4.3 Kernel Function Per-Thread Input Example

The following example shows how to use the stage_in attribute to pass per-thread inputs. The stage_in attribute in a kernel function allows you to decouple the data type for declaring the per-thread inputs in the function from the actual data type used to store the per-thread inputs.

struct PerThreadInput {
float4 a [[attribute(0)]];  
float3 b [[attribute(1)]];  
half4 c [[attribute(2)]];  
half2 d [[attribute(3)]];  
};

kernel void
my_kernel(PerThreadInput thread_input [[stage_in]],
           ...
           uint t_id [[thread_position_in_grid]])
{
...
}

5.3 Storage Class Specifiers

Metal supports the static and extern storage class specifiers. Metal does not support the thread_local storage class specifiers.

You can only use the extern storage-class specifier for functions and variables declared in program scope or for variables declared inside a function. The static storage-class specifier is only for device variables declared in program scope (see section 4.2) and is not for variables declared inside a graphics or kernel function. The following example incorrectly uses the static specifier for the variables b and c declared inside a kernel function.

extern constant float4 noise_table[256];
static constant float4 color_table[256] = {...}; // Here, static is ok.

extern void my_foo(texture2d<float> img);
extern void my_bar(device float *a);

[[kernel]] void
my_kernel(texture2d<float> img [[texture(0)]],
           device float *ptr [[buffer(0)]])
{
    extern constant float4 a;
    static constant float4 b;  // Here, static is an error.
    static float c;            // Here, static is an error.

    ...
    my_foo(img);
    ...
    my_bar(ptr);
    ...
}

5.4 Sampling and Interpolation Attributes

Sampling and interpolation attributes are used with inputs to fragment functions declared with the stage_in attribute except for members of type interpolant<T,P>. The attribute
determines what sampling method the fragment function uses and how the interpolation is performed, including whether to use perspective-correct interpolation, linear interpolation, or no interpolation.

The sampling and interpolation attribute can be specified on any `stage_in` structure member whose type is scalar and vector. The sampling and interpolation attributes supported are:

- `center_perspective`
- `center_no_perspective`
- `centroid_perspective`
- `centroid_no_perspective`
- `sample_perspective`
- `sample_no_perspective`
- `flat`

`center_perspective` is the default sampling and interpolation attribute, with the following exceptions:

- For a variable with the `[[position]]` attribute, the only valid sampling and interpolation attribute is `center_no_perspective`.
- For an integer variable, the only valid sampling and interpolation attribute is `flat`.

A perspective attribute (`center_perspective`, `centroid_perspective`, or `sample_perspective`) indicates the values across a primitive are interpolated in a perspective-correct manner. A nonperspective attribute (`center_no_perspective`, `centroid_no_perspective`, or `sample_no_perspective`) indicates the values across a primitive are linearly interpolated in screen coordinates.

The center attribute variants (`center_perspective` and `center_no_perspective`) cause sampling to use the center of each pixel.

The sampling attribute variants (`sample_perspective` and `sample_no_perspective`) cause interpolation at a sample location rather than at the pixel center. With one of these attributes, the fragment function (or code blocks in the fragment function) that use these variables execute per-sample rather than per-fragment.

If a centroid attribute variant is specified (`centroid_perspective` and `centroid_no_perspective`), the interpolation point sampled needs to be within both the primitive and the centroid of the pixel.

The following example demonstrates how to specify the interpolation of data for different members of a user-defined structure:

```cpp
struct FragmentInput {
    float4 pos [[center_no_perspective]];
    float4 color [[center_perspective]];
    float2 texcoord;
    int index [[flat]];
    float f [[sample_perspective]];
    interpolant<float4, interpolation::perspective> icolor;
};
```
In Metal 2.4 or later, the sample and interpolation attribute can also be specified on any `stage_in` structure member whose type is structure. All the members in the structure inherit the specified sampling and interpolation qualifiers. Field declarations in a structure where sampling and interpolation qualifiers have been inherited are valid only if one of the following is true:

- The type of field is compatible with the inherited qualifiers.
- The field declaration does not have a sampling, and interpolation qualifiers attribute.
- The field declaration has the same sampling, and interpolation qualifiers attribute as the inherited one.

The following example demonstrates how to specify the interpolation on structure types.

```metal
struct VOut {
    float4 pos [[position]];
}

struct POut {
    float4 color0;
    float4 color1;
};

[[mesh]] void mesh_function(mesh<VOut, POut, 3, 1,
                             topology::triangle> m)

struct FragmentInput {
    VOut    vin;
    POut    pin [[center _perspective]];
};
```

### 5.5 Per-Fragment Function versus Per-Sample Function

The fragment function is typically executed per-fragment. The sampling attribute identifies if fragment input interpolation will be per-sample or per-fragment. Similarly, the `[[sample_id]]` attribute identifies the current sample index, and the `[[color(m)]]` attribute identifies the destination fragment color or sample color (for a multisampled color attachment) value. If you use any of these attributes with arguments to a fragment function, the fragment function may execute per-sample instead of per-pixel. (The implementation may decide to only execute the code that depends on the per-sample values to execute per-sample and the rest of the fragment function may execute per-fragment.)

Only the inputs with `sample` access specified (or declared with the `[[sample_id]]` or `[[color(m)]]` attribute) differ between invocations per-fragment or per-sample, whereas other inputs still interpolate at the pixel center.

The following example uses the `[[color(m)]]` attribute to specify that this fragment function executes on a per-sample basis:
[[fragment]] float4
my_fragment(float2 tex_coord [[stage_in]],
texture2d<float> img [[texture(0)]],
sampler s [[sampler(0)]],
float4 framebuffer [[color(0)]]
{
    return c = mix(img.sample(s, tex_coord), framebuffer, mix_factor);
}

5.6 Imageblock Attributes

iOS: Support for imageblock since Metal 2.
macOS: Support for imageblock since Metal 2.3.

This section and its subsections describe several attributes for imageblocks, including the
[[imageblock_data(type)]] attribute that specifies input and output imageblock with an
explicit imageblock layout for a fragment function.

5.6.1 Matching Data Members of Master and View Imageblocks

You can use the [[user(name)]] attribute to specify an attribute name for a data member of
the imageblock data type for a fragment function. If the imageblock structure specified in a
fragment function is a subset of the master explicit imageblock structure, the following rules
match data members declared in the imageblock structure used in a fragment function with
corresponding data members declared in the master explicit imageblock structure:

- Every attribute name given by [[user(name)]] needs to be unique for each data
  member in the imageblock.
- The attribute name given by [[user(name)]] for a data member needs to match with
  a data member declared in the master explicit imageblock structure, and their associated
data types needs to also match.
- If the [[user(name)]] attribute is not specified, the data member name and type
  declared in the imageblock data type for a fragment function and the master imageblock
  structure needs to match. Additionally, the data member cannot be within a nested
  structure that is either within the view imageblock structure or within the master
  imageblock structure.

The following example shows the [[user(name)]] attribute in declarations of data members
in master and view imageblock structures:

// The explicit layout imageblock data master structure.
struct IM {
    rgba8unorm<half4> a [[user(my_a), raster_order_group(0)]];  
    rgb9e5<float4> b [[user(my_b), raster_order_group(0)]];    
    int c [[user(my_c), raster_order_group(0)]];             
    float d [[user(my_d), raster_order_group(0)]];             
}
};

// The explicit layout imageblock data view structure for input.
struct IVIn {
    rgb9e5<float4> x [[user(my_b)]]; // Maps to IM::b
    float y [[user(my_d)]]; // Maps to IM::d
};

// The explicit layout imageblock data view structure for output.
struct IVOut {
    int z [[user(my_c)]]; // Maps to IM::c
};

// The fragment return structure.
struct FragOut {
    // IVOut is a view of the master IM.
    IVOut i [[imageblock_data(IM)]];
};

// IVIn is a view of the master IM.
[[fragment]] FragOut
my_fragment(IVIn i [[imageblock_data(IM)]], ...) {
    FragOut fragOut;
    ... = i.x;
    ... = i.y;
    fragOut.i.z = ...
    return fragOut;
}

The following example shows the declaration of data members in master and view imageblock structures without the [[user(name)]] attribute:

struct IM {
    rgba8unorm<half4> a [[raster_order_group(0)]];
    rgb9e5<float4> b [[raster_order_group(0)]];
    int c [[raster_order_group(0)]];
    float d [[raster_order_group(0)]];
};

struct IVIn {
    rgb9e5<float4> b; // Maps to IM::b
    float d; // Maps to IM::d
};

struct IVOut {
    int c; // Maps to IM::c
};

struct FragOut {
    IVOut i [[imageblock_data(IM)]];
}
You can declare nested structures in the master imageblock and view imageblock structures. The following example shows how to use nested structures in an imageblock with data members declared with the `[[user(name)]]` attribute:

```cpp
struct A {
    rgba8unorm<half4> a [[user(A_a)]];
    rgb9e5<float4> b [[user(A_b)]];
};

struct B {
    int a [[user(B_a), raster_order_group(1)]];
    float b [[user(B_b), raster_order_group(2)]];
};

struct IM {
    A a [[user(A), raster_order_group(0)]];
    B b [[user(B)]];
};

struct IVIn {
    A x [[user(A)]]; // Maps to IM::a
};

struct IVOut {
    B y [[user(B)]]; // Maps to IM::b
    rgb9e5<float4> z [[user(A_b)]]; // Maps to IM::A::b
};

struct FragOut {
    IVOut i [[imageblock_data(IM)]];
};

fragment FragOut
my_fragment(IVIn i [[imageblock_data(IM)]], ...) {
    FragOut fragOut;
    ... = i.b;
    ... = i.d;
    fragOut.i.c = ...;
    return fragOut;
}
```
fragOut.i.z = ...
return fragOut;

Each field of a view structure must correspond to exactly one master structure field. A master structure field can refer to a top-level structure field as well as a field within a nested structure. It is illegal for two or more view structure fields to alias the same master structure field.

Example of Illegal Use:

```c
struct M {
    struct A {
        int a [[user(x)]];  
    }
    b [[user(y), raster_order_group(0)]];
};

struct V {
    int a [[user(x)]];  
    M::A b [[user(y)]];  // Illegal: b aliases with a
};

fragment void
f(V i [[imageblock_data(M)]])
{
...
```

Explicit imageblock types cannot have data members declared with the [[color(n)]] attribute.

### 5.6.2 Imageblocks and Raster Order Groups

In a kernel function, a [[raster_order_group(index)]] attribute specified on data members of an imageblock is ignored.

In a fragment function, you must specify the [[raster_order_group(index)]] attribute for data members of the master explicit imageblock data structure.

If the master explicit imageblock structure contains data members that are structures, you can specify the [[raster_order_group(index)]] attribute for all data members in the nested structure or just the nested structure. If you specify the [[raster_order_group(index)]] attribute for the nested structure, then it applies to all data members of the nested structure, and no data member in the nested structure can have the [[raster_order_group(index)]] attribute declared.

You optionally may specify the [[raster_order_group(index)]] attribute for data members of an imageblock view structure, but the [[raster_order_group(index)]] must match the same [[raster_order_group(index)]] specified on the data member of the master explicit imageblock structure.
The following example shows how you can specify the `[[raster_order_group(index)]]` attribute for data members of a master imageblock. Since the `[[raster_order_group(index)]]` attribute specifies the `S` structure member of the `gBufferData` structure, you cannot use this attribute on any members of the `S` structure.

```c
struct S {
    rgb9e5<half3> normal;
    float factor;
};

struct gBufferData {
    half3 color [[raster_order_group(0)]];  // This causes a compilation error.
    S s [[raster_order_group(1)]];  // This causes a compilation error.
    rgb11b10f<half3> lighting [[raster_order_group(2)]];  // This causes a compilation error.
};
```

Data members declared as an array have a single raster order group associated with all members of the array. The following example shows how you can specify the `[[raster_order_group(index)]]` attribute for a data member of a master imageblock that is an array of a structure type.

```c
struct S {
    rgb9e5<half3> normal;
    float factor;
};

struct IM {
    half3 color [[raster_order_group(0)]];  // This causes a compilation error.
    S s [[raster_order_group(1)][2];  // This causes a compilation error.
    rgb11b10f<half3> lighting [[raster_order_group(2)]];  // This causes a compilation error.
};
```

The following example shows an incorrect use of the `[[raster_order_group(index)]]` attribute where data member `s` is an array of a structure of type `S` with members that specify raster order groups that result in a compilation error.

```c
struct S {
    rgb9e5<half3> normal [[raster_order_group(0)]];  // This causes a compilation error.
    float factor [[raster_order_group(1)]];  // This causes a compilation error.
};

struct IM {
    half3 color [[raster_order_group(0)]];  // This causes a compilation error.
    S s[2];  // This causes a compilation error.
    rgb11b10f<half3> lighting [[raster_order_group(2)]];  // This causes a compilation error.
};
```
5.6.3 Imageblock Layouts for Fragment Functions

In a fragment function, you can access the imageblock in two ways:

- As a color attachment, where the storage layout of the imageblock is not known in the fragment function. An *implicit imageblock layout* uses the existing color attachment attribute. (For more about the implicit imageblock layout, see section 5.6.3.1.)

- As a structure for declaring the imageblock data where the fragment function explicitly specifies the storage layout of the imageblock. (For more about the *explicit imageblock layout*, see section 5.6.3.2.)

5.6.3.1 Implicit Imageblock Layout for Fragment Functions

You can access the imageblock data (all the data members in the imageblock associated with a pixel) in a fragment function. Metal creates an implicit imageblock that matches the behavior of color attachments (for input to and output from a fragment function). In this mode, the types associated with the color attachments, as described in the fragment function, are the ALU types (that is, the types used to perform computations in the fragment function). The Metal runtime defines the actual pixel storage format.

When accessing the imageblock data as color attachments, you cannot declare the pixel storage types described in section 2.7 in the imageblock slice structure.

For an imageblock data implicit layout of type \( T \), \( T \) is a structure where each member satisfies one of the following:

- Have a color attachment (see the \([[[\text{color}(m)]]]\) attribute in Table 5.5 of section 5.2.3.4). The color index \( m \) needs to be unique for each member (and sub-member) of \( T \).

- Be a structure type with members that satisfy the constraint on the list.

5.6.3.2 Explicit Imageblock Layout for Fragment Functions

The imageblock data with *explicit* layout has its layout declared in the shading function, not via the runtime as is done for color attachments. You declare the imageblock data for an explicit layout as a structure. Each data member of the per-fragment imageblock data can be:

- a scalar or vector, integer or floating-point data type,
- one of the pixel data types described in section 2.7,
- an array of these types,
- or a structure built with these types.

The data members of the imageblock structure use the appropriate alignment rules for each data member type declared in the structure to determine the actual structure layout and size.

A fragment function can read one or more data members in the per-fragment imageblock data and write to one or more data members in the per-fragment imageblock data. You can declare the input and output imageblock data to a fragment function as a structure. The input and output imageblock structures can be the fully explicit imageblock structure (referred to as the master explicit imageblock structure), or be a subset of the master explicit imageblock structure (referred to as the imageblock view structure). For the latter, use the
[[imageblock_data(type)]] attribute with the input and output imageblock data structure specified on a fragment function, where type specifies the fully explicit imageblock data structure.

If you specify the [[imageblock_data]] attribute on the input argument or output structure element without type, by default the fragment function uses the master explicit imageblock data structure on the input or output.

Example:

```
struct I {
    float a [[raster_order_group(0)]]; // implicit imageblock data
};

struct FragOut {
    float c [[color(0)]]; // explicit imageblock data
    I i [[imageblock_data]]; // explicit imageblock data
};

fragment FragOut
    my_fragment(I i [[imageblock_data]])
    {
        FragOut fragOut;
        ...
        return fragOut;
    }
```

Fragment functions can access both an implicit imageblock and an explicit imageblock as separate input arguments, or as fields in a return structure.

Example:

```
struct I {
    float a [[raster_order_group(0)]]; // implicit imageblock data
};

struct FragOut {
    float c [[color(0)]]; // explicit imageblock data
    I i [[imageblock_data]]; // explicit imageblock data
};

[[fragment]] FragOut
    my_fragment(I i [[imageblock_data]], float c [[color(0)]])
    {
        FragOut fragOut;
        ...
        return fragOut;
    }
```
By default, the explicit imageblock storage is separate from the storage of the implicit imageblock. To share storage between the explicit imageblock and implicit imageblock, see section 5.6.5.

### 5.6.4 Imageblock Layouts in Kernel Functions

The `imageblock<T>` type (defined in the header `<metal_imageblocks.h>`) can only be used for arguments declared in a kernel function or in a user function that is called by a kernel function. Only a kernel function can have an argument declared as an `imageblock<T>` type. The data in an imageblock is visible only to threads in a threadgroup.

This imageblock argument to a kernel function is declared as the following templated type:

```cpp
class imageblock_layout_explicit;
class imageblock_layout_implicit;
template<typename T, typename L>
struct imageblock;
```

With the following restrictions:

- `L` is either `imageblock_layout_explicit` or `imageblock_layout_implicit`.
- `T` is a structure; members of `T` can be any of the following:
  - scalars
  - vectors and packed vectors
  - pixel data types
  - an array with elements that are one of the types on this list
  - a structure with members that are one of the types on this list

For an imageblock with implicit layout (`imageblock_layout_implicit`), each member of the structure may have a color attachment (see the `[[color(m)]]` attribute in Table 5.5 of section 5.2.3.4). The color index `m` needs to be unique for each member (and sub-member) of `T`.

If you do not specify an imageblock layout, the compiler deduces the layout based on `T`. If `T` is not compatible with an implicit or explicit imageblock, a compiler error occurs.

Both explicit and implicit imageblocks can be arguments to a kernel function. This also makes it easy to share explicit and implicit imageblock structures between fragment and kernel functions. By default, the explicit imageblock storage is separate from the storage of the implicit imageblock. To share storage between the explicit imageblock and implicit imageblock, see section 5.6.5.
5.6.5  Aliasing Explicit and Implicit Imageblocks

By default, explicit and implicit imageblocks do not alias. To alias the allocation of an explicit imageblock with the implicit imageblock fully or partially, you can use the following attributes to specify an explicit imageblock:

[[alias_implicit_imageblock]]
[[alias_implicit_imageblock_color(n)]]

The [[alias_implicit_imageblock]] attribute specifies that the explicit imageblock allocation completely aliases the implicit imageblock.

The [[alias_implicit_imageblock_color(n)]] attribute specifies that the explicit imageblock allocation aliases the implicit imageblock starting at a specific color attachment given by color(n). If n is a value that is between the smallest and largest declared attachments, inclusive, but n references an undeclared attachment, then a compile-time error occurs. If n is a value that exceeds the number of declared attachments, then compilation succeeds, but the attribute is ignored.

The behavior of accessing data members of an aliased implicit imageblock with an explicit imageblock is undefined if the kernel or fragment function modifies the aliased imageblock data members using the explicit imageblock and its associated member functions.

Example:

```
struct I {
    rgba8unorm<half4> a;
    rgb9e5<float4> b;
    int c;
    float d;
};

struct FragOut {
    float4 finalColor [[color(0)]];
    I i [[imageblock_data, alias_implicit_imageblock_color(1)]];
};

[[fragment]] FragOut
my_fragment(I i [[imageblock_data]], ...) {
    FragOut fragOut;
    ...
    return fragOut;
}
```

5.6.6  Imageblocks and Function Constants

Do not use [[function_constant(name)]] with data members of an imageblock structure either as input to or as returned output from a fragment or kernel function.
5.7 Graphics Function – Signature Matching

A graphics function signature is a list of parameters that are either input to or output from a graphics function.

5.7.1 Vertex – Fragment Signature Matching

You can pass two kinds of data between a vertex and fragment function: user-defined and built-in variables.

You can declare the per-instance input to a fragment function with the [[stage_in]] attribute. These are output by an associated vertex function.

You can declare built-in variables with one of the attributes defined in section 5.2.3. Examples of variables that use these attributes are:

- the vertex function output (with the [[position]], [[point_size]], or [[clip_distance]] attribute),
- the rasterizer output (with the [[point_coord]], [[frontFacing]], [[sample_id]], or [[sample_mask]] attribute),
- or fragment function input that refers to a framebuffer color value (with [[color]]).

Always return a built-in variable that specifies the [[position]] attribute. For built-in variables with either the [[point_size]] or [[clip_distance]] attribute, that attribute must also specify the corresponding vertex function output. If they are used and read in a fragment function, the shader has undefined behavior.

You may also declare built-in variables that are rasterizer output or refer to a framebuffer color value as the fragment function input with the appropriate attribute.

You can also use the attribute [[user(name)]] syntax to specify an attribute name for any user-defined variable.

A vertex function and a fragment function have matching signatures if:

- There is no input argument with the [[stage_in]] attribute declared in the fragment function.
- For a fragment function argument declared with [[stage_in]], each element in the type associated with this argument can be one of the following: a built-in variable generated by the rasterizer, a framebuffer color value passed as input to the fragment function, or a user-generated output from a vertex function. For built-in variables generated by the rasterizer or framebuffer color values, there is no requirement to associate a matching type with elements of the vertex return type. For elements that are user-generated outputs, the following rules apply:

If you specify an attribute name for an element using [[user(name)]], the attribute name must match with an element in the return type of the vertex function. If you do not specify the [[user(name)]] attribute name, then the argument name and types must match. In either case, their corresponding data types must also match or the fragment function argument type needs to be interpolant<T,P>, where T is the element’s type in the vertex return type.

Below is an example of using compatible signatures together (my_vertex and my_fragment, or my_vertex and my_fragment2) to render a primitive:
struct VertexOutput {
    float4 position [[position]];
    float3 normal;
    float2 texcoord;
};

vertex VertexOutput
my_vertex(...) {
    VertexOutput v;
    ...
    return v;
}

fragment float4
my_fragment(VertexOutput f [[stage_in]], ...) {
    float4 clr;
    ...
    return clr;
}

fragment float4
my_fragment2(VertexOutput f [[stage_in]], bool is_front_face [[front_facing]], ...) {
    float4 clr;
    ...
    return clr;
}

The following is an example of compatible signatures:

struct VertexOutput {
    float4 position [[position]];
    float3 vertex_normal [[user(normal)]];
    float2 texcoord [[user(texturecoord)]];
};

struct FragInput {
    float3 frag_normal [[user(normal)]];
    float4 position [[position]];
    float4 framebuffer_color [[color(0)]];
    bool is_front_face [[front_facing]];
};

vertex VertexOutput
my_vertex(...) {

VertexOutput v;
...
return v;
}

fragment float4
my_fragment(FragInput f [[stage_in]], ...) {
  float4 clr;
  ...
  return clr;
}

The following is an example of compatible signatures:

struct VertexOutput {
  float4 position [[position]];
  float3 normal;
  float2 texcoord;
};

vertex VertexOutput
my_vertex(…)
{
  VertexOutput v;
  ...
  return v;
}

fragment float4
my_fragment(float4 p [[position]], ...) {
  float4 clr;
  ...
  return clr;
}

Below is an example of incompatible signatures. The data type of normal in VertexOutput (float3) does not match the type of normal in FragInput (half3):

struct VertexOutput {
  float4 position [[position]];
  float3 normal;
  float2 texcoord;
};
struct FragInput {
    float4 position [[position]];
    half3 normal;
};

vertex VertexOutput
my_vertex(…)
{
    VertexOutput v;
    ...
    return v;
}

fragment float4
my_fragment(FragInput f [[stage_in]], …)
{
    float4 clr;
    ...
    return clr;
}

Below is another example of incompatible signatures. The attribute index of normal in
VertexOutput (normal) does not match the index of normal in FragInput (foo):

struct VertexOutput {
    float4 position [[position]];
    float3 normal [[user(normal)]];
    float2 texcoord [[user(texturecoord)]];
};

struct FragInput {
    float3 normal [[user(foo)]];
    float4 position [[position]];
};

vertex VertexOutput
my_vertex_shader(…)
{
    VertexOutput v;
    ...
    return v;
}

fragment float4
my_fragment_shader(FragInput f [[stage_in]], …)
{
    float4 clr;
    ...
    return clr;
5.7.2 Mesh – Fragment Signature Matching

You can pass the two kinds of data from vertex (V) and primitive (P) of mesh<V, P, NV, NP, t> from the mesh function to the fragment function: user-defined and built-in variables. The per-vertex mesh outputs defined in vertex (V) are always interpolated, whereas the per-primitive mesh outputs defined in primitive (P) are never interpolated. Due to this difference, the rules for signature matching of user-generated output have been adjusted from those described in section 5.7.1 Vertex - Fragment signature matching.

A given fragment input matches a user-generated mesh output from vertex (V) and primitive (P) if the following is true: If you specify an attribute name for an element using [[user(name)]]; the attribute name must match with an element in the return type of the mesh output. If you do not specify the [[user(name)]] attribute name, then the argument name and types must match. In either case, their corresponding data types must also match, or the fragment function argument type needs to be interpolant<T, P>, where T is the element's type in the vertex return type.

A mesh function and a fragment function have matching signatures for user-generated inputs with user-generated mesh outputs if:

- For a given user-generated fragment input with a flat interpolation:
  - There is a matching per-primitive mesh output, and the output is propagated to the fragment input without interpolation.
  - There is a matching per-vertex mesh output, and the output is propagated to the fragment input without interpolation.

- For a given user-generated fragment input with a non-flat interpolation:
  - There is a matching per-primitive mesh output, and the output is propagated to the fragment input without interpolation.
  - There is a matching per-vertex mesh output, and the output is interpolated across the primitive in the same method as non-flat vertex outputs are interpolated.

5.8 Program Scope Function Constants

All OS: function constants supported since Metal 1.2. Since Metal 2, support for using a function constant to specify the binding number for a resource (see section 5.8.1.4) to specify the index for the color() or raster_order_group attributes (section 5.8.1.5), and to identify that a structure element is optional (section 5.8.1.6).

Function constants enable the generation of multiple variants of a function. Without using function constants, you can compile one function many times with different preprocessor macro defines to enable different features (an ubershader). Using preprocessor macros for ubershaders with offline compiling can result in a large number of variants and a significant increase in the size of the shading function library assets. Function constants provide the same
ease of use as preprocessor macros but moves the generation of the specific variants to the creation of the pipeline state, so you don’t have to compile the variants offline.

5.8.1 Specifying Program Scope Function Constants

Program scope variables declared with (or initialized with) the following attribute are function constants:

```
[[function_constant(index)]]
```

The value `index` needs to be between 0 and 65535.

In Metal, function constants can:

- control code paths that get compiled.
- specify the optional arguments of a function (graphics, kernel, or user functions).
- specify optional elements of a structure with the `[[stage_in]]` attribute.

Function constants are not initialized in the Metal function source. Instead, their values are specified during the creation of the render or compute pipeline state. The `index` value specifies a location index that can refer to the function constant variable (instead of by its name) in the runtime.

Examples:

```swift
constant int a [[function_constant(0)]];
constant bool b [[function_constant(2)]];
```

Functions constants can only be a scalar or vector type. Using a user-defined type or an array of a scalar or vector type for a function constant results in a compilation error.

The value of function constants `a` and `b` are specified during the creation of the render or compute pipeline state.

You can also use function constants to initialize variables in program scope declared in the constant address space.

Examples:

```swift
constant int a [[function_constant(0)]];
constant bool b [[function_constant(2)]];
constant bool c = ((a == 1) && b);
constant int d = (a * 4);
```

You can use the following built-in function to determine if a function constant has been defined and is available. `name` refers to the function constant variable.

```swift
bool is_function_constant_defined(name)
```

Returns `true` if the function constant variable is defined and `false` otherwise.

If a function constant variable value is not defined during the creation of the pipeline state and if the graphics or kernel function specified with the render or compute pipeline state uses these function constants, the behavior is the same as when the value of `is_function_constant_defined(name)` is `false`. 
5.8.1.1 Function Constants to Control Code Paths to Compile

Consider the following function which uses preprocessor macros for function constants:

```c
struct VertexOutput {
    float4 position [[position]];
    float4 color;
};

struct VertexInput {
    float4 position [[attribute(0)]];
    float4 offset   [[attribute(1)]];
    float4 color    [[attribute(2)]];
};

vertex VertexOutput myVertex(VertexInput vIn [[stage_in]])
{
    VertexOutput vOut;
    vOut.position = vIn.position;
    #ifdef OFFSET_DEFINED
        vOut.position += vIn.offset;
    #endif

    #ifdef COLOR_DEFINED
        vOut.color = vIn.color;
    #else
        vOut.color = float4(0.0f);
    #endif

    return vOut;
}
```

The corresponding function written using function constant variables is:

```c
constant bool offset_defined [[function_constant(0)]];
constant bool color_defined [[function_constant(1)]];

vertex VertexOutput myVertex(VertexInput vIn [[stage_in]])
{
    VertexOutput vOut;
    vOut.position = vIn.position;
    if (offset_defined)
        vOut.position += vIn.offset;
```
if (color_defined)
    vOut.color = vIn.color;
else
    vOut.color = float4(0.0f);
return vOut;

5.8.1.2 Function Constants when Declaring the Arguments of Functions

You can declare an argument to a graphics, kernel, or other user function with the
[[function_constant(name)]] attribute to identify that the argument is optional. The
name attribute refers to a function constant variable. If the value of the function constant
variable given by name is non-zero or true (determined during creation of the pipeline
state), the declaration of the argument is in the function signature. If the value of the function
constant variable given by name is 0 or false, the argument is not declared in the function
signature. If name refers to a function constant variable that has not been defined (determined
during the creation of the pipeline state), the behavior is the same as if the value of
is_function_constant_defined(name) is false.

Consider the following fragment function that uses preprocessor macros in its function
declaration:

fragment half4
myFragment(
    constant GlobalUniformData *globalUniform [[buffer(0)]]),
    constant RenderUniformData_ModelWithLightmap
        *renderUniform    [[buffer(1)]]),
    constant MaterialUniformData
        *materialUniform [[buffer(2)]]),
texture2d<float> DiffuseTexture [[texture(0)]],
texture2d<float> LightmapTexture [[texture(1)]],
texture2d<float> FogTexture [[texture(3)]],
#ifdef MED_QUALITY
texture2d<float> LookupTexture [[texture(4)]],
#endif
#ifdef REALTIME_SHADOW
texture2d<float> RealtimeShadowMapTexture [[texture(10)]],
#endif
    sampler DiffuseTextureSampler [[sampler(0)]],
sampler LightmapTextureSampler [[sampler(1)]],
sampler FogTextureSampler [[sampler(3)]],
#ifdef MED_QUALITY
    sampler LookupTextureSampler [[sampler(4)]],
#endif
#ifdef REALTIME_SHADOW
    sampler RealtimeShadowMapTextureSampler [[sampler(10)]],
#endif
    VertexOutput fragIn [[stage_in]])
Here is the corresponding fragment function, after using function constants instead of `#ifdef` statements to rewrite the previous code:

```cpp
constant bool realtime_shadow [[function_constant(0)]];  
constant bool med_quality [[function_constant(1)]];    
constant bool med_quality_defined =               
  is_function_constant_defined(med_quality);      
constant bool realtime_shadow_defined =           
  is_function_constant_defined(realtime_shadow);  

fragment half4
myFragment(
  constant GlobalUniformData *globalUniform [[buffer(0)]],  
  constant RenderUniformData_ModelWithLightmap         
    *renderUniform [[buffer(1)]],                        
  constant MaterialUniformData *materialUniform [[buffer(2)]], 
  texture2d<float> DiffuseTexture [[texture(0)]],      
  texture2d<float> LightmapTexture [[texture(1)]],     
  texture2d<float> FogTexture [[texture(3)]],          
  texture2d<float> LookupTexture [[texture(4),       
    function_constant(med_quality_defined)]],         
  texture2d<float> RealtimeShadowMapTexture [[texture(10), 
    function_constant(realtime_shadow_defined)]],      
  sampler DiffuseTextureSampler [[sampler(0)]],       
  sampler LightmapTextureSampler [[sampler(1)]],      
  sampler FogTextureSampler [[sampler(3)]],           
  sampler LookupTextureSampler [[sampler(4),         
    function_constant(med_quality_defined)]],         
  sampler RealtimeShadowMapTextureSampler [[sampler(10), 
    function_constant(realtime_shadow_defined)]],      
  VertexOutput fragIn [[stage_in]])
```

Below is another example that shows how to use function constants with arguments to a function:

```cpp
constant bool hasInputBuffer [[function_constant(0)]];  

kernel void kernelOptionalBuffer(
  device int *input [[buffer(0), function_constant(hasInputBuffer)]],  
  device int *output [[buffer(1)]],  
  uint tid [[thread_position_in_grid]])
{
  if (hasInputBuffer)
    output[tid] = input[0] * tid;
  else
    output[tid] = tid;
}  ```
5.8.1.3  Function Constants for Elements of an Input Assembly Structure

You can use the `[[function_constant(name)]]` attribute to specify elements of an input assembly structure (declared with the `[[stage_in]]` attribute) as optional. If the value of the function constant variable given by `name` is non-zero or `true` (determined during the creation of the render or compute pipeline state), the element in the structure is declared in the function signature. If the value of the function constant variable given by `name` is `0` or `false`, the element is not declared in the structure.

Example:

```cpp
constant bool offset_defined [[function_constant(0)]];
call constant bool color_defined [[function_constant(1)]];

struct VertexOutput {
    float4 position [[position]];
    float4 color;
};
struct VertexInput {
    float4 position [[attribute(0)]];
    float4 offset   [[attribute(1),
                    function_constant(offset_defined)]];
    float4 color    [[attribute(2),
                    function_constant(color_defined)]];
};

type VertexOutput

myVertex(VertexInput vIn [[stage_in]])
{
    VertexOutput vOut;
    vOut.position = vIn.position;
    if (offset_defined)
        vOut.position += vIn.offset;
    if (color_defined)
        vOut.color = vIn.color;
    else
        vOut.color = float4(0.0f);
    return vOut;
}
```

5.8.1.4  Function Constants for Resource Bindings

All OS: Using a function constant to specify resource bindings supported since Metal 2.

An argument to a graphics or kernel functions that is a resource (buffer, texture, or sampler) can use a function constant to specify its binding number. The function constant needs to be a scalar integer type.

Example:
constant int indexA [[function_constant(0)]];  
constant int indexB = indexA + 2; 
constant int indexC [[function_constant(1)]];  
constant int indexD [[function_constant(2)]];  

[[kernel]] void 
my_kernel(constant UserParams& params [[buffer(indexA)]],  
device T * p [[buffer(indexB)]],  
texture2d<float> texA [[texture(indexC)]],  
sampler s [[sampler(indexD)]], ...
{
... 
}

5.8.1.5 Function Constants for Color Attachments and Raster Order Groups

All OS: using a function constant to specify a color attachment or raster order group attribute index supported since Metal 2.

The [[color(n)]] or [[raster_order_group(index)]] index can also be a function constant. The function constant used needs to be a scalar integer type.

Example:
constant int colorAttachment0 [[function_constant(0)]];  
constant int colorAttachment1 [[function_constant(1)]];  
constant int group0 [[function_constant(2)]];  

struct FragmentOutput {  
    float4 color0 [[color(colorAttachment0)]];  
    float4 color1 [[color(colorAttachment1)]];  
};

[[fragment]] FragmentOutput  
my_fragment(texture2d<float> texA [[texture(0),  
raster_order_group(group0)]]);  
{
... 
}

5.8.1.6 Function Constants with Elements of a Structure

All OS: Using a function constant to identify that a structure element is optional; supported since Metal 2.

To identify that an element of a structure is optional, you can specify the [[function_constant(name)]] attribute with elements of a structure that is the return type of a graphics or user function or is passed by value as an argument to a kernel, graphics, or user function. The behavior is similar to function constants for elements with the [[stage_in]] attribute, as described in section 5.8.1.3.

If the value of the function constant variable given by name is non-zero or true (determined during the render or compute pipeline state creation), the element in the structure is declared in the function signature. If the value of the function constant variable given by name is 0 or false, the element is not considered to be declared in the structure. If name refers to a
function constant variable that is undefined, the behavior is the same as if 
\texttt{is\_function\_constant\_defined(name) returns false}.

### 5.9 Program Scope Global Built-ins

Starting in Metal 3.1, developers can define global variables using attributes defined in Table 5.8 and use them in a kernel (including tile), mesh, or object context. The global variables cannot be used in a dynamic library or a separately compiled binary function.

Example:

```c
uint2 gid [[thread_position_in_grid]];

float4 get_color(texture2d<float> texInput, sampler s) {
    return texInput.sample(s, float2(gid));
}

[[kernel]] void my_kernel(texture2d<float> texInput, sampler s, ...){
    auto color = get_color(texInput, s);
    ...
}
```

### 5.10 Per-Primitive Viewport and Scissor Rectangle Index Selection

macOS: Support for the \texttt{viewport\_array\_index} attribute since Metal 2.
iOS: Support for the \texttt{viewport\_array\_index} attribute since Metal 2.1.

The \texttt{[[viewport\_array\_index]]} attribute supports built-in variables as both vertex output and fragment input. With \texttt{[[viewport\_array\_index]]}, the vertex function output specifies the rasterization viewport and scissor rectangle from the arrays specified by the \texttt{setViewports\_count:} and \texttt{setScissorRects\_count:} framework calls, respectively.

The unclamped value of the vertex function output for \texttt{[[viewport\_array\_index]]} is provided as input to the fragment function, even if the value is out of range.

The behavior of the fragment function with an unclamped \texttt{[[viewport\_array\_index]]} value depends upon the implementation. Either Metal can render every primitive to viewport/scissor rectangle \(0\), regardless of the passed value, or Metal can render to the \(n\)th viewport/scissor rectangle, where \(n\) is the clamped value. (Hardware that does not support this feature acts as only one viewport and one scissor rectangle are permitted, so the value for \texttt{[[viewport\_array\_index]]} is 0.)

You can specify \texttt{[[viewport\_array\_index]]} in a post-tessellation vertex function. You cannot specify \texttt{[[viewport\_array\_index]]} in the tessellation factor buffer.
Specifying \([[\text{viewport\_array\_index}]\]\) as fragment function input counts against the number of input assembly components available. (Input assembly components are the fragment function inputs declared with the \text{stage\_in} qualifier.)

You must return the same value of \([[\text{viewport\_array\_index}]\]\) for every vertex in a primitive. If the values differ, the behavior and the value passed to the fragment function are undefined. The same behavior applies to primitives generated by tessellation.

5.11 Additional Restrictions

MSL functions and arguments have these additional restrictions:

- Writes to a buffer from a vertex function are not guaranteed to be visible to reads from the associated fragment function of a given primitive.
- If a vertex function does writes to one or more buffers or textures, its return type needs to be \text{void}.
- The return type of a vertex or fragment function cannot include an element that is a packed vector type, matrix type, a structure type, a reference, or a pointer to a type.
- The number of inputs to a fragment function declared with the \text{stage\_in} attribute is limited. The input limits differ for different feature sets. The Metal Feature Set Tables at \url{https://developer.apple.com/metal/Metal-Feature-Set-Tables.pdf} lists the specific limits below “Implementation Limits by GPU Family”. (An input vector counts as \(n\) input scalars, where \(n\) is the number of components in the vector.)
- The argument type for arguments to a graphics or kernel function cannot be a derived class. Also the type of an argument to a graphics function that is declared with the \text{stage\_in} attribute cannot be a derived class.
6 Metal Standard Library

This chapter describes functions in the Metal Standard Library (MSLib).

6.1 Namespace and Header Files

The MSLib functions and enumerations are declared in the metal namespace. In addition to the header files described in the MSLib functions, the <metal_stdlib> header is available and can access all the functions supported by the MSLib.

6.2 Common Functions

The functions in Table 6.1 are defined in the header <metal_common>. T is one of the scalar or vector half or float floating-point types.

<table>
<thead>
<tr>
<th>Built-in Common Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T clamp(T x, T minval, T maxval)</td>
<td>Returns fmin(fmax(x, minval), maxval). Results are undefined if minval &gt; maxval.</td>
</tr>
<tr>
<td>T mix(T x, T y, T a)</td>
<td>Returns the linear blend of x and y implemented as: x + (y - x) * a. a needs to be a value in the range 0.0 to 1.0. If a is not in the range 0.0 to 1.0, the return values are undefined.</td>
</tr>
<tr>
<td>T saturate(T x)</td>
<td>Clamp the specified value within the range of 0.0 to 1.0.</td>
</tr>
<tr>
<td>T sign(T x)</td>
<td>Returns 1.0 if x &gt; 0, -0.0 if x = -0.0, +0.0 if x = +0.0, or -1.0 if x &lt; 0. Returns 0.0 if x is a NaN.</td>
</tr>
</tbody>
</table>
**Built-in Common Functions**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
</table>
| T smoothstep(T edge0, T edge1, T x) | Returns 0.0 if \( x \leq edge0 \) and 1.0 if \( x \geq edge1 \) and performs a smooth Hermite interpolation between 0 and 1 when \( edge0 < x < edge1 \). This is useful in cases where you want a threshold function with a smooth transition. This is equivalent to: \[
    t = \text{clamp}((x - edge0)/(edge1 - edge0), 0, 1);
    \text{return } t * t * (3 - 2 * t);
\]
| Results are undefined if \( edge0 >= edge1 \) or if \( x, edge0, \) or \( edge1 \) is a NaN. |
| T step(T edge, T x)                | Returns 0.0 if \( x < edge \), otherwise it returns 1.0. |

For single precision floating-point, Metal also supports a precise and fast variant of the following common functions: clamp and saturate. The difference between the Fast and precise function variants handle NaNs differently. In the fast variant, the behavior of NaNs is undefined, whereas the precise variants follow the IEEE 754 rules for NaN handling. The `ffast-math` compiler option (refer to section 1.5.3) selects the appropriate variant when compiling the Metal source. In addition, the `metal::precise` and `metal::fast` nested namespaces provide an explicit way to select the fast or precise variant of these common functions.

### 6.3 Integer Functions

The integer functions in Table 6.2 are defined in the header `<metal_integer>`. \( T \) is one of the scalar or vector integer types. \( Tu \) is the corresponding unsigned scalar or vector integer type. \( T32 \) is one of the scalar or vector 32-bit `int` or `uint` types.

<table>
<thead>
<tr>
<th>Built-in Integer Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T abs(T x)</td>
<td>Returns (</td>
</tr>
<tr>
<td>Tu absdiff(T x, T y)</td>
<td>Returns (</td>
</tr>
<tr>
<td>T addsat(T x, T y)</td>
<td>Returns ( x + y ) and saturates the result.</td>
</tr>
<tr>
<td>T clamp(T x, T minval, T maxval)</td>
<td>Returns ( \min(\max(x, \text{minval}), \text{maxval}) ). Results are undefined if ( \text{minval} &gt; \text{maxval} ).</td>
</tr>
<tr>
<td>Built-in Integer Functions</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>T clz(T x)</td>
<td>Returns the number of leading 0-bits in x, starting at the most significant bit position. If x is 0, returns the size in bits of the type of x or component type of x, if x is a vector.</td>
</tr>
<tr>
<td>T ctz(T x)</td>
<td>Returns the count of trailing 0-bits in x. If x is 0, returns the size in bits of the type of x or if x is a vector, the component type of x.</td>
</tr>
<tr>
<td>T extract_bits(T x,</td>
<td>Extract bits [offset, offset+bits−1] from x, returning them in the least significant bits of the result.</td>
</tr>
<tr>
<td>uint offset,</td>
<td>For unsigned data types, the most significant bits of the result are set to zero. For signed data types, the most significant bits are set to the value of bit offset+bits−1.</td>
</tr>
<tr>
<td>uint bits)</td>
<td>If bits is zero, the result is zero. If the sum of offset and bits is greater than the number of bits used to store the operand, the result is undefined.</td>
</tr>
<tr>
<td>All OS: Since Metal 1.2.</td>
<td></td>
</tr>
<tr>
<td>T hadd(T x, T y)</td>
<td>Returns (x + y) &gt;&gt; 1. The intermediate sum does not modulo overflow.</td>
</tr>
<tr>
<td>T insert_bits(T base,</td>
<td>Returns the insertion of the bits least-significant bits of insert into base.</td>
</tr>
<tr>
<td>T insert,</td>
<td>The result has bits [offset, offset+bits−1] taken from bits [0, bits−1] of insert, and all other bits are taken directly from the corresponding bits of base. If bits is zero, the result is base. If the sum of offset and bits is greater than the number of bits used to store the operand, the result is undefined.</td>
</tr>
<tr>
<td>uint offset,</td>
<td></td>
</tr>
<tr>
<td>uint bits)</td>
<td></td>
</tr>
<tr>
<td>All OS: Since Metal 1.2.</td>
<td></td>
</tr>
<tr>
<td>T32 mad24(T32 x, T32 y, T32 z)</td>
<td>Uses mul24 to multiply two 24-bit integer values x and y, adds the 32-bit integer result to the 32-bit integer z, and returns that sum.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>T madhi(T a, T b, T c)</td>
<td>Returns mulhi(a, b) + c.</td>
</tr>
<tr>
<td>T madsat(T a, T b, T c)</td>
<td>Returns a * b + c and saturates the result.</td>
</tr>
<tr>
<td>T max(T x, T y)</td>
<td>Returns y if x &lt; y, otherwise it returns x.</td>
</tr>
<tr>
<td>T max3(T x, T y, T z)</td>
<td>Returns max(x, max(y, z)).</td>
</tr>
<tr>
<td>All OS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>Built-in Integer Functions</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>T (\text{median3}(T\ x,\ T\ y,\ T\ z))</td>
<td>Return the middle value of (x), (y), and (z).</td>
</tr>
<tr>
<td>All OS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>T (\text{min}(T\ x,\ T\ y))</td>
<td>Returns (y) if (y &lt; x), otherwise it returns (x).</td>
</tr>
<tr>
<td>T (\text{min3}(T\ x,\ T\ y,\ T\ z))</td>
<td>Returns (\text{min}(x,\ \text{min}(y,\ z))).</td>
</tr>
<tr>
<td>All OS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>T32 (\text{mul24}(T32\ x,\ T32\ y))</td>
<td>Multiplies two 24-bit integer values (x) and (y) and returns the 32-bit integer result. (x) and (y) are 32-bit integers but only the low 24 bits perform the multiplication. (See details following this table.)</td>
</tr>
<tr>
<td>All OS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>T (\text{mulhi}(T\ x,\ T\ y))</td>
<td>Computes (x \ast y) and returns the high half of the product of (x) and (y).</td>
</tr>
<tr>
<td>T (\text{popcount}(T\ x))</td>
<td>Returns the number of non-zero bits in (x).</td>
</tr>
<tr>
<td>T (\text{reverse_bits}(T\ x))</td>
<td>Returns the reversal of the bits of (x). The bit numbered (n) of the result is taken from bit ((\text{bits} – 1) – n) of (x), where \text{bits} is the total number of bits used to represent (x).</td>
</tr>
<tr>
<td>All OS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>T (\text{rhadd}(T\ x,\ T\ y))</td>
<td>Returns ((x + y + 1) &gt;&gt; 1). The intermediate sum does not modulo overflow.</td>
</tr>
<tr>
<td>T (\text{rotate}(T\ v,\ T\ i))</td>
<td>For each element in (v), the bits are shifted left by the number of bits given by the corresponding element in (i). Bits shifted off the left side of the element are shifted back in from the right.</td>
</tr>
<tr>
<td>T (\text{subsat}(T\ x,\ T\ y))</td>
<td>Returns (x – y) and saturates the result.</td>
</tr>
</tbody>
</table>

The \(\text{mul24}\) function only operates as described if \(x\) and \(y\) are signed integers and \(x\) and \(y\) are in the range \([-2^{23},\ 2^{23} – 1]\), or if \(x\) and \(y\) are unsigned integers and \(x\) and \(y\) are in the range \([0,\ 2^{24} – 1]\). If \(x\) and \(y\) are not in this range, the multiplication result is implementation-defined.

### 6.4 Relational Functions

The relational functions in Table 6.3 are defined in the header `<metal_relational>`. \(T\) is one of the scalar or vector floating-point types including \(\text{bf}float\) types. \(T_i\) is one of the scalar or vector integer or boolean types. \(T_b\) only refers to the scalar or vector Boolean types.
### Table 6.3. Relational functions in the Metal standard library

<table>
<thead>
<tr>
<th>Built-in Relational Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bool all(Tb x)</td>
<td>Returns true only if all components of x are true.</td>
</tr>
<tr>
<td>bool any(Tb x)</td>
<td>Returns true only if any component of x are true.</td>
</tr>
<tr>
<td>Tb isfinite(T x)</td>
<td>Test for finite value.</td>
</tr>
<tr>
<td>Tb isinf(T x)</td>
<td>Test for infinity value (positive or negative).</td>
</tr>
<tr>
<td>Tb isnan(T x)</td>
<td>Test for a NaN.</td>
</tr>
<tr>
<td>Tb isnormal(T x)</td>
<td>Test for a normal value.</td>
</tr>
<tr>
<td>Tb isordered(T x, T y)</td>
<td>Test if arguments are ordered. isordered() takes arguments x and y and returns the result (x == x) &amp;&amp; (y == y).</td>
</tr>
<tr>
<td>Tb isunordered(T x, T y)</td>
<td>Test if arguments are unordered. isunordered() takes arguments x and y and returns true if x or y is NaN; otherwise returns false.</td>
</tr>
<tr>
<td>Tb not(Tb x)</td>
<td>Returns the component-wise logical complement of x.</td>
</tr>
<tr>
<td>T select(T a, T b, Tb c)</td>
<td>For each component of a vector type, result[i] = c[i] ? b[i] : a[i]</td>
</tr>
<tr>
<td>Ti select(Ti a, Ti b, Tb c)</td>
<td>For a scalar type, result = c ? b : a</td>
</tr>
<tr>
<td>Tb signbit(T x)</td>
<td>Test for sign bit. Returns true if the sign bit is set for the floating-point value in x; otherwise returns false.</td>
</tr>
</tbody>
</table>

### 6.5 Math Functions

The math functions in Table 6.4 are defined in the header `<metal_math>`. T is one of the scalar or vector half or float floating-point types. Ti refers only to the scalar or vector integer types.

<table>
<thead>
<tr>
<th>Built-in Math Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T acos(T x)</td>
<td>Compute arc cosine of x.</td>
</tr>
<tr>
<td>T acosh(T x)</td>
<td>Compute inverse hyperbolic cosine of x.</td>
</tr>
<tr>
<td>Built-in Math Functions</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>T asin(T x)</td>
<td>Compute arc sine function of x.</td>
</tr>
<tr>
<td>T asinh(T x)</td>
<td>Compute inverse hyperbolic sine of x.</td>
</tr>
<tr>
<td>T atan(T y_over_x)</td>
<td>Compute arc tangent of x.</td>
</tr>
<tr>
<td>T atan2(T y, T x)</td>
<td>Compute arc tangent of y over x.</td>
</tr>
<tr>
<td>T atanh(T x)</td>
<td>Compute hyperbolic arc tangent of x.</td>
</tr>
<tr>
<td>T ceil(T x)</td>
<td>Round x to integral value using the round to positive infinity rounding mode.</td>
</tr>
<tr>
<td>T copysign(T x, T y)</td>
<td>Return x with its sign changed to match the sign of y.</td>
</tr>
<tr>
<td>T cos(T x)</td>
<td>Compute cosine of x.</td>
</tr>
<tr>
<td>T cosh(T x)</td>
<td>Compute hyperbolic cosine of x.</td>
</tr>
<tr>
<td>T cospi(T x)</td>
<td>Compute cos(πx).</td>
</tr>
<tr>
<td>T divide(T x, T y)</td>
<td>Compute x / y.</td>
</tr>
<tr>
<td>T exp(T x)</td>
<td>Exponential base e function.</td>
</tr>
<tr>
<td>T exp2(T x)</td>
<td>Exponential base 2 function.</td>
</tr>
<tr>
<td>T exp10(T x)</td>
<td>Exponential base 10 function.</td>
</tr>
<tr>
<td>T fabs(T x)</td>
<td>Compute absolute value of a floating-point number.</td>
</tr>
<tr>
<td>T abs(T x)</td>
<td></td>
</tr>
<tr>
<td>T fdim(T x, T y)</td>
<td>x – y if x &gt; y; +0 if x &lt;= y.</td>
</tr>
<tr>
<td>T floor(T x)</td>
<td>Round x to integral value using the round to negative infinity rounding mode.</td>
</tr>
<tr>
<td>T fma(T a, T b, T c)</td>
<td>Returns the correctly rounded floating-point representation of the sum of c with the infinitely precise product of a and b. Rounding of intermediate products shall not occur. Edge case behavior is per the IEEE 754-2008 standard.</td>
</tr>
<tr>
<td>T fmax(T x, T y)</td>
<td>Returns y if x &lt; y, otherwise returns x. If one argument is a NaN, fmax() returns the other argument. If both arguments are NaNs, fmax() returns a NaN. If x and y are denormals and the GPU doesn’t support denormals, either value may be returned.</td>
</tr>
<tr>
<td>T max(T x, T y)</td>
<td></td>
</tr>
</tbody>
</table>
## Built-in Math Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>fmax3(T x, T y, T z)</code></td>
<td>Returns <code>fmax(x, fmax(y, z))</code>.</td>
</tr>
<tr>
<td><code>max3(T x, T y, T z)</code></td>
<td>Returns the middle value of <code>x</code>, <code>y</code>, and <code>z</code>. (If one or more values are NaN, see discussion after this table.)</td>
</tr>
<tr>
<td><code>median3(T x, T y, T z)</code></td>
<td>Returns <code>fmin(x, fmin(y, z))</code>.</td>
</tr>
<tr>
<td><code>median3(T x, T y, T z)</code></td>
<td>Returns <code>fmin(x, fmin(y, z))</code>.</td>
</tr>
<tr>
<td><code>fmin(T x, T y)</code></td>
<td>Returns <code>y</code> if <code>y &lt; x</code>, otherwise it returns <code>x</code>. If one argument is a NaN, <code>fmin()</code> returns the other argument. If both arguments are NaNs, <code>fmin()</code> returns a NaN. If <code>x</code> and <code>y</code> are denormals and the GPU doesn’t support denormals, either value may be returned.</td>
</tr>
<tr>
<td><code>mod(T x, T y)</code></td>
<td>Returns <code>x - y * trunc(x/y)</code>.</td>
</tr>
<tr>
<td><code>frac(T x)</code></td>
<td>Returns the fractional part of <code>x</code> that is greater than or equal to 0 or less than 1.</td>
</tr>
<tr>
<td><code>frexp(T x, Ti &amp;exponent)</code></td>
<td>Extract mantissa and exponent from <code>x</code>. For each component the mantissa returned is a float with magnitude in the interval <code>[1/2, 1)</code> or 0. Each component of <code>x</code> equals mantissa returned * 2^exp.</td>
</tr>
<tr>
<td><code>ilogb(T x)</code></td>
<td>Return the exponent as an integer value.</td>
</tr>
<tr>
<td><code>ldexp(T x, Ti k)</code></td>
<td>Multiply <code>x</code> by 2 to the power <code>k</code>.</td>
</tr>
<tr>
<td><code>log(T x)</code></td>
<td>Compute the natural logarithm of <code>x</code>.</td>
</tr>
<tr>
<td><code>log2(T x)</code></td>
<td>Compute the base 2 logarithm of <code>x</code>.</td>
</tr>
<tr>
<td><code>log10(T x)</code></td>
<td>Compute the base 10 logarithm of <code>x</code>.</td>
</tr>
<tr>
<td><code>modf(T x, T &amp;intval)</code></td>
<td>Decompose a floating-point number. The <code>modf</code> function breaks the argument <code>x</code> into integral and fractional parts, each of which has the same sign as the argument. Returns the fractional value. The integral value is returned in <code>intval</code>.</td>
</tr>
<tr>
<td><code>nextafter(T x, T y)</code></td>
<td>Return next representable floating-point value after <code>x</code> in the direction of <code>y</code>. If <code>x</code> equals <code>y</code>, return</td>
</tr>
</tbody>
</table>
### Built-in Math Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>T pow(T x, T y)</code></td>
<td>Compute <code>x</code> to the power <code>y</code>.</td>
</tr>
<tr>
<td><code>T powr(T x, T y)</code></td>
<td>Compute <code>x</code> to the power <code>y</code>, where <code>x</code> is &gt;= 0.</td>
</tr>
<tr>
<td><code>T rint(T x)</code></td>
<td>Round <code>x</code> to integral value using round ties to even rounding mode in floating-point format.</td>
</tr>
<tr>
<td><code>T round(T x)</code></td>
<td>Return the integral value nearest to <code>x</code>, rounding halfway cases away from zero.</td>
</tr>
<tr>
<td><code>T rsqrt(T x)</code></td>
<td>Compute inverse square root of <code>x</code>.</td>
</tr>
<tr>
<td><code>T sin(T x)</code></td>
<td>Compute sine of <code>x</code>.</td>
</tr>
<tr>
<td><code>T sincos(T x, T &amp;cosval)</code></td>
<td>Compute sine and cosine of <code>x</code>. Return the computed sine in the function return value, and return the computed cosine in <code>cosval</code>.</td>
</tr>
<tr>
<td><code>T sinh(T x)</code></td>
<td>Compute hyperbolic sine of <code>x</code>.</td>
</tr>
<tr>
<td><code>T sinpi(T x)</code></td>
<td>Compute <code>sin(\pi x)</code>.</td>
</tr>
<tr>
<td><code>T sqrt(T x)</code></td>
<td>Compute square root of <code>x</code>.</td>
</tr>
<tr>
<td><code>T tan(T x)</code></td>
<td>Compute tangent of <code>x</code>.</td>
</tr>
<tr>
<td><code>T tanh(T x)</code></td>
<td>Compute hyperbolic tangent of <code>x</code>.</td>
</tr>
<tr>
<td><code>T tanpi(T x)</code></td>
<td>Compute <code>tan(\pi x)</code>.</td>
</tr>
<tr>
<td><code>T trunc(T x)</code></td>
<td>Round <code>x</code> to integral value using the round toward zero rounding mode.</td>
</tr>
</tbody>
</table>

For `fmedian3`, if all values are NaN, return NaN. Otherwise, treat NaN as missing data and remove it from the set. If two values are NaN, return the non-NaN value. If one of the values is NaN, the function can return either non-NaN value.

For single precision floating-point, Metal supports two variants of the math functions listed in Table 6.4: the precise and the fast variants. The `ffast_math` compiler option (refer to section 1.5.3) selects the appropriate variant when compiling the Metal source. In addition, the `metal::precise` and `metal::fast` nested namespaces provide an explicit way to select the fast or precise variant of these math functions for single precision floating-point.

Examples:
```c
float x;
float a = sin(x); // Use fast or precise version of sin based on
```
// whether you specify -ffast-math as compile option.

float b = fast::sin(x); // Use fast version of sin().
float c = precise::cos(x); // Use precise version of cos().

All OS: Constants listed in Table 6.5 and Table 6.6 are supported since Metal 1.2.
Table 6.5 lists available symbolic constants with values of type float that are accurate within the precision of a single-precision floating-point number.

Table 6.5. Constants for single-precision floating-point math functions

<table>
<thead>
<tr>
<th>Constant Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAXFLOAT</td>
<td>Value of maximum non-infinite single precision floating-point number.</td>
</tr>
<tr>
<td>HUGE_VALF</td>
<td>A positive float constant expression. HUGE_VALF evaluates to +infinity.</td>
</tr>
<tr>
<td>INFINITY</td>
<td>A constant expression of type float representing positive or unsigned infinity.</td>
</tr>
<tr>
<td>NAN</td>
<td>A constant expression of type float representing a quiet NaN.</td>
</tr>
<tr>
<td>M_E_F</td>
<td>Value of e</td>
</tr>
<tr>
<td>M_LOG2E_F</td>
<td>Value of log₂e</td>
</tr>
<tr>
<td>M_LOG10E_F</td>
<td>Value of log₁₀e</td>
</tr>
<tr>
<td>M_LN2_F</td>
<td>Value of logₑ2</td>
</tr>
<tr>
<td>M_LN10_F</td>
<td>Value of logₑ₁₀</td>
</tr>
<tr>
<td>M_PI_F</td>
<td>Value of π</td>
</tr>
<tr>
<td>M_PI_2_F</td>
<td>Value of π / 2</td>
</tr>
<tr>
<td>M_PI_4_F</td>
<td>Value of π / 4</td>
</tr>
<tr>
<td>M_1_PI_F</td>
<td>Value of 1 / π</td>
</tr>
<tr>
<td>M_2_PI_F</td>
<td>Value of 2 / π</td>
</tr>
<tr>
<td>M_2_SQRTPI_F</td>
<td>Value of 2 / √π</td>
</tr>
<tr>
<td>M_SQRT2_F</td>
<td>Value of √2</td>
</tr>
<tr>
<td>M_SQRT1_2_F</td>
<td>Value of 1 / √2</td>
</tr>
</tbody>
</table>
Table 6.6 lists available symbolic constants with values of type `half` that are accurate within the precision of a half-precision floating-point number.

<table>
<thead>
<tr>
<th>Constant Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAXHALF</td>
<td>Value of maximum non-infinite half precision floating-point number.</td>
</tr>
<tr>
<td>HUGE_VALH</td>
<td>A positive half constant expression. HUGE_VALH evaluates to +infinity.</td>
</tr>
<tr>
<td>M_E_H</td>
<td>Value of e</td>
</tr>
<tr>
<td>M_LOG2E_H</td>
<td>Value of log₂e</td>
</tr>
<tr>
<td>M_LOG10E_H</td>
<td>Value of log₁₀e</td>
</tr>
<tr>
<td>M_LN2_H</td>
<td>Value of logₑ2</td>
</tr>
<tr>
<td>M_LN10_H</td>
<td>Value of logₑ10</td>
</tr>
<tr>
<td>M_PI_H</td>
<td>Value of π</td>
</tr>
<tr>
<td>M_PI_2_H</td>
<td>Value of π / 2</td>
</tr>
<tr>
<td>M_PI_4_H</td>
<td>Value of π / 4</td>
</tr>
<tr>
<td>M_1_PI_H</td>
<td>Value of 1 / π</td>
</tr>
<tr>
<td>M_2_PI_H</td>
<td>Value of 2 / π</td>
</tr>
<tr>
<td>M_2_SQRTPI_H</td>
<td>Value of 2 / √π</td>
</tr>
<tr>
<td>M_SQRT2_H</td>
<td>Value of √2</td>
</tr>
<tr>
<td>M_SQRT1_2_H</td>
<td>Value of 1 / √2</td>
</tr>
</tbody>
</table>

Table 6.7 lists available symbolic constants with values of type `bfloat` that are accurate within the precision of a brain floating-point number.
### Table 6.7. Constants for brain floating-point math functions

<table>
<thead>
<tr>
<th>Constant Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAXBFLOAT</td>
<td>Value of maximum non-infinite bfloat floating-point number.</td>
</tr>
<tr>
<td>HUGE_VALBF</td>
<td>A positive half constant expression. HUGE_VALBF evaluates to +infinity.</td>
</tr>
<tr>
<td>M_E_BF</td>
<td>Value of e</td>
</tr>
<tr>
<td>M_LOG2E_BF</td>
<td>Value of ( \log_2 e )</td>
</tr>
<tr>
<td>M_LOG10E_BF</td>
<td>Value of ( \log_{10} e )</td>
</tr>
<tr>
<td>M_LN2_BF</td>
<td>Value of ( \log_e 2 )</td>
</tr>
<tr>
<td>M_LN10_BF</td>
<td>Value of ( \log_e 10 )</td>
</tr>
<tr>
<td>M_PI_BF</td>
<td>Value of ( \pi )</td>
</tr>
<tr>
<td>M_PI_2_BF</td>
<td>Value of ( \pi / 2 )</td>
</tr>
<tr>
<td>M_PI_4_BF</td>
<td>Value of ( \pi / 4 )</td>
</tr>
<tr>
<td>M_1_PI_BF</td>
<td>Value of ( 1 / \pi )</td>
</tr>
<tr>
<td>M_2_PI_BF</td>
<td>Value of ( 2 / \pi )</td>
</tr>
<tr>
<td>M_2_SQRTPI_BF</td>
<td>Value of ( 2 / \sqrt{\pi} )</td>
</tr>
<tr>
<td>M_SQRT2_BF</td>
<td>Value of ( \sqrt{2} )</td>
</tr>
<tr>
<td>M_SQRT1_2_BF</td>
<td>Value of ( 1 / \sqrt{2} )</td>
</tr>
</tbody>
</table>

### 6.6 Matrix Functions

The functions in Table 6.8 are defined in the header `<metal_matrix>`. \( T \) is float or half.

### Table 6.8. Matrix functions in the Metal standard library

<table>
<thead>
<tr>
<th>Built-in Matrix Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>float determinant(floatnxn) half determinant(halfnxn)</td>
<td>Compute the determinant of the matrix. The matrix needs to be a square matrix.</td>
</tr>
<tr>
<td>floatmxn transpose(floatnxm) halfmxn transpose(halfnxm)</td>
<td>Transpose a matrix.</td>
</tr>
</tbody>
</table>
Example:
float4x4 mA;
float det = determinant(mA);

6.7 SIMD-group Matrix Functions

The SIMD-group Matrix functions are defined in the header `<metal_simdgroup_matrix>.

6.7.1 Creating, Loading, and Storing Matrix Elements

Metal Shading Library supports the following functions to initialize a SIMD-group matrix with a value, load data from threadgroup or device memory, and store data to threadgroup or device memory.

**Table 6.9. SIMD-group matrix load and stores**

<table>
<thead>
<tr>
<th>Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>simdgroup_matrix&lt;T,Cols,Rows&gt;(T dval)</td>
<td>Creates a diagonal matrix with the given value.</td>
</tr>
<tr>
<td>simdgroup_matrix&lt;T,Cols,Rows&gt; make_filled_simdgroup_matrix(T value)</td>
<td>Initializes a SIMD-group matrix filled with the given value.</td>
</tr>
<tr>
<td>void simdgroup_load( thread simdgroup_matrix&lt;T,Cols,Rows&gt;&amp; d, const threadgroup T *src, ulong elements_per_row = Cols, ulong2 matrix_origin = 0, bool transpose_matrix = false)</td>
<td>Loads data from threadgroup memory into a SIMD-group matrix. The <code>elements_per_row</code> parameter indicates the number of elements in the source memory layout.</td>
</tr>
<tr>
<td>void simdgroup_load( thread simdgroup_matrix&lt;T,Cols,Rows&gt;&amp; d, const device T *src, ulong elements_per_row = Cols, ulong2 matrix_origin = 0, bool transpose_matrix = false)</td>
<td>Loads data from device memory into a SIMD-group matrix. The <code>elements_per_row</code> parameter indicates the number of elements in the source memory layout.</td>
</tr>
<tr>
<td>void simdgroup_store( thread simdgroup_matrix&lt;T,Cols,Rows&gt; a, const threadgroup T *dst,</td>
<td>Stores data from a SIMD-group matrix into threadgroup memory. The <code>elements_per_row</code></td>
</tr>
</tbody>
</table>
### Functions

<table>
<thead>
<tr>
<th>Description</th>
<th>Functions</th>
</tr>
</thead>
</table>
| Stores data from a SIMD-group matrix into device memory. The elements_per_row parameter indicates the number of elements in the destination memory layout. | void simdgroup_store(
  thread simdgroup_matrix<T,Cols,Rows> a,
  const device T *dst,
  ulong elements_per_row = Cols,
  ulong2 matrix_origin = 0,
  bool transpose_matrix = false)

Parameter indicates the number of elements in the destination memory layout. | ulong elements_per_row = Cols,
ulong2 matrix_origin = 0,
bool transpose_matrix = false) |

---

### 6.7.2 Matrix Operations

SIMD-group matrices support multiply-accumulate and multiple operations.

**Table 6.10. SIMD-group operations**

<table>
<thead>
<tr>
<th>Operations</th>
<th>Description</th>
</tr>
</thead>
</table>
| void simdgroup_multiply_accumulate(
  thread simdgroup_matrix<T,Cols,Rows>& d,
  thread simdgroup_matrix<T,K,Rows>& a,
  thread simdgroup_matrix<T,Cols,K>& b,
  thread simdgroup_matrix<T,Cols,Rows>& c)
| Returns \(d = a \times b + c\) |
| void simdgroup_multiply(
  thread simdgroup_matrix<T,Cols,Rows>& d,
  thread simdgroup_matrix<T,K,Rows>& a,
  thread simdgroup_matrix<T,Cols,K>& b)
| Returns \(d = a \times b\) |
| * | Returns \(a \times b\) |

Here is an example of how to use SIMD-group matrices.

```c
kernel void float_matmad(device float *pMatA, device float *pMatB
  device float *pMatC, device float *pMatR)
{ 
  simdgroup_float8x8 sgMatA;
  simdgroup_float8x8 sgMatB;
  simdgroup_float8x8 sgMatC;
  simdgroup_float8x8 sgMatR;

  simdgroup_load(sgMatA, pMatA);
  simdgroup_load(sgMatB, pMatB);
  simdgroup_load(sgMatC, pMatC);
  simdgroup_load(sgMatR, pMatR);
```
simdgroup_multiply_accumulate(sgMatR, sgMatA, sgMatB, sgMatC);
simdgroup_store(sgMatR, pMatR);

6.8 Geometric Functions

The functions in Table 6.11 are defined in the header <metal_geometric>. T is a vector floating-point type (floatn or halfn). Ts refers to the corresponding scalar type. (If T is floatn, the scalar type Ts is float. If T is halfn, Ts is half.)

<table>
<thead>
<tr>
<th>Built-in Geometric Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T cross(T x, T y)</td>
<td>Return the cross product of x and y. T needs to be a 3-component vector type.</td>
</tr>
<tr>
<td>Ts distance(T x, T y)</td>
<td>Return the distance between x and y, which is length(x–y)</td>
</tr>
<tr>
<td>Ts distance_squared(T x, T y)</td>
<td>Return the square of the distance between x and y.</td>
</tr>
<tr>
<td>Ts dot(T x, T y)</td>
<td>Return the dot product of x and y, which is ( x[0] \times y[0] + x[1] \times y[1] + \ldots )</td>
</tr>
<tr>
<td>T faceforward(T N, T I, T Nref)</td>
<td>If ( \text{dot}(\text{Nref}, \ I) &lt; 0.0 ) return N, otherwise return (-N).</td>
</tr>
<tr>
<td>Ts length(T x)</td>
<td>Return the length of vector x, which is ( \sqrt{x[0]^2 + x[1]^2 + \ldots} )</td>
</tr>
<tr>
<td>Ts length_squared(T x)</td>
<td>Return the square of the length of vector x, which is ( (x[0]^2 + x[1]^2 + \ldots) )</td>
</tr>
<tr>
<td>T normalize(T x)</td>
<td>Return a vector in the same direction as x but with a length of 1.</td>
</tr>
<tr>
<td>T reflect(T I, T N)</td>
<td>For the incident vector I and surface orientation N, compute normalized N (NN), and return the reflection direction: ( I - 2 \times \text{dot}(NN, I) \times NN ).</td>
</tr>
</tbody>
</table>
### Built-in Geometric Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>T refract(T I, T N, Ts eta)</code></td>
<td>For the incident vector I and surface normal N, and the ratio of indices of refraction eta, return the refraction vector. The input parameters for the incident vector I and the surface normal N needs to already be normalized to get the desired results.</td>
</tr>
</tbody>
</table>

For single precision floating-point, Metal also supports a precise and fast variant of the following geometric functions: `distance`, `length`, and `normalize`. To select the appropriate variant when compiling the Metal source, use the `ffast-math` compiler option (refer to section 1.5.3). In addition, the `metal::precise` and `metal::fast` nested namespaces are also available and provide an explicit way to select the fast or precise variant of these geometric functions.

### 6.9 Synchronization and SIMD-group Functions

You can use synchronization and SIMD-group functions in:

- kernel functions
- fragment functions
- visible functions that kernel or fragment functions call

#### 6.9.1 Threadgroup and SIMD-group Synchronization Functions

The `<metal_compute>` header defines the synchronization functions in Table 6.12, which lists threadgroup and SIMD-group synchronization functions it supports.

<table>
<thead>
<tr>
<th>Built-in Threadgroup Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>void threadgroup_barrier(mem_flags flags)</code></td>
<td>All threads in a threadgroup executing the kernel, fragment, mesh, or object need to execute this function before any thread can continue execution beyond the threadgroup_barrier.</td>
</tr>
<tr>
<td><code>void simdgroup_barrier(mem_flags flags)</code></td>
<td>All threads in a SIMD-group executing the kernel, fragment, mesh, or object need to execute this function before any thread can continue execution beyond the simdgroup_barrier.</td>
</tr>
</tbody>
</table>

A barrier function (`threadgroup_barrier` or `simdgroup_barrier`) acts as an execution and memory barrier. All threads in a threadgroup (or SIMD-group) executing the kernel need to encounter the `threadgroup_barrier` (or `simdgroup_barrier`) function.
If `threadgroup_barrier` (or `simdgroup_barrier`) is inside a conditional statement and if any thread enters the conditional statement and executes the barrier function, then all threads in the threadgroup (or SIMD-group) need to enter the conditional and execute the barrier function.

If `threadgroup_barrier` (or `simdgroup_barrier`) is inside a loop, for each iteration of the loop, all threads in the threadgroup (or SIMD-group) need to execute the barrier function before any threads continue execution beyond the barrier function.

The `threadgroup_barrier` (or `simdgroup_barrier`) function can also queue a memory fence (for reads and writes) to ensure the correct ordering of memory operations to threadgroup or device memory.

Table 6.13 describes the bit field values for the `mem_flags` argument to `threadgroup_barrier` and `simdgroup_barrier`. The `mem_flags` argument ensures the correct memory is in the correct order between threads in the threadgroup or simdgroup (for `threadgroup_barrier` or `simdgroup_barrier`) respectively.

### Table 6.13. Memory flag enumeration values for barrier functions

<table>
<thead>
<tr>
<th><code>mem_flags</code></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mem_none</td>
<td>The flag sets <code>threadgroup_barrier</code> or <code>simdgroup_barrier</code> to only act as an execution barrier and doesn’t apply a Memory fence.</td>
</tr>
<tr>
<td>mem_device</td>
<td>The flag ensures the GPU correctly orders the memory operations to device memory for threads in the threadgroup or simdgroup.</td>
</tr>
<tr>
<td>mem_threadgroup</td>
<td>The flag ensures the GPU correctly orders the memory operations to threadgroup memory for threads in a threadgroup or simdgroup.</td>
</tr>
<tr>
<td>mem_texture</td>
<td>The flag ensures the GPU correctly orders the memory operations to texture memory for threads in a threadgroup or simdgroup for a texture with the <code>read_write</code> access qualifier.</td>
</tr>
<tr>
<td>block</td>
<td>The flag ensures the GPU correctly orders the memory operations to threadgroup imageblock memory for threads in a threadgroup or simdgroup.</td>
</tr>
<tr>
<td>mem_object_data</td>
<td>The flag ensures the GPU correctly orders the memory operations to <code>object_data</code> memory for threads in the threadgroup or simdgroup.</td>
</tr>
</tbody>
</table>

### 6.9.2 SIMD-group Functions

The `<metal_simdgroup>` header defines the SIMD-group functions for kernel and fragment functions. macOS supports SIMD-group functions in Metal 2 and later, and iOS supports most
SIMD-group functions in Metal 2.2 and later. Table 6.14 and Table 6.15 lists the SIMD-group functions and their availabilities in iOS and macOS. See the Metal Feature Set Tables to determine which GPUs support each table.

SIMD-group functions allow threads in a SIMD-group (see section 4.4.1) to share data without using threadgroup memory or requiring any synchronization operations, such as a barrier.

An active thread is a thread that is executing. An inactive thread is a thread that is not executing. For example, a thread may not be active due to flow control or when a task has insufficient work to fill the group. A thread needs to only read data from another active thread in the SIMD-group.

Helper threads may also be active and inactive. For example, if a helper thread finishes executing, it becomes an inactive helper thread. Helper threads for SIMD-group functions can be active or inactive. Use simd_is_helper_thread() (see Table 6.14) to inspect whether a thread is a helper thread.

Table 6.14 uses the placeholder T to represent a scalar or vector of any integer or floating-point type, except:

- bool
- void
- size_t
- ptrdiff_t

For bitwise operations, Ti needs to be an integer scalar or vector.

See 6.9.2.1 after the table for examples that use SIMD-group functions.

### Table 6.14. SIMD-group permute functions in the Metal standard library

<table>
<thead>
<tr>
<th>Built-in SIMD-group Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>simd_vote</td>
<td>Returns a simd_vote mask that represents the active threads. This function is equivalent to simd_ballot(true) and sets the bits that represent active threads to 1, and inactive Threads to 0.</td>
</tr>
<tr>
<td>simd_active_threads_mask()</td>
<td></td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.2.</td>
<td></td>
</tr>
<tr>
<td>bool simd_all(bool expr)</td>
<td>Returns true if all active threads evaluate expr to true.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.2.</td>
<td></td>
</tr>
<tr>
<td>bool simd_any(bool expr)</td>
<td>Returns true if at least one active thread evaluates Expr to true.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.2.</td>
<td></td>
</tr>
<tr>
<td><strong>Built-in SIMD-group Functions</strong></td>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td><strong>simd_vote simd_ballot (bool expr)</strong>&lt;br&gt;macOS: Since Metal 2.1.&lt;br&gt;iOS: Since Metal 2.2.</td>
<td>Returns a wrapper type — see the <code>simd_vote</code> example — around a bitmask of the evaluation of the Boolean expression for all active threads in the SIMD-group for which <code>expr</code> is true. The function sets the bits that correspond to inactive threads to 0.</td>
</tr>
<tr>
<td><strong>T simd_broadcast(T data, ushort broadcast_lane_id)</strong>&lt;br&gt;macOS: Since Metal 2.&lt;br&gt;iOS: Since Metal 2.2.</td>
<td>Broadcasts <code>data</code> from the thread whose SIMD lane ID is equal to <code>broadcast_lane_id</code>. The specification doesn’t define the behavior when <code>broadcast_lane_id</code> isn’t a valid SIMD lane ID or isn’t the same for all threads in a SIMD-group.</td>
</tr>
<tr>
<td><strong>T simd_broadcast_first(T data)</strong>&lt;br&gt;macOS: Since Metal 2.1.&lt;br&gt;iOS: Since Metal 2.2.</td>
<td>Broadcasts <code>data</code> from the first active thread — the active thread with the smallest index — in the SIMD-group to all active threads.</td>
</tr>
<tr>
<td><strong>bool simd_is_first()</strong>&lt;br&gt;macOS: Since Metal 2.1.&lt;br&gt;iOS: Since Metal 2.2.</td>
<td>Returns <code>true</code> if the current thread is the first active thread — the active thread with the smallest index — in the current SIMD-group; otherwise <code>false</code>.</td>
</tr>
<tr>
<td><strong>T simd_shuffle(T data, ushort simd_lane_id)</strong>&lt;br&gt;macOS: Since Metal 2.&lt;br&gt;iOS: Since Metal 2.2.</td>
<td>Returns <code>data</code> from the thread whose SIMD lane ID is <code>simd_lane_id</code>. The <code>simd_lane_id</code> needs to be a valid SIMD lane ID but doesn’t have to be the same for all threads in the SIMD-group.</td>
</tr>
<tr>
<td><strong>T simd_shuffle_and_fill_down(T data, T filling_data, ushort delta)</strong>&lt;br&gt;All OS: Since Metal 2.4.</td>
<td>Returns <code>data</code> or <code>filling_data</code> from the thread whose SIMD lane ID is the sum of the caller’s SIMD lane ID and <code>delta</code>. If the sum is greater than the SIMD-group size, the function copies values from the lower <code>delta</code> lanes of <code>filling_data</code> into the upper <code>delta</code> lanes of <code>data</code>. The value for <code>delta</code> needs to be the same for all threads in a SIMD-group.</td>
</tr>
<tr>
<td>Built-in SIMD-group Functions</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>T simd_shuffle_and_fill_down(T data, T filling_data, ushort delta, ushort modulo)</td>
<td>Returns data or filling data for each vector from the thread whose SIMD lane ID is the sum of the caller’s SIMD lane ID and delta. If the sum is greater than modulo, the function copies values from the lower delta lanes of filling_data into the upper delta lanes of data. The value of delta needs to be the same for all threads in a SIMD-group. The modulo parameter defines the vector width that splits the SIMD-group into separate vectors.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4.</td>
<td></td>
</tr>
<tr>
<td>T simd_shuffle_and_fill_up(T data, T filling_data, ushort delta)</td>
<td>Returns data or filling data from the thread whose SIMD lane ID is the difference from the caller’s SIMD lane ID minus delta. If the difference is negative, the operation copies values from the upper delta lanes of filling_data to the lower delta lanes of data. The value of delta needs to be the same for all threads in a SIMD-group.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4.</td>
<td></td>
</tr>
<tr>
<td>T simd_shuffle_and_fill_up(T data, T filling_data, ushort delta, ushort modulo)</td>
<td>Returns data or filling data for each vector from the thread whose SIMD lane ID is the difference from the caller’s SIMD lane ID minus delta. If the difference is negative, the operation copies values from the upper delta lanes of filling_data to the lower delta lanes of data. The value of delta needs to be the same for all threads in a SIMD-group. The modulo parameter defines the vector width that splits the SIMD-group into separate vectors.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4.</td>
<td></td>
</tr>
<tr>
<td>T simd_shuffle_down(T data, ushort delta)</td>
<td>Returns data from the thread whose SIMD lane ID is the sum of caller’s SIMD lane ID and delta. The value for delta needs to be the same for all threads in the SIMD-group. This function doesn’t modify the upper delta lanes of data because it doesn’t wrap values around the SIMD-group.</td>
</tr>
</tbody>
</table>
### Built-in SIMD-group Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T \text{ simd_shuffle_rotate_down}(T \text{ data}, \text{ ushort delta}) )</td>
<td>Returns data from the thread whose SIMD lane ID is the sum of caller’s SIMD lane ID and delta. The value for delta needs to be the same for all threads in the SIMD-group. This function wraps values around the SIMD-group.</td>
</tr>
<tr>
<td>( T \text{ simd_shuffle_rotate_up}(T \text{ data}, \text{ ushort delta}) )</td>
<td>Returns data from the thread whose SIMD lane ID is the difference from the caller’s SIMD lane ID minus delta. The value of delta needs to be the same for all threads in a SIMD-group. This function wraps values around the SIMD-group.</td>
</tr>
<tr>
<td>( T \text{ simd_shuffle_up}(T \text{ data}, \text{ ushort delta}) )</td>
<td>Returns data from the thread whose SIMD lane ID is the difference from the caller’s SIMD lane ID minus delta. The value of delta needs to be the same for all threads in a SIMD-group. This function doesn’t modify the lower delta lanes of data because it doesn’t wrap values around the SIMD-group.</td>
</tr>
<tr>
<td>( T \text{ i simd_shuffle_xor}(T \text{ i value}, \text{ ushort mask}) )</td>
<td>Returns data from the thread whose SIMD lane ID is equal to the bitwise XOR (^) of the caller’s SIMD lane ID and mask. The value of mask needs to be the same for all threads in a SIMD-group.</td>
</tr>
</tbody>
</table>

**Table 6.15. SIMD-group reduction functions in the Metal standard library**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T \text{ i simd_and}(T \text{ i data}) )</td>
<td>Returns the bitwise AND (&amp;) of data across all active threads in the SIMD-group and broadcasts the result to all active threads in the SIMD-group.</td>
</tr>
<tr>
<td>Built-in SIMD-group Functions</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>bool simd_is_helper_thread()</td>
<td>Returns <code>true</code> if the current thread is a helper thread; otherwise, <code>false</code>. You need to call this function from a fragment function or another function that your fragment function calls; otherwise, it may trigger a compile-time error.</td>
</tr>
<tr>
<td>T simd_max(T data)</td>
<td>Returns <code>data</code> with the highest value from across all active threads in the SIMD-group and broadcasts that value to all active threads in the SIMD-group.</td>
</tr>
<tr>
<td>T simd_min(T data)</td>
<td>Returns <code>data</code> with the lowest value from across all active threads in the SIMD-group and broadcasts that value to all active threads in the SIMD-group.</td>
</tr>
<tr>
<td>Ti simd_or(Ti data)</td>
<td>Returns the bitwise OR ((|)) of <code>data</code> across all active threads in the SIMD-group and broadcasts the result to all active threads in the SIMD-group.</td>
</tr>
<tr>
<td>T simd_prefix_exclusive_product(T data)</td>
<td>For a given thread, returns the product of the input values in <code>data</code> for all active threads with a lower index in the SIMD-group. The first thread in the group, returns (T(1)).</td>
</tr>
<tr>
<td>T simd_prefix_exclusive_sum (T data)</td>
<td>For a given thread, returns the sum of the input values in <code>data</code> for all active threads with a lower index in the SIMD-group. The first thread in the group, returns (T(0)).</td>
</tr>
<tr>
<td>T simd_prefix_inclusive_product(T data)</td>
<td>For a given thread, returns the product of the input values in <code>data</code> for all active threads with a lower or the same index in the SIMD-group.</td>
</tr>
<tr>
<td>T simd_prefix_inclusive_sum (T data)</td>
<td>For a given thread, returns the sum of the input values in <code>data</code> for all active threads with a lower or the same index in the SIMD-group.</td>
</tr>
</tbody>
</table>
### Built-in SIMD-group Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T simd_product(T data)</td>
<td>Returns the product of the input values in <code>data</code> across all active threads in the SIMD-group and broadcasts the result to all active threads in the SIMD-group.</td>
</tr>
<tr>
<td>T simd_sum(T data)</td>
<td>Returns the sum of the input values in <code>data</code> across all active threads in the SIMD-group and broadcasts the result to all active threads in the SIMD-group.</td>
</tr>
<tr>
<td>Ti simd_xor(Ti data)</td>
<td>Returns the bitwise XOR (^) of <code>data</code> across all active threads in the SIMD-group and broadcasts the result to all active threads in the SIMD-group.</td>
</tr>
</tbody>
</table>

### 6.9.2.1 Examples

To demonstrate the shuffle functions, start with this SIMD-group's initial state:

<table>
<thead>
<tr>
<th>SIMD Lane ID</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>i</td>
<td>j</td>
<td>k</td>
<td>l</td>
<td>m</td>
<td>n</td>
<td>o</td>
<td>p</td>
</tr>
</tbody>
</table>

The `simd_shuffle_up()` function shifts each SIMD-group upward by `delta` threads. For example, with a `delta` value of 2, the function:
- Shifts the SIMD lane IDs down by two
- Marks the lower two lanes as invalid

<table>
<thead>
<tr>
<th>Computed SIMD Lane ID</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>valid</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>data</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>i</td>
<td>j</td>
<td>k</td>
<td>l</td>
<td>m</td>
<td>n</td>
</tr>
</tbody>
</table>

The `simd_shuffle_up()` function is a non-wrapping operation that doesn't affect the lower `delta` lanes.

Similarly `simd_shuffle_down()` function shifts each SIMD-group downward by the `delta` threads. Starting with the same initial SIMD-group state, with a `delta` value of 2, the function:
- Shifts the SIMD lane IDs up by two
- Marks the upper two lanes as invalid

<table>
<thead>
<tr>
<th>Computed SIMD Lane ID</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>valid</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>i</td>
<td>j</td>
<td>k</td>
<td>l</td>
<td>m</td>
<td>n</td>
<td>o</td>
<td>p</td>
<td>o</td>
<td>p</td>
</tr>
</tbody>
</table>
The `simd_shuffle_down()` function is a non-wrapping operation that doesn’t affect the upper delta lanes.

To demonstrate the shuffle-and-fill functions, start this SIMD-group’s initial state:

<table>
<thead>
<tr>
<th>SIMD Lane ID</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
<td>y</td>
<td>z</td>
</tr>
<tr>
<td>filling</td>
<td>fa</td>
<td>fb</td>
<td>fc</td>
<td>fd</td>
<td>fe</td>
<td>ff</td>
<td>fg</td>
<td>fh</td>
<td>fs</td>
<td>ft</td>
<td>fu</td>
<td>fv</td>
<td>fw</td>
<td>fx</td>
<td>fy</td>
<td>fz</td>
</tr>
</tbody>
</table>

The `simd_shuffle_and_fill_up()` function shifts each SIMD-group upward by delta threads — similar to `simd_shuffle_up()` — and assigns the values from the upper filling lanes to the lower data lanes by wrapping the SIMD lane IDs. For example, with a delta value of 2, the function:

- Shifts the SIMD lane IDs down by two
- Assigns the upper two lanes of `filling` to the lower two lanes of `data`

<table>
<thead>
<tr>
<th>Computed SIMD Lane ID</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>fy</td>
<td>fz</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
</tr>
</tbody>
</table>

The `simd_shuffle_and_fill_up()` function with the modulo parameter splits the SIMD-group into vectors, each with size modulo, and shifts each vector by the delta threads. For example, with a modulo value of 8 and a delta value of 2, the function:

- Shifts the SIMD lane IDs down by two
- Assigns the upper two lanes of each vector in `filling` to the lower two lanes of each vector in `data`

<table>
<thead>
<tr>
<th>Computed SIMD Lane ID</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>fg</td>
<td>fh</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>fy</td>
<td>fz</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
</tr>
</tbody>
</table>

The `simd_shuffle_and_fill_down()` function shifts each SIMD-group downward by delta threads — similar to `simd_shuffle_down()` — and assigns the values from the lower filling lanes to the upper data lanes by wrapping the SIMD lane IDs. For example, with a delta value of 2, the function:

- Shifts the SIMD lane IDs up by two
- Assigns the lower two lanes of `filling` to the upper two lanes of `data`

<table>
<thead>
<tr>
<th>Computed SIMD Lane ID</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td>fa</td>
<td>fb</td>
</tr>
</tbody>
</table>

The `simd_shuffle_and_fill_down()` function with the modulo parameter splits the SIMD-group into vectors, each with size modulo and shifts each vector by the delta threads. For example, with a modulo value of 8 and a delta value of 2, the function:

- Shifts the SIMD lane IDs up by two
• Assigns the lower two lanes of each vector in `filling` to the upper two lanes of each vector in `data`

<table>
<thead>
<tr>
<th>Computed SIMD Lane ID</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>fa</td>
<td>fb</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td>fs</td>
<td>ft</td>
</tr>
</tbody>
</table>

Below is an example of how to use these SIMD functions to perform a reduction operation:

```c
kernel void
reduce(const device int *input [[buffer(0)]],
    device atomic_int *output [[buffer(1)]],
    threadgroup int *ldata [[threadgroup(0)]],
    uint gid [[thread_position_in_grid]],
    uint lid [[thread_position_in_threadgroup]],
    uint lsize [[threads_per_threadgroup]],
    uint simd_size [[threads_per_simdgroup]],
    uint simd_lane_id [[thread_index_in_simdgroup]],
    uint simd_group_id [[simdgroup_index_in_threadgroup]])
{
    // Perform the first level of reduction.
    // Read from device memory, write to threadgroup memory.
    int val = input[gid] + input[gid + lsize];
    for (uint s=lsize/simd_size; s>simd_size; s/=simd_size)
    {
        // Perform per-SIMD partial reduction.
        for (uint offset=simd_size/2; offset>0; offset/=2)
            val += simd_shuffle_down(val, offset);
        // Write per-SIMD partial reduction value to threadgroup memory.
        if (simd_lane_id == 0)
            ldata[simd_group_id] = val;
        // Wait for all partial reductions to complete.
        threadgroup_barrier(mem_flags::mem_threadgroup);
        val = (lid < s) ? ldata[lid] : 0;
    }
    // Perform final per-SIMD partial reduction to calculate
    // the threadgroup partial reduction result.
    for (uint offset=simd_size/2; offset>0; offset/=2)
        val += simd_shuffle_down(val, offset);
    // Atomically update the reduction result.
    if (lid == 0)
        atomic_fetch_add_explicit(output, val,
            memory_order_relaxed);
}
```
The `simd_active_threads_mask` and `simd_ballot` function uses the `simd_vote` wrapper type (see below), which can be explicitly cast to its underlying type represented by `vote_t`.

class simd_vote {
public:
    explicit constexpr simd_vote(vote_t v = 0);
    explicit constexpr operator vote_t() const;

    // Returns true if all bits corresponding to threads in the
    // SIMD-group are set.
    // You can use all() with the return value of simd_ballot(expr)
    // to determine if all threads are active.
    bool all() const;

    // Returns true if any bit corresponding to a valid thread in
    // the SIMD-group is set.
    // You can use any() with the return value of simd_ballot(expr)
    // to determine if at least one thread is active.
    bool any() const;

private:
    // bit i in v represents the 'vote' for the thread in the
    // SIMD-group at index i
    uint64_t v;
};

Note that `simd_all(expr)` is different from `simd_ballot(expr).all()`:

- `simd_all(expr)` returns true if all active threads evaluate `expr` to true.
- `simd_ballot(expr).all()` returns true if all threads were active and evaluated `expr` to true. (`simd_vote::all()` does not look at which threads are active.)

The same logic applies to `simd_any`, `simd_vote::any()`, and to the equivalent quad functions listed in section 6.9.3.

On hardware with fewer than 64 threads in a SIMD-group, the value of the top bits in `simd_vote` is undefined. In particular, since you can initialize these bits, do not assume that the top bits are set to 0.

6.9.3 Quad-group Functions

macOS: Quad-group functions supported since Metal 2.1.

iOS: Some quad-group functions supported since Metal 2, including `quad_broadcast`, `quad_shuffle`, `quad_shuffle_up`, `quad_shuffle_down`, and `quad_shuffle_xor`.

A quad-group function is a SIMD-group function (see section 6.9.2) with an execution width of 4. The active and inactive thread terminology is the same as in section 6.9.2.
Helper threads only execute to compute gradients for quad-groups in a fragment shader and then become inactive.

Kernels and fragment functions can call the quad-group functions listed in Table 6.16 and Table 6.17. Threads may only read data from another active thread in a quad-group. See the [Metal Feature Set Tables](https://developer.apple.com/metal/) to determine which GPUs support each table.

The placeholder `T` for Table 6.16 and Table 6.17 represents a scalar or vector of any integer or floating-point type, except:

- `bool`
- `void`
- `size_t`
- `ptrdiff_t`

For bitwise operations, `T` needs to be an integer scalar or vector.

### Table 6.16. Quad-group permute functions in the Metal standard library

<table>
<thead>
<tr>
<th>Built-in Quad-group Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>quad_vote</code> <code>quad_ballot (bool expr)</code></td>
<td>Returns a wrapper type — see the <a href="https://developer.apple.com/metal/">quad_vote example</a> — around a bitmask of the evaluation of the Boolean expression for all active threads in the quad-group for which <code>expr</code> is true. The function sets the bits that correspond to inactive threads to 0.</td>
</tr>
<tr>
<td><code>T quad_broadcast(T data, ushort broadcast_lane_id)</code></td>
<td>Broadcasts data from the thread whose quad lane ID is <code>broadcast_lane_id</code>. The value for <code>broadcast_lane_id</code> needs to be a valid quad lane ID that’s the same for all threads in a quad-group.</td>
</tr>
<tr>
<td><code>T quad_broadcast_first(T data)</code></td>
<td>Broadcasts data from the first active thread — the active thread with the smallest index — in the quad-group to all active threads.</td>
</tr>
<tr>
<td><code>T quad_shuffle(T data, ushort quad_lane_id)</code></td>
<td>Returns data from the thread whose quad lane ID is the sum of the caller’s quad lane ID and <code>delta</code>. The value for <code>quad_lane_id</code> needs to be a valid lane ID and may differ from other threads in the quad-group.</td>
</tr>
</tbody>
</table>

2023-06-02 | Copyright © 2023 Apple Inc. | All Rights Reserved.

Page 181 of 287
<table>
<thead>
<tr>
<th>Built-in Quad-group Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T quad_shuffle_and_fill_down(T data, T filling_data, ushort delta)</td>
<td>Returns data or filling_data from the thread whose quad lane ID is the sum of the caller's quad lane ID and delta. If the sum is greater than the quad-group size, the function copies values from the lower delta lanes of filling_data into the upper delta lanes of data. The value for delta needs to be the same for all threads in a quad-group.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4</td>
<td></td>
</tr>
<tr>
<td>T quad_shuffle_and_fill_down(T data, T filling_data, ushort delta, ushort modulo)</td>
<td>Returns data or filling_data for each vector, from the thread whose quad lane ID is the sum of caller's quad lane ID and delta. If the sum is greater than the quad-group size, the function copies values from the lower delta lanes of filling_data into the upper delta lanes of data. The value of delta needs to be the same for all threads in a quad-group. The modulo parameter defines the vector width that splits the quad-group into separate vectors.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4</td>
<td></td>
</tr>
<tr>
<td>T quad_shuffle_and_fill_up(T data, T filling_data, ushort delta)</td>
<td>Returns data or filling_data from the thread whose quad lane ID is the difference from the caller's quad lane ID minus delta. If the difference is negative, the operation copies values from the upper delta lanes of filling_data to the lower delta lanes of data. If the difference is negative, the function shuffles data from filling_data into the lower delta lanes. The value of delta needs to be the same for all threads in a quad-group.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4</td>
<td></td>
</tr>
<tr>
<td>T quad_shuffle_and_fill_up(T data, T filling_data, ushort delta, ushort modulo)</td>
<td>Returns data or filling_data for each vector from the thread whose quad lane ID is the difference from the caller's quad lane ID minus delta. If the difference is negative, the operation copies values from the upper delta lanes of filling_data to the lower delta lanes of data. The value of delta needs to be the same for all threads in a quad-group. The modulo parameter defines the width that splits the quad-group into separate vectors.</td>
</tr>
<tr>
<td>All OS: Since Metal 2.4</td>
<td></td>
</tr>
<tr>
<td>Built-in Quad-group Functions</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>T quad_shuffle_down(T data, ushort delta)</td>
<td>Returns data from the thread whose quad lane ID is the sum of the caller’s quad lane ID and delta. The value for delta needs to be the same for all threads in a quad-group. The function doesn’t modify the upper delta lanes of data because it doesn’t wrap values around the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.</td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.</td>
<td></td>
</tr>
<tr>
<td>T quad_shuffle_rotate_down(T data, ushort delta)</td>
<td>Returns data from the thread whose quad lane ID is the sum of the caller’s quad lane ID and delta. The value for delta needs to be the same for all threads in a quad-group. This function wraps values around the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.2.</td>
<td></td>
</tr>
<tr>
<td>T quad_shuffle_rotate_up(T data, ushort delta)</td>
<td>Returns data from the thread whose quad lane ID is the difference from the caller’s quad lane ID minus delta. The value for delta needs to be the same for all threads in a quad-group. This function wraps values around the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1.</td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.2.</td>
<td></td>
</tr>
<tr>
<td>T quad_shuffle_up(T data, ushort delta)</td>
<td>Returns data from thread whose quad lane ID is the difference from the caller’s quad lane ID minus delta. The value for delta needs to be the same for all threads in a quad-group. This function doesn’t modify the lower delta lanes of data because it doesn’t wrap values around the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.</td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.</td>
<td></td>
</tr>
<tr>
<td>T quad_shuffle_xor(T value, ushort mask)</td>
<td>Returns data from the thread whose quad lane ID is a bitwise XOR (^) of the caller’s quad lane ID and mask. The value of mask needs to be the same for all threads in a quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.</td>
<td></td>
</tr>
<tr>
<td>iOS: Since Metal 2.</td>
<td></td>
</tr>
</tbody>
</table>
### Table 6.17. Quad-group reduction functions in the Metal standard library

<table>
<thead>
<tr>
<th>Built-in Quad-group Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>quad_vote quad_active_threads_mask()</td>
<td>Returns a quad_vote mask that represents the active threads. The function is equivalent to \texttt{quad_ballot(true)} and sets the bits that represent active threads to 1 and inactive threads to 0.</td>
</tr>
<tr>
<td>bool quad_all(bool expr)</td>
<td>Returns true if all active threads evaluate \texttt{expr} to true.</td>
</tr>
<tr>
<td>T quad_and(T data)</td>
<td>Returns the bitwise AND (&amp;) of data across all active threads in the quad-group and broadcasts the result to all active threads in the quad-group.</td>
</tr>
<tr>
<td>bool quad_any(bool expr)</td>
<td>Returns true if at least one active thread evaluates \texttt{expr} to true.</td>
</tr>
<tr>
<td>bool quad_is_first()</td>
<td>Returns true if the current thread is the first active thread — the active thread with the smallest index — in the current quad-group; otherwise false.</td>
</tr>
<tr>
<td>bool quad_is_helper_thread()</td>
<td>Returns true if the current thread is a helper thread; otherwise, false. You need to call this function from a fragment function or another function that your fragment function calls; otherwise, it may trigger a compile-time error.</td>
</tr>
<tr>
<td>T quad_max(T data)</td>
<td>Returns data with the highest value from across all active threads in the quad-group and broadcasts that value to all active threads in the quad-group.</td>
</tr>
<tr>
<td>T quad_min(T data)</td>
<td>Returns data with the lowest value from across all active threads in the quad-group and broadcasts that value to all active threads in the quad-group.</td>
</tr>
<tr>
<td>Built-in Quad-group Functions</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>$\text{T quad_or(T data)}$</td>
<td>Returns the bitwise OR (</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1. iOS: Since Metal 2.3.</td>
<td></td>
</tr>
<tr>
<td>$\text{T quad_prefix_exclusive_product (T data)}$</td>
<td>For a given thread, returns the product of the input values in data for all active threads with a lower index in the quad-group. For the first thread in the group, return $\text{T}(1)$.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1. iOS: Since Metal 2.3.</td>
<td></td>
</tr>
<tr>
<td>$\text{T quad_prefix_exclusive_sum (T data)}$</td>
<td>For a given thread, returns the sum of the input values in data for all active threads with a lower index in the quad-group. For the first thread in the group, return $\text{T}(0)$.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1. iOS: Since Metal 2.3.</td>
<td></td>
</tr>
<tr>
<td>$\text{T quad_prefix_inclusive_product (T data)}$</td>
<td>For a given thread, returns the product of the input values in data for all active threads with a lower or the same index in the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1. iOS: Since Metal 2.3.</td>
<td></td>
</tr>
<tr>
<td>$\text{T quad_prefix_inclusive_sum (T data)}$</td>
<td>For a given thread, returns the sum of the input values in data for all active threads with a lower or the same index in the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1. iOS: Since Metal 2.3.</td>
<td></td>
</tr>
<tr>
<td>$\text{T quad_product(T data)}$</td>
<td>Returns the product of the input values in data across all active threads in the quad-group and broadcasts the result to all active threads in the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1. iOS: Since Metal 2.3.</td>
<td></td>
</tr>
<tr>
<td>$\text{T quad_sum(T data)}$</td>
<td>Returns the sum of the input values in data across all active threads in the quad-group and broadcasts the result to all active threads in the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1. iOS: Since Metal 2.3.</td>
<td></td>
</tr>
<tr>
<td>$\text{T quad_xor(T data)}$</td>
<td>Returns the bitwise XOR (^) of data across all active threads in the quad-group and broadcasts the result to all active threads in the quad-group.</td>
</tr>
<tr>
<td>macOS: Since Metal 2.1. iOS: Since Metal 2.3.</td>
<td></td>
</tr>
</tbody>
</table>

In a kernel function, quads divide across the SIMD-group. In a fragment function, the lane ID represents the fragment location in a 2 x 2 quad:

- Lane ID 0 is the upper-left pixel
- Lane ID 1 is the upper-right pixel
• Lane ID 2 is the lower-left pixel
• Lane ID 3 is the lower-right pixel

To demonstrate the shuffle functions, start with this quad-group’s initial state:

<table>
<thead>
<tr>
<th>Quad Lane ID</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

The quad_shuffle_up() function shifts each quad-group upward by delta threads. For example, with a delta value of 2, the function:

• Shifts the quad lane IDs down by two
• Marks the lower two lanes as invalid

<table>
<thead>
<tr>
<th>Computed Quad Lane ID</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>valid</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>data</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>b</td>
</tr>
</tbody>
</table>

The quad_shuffle_up() function is a non-wrapping operation that doesn’t affect the lower delta lanes.

Similarly, quad_shuffle_down() function shifts each quad-group downward by delta threads. Starting with the same initial quad-group state, with a delta of 2, the function:

• Shifts the quad lane IDs up by two
• Marks the upper two lanes as invalid

<table>
<thead>
<tr>
<th>Computed Quad Lane ID</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>valid</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data</td>
<td>c</td>
<td>d</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

The quad_shuffle_down() function is a non-wrapping operation that doesn’t affect the upper delta lanes.

To demonstrate the shuffle-and-fill functions, start this quad-group’s initial state:

<table>
<thead>
<tr>
<th>Quad Lane ID</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>filling</td>
<td>fa</td>
<td>fb</td>
<td>fc</td>
<td>fd</td>
</tr>
</tbody>
</table>

The quad_shuffle_and_fill_up() function shifts each quad-group upward by the delta threads — similar to quad_shuffle_up() — and assigns the values from the upper filling lanes to the lower data lanes by wrapping the quad lane IDs. For example, with a delta value of 2, the function:

• Shifts the quad lane IDs down by two
• Assigns the upper two lanes of \textit{filling} to the lower two lanes of \textit{data}

<table>
<thead>
<tr>
<th>Computed Quad Lane ID</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>fc</td>
<td>fd</td>
<td>a</td>
<td>b</td>
</tr>
</tbody>
</table>

The \texttt{quad_shuffle_and_fill_up()} function with the \texttt{modulo} parameter splits the quad-group into vectors, each with size \texttt{modulo} and shifts each vector by the \texttt{delta} threads. For example, with a \texttt{modulo} value of 2 and a \texttt{delta} value of 1, the function:

• Shifts the quad lane IDs down by one
• Assigns the upper lane of each vector in \textit{filling} to the lower lane of each vector in \textit{data}

<table>
<thead>
<tr>
<th>Computed Quad Lane ID</th>
<th>-1</th>
<th>0</th>
<th>-1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>fb</td>
<td>a</td>
<td>fd</td>
<td>c</td>
</tr>
</tbody>
</table>

The \texttt{quad_shuffle_and_fill_down()} function shifts each quad-group downward by \texttt{delta threads} — similar to \texttt{quad_shuffle_down()} — and assigns the values from the lower \texttt{filling} lanes to the upper \texttt{data} lanes by wrapping the quad lane IDs. For example, with a \texttt{delta} value of 2, the function:

• Shifts the quad lane IDs up by two
• Assigns the lower two lanes of \textit{filling} to the upper two lanes of \textit{data}

<table>
<thead>
<tr>
<th>Computed Quad Lane ID</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>c</td>
<td>d</td>
<td>fa</td>
<td>fb</td>
</tr>
</tbody>
</table>

The \texttt{quad_shuffle_and_fill_down()} function with the \texttt{modulo} parameter splits the quad-group into vectors, each with size \texttt{modulo} and shifts each vector by the \texttt{delta} threads. For example, with a \texttt{modulo} value of 2 and a \texttt{delta} value of 1, the function:

• Shifts the quad lane IDs up by one
• Assigns the lower lane of each vector in \textit{filling} to the upper lane of each vector in \textit{data}

<table>
<thead>
<tr>
<th>Computed Quad Lane ID</th>
<th>1</th>
<th>2</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>b</td>
<td>fa</td>
<td>d</td>
<td>fc</td>
</tr>
</tbody>
</table>

The \texttt{quad_ballot} function uses the \texttt{quad_vote} wrapper type, which can be explicitly cast to its underlying type. (In the following example, note use of \texttt{vote_t} to represent an underlying type, XXX.)
class quad_vote {
public:
    typedef XXX vote_t;
    explicit constexpr quad_vote(vote_t v = 0);
    explicit constexpr operator vote_t() const;

    // Returns true if all bits corresponding to threads in the
    // quad-group (the four bottom bits) are set.
    bool all() const;

    // Returns true if any bit corresponding to a thread in the
    // quad-Group is set.
    bool any() const;
};

The quad_vote constructor masks out the top bits (that is, other than the four bottom bits). Hence, the non-bottom-four bits are guaranteed to be unset when cast to vote_t.

6.10 Graphics Functions

The graphics functions in this section and its subsections are defined in the header <metal_graphics>. You can only call these graphics functions from a vertex function or a fragment function.

6.10.1 Fragment Functions

You can only call the functions in this section (listed in Table 6.18, Table 6.19, and Table 6.20) inside a fragment function (see section 5.1.2) or inside a function called from a fragment function. Otherwise the behavior is undefined and may result in a compile-time error.

Fragment function helper threads may be created to help evaluate derivatives (explicit or implicit) for use with a fragment thread(s). Fragment function helper threads execute the same code as the non-helper fragment threads, but do not have side effects that modify the render target(s) or any other memory that can be accessed by the fragment function. In particular:

- Fragments corresponding to helper threads are discarded when the fragment function execution is complete without any updates to the render target(s).
- Stores and atomic operations to buffers and textures performed by helper threads have no effect on the underlying memory associated with the buffer or texture.

6.10.1.1 Fragment Functions – Derivatives

Metal includes the functions in Table 6.18 to compute derivatives. T is one of float, float2, float3, float4, half, half2, half3, or half4.

Derivatives are undefined within non-uniform control flow.
Note: Prior to Metal 2.3, discard_fragment could make the control flow non-uniform. Since Metal 2.3, discard_fragment does not affect whether the control flow is considered non-uniform or not. See Section 6.10.1.3 for more information.

<table>
<thead>
<tr>
<th>Built-in fragment functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T dfdx(T p)</td>
<td>Returns a high precision partial derivative of the specified value with respect to the screen space x coordinate.</td>
</tr>
<tr>
<td>T dfdy(T p)</td>
<td>Returns a high precision partial derivative of the specified value with respect to the screen space y coordinate.</td>
</tr>
<tr>
<td>T fwidth(T p)</td>
<td>Returns the sum of the absolute derivatives in x and y using local differencing for p; that is, fabs(dfdx(p)) + fabs(dfdy(p))</td>
</tr>
</tbody>
</table>

6.10.1.2 Fragment Functions – Samples

Metal includes the per-sample functions listed in Table 6.19. `get_num_samples` and `get_sample_position` return the number of samples for the color attachment and the sample offsets for a given sample index. For example, for transparency super-sampling, these functions can be used to shade per-fragment but do the alpha test per-sample.

<table>
<thead>
<tr>
<th>Built-in fragment functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>uint get_num_samples()</td>
<td>Returns the number of samples for the multisampled color attachment.</td>
</tr>
<tr>
<td>float2 get_sample_position(uint index)</td>
<td>Returns the normalized sample offset (x, y) for a given sample index index. Values of x and y are in [0.0 ... 1.0].</td>
</tr>
</tbody>
</table>

If you have customized sample positions (set with the `setSamplePositions:count:` method of `MTLRenderPassDescriptor`), `get_sample_position(index)` returns the position programmed for the specified index.

6.10.1.3 Fragment Functions – Flow Control

The Metal function in Table 6.20 terminates a fragment.
Table 6.20. Fragment flow control function in the Metal standard library

<table>
<thead>
<tr>
<th>Built-in fragment functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>void discard_fragment(void)</td>
<td>Marks the current fragment as terminated and discards this fragment's output of the fragment function.</td>
</tr>
</tbody>
</table>

Writes to a buffer or texture from a fragment thread made before calling `discard_fragment` are not discarded.

Multiple fragment threads or helper threads associated with a fragment thread execute together to compute derivatives. Prior to Metal 2.3, if any (but not all) of these threads executes the `discard_fragment` function, the thread is terminated and the behavior of any derivative computations (explicit or implicit) is undefined. Since Metal 2.3, `discard_fragment` marks the fragment as terminated while continuing to execute in parallel and has no effect on whether derivatives are defined. Even though execution continues, the write behavior remains the same as before. The fragment will discard the fragment output and discard all writes to buffer or texture after `discard_fragment`.

### 6.11 Pull-Model interpolation

All OS: Pull-model interpolation is supported since Metal 2.3.

The interpolant type `interpolant<T,P>` (section 2.18) and associated methods are defined in `<metal_interpolate>`. In a fragment function, you explicitly interpolate the values of a `interpolant<T,P>` type by invoking its methods, as shown below. The interpolant may be sampled and interpolated multiple times, in different modes, and may be passed to other functions to be sampled and interpolated there. Perspective correctness is fixed across all interpolations of the argument by the value of `P` in its type.

<table>
<thead>
<tr>
<th>Interpolant method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T interpolate_at_center()</td>
<td>Sample shader input at the center of a pixel, returning the same value as if the input had type <code>T</code> with <code>[[center_perspective]]</code> or <code>[[center_no_perspective]]</code>.</td>
</tr>
<tr>
<td>T interpolate_at_centroid()</td>
<td>Sample shader input within the covered area of the pixel, returning the same value as if the input had type <code>T</code> with <code>[[centroid_perspective]]</code> or <code>[[centroid_no_perspective]]</code>.</td>
</tr>
<tr>
<td>T interpolate_at_offset(float2 offset)</td>
<td>Sample shader input at a specified window-coordinate offset from a pixel's top-left corner.</td>
</tr>
</tbody>
</table>
Allowable offset components are in the range \([0.0, 1.0)\) along a 1/16 pixel grid.

| \(T\) interpolate_at_sample(uint sample) | Sample shader input at the location of the specified sample index, returning the same value as if the input had type \(T\) with [[sample_perspective]] or [[sample_no_perspective]] and was in the specified per-sample evaluation of the shader. If a sample of the given index does not exist, the position of interpolation is undefined. |

6.12 Texture Functions

The texture member functions, defined in the header `<metal_texture>`, listed in this section and its subsections for different texture types include:

- `sample` - `sample` from a texture,
- `sample_compare` - `sample compare` from a texture,
- `gather` - `gather` from a texture,
- `gather_compare` - `gather compare` from a texture,
- `read` - `sampler-less read` from a texture,
- `write` - `write` to a texture,
- `texture query` (such as `get_width`, `get_height`, `get_num_mip_levels`, `get_array_size`), and
- `texture fence`.

Metal 3.1 introduces new atomic texture member functions supported on 1D texture, 1D texture array, 2D texture, 2D texture array, 3D texture, and texture buffer for `int`, `uint`, `ulong` color types:

- `atomic_load` - `atomic load` from a texture,
- `atomic_store` - `atomic store` to a texture,
- `atomic_exchange` - `atomic exchange` a value for a texture,
- `atomic_compare_exchange_weak` - `atomic compare and exchange` in a texture,
- `atomic_fetch_op_explicit` - `atomic fetch` and modify where `op` can be `add`, and, `max`, `min`, `or`, `sub`, or `xor` for `int` and `uint` color type.
- `atomic_max` - `atomic max` in a texture for `ulong` color type.
- `atomic_min` - `atomic min` in a texture for `ulong` color type.

The `texture sample, sample_compare, gather, and gather_compare` functions take an offset argument for a 2D texture, 2D texture array, and 3D texture. The offset is an integer.
value applied to the texture coordinate before looking up each pixel. This integer value can be in the range -8 to +7. The default value is 0.

The texture sample, sample_compare, gather, and gather_compare functions require that you declare the texture with the sample access attribute. The texture read functions require that you declare the texture with the sample, read, or read_write access attribute. The texture write functions require that you declare the texture with the write or read_write access attribute. (For more about access attributes, see section 2.9.)

The texture sample_compare and gather_compare functions are only available for depth texture types.

compare_func sets the comparison test for the sample_compare and gather_compare functions. For more about compare_func, see section 2.10.

Overloaded variants of the texture sample and sample_compare functions with an lod_options argument are available for a 2D texture, 2D texture array, 2D depth texture array, 3D texture, cube texture, cube texture array, cube depth texture, and cube depth texture array. (LOD/lod is short for level-of-detail.) The values for lod_options are:

- level(float lod) - sample from the specified mipmap level
- bias(float value) - apply the specified bias to a mipmap level before sampling
- gradient*(T dPdx, T dPdy) - apply the specified gradients with respect to the x and y directions. The texture type changes the name and the arguments; for example, for 3D textures, the name is gradient3d and the arguments are float3 type.
- min_lod_clamp(float lod) - specify lowest mipmap level for sampler access, which restricts sampler access to a range of mipmap levels. (All OS: Support since Metal 2.2.)

On macOS up to Metal 2.2, for sample_compare functions, bias and gradient* are not supported, and lod needs to be a zero constant. In Metal 2.3, this restriction is lifted for some hardware.

Since Metal 2.2, you can specify a LOD range for a sampler. You can either specify a minimum and maximum mipmap level or use min_lod_clamp to specify just the minimum mipmap level of an open range. When the sampler determines which mipmaps to sample, the selection is clamped to the specified range.

Clamping the LOD is useful where some of the texture data is not available all the time (for example, texture streaming). You can create a texture with all the necessary mipmaps and then can stream image data starting from the smallest mipmaps. When the GPU samples the texture, it clamps to the mipmaps that already have valid data. When you copy larger mipmaps into the texture, you reduce the minimum LOD level. As new data becomes ready, you can change the LOD clamp, which changes the sampling resolution.

The texture sample and sample_compare functions that do not take an explicit LOD or gradients have a default LOD of 0. The gather and gather_compare functions called from kernel or vertex functions also have a default LOD of 0.

For the gather and gather_compare functions, place the four samples that contribute to filtering into x y z w components in counter-clockwise order, starting with the sample to the lower-left of the queried location. This is the same as nearest sampling with unnormalized texture coordinate deltas at the following locations: (-,+), (+,+), (+,-),(-,-), where the magnitude of the deltas are always half a pixel.
A `read` from or `write` to a texture is out-of-bounds if and only if any of these conditions is met:

- the coordinates accessed are out-of-bounds,
- the level of detail argument is out-of-bounds,
- the texture is a texture array (`texture?d_array` type), and the array slice argument is out-of-bounds,
- the texture is a `texturecube` or `texturecube_array` type, and the face argument is out-of-bounds, or
- the texture is a multisampled texture, and the sample argument is out-of-bounds.

For all texture types, an out-of-bounds `write` to a texture is ignored.

For all texture types:

- for components specified in a pixel format, an out-of-bounds `read` returns a color with components with the value zero.
- for components unspecified in a pixel format, an out-of-bounds `read` returns the default value.

For unspecified color components in a pixel format, the default values are:

- zero, for components other than alpha.
- one, for the alpha component.

In a pixel format with integer components, the alpha default value is represented as the integral value 0x1. For a pixel format with floating-point or normalized components, the alpha default value is represented as the floating-point value 1.0.

For example, for a texture with the `MTLPixelFormatR8Uint` pixel format, the default values for unspecified integer components are G = 0, B = 0, and A = 1. For a texture with the `MTLPixelFormatR8Unorm` pixel format, the default values for unspecified normalized components are G = 0.0, B = 0.0, and A = 1.0. For a texture with depth or stencil pixel format (such as `MTLPixelFormatDepth24Unorm_Stencil18` or `MTLPixelFormatStencil18`), the default value for an unspecified component is undefined.

On macOS up to Metal 2.2, for texture `write` functions, `lod` needs to be 0. In Metal 2.3, this restriction is lifted for some hardware.

The following texture member functions are available to support sparse textures:

- macOS: sparse texture functions supported since Metal 2.3.
- iOS: sparse texture functions supported since Metal 2.2.

- `sparse_sample` - sample from a sparse texture,
- `sparse_sample_compare` - sample compare from a sparse texture,
- `sparse_gather` - gather from a sparse texture, and
- `sparse_gather_compare` - gather compare from a sparse texture.

These sparse texture member functions return a `sparse_color` structure that contains one or more color values and a residency flag. If any of the accessed pixels is not mapped, `resident` is set to `false`. 
template <typename T>
struct sparse_color {
public:
  constexpr sparse_color(T value, bool resident) thread;

  // Indicates whether all memory addressed to retrieve the value was mapped.
  constexpr bool resident() const thread;

  // Retrieve the color value.
  constexpr T const value() const thread;
};

For a sparse texture, to specify the minimum LOD level that the sampler can access, use min_lod_clamp.

Note:
For sections 6.12.1 through 6.12.16, the following abbreviations are used for the data types of function arguments and return values:

Tv denotes a 4-component vector type based on the templated type <T> for declaring the texture type:

- If T is float, Tv is float4.
- If T is half, Tv is half4.
- If T is int, Tv is int4.
- If T is uint, Tv is uint4.
- If T is short, Tv is short4.
- If T is ushort, Tv is ushort4.
- If T is ulong, Tv is ulong4  (since Metal 3.1)

Metal does not support sampling of textures when T is ulong. Note that not all operations are supported on all types.

Since Metal 3.1, texture support atomic functions for element T where T is int, uint, or ulong:

- When the element T is int or uint, the texture on the Metal needs to be either MTLPixelFormatR32Uint, or MTLPixelFormatR32Sint,
- When the element T is ulong, the texture on the Metal needs to be MTLPixelFormatRG32Uint.

The semantics of the atomic texture functions are the same as the atomic functions defined in Sec 6.15.

sparse_color-Tv denotes a sparse_color structure that contains a 4-component vector of color values, based on the templated type <T>, and a residency flag. These represent the return values of many sparse texture member functions.
sparse_color-T denotes a sparse_color structure that contains a single value, based on the templated type <T>, and a residency flag. T typically represents a depth value that a sparse texture member function returns.

The following functions can be used to return the LOD (mip level) computation result for a simulated texture fetch:

macOS: sparse texture functions supported since Metal 2.2.
iOS: sparse texture functions supported since Metal 2.3.

calculate_unclamped_lod - calculates the level of detail that would be sampled for the given coordinates, ignoring any sampler parameter. The fractional part of this value contains the mip level blending weights, even if the sampler indicates a nearest Mip selection.

calculate_clamped_lod - similar to the calculate_unclamped_lod, but additionally clamps the LOD to stay:
- within the texture mip count limits,
- within the sampler's lod_clamp min and max values
- less than or equal to the sampler's max_anisotropy value

6.12.1 1D Texture

This member function samples from a 1D texture.

Tv sample(sampler s, float coord) const

These member functions perform sampler-less reads from a 1D texture. Since mipmaps are not supported for 1D textures, lod needs to be 0.

Tv read(uint coord, uint lod = 0) const
Tv read(ushort coord, ushort lod = 0) const // All OS: Since Metal 1.2.

These member functions can write to a 1D texture. Since mipmaps are not supported for 1D textures, lod needs to be 0.

void write(Tv color, uint coord, uint lod = 0)
void write(Tv color, ushort coord,
          ushort lod = 0) // All OS: Since Metal 1.2.

These member functions query a 1D texture. Since mipmaps are not supported for 1D textures, get_num_mip_levels() always return 0, and lod needs to be 0 for get_width():

uint get_width(uint lod = 0) const
uint get_num_mip_levels() const

This member function samples from a sparse 1D texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_sample(sampler s, float coord) const

These member functions perform a sampler-less read from a sparse 1D texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS. Since mipmaps are not supported for 1D textures, lod needs to be 0.
sparse_color-Tv sparse_read(ushort coord, ushort lod = 0) const
sparse_color-Tv sparse_read(uint coord, uint lod = 0) const

These member functions execute an atomic load from a 1D texture starting with Metal 3.1:
Tv atomic_load(uint coord) const
Tv atomic_load(ushort coord) const

These member functions execute an atomic store to a 1D texture starting with Metal 3.1:
void atomic_store(Tv color, uint coord) const
void atomic_store (Tv color, ushort coord) const

These member functions execute an atomic compare and exchange to a 1D texture starting with Metal 3.1:
bool atomic_compare_exchange_weak(uint coord, thread Tv *expected, Tv desired) const
bool atomic_compare_exchange_weak(ushort coord, thread Tv *expected, Tv desired) const

These member functions execute an atomic exchange to a 1D texture starting with Metal 3.1:
Tv atomic_exchange(uint coord, Tv desired) const
Tv atomic_exchange(ushort coord, Tv desired) const

These member functions execute an atomic fetch and modify to a 1D texture starting with Metal 3.1 where op is add, and, max, min, or, sub, or xor for int and uint color type:
Tv atomic_fetch_op(uint coord, Tv operand)
Tv atomic_fetch_op(ushort coord, Tv operand) const

These member functions execute an atomic min or max to a 1D texture starting with Metal 3.1:
void atomic_min(uint coord, ulong4 operand)
void atomic_min(ushort coord, ulong4 operand)
void atomic_max(uint coord, ulong4 operand)
void atomic_max(ushort coord, ulong4 operand)

6.12.2 1D Texture Array

This member function samples from a 1D texture array:
Tv sample(sampler s, float coord, uint array) const

These member functions perform sampler-less reads from a 1D texture array. Since mipmaps are not supported for 1D textures, lod must be 0.
Tv read(uint coord, uint array, uint lod = 0) const
Tv read(ushort coord, ushort array, ushort lod = 0) const // All OS: Since Metal 1.2.

These member functions write to a 1D texture array. Since mipmaps are not supported for 1D textures, Lod must be 0.
void write(Tv color, uint coord, uint array, uint lod = 0)
void write(Tv color, ushort coord, ushort array, ushort lod = 0) // All OS: Since Metal 1.2.

These member functions query a 1D texture array. Since mipmaps are not supported for 1D textures, get_num_mip_levels() always return 0, and lod must be 0 for get_width().
uint get_width(uint lod = 0) const
uint get_array_size() const
uint get_num_mip_levels() const

This function samples from a sparse 1D texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_sample(sampler s, float coord, uint array) const

These functions perform a sampler-less read from a sparse 1D texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS. Since mipmaps are not supported for 1D texture arrays, Lod must be 0.
sparse_color-Tv sparse_read(ushort coord, ushort array, ushort lod = 0) const
sparse_color-Tv sparse_read(uint coord, uint array, uint lod = 0) const
These member functions execute an atomic load from a 1D texture array starting with Metal 3.1:
Tv atomic_load(uint coord, uint array) const
Tv atomic_load(ushort coord, ushort array) const

These member functions execute an atomic store to a 1D texture array starting with Metal 3.1:
void atomic_store(Tv color, uint coord, uint array) const
void atomic_store (Tv color, ushort coord, ushort array) const

These member functions execute an atomic compare and exchange to a 1D texture array starting with Metal 3.1:
bool atomic_compare_exchange_weak(uint coord, uint array, thread Tv *expected,
                                     Tv desired) const
bool atomic_compare_exchange_weak(ushort coord, ushort array, thread Tv *expected,
                                     Tv desired) const

These member functions execute an atomic exchange to a 1D texture array starting with Metal 3.1:
Tv atomic_exchange(uint coord, uint array, Tv desired) const
Tv atomic_exchange(ushort coord, ushort array, Tv desired) const

These member functions execute an atomic fetch and modify to a 1D texture array starting with Metal 3.1 where op is add, and, max, min, or, sub, or xor:
Tv atomic_fetch_op(uint coord, uint array, Tv operand)
Tv atomic_fetch_op(ushort coord, ushort array, Tv operand) const

These member functions execute an atomic min or max to a 1D texture array starting with Metal 3.1:
void atomic_min(uint coord, uint array, ulong4 operand)
void atomic_min(ushort coord, ushort array, ulong4 operand)
void atomic_max(uint coord, uint array, ulong4 operand)
void atomic_max(ushort coord, ushort array, ulong4 operand)
6.12.3 2D Texture

For the functions in this section, the following data types and corresponding constructor functions can specify sampling options (lod_options):

- bias(float value)
- level(float lod)
- gradient2d(float2 dPdx, float2 dPdy)
- min_lod_clamp(float lod) // All OS: support since Metal 2.2

These member functions sample from a 2D texture:

```c
Tv sample(sampler s, float2 coord, int2 offset = int2(0)) const
Tv sample(sampler s, float2 coord, lod_options options, int2 offset = int2(0)) const
Tv sample(sampler s, float2 coord, bias bias_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
Tv sample(sampler s, float2 coord, gradient2d grad_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
```

These member functions perform sampler-less reads from a 2D texture:

```c
Tv read(uint2 coord, uint lod = 0) const
Tv read(ushort2 coord, ushort lod = 0) const // All OS: Since Metal 1.2.
```

These member functions write to a 2D texture. On macOS up to Metal 2.2, lod must be 0. In Metal 2.3, this restriction is lifted for some hardware.

```c
void write(Tv color, uint2 coord, uint lod = 0)
void write(Tv color, ushort2 coord, ushort lod = 0) // All OS: Since Metal 1.2.
```

This member functions gathers four samples for bilinear interpolation when sampling a 2D texture:

```c
eenum class component {x, y, z, w};
Tv gather(sampler s, float2 coord, int2 offset = int2(0), component c = component::x) const
```

These member functions query a 2D texture query:

```c
uint get_width(uint lod = 0) const
```
uint get_height(uint lod = 0) const
uint get_num_mip_levels() const

These member functions sample from a sparse 2D texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_sample(sampler s, float2 coord,
     int2 offset = int2(0)) const
sparse_color-Tv sparse_sample(sampler s, float2 coord, bias options,
     int2 offset = int2(0)) const
sparse_color-Tv sparse_sample(sampler s, float2 coord,
     level options,
     int2 offset = int2(0)) const
sparse_color-Tv sparse_sample(sampler s, float2 coord,
     min_lod_clamp min_lod_clamp_options,
     int2 offset = int2(0)) const
sparse_color-Tv sparse_sample(sampler s, float2 coord,
     bias bias_options,
     min_lod_clamp min_lod_clamp_options,
     int2 offset = int2(0)) const
sparse_color-Tv sparse_sample(sampler s, float2 coord,
     gradient2d grad_options,
     int2 offset = int2(0)) const
sparse_color-Tv sparse_sample(sampler s, float2 coord,
     gradient2d grad_options,
     min_lod_clamp min_lod_clamp_options,
     int2 offset = int2(0)) const

These member functions perform a sampler-less read from a sparse 2D texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_read(ushort2 coord, ushort lod = 0) const
sparse_color-Tv sparse_read(uint2 coord, uint lod = 0) const

This member function gathers four samples for bilinear interpolation from a sparse 2D texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_gather(sampler s, float2 coord,
     int2 offset = int2(0),
     component c = component::x) const

These member functions simulate a texture fetch and return the LOD (mip level) computation result starting with Metal 2.3 in iOS and Metal 2.2 in macOS.
float calculate_clamped_lod(sampler s, float2 coord);
float calculate_unclamped_lod(sampler s, float2 coord);

These member functions execute an atomic load from a 2D texture starting with Metal 3.1:
Tv atomic_load(uint2 coord) const
Tv atomic_load(ushort2 coord) const

These member functions execute an atomic store to a 2D texture starting with Metal 3.1:
void atomic_store(Tv color, uint2 coord) const
void atomic_store(Tv color, ushort2 coord) const

These member functions execute an atomic compare and exchange to a 2D texture starting with Metal 3.1:
bool atomic_compare_exchange_weak(uint2 coord, thread Tv *expected, Tv desired) const
bool atomic_compare_exchange_weak(ushort2 coord, thread Tv *expected, Tv desired) const

These member functions execute an atomic exchange to a 2D texture starting with Metal 3.1:
Tv atomic_exchange(uint2 coord, Tv desired) const
Tv atomic_exchange(ushort2 coord, Tv desired) const

These member functions execute an atomic fetch and modify to a 2D texture starting with Metal 3.1 where op is add, and, max, min, or, sub, or xor for int and uint color type:
Tv atomic_fetch_op(uint2 coord, Tv operand)
Tv atomic_fetch_op(ushort2 coord, Tv operand) const

These member functions execute an atomic min or max to a 2D texture starting with Metal 3.1:
void atomic_min(uint2 coord, ulong4 operand)
void atomic_min(ushort2 coord, ulong4 operand)
void atomic_max(uint2 coord, ulong4 operand)
void atomic_max(ushort2 coord, ulong4 operand)
6.12.3.1 2D Texture Sampling Example

The following code shows several uses of the 2D texture sample function, depending upon its arguments:

```cpp
texture2d<float> tex;
sampler s;
float2 coord;
int2 offset;
float lod;

// No optional arguments.
float4 clr = tex.sample(s, coord);

// Sample using a mipmap level.
clr = tex.sample(s, coord, level(lod));

// Sample with an offset.
clr = tex.sample(s, coord, offset);

// Sample using a mipmap level and an offset.
clr = tex.sample(s, coord, level(lod), offset);
```

6.12.4 2D Texture Array

For the functions in this section, the following data types and corresponding constructor functions can specify sampling options (lod_options):

- bias(float value)
- level(float lod)
- gradient2d(float2 dPdx, float2 dPdy)
- min_lod_clamp(float lod) // All OS: support since Metal 2.2

These member functions sample from a 2D texture array:

```cpp
Tv sample(sampler s, float2 coord, uint array, int2 offset = int2(0)) const
Tv sample(sampler s, float2 coord, uint array, lod_options options, int2 offset = int2(0)) const
Tv sample(sampler s, float2 coord, uint array, bias bias_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
Tv sample(sampler s, float2 coord, uint array, gradient2d grad_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
```
These member functions perform sampler-less reads from a 2D texture array:

```cpp
Tv read(uint2 coord, uint array, uint lod = 0) const
Tv read(ushort2 coord, ushort array,
       ushort lod = 0) const // All OS: Since Metal 1.2.
```

These member functions write to a 2D texture array. On macOS up to Metal 2.2, `lod` must be 0. In Metal 2.3, this restriction is lifted for some hardware.

```cpp
void write(Tv color, uint2 coord, uint array, uint lod = 0)
void write(Tv color, ushort2 coord, ushort array,
          ushort lod = 0) // All OS: Since Metal 1.2.
```

This member functions gathers four samples for bilinear interpolation when sampling a 2D texture array:

```cpp
Tv gather(sampler s, float2 coord, uint array,
          int2 offset = int2(0),
          component c = component::x) const
```

These member functions query a 2D texture array:

```cpp
uint get_width(uint lod = 0) const
uint get_height(uint lod = 0) const
uint get_array_size() const
uint get_num_mip_levels() const
```

These member functions sample from a sparse 2D texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```cpp
sparse_color–Tv sparse_sample(sampler s, float2 coord, uint array,
                               int2 offset = int2(0)) const
sparse_color–Tv sparse_sample(sampler s, float2 coord, uint array,
                               bias options,
                               int2 offset = int2(0)) const
sparse_color–Tv sparse_sample(sampler s, float2 coord, uint array,
                               level options,
                               int2 offset = int2(0)) const
sparse_color–Tv sparse_sample(sampler s, float2 coord, uint array,
                               min_lod_clamp min_lod_clamp_options,
                               int2 offset = int2(0)) const
sparse_color–Tv sparse_sample(sampler s, float2 coord, uint array,
                               bias bias_options,
                               min_lod_clamp bias_options,
                               min_lod_clamp bias_options,
                               int2 offset = int2(0)) const
```
min_lod_clamp min_lod_clamp_options,
int2 offset = int2(0)) const
sparse_color-Tv sparse_sample(sampler s, float2 coord, uint array,
gradient2d options,
int2 offset = int2(0)) const
sparse_color-Tv sparse_sample(sampler s, float2 coord, uint array,
gradient2d grad_options,
min_lod_clamp min_lod_clamp_options,
int2 offset = int2(0)) const
These functions perform a sampler-less read from a sparse 2D texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_read(ushort2 coord, ushort array,
ushort lod = 0) const
sparse_color-Tv sparse_read(uint2 coord, uint array,
uint lod = 0) const
This function gathers four samples for bilinear interpolation from a sparse 2D texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_gather(sampler s, float2 coord, uint array,
int2 offset = int2(0),
component c = component::x) const
These member functions simulate a texture fetch and return the LOD (mip level) computation result starting with Metal 2.3 in iOS and Metal 2.2 in macOS.
float calculate_clamped_lod(sampler s, float2 coord);
float calculate_unclamped_lod(sampler s, float2 coord);
These member functions execute an atomic load from a 2D texture array starting with Metal 3.1:
Tv atomic_load(uint2 coord, uint array) const
Tv atomic_load(ushort2 coord, ushort array) const
These member functions execute an atomic store to a 2D texture array starting with Metal 3.1:
void atomic_store(Tv color, uint2 coord, uint array) const
void atomic_store(Tv color, ushort2 coord, ushort array) const
These member functions execute an atomic compare and exchange to a 2D texture array starting with Metal 3.1:
bool atomic_compare_exchange_weak(uint2 coord, uint array,
    thread Tv *expected,
    Tv desired) const

bool atomic_compare_exchange_weak(ushort2 coord, ushort array,
    thread Tv *expected,
    Tv desired) const

These member functions execute an atomic exchange to a 2D texture array starting with Metal 3.1:
Tv atomic_exchange(uint2 coord, uint array, Tv desired) const
Tv atomic_exchange(ushort2 coord, ushort array, Tv desired) const

These member functions execute an atomic fetch and modify to a 2D texture array starting with Metal 3.1 where op is add, and, max, min, or, sub, or xor for int and uint color type:
Tv atomic_fetch_op(uint2 coord, uint array, Tv operand)
Tv atomic_fetch_op(ushort2 coord, ushort array, Tv operand) const

These member functions execute an atomic min or max to a 2D texture array starting with Metal 3.1:
void atomic_min(uint2 coord, uint array, ulong4 operand)
void atomic_min(ushort2 coord, ushort array, ulong4 operand)
void atomic_max(uint2 coord, uint array, ulong4 operand)
void atomic_max(ushort2 coord, ushort array, ulong4 operand)

6.12.5  3D Texture

For the functions in this section, the following data types and corresponding constructor functions can specify sampling options (lod_options):
bias(float value)
level(float lod)
gradients3d(float3 dPdx, float3 dPdy)
min_lod_clamp(float lod) // All OS: support since Metal 2.2

These member functions sample from a 3D texture:
Tv sample(sampler s, float3 coord, int3 offset = int3(0)) const
Tv sample(sampler s, float3 coord, lod_options options,
    int3 offset = int3(0)) const
Tv sample(sampler s, float3 coord, bias bias_options,
   min_lod_clamp min_lod_clamp_options,
   int3 offset = int3(0)) const

Tv sample(sampler s, float3 coord, gradient3d grad_options,
   min_lod_clamp min_lod_clamp_options,
   int3 offset = int3(0)) const

These member functions perform sampler-less reads from a 3D texture:
Tv read(uint3 coord, uint lod = 0) const
Tv read(ushort3 coord, ushort lod = 0) const // All OS: Since Metal 1.2.

These member functions write to a 3D texture. On macOS up to Metal 2.2, lod must be 0. In Metal 2.3, this restriction is lifted for Some hardware.
void write(Tv color, uint3 coord, uint lod = 0)
void write(Tv color, ushort3 coord,
   ushort lod = 0) // All OS: Since Metal 1.2.

These member functions query a 3D texture:
uint get_width(uint lod = 0) const
uint get_height(uint lod = 0) const
uint get_depth(uint lod = 0) const
uint get_num_mip_levels() const

These functions sample from a sparse 3D texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_sample(sampler s, float3 coord,
   int3 offset = int3(0)) const
sparse_color-Tv sparse_sample(sampler s, float3 coord, bias options,
   int3 offset = int3(0)) const
sparse_color-Tv sparse_sample(sampler s, float3 coord,
   level options,
   int3 offset = int3(0)) const
sparse_color-Tv sparse_sample(sampler s, float3 coord,
   min_lod_clamp min_lod_clamp_options, int3 offset = int3(0)) const
sparse_color-Tv sparse_sample(sampler s, float3 coord,
   bias bias_options,
   min_lod_clamp min_lod_clamp_options,
   int3 offset = int3(0)) const
sparse_color-Tv sparse_sample(sampler s, float3 coord,
    gradient3d grad_options,
    int3 offset = int3(0)) const
sparse_color-Tv sparse_sample(sampler s, float3 coord,
    gradient3d grad_options,
    min_lod_clamp min_lod_clamp_options,
    int3 offset = int3(0)) const

These member functions perform a sampler-less read from a sparse 3D texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_read(uint3 coord, uint lod = 0) const
sparse_color-Tv sparse_read(ushort3 coord, ushort lod = 0) const

These member functions simulate a texture fetch and return the LOD (mip level) computation result starting with Metal 2.3 in iOS and Metal 2.2 in macOS.
float calculate_clamped_lod(sampler s, float3 coord)
float calculate_unclamped_lod(sampler s, float3 coord)

These member functions execute an atomic load from a 3D texture starting with Metal 3.1:
Tv atomic_load(uint3 coord) const
Tv atomic_load(ushort3 coord) const

These member functions execute an atomic store to a 3D texture starting with Metal 3.1:
void atomic_store(Tv color, uint3 coord) const
void atomic_store(Tv color, ushort3 coord) const

These member functions execute an atomic compare and exchange to a 3D texture starting with Metal 3.1:
bool atomic_compare_exchange_weak(uint3 coord, thread Tv *expected,
    Tv desired) const
bool atomic_compare_exchange_weak(ushort3 coord, thread Tv *expected,
    Tv desired) const

These member functions execute an atomic exchange to a 3D texture starting with Metal 3.1:
Tv atomic_exchange(uint3 coord, Tv desired) const
Tv atomic_exchange(ushort3 coord, Tv desired) const
These member functions execute an atomic fetch and modify to a 3D texture starting with Metal 3.1 where \( op \) is \( \text{add}, \text{and}, \text{max}, \text{min}, \text{or}, \text{sub}, \text{or} \text{xor} \) for \( \text{int} \) and \( \text{uint} \) color type:

\[
\text{Tv atomic_fetch_{op}(uint3 coord, Tv operand)}
\]

\[
\text{Tv atomic_fetch_{op}(us long3 coord, Tv operand) const}
\]

These member functions execute an atomic min or max to a 3D texture starting with Metal 3.1:

\[
\text{void atomic_{min}(uint3 coord, ulong4 operand)}
\]

\[
\text{void atomic_{min}(ushort3 coord, ulong4 operand)}
\]

\[
\text{void atomic_{max}(uint3 coord, ulong4 operand)}
\]

\[
\text{void atomic_{max}(ushort3 coord, ulong4 operand)}
\]

### 6.12.6 Cube Texture

For the functions in this section, the following data types and corresponding constructor functions can specify sampling options (lod_options):

bias(float value)

level(float lod)

gradientcube(float3 dPdx, float3 dPdy)

min_lod_clamp(float lod) // All OS: support since Metal 2.2

These member functions sample from a cube texture:

\[
\text{Tv sample(sampler s, float3 coord) const}
\]

\[
\text{Tv sample(sampler s, float3 coord, lod_options options) const}
\]

\[
\text{Tv sample(sampler s, float3 coord, bias bias\_options,}
\text{ min\_lod\_clamp min\_lod\_clamp\_options) const}
\]

\[
\text{Tv sample(sampler s, float3 coord, gradientcube grad\_options,}
\text{ min\_lod\_clamp min\_lod\_clamp\_options) const}
\]

Table 6.21 describes the cube face and the number used to identify the face.

<table>
<thead>
<tr>
<th>Face Number</th>
<th>Cube face</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Positive X</td>
</tr>
<tr>
<td>1</td>
<td>Negative X</td>
</tr>
<tr>
<td>2</td>
<td>Positive Y</td>
</tr>
<tr>
<td>Face Number</td>
<td>Cube face</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>3</td>
<td>Negative Y</td>
</tr>
<tr>
<td>4</td>
<td>Positive Z</td>
</tr>
<tr>
<td>5</td>
<td>Negative Z</td>
</tr>
</tbody>
</table>

This member function gathers four samples for bilinear interpolation when sampling a cube texture:

\[
Tv \text{ gather(sampler s, float3 coord, component c = component::x)} \text{ const}
\]

These member functions perform sampler-less reads from a cube texture:

\[
Tv \text{ read(uint2 coord, uint face, uint lod = 0)} \text{ const}
\]

\[
Tv \text{ read(ushort2 coord, ushort face, ushort lod = 0)} \text{ const} // \text{ All OS: Since Metal 1.2.}
\]

These member functions write to a cube texture. On macOS up to Metal 2.2, lod must be 0. In Metal 2.3, this restriction is lifted for some hardware.

\[
\text{void write(Tv color, uint2 coord, uint face, uint lod = 0)}
\]

\[
\text{void write(Tv color, ushort2 coord, ushort face, ushort lod = 0)} // \text{ All OS: Since Metal 1.2.}
\]

These member functions query a cube texture:

\[
\text{uint get_width(uint lod = 0)} \text{ const}
\]

\[
\text{uint get_height(uint lod = 0)} \text{ const}
\]

\[
\text{uint get_num_mip_levels() const}
\]

These member functions sample from a sparse cube texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

\[
\text{sparse_color-Tv sparse_sample(sampler s, float3 coord) const}
\]

\[
\text{sparse_color-Tv sparse_sample(sampler s, float3 coord, bias options) const}
\]

\[
\text{sparse_color-Tv sparse_sample(sampler s, float3 coord, level options) const}
\]

\[
\text{sparse_color-Tv sparse_sample(sampler s, float3 coord, min_lod_clamp min_lod_clamp_options) const}
\]
sparse_color-Tv sparse_sample(sampler s, float3 coord,
    bias bias_options,
    min_lod_clamp min_lod_clamp_options) const
sparse_color-Tv sparse_sample(sampler s, float3 coord,
    gradientcube grad_options) const
sparse_color-Tv sparse_sample(sampler s, float3 coord,
    gradientcube grad_options,
    min_lod_clamp min_lod_clamp_options) const

These member functions perform a sampler-less read from a sparse cube texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_read(ushort3 coord, ushort lod = 0) const
sparse_color-Tv sparse_read(uint3 coord, uint lod = 0) const

This member function gathers four samples for bilinear interpolation from a sparse cube texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_gather(sampler s, float3 coord,
    component c = component::x) const

These member functions simulate a texture fetch and return the LOD (mip level) computation result starting with Metal 2.3 in iOS and Metal 2.2 in macOS.

float calculate_clamped_lod(sampler s, float3 coord);
float calculate_unclamped_lod(sampler s, float3 coord);

### 6.12.7 Cube Texture Array

For the functions in this section, the following data types and corresponding constructor functions can specify sampling options (lod_options):

- bias(float value)
- level(float lod)
- gradientcube(float3 dPdx, float3 dPdy)
- min_lod_clamp(float lod) // All OS: support since Metal 2.2

These member functions sample from a cube texture array:

Tv sample(sampler s, float3 coord, uint array) const
Tv sample(sampler s, float3 coord, uint array,
    lod_options options) const
Tv sample(sampler s, float3 coord, uint array, bias bias_options,
  min_lod_clamp min_lod_clamp_options) const
Tv sample(sampler s, float3 coord, uint array,
  gradientcube grad_options,
  min_lod_clamp min_lod_clamp_options) const

This member function gathers four samples for bilinear interpolation when sampling a cube texture array:
Tv gather(sampler s, float3 coord, uint array,
component c = component::x) const

These member functions perform sampler-less reads from a cube texture array:
Tv read(uint2 coord, uint face, uint array, uint lod = 0) const
Tv read(ushort2 coord, ushort face, ushort array,
  ushort lod = 0) const // All OS: Since Metal 1.2.

These member functions write to a cube texture array. On macOS up to Metal 2.2, lod must be 0. In Metal 2.3, this restriction is lifted for Some hardware.
void write(Tv color, uint2 coord, uint face, uint array,
  uint lod = 0)
void write(Tv color, ushort2 coord, ushort face, ushort array,
  ushort lod = 0) // All OS: Since Metal 1.2.

These member functions query a cube texture array:
uint get_width(uint lod = 0) const
uint get_height(uint lod = 0) const
uint get_array_size() const
uint get_num_mip_levels() const

These member functions sample from a sparse cube texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_sample(sampler s, float3 coord,
  uint array) const
sparse_color-Tv sparse_sample(sampler s, float3 coord, uint array,
  bias options) const
sparse_color-Tv sparse_sample(sampler s, float3 coord, uint array,
  level options) const
sparse_color-Tv sparse_sample(sampler s, float3 coord, uint array,  
    min_lod_clamp min_lod_clamp_options) const
sparse_color-Tv sparse_sample(sampler s, float3 coord, uint array,  
    bias bias_options,  
    min_lod_clamp min_lod_clamp_options) const
sparse_color-Tv sparse_sample(sampler s, float3 coord, uint array,  
    gradientcube options) const
sparse_color-Tv sparse_sample(sampler s, float3 coord, uint array,  
    gradientcube grad_options,  
    min_lod_clamp min_lod_clamp_options) const

These member functions perform a sampler-less read from a sparse 2D texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_read(ushort2 coord, ushort face,  
    ushort array, ushort lod = 0) const
sparse_color-Tv sparse_read(uint2 coord, uint face,  
    uint array, uint lod = 0) const

This member function gathers four samples for bilinear interpolation from a sparse cube texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_gather(sampler s, float3 coord, uint array,  
    component c = component::x) const

These member functions simulate a texture fetch and return the LOD (mip level) computation result starting with Metal 2.3 in iOS and Metal 2.2 in macOS.

float calculate_clamped_lod(sampler s, float3 coord);
float calculate_unclamped_lod(sampler s, float3 coord);

6.12.8 2D Multisampled Texture

These member functions perform sampler-less reads from a 2D multisampled texture:
Tv read(uint2 coord, uint sample) const
Tv read(ushort2 coord, ushort sample) const // All OS: Since Metal 1.2.

If you have customized sample positions (set with the `setSamplePositions:count:` method of `MTLRenderPassDescriptor`), then `read(coord, sample)` returns the data for the sample at the programmed sample position.

These member functions query a 2D multisampled texture:
uint get_width() const
uint get_height() const
uint get_num_samples() const

These member functions perform a sampler-less read from a sparse 2D multisampled texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_read(ushort2 coord, ushort sample) const
sparse_color-Tv sparse_read(uint2 coord, uint sample) const

6.12.9  2D Multisampled Texture Array
macOS: Since Metal 2.
iOS: Since Metal 2.3.
The following member functions can perform sampler-less reads from a 2D multisampled texture array:
Tv read(uint2 coord, uint array, uint lod = 0) const
Tv read(ushort2 coord, ushort array, ushort lod = 0) const

These member functions query a 2D multisampled texture array:
uint get_width(uint lod = 0) const
uint get_height(uint lod = 0) const
uint get_array_size() const
uint get_num_mip_levels() const

These functions perform a sampler-less read from a sparse 2D multisampled texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-Tv sparse_read(ushort2 coord, ushort array,
    ushort sample) const
sparse_color-Tv sparse_read(uint2 coord, uint array,
    uint sample) const

6.12.10  2D Depth Texture
For the functions in this section, the following data types and corresponding constructor functions can specify sampling options (lod_options):
bias(float value)
level(float lod)
gradient2d(float2 dPdx, float2 dPdy)
min_lod_clamp(float lod) // All OS: support since Metal 2.2

These member functions sample from a 2D depth texture:
T sample(sampler s, float2 coord, int2 offset = int2(0)) const
T sample(sampler s, float2 coord, lod_options options,
    int2 offset = int2(0)) const
T sample(sampler s, float2 coord, bias bias_options,
    min_lod_clamp min_lod_clamp_options,
    int2 offset = int2(0)) const
T sample(sampler s, float2 coord, gradient2d grad_options,
    min_lod_clamp min_lod_clamp_options,
    int2 offset = int2(0)) const

These member functions sample from a 2D depth texture and compare a single component against the comparison value:
T sample_compare(sampler s, float2 coord, float compare_value,
    int2 offset = int2(0)) const
T sample_compare(sampler s, float2 coord, float compare_value,
    lod_options options, int2 offset = int2(0)) const
T must be a float type.
sample_compare performs a comparison of the compare_value value against the pixel value (1.0 if the comparison passes and 0.0 if it fails). These comparison result values per-pixel are then blended together as in normal texture filtering and the resulting value between 0.0 and 1.0 is returned. On macOS up to Metal 2.2, the supported lod_options values are level and min_lod_clamp (the latter, since Metal 2.2); lod must be a zero constant. In Metal 2.3, this restriction for lod_options is lifted for Apple silicon.

These member functions perform sampler-less reads from a 2D depth texture:
T read(uint2 coord, uint lod = 0) const
T read(ushort2 coord, ushort lod = 0) const // All OS: Since Metal 1.2.

This built-in function gathers four samples for bilinear interpolation when sampling a 2D depth texture:
Tv gather(sampler s, float2 coord, int2 offset = int2(0)) const

This member function gathers four samples for bilinear interpolation when sampling a 2D depth texture and comparing these samples with a specified comparison value (1.0 if the comparison passes and 0.0 if it fails).
Tv gather_compare(sampler s, float2 coord, float compare_value, int2 offset = int2(0)) const

T must be a float type.

The following member functions query a 2D depth texture:

uint get_width(uint lod = 0) const
uint get_height(uint lod = 0) const
uint get_num_mip_levels() const

These member functions sample from a sparse 2D depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-T sparse_sample(sampler s, float2 coord, int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, bias options, int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, level options, int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, bias bias_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, gradient2d grad_options, int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, gradient2d grad_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const

These member functions sample from a sparse 2D depth texture and compare a single component against a comparison value starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-T sparse_sample_compare(sampler s, float2 coord, float compare_value, int2 offset = int2(0)) const
sparse_color-T sparse_sample_compare(sampler s, float2 coord, float compare_value,
These member functions perform a sampler-less read from a sparse 2D depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```
sparse_color-T sparse_sample_compare(sampler s, float2 coord, float compare_value, level options, int2 offset = int2(0)) const
sparse_color-T sparse_sample_compare(sampler s, float2 coord, float compare_value, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
sparse_color-T sparse_sample_compare(sampler s, float2 coord, float compare_value, bias bias_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
sparse_color-T sparse_sample_compare(sampler s, float2 coord, float compare_value, gradient2d grad_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
sparse_color-T sparse_sample_compare(sampler s, float2 coord, float compare_value, gradient2d grad_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
```

This member function gathers four samples for bilinear interpolation from a sparse 2D depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```
sparse_color-TV sparse_gather(sampler s, float2 coord, int2 offset = int2(0), component c = component::x) const
```

This member function gathers those samples and compare them against a comparison value from a sparse 2D depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```
sparse_color-TV sparse_gather_compare(sampler s, float2 coord, float compare_value, int2 offset = int2(0)) const
```

These member functions simulate a texture fetch and return the LOD (mip level) computation result starting with Metal 2.3 in iOS and Metal 2.2 in macOS.

```
float calculate_clamped_lod(sampler s, float2 coord);
```
float calculate_unclamped_lod(sampler s, float2 coord);

### 6.12.11 2D Depth Texture Array

The member functions in this section use the following data types and constructor functions to set the sampling option fields of their lod_options parameter.

- bias(float value)
- level(float lod)
- gradient2d(float2 dPdx, float2 dPdy)
- min_lod_clamp(float lod) // All OS: support since Metal 2.2

These member functions sample from a 2D depth texture array.

- \( T \) sample(sampler s, float2 coord, uint array, int2 offset = int2(0)) const
- \( T \) sample(sampler s, float2 coord, uint array, lod_options options, int2 offset = int2(0)) const
- \( T \) sample(sampler s, float2 coord, uint array, bias bias_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const
- \( T \) sample(sampler s, float2 coord, uint array, gradient2d grad_options, min_lod_clamp min_lod_clamp_options, int2 offset = int2(0)) const

These member functions sample from a 2D depth texture array and compare a single component to a value where \( T \) is a float type.

- \( T \) sample_compare(sampler s, float2 coord, uint array, float compare_value, int2 offset = int2(0)) const
- \( T \) sample_compare(sampler s, float2 coord, uint array, float compare_value, lod_options options, int2 offset = int2(0)) const

The lod_options fields support are:

- **level**
- **bias** for all iOS Metal versions and macOS Metal 2.3 and later on Apple silicon
- **gradient** for iOS Metal versions and macOS Metal 2.3 and later on Apple silicon
- **min_lod_clamp** for Metal 2.2 and later
  - Must be 0 for Metal 2.2
  - Can be any value for all iOS Metal versions and macOS Metal 2.3 and later on Apple silicon
These member functions read from a 2D depth texture array without using a sampler.

T read(uint2 coord, uint array, uint lod = 0) const
T read(ushort2 coord, ushort array,       
    ushort lod = 0) const // All OS: Since Metal 1.2.

This member function gathers four samples for bilinear interpolation when sampling a 2D depth texture array.

Tv gather(sampler s, float2 coord, uint array,       
    int2 offset = int2(0)) const

This member function gathers four samples for bilinear interpolation when sampling a 2D depth texture array and compares them to a value where Tv is a float vector type.

Tv gather_compare(sampler s, float2 coord, uint array,       
    float compare_value, int2 offset = int2(0)) const

The following member functions query a 2D depth texture array.

uint get_width(uint lod = 0) const
uint get_height(uint lod = 0) const
uint get_array_size() const
uint get_num_mip_levels() const

These member functions sample from a sparse 2D depth texture array, starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-T sparse_sample(sampler s, float2 coord, uint array,       
    int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, uint array,       
    bias options,       
    int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, uint array,       
    level options,       
    int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, uint array,       
    min_lod_clamp min_lod_clamp_options,       
    int2 offset = int2(0)) const
sparse_color-T sparse_sample(sampler s, float2 coord, uint array,       
    bias bias_options,       
    min_lod_clamp min_lod_clamp_options,       
    int2 offset = int2(0)) const
These functions sample from a sparse 2D depth texture array and compare a single component to a comparison value, starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

These functions read from a sparse 2D depth texture array without a sampler, starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
sparse_color-T sparse_read(uint2 coord, uint array, uint lod = 0) const

This function gathers four samples for bilinear interpolation from a sparse 2D depth texture array, starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_gather(sampler s, float2 coord, uint array, int2 offset = int2(0), component c = component::x) const

This function gathers those samples and compares them against a value from a sparse 2D depth texture array, starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-Tv sparse_gather_compare(sampler s, float2 coord, uint array, float compare_value, int2 offset = int2(0)) const

These functions simulate a texture fetch and return a LOD (mip level) computation result, starting with Metal 2.3 in iOS and Metal 2.2 in macOS.

float calculate_clamped_lod(sampler s, float2 coord);
float calculate_unclamped_lod(sampler s, float2 coord);

### 6.12.12 2D Multisampled Depth Texture

The following member functions can perform sampler-less reads from a 2D multisampled depth texture:

T read(uint2 coord, uint sample) const
T read(ushort2 coord, ushort sample) const // All OS: Since Metal 1.2.

The following member functions query a 2D multisampled depth texture:

uint get_width() const
uint get_height() const
uint get_num_samples() const

These member functions perform a sampler-less read from a sparse 2D multisampled depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-T sparse_read(ushort2 coord, ushort sample) const
sparse_color-T sparse_read(uint2 coord, uint sample) const
6.12.13  2D Multisampled Depth Texture Array

macOS: Since Metal 2.
iOS: Since Metal 2.3.

The following member functions perform sampler-less reads from a 2D multisampled depth texture array:

Tv read(uint2 coord, uint array, uint lod = 0) const
Tv read(ushort2 coord, ushort array, ushort lod = 0) const

The following member functions query a 2D multisampled depth texture array:

uint get_width(uint lod = 0) const
uint get_height(uint lod = 0) const
uint get_array_size() const
uint get_num_mip_levels() const

These member functions perform a sampler-less read from a sparse 2D multisampled depth texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-T sparse_read(ushort2 coord, ushort array, ushort sample) const
sparse_color-T sparse_read(uint2 coord, uint array, uint sample) const

6.12.14  Cube Depth Texture

For the functions in this section, the following data types and corresponding constructor functions can specify sampling options (lod_options):

bias(float value)
level(float lod)
gradientcube(float3 dPdx, float3 dPdy)
min_lod_clamp(float lod) // All OS: support since Metal 2.2

The following member functions sample from a cube depth texture:

T sample(sampler s, float3 coord) const
T sample(sampler s, float3 coord, lod_options options) const
T sample(sampler s, float3 coord, bias bias_options,
          min_lod_clamp min_lod_clamp_options) const
The following member functions sample from a cube depth texture and compare a single component against the specified comparison value:

T sample_compare(sampler s, float3 coord, float compare_value) const
T sample_compare(sampler s, float3 coord, float compare_value, lod_options options) const

T must be a float type. On macOS up to Metal 2.2, the supported lod_options values are level and min_lod_clamp (the latter, since Metal 2.2), and lod must be a zero constant. In Metal 2.3, this restriction for lod_options is lifted for some hardware.

The following member functions perform sampler-less reads from a cube depth texture:

T read(uint2 coord, uint face, uint lod = 0) const
T read(ushort2 coord, ushort face, ushort lod = 0) const // All OS: Since Metal 1.2.

This member function gathers four samples for bilinear interpolation when sampling a cube depth texture:

Tv gather(sampler s, float3 coord) const

This member function gathers four samples for bilinear interpolation when sampling a cube texture and comparing these samples with a specified comparison value:

Tv gather_compare(sampler s, float3 coord, float compare_value) const

T must be a float type.

The following member functions query a cube depth texture:

uint get_width(uint lod = 0) const
uint get_height(uint lod = 0) const
uint get_num_mip_levels() const

These member functions sample from a sparse cube depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-T sparse_sample(sampler s, float3 coord) const
sparse_color-T sparse_sample(sampler s, float3 coord,
These member functions sample from a sparse cube depth texture and compare a single component against a comparison value starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

These member functions perform a sampler-less read from a sparse cube depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.
This member function gathers four samples for bilinear interpolation from a sparse cube depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```c
sparse_color-Tv sparse_gather(sampler s, float3 coord) const
```

This member function gathers those samples and compare them against a comparison value from a sparse cube depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```c
sparse_color-Tv sparse_gather_compare(sampler s, float3 coord, float compare_value) const
```

These member functions simulate a texture fetch and return the LOD (mip level) computation result starting with Metal 2.3 in iOS and Metal 2.2 in macOS.

```c
float calculate_clamped_lod(sampler s, float3 coord);
float calculate_unclamped_lod(sampler s, float3 coord);
```

### 6.12.15 Cube Depth Texture Array

For the functions in this section, the following data types and corresponding constructor functions can specify sampling options (lod_options):

- bias(float value)
- level(float lod)
- gradientcube(float3 dPdx, float3 dPdy)
- min_lod_clamp(float lod) // All OS: support since Metal 2.2

These member functions sample from a cube depth texture array:

```c
T sample(sampler s, float3 coord, uint array) const
T sample(sampler s, float3 coord, uint array, lod_options options) const
T sample(sampler s, float3 coord, uint array, bias bias_options,
         min_lod_clamp min_lod_clamp_options) const
T sample(sampler s, float3 coord, uint array, gradientcube grad_options,
         min_lod_clamp min_lod_clamp_options) const
```

These member functions sample from a cube depth texture and compare a single component against the specified comparison value:

```c
T sample_compare(sampler s, float3 coord, uint array, float compare_value) const
```
T sample_compare(sampler s, float3 coord, uint array, float compare_value, lod_options options) const

_T must be a float type. On macOS up to Metal 2.2, the supported _lod_options values are level and _min_lod_clamp (the latter, since Metal 2.2), and _lod must be a zero constant. In Metal 2.3, this restriction for _lod_options is lifted for some hardware._

These member functions perform sampler-less reads from a cube depth texture array:

_T read(uint2 coord, uint face, uint array, uint _lod = 0) const
_T read(ushort2 coord, ushort face, ushort array, ushort _lod = 0) const // All OS: Since Metal 1.2.

This member function gathers four samples for bilinear interpolation when sampling a cube depth texture:

Tv gather(sampler s, float3 coord, uint array) const

This member function gathers four samples for bilinear interpolation when sampling a cube depth texture and comparing these samples with a specified comparison value:

Tv gather_compare(sampler s, float3 coord, uint array, float compare_value) const

_T must be a float type._

These member functions query a cube depth texture:

uint get_width(uint _lod = 0) const
uint get_height(uint _lod = 0) const
uint get_array_size() const
uint get_num_mip_levels() const

These member functions sample from a sparse cube depth texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

sparse_color-_T sparse_sample(sampler s, float3 coord, uint array) const
sparse_color-_T sparse_sample(sampler s, float3 coord, uint array, bias options) const
sparse_color-_T sparse_sample(sampler s, float3 coord, uint array, level options) const
sparse_color-_T sparse_sample(sampler s, float3 coord, uint array, _min_lod_clamp min_lod_clamp_options) const
sparse_color-_T sparse_sample(sampler s, float3 coord, uint array, bias bias_options,
These member functions sample from a sparse cube depth texture array and compare a single component against a comparison value starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```
sparse_color-T sparse_sample(sampler s, float3 coord, uint array, gradientcube grad_options) const
sparse_color-T sparse_sample(sampler s, float3 coord, uint array, gradientcube grad_options, min_lod_clamp min_lod_clamp_options) const
```

These member functions perform a sampler-less read from a sparse cube depth texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```
sparse_color-T sparse_read(ushort2 coord, ushort face, ushort array, ushort lod = 0) const
sparse_color-T sparse_read(uint2 coord, uint face, uint array, uint lod = 0) const
```
This member function gathers four samples for bilinear interpolation from a sparse cube depth texture array starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```
sparse_color-Tv sparse_gather(sampler s, float3 coord, uint array) const
```

This member function gathers those samples and compare them against a comparison value from a sparse 2D depth texture starting with Metal 2.2 in iOS and Metal 2.3 in macOS.

```
sparse_color-Tv sparse_gather_compare(sampler s, float3 coord, uint array, float compare_value) const
```

These member functions simulate a texture fetch and return the LOD (mip level) computation result starting with Metal 2.3 in iOS and Metal 2.2 in macOS.

```
float calculate_clamped_lod(sampler s, float3 coord);
float calculate_unclamped_lod(sampler s, float3 coord);
```

### 6.12.16 Texture Buffer Functions

All OS: texture buffers and these functions supported since Metal 2.1.

The following member functions can read from and write to an element in a texture buffer (also see section 2.9.1):

```
Tv read(uint coord) const;
void write(Tv color, uint coord);
```

These member functions execute an atomic load from a texture buffer starting with Metal 3.1:

```
Tv atomic_load(uint coord) const
Tv atomic_load(ushort coord) const
```

These member functions execute an atomic store to a texture buffer starting with Metal 3.1:

```
void atomic_store(Tv color, uint coord) const
void atomic_store (Tv color, ushort coord) const
```

These member functions execute an atomic compare and exchange to a texture buffer starting with Metal 3.1:

```
bool atomic_compare_exchange_weak(uint coord, thread Tv *expected, Tv desired) const
bool atomic_compare_exchange_weak(ushort coord, thread Tv *expected,
Tv desired) const

These member functions execute an atomic exchange to a texture buffer starting with Metal 3.1:
Tv atomic_exchange(uint coord, Tv desired) const
Tv atomic_exchange(ushort coord, Tv desired) const

These member functions execute an atomic fetch and modify to a texture buffer starting with Metal 3.1 where op is add, and, max, min, or, sub, or xor for int and uint color type:
Tv atomic_fetch_op(uint coord, Tv operand) const
Tv atomic_fetch_op(ushort coord, Tv operand) const

These member functions execute an atomic min or max to a texture buffer starting with Metal 3.1:
void atomic_min(uint coord, ulong4 operand)
void atomic_min(ushort coord, ulong4 operand)
void atomic_max(uint coord, ulong4 operand)
void atomic_max(ushort coord, ulong4 operand)

The following example uses the read method to access a texture buffer:
kernel void
myKernel(texture_buffer<float, access::read> myBuffer) { 
       uint index = ...;
       float4 value = myBuffer.read(index);
   }

Use the following method to query the number of elements in a texture buffer:
uint get_width() const;

6.12.17 Texture Synchronization Functions
All OS: Texture synchronization functions supported since Metal 1.2.
The texture fence() member function ensures that writes to the texture by a thread become visible to subsequent reads from that texture by the same thread (the thread that is performing the write). Texture types (including texture buffers) that you can declare with the access::read_write attribute support the Fence function.
void fence()
The following example shows how to use a texture fence function to make sure that writes to a texture by a thread are visible to later reads to the same location by the same thread:

```cpp
kernel void
my_kernel(texture2d<float, access::read_write> texA,
       ...
ushort2 gid [[thread_position_in_grid]])
{
    float4 clr = ...;
texA.write(clr, gid);
    ...
    // Use fence to ensure that writes by thread become
    // visible to later reads by the thread.
texA.fence();
    clr_new = texA.read(gid);
    ...
}
```

### 6.12.18 Null Texture Functions

All OS: null texture functions supported since Metal 1.2.

macOS: null texture functions supported for `texture2d_ms_array` and `depth2d_ms_array` since Metal 2.

Use the following functions to determine if a texture is a null texture. If the texture is a null texture, `is_null_texture` returns true; otherwise it returns false.

```cpp
bool is_null_texture(texture1d<T, access>);
bool is_null_texture(texture1d_array<T, access>);
bool is_null_texture(texture2d<T, access>);
bool is_null_texture(texture2d_array<T, access>);
bool is_null_texture(texture3d<T, access>);
bool is_null_texture(texturecube<T, access>);
bool is_null_texture(texturecube_array<T, access>);
bool is_null_texture(texture2d_ms<T, access>);
// texture2d_ms_array is macOS only, since Metal 2
bool is_null_texture(texture2d_ms_array<T, access>);
bool is_null_texture(depth2d<T, access>);
bool is_null_texture(depth2d_array<T, access>);
bool is_null_texture(depthcube<T, access>);
bool is_null_texture(depthcube_array<T, access>);
```
bool is_null_texture(depth2d_ms<T, access>);
// depth2d_ms_array is macOS only, since Metal 2
bool is_null_texture(depth2d_ms_array<T, access>);

The behavior of calling any texture member function with a null texture is undefined.

### 6.13 Imageblock Functions

macOS: Imageblocks supported since Metal 2.3.

iOS: imageblocks supported since Metal 2.

This section lists the Metal member functions for imageblocks. (For more about the imageblock data type, see sections 2.11 and 5.6.)

The following member functions query information about the imageblock:

```cpp
ushort get_width() const;
ushort get_height() const;
ushort get_num_samples() const;
```

Use the following member function to query the number of unique color entries for a specific location given by an (x, y) coordinate inside the imageblock:

```cpp
ushort get_num_colors(ushort2 coord) const;
```

The following member function returns the color coverage mask (that is, whether a given color covers one or more samples in the imageblock). Each sample is identified by its bit position in the return value. If a bit is set, then this indicates that this sample uses the color index.

```cpp
ushort get_color_coverage_mask(ushort2 coord, ushort color_index) const;
```

`color_index` is a value from 0 to `get_num_colors() - 1`.

#### 6.13.1 Functions for Imageblocks with Implicit Layout

Use the following functions to read or write an imageblock at pixel rate for a given (x, y) coordinate inside the imageblock:

```cpp
T read(ushort2 coord) const;
void write(T data, ushort2 coord);
```

Use the following member function to read or write an imageblock at sample or color rate. `coord` specifies the (x, y) coordinate inside the imageblock, and `index` is the sample or color index.

```cpp
enum class imageblock_data_rate { color, sample };
```
T read(ushort2 coord, ushort index, 
    imageblock_data_rate data_rate) const;

void write(T data, ushort2 coord, ushort index, 
    imageblock_data_rate data_rate);

Example:

struct Foo {
    float4 a [[color(0)]]; 
    int4 b [[color(1)]]; 
};

kernel void
my_kernel(imageblock<Foo, imageblock_layout_implicit> img_blk, 
    ushort2 lid [[thread_position_in_threadgroup]] ...) 
{
    ... 
    Foo f = img_blk.read(lid); float4 r = f.a; 
    ... 
    f.a = r; 
    ... 
    img_blk.write(f, lid);
}

Use the following member function to write an imageblock with a color coverage mask. You must use this member function when writing to an imageblock at color rate:

void write(T data, ushort2 coord, ushort color_coverage_mask);

Use the following member functions to get a region of a slice for a given data member in the imageblock. You use these functions to write data associated with a specific data member described in the imageblock for all threads in the threadgroup to a specified region in a texture. color_index refers to the data member declared in the structure type specified in imageblock<T> with the [[color(n)]] attribute where n is color_index.size is the actual size of the copied slice.

const imageblock_slice<E, imageblock_layout_implicit> slice(ushort color_index) const;
const imageblock_slice<E, imageblock_layout_implicit> slice(ushort color_index, ushort2 size) const;

The region to copy has an origin of (0,0). The slice(...) member function that does not have the argument size copies the entire width and height of the imageblock.
6.13.2 Functions for Imageblocks with Explicit Layout

Use the following member functions to get a reference to the imageblock data for a specific location given by an \((x, y)\) coordinate inside the imageblock. Use these member functions when reading or writing data members in an imageblock at pixel rate.

\[
\text{threadgroup_imageblock } T* \text{ data(ushort2 coord);} \\
\text{const threadgroup_imageblock } T* \text{ data(ushort2 coord) const;}
\]

Use the following member functions to get a reference to the imageblock data for a specific location given by an \((x, y)\) coordinate inside the imageblock and a sample or color index. Use these member functions when reading or writing data members in an imageblock at sample or color rate. \(T\) is the type specific in the imageblock<T> templated declaration. coord is the coordinate in the imageblock, and index is the sample or color index for a multisampled imageblock. data_rate specifies whether the index is a color or sample index. If coord refers to a location outside the imageblock dimensions or if index is an invalid index, the behavior of data() is undefined.

\[
\text{enum class imageblock_data_rate \{ color, sample \};} \\
\text{threadgroup_imageblock } T* \text{ data(ushort2 coord, ushort index, imageblock_data_rate data_rate);} \\
\text{const threadgroup_imageblock } T* \text{ data(ushort2 coord, ushort index, imageblock_data_rate data_rate) const;}
\]

Calling the data(coord) member function for an imageblock that stores pixels at sample or color rate is equivalent to calling data(coord, 0, imageblock_data_rate::sample).

Example:

\[
\text{struct Foo \{} \\
\quad \text{rgba8unorm<half4> a;} \\
\quad \text{int b;} \\
\text{\};}
\]

\[
\text{kernel void my_kernel(imageblock<Foo> img_blk,} \\
\quad \text{ushort2 lid [[thread_position_in_threadgroup]] \ldots} \\
\text{\{} \\
\quad \text{\ldots} \\
\quad \text{threadgroup_imageblock Foo* f = img_blk.data(lid);} \\
\quad \text{half4 r = f->a;} \\
\quad \text{f->a = r;} \\
\quad \text{\ldots} \\
\text{\}}
\]

Use the following write member function to write an imageblock with a color coverage mask. You must use this member function when writing to an imageblock at color rate.

\[
\text{void write(T data, ushort2 coord, ushort color_coverage_mask);} \\
\]
Use the following `slice` member functions to get a region of a slice for a given data member in the `imageblock` structure. You use this function to write data associated with a specific data member described in the `imageblock` structure for all threads in the threadgroup to a specified region in a texture.

`data_member` is a data member declared in the structure type specified in `imageblock<T>`. `size` is the actual size of the copied slice.

```cpp
const imageblock_slice<E, imageblock_layout_explicit>
slice(const threadgroup_imageblock E& data_member) const;
const imageblock_slice<E, imageblock_layout_explicit>
slice(const threadgroup_imageblock E& data_member, ushort2 size) const;
```

The region to copy has an origin of (0,0). The `slice(…)` member function that does not have the argument `size` copies the entire width and height of the `imageblock`.

### 6.13.3 Writing an `imageblock` Slice to a Region in a Texture

Use the following `write(…)` member function in these texture types to write pixels associated with a slice in the `imageblock` to a texture starting at location given by `coord`. Note that the `write` from `imageblock_slice` to a texture must have matching MSAA modes or the result is undefined.

#### For a 1D texture:

```cpp
void write(imageblock_slice<E, imageblock_layout_explicit> slice, uint coord, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, ushort coord, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, uint coord, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, ushort coord, ushort lod = 0);
```

#### For a 1D texture array:

```cpp
void write(imageblock_slice<E, imageblock_layout_explicit> slice, uint coord, uint array, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, ushort coord, ushort array, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, uint coord, uint array, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, ushort coord, ushort array, ushort lod = 0);
```

#### For a 2D texture:

```cpp
void write(imageblock_slice<E, imageblock_layout_explicit> slice, uint coord, uint array, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, ushort coord, ushort array, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, uint coord, uint array, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, ushort coord, ushort array,ushort lod = 0);
```
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  uint2 coord, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  ushort2 coord, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, 
  uint2 coord, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, 
  ushort2 coord, ushort lod = 0);

For a 2D MSAA texture:
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  uint2 coord, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  ushort2 coord, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, 
  uint2 coord, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, 
  ushort2 coord, ushort lod = 0);

For a 2D texture array:
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  uint2 coord, uint array, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  ushort2 coord, ushort array, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, 
  uint2 coord, uint array, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, 
  ushort2 coord, ushort array, ushort lod = 0);

For a cube texture:
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  uint2 coord, uint face, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  ushort2 coord, ushort face, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, 
  uint2 coord, uint face, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, 
  ushort2 coord, ushort face, ushort lod = 0);

For a cube texture array:
void write(imageblock_slice<E, imageblock_layout_explicit> slice, 
  uint2 coord, uint face, uint array, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, ushort2 coord, ushort face, ushort array, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, uint2 coord, uint face, uint array, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, ushort2 coord, ushort face, ushort array, ushort lod = 0);

For a 3D texture:
void write(imageblock_slice<E, imageblock_layout_explicit> slice, uint3 coord, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_explicit> slice, ushort3 coord, ushort lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, uint3 coord, uint lod = 0);
void write(imageblock_slice<E, imageblock_layout_implicit> slice, ushort3 coord, ushort lod = 0);

Example:

struct Foo {
    half4 a;
    int b;
    float c;
};

kernel void
my_kernel(texture2d<half> src [[ texture(0) ]],
    texture2d<half, access::write> dst [[ texture(1) ]],
    imageblock<Foo> img_blk,
    ushort2 lid [[ thread_position_in_threadgroup ]],
    ushort2 gid [[ thread_position_in_grid ]])
{
    // Read the pixel from the input image using the thread ID.
    half4 clr = src.read(gid);

    // Get the image slice.
    threadgroup_imageblock Foo* f = img_blk.data(lid);
    // Write the pixel in the imageblock using the thread ID in
    // threadgroup.
    f->a = clr;

    // A barrier to make sure all threads finish writing to the
    // imageblock.
    // In this case, each thread writes to its location in the
    // imageblock so a barrier is not necessary.
threadgroup_barrier(mem_flags::mem_threadgroup_imageblock);

// Process the pixels in imageblock, and update the elements in
// slice.
process_pixels_in_imageblock(img_blk, gid, lid);

// A barrier to make sure all threads finish writing to the
// elements in the imageblock.
threadgroup_barrier(mem_flags::mem_threadgroup_imageblock);

// Write a specific element in an imageblock to the output
// image. Only one thread in the threadgroup performs the
// imageblock write.
if (lid.x == 0 && lid.y == 0)
    dst.write(img_blk.slice(f->a), gid);
}

### 6.14 Pack and Unpack Functions

This section lists the Metal functions, defined in the header `<metal_pack>`, for converting a vector floating-point data to and from a packed integer value. Refer to subsections of section 7.7 for details on how to convert from an 8-, 10-, or 16-bit signed or unsigned integer value to a normalized single- or half-precision floating-point value and vice-versa.

#### 6.14.1 Unpack and Convert Integers to a Floating-Point Vector

Table 6.22 lists functions that unpack multiple values from a single unsigned integer and then converts them into floating-point values that are stored in a vector.

<table>
<thead>
<tr>
<th>Built-in Unpack Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>float4 unpack_unorm4x8_to_float(uint x)</td>
<td>Unpack a 32-bit unsigned integer into four 8-bit signed or unsigned integers and then convert each 8-bit signed or unsigned integer value to a normalized single- or half-precision floating-point value to generate a 4-component vector.</td>
</tr>
<tr>
<td>float4 unpack_snorm4x8_to_float(uint x)</td>
<td></td>
</tr>
<tr>
<td>half4 unpack_unorm4x8_to_half(uint x)</td>
<td></td>
</tr>
<tr>
<td>half4 unpack_snorm4x8_to_half(uint x)</td>
<td></td>
</tr>
</tbody>
</table>
### Built-in Unpack Functions

<table>
<thead>
<tr>
<th>Function Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>float4 unpack_unorm4x8_srgb_to_float(uint x) half4 unpack_unorm4x8_srgb_to_half(uint x)</td>
<td>Unpack a 32-bit unsigned integer into four 8-bit signed or unsigned integers and then convert each 8-bit signed or unsigned integer value to a normalized single- or half-precision floating-point value to generate a 4-component vector. The r, g, and b color values are converted from sRGB to linear RGB.</td>
</tr>
<tr>
<td>float2 unpack_unorm2x16_to_float(uint x) float2 unpack_snorm2x16_to_float(uint x) half2 unpack_unorm2x16_to_half(uint x) half2 unpack_snorm2x16_to_half(uint x)</td>
<td>Unpack a 32-bit unsigned integer into two 16-bit signed or unsigned integers and then convert each 16-bit signed or unsigned integer value to a normalized single- or half-precision floating-point value to generate a 2-component vector.</td>
</tr>
<tr>
<td>float4 unpack_unorm10a2_to_float(uint x) float3 unpack_unorm565_to_float(ushort x) half4 unpack_unorm10a2_to_half(uint x) half3 unpack_unorm565_to_half(ushort x)</td>
<td>Convert a 10a2 (1010102) or 565 color value to the corresponding normalized single- or half-precision floating-point vector.</td>
</tr>
</tbody>
</table>

When converting from a 16-bit unsigned normalized or signed normalized value to a half-precision floating point, the unpack_unorm2x16_to_half and unpack_snorm2x16_to_half functions may lose precision.

#### 6.14.2 Convert Floating-Point Vector to Integers, then Pack the Integers

Table 6.23 lists functions that start with a floating-point vector, converts the components into integer values, and then packs the multiple values into a single unsigned integer.

### Table 6.23. Pack functions

<table>
<thead>
<tr>
<th>Built-in Pack Functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>uint pack_float_to_unorm4x8(float4 x) uint pack_float_to_snorm4x8(float4 x) uint pack_half_to_unorm4x8(half4 x) uint pack_half_to_snorm4x8(half4 x)</td>
<td>Convert a four-component vector normalized single- or half-precision floating-point value to four 8-bit integer values and pack these 8-bit integer values into a 32-bit unsigned integer.</td>
</tr>
</tbody>
</table>
## 6.15 Atomic Functions

The Metal programming language implements a subset of the C++14 atomics and synchronization operations. Metal atomic functions must operate on Metal atomic data, as described in section 2.6.

Atomic operations play a special role in making assignments in one thread visible to another thread. A synchronization operation on one or more memory locations is either an acquire operation, a release operation, or both. A synchronization operation without an associated memory location is a fence and can be either an acquire fence, a release fence, or both. In addition, there are relaxed atomic operations that are not synchronization operations.

There are only a few kinds of operations on atomic types, although there are many instances of those kinds. This section specifies each general kind.

Atomic functions are defined in the header `<metal_atomic>`.

### 6.15.1 Memory Order

The enumeration `memory_order` specifies the detailed regular (non-atomic) memory synchronization operations (see section 29.3 of the C++14 specification) and may provide for operation ordering. For atomic operations, `memory_order_relaxed` is the only enumeration value. With `memory_order_relaxed`, there are no synchronization or ordering constraints; the operation only requires atomicity. These operations do not order memory, but they guarantee atomicity and modification order consistency. A typical use for relaxed memory...
ordering is updating counters, such as reference counters since this only requires atomicity, but neither ordering nor synchronization.

6.15.2 Atomic Functions

In addition, accesses to atomic objects may establish interthread synchronization and order non-atomic memory accesses as specified by `memory_order`.

In the atomic functions described in the subsections of this section:

- A refers to one of the atomic types.
- C refers to its corresponding non-atomic type.
- M refers to the type of the other argument for arithmetic operations. For atomic integer types, M is C.

Note that each atomic function may support only some types. In each section below, we indicate which type A is supported.

**All OS:** Functions listed with names that end with `_explicit` (such as `atomic_store_explicit` or `atomic_load_explicit`) supported since Metal 1 unless otherwise indicated. Metal 3 supports the `atomic_float` for device memory only.

**iOS:** Support for the `atomic_store`, `atomic_load`, `atomic_exchange`, `atomic_compare_exchange_weak`, `atomic_fetch_key` functions since Metal 2.

6.15.2.1 Atomic Store Functions

These functions atomically replace the value pointed to by `object` with `desired`. These functions support atomic types A of `atomic_int`, `atomic_uint`, `atomic_bool`, and `atomic_float`.

**All OS:** Support for the `atomic_store_explicit` function with `memory_order_relaxed` supported, as indicated.

```c
void atomic_store_explicit(threadgroup A* object, C desired, memory_order order) // All OS: Since Metal 2.
void atomic_store_explicit(volatile threadgroup A* object, C desired, memory_order order) // All OS: Since Metal 1.
void atomic_store_explicit(device A* object, C desired, memory_order order) // All OS: Since Metal 2.
void atomic_store_explicit(volatile device A* object, C desired, memory_order order) // All OS: Since Metal 1.
```
6.15.2.2 Atomic Load Functions

These functions atomically obtain the value pointed to by object. These functions support atomic types A of atomic_int, atomic_uint, atomic_bool, and atomic_float.

All OS: Support for the atomic_load_explicit function with memory_order_relaxed supported, as indicated.

C atomic_load_explicit(const threadgroup A* object, memory_order order) // All OS: Since Metal 2.
C atomic_load_explicit(const volatile threadgroup A* object, memory_order order) // All OS: Since Metal 1.
C atomic_load_explicit(const device A* object, memory_order order) // All OS: Since Metal 2.
C atomic_load_explicit(const volatile device A* object, memory_order order) // All OS: Since Metal 1.

6.15.2.3 Atomic Exchange Functions

These functions atomically replace the value pointed to by object with desired and return the value object previously held. These functions support atomic types A of atomic_int, atomic_uint, atomic_bool, and atomic_float.

All OS: Support for the atomic_exchange_explicit function with memory_order_relaxed supported, as indicated.

C atomic_exchange_explicit(threadgroup A* object, C desired, memory_order order) // All OS: Since Metal 2.
C atomic_exchange_explicit(volatile threadgroup A* object, C desired, memory_order order) // All OS: Since Metal 1.
C atomic_exchange_explicit(device A* object, C desired, memory_order order) // All OS: Since Metal 2.
C atomic_exchange_explicit(volatile device A* object, C desired, memory_order order) // All OS: Since Metal 1.

6.15.2.4 Atomic Compare and Exchange Functions

These compare-and-exchange functions atomically compare the value in *object with the value in *expected. If those values are equal, the compare-and-exchange function performs a read-modify-write operation to replace *object with desired. Otherwise if those values are not equal, the compare-and-exchange function loads the actual value from *object into *expected. If the underlying atomic value in *object was successfully changed, the compare-and-exchange function returns true; otherwise it returns false. These functions
support atomic types A of atomic_int, atomic_uint, atomic_bool, and atomic_float.

Copying is performed in a manner similar to std::memcpy. The effect of a compare-and-exchange function is:

```c
if (memcmp(object, expected, sizeof(*object)) == 0) {
    memcpy(object, &desired, sizeof(*object));
} else {
    memcpy(expected, object, sizeof(*object));
}
```

All OS: Support for the atomic_compare_exchange_weak_explicit function supported as indicated; support for memory_order_relaxed for indicating success and failure. If the comparison is true, the value of success affects memory access, and if the comparison is false, the value of failure affects memory access.

```c
bool atomic_compare_exchange_weak_explicit(threadgroup A* object,
    C *expected, C desired, memory_order success,
    memory_order failure) // All OS: Since Metal 2.

bool atomic_compare_exchange_weak_explicit(volatile threadgroup A* object,
    C *expected, C desired, memory_order success,
    memory_order failure) // All OS: Since Metal 1.

bool atomic_compare_exchange_weak_explicit(device A* object,
    C *expected, C desired, memory_order success,
    memory_order failure) // All OS: Since Metal 2.

bool atomic_compare_exchange_weak_explicit(volatile device A* object,
    C *expected, C desired, memory_order success,
    memory_order failure) // All OS: Since Metal 1.
```

6.15.2.5 Atomic Fetch and Modify Functions

All OS: The following atomic fetch and modify functions are supported, as indicated.

The only supported value for order is memory_order_relaxed.

```c
C atomic_fetch_key_explicit(threadgroup A* object,
    M operand,
    memory_order order) // All OS: Since Metal 2.

C atomic_fetch_key_explicit(volatile threadgroup A* object,
    M operand,
    memory_order order) // All OS: Since Metal 1.

C atomic_fetch_key_explicit(device A* object,
    M operand,
    memory_order order) // All OS: Since Metal 2.
```
C atomic_fetch_key_explicit(volatile device A* object,
    M operand,
    memory_order order) // All OS: Since Metal 1.

The key in the function name is a placeholder for an operation name listed in the first column of Table 6.24, such as atomic_fetch_add_explicit. The operations detailed in Table 6.24 are arithmetic and bitwise computations. The function atomically replaces the value pointed to by object with the result of the specified computation (third column of Table 6.24). The function returns the value that object held previously. There are no undefined results.

These functions are applicable to any atomic object of type atomic_int, and atomic_uint. Add and sub are supported for atomic_float.

Table 6.24. Atomic operations

<table>
<thead>
<tr>
<th>Key</th>
<th>Operator</th>
<th>Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>add</td>
<td>+</td>
<td>Addition</td>
</tr>
<tr>
<td>and</td>
<td>&amp;</td>
<td>Bitwise and</td>
</tr>
<tr>
<td>max</td>
<td>max</td>
<td>Compute max</td>
</tr>
<tr>
<td>min</td>
<td>min</td>
<td>Compute min</td>
</tr>
<tr>
<td>or</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sub</td>
<td>−</td>
<td>Subtraction</td>
</tr>
<tr>
<td>xor</td>
<td>^</td>
<td>Bitwise exclusive or</td>
</tr>
</tbody>
</table>

These operations are atomic read-modify-write operations. For signed integer types, the arithmetic operation uses two’s complement representation with silent wrap-around on overflow.

6.15.2.6 Atomic Modify Functions (64 Bits)

All OS: Metal supports the following atomic modify functions since Metal 2.4. See the Metal Feature Set Tables to determine which GPUs support this feature.

These functions are applicable to any atomic object of type atomic_ulong. The only supported value for order is memory_order_relaxed.

void atomic_key_explicit(device A* object,
    M operand,
    memory_order order)

void atomic_key_explicit(volatile device A* object,
    M operand,
    memory_order order)
The key in the function name is a placeholder for an operation name listed in the first column of Table 6.25, such as atomic_max_explicit. The operations detailed in Table 6.25 are arithmetic. The function atomically replaces the value pointed to by \texttt{object} with the result of the specified computation (third column of Table 6.25). The function returns \texttt{void}. There are no undefined results.

<table>
<thead>
<tr>
<th>Key</th>
<th>Operator</th>
<th>Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>max</td>
<td>Compute max</td>
</tr>
<tr>
<td>min</td>
<td>min</td>
<td>Compute min</td>
</tr>
</tbody>
</table>

These operations are atomic read-modify-write operations.

6.16 Encoding Commands for Indirect Command Buffers

Indirect Command Buffers (ICBs) support the encoding of Metal commands into a Metal buffer for repeated use. Later, you can submit these encoded commands to the CPU or GPU for execution. ICBs for both render and compute commands use the \texttt{command_buffer} type to encode commands into an ICB object (represented in the Metal framework by \texttt{MTLIndirectCommandBuffer}):

```c
struct command_buffer {
    size_t size() const;
};
```

An ICB can contain either render or compute commands but not both. Execution of compute commands from a render encoder is illegal. So is execution of render commands from a compute encoder.

6.16.1 Encoding Render Commands in Indirect Command Buffers

All OS: Indirect Command Buffers for render commands are supported since Metal 2.1. ICBs allow the encoding of draw commands into a Metal buffer for subsequent execution on the GPU.

In a shading language function, use the \texttt{command_buffer} type to encode commands for ICBs into a Metal buffer object that provides indexed access to a \texttt{render_command} structure.

```c
struct arguments {
    command_buffer cmd_buffer;
};

kernel void producer(device arguments &args,
                      ushort cmd_idx [[thread_position_in_grid]])
{
```
render_command cmd(args.cmd_buffer, cmd_idx);
...
}

render_command can encode any draw command type. The following public interface for render_command is defined in the header <metal_command_buffer>. To pass render_pipeline_state objects to your shader, use argument buffers. Within an argument buffer, the pipeline state can be passed as scalars or in an array.

set_render_pipeline_state(...) and render pipeline states are available on macOS since Metal 2.1 and on iOS since Metal 2.2.

enum class primitive_type { point, line, line_strip, triangle, triangle_strip };

struct render_command {
    public:
        explicit render_command(command_buffer icb, unsigned cmd_index);
        void set_render_pipeline_state(
            render_pipeline_state pipeline_state);

        template <typename T …>
        void set_vertex_buffer(device T *buffer, uint index);
        template <typename T …>
        void set_vertex_buffer(constant T *buffer, uint index);

        // Metal 3.1: Supported passing vertex strides
        template <typename T …>
        void set_vertex_buffer(device T *buffer, size_t stride,
            uint index);
        template <typename T …>
        void set_vertex_buffer(constant T *buffer, size_t stride,
            uint index);

        template <typename T …>
        void set_fragment_buffer(device T *buffer, uint index);
        template <typename T …>
        void set_fragment_buffer(constant T *buffer, uint index);

        void draw_primitives(primitive_type type, uint vertex_start,
            uint vertex_count, uint instance_count,
            uint base_instance);

        // Overloaded draw_indexed_primitives based on index_buffer
        void draw_indexed_primitives(primitive_type type,
            uint index_count,
            device ushort *index_buffer,
void draw_indexed_primitives(primitive_type type,
    uint index_count,
    device uint *index_buffer,
    uint instance_count,
    uint base_vertex,
    uint base_instance);

void draw_indexed_primitives(primitive_type type,
    uint index_count,
    constant ushort *index_buffer,
    uint instance_count,
    uint base_vertex,
    uint base_instance);

void draw_indexed_primitives(primitive_type type,
    uint index_count,
    constant uint *index_buffer,
    uint instance_count,
    uint base_vertex,
    uint base_instance);

// Overloaded draw_patches based on patch_index_buffer and
tessellation_factor_buffer
void draw_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    uint instance_count, uint base_instance,
    const device MTLQuadTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    uint instance_count, uint base_instance,
    const device
    MTLTriangleTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
uint instance_count, uint base_instance,
        constant MTLQuadTessellationFactorsHalf
        *tessellation_factor_buffer,
        uint instance_stride = 0);

void draw_patches(uint number_of_patch_control_points,
        uint patch_start, uint patch_count,
        const device uint *patch_index_buffer,
        uint instance_count, uint base_instance,
        constant MTLTriangleTessellationFactorsHalf
        *tessellation_factor_buffer,
        uint instance_stride = 0);

void draw_patches(uint number_of_patch_control_points,
        uint patch_start, uint patch_count,
        const uint *patch_index_buffer,
        uint instance_count, uint base_instance,
        const device MTLQuadTessellationFactorsHalf
        *tessellation_factor_buffer,
        uint instance_stride = 0);

void draw_patches(uint number_of_patch_control_points,
        uint patch_start, uint patch_count,
        const uint *patch_index_buffer,
        uint instance_count, uint base_instance,
        const device
        MTLTriangleTessellationFactorsHalf
        *tessellation_factor_buffer,
        uint instance_stride = 0);

void draw_patches(uint number_of_patch_control_points,
        uint patch_start, uint patch_count,
        const uint *patch_index_buffer,
        uint instance_count, uint base_instance,
        constant MTLQuadTessellationFactorsHalf
        *tessellation_factor_buffer,
        uint instance_stride = 0);

void draw_patches(uint number_of_patch_control_points,
        uint patch_start, uint patch_count,
        constant uint *patch_index_buffer,
        uint instance_count, uint base_instance,
        constant MTLQuadTessellationFactorsHalf
        *tessellation_factor_buffer,
        uint instance_stride = 0);

// Overloaded draw_indexed_patches based on patch_index_buffer,
// control_point_index_buffer and tessellation_factor_buffer

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    const device void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    const device MTLQuadTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    const device void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    const device MTLTriangleTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    const device void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    constant MTLQuadTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    const device void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    constant MTLTriangleTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    constant void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    const device MTLQuadTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    const device void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    constant MTLQuadTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    const device uint *patch_index_buffer,
    constant void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    constant MTLTriangleTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    constant uint *patch_index_buffer,
    const device void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    constant MTLTriangleTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    constant uint *patch_index_buffer,
    const device void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    constant device MTLTriangleTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
    uint patch_start, uint patch_count,
    constant uint *patch_index_buffer,
    const device void *control_point_index_buffer,
    uint instance_count, uint base_instance,
    constant device MTLTriangleTessellationFactorsHalf
    *tessellation_factor_buffer,
    uint instance_stride = 0);
uint instance_count, uint base_instance,
constant MTLQuadTessellationFactorsHalf
 *tessellation_factor_buffer,
uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
 uint patch_start, uint patch_count,
constant uint *patch_index_buffer,
const device void *control_point_index_buffer,
uint instance_count, uint base_instance,
constant MTLTriangleTessellationFactorsHalf
 *tessellation_factor_buffer,
uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
 uint patch_start, uint patch_count,
constant uint *patch_index_buffer,
constant void *control_point_index_buffer,
uint instance_count, uint base_instance,
const device MTLQuadTessellationFactorsHalf
 *tessellation_factor_buffer,
uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
 uint patch_start, uint patch_count,
constant uint *patch_index_buffer,
constant void *control_point_index_buffer,
uint instance_count, uint base_instance,
constant MTLQuadTessellationFactorsHalf
 *tessellation_factor_buffer,
uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
 uint patch_start, uint patch_count,
constant uint *patch_index_buffer,
constant void *control_point_index_buffer,
uint instance_count, uint base_instance,
constant MTLQuadTessellationFactorsHalf
 *tessellation_factor_buffer,
uint instance_stride = 0);

void draw_indexed_patches(uint number_of_patch_control_points,
 uint patch_start, uint patch_count,
constant uint *patch_index_buffer,
constant void *control_point_index_buffer,
uint instance_count, uint base_instance,
constant MTLQuadTessellationFactorsHalf
 *tessellation_factor_buffer,
// Reset the entire command. After reset(), without further
// modifications, execution of this command shall not perform
// any action.
void reset();

// Copy the content of the `source` command into this command.
void copy_command(render_command source);
};

When accessing command_buffer, Metal does not check whether the access is within
bounds. If an access is beyond the capacity of the buffer, the behavior is undefined.

The exposed methods in render_command mirror the interface of
MTLIndirectRenderCommand and are similar to MTLRenderCommandEncoder. Notable
differences with MTLRenderCommandEncoder are:

- Calls to draw* methods in render_command encode the actions taken by the
  command. If multiple calls are made, only the last one takes effect.
- The tessellation arguments are passed directly in render_command::draw_patches
  and render_command::draw_indexed_patches. Other calls do not set up the
tessellation arguments.

6.16.2 Encoding Compute Commands in Indirect Command Buffers

iOS: Indirect Command Buffers for compute commands are supported since Metal 2.2.
macOS: Indirect Command Buffers for compute commands are supported since Metal 2.3.

ICBs allow the encoding of dispatch commands into a Metal buffer for subsequent execution on
the GPU.

In a shading language function, use the command_buffer type to encode commands for ICBs
into a Metal buffer object that provides indexed access to a compute_command structure.

```c
struct arguments {
    command_buffer cmd_buffer;
};

[[kernel]] void producer(device arguments &args,
                          ushort cmd_idx [[thread_position_in_grid]])
{
    compute_command cmd(args.cmd_buffer, cmd_idx);
    ...
}
```

compute_command can encode any dispatch command type. The following public interface
for compute_command is defined in the header <metal_command_buffer>. The
compute_pipeline_state type represents compute pipeline states, which can only be
passed to shaders through argument buffers. Within an argument buffer, the pipeline state can be passed as scalars or in an array.

```cpp
struct compute_command {
    public:
        explicit compute_command(command_buffer icb,
                                  unsigned cmd_index);

        void set_compute_pipeline_state(
            compute_pipeline_state pipeline);

        template <typename T …>
        void set_kernel_buffer(device T *buffer, uint index);
        template <typename T …>
        void set_kernel_buffer(constant T *buffer, uint index);

        // Metal 3.1: Supports passing kernel strides
        template <typename T …>
        void set_kernel_buffer(device T *buffer, size_t stride,
                                uint index);
        template <typename T …>
        void set_kernel_buffer(constant T *buffer, size_t stride,
                                uint index);

        void set_barrier();
        void clear_barrier();

        void concurrent_dispatch_threadgroups(
            uint3 threadgroups_per_grid,
            uint3 threads_per_threadgroup);
        void concurrent_dispatch_threads(uint3 threads_per_grid,
                                           uint3 threads_per_threadgroup);

        void set_threadgroup_memory_length(uint length, uint index);
        void set_stage_in_region(uint3 origin, uint3 size);

        // Reset the entire command. After reset(), without further
        // modifications, execution of this command shall not perform
        // any action.
        void reset();

        // Copy the content of the `source` command into this command.
        void copy_command(compute_command source);
};
```

When accessing `command_buffer`, Metal does not check whether the access is within bounds. If an access is beyond the capacity of the buffer, the behavior is undefined.

The exposed methods in `compute_command` mirror the interface of `MTLIndirectComputeCommand` and are similar to `MTLComputeCommandEncoder`. 
In an ICB, dispatches are always concurrent. Calls to the concurrent_dispatch* methods in compute_command encode the actions taken by the command. If multiple calls are made, only the last one takes effect.

The application is responsible for putting barriers where they are needed. Barriers encoded in an ICB do not affect the parent encoder.

The CPU may have initialized individual commands within a command_buffer before the command_buffer is passed as an argument to a shader. If the CPU has not already initialized a command, you must reset that command before using it.

### 6.16.3 Copying Commands of an Indirect Command Buffer

Copying a command structure (either render_command or compute_command) via operator= does not copy the content of the command, but only makes the destination command point to the same buffer and index as the source command. To copy the content of the command, call the copy_command functions listed in sections 6.16.1 and 6.16.2.

Copying is only supported between commands pointing to compatible command buffers. Two command buffers are compatible only if they have matching ICB descriptors (MTLIndirectCommandBufferDescriptor objects). The commands themselves must also refer to valid indexes within the buffers. The following example illustrates using copy_command to copy the content of a render command from cmd0 to cmd1:

```plaintext
struct arguments {
    command_buffer cmd_buffer;
    render_pipeline_state pipeline_state_0;
    render_pipeline_state pipeline_state_1;
};

[[kernel]] void producer(device arguments &args) {
    render_command cmd0(args.cmd_buffer, 0);
    render_command cmd1(args.cmd_buffer, 1);
    cmd0.set_render_pipeline_state(args.pipeline_state_0);

    // Make the command at index 1 point to command at index 0.
    cmd1 = cmd0;

    // Change the pipeline state for the command at index 0 in the // buffer.
    cmd1.set_render_pipeline_state(args.pipeline_state_0);

    // The command at index 1 in the buffer is not yet modified.
    cmd1 = render_command(args.cmd_buffer, 1);

    // Copy the content of the command at index 0 to command at // index 1.
    cmd1.copy_command(cmd0);
}
```
### 6.17 Variable Rasterization Rate

iOS: Variable Rasterization Rate and the Rasterization Rate Map supported since Metal 2.2.
macOS: Variable Rasterization Rate and the Rasterization Rate Map since Metal 2.3.

Variable rasterization rate (VRR) can reduce the shading cost of high-resolution rendering by reducing the fragment shader invocation rate based on screen position. VRR is especially useful to avoid oversampling peripheral information in AR/VR applications.

To support VRR in a shading language function, use the `rasterization_rate_map_decoder` structure to describe the mapping of per-layer rasterization rate data. Each layer contains minimum quality values in screen space and can have a different physical fragment space dimension. For AR/VR, these quality values are based on the lens transform and/or eye-tracking information.

```c
struct rasterization_rate_map_data;

struct rasterization_rate_map_decoder {
    explicit rasterization_rate_map_decoder(
        constant rasterization_rate_map_data &data) thread;

    float2 map_screen_to_physical_coordinates(float2 screen_coordinates,
        uint layer_index = 0) const thread;
    uint2 map_screen_to_physical_coordinates(uint2 screen_coordinates,
        uint layer_index = 0) const thread;
    float2 map_physical_to_screen_coordinates(float2 physical_coordinates,
        uint layer_index = 0) const thread;
    uint2 map_physical_to_screen_coordinates(uint2 physical_coordinates,
        uint layer_index = 0) const thread;
};
```

The VRR map describes the mapping between screen space and physical fragment space and enables conversion of the rendering results back to the desired screen resolution. To convert between screen space and physical fragment space in the shader, the app must call the `copyParameterDataToBuffer:offset:` method of `MTLRasterizationRateMap` to fill the buffer with map data before using any of the conversion functions in the `rasterization_rate_map_decoder` structure. Passing anything other than a pointer to the data exported by the `copyParameterDataToBuffer:offset:` method has an undefined behavior.

The following example shows how the app must pass the `rasterization_rate_map_data` at the shader bind point to the constructor of the `rasterization_rate_map_decoder` structure:

```c
[[fragment]] float4 fragment_shader(/* other arguments */
    constant rasterization_rate_map_data &data [[buffer(0)]]) {
    float2 screen_coords = ...;
    rasterization_rate_map_decoder map(data);
    float2 physical_coords =
        map.map_screen_to_physical_coordinates(screen_coords);
    ...
}
```
Alternately, the app can compute the offset where the compiled data is stored and use an explicit cast or pointer arithmetic to form the data for a valid `rasterization_rate_map_data`. Since `rasterization_rate_map_data` is an incomplete type, some operations on it are inherently forbidden (such as pointer arithmetic on the pointer type or `sizeof`).

### 6.18 Ray Tracing Functions

All OS: Ray tracing functions are supported since Metal 2.3. The ray tracing functions and types are defined in `<metal_raytracing>` in the namespace `metal::raytracing`. In Metal 2.3, they are only supported in a compute function (kernel functions) except where noted below. As of Metal 2.4, they are also supported in vertex, fragment, and tile functions.

#### 6.18.1 Acceleration Structure Functions

In Metal 2.3, call one of the following functions to check if an acceleration structure (see section 2.17.6) is null:

```cpp
bool is_null_primitive_acceleration_structure(primitive_acceleration_structure)
```

```cpp
bool is_null_instance_acceleration_structure(instance_acceleration_structure)
```

In Metal 2.4, call the following function to check if an acceleration structure is null.

```cpp
bool is_null_acceleration_structure(acceleration_structure<intersection_tags...>)
```

In Metal 3.1. Metal supports iterate over the acceleration structure referenced by an instance acceleration structure using the following functions:

Call the following function to query the number of instances in an instance acceleration structure.

```cpp
uint get_instance_count() const
```
Call the following function to retrieve the acceleration structure referenced by an instance contained in an instance acceleration structure. The return type is the templatized type defined in section 2.17.6.

```cpp
template <typename... intersection_tags>
acceleration_structure< intersection_tags...>
```

If the declared return type does not match the acceleration structure type reference by the instance contained in an instance acceleration structure, then the results are undefined. Instance acceleration structures that do not use instance and/or primitive motion tags can be returned as an acceleration structure type that does contain those tags. For example, an instance acceleration structure without any motion (instance or primitive) can be returned as:

- `acceleration_structure<instancing>`
- `acceleration_structure<instancing, instance_motion>`
- `acceleration_structure<instancing, primitive_motion>`
- `acceleration_structure<instancing, primitive_motion, instance_motion>`

This capability allows developers to avoid providing a dedicated intersector for each set of tags when working with multiple acceleration structure types at the potential performance cost due to traversing an acceleration structure that does not require those tags.

### 6.18.2 Intersector Intersect Functions

After creating the `intersector<intersection_tags...>` object (see section 2.17.5), you can call one of the following `intersect` functions based on the value of the `intersection_tags`.

<table>
<thead>
<tr>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>result_type intersect(...parameters...)</code>.</td>
</tr>
</tbody>
</table>

Table 6.27 shows the possible parameters for `intersect` function. All `intersect` functions must have `ray` and `accel_struct` parameter. The other parameters are optional.
### Table 6.27. Intersect functions input parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ray</td>
<td>Ray properties</td>
</tr>
<tr>
<td>accel_struct</td>
<td>Acceleration structure of type acceleration_structure&lt;intersection_tags...&gt;.</td>
</tr>
<tr>
<td>mask</td>
<td>Intersection mask to be AND'd with instance mask defined in the Metal API MTLAccelerationStructureInstanceDescriptor. Instances with non-overlapping masks will be skipped.</td>
</tr>
<tr>
<td>time</td>
<td>All OS: Since Metal 2.4.</td>
</tr>
<tr>
<td></td>
<td>The time associated with the ray. The parameter exists if the intersection_tags have primitive_motion or instance_motion.</td>
</tr>
<tr>
<td>func_table</td>
<td>Intersection function table of type intersection_function_table&lt;intersection_tags...&gt;. See section 2.17.3.</td>
</tr>
<tr>
<td>payload</td>
<td>User payload object, which is passed by reference. When the user calls intersect(), the payload parameter is copied to the ray_data address space and passed to the intersection function. The result is copied on the exit of the intersection function (section 5.1.6) and the payload object is updated.</td>
</tr>
</tbody>
</table>

The `result_type` is

```cpp
using result_type = intersection_result<intersection_tags...>;
```

The following set of intersect functions are available only if `intersection_tags` does not have instancing:

```cpp
result_type intersect(
    ray ray,
    primitive_acceleration_structure accel_struct) const;
```

```cpp
result_type intersect(
    ray ray,
    primitive_acceleration_structure accel_struct,
    intersection_function_table<intersection_tags...> func_table) const;
```
template <typename T>
result_type
intersect(
    ray ray,
    primitive_acceleration_structure accel_struct,
    intersection_function_table<intersection_tags...> func_table,
    thread T &payload) const;

The following set of intersect functions are available only if intersection_tags have instancing.

result_type
intersect(
    ray ray,
    instance_acceleration_structure accel_struct,
    uint mask = ~0U) const;

result_type
intersect(
    ray ray,
    instance_acceleration_structure accel_struct,
    intersection_function_table<intersection_tags...> func_table) const;

template <typename T>
result_type
intersect(
    ray ray,
    instance_acceleration_structure accel_struct,
    intersection_function_table<intersection_tags...> func_table,
    thread T &payload) const;

result_type
intersect(
    ray ray,
    instance_acceleration_structure accel_struct,
    uint mask,
    intersection_function_table<intersection_tags...> func_table) const;

template <typename T>
result_type
intersect(
As of Metal 2.4, the following set of intersect functions are available if `intersection_tags` have `primitive_motion` or `instance_motion`.

```cpp
template <typename T, intersection_tags...>
result_type intersect(
  ray ray,
  instance_acceleration_structure accel_struct,
  uint mask,
  intersection_function_table<intersection_tags...> func_table,
  thread T &payload) const;
```

```cpp
template <typename T, intersection_tags...>
result_type intersect(
  ray ray,
  acceleration_structure<intersection_tags...> accel_struct,
  float time) const;
```

```cpp
template <typename T, intersection_tags...>
result_type intersect(
  ray ray,
  acceleration_structure<intersection_tags...> accel_struct,
  float time,
  intersection_function_table<intersection_tags...> func_table) const;
```

As of Metal 2.4, the following set of intersect functions are available only if `intersection_tags` have `instancing` and either `primitive_motion` or `instance_motion`.

```cpp
template <typename T, intersection_tags...>
result_type intersect(
  ray ray,
  acceleration_structure<intersection_tags...> accel_struct,
  uint mask = ~0U,
  float time = 0.0f) const;
```
template <typename T, intersection_tags...>
  result_type
  intersect(
    ray ray,
    acceleration_structure< intersection_tags...> accel_struct,
    uint mask,
    float time,
    intersection_function_table<intersection_tags...> func_table)
  const;

template <typename T, intersection_tags...>
  result_type
  intersect(
    ray ray,
    acceleration_structure< intersection_tags...> accel_struct,
    uint mask,
    float time,
    intersection_function_table<intersection_tags...> func_table,
    thread T &payload)
  const;

6.18.3 Intersector functions to control traversal behavior

All OS: Metal 3.1 adds support for curves.

To override the default behavior of the traversal, one can use the following member functions of intersector<intersection_tags...> object.
Table 6.28. Intersect functions to control traversal

<table>
<thead>
<tr>
<th>Functions to control traversal behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>void set_triangle_front_facing_winding(winding)</td>
</tr>
<tr>
<td>void set_geometry_cull_mode(geometry_cull_mode)</td>
</tr>
<tr>
<td>void set_opacity_cull_mode(opacity_cull_mode)</td>
</tr>
<tr>
<td>void force_opacity(forced_opacity)</td>
</tr>
<tr>
<td>void assume_geometry_type(geometry_type)</td>
</tr>
<tr>
<td>void assume_identity_transforms(bool)</td>
</tr>
<tr>
<td>void accept_any_intersection(bool)</td>
</tr>
</tbody>
</table>

Triangles have two sides or "faces". The front facing winding determines which triangle face is considered the "front" face when viewed from the ray origin. If the vertices appear in clockwise order when viewed from the ray origin and the front facing winding is clockwise, then the visible face is the front face. The other face is the back face. If the front facing winding is counterclockwise, then the opposite is true. Use the following function to change the default winding (clockwise):

```cpp
tenum class winding {
    clockwise,
    counterclockwise
};
void set_triangle_front_facing_winding(winding w);
```

To change the default triangle cull mode (none), use the following function.

```cpp
tenum class triangle_cull_mode {
    none,
    front,
    back
};
void set_triangle_cull_mode(triangle_cull_mode tcm);
```

If the cull mode is set to front, then triangles whose front face is visible from the ray origin are not considered for intersection. Otherwise, if the cull mode is set to back, then triangles whose back face is visible from the ray origin are not considered for intersection.

The following function may be used to set the intersector to cull all bounding box or triangle primitives from the set of candidate geometries. The default geometry cull mode is none.

```cpp
tenum class geometry_cull_mode {
    none,
```
triangle,
bounding_box,
curve // Since Metal 3.1
};

void set_geometry_cull_mode(geometry_cull_mode gcm);

The default opacity cull mode is none. Use the following function to change the opacity. See
below on how opacity will affect triangle and bounding box primitives.
enum class opacity_cull_mode {
  none,
  opaque,
  non_opaque
};
void set_opacity_cull_mode(opacity_cull_mode ocm);

Call the following function to override per-instance and per-geometry setting of forced
capacity. The default is none.
enum class forced_opacity {
  none,
  opaque,
  non_opaque
};
void force_opacity(forced_opacity fo);

Triangle primitives may also be culled based on their opacity: An opaque triangle will not run
any intersection function. A non_opaque triangle will run its intersection function to accept or
reject the hit.

The PrimitiveAccelerationStructure encodes if the triangle is opaque or
non_opaque by declaring MTLAccelerationStructureGeometryFlagOpaque. The
opaqueness can be overridden by calling intersector.force_opacity(). If used, this
takes precedence over the per-instance opaqueness flags
(MTLAccelerationStructureInstanceFlagOpaque and
MTLAccelerationStructureInstanceFlagNonOpaque), which in turn takes
precedence over the per-geometry opaqueness.

For custom bounding box primitives, the opaqueness will be evaluated in the same way as
described for triangles (first intersector.set_opacity_cull_mode(), then
InstanceFlags, then GeometryFlags). The opaque parameter informs the bounding box
intersection program the resolved opaqueness state. The intersection function may then use
this to influence its evaluation of if a hit is encountered or not.

intersector.set_opacity_cull_mode() will skip over primitive types based on their
opaqueness.
If `intersector.force_opacity()` is set to opaque or nonOpaque then `intersector.set_opacity_cull_mode()` must be none. The reverse is also true: Opacity Override and Opacity culling cannot be mixed. The results of illegal combinations are undefined.

Use the following functions to declare if the acceleration structure contains a triangle, bounding box, and/or curve geometry. The default geometry is `geometry_type::triangle | geometry_type::bounding_box`. By default, Metal assumes acceleration structure will not contain curve geometry to improve performance. Call `assume_geometry_type` with a value that includes `geometry_type::curve` to enable curves to be intersected in an intersect call or intersection query step.

```cpp
e num class geometry_type {
  none,
  triangle,
  bounding_box,
  curve,    // Since Metal 3.1
  all
};
void assume_geometry_type(geometry_type gt)
```

To set the intersector object to assume identify transforms, call the following function with the value `true`. The default is `false`.

```cpp
void assume_identity_transforms(bool value);
```

To set the intersector object to immediately return the first intersection it finds, call the following function with the value `true`. The default is `false`. One use of this function is when you only need to know if one point is visible from another, such as when rendering shadows or ambient occlusion.

```cpp
void accept_any_intersection(bool value);
```

Starting from Metal 3.1, use the following functions to add hints to the `intersector` and `intersection_query` to specify the curve basis, the number of control points, and the curve type to optimize traversal for specific curve types. Note that `curve_basis` is a enum and not a bitmask.

```cpp
e num class curve_basis {
  bspline,
  catmull_rom,
  linear,
  bezier,
  all,
};
```
enum class curve_type {
    round,
    flat,
    all,
};

Use the following function to set the curve basis function to assume. Defaults to `curve_basis::all`, meaning that all curve basis functions will be enabled.

```cpp
void assume_curve_basis(curve_basis cb)
```

Use the following function to set the curve type to assume. Defaults to `curve_type::all`, meaning that both curve types will be enabled.

```cpp
void assume_curve_type(curve_type ct)
```

Use the following function to set the number of curve control points to assume. Defaults to 0, meaning that any number of control points, as appropriate for the assumed curve basis (if any), will be enabled. Other valid options are 2, 3, or 4, depending on the curve basis.

```cpp
void assume_curve_control_point_count(uint n)
```

### 6.18.4 Intersection Query Functions

**All OS: Intersection Query is supported since Metal 2.4**

**All OS: Intersection Query supports curves since Metal 3.1**

To start traversals and query traversal specific information, create an intersection query object (see section 2.17.7) with a non-default constructor or first call `reset(...)`. If not called in this sequence, the behavior is undefined.

Table 6.29, Table 6.31, and Table 6.32 show the list of functions that can be called depending on the geometry type encountered during the traversal, assuming `next()` has returned `true`. Note that some functions come in pairs: a candidate and a committed primitive. When `next()` is called for the first time, the primitive reported after the traversal is always a candidate until the user commits the primitive by calling `commit_triangle_intersection()`, `commit_bounding_box_intersection()`, or `commit_curve_intersection()` on the query object. Note that opaque triangles, tested without user intersection, commit automatically when intersected.
### Table 6.29. Intersection query functions

<table>
<thead>
<tr>
<th>Functions</th>
<th>Triangle</th>
<th>Bounding</th>
<th>Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>void reset(...)</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><code>bool next()</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><code>void abort()</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><code>intersection_type get_candidate_intersection_type()</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><code>intersection_type get_committed_intersection_type()</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><code>void commit_triangle_intersection()</code></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>void commit_bounding_box_intersection(float distance)</code></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td><code>void commit_curve_intersection()</code></td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>

Since Metal 3.1

Starting with Metal 3.1, intersection query supports the following functions when specified with the `max_levels<Count>` intersection tags.

### Table 6.30. Intersection query functions with max_levels<Count>

<table>
<thead>
<tr>
<th>Functions</th>
<th>Triangle</th>
<th>Bounding</th>
<th>Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>uint get_candidate_instance_count()</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Since Metal 3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>uint get_candidate_instance_id(uint depth)</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Since Metal 3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>uint get_candidate_user_instance_id(uint depth)</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Since Metal 3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>uint get_committed_instance_count()</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Since Metal 3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>uint get_committed_instance_id(uint depth)</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Since Metal 3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>uint get_committed_user_instance_id(uint depth)</code></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Since Metal 3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 6.31. Intersection query ray value functions

<table>
<thead>
<tr>
<th>Ray Values Functions</th>
<th>Triangle</th>
<th>Bounding</th>
<th>Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>float3 get_world_space_ray_origin()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float3 get_world_space_ray_direction()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float get_ray_min_distance()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>intersection_params get_intersection_params()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

### Table 6.32. Intersection query candidate value functions

<table>
<thead>
<tr>
<th>Candidate Intersections Value Functions</th>
<th>Triangle</th>
<th>Bounding</th>
<th>Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>float get_candidate_triangle_distance()</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uint get_candidate_instance_id()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>uint get_candidate_user_instance_id()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>uint get_candidate_geometry_id()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>uint get_candidate_primitive_id()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float2 get_candidate_triangle_barycentric_coord()</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bool is_candidate_non_opaque_bounding_box()</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>bool is_candidate_triangle_front_facing()</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>float4x3 get_candidate_object_to_world_transform()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float4x3 get_candidate_world_to_object_transform()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float3 get_candidate_ray_origin()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float3 get_candidate_ray_direction()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>const device void * get_candidate_primitive_data()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Since Metal 3
## Table 6.33. Intersect query committed value functions

<table>
<thead>
<tr>
<th>Committed Intersections Value Functions</th>
<th>Triangle</th>
<th>Bounding</th>
<th>Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>float get_committed_distance()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>uint get_committed_instance_id()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>uint get_committed_user_instance_id()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>uint get_committed_geometry_id()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>uint get_committed_primitive_id()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float2 get_committed_triangle_barycentric_coord()</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bool is_committed_triangle_front_facing()</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>float4x3 get_committed_object_to_world_transform()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float4x3 get_committed_world_to_object_transform()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float3 get_committed_ray_origin()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>float3 get_committed_ray_direction()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>const device void * get_committed_primitive_data()</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>Since Metal 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>float get_candidate_curve_parameter()</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td><strong>Since Metal 3.1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>float get_committed_curve_parameter()</td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>

Call the following function to query the distance of a candidate triangle hit that needs consideration.

```c
float get_candidate_triangle_distance()
```

Call the following function to query the distance of the currently committed hit.

```c
float get_committed_distance()
```
Call the following function to query the top level structure instance ID for the current candidate hit.

```c
uint get_candidate_instance_id()
```

Call the following function to query user instance ID provided by user on the bottom level acceleration structure for the current candidate hit.

```c
uint get_candidate_user_instance_id()
```

Call the following function to query the bottom level structure geometry ID for the current candidate hit.

```c
uint get_candidate_geometry_id()
```

Call the following function to query the bottom level structure primitive ID within the geometry for the current candidate hit.

```c
uint get_candidate_primitive_id()
```

Call the following function to query the top level structure instance ID for the current committed hit.

```c
uint get_committed_instance_id()
```

Call the following function to query user instance ID provided by user on the bottom level acceleration structure for the current committed hit.

```c
uint get_committed_user_instance_id()
```

Call the following function to query the bottom level structure geometry ID for the current committed hit.

```c
uint get_committed_geometry_id()
```

Call the following function to query the bottom level structure primitive ID within the geometry for the current committed hit.

```c
uint get_committed_primitive_id()
```

Call the following function to query the ray origin in object space for the current hit candidate.

```c
float3 get_candidate_ray_origin()
```
Call the following function to query the ray direction in object space for the current hit candidate.

```c
float3 get_candidate_ray_direction()
```

Call the following function to query the ray origin in object space for the current committed hit.

```c
float3 get_committed_ray_origin()
```

Call the following function to query the ray direction in object space for the current committed hit.

```c
float3 get_committed_ray_direction()
```

Call the following function to query the matrix for transforming ray origin/direction of current hit candidate from object-space to world-space.

```c
float4x3 get_candidate_object_to_world_transform()
```

Call the following function to query the matrix for transforming ray origin/direction of current committed hit from world-space to object-space.

```c
float4x3 get_committed_object_to_world_transform()
```

Call the following function to query the matrix for transforming ray origin/direction of current candidate hit from world-space to object-space.

```c
float4x3 get_candidate_world_to_object_transform()
```

Call the following function to query the matrix for transforming ray origin/direction of current committed hit from object-space to world-space.

```c
float4x3 get_committed_world_to_object_transform()
```

Call the following function to query the candidate hit location barycentric coordinates. Valid when `get_candidate_intersection_type()` returns triangle.

```c
float2 get_candidate_triangle_barycentric_coord()
```

For vertex attributes v0, v1, and v2, the value at the specified barycentric point is:

```c
v1 * barycentric_coord.x +
v2 * barycentric_coord.y +
v0 * (1.0f - (barycentric_coord.x + barycentric_coord.y))
```

Call the following function to query the committed hit location barycentric coordinates. Valid when `get_committed_intersection_type()` returns triangle.

```c
float2 get_committed_triangle_barycentric_coord()
```
Call the following function to query if the hit triangle candidate is front or back facing. Returns true if it is front face and false if it is back face. Valid when get_candidate_intersection_type() returns triangle.

    bool is_candidate_triangle_front_facing()

Call the following function to query if the committed hit is front or back facing. Returns true if it is front face and false if it is back face. Valid when get_committed_intersection_type() returns triangle.

    bool is_committed_triangle_front_facing()

Call the following function to query the per-primitive data for the current candidate primitive.

    const device void *get_candidate_primitive_data()

Call the following function to query the per-primitive data for the current committed hit.

    const device void *get_committed_primitive_data()

Starting with Metal 3.1, the following two functions can be called when get_candidate_intersection_type() returns curve and the intersection tag has curve_data:

Call the following to query the curve parameter for the current candidate curve.

    float get_candidate_curve_parameter()

Call the following to query the curve parameter for the current committed intersection. Valid when get_candidate_intersection_type() returns curve.

    float get_committed_curve_parameter()

Starting with Metal 3.1, the rest of the functions in this section can be called when the intersection tag has max_levels<Count>:

Call the following function to query the number of instances in the candidate intersection.

    uint get_candidate_instance_count()

Call the following function to query the instance ID at level depth in the candidate intersection.

    uint get_candidate_instance_id(uint depth)

Call the following function to query the user instance ID at level depth in the candidate intersection.

    uint get_candidate_user_instance_id(uint depth)
Call the following function to query the number of instances in the committed intersection.

```c
uint get_committed_instance_count()
```

Call the following function to query the instance ID at level `depth` in the committed intersection.

```c
uint get_committed_instance_id(uint depth)
```

Call the following function to query the user instance ID at level `depth` in the committed intersection.

```c
uint get_committed_user_instance_id(uint depth);
```

### 6.18.5 Indirect Instance Descriptors

Starting in Metal 3.1, developers can fill out indirect instance descriptors from the GPU. Metal provides the following type definitions:

```c
enum MTLAccelerationStructureInstanceOptions : uint {
    MTLAccelerationStructureInstanceOptionNone = 0,
    MTLAccelerationStructureInstanceOptionDisableTriangleCulling = (1 << 0),
    MTLAccelerationStructureInstanceOptionTriangleFrontFacingWindingCounterClockwise = (1 << 1),
    MTLAccelerationStructureInstanceOptionOpaque = (1 << 2),
    MTLAccelerationStructureInstanceOptionNonOpaque = (1 << 3),
};
```

```c
typedef packed_float3 MTLPackedFloat3;
typedef packed_float3 MTLPackedFloat4x3[4];
```

```c
struct MTLAccelerationStructureInstanceDescriptor {
    MTLPackedFloat4x3 transformationMatrix;
    MTLAccelerationStructureInstanceOptions options;
    uint mask;
    uint intersectionFunctionTableOffset;
    uint accelerationStructureIndex;
};
```

```c
struct MTLAccelerationStructureUserIDInstanceDescriptor {
    MTLPackedFloat4x3 transformationMatrix;
```
MTLAccelerationStructureInstanceOptions options;
  uint mask;
  uint intersectionFunctionTableOffset;
  uint accelerationStructureIndex;
  uint userID;
};

To facilitate filing out the descriptor, Metal provides an implicit conversion from
acceleration_structure<intersection_tags...> to MTLResourceID.
acceleration_structure<primitive_motion> primitiveAStruct = ...;
MTLResourceID resource_id = primitiveAStruct;

6.18.6 Curve Utility Functions

Starting in Metal 3.1, Metal provides a set of curve utility functions defined in the header
<metal_curves>. The following abbreviations are used:

Ps is float or half.
P is a scalar or a vector of Ps. If Ps is float, P is float4.

The functions return the position or the first or second derivative on a curve given a curve
parameter t, and control points p0, p1, etc. As shown in Table 6.34, the functions support
quadratic Bézier, cubic Bézier, quadratic B-Spline, cubic B-Spline, cubic Hermite, and Catmull-Rom curves.

Table 6.34. Curve Utility Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P bezier( Ps_t, P p0, P p1, P p2)</td>
<td>Returns the position on a quadratic Bézier curve</td>
</tr>
<tr>
<td>P bezier_derivative( Ps_t, P p0, P p1, P p2)</td>
<td>Returns the first derivative on a quadratic Bézier curve</td>
</tr>
<tr>
<td>P bezier_second_derivative( Ps_t, P p0, P p1, P p2)</td>
<td>Returns the second derivative on a quadratic Bézier curve</td>
</tr>
<tr>
<td>P bezier( Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the position on a cubic Bézier curve</td>
</tr>
<tr>
<td>P bezier_derivative( Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the first derivative on a cubic Bézier curve</td>
</tr>
<tr>
<td>P bezier_second_derivative( Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the second derivative on a cubic Bézier curve</td>
</tr>
<tr>
<td>P bspline( Ps_t, P p0, P p1, P p2)</td>
<td>Returns the position on a quadratic B-spline curve</td>
</tr>
<tr>
<td>Function</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>P bspline_derivative(Ps_t, P p0, P p1, P p2)</td>
<td>Returns the first derivative on a quadratic B-spline curve</td>
</tr>
<tr>
<td>P bspline_second_derivative(Ps_t, P p0, P p1, P p2)</td>
<td>Returns the second derivative on a quadratic B-spline curve</td>
</tr>
<tr>
<td>P bspline(Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the position on a cubic B-spline curve</td>
</tr>
<tr>
<td>P bspline_derivative(Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the first derivative on a cubic B-spline curve</td>
</tr>
<tr>
<td>P bspline_second_derivative(Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the second derivative on a cubic B-spline curve</td>
</tr>
<tr>
<td>P hermite(Ps_t, P p0, P p1, P p2, P p3, P m0, P m1)</td>
<td>Returns the position on a cubic Hermite curve</td>
</tr>
<tr>
<td>P hermite_derivative(Ps_t, P p0, P p1, P m0, P m1)</td>
<td>Returns the first derivative on a cubic Hermite curve</td>
</tr>
<tr>
<td>P hermite_second_derivative(Ps_t, P p0, P p1, P m0, P m1)</td>
<td>Returns the second derivative on a cubic Hermite curve</td>
</tr>
<tr>
<td>P catmull_rom(Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the position on a Catmull-Rom curve</td>
</tr>
<tr>
<td>P catmull_rom_derivative(Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the first derivative on a Catmull-Rom curve</td>
</tr>
<tr>
<td>P catmull_rom_second_derivative(Ps_t, P p0, P p1, P p2, P p3)</td>
<td>Returns the second derivative on a Catmull-Rom curve</td>
</tr>
</tbody>
</table>
7 Numerical Compliance

This chapter covers how Metal represents floating-point numbers with regard to accuracy in mathematical operations. Metal is compliant to a subset of the IEEE 754 standard.

7.1 INF, NaN, and Denormalized Numbers

INF must be supported for single-precision, half-precision, and brain floating-point numbers.

NaNs must be supported for single-precision, half-precision, and brain floating-point numbers (with fast math disabled). If fast math is enabled the behavior of handling NaN or INF (as inputs or outputs) is undefined. Signaling NaNs are not supported.

Denormalized single-precision, half-precision, or brain floating-point numbers passed as input to or produced as the output of single-precision, half-precision, or brain floating-point arithmetic operations may be flushed to zero.

7.2 Rounding Mode

Either round ties to even or round toward zero rounding mode may be supported for single-precision, half-precision, and brain floating-point operations.

7.3 Floating-Point Exceptions

Floating-point exceptions are disabled in Metal.

7.4 Relative Error as ULPs

Table 7.1 describes the minimum accuracy of single-precision floating-point basic arithmetic operations and math functions given as ULP values. The reference value used to compute the ULP value of an arithmetic operation is the infinitely precise result.

<table>
<thead>
<tr>
<th>Math Function</th>
<th>Minimum Accuracy (ULP Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x + y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x - y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x * y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>1.0 / x</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x / y</td>
<td>Correctly rounded</td>
</tr>
</tbody>
</table>

Table 7.1. Accuracy of single-precision floating-point operations and functions
<table>
<thead>
<tr>
<th>Math Function</th>
<th>Minimum Accuracy (ULP Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>acos</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>acosh</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>asin</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>asinh</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>atan</td>
<td>&lt;= 5 ulp</td>
</tr>
<tr>
<td>atan2</td>
<td>&lt;= 6 ulp</td>
</tr>
<tr>
<td>atanh</td>
<td>&lt;= 5 ulp</td>
</tr>
<tr>
<td>ceil</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>copysign</td>
<td>0 ulp</td>
</tr>
<tr>
<td>cos</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>cosh</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>cospi</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>exp</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>exp2</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>exp10</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>fabs</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fdim</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>floor</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>fma</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>fmax</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fmin</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fmod</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fract</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>frexp</td>
<td>0 ulp</td>
</tr>
<tr>
<td>ilogb</td>
<td>0 ulp</td>
</tr>
<tr>
<td>ldexp</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>log</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>log2</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>Math Function</td>
<td>Minimum Accuracy (ULP Values)</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>log10</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>modf</td>
<td>0 ulp</td>
</tr>
<tr>
<td>nextafter</td>
<td>0 ulp</td>
</tr>
<tr>
<td>pow</td>
<td>&lt;= 16 ulp</td>
</tr>
<tr>
<td>powr</td>
<td>&lt;= 16 ulp</td>
</tr>
<tr>
<td>rint</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>round</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>rsqrt</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>sin</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>sincos</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>sinh</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>sinpi</td>
<td>&lt;= 4 ulp</td>
</tr>
<tr>
<td>sqrt</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>tan</td>
<td>&lt;= 6 ulp</td>
</tr>
<tr>
<td>tanpi</td>
<td>&lt;= 6 ulp</td>
</tr>
<tr>
<td>tanh</td>
<td>&lt;= 5 ulp</td>
</tr>
<tr>
<td>trunc</td>
<td>Correctly rounded</td>
</tr>
</tbody>
</table>

Table 7.2 describes the minimum accuracy of single-precision floating-point arithmetic operations given as ULP values with fast math enabled (which is the default unless you specify `-fno-fast-math` as a compiler option).

**Table 7.2. Accuracy of single-precision operations and functions with fast math enabled**

<table>
<thead>
<tr>
<th>Math Function</th>
<th>Minimum Accuracy (ULP Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x + y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x - y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x * y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>1.0 / x</td>
<td>&lt;= 1 ulp for x in the domain of $2^{-126}$ to $2^{126}$</td>
</tr>
<tr>
<td>Math Function</td>
<td>Minimum Accuracy (ULP Values)</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>x / y</td>
<td>$\leq 2.5 \text{ ulp for } y$ in the domain of $2^{-126}$ to $2^{126}$</td>
</tr>
<tr>
<td>acos(x)</td>
<td>$\leq 5 \text{ ulp for } x$ in the domain $[-1, 1]$</td>
</tr>
<tr>
<td>acosh(x)</td>
<td>Implemented as $\log(x + \sqrt{x * x - 1.0})$</td>
</tr>
<tr>
<td>asin(x)</td>
<td>$\leq 5 \text{ ulp for } x$ in the domain $[-1, 1]$ and $</td>
</tr>
<tr>
<td>asinh(x)</td>
<td>Implemented as $\log(x + \sqrt{x * x + 1.0})$</td>
</tr>
<tr>
<td>atan(x)</td>
<td>$\leq 5 \text{ ulp}$</td>
</tr>
<tr>
<td>atanh(x)</td>
<td>Implemented as $0.5 \times \left( \log\left(\frac{1.0 + x}{1.0 - x}\right) \right)$</td>
</tr>
<tr>
<td>atan2(y, x)</td>
<td>Implemented as if $x &gt; 0$, $\tan(\frac{y}{x})$, if $x &lt; 0$ and $y &gt; 0$, $\tan(\frac{y}{x}) + M_{\text{PI}}$, if $x &lt; 0$ and $y &lt; 0$, $\tan(\frac{y}{x}) - M_{\text{PI}}$, and if $x = 0$ and $y = 0$, is undefined.</td>
</tr>
<tr>
<td>cos(x)</td>
<td>For $x$ in the domain $[-\pi, \pi]$, the maximum absolute error is $\leq 2^{-13}$ and larger otherwise.</td>
</tr>
<tr>
<td>cosh(x)</td>
<td>Implemented as $0.5 \times (\exp(x) + \exp(-x))$</td>
</tr>
<tr>
<td>cospi(x)</td>
<td>For $x$ in the domain $[-1, 1]$, the maximum absolute error is $\leq 2^{-13}$ and larger otherwise.</td>
</tr>
<tr>
<td>exp(x)</td>
<td>$\leq 3 + \text{floor}(\text{fabs}(2 * x)) \text{ ulp}$</td>
</tr>
<tr>
<td>exp2(x)</td>
<td>$\leq 3 + \text{floor}(\text{fabs}(2 * x)) \text{ ulp}$</td>
</tr>
<tr>
<td>exp10(x)</td>
<td>Implemented as $\exp(2 \times \log(10))$</td>
</tr>
<tr>
<td>fabs</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fdim</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>floor</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>fma</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>fmax</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fmin</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fmod</td>
<td>Undefined</td>
</tr>
<tr>
<td>fract</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>frexp</td>
<td>0 ulp</td>
</tr>
<tr>
<td>ilogb</td>
<td>0 ulp</td>
</tr>
<tr>
<td>Math Function</td>
<td>Minimum Accuracy (ULP Values)</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>ldexp</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>log(x)</td>
<td>For x in the domain [0.5, 2], the maximum absolute error is &lt;= 2⁻²¹; otherwise if x &gt; 0 the maximum error is &lt;= 3 ulp; otherwise the results are undefined.</td>
</tr>
<tr>
<td>log2(x)</td>
<td>For x in the domain [0.5, 2], the maximum absolute error is &lt;= 2⁻²²; otherwise if x &gt; 0 the maximum error is &lt;= 2 ulp; otherwise the results are undefined.</td>
</tr>
<tr>
<td>log10(x)</td>
<td>Implemented as log2(x) * log10(2)</td>
</tr>
<tr>
<td>modf</td>
<td>0 ulp</td>
</tr>
<tr>
<td>pow(x, y)</td>
<td>Implemented as exp2(y * log2(x)). Undefined for x = 0 and y = 0.</td>
</tr>
<tr>
<td>powr(x, y)</td>
<td>Implemented as exp2(y * log2(x)). Undefined for x = 0 and y = 0.</td>
</tr>
<tr>
<td>rint</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>round(x)</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>rsqrt</td>
<td>&lt;= 2 ulp</td>
</tr>
<tr>
<td>sin(x)</td>
<td>For x in the domain [-pi, pi], the maximum absolute error is &lt;= 2⁻¹³ and larger otherwise.</td>
</tr>
<tr>
<td>sinh(x)</td>
<td>Implemented as 0.5 * (exp(x) - exp(-x))</td>
</tr>
<tr>
<td>sincos(x)</td>
<td>ULP values as defined for sin(x) and cos(x)</td>
</tr>
<tr>
<td>sinpi(x)</td>
<td>For x in the domain [-1, 1], the maximum absolute error is &lt;= 2⁻¹³ and larger otherwise.</td>
</tr>
<tr>
<td>sqrt(x)</td>
<td>Implemented as x * rsqrt(x) with special cases handled correctly.</td>
</tr>
<tr>
<td>tan(x)</td>
<td>Implemented as sin(x) * (1.0 / cos(x))</td>
</tr>
<tr>
<td>tanh(x)</td>
<td>Implemented as (t - 1.0)/(t + 1.0), where t = exp(2.0 * x)</td>
</tr>
<tr>
<td>tanpi(x)</td>
<td>Implemented as tan(x * pi)</td>
</tr>
<tr>
<td>trunc</td>
<td>Correctly rounded</td>
</tr>
</tbody>
</table>

Table 7.3 describes the minimum accuracy of half-precision floating-point basic arithmetic operations and math functions given as ULP values. Table 7.3 applies to iOS and macOS, starting with MTLGPUFamilyApple4 hardware.
### Table 7.3. Accuracy of half-precision floating-point operations and functions

<table>
<thead>
<tr>
<th>Math Function</th>
<th>Minimum Accuracy (ULP Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x + y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x - y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x * y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>1.0 / x</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x / y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>acos(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>acosh(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>asin(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>asinh(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>atan(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>atanh(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>atan2(y, x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>cos(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>cosh(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>cospi(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>exp(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>exp2(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>exp10(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>fabs</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fdim</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>floor</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>fma</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>fmax</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fmin</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fmod</td>
<td>0 ulp</td>
</tr>
<tr>
<td>fract</td>
<td>Correctly rounded</td>
</tr>
</tbody>
</table>
### Math Function

<table>
<thead>
<tr>
<th>Math Function</th>
<th>Minimum Accuracy (ULP Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>frexp</td>
<td>0 ulp</td>
</tr>
<tr>
<td>ilogb</td>
<td>0 ulp</td>
</tr>
<tr>
<td>ldexp</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>log(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>log2(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>log10(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>modf</td>
<td>0 ulp</td>
</tr>
<tr>
<td>nextafter</td>
<td>0 ulp</td>
</tr>
<tr>
<td>rint</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>round(x)</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>rsqrt</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>sin(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>sinh(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>sincos(x)</td>
<td>ULP values as defined for sin(x) and cos(x)</td>
</tr>
<tr>
<td>sinpi(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>sqrt(x)</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>tan(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>tanh(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>tanpi(x)</td>
<td>&lt;= 1 ulp</td>
</tr>
<tr>
<td>trunc</td>
<td>Correctly rounded</td>
</tr>
</tbody>
</table>

Table 7.4 describes the minimum accuracy of brain floating-point basic arithmetic operations and math functions given as ULP values. Table 7.4 applies to all OS, starting with MTLGPUFamilyApple6 or MTLGPUFamilyMac2 hardware.

**Table 7.4. Accuracy of brain floating-point operations and functions**

<table>
<thead>
<tr>
<th>Math Function</th>
<th>Minimum Accuracy (ULP Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x + y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x - y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>Math Function</td>
<td>Minimum Accuracy (ULP Values)</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>x * y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>1.0 / x</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x / y</td>
<td>Correctly rounded</td>
</tr>
</tbody>
</table>

Table 7.5. Accuracy of brain floating-point operations and functions with fast math enabled

<table>
<thead>
<tr>
<th>Math Function</th>
<th>Minimum Accuracy (ULP Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x + y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x - y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>x * y</td>
<td>Correctly rounded</td>
</tr>
<tr>
<td>1.0 / x</td>
<td>&lt;= 0.6 ulp for x in the domain of $2^{-126}$ to $2^{126}$</td>
</tr>
<tr>
<td>x / y</td>
<td>&lt;= 0.6 ulp for y in the domain of $2^{-126}$ to $2^{126}$</td>
</tr>
</tbody>
</table>

Even though the precision of individual math operations and functions are specified in Table 7.1, Table 7.2, Table 7.3, Table 7.4, and Table 7.5, the Metal compiler, in fast math mode (see section 1.5.5), may do various optimization like reassociate floating-point operations that may dramatically change results in floating-point. Reassociation may change or ignore the sign of zero, allow optimizations to assume the arguments and result are not NaN or +/-INF, inhibit or create underflow or overflow and thus cannot be in code that relies on rounding behavior such as $(x + 2^{52}) - 2^{52}$, or ordered floating-point comparisons.

The ULP is defined as follows:

If $x$ is a real number that lies between two finite consecutive floating-point numbers $a$ and $b$, without being equal to one of them, then $\text{ulp}(x) = |b - a|$, otherwise $\text{ulp}(x)$ is the distance between the two non-equal finite floating-point numbers nearest $x$. Moreover, $\text{ulp}(\text{NaN})$ is NaN.

7.5 Edge Case Behavior in Flush to Zero Mode

If denormalized values are flushed to zero, then a function may return one of four results:

1. Any conforming result for non-flush-to-zero mode.
2. If the result given by step 1 is a subnormal before rounding, it may be flushed to zero.
3. Any non-flushed conforming result for the function if one or more of its subnormal operands are flushed to zero.
4. If the result of step 3 is a subnormal before rounding, the result may be flushed to zero.
In each of the above cases, if an operand or result is flushed to zero, the sign of the zero is undefined.

7.6 Conversion Rules for Floating-Point and Integer Types

When converting from a floating point type to an integer, the conversion uses round toward zero rounding mode. Use the "round ties to even" or "round toward zero" rounding mode for conversions from a floating-point or integer type to a floating-point type.

The conversions from half and bfloat to float are lossless. Conversions from float to half or to bfloat round the mantissa using the round ties to even rounding mode. When converting a float to a half, denormalized numbers generated for the half data type may not be flushed to zero.

When converting a floating-point type to an integer type, if the floating-point value is NaN, the resulting integer is 0.

Note that fast math does not change the accuracy of conversion operations.

7.7 Texture Addressing and Conversion Rules

The texture coordinates specified to the sample, sample_compare, gather, gather_compare, read, and write functions cannot be INF or NaN. In addition, for the texture read and write functions, the texture coordinates must refer to a region inside the texture. If the texture coordinates are outside the bounds of the texture, the read and write function behavior is undefined.

The following sections discuss the application of conversion rules when reading and writing textures in a graphics or kernel function. When performing a multisample resolve operation, these conversion rules do not apply.

7.7.1 Conversion Rules for Normalized Integer Pixel Data Types

This section discusses converting normalized integer pixel data types to floating-point values and vice-versa.

7.7.1.1 Converting Normalized Integer Pixel Data Types to Floating-Point Values

For textures that have 8-, 10-, or 16-bit normalized unsigned integer pixel values, the texture sample and read functions convert the pixel values from an 8- or 16-bit unsigned integer to a normalized single- or half-precision floating-point value in the range \([0.0 \ ... \ 1.0]\).

For textures that have 8- or 16-bit normalized signed integer pixel values, the texture sample and read functions convert the pixel values from an 8- or 16-bit signed integer to a normalized single- or half-precision floating-point value in the range \([-1.0 \ ... \ 1.0]\).
These conversions are performed as listed in the second column of Table 7.6. The precision of the conversion rules is guaranteed to be <= 1.5 ulp, except for the cases described in the “Corner Cases” column.

**Table 7.6. Conversion to a normalized float value**

<table>
<thead>
<tr>
<th>Convert from</th>
<th>Conversion Rule to Normalized Float</th>
<th>Corner Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-bit normalized unsigned integer</td>
<td>float(c)</td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 must convert to 1.0</td>
</tr>
<tr>
<td>2-bit normalized unsigned integer</td>
<td>float(c) / 3.0</td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 must convert to 1.0</td>
</tr>
<tr>
<td>4-bit normalized unsigned integer</td>
<td>float(c) / 15.0</td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15 must convert to 1.0</td>
</tr>
<tr>
<td>5-bit normalized unsigned integer</td>
<td>float(c) / 31.0</td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31 must convert to 1.0</td>
</tr>
<tr>
<td>6-bit normalized unsigned integer</td>
<td>float(c) / 63.0</td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>63 must convert to 1.0</td>
</tr>
<tr>
<td>8-bit normalized unsigned integer</td>
<td>float(c) / 255.0</td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>255 must convert to 1.0</td>
</tr>
<tr>
<td>10-bit normalized unsigned integer</td>
<td>float(c) / 1023.0</td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1023 must convert to 1.0</td>
</tr>
<tr>
<td>16-bit normalized unsigned integer</td>
<td>float(c) / 65535.0</td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>65535 must convert to 1.0</td>
</tr>
<tr>
<td>8-bit normalized signed integer</td>
<td>max(-1.0, float(c)/127.0)</td>
<td>-128 and -127 must convert to -1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>127 must convert to 1.0</td>
</tr>
<tr>
<td>16-bit normalized signed integer</td>
<td>max(-1.0, float(c)/32767.0)</td>
<td>-32768 and -32767 must convert to -1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 must convert to 0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32767 must convert to 1.0</td>
</tr>
</tbody>
</table>

**7.7.1.2 Converting Floating-Point Values to Normalized Integer Pixel Data Types**

For textures that have 8-, 10-, or 16-bit normalized unsigned integer pixel values, the texture write functions convert the single- or half-precision floating-point pixel value to an 8- or 16-bit unsigned integer.

For textures that have 8- or 16-bit normalized signed integer pixel values, the texture write functions convert the single- or half-precision floating-point pixel value to an 8- or 16-bit signed integer.
NaN values are converted to zero.

Conversions from floating-point values to normalized integer values are performed as listed in Table 7.7.

**Table 7.7. Conversion from floating-point to a normalized integer value**

<table>
<thead>
<tr>
<th>Convert to</th>
<th>Conversion Rule to Normalized Integer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-bit normalized unsigned integer</td>
<td>( x = \min(\max(f, 0.0), 1.0) ) ( i0:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>2-bit normalized unsigned integer</td>
<td>( x = \min(\max(f \times 3.0, 0.0), 3.0) ) ( i1:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>4-bit normalized unsigned integer</td>
<td>( x = \min(\max(f \times 15.0, 0.0), 15.0) ) ( i3:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>5-bit normalized unsigned integer</td>
<td>( x = \min(\max(f \times 31.0, 0.0), 31.0) ) ( i4:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>6-bit normalized unsigned integer</td>
<td>( x = \min(\max(f \times 63.0, 0.0), 63.0) ) ( i5:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>8-bit normalized unsigned integer</td>
<td>( x = \min(\max(f \times 255.0, 0.0), 255.0) ) ( i7:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>10-bit normalized unsigned integer</td>
<td>( x = \min(\max(f \times 1023.0, 0.0), 1023.0) ) ( i9:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>16-bit normalized unsigned integer</td>
<td>( \text{result} = \min(\max(f \times 65535.0, 0.0), 65535.0) ) ( i15:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>8-bit normalized signed integer</td>
<td>( \text{result} = \min(\max(f \times 127.0, -127.0), 127.0) ) ( i7:0 = \text{intRTNE}(x) )</td>
</tr>
<tr>
<td>16-bit normalized signed integer</td>
<td>( \text{result} = \min(\max(f \times 32767.0, -32767.0), 32767.0) ) ( i15:0 = \text{intRTNE}(x) )</td>
</tr>
</tbody>
</table>

In Metal 2, all conversions to and from unorm data types shall be correctly rounded.

### 7.7.2 Conversion Rules for Half-Precision Floating-Point Pixel Data Type

For textures that have half-precision floating-point pixel color values, the conversions from `half` to `float` are lossless. Conversions from `float` to `half` round the mantissa using the round ties to even rounding mode. Denormalized numbers for the `half` data type which may be generated when converting a `float` to a `half` may not be flushed to zero. A `float` NaN may be converted to an appropriate NaN or be flushed to zero in the `half` type. A `float` INF must be converted to an appropriate INF in the `half` type.
7.7.3 Conversion Rules for Single-Precision Floating-Point Pixel Data Type

The following rules apply for reading and writing textures that have single-precision floating-point pixel color values:

- NaNs may be converted to a NaN value(s) or be flushed to zero.
- INFs must be preserved.
- Denormalized numbers may be flushed to zero.
- All other values must be preserved.

7.7.4 Conversion Rules for 10- and 11-bit Floating-Point Pixel Data Type

The floating-point formats use 5 bits for the exponent, with 5 bits of mantissa for 10-bit floating-point types, or 6 bits of mantissa for 11-bit floating-point types with an additional hidden bit for both types. There is no sign bit. The 10- and 11-bit floating-point types preserve denormals.

These floating-point formats use the following rules:

- If the exponent and mantissa are 0, the floating-point value is 0.0.
- If the exponent is 31 and the mantissa is != 0, the resulting floating-point value is a NaN.
- If the exponent is 31 and the mantissa is 0, the resulting floating-point value is positive infinity.
- If 0 <= exponent <= 31, the floating-point value is $2^{\text{exponent} - 15} \times (1 + \text{mantissa}/N)$.
- If the exponent is 0 and the mantissa is != 0, the floating-point value is a denormal number given as $2^{\text{exponent} - 14} \times \text{mantissa} / N$. If mantissa is 5 bits, N is 32; if mantissa is 6 bits, N is 64.

Conversion of a 10- or 11-bit floating-point pixel data type to a half- or single-precision floating-point value is lossless. Conversion of a half or single precision floating-point value to a 10- or 11-bit floating-point value must be <= 0.5 ULP. Any operation that results in a value less than zero for these floating-point types is clamped to zero.

7.7.5 Conversion Rules for 9-bit Floating-Point Pixel Data Type with a 5-bit Exponent

The RGB9E5_SharedExponent shared exponent floating-point format uses 5 bits for the exponent and 9 bits for the mantissa. There is no sign bit.

Conversion from this format to a half- or single-precision floating-point value is lossless and computed as $2^{\text{shared exponent} - 15} \times (\text{mantissa}/512)$ for each color channel.

Conversion from a half or single precision floating-point RGB color value to this format is performed as follows, where N is the number of mantissa bits per component (9), B is the exponent bias (15) and E_max is the maximum allowed biased exponent value (31).

- Clamp the r, g, and b components (in the process, mapping NaN to zero) as follows:

\[
\begin{align*}
rc &= \max(0, \min(\text{sharedexpmax}, r)) \\
gc &= \max(0, \min(\text{sharedexpmax}, g)) \\
bc &= \max(0, \min(\text{sharedexpmax}, b))
\end{align*}
\]

Where \(\text{sharedexpmax} = ((2^N - 1)/2^N) \times 2(E_{\text{max}} - B)\).

- Determine the largest clamped component maxc:

\[
\text{maxc} = \max(rc, gc, bc)
\]
• Compute a preliminary shared exponent \( expp \)
  \[
  expp = \max(-B - 1, \ \ \ \ \ \ \ \ \ \ \ \ \floor{\log2(\maxc)} + 1 + B
  \]

• Compute a refined shared exponent \( exps \)
  \[
  \maxs = \floor{\frac{\maxc}{2^{expp - B - N}} + 0.5f} \\
  exps = expp, \ \text{if} \ \ 0 \leq \maxs < 2N, \ \text{and} \ \ exps = expp + 1, \ \text{if} \ \ maxs = 2N.
  \]

• Finally, compute three integer values in the range 0 to \( 2N - 1 \):
  \[
  rs = \floor{\frac{rc}{2^{expp - B - N}} + 0.5f} \\
  gs = \floor{\frac{gc}{2^{expp - B - N}} + 0.5f} \\
  bs = \floor{\frac{bc}{2^{expp - B - N}} + 0.5f}
  \]

Conversion of a half- or single-precision floating-point color values to the MTLPixelFormatRGB9E5Float shared exponent floating-point value is \( \leq 0.5 \) ULP.

### 7.7.6 Conversion Rules for Signed and Unsigned Integer Pixel Data Types

For textures that have an 8- or 16-bit signed or unsigned integer pixel values, the texture sample and read functions return a signed or unsigned 32-bit integer pixel value. The conversions described in this section must be correctly saturated.

Writes to these integer textures perform one of the conversions listed in Table 7.8.

<table>
<thead>
<tr>
<th>Convert From</th>
<th>To</th>
<th>Conversion Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>32-bit signed integer</td>
<td>8-bit signed integer</td>
<td>( \text{result} = \text{convert} _\text{char} _\text{saturate}(\text{val}) )</td>
</tr>
<tr>
<td>32-bit signed integer</td>
<td>16-bit signed integer</td>
<td>( \text{result} = \text{convert} _\text{short} _\text{saturate}(\text{val}) )</td>
</tr>
<tr>
<td>32-bit unsigned integer</td>
<td>8-bit unsigned integer</td>
<td>( \text{result} = \text{convert} _\text{uchar} _\text{saturate}(\text{val}) )</td>
</tr>
<tr>
<td>32-bit unsigned integer</td>
<td>16-bit unsigned integer</td>
<td>( \text{result} = \text{convert} _\text{ushort} _\text{saturate}(\text{val}) )</td>
</tr>
</tbody>
</table>

### 7.7.7 Conversion Rules for sRGBA and sBGRA Textures

Conversion from sRGB space to linear space is automatically done when sampling from an sRGB texture. The conversion from sRGB to linear RGB is performed before the filter specified in the sampler specified when sampling the texture is applied. If the texture has an alpha channel, the alpha data is stored in linear color space.

Conversion from linear to sRGB space is automatically done when writing to an sRGB texture. If the texture has an alpha channel, the alpha data is stored in linear color space.

2023-06-02 | Copyright © 2023 Apple Inc. | All Rights Reserved.

Page 285 of 287
The following is the conversion rule for converting a normalized 8-bit unsigned integer from an sRGB color value to a floating-point linear RGB color value (call it \( c \)):

\[
\text{if } (c \leq 0.04045) \\
\quad \text{result} = c / 12.92; \\
\text{else} \\
\quad \text{result} = \text{powr}((c + 0.055) / 1.055, 2.4); 
\]

The precision of the above conversion must ensure that the delta between the resulting infinitely precise floating point value when converting result back to an unnormalized sRGB value but without rounding to an 8-bit unsigned integer value (call it \( r \)) and the original sRGB 8-bit unsigned integer color value (call it \( r_{\text{orig}} \)) is \( \leq 0.5 \); for example:

\[
\text{fabs}(r - r_{\text{orig}}) \leq 0.5 
\]

Use the following rules for converting a linear RGB floating-point color value (call it \( c \)) to a normalized 8-bit unsigned integer sRGB value:

\[
\text{if } (\text{isnan}(c)) \quad c = 0.0; \\
\text{if } (c > 1.0) \quad c = 1.0; \\
\text{else if } (c < 0.0) \quad c = 0.0; \\
\text{else if } (c < 0.0031308) \quad c = 12.92 \times c; \\
\text{else} \quad c = 1.055 \times \text{powr}(c, 1.0/2.4) - 0.055; 
\]

// Convert to integer scale: \( c = c \times 255.0 \) \\
// Convert to integer: \( c = c + 0.5 \) \\
// Drop the decimal fraction. The remaining floating-point(integral) value \\
// is converted directly to an integer. 

The precision of the above conversion shall be:

\[
\text{fabs(reference result - integer result)} < 1.0. 
\]