#### Neural Networks and Accelerate

Session 715

Eric Bainville Core OS, Vector and Numerics Group Steve Canon Core OS, Vector and Numerics Group

### Performance Libraries for CPU

That thing we do in the Vector and Numerics group

Accelerate - Image processing: vlmage

Accelerate - Image processing: vlmage

Accelerate - Signal processing: vDSP

Accelerate - Image processing: vlmage

Accelerate - Signal processing: vDSP

Accelerate - Linear algebra: BLAS, SparseBLAS, LAPACK, Linear Algebra

Accelerate - Image processing: vlmage

Accelerate - Signal processing: vDSP

Accelerate - Linear algebra: BLAS, SparseBLAS, LAPACK, Linear Algebra

Vector extensions: simd

Accelerate - Image processing: vlmage

Accelerate - Signal processing: vDSP

Accelerate - Linear algebra: BLAS, SparseBLAS, LAPACK, Linear Algebra

Vector extensions: simd

Lossless compression: Compression

Accelerate - Image processing: vlmage

Accelerate - Signal processing: vDSP

Accelerate - Linear algebra: BLAS, SparseBLAS, LAPACK, Linear Algebra

Vector extensions: simd

Lossless compression: Compression

Optimized for all supported CPUs

Lossless compression: Compression

Lossless compression: Compression

Accelerate - Machine learning: BNNS

Lossless compression: Compression

Accelerate - Machine learning: BNNS

Accelerate - Numerical integration: Quadrature

Lossless compression: Compression

Accelerate - Machine learning: BNNS

Accelerate - Numerical integration: Quadrature

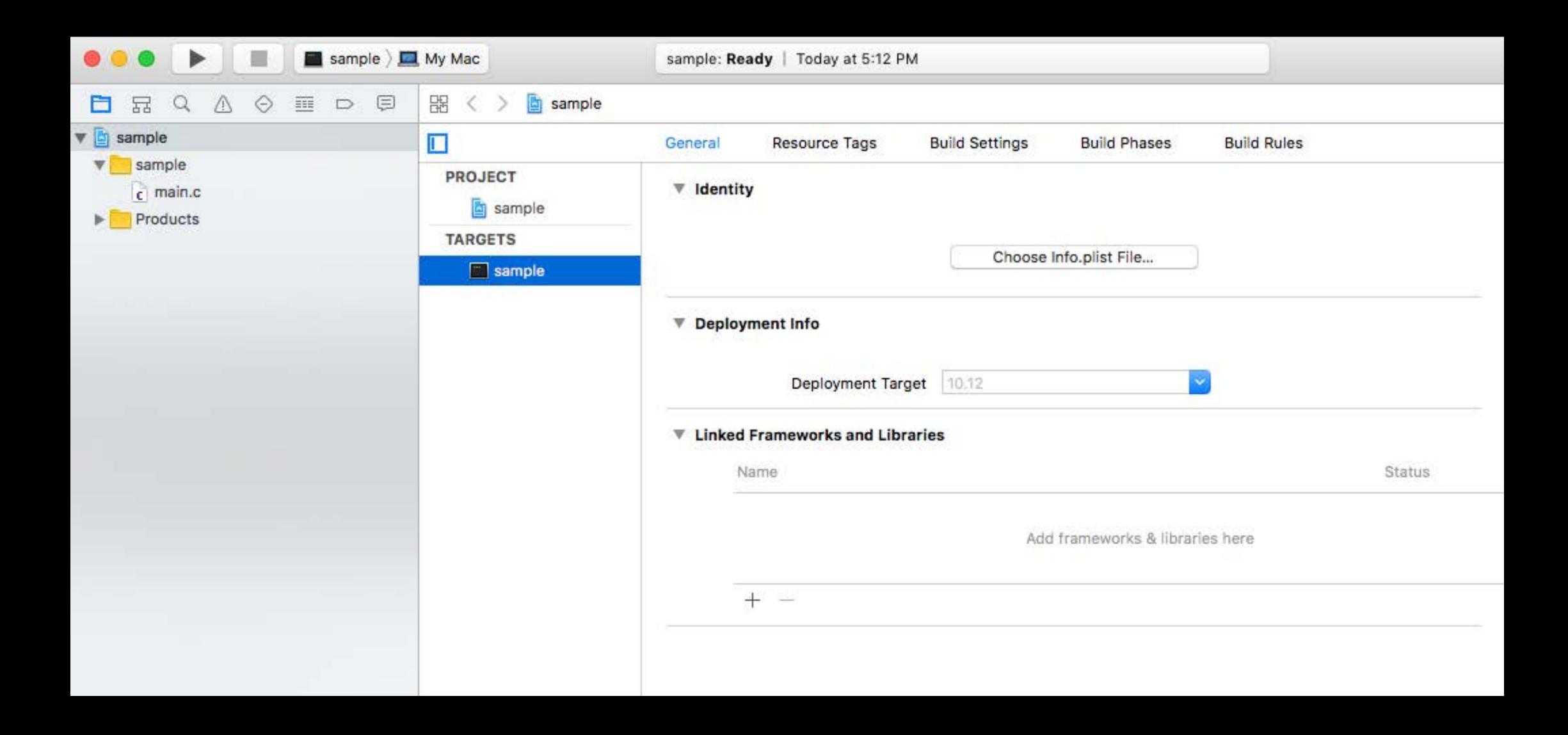
Vector extensions: simd

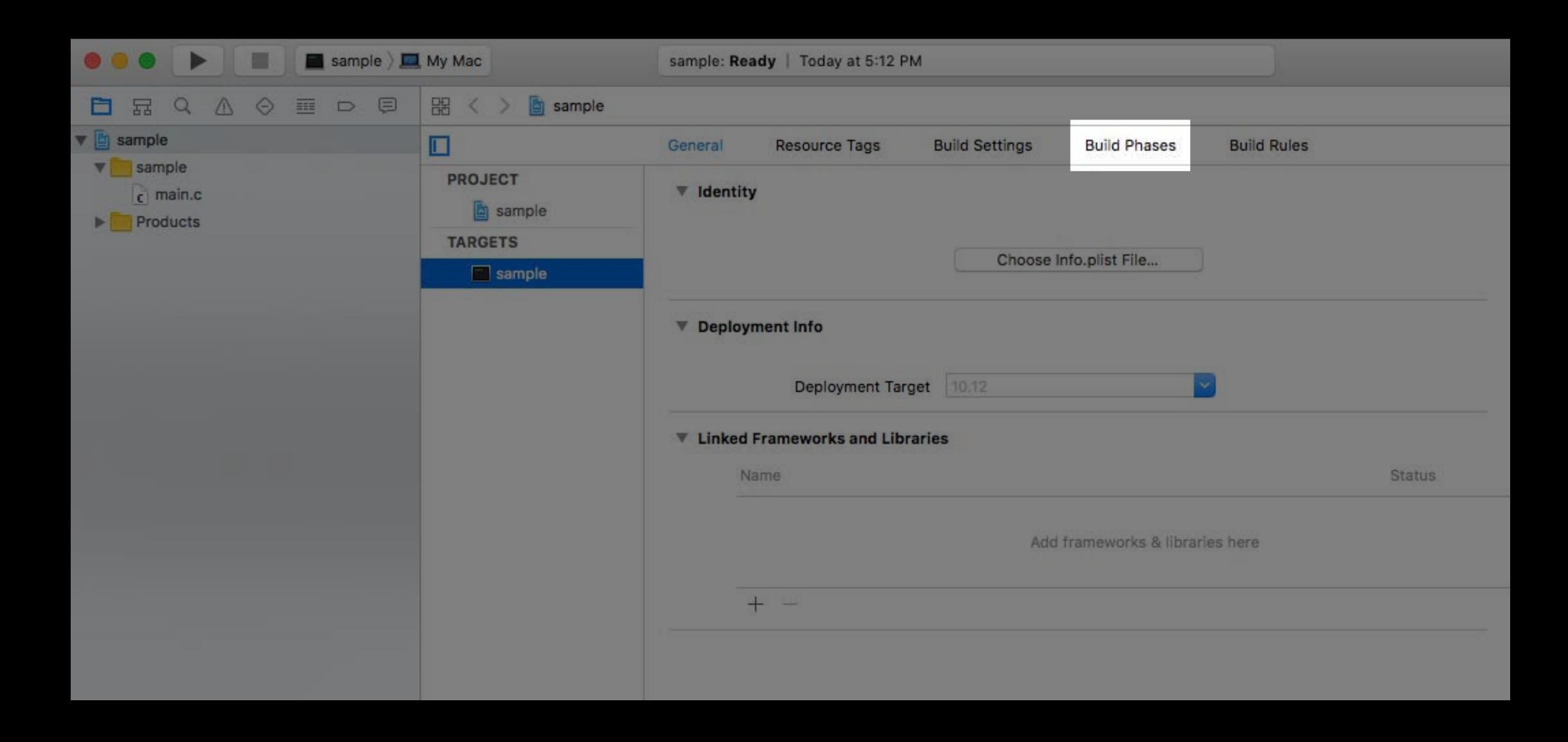
## Using Accelerate

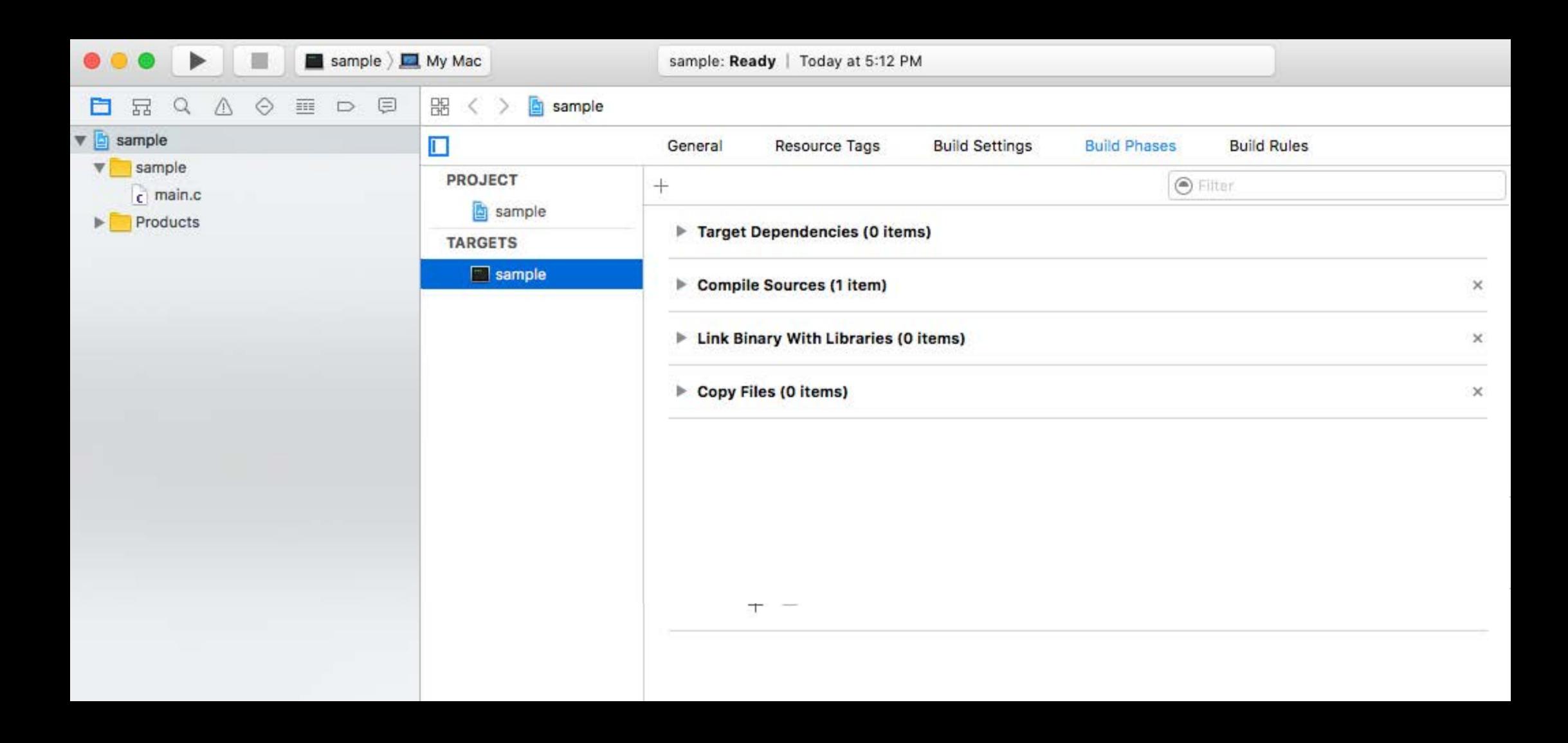
```
// Swift
import Accelerate

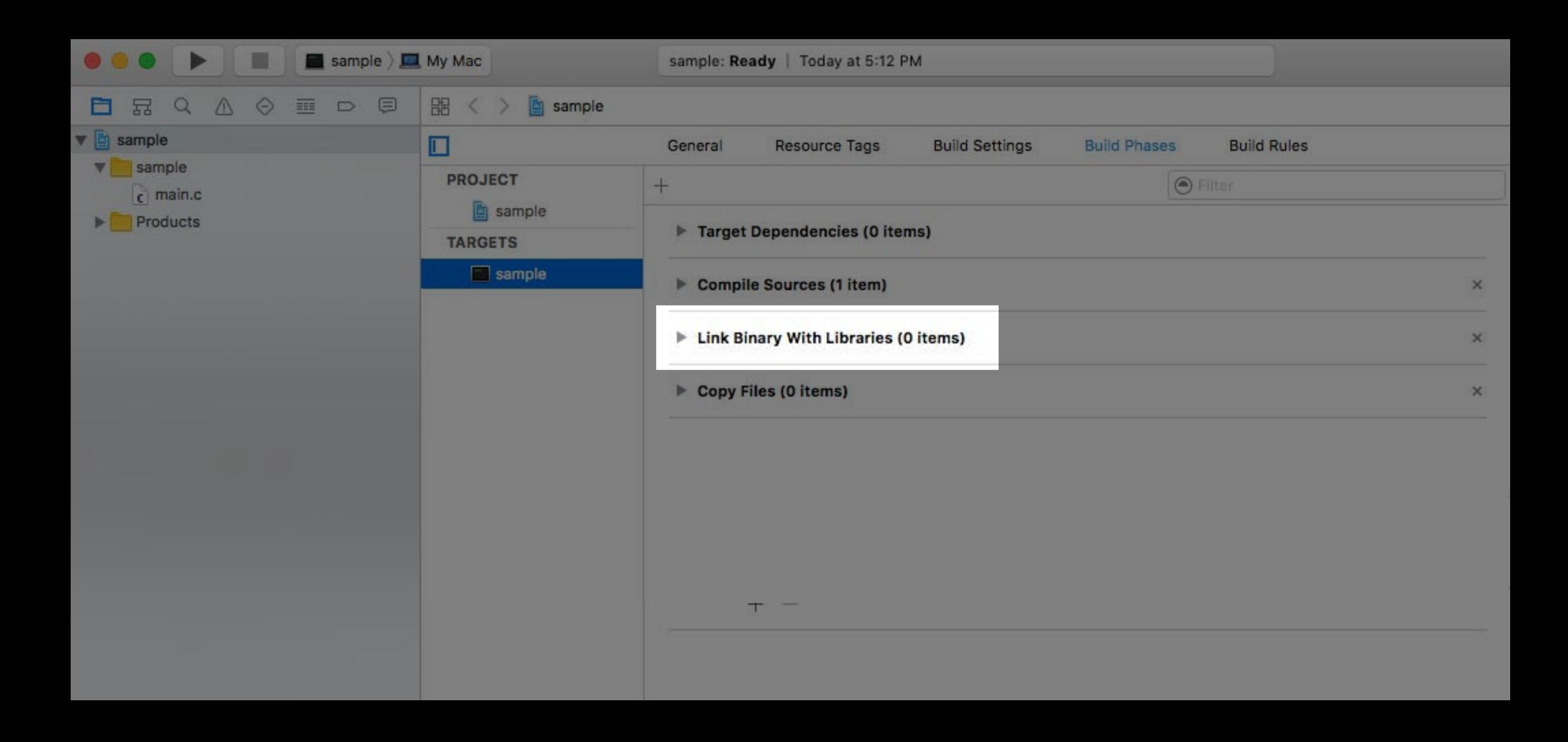
// C / C++
#include <Accelerate/Accelerate.h>

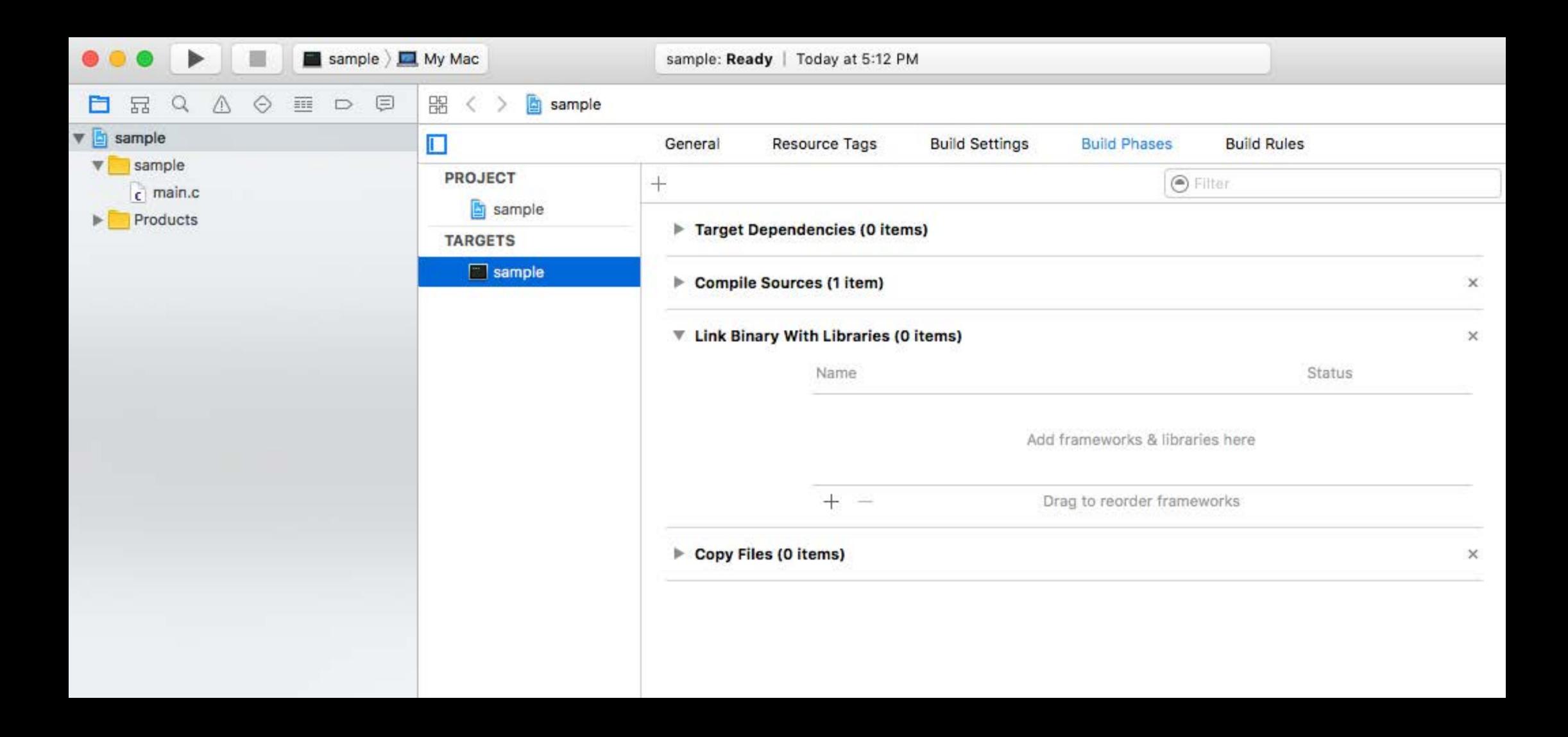
// Objective-C
@import Accelerate
```

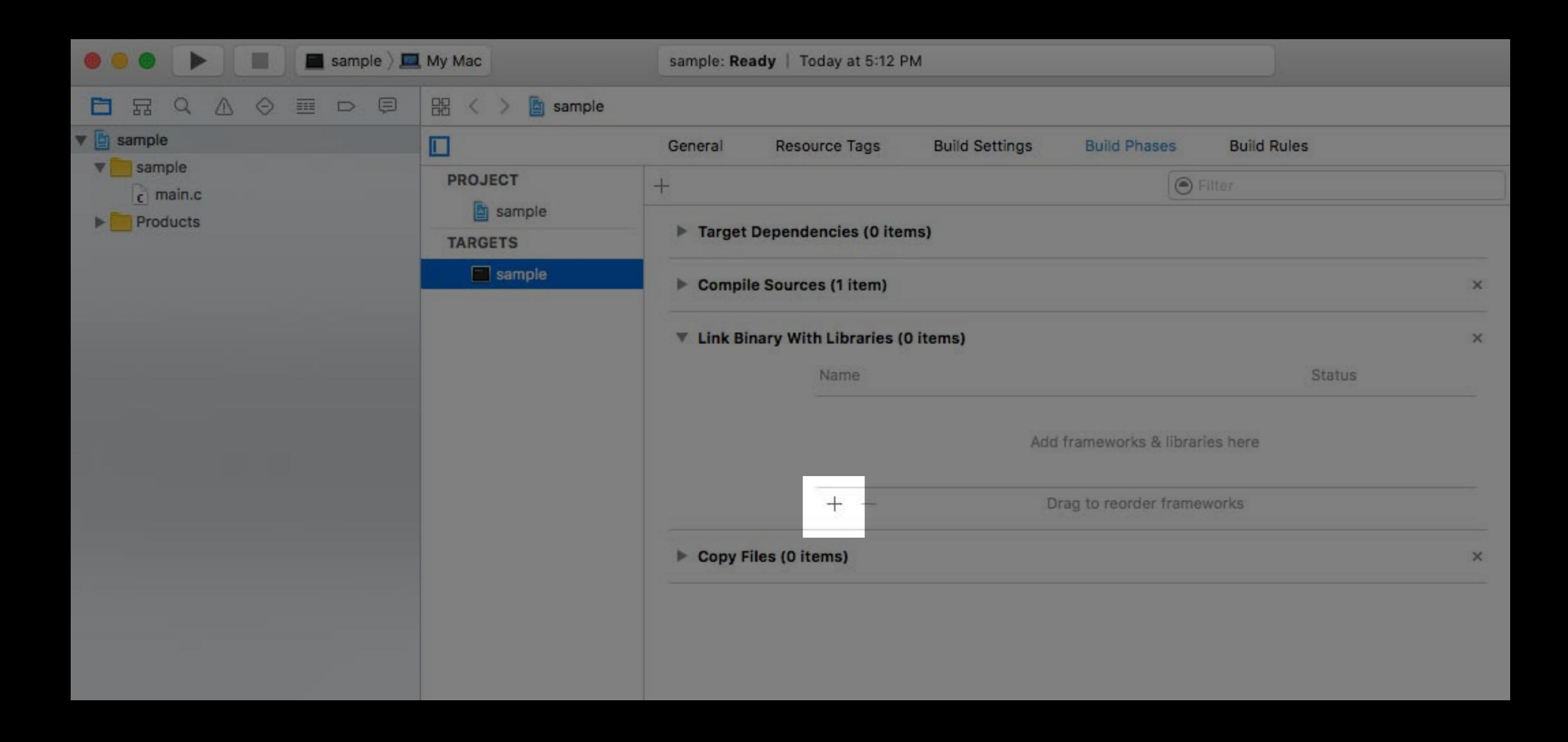


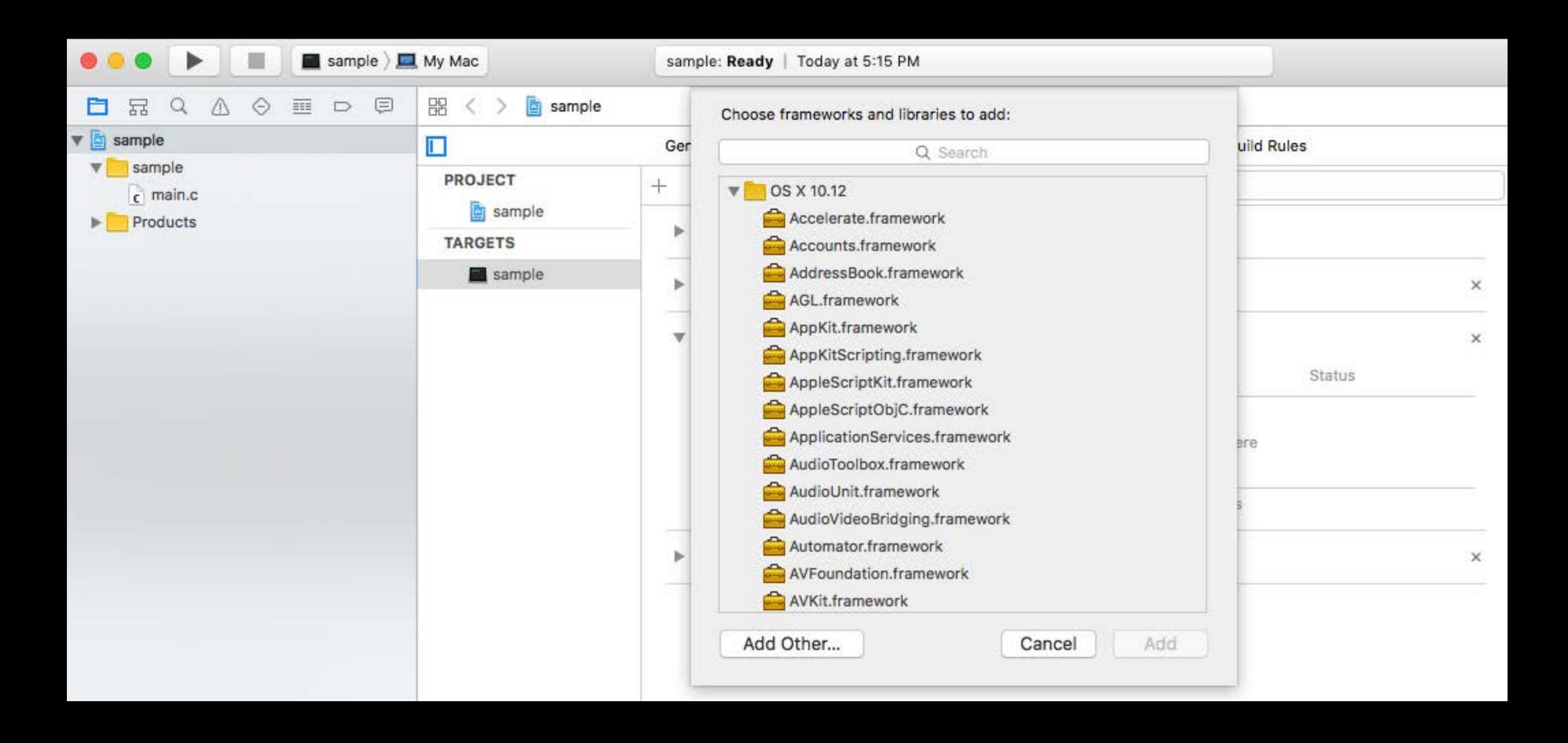


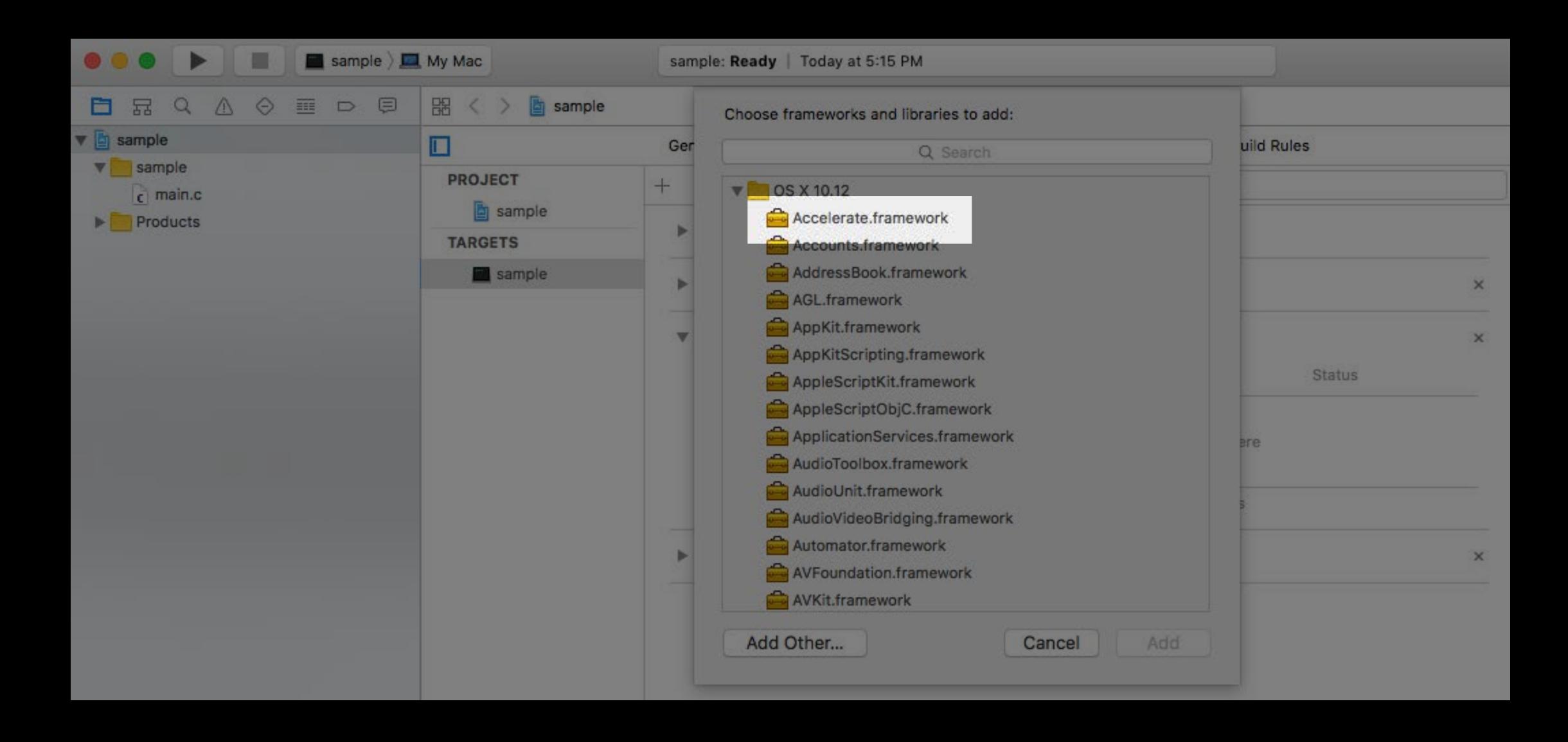












Remember last year?

#### LEMPEL ZIV FINITE STATE ENTROPY



# Compression LZFSE

LZFSE is now Open Source

LZFSE is now Open Source

Hosted on github.com/lzfse

LZFSE is now Open Source

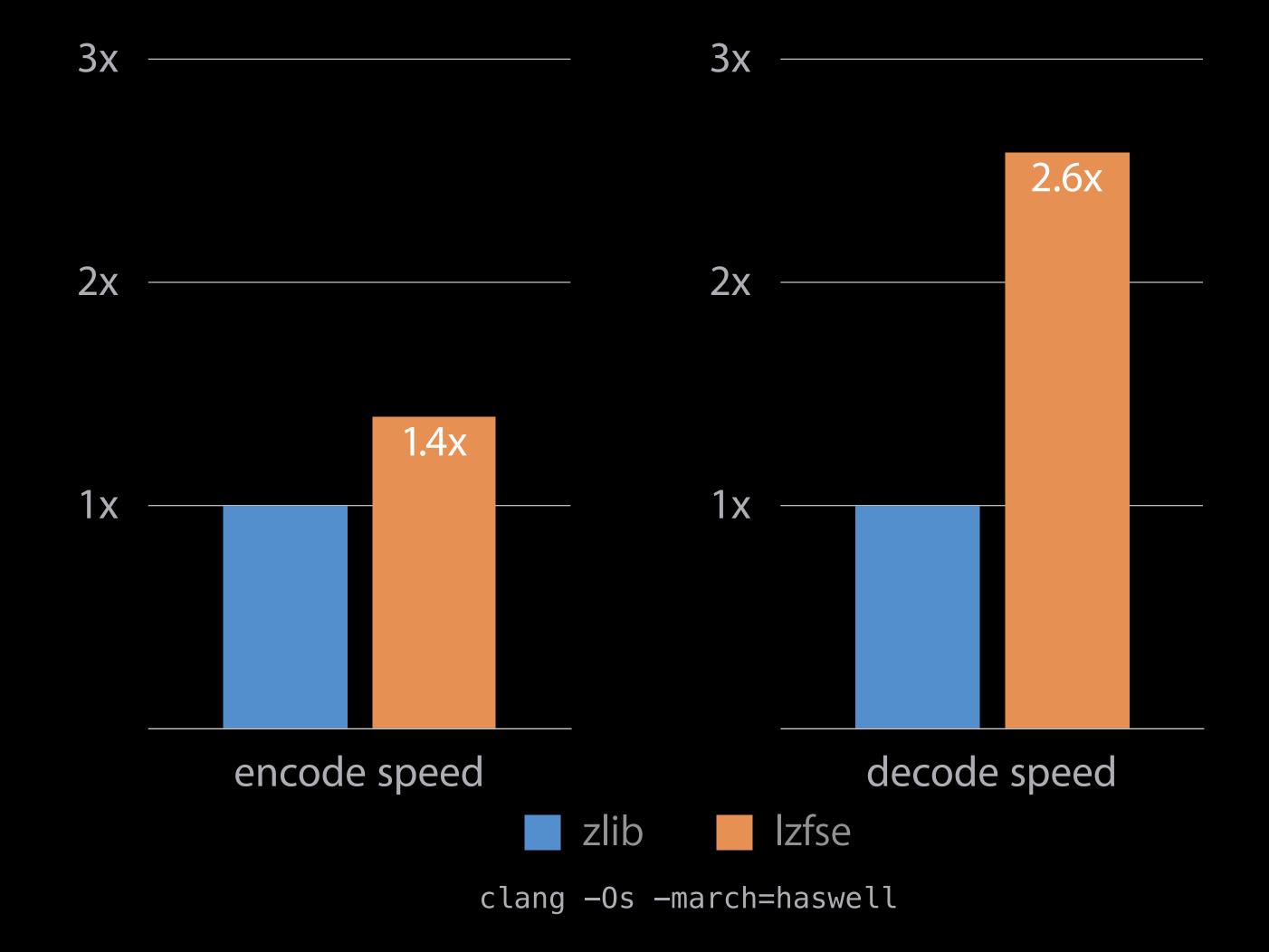
Hosted on github.com/lzfse

BSD license

LZFSE is now Open Source

Hosted on github.com/lzfse

BSD license



### BNNS

Basic Neural Network Subroutines

#### BNNS

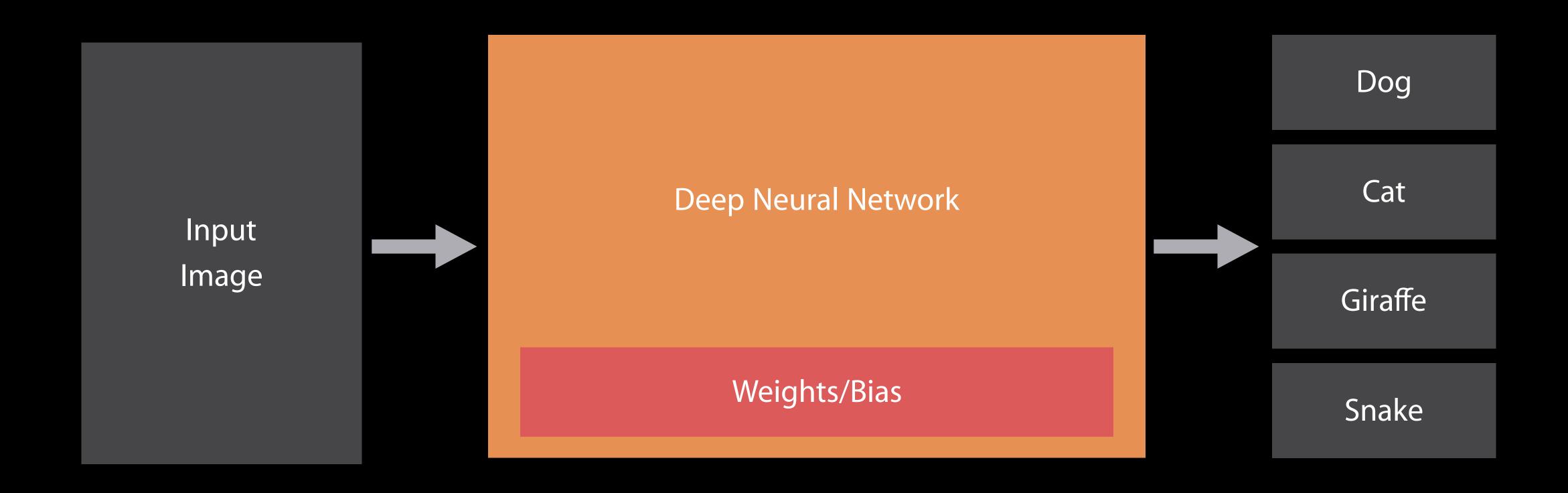


New in Accelerate

BNNS = Basic Neural Network Subroutines BLAS = Basic Linear Algebra Subroutines

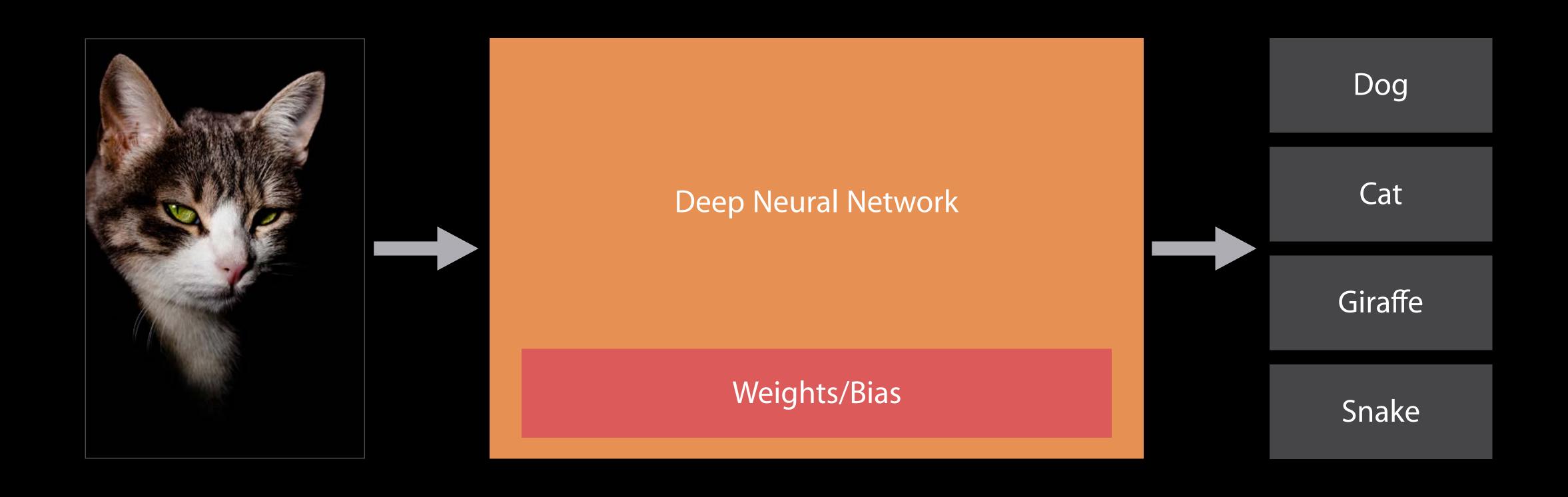
### Deep Neural Network

#### Training



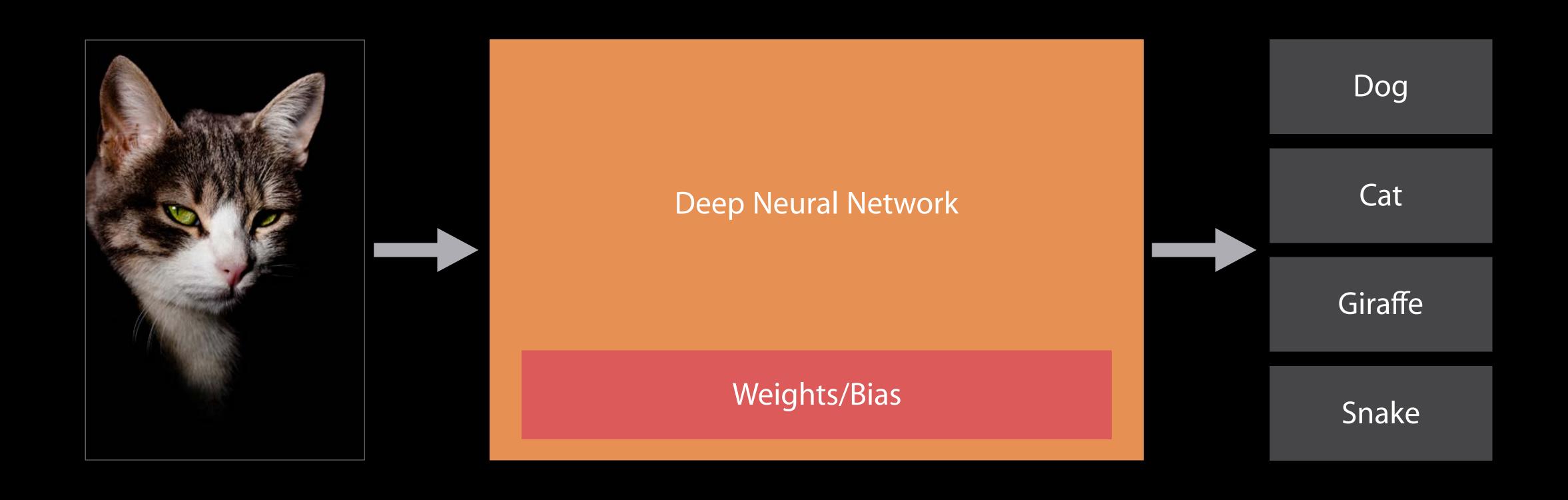
### Deep Neural Network

#### Training

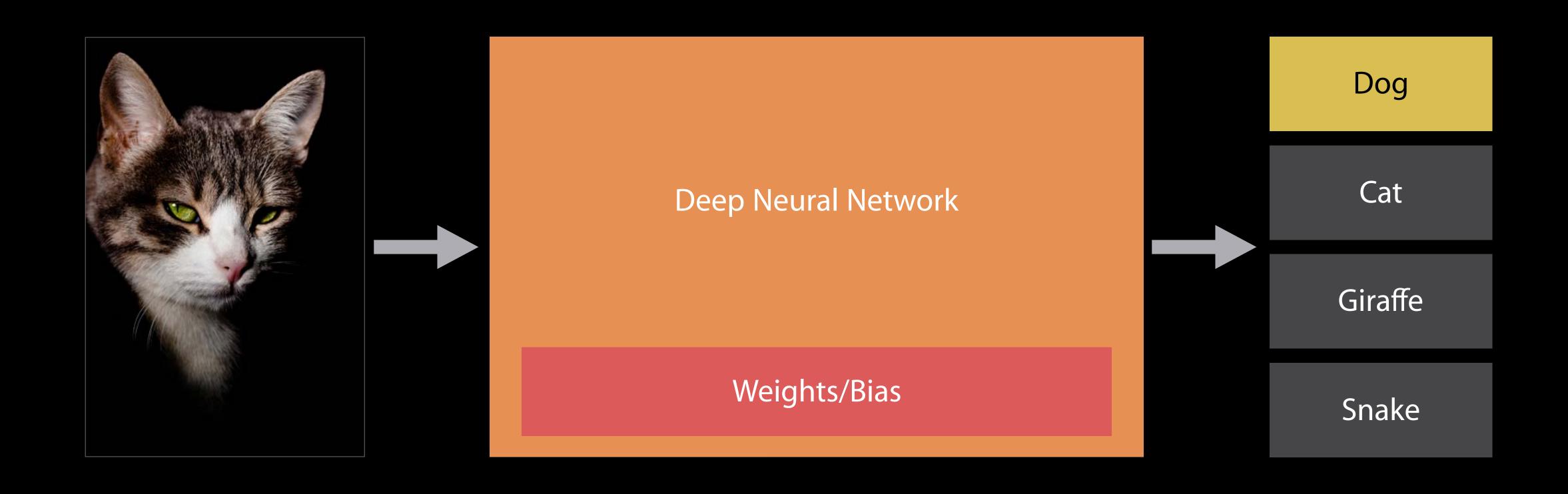


### Deep Neural Network

Training



Training



#### Training



Deep Neural Network

Weights/Bias

Dog

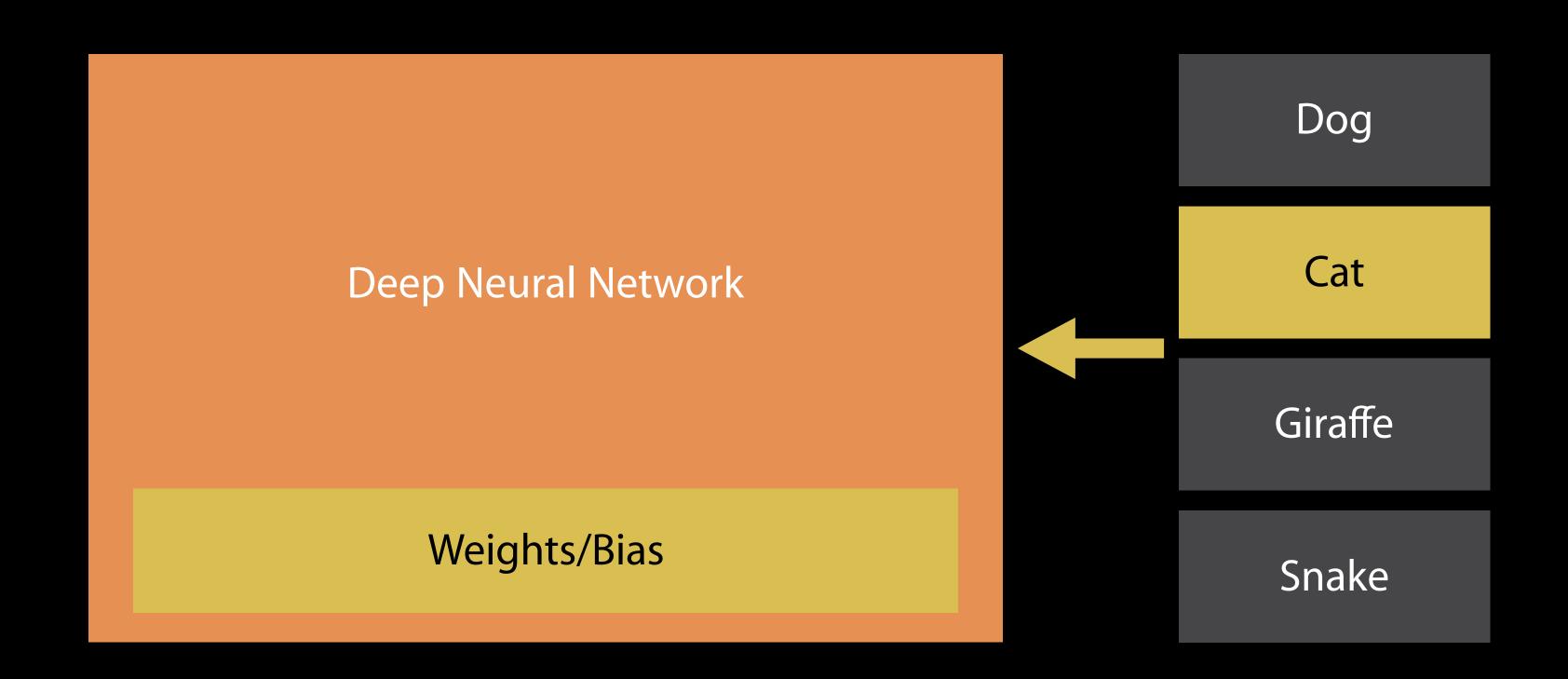
Cat

Giraffe

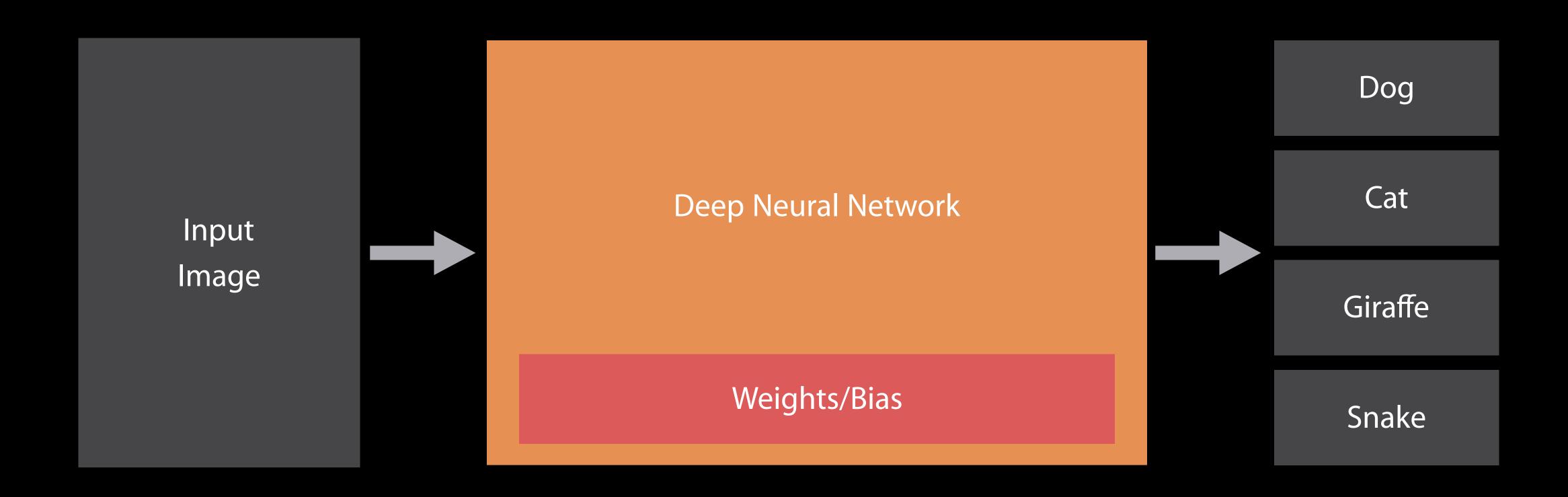
Snake

#### Training

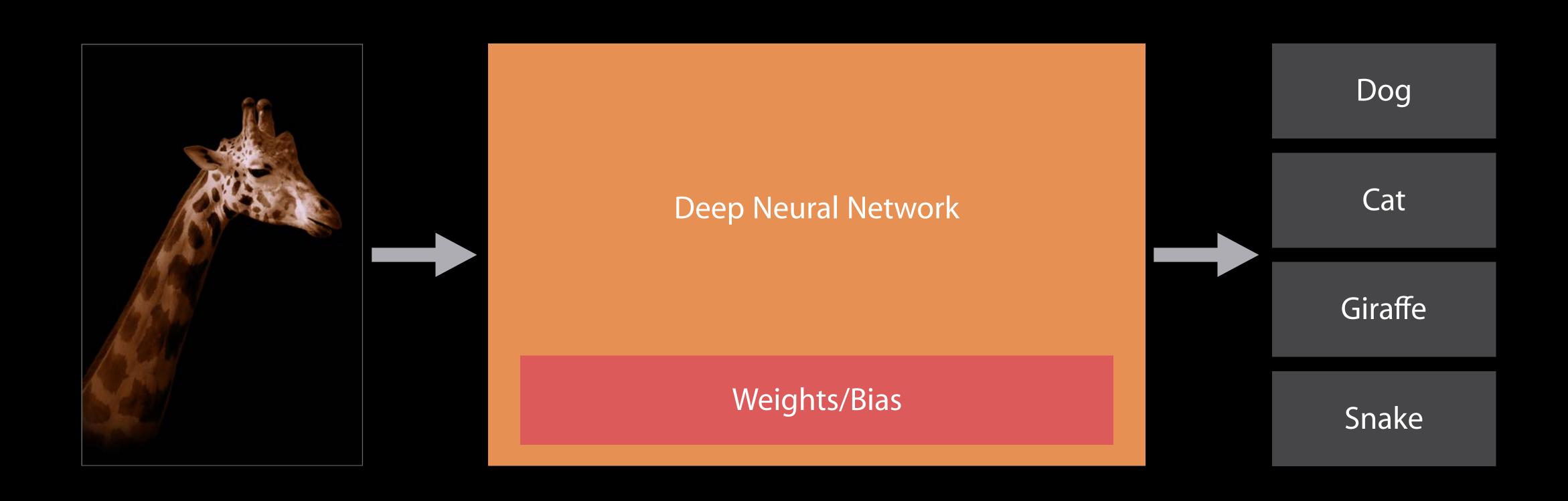




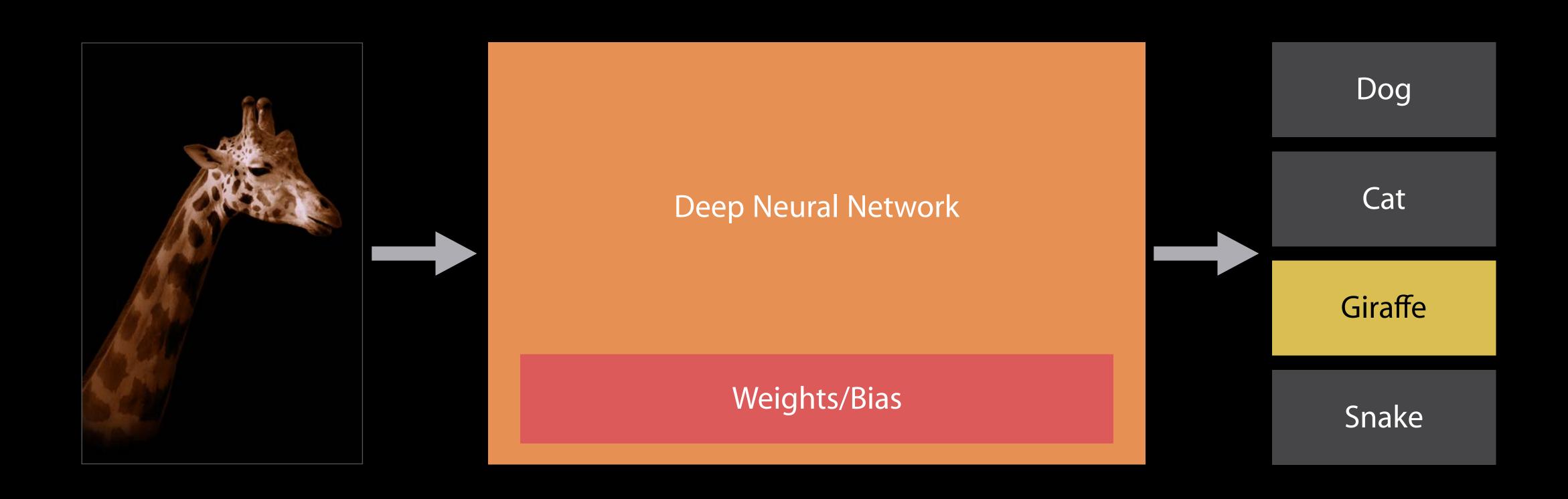
Inference

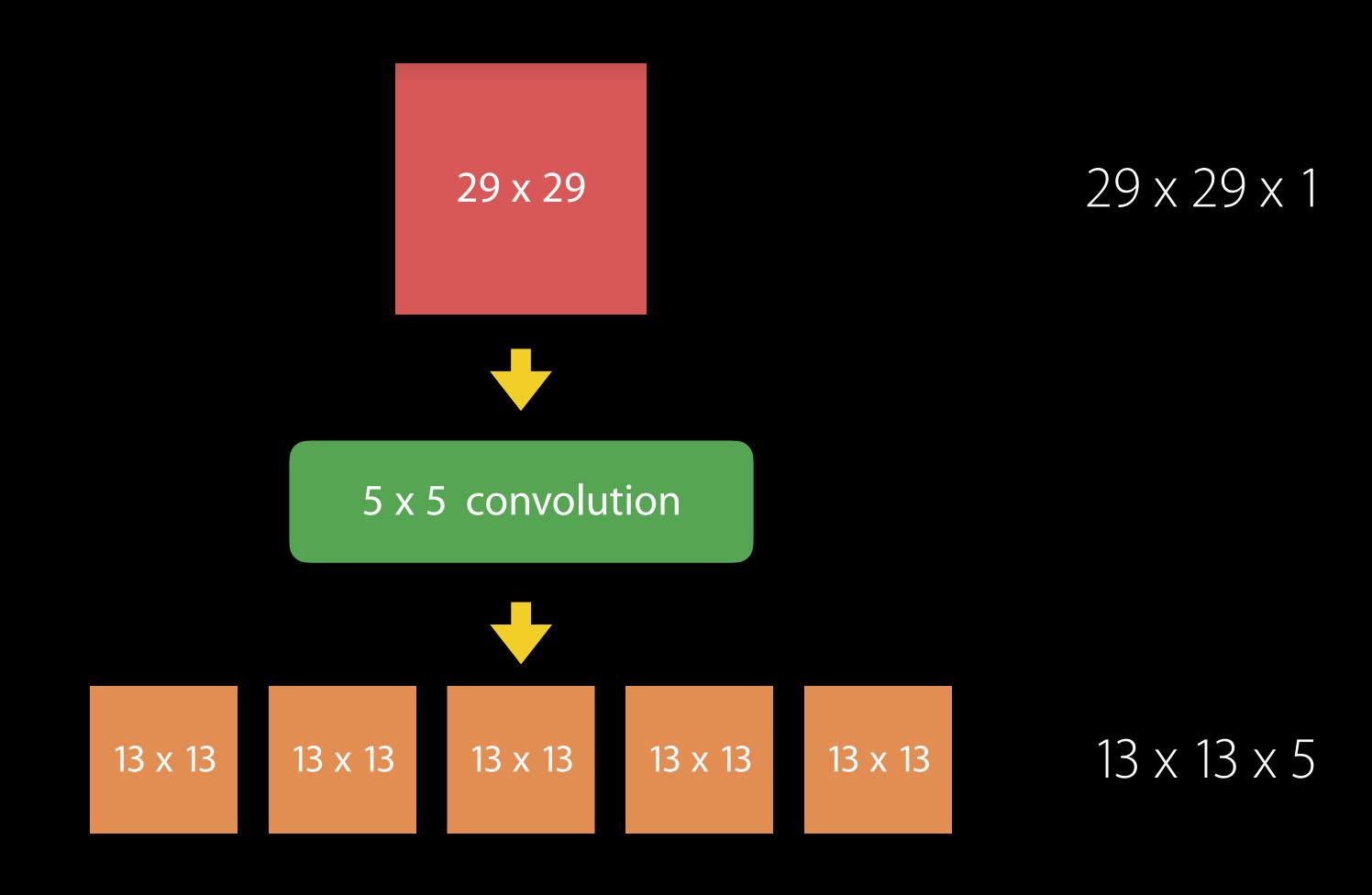


#### Inference

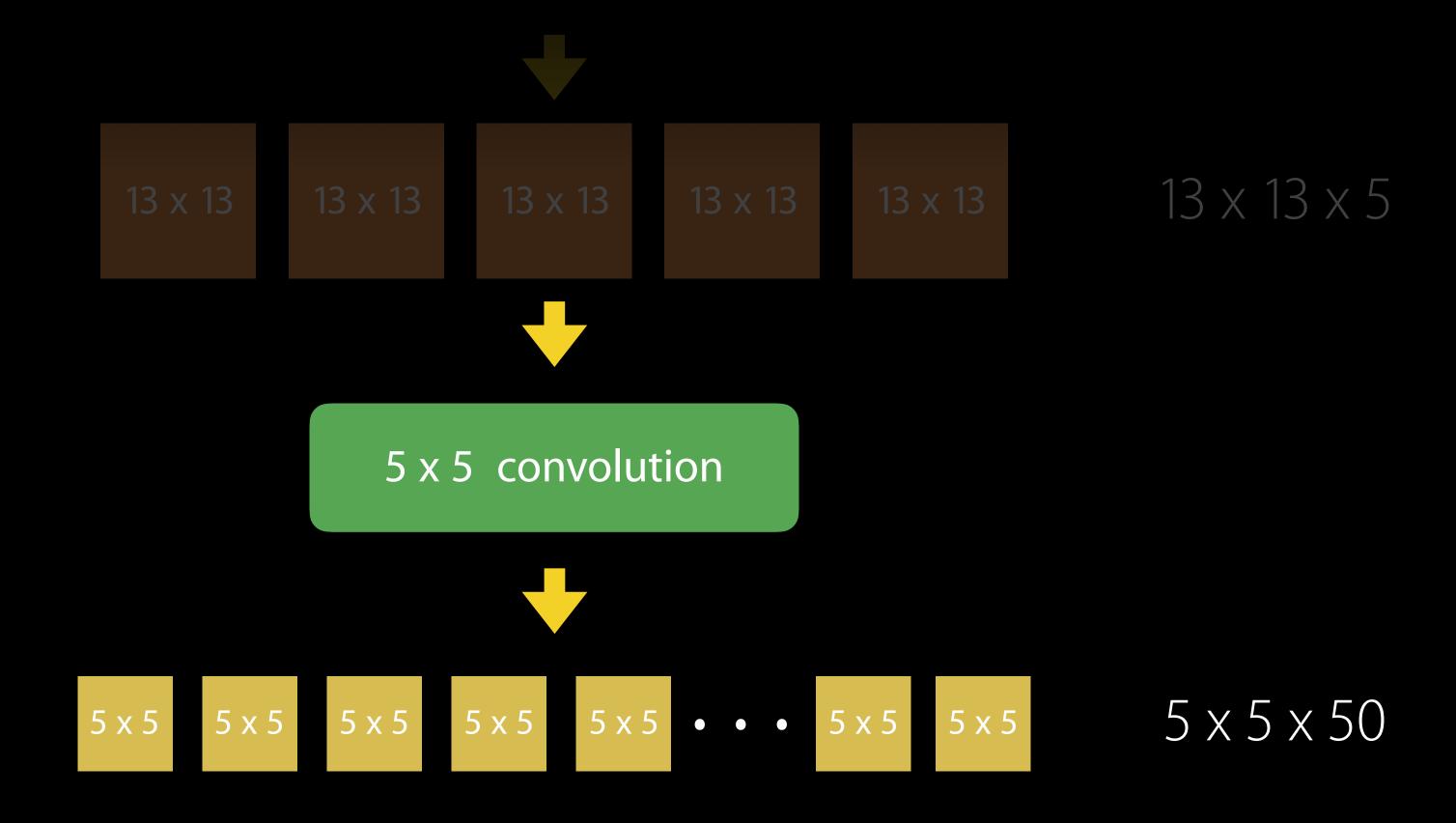


#### Inference

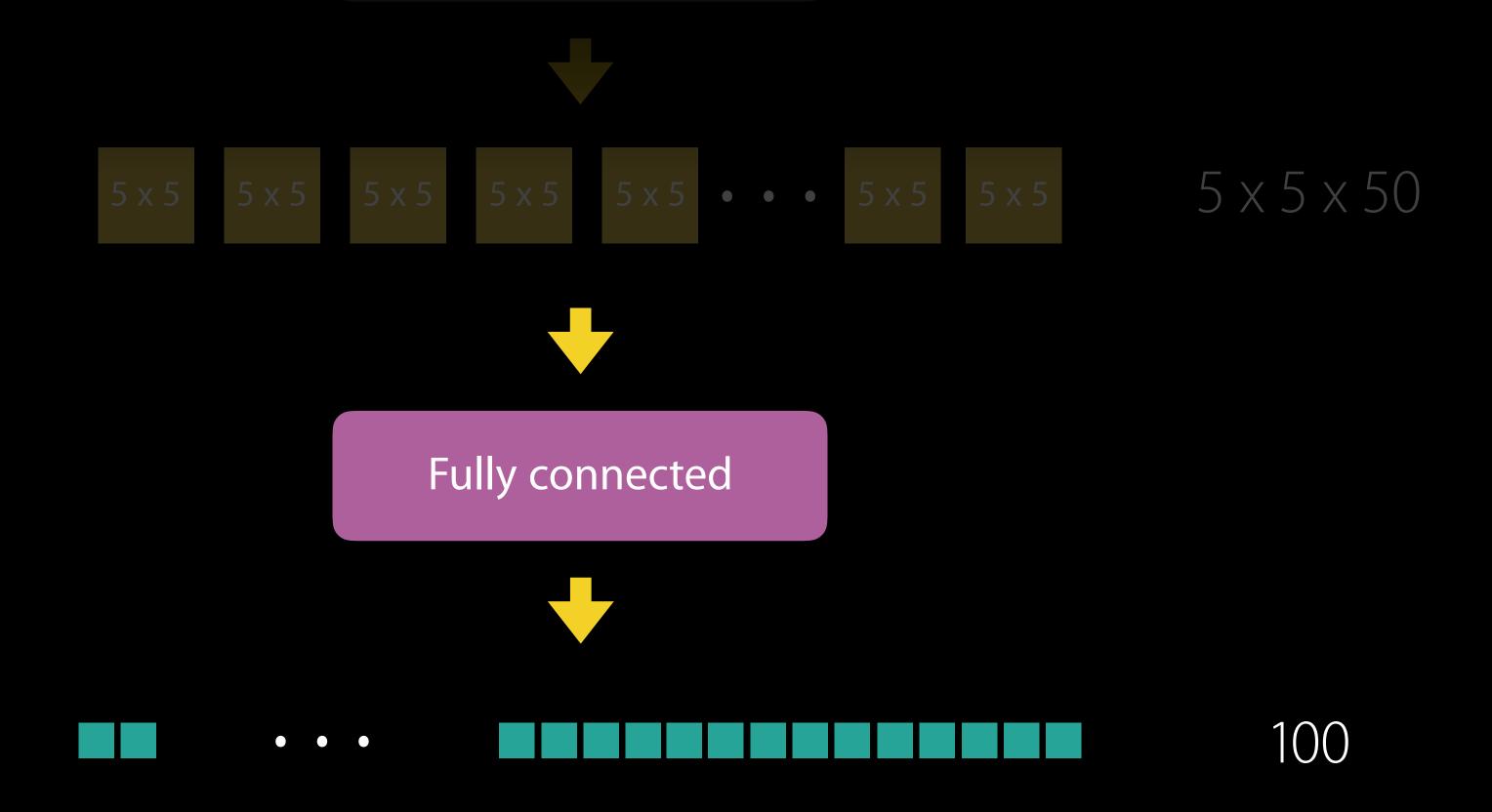




Example





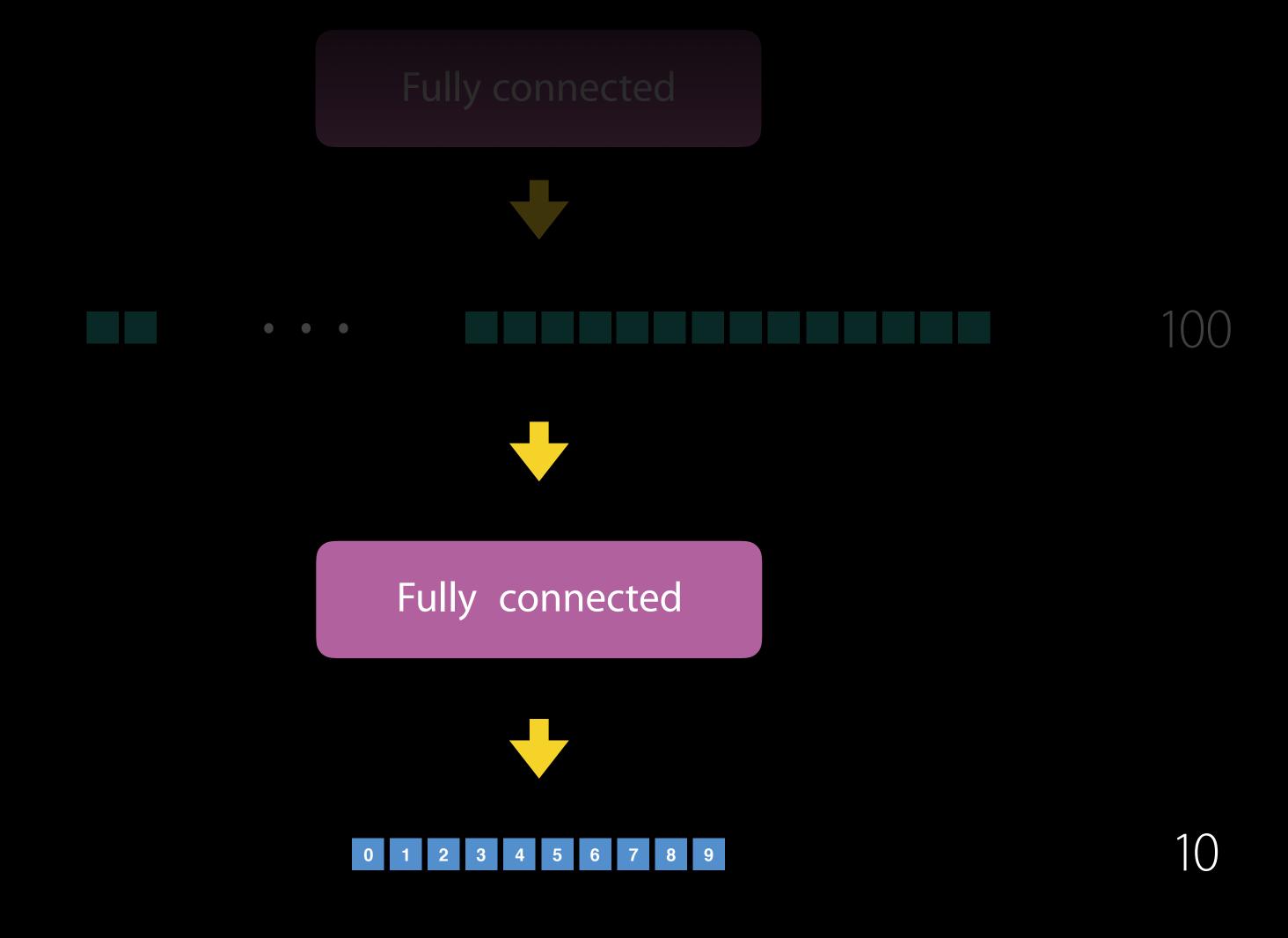


Example

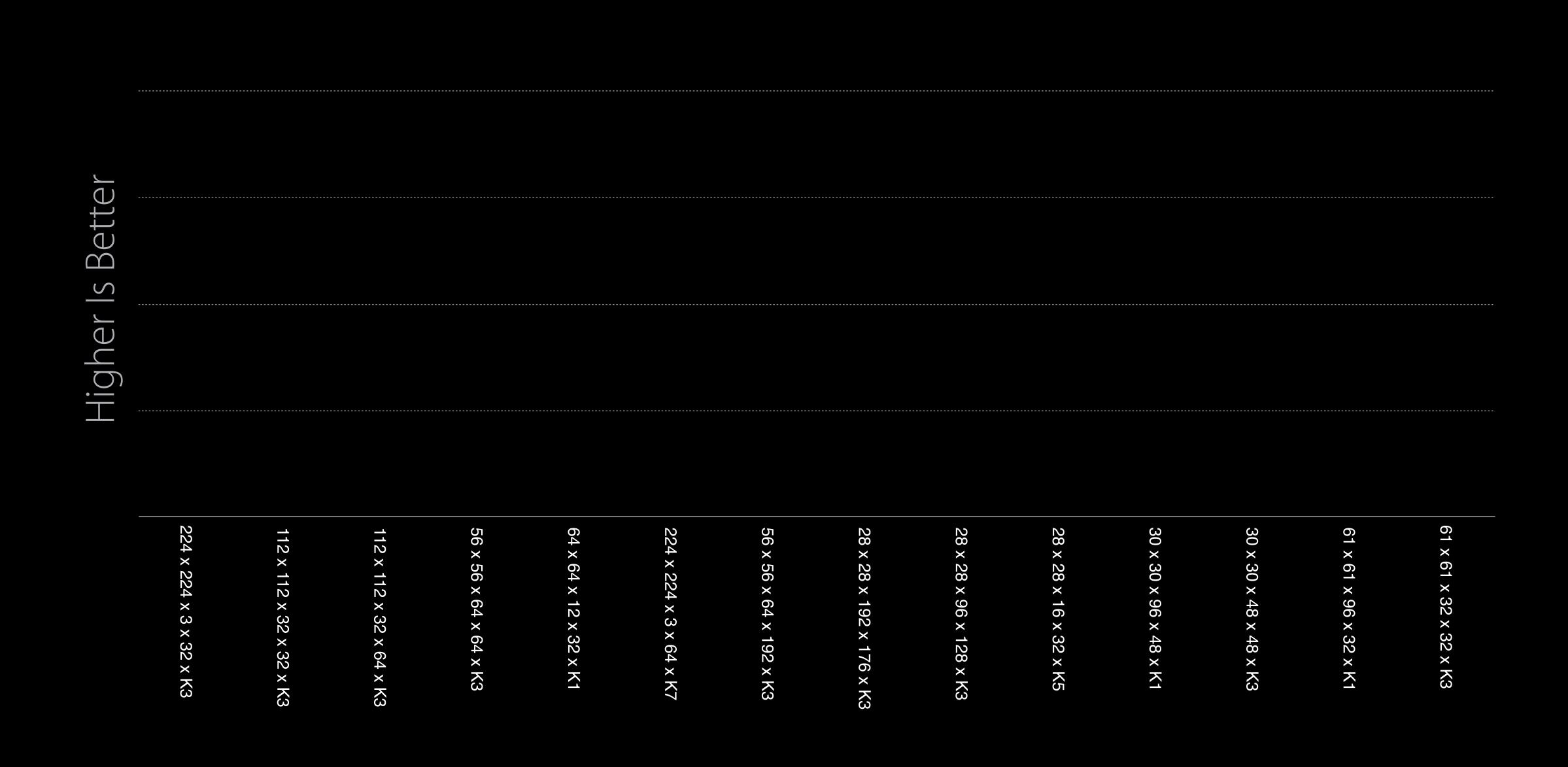
Fully connected



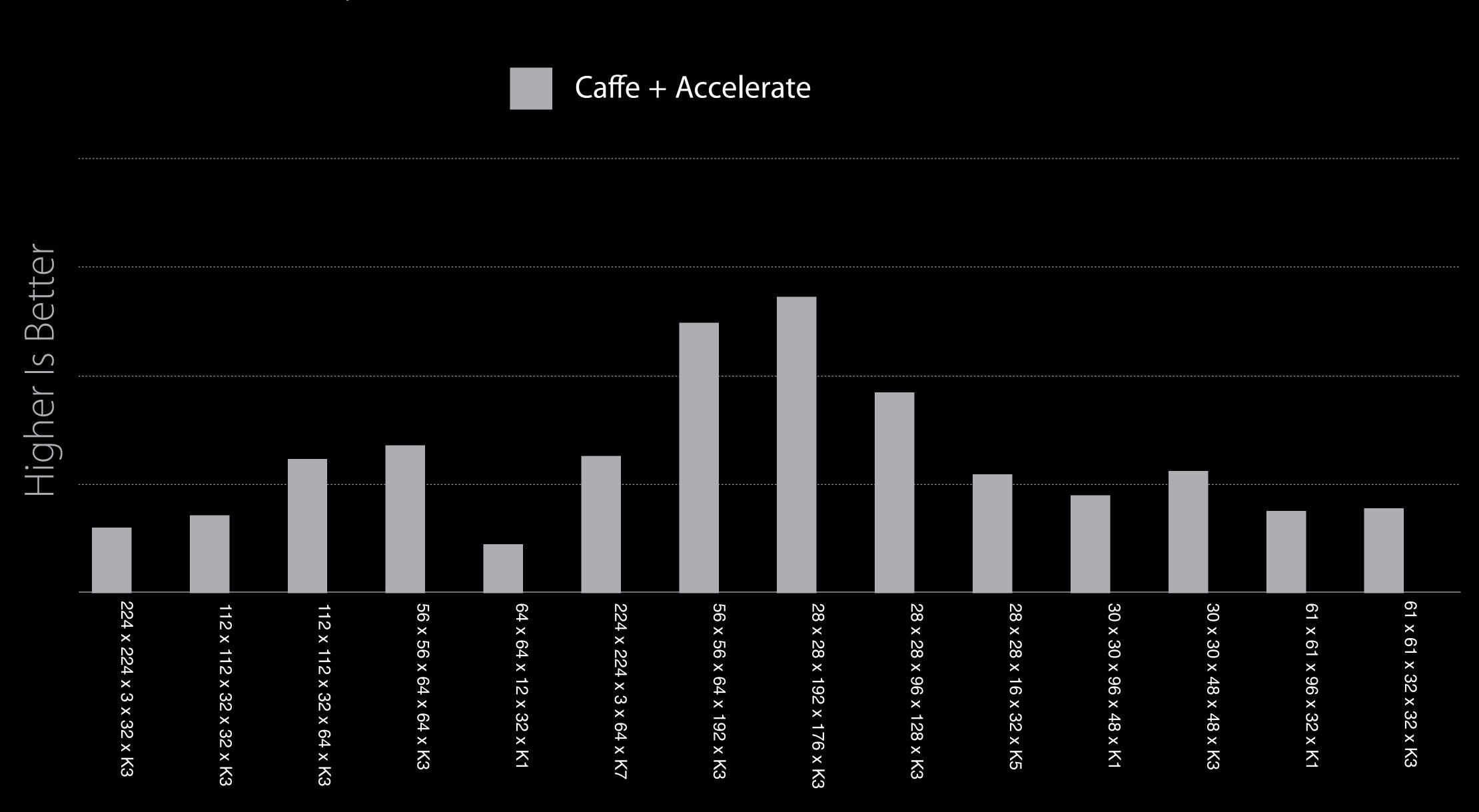
100



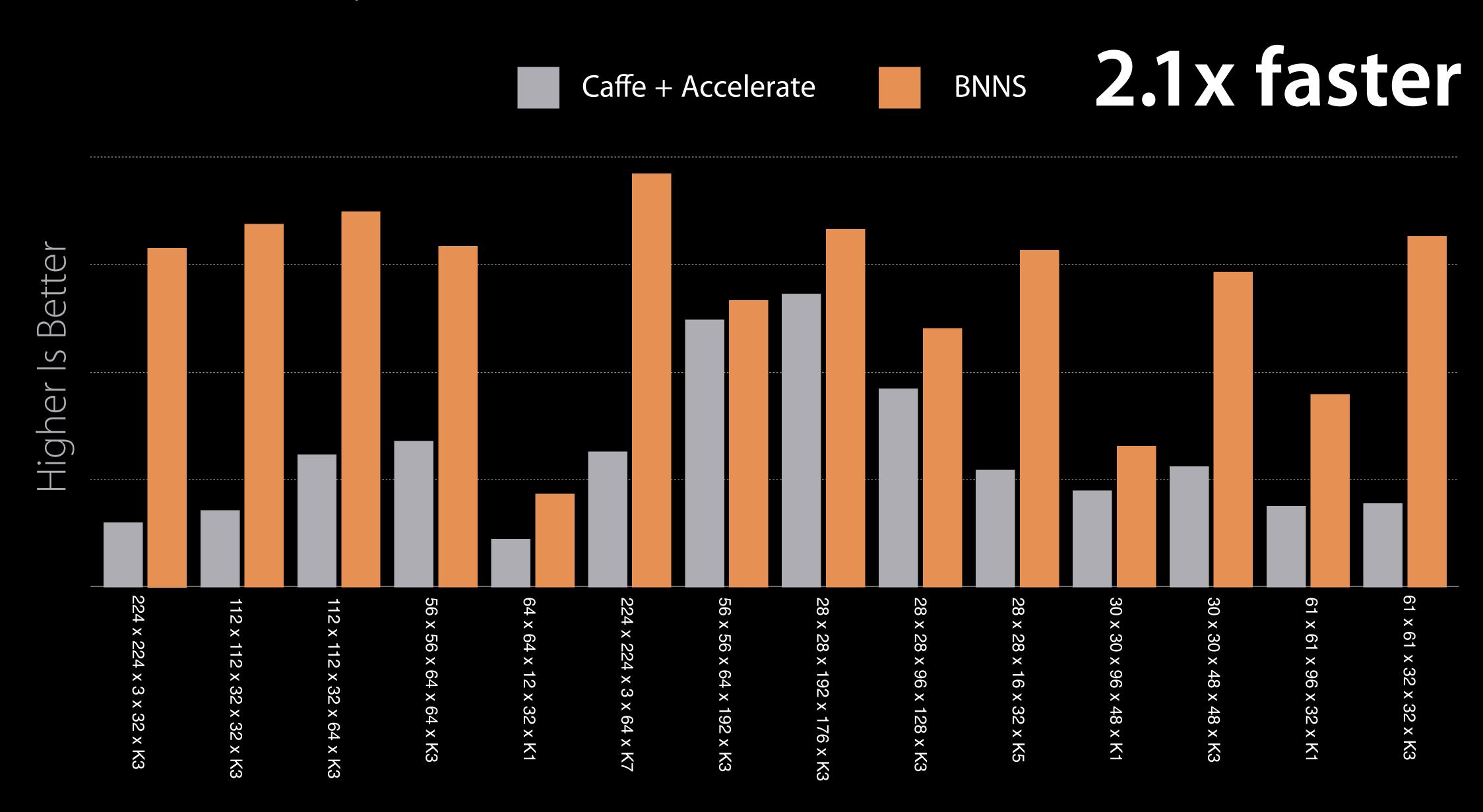
#### Performance, four core Haswell MacBook Pro



#### Performance, four core Haswell MacBook Pro



#### Performance, four core Haswell MacBook Pro



Features

Features

Low-level compute functions for CPU

#### Features

Low-level compute functions for CPU Inference only

#### Features

Low-level compute functions for CPU

Inference only

#### Features

Low-level compute functions for CPU

Inference only

Convolution layers

Pooling layers

#### Features

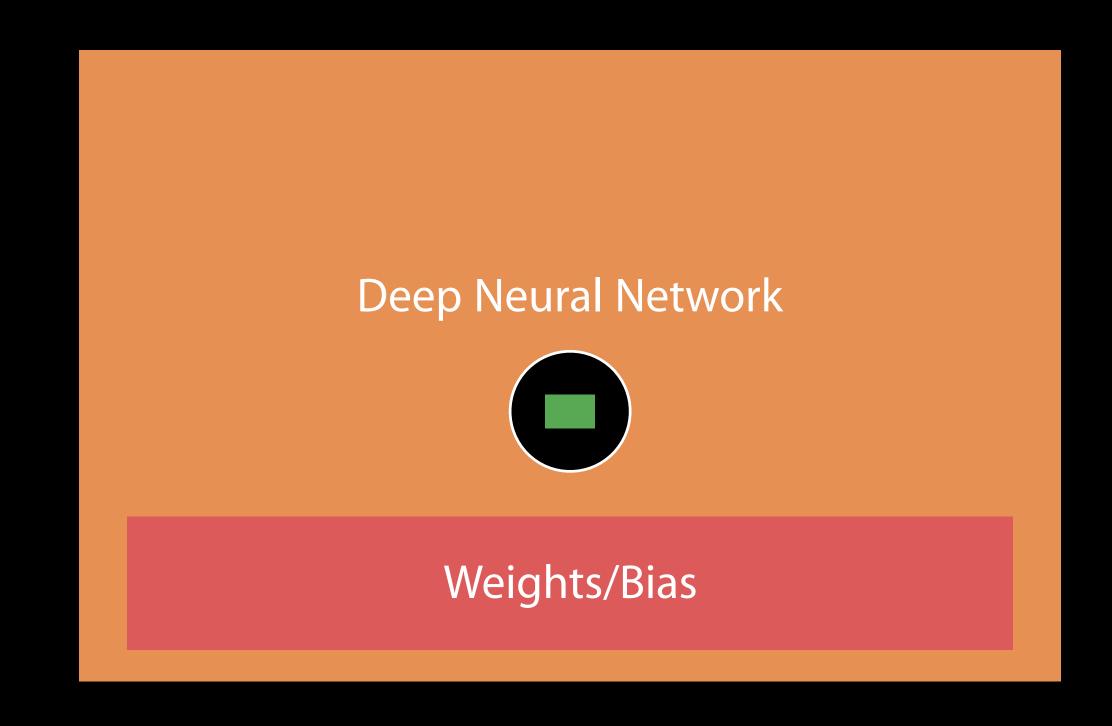
Low-level compute functions for CPU

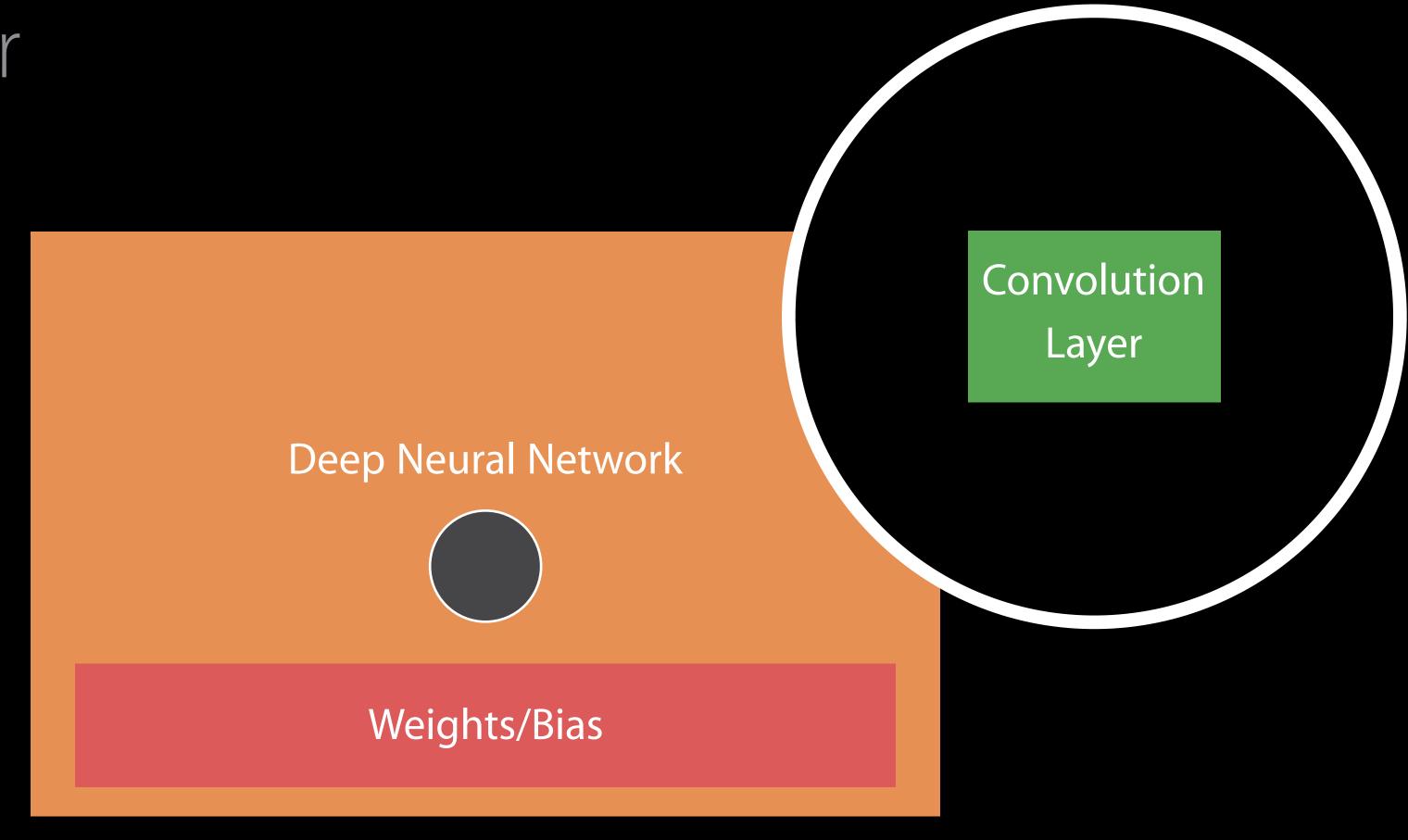
Inference only

Convolution layers

Pooling layers

Fully connected layers





Input image

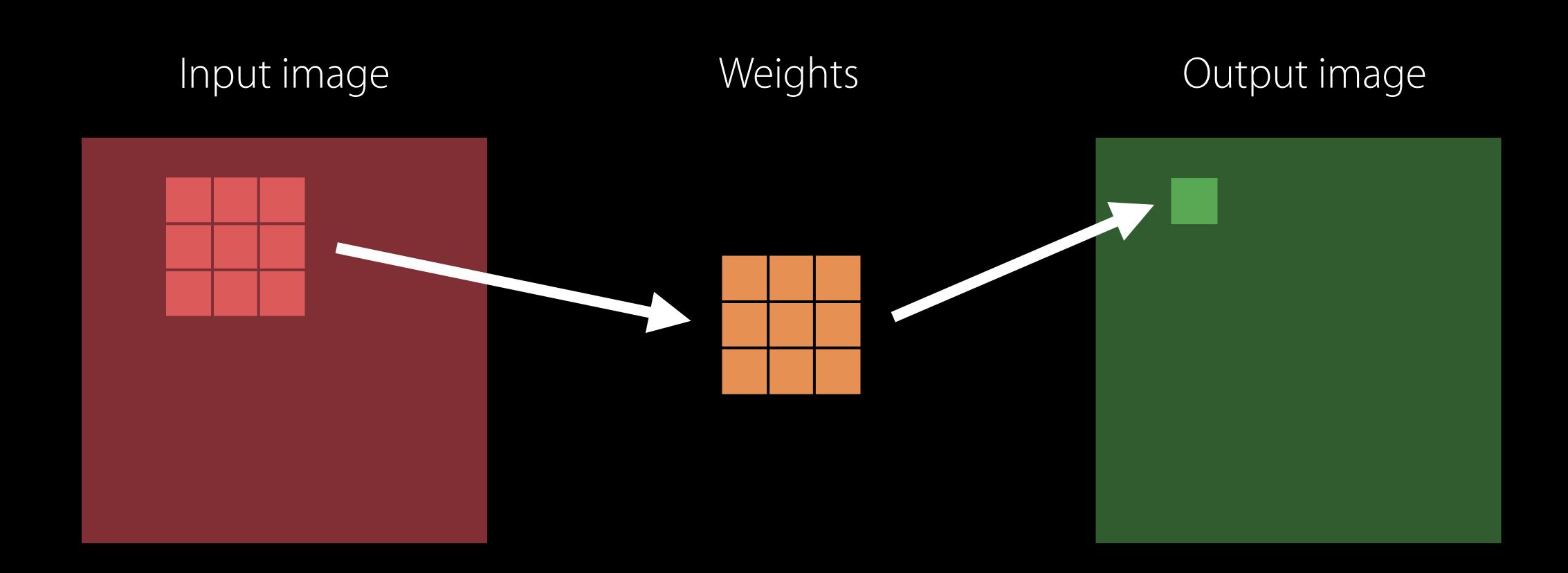
Weights

Output image

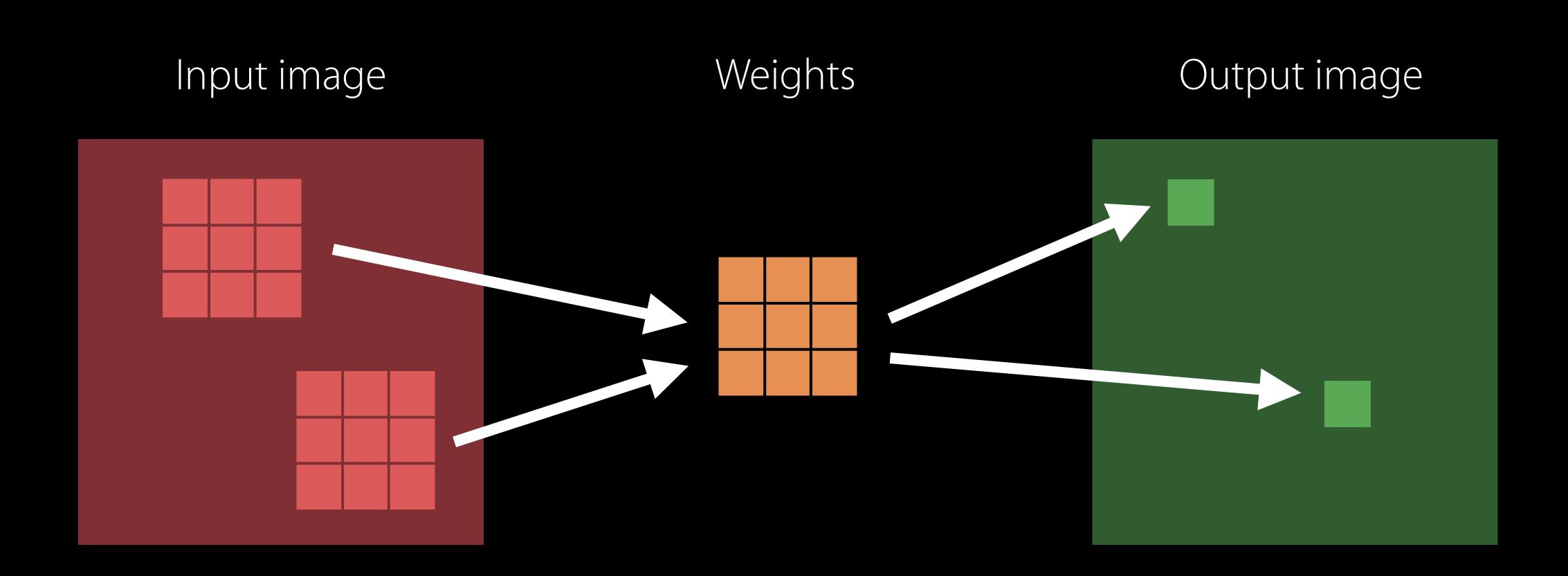
Input image

Weights

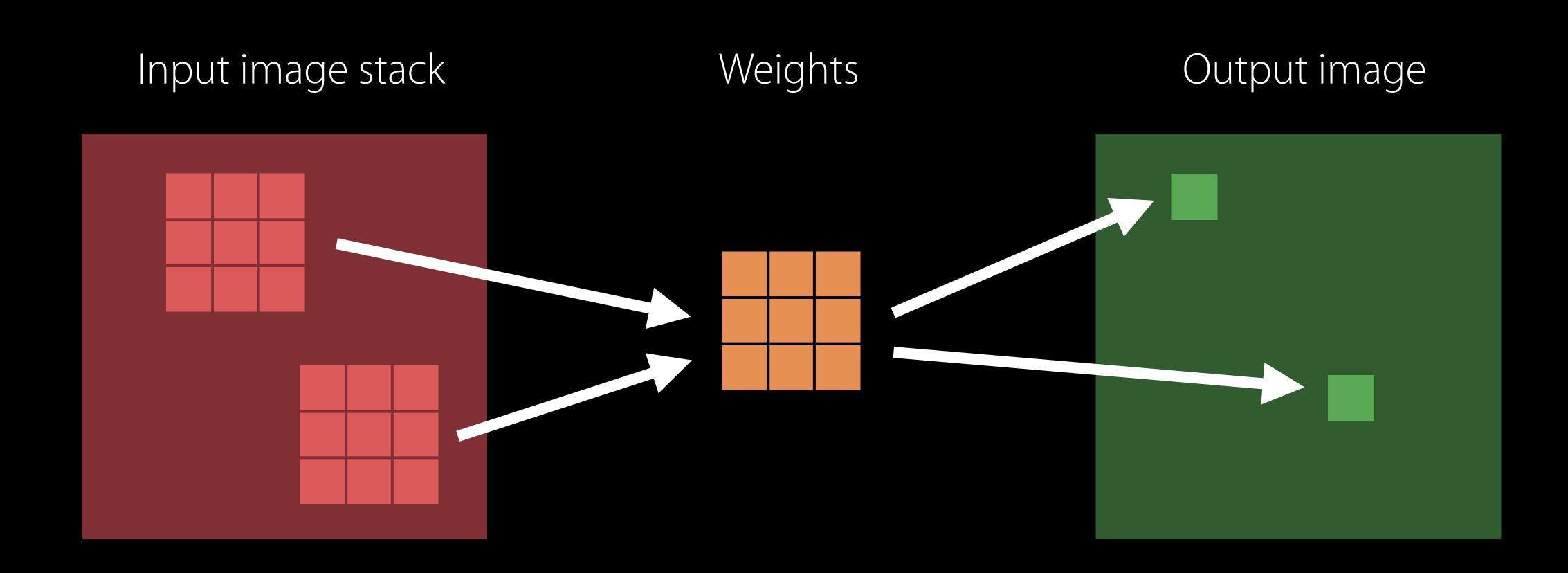
Output image

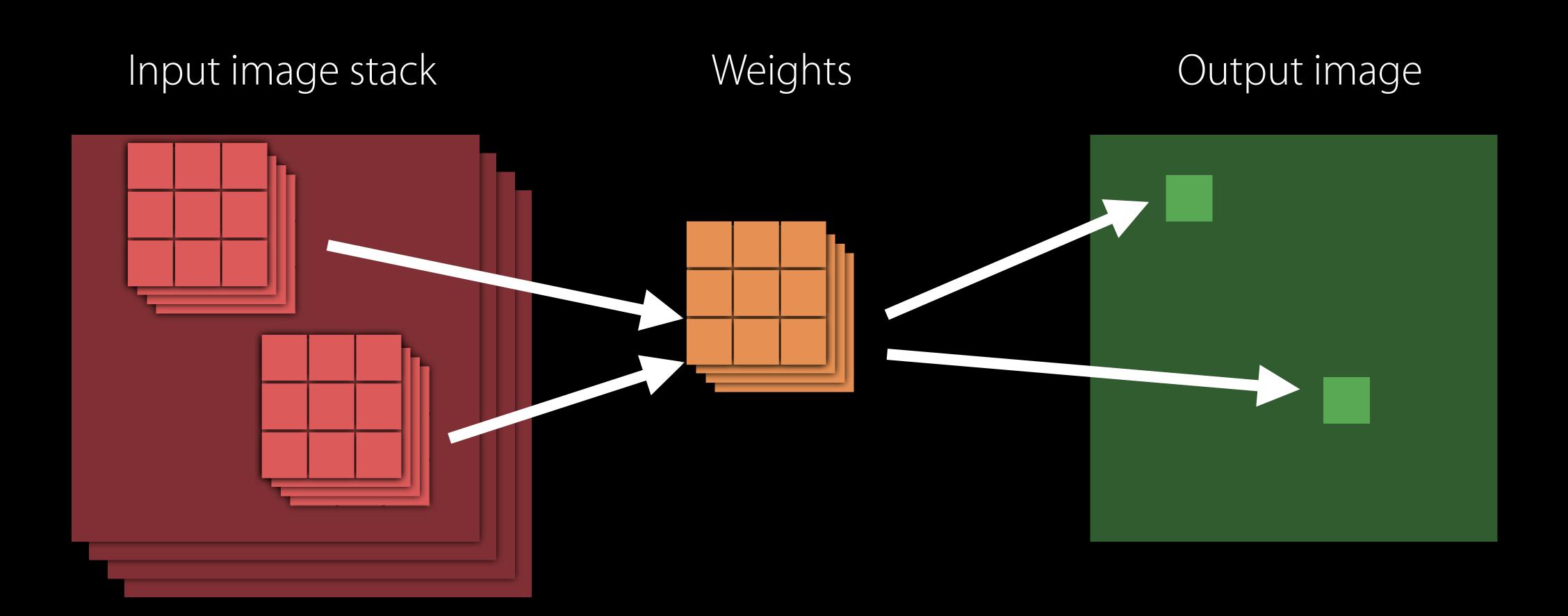


$$O(x,y) = \sum_{kx,ky} W(kx,ky)I(x+kx,y+ky)$$

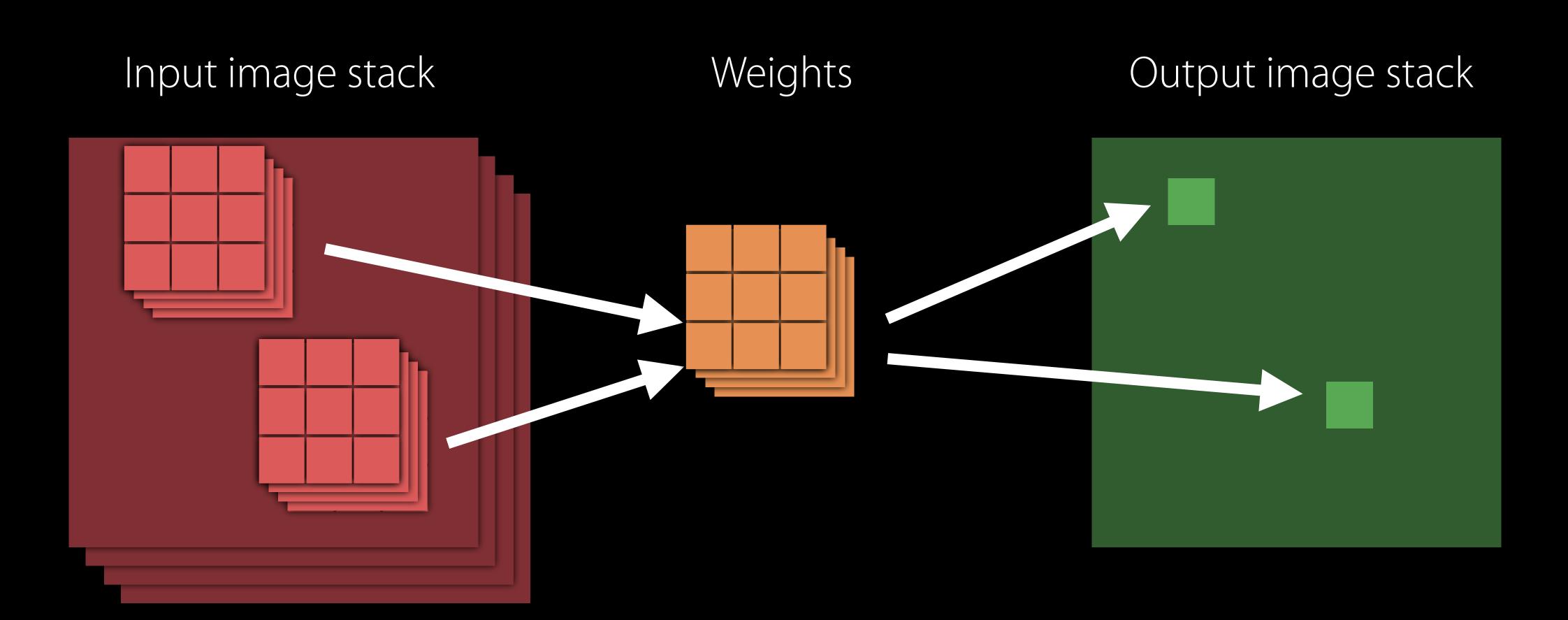


$$O(x,y) = \sum_{kx,ky} W(kx,ky)I(x+kx,y+ky)$$

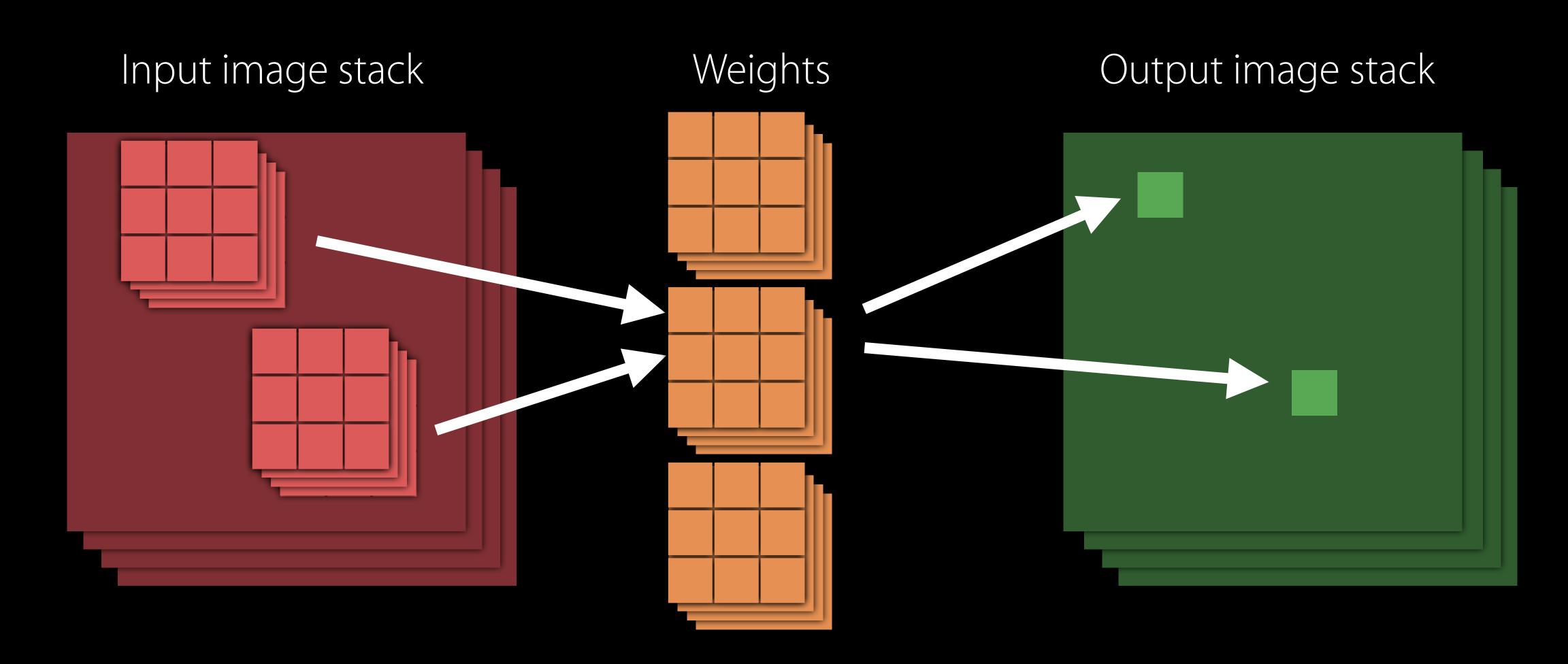




$$O(x,y) = \sum_{kx,ky,ic} W(kx,ky,ic)I(x+kx,y+ky,ic)$$



$$O(x,y) = \sum_{kx,ky,ic} W(kx,ky,ic)I(x+kx,y+ky,ic)$$



$$O(x, y, oc) = \sum_{kx, ky, ic} W(kx, ky, ic, oc) I(x + kx, y + ky, ic)$$

# Convolution Layer Example

Input image stack: 224 x 224 x 64

Output image stack: 222 x 222 x 96

Weights: 3 x 3 x 64 x 96

Floating point operations: 5.45 billion

All layers: 1-2 trillion

#include <Accelerate/Accelerate.h>

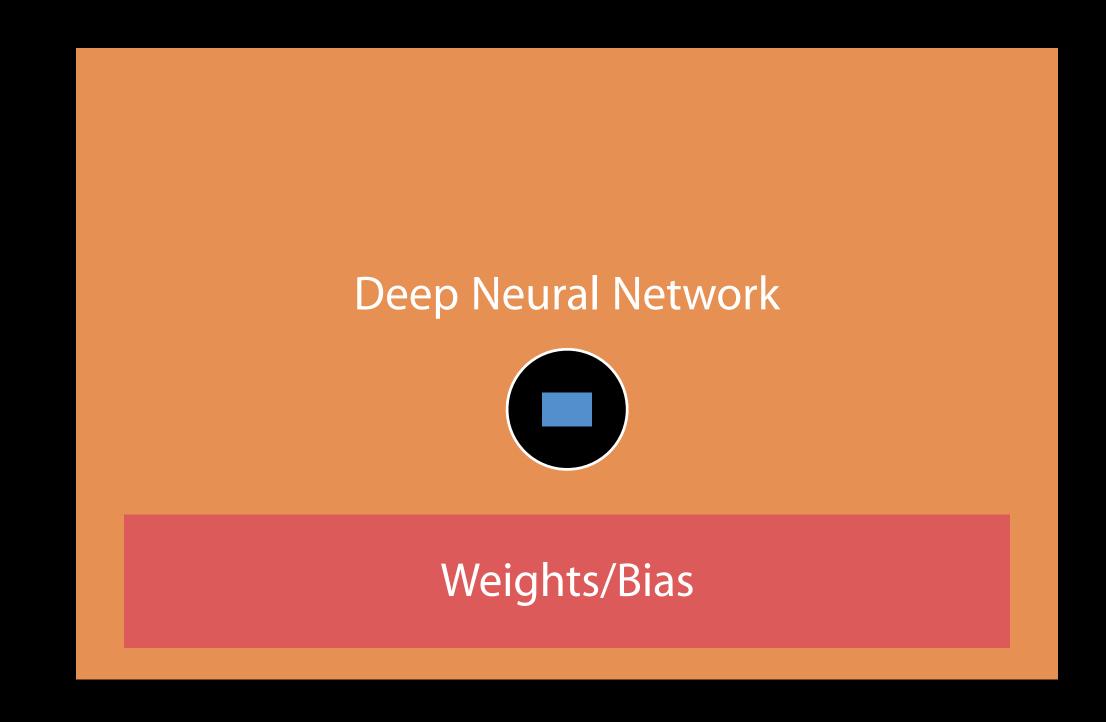
```
#include <Accelerate/Accelerate.h>
```

```
// Describe convolution layer
BNNSConvolutionLayerParameters conv = {
    .k_{width} = 3,
                                                 // kernel height
                                                 // kernel width
    .k_height = 3,
    x_padding = 0,
                                                 // X padding
    y_padding = 0,
                                                 // Y padding
    x_stride = 1,
                                                 // X stride
    y_stride = 1,
                                                 // Y stride
    in_{channels} = 64,
                                                 // input channels
    \bulletout_channels = 96,
                                                 // output channels
    weights = {
        data_type = BNNSDataTypeFloat16,
                                               // weights storage type
        .data = weights
                                                 // pointer to weights data
```

```
#include <Accelerate/Accelerate.h>
// Create convolution layer filter
BNNSFilter filter = BNNSFilterCreateConvolutionLayer(
                   // BNNSImageStackDescriptor for input stack
    &in_stack,
   &out_stack,
                   // BNNSImageStackDescriptor for output stack
                   // BNNSConvolutionLayerParameters
   &conv,
   NULL);
                   // BNNSFilterParameters (NULL = defaults)
// Use the filter ...
// Destroy filter
BNNSFilterDestroy(filter);
```

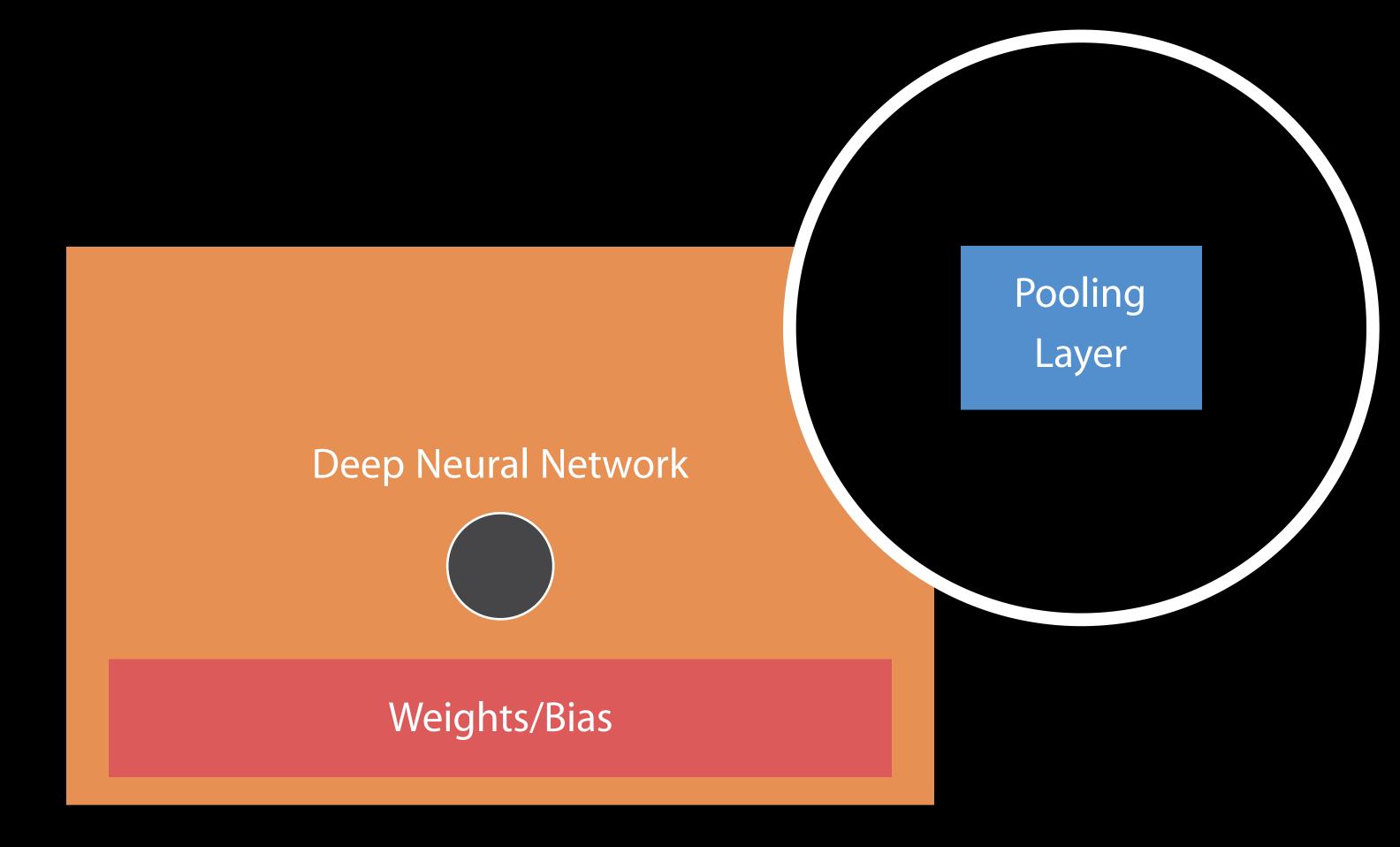
## Deep Neural Network

Pooling layer

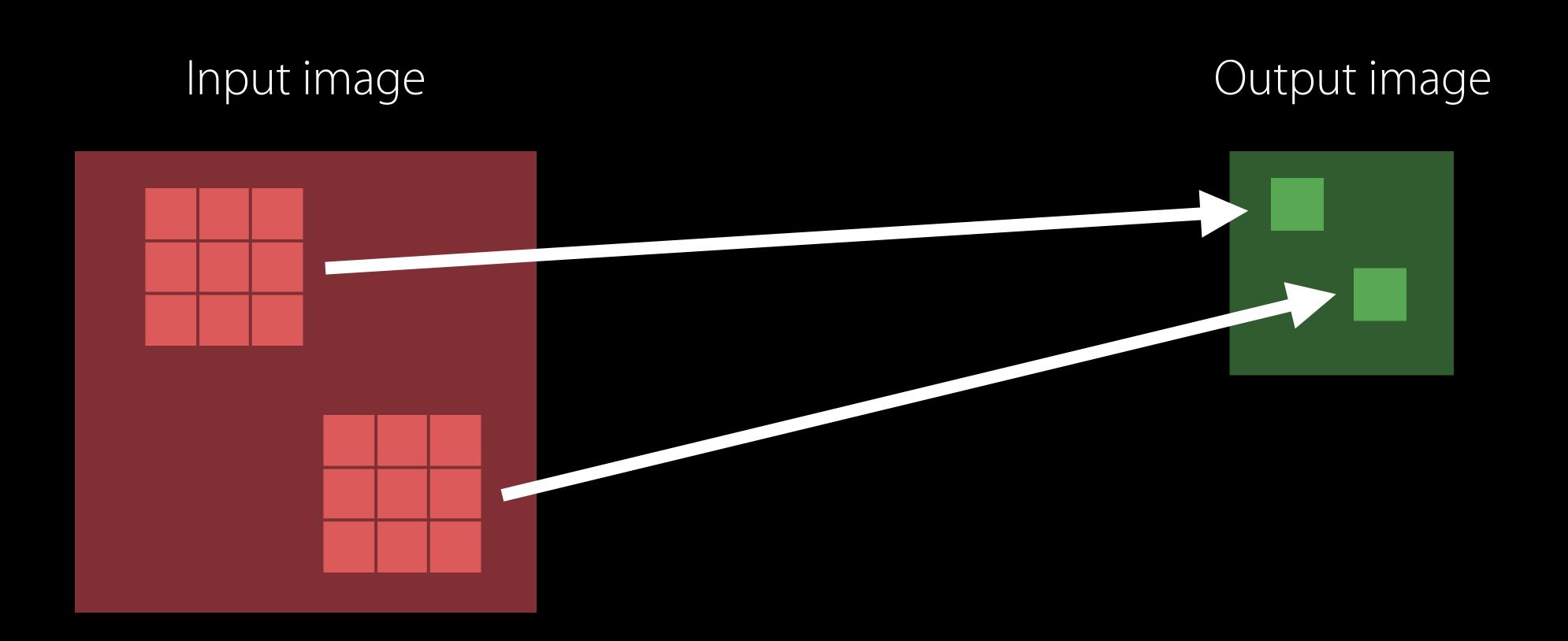


## Deep Neural Network

Pooling layer



## Pooling Layer



$$O(x, y, c) = \max_{i,j \le k} I(s_x \cdot x + i, s_y \cdot y + j, c)$$

```
// Describe pooling layer
BNNSPoolingLayerParameters pool = {
    .k_{width} = 3,
                                                 // kernel height
    .k_height = 3,
                                                 // kernel width
    x_padding = 1,
                                                 // X padding
    y_padding = 1,
                                                 // Y padding
    x_stride = 2,
                                                 // X stride
    y_stride = 2
                                                 // Y stride
    in_{channels} = 64,
                                                 // input channels
    \bulletout_channels = 64,
                                                 // output channels
    pooling_function = BNNSPoolingFunctionMax // pooling function
};
```

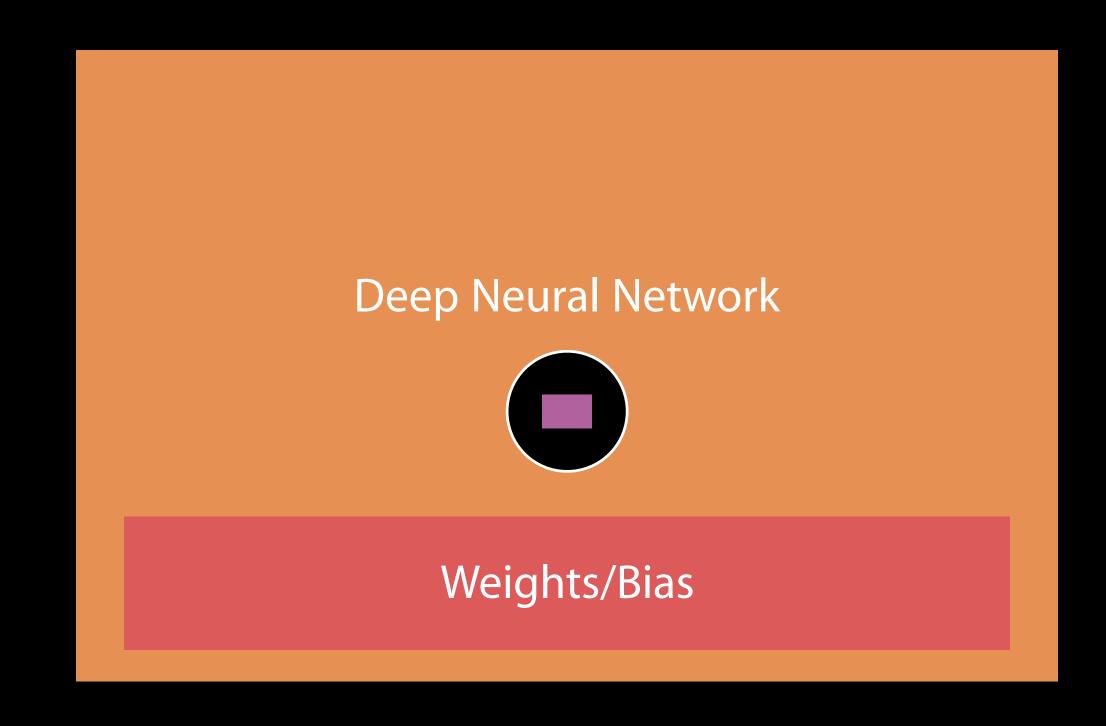
#include <Accelerate/Accelerate.h>

```
// Create pooling layer filter
BNNSFilter filter = BNNSFilterCreatePoolingLayer(
                  // BNNSImageStackDescriptor for input stack
   &in_stack,
                  // BNNSImageStackDescriptor for output stack
   &out_stack,
   &pool,
                  // BNNSPoolingLayerParameters
   NULL);
                  // BNNSFilterParameters (NULL = defaults)
// Use the filter ...
// Destroy filter
BNNSFilterDestroy(filter);
```

#include <Accelerate/Accelerate.h>

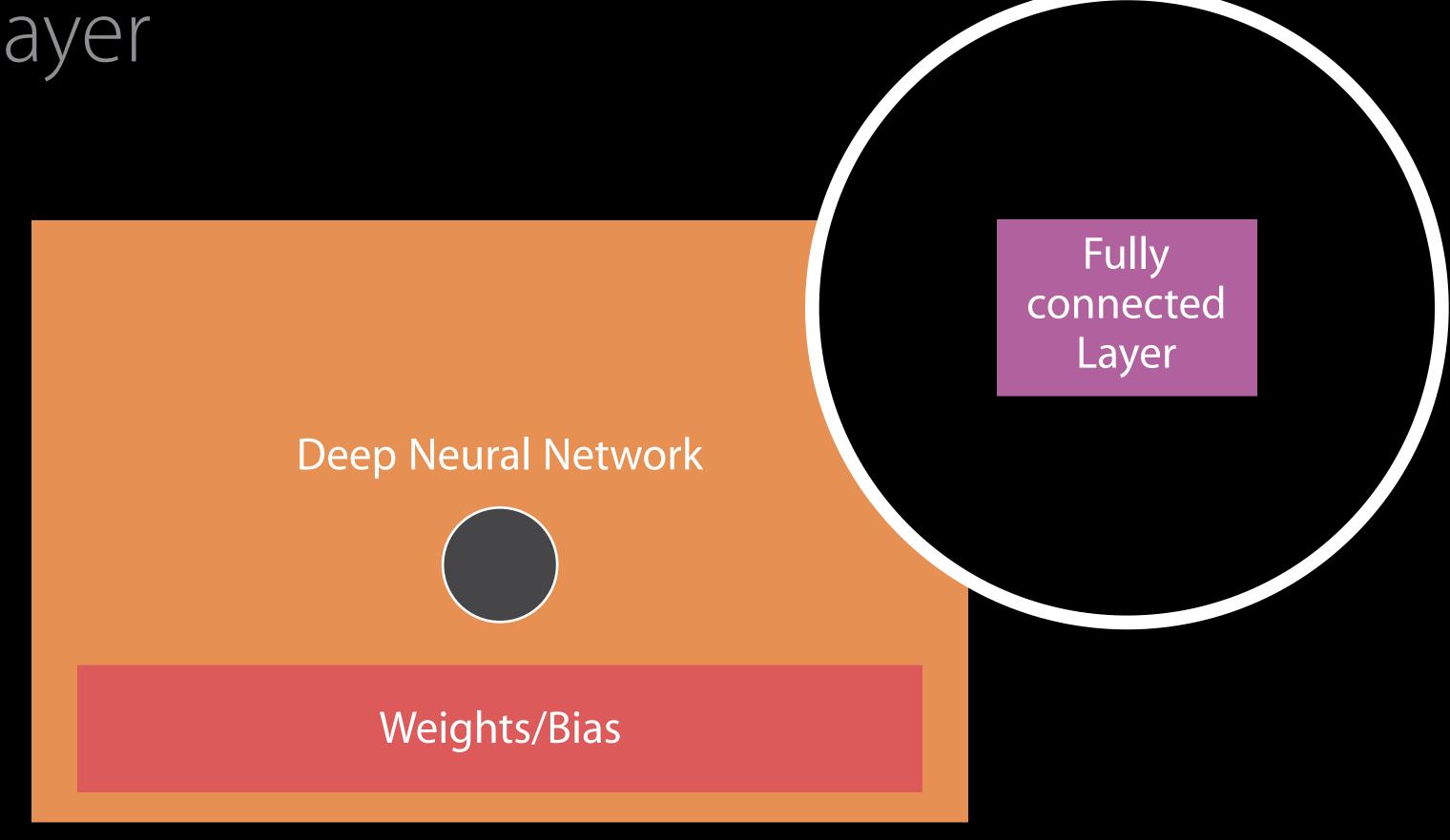
## Deep Neural Network

Fully connected layer



## Deep Neural Network

Fully connected layer

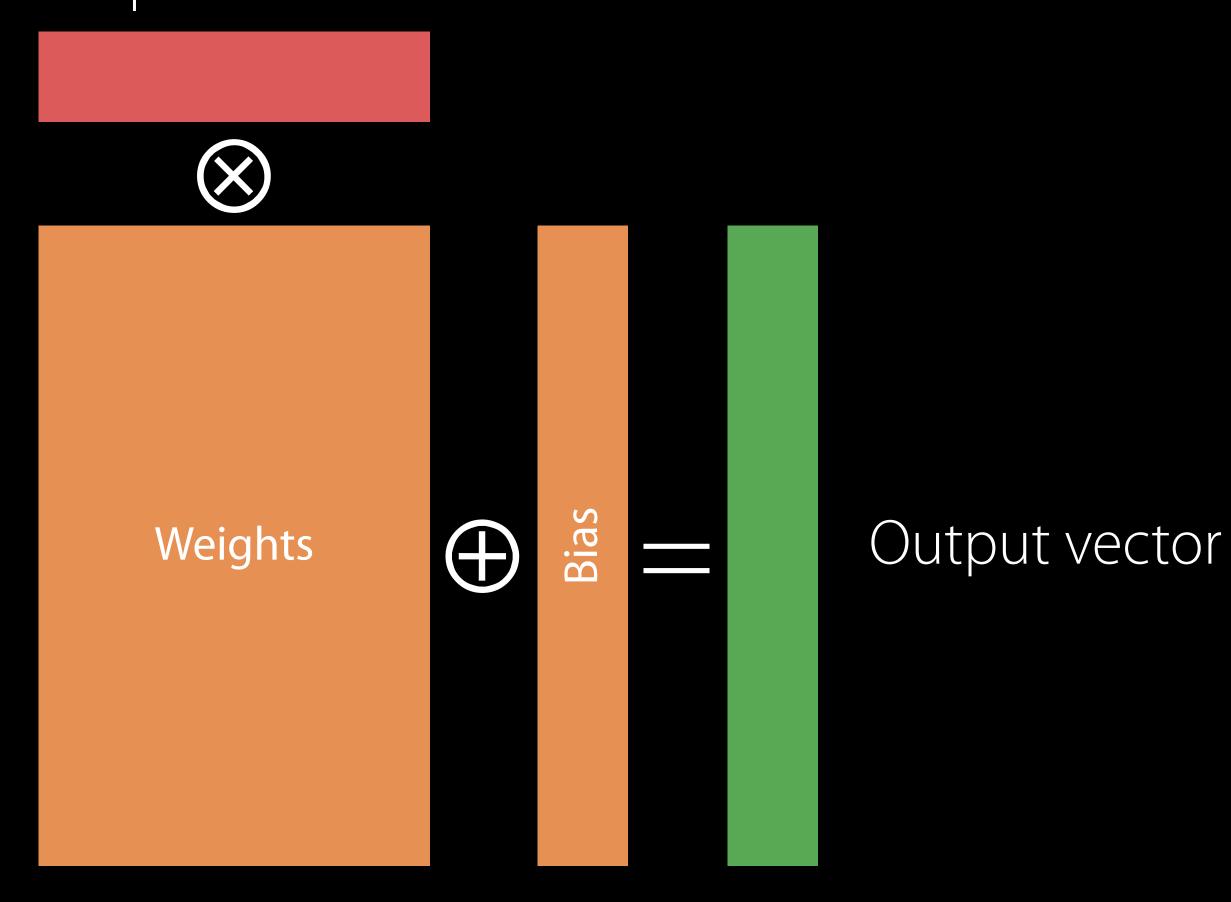


## Fully Connected Layer

Input vector

## Fully Connected Layer

Input vector



$$O(i) = \sum_{j} W(i,j)I(j) + B(i)$$

```
#include <Accelerate/Accelerate.h>
```

```
// Create fully connected layer filter
BNNSFilter filter = BNNSFilterCreateFullyConnectedLayer(
   &in_vec,
                   // BNNSVectorDescriptor for input vector
                   // BNNSVectorDescriptor for output vector
   &out_vec,
   &full,
                   // BNNSFullyConnectedLayerParameters
   NULL);
                   // BNNSFilterParameters (NULL = defaults)
// Use the filter ...
// Destroy filter
BNNSFilterDestroy(filter);
```

#include <Accelerate/Accelerate.h>

// pointer to output data

out);

#### #include <Accelerate/Accelerate.h>

Low-level compute functions for neural networks

Low-level compute functions for neural networks

Fast and energy-efficient inference

Low-level compute functions for neural networks

Fast and energy-efficient inference

Multiple storage types

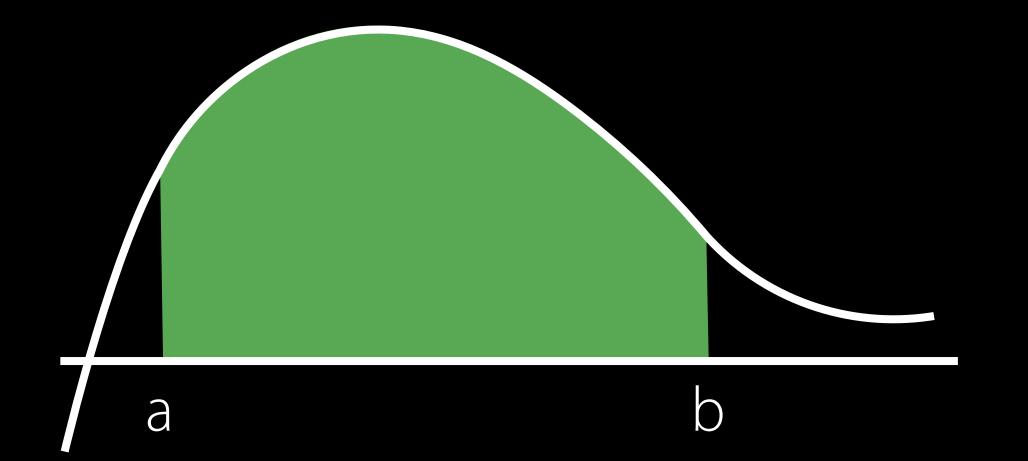
## Quadrature

Numerical integration

## Quadrature

#### Numerical integration

$$\int_{a}^{b} f(x) \, \mathrm{d}x$$



```
#include <Accelerate/Accelerate.h> // Quadrature is part of Accelerate
// Describe the function to integrate
quadrature_integrate_function fun = {
                                          // evaluation callback
    fun = f
};
// Evaluates the function at n points x[i] -> y[i]
void f(void *arg, size_t n, const double *x, double *y)
    for (size_t i=0; i<n; i++) {</pre>
        y[i] = 1.0 / (1.0 + x[i] * x[i]);
```

```
#include <Accelerate/Accelerate.h> // Quadrature is part of Accelerate
// Describe the integration method and parameters
quadrature_integrate_options opt = {
                                                // integration algorithm
    integrator = QUADRATURE_INTEGRATE_QAG,
    .abs_tolerance = 1.0e-8,
                                                 // requested tolerance
                                                 // max number of intervals for QAG
    max_intervals = 12
};
// QNG
        simple non-adaptive integrator
        simple globally adaptive integrator
// QAG
        globally adaptive integrator with convergence acceleration
// QAGS
```

```
// Compute the integral
quadrature_status status;
double est_error;
double result = quadrature_integrate(
   &fun,
               // quadrature_integrate_function, function to integrate
   -1.0, // a, first bound of interval
   2.0,
          // b, second bound of interval
               // quadrature_integrate_options, integration method and options
   &opt,
   &status,
               // quadrature_status, receives success/failure code
               // double, receives the estimated absolute error
   &est_error,
   0, NULL);
               // optional args
```

#include <Accelerate/Accelerate.h>

## Simo

Vector and geometry operations

Steve Canon Core OS, Vector and Numerics Group

## simd

#### Simo

Geometric operations on vectors and matrices for C, Objective-C, C++, and Swift

#### simd

Geometric operations on vectors and matrices for C, Objective-C, C++, and Swift Closely mirrors Metal shading language

# simd Types

## simd

Types

Vectors of floats, doubles, signed and unsigned integers of length 2, 3, and 4

#### simo

#### Types

Vectors of floats, doubles, signed and unsigned integers of length 2, 3, and 4 Matrices of floats and doubles, of size NxM, where N and M are 2, 3, or 4

## simd Operations

## simd

Operations

Arithmetic operators on vectors and matrices

## sima

Operations

Arithmetic operators on vectors and matrices Geometry and shader functions

```
// myCode.m:
@import simd;
vector_float3 reflect(vector_float3 x, vector_float3 n) {
    return x - 2*vector_dot(x,n)*n;
// myCode.cpp:
#include <simd/simd.h>
using namespace simd;
float3 reflect(float3 x, float3 n) {
    return x - 2*dot(x,n)*n;
// myCode.swift:
import simd
func reflect(x: float3, n: float3) -> float3 {
    return x - 2*dot(x,n)*n
```

```
// myCode.m:
@import simd;
vector_float3 reflect(vector_float3 x, vector_float3 n) {
    return x - 2*vector_dot(x,n)*n;
// myCode.cpp:
#include <simd/simd.h>
using namespace simd;
float3 reflect(float3 x, float3 n) {
    return x - 2*dot(x,n)*n;
// myCode.swift:
import simd
func reflect(x: float3, n: float3) -> float3 {
    return x - 2*dot(x,n)*n
```

### simd

Interoperation between languages

Interoperation between languages

Vector types are compiler extensions in C, Objective-C, and C++

Interoperation between languages

Vector types are compiler extensions in C, Objective-C, and C++ Swift vector types are structs

#### Interoperation between languages

Vector types are compiler extensions in C, Objective-C, and C++

Swift vector types are structs

The compiler maps between corresponding vector types for you

```
// myHeader.h:
@import simd;

vector_float3 someFunction(vector_float3 x, vector_float3 y);

// myCode.swift:
import simd

let x = float3(1,2,3)
let y = float3(0,0,1)
// Vector types are bridged automatically.
let z = someFunction(x, y)
```

```
// myHeader.h:
@import simd;

vector_float3 someFunction(vector_float3 x, vector_float3 y);

// myCode.swift:
import simd
```

let x = float3(1,2,3)

let y = float3(0,0,1)

let z = someFunction(x, y)

// Vector types are bridged automatically.

```
// myHeader.h:
@import simd;

vector_float3 someFunction(vector_float3 x, vector_float3 y);

// myCode.swift:
import simd

let x = float3(1,2,3)
let y = float3(0,0,1)

// Vector types are bridged automatically.
let z = someFunction(x, y)
```

Interoperation between languages

Interoperation between languages

Swift matrix types are layout-compatible with C matrix types

```
import simd

// Use initializer to convert C matrix to Swift matrix.
let mat = float4x4(CFunctionReturningMatrix())

// Use cmatrix property to convert Swift matrix to C matrix.
let result = CFunctionConsumingMatrix(mat.cmatrix)
```

```
import simd
```

```
// Use initializer to convert C matrix to Swift matrix.
let mat = float4x4(CFunctionReturningMatrix())

// Use cmatrix property to convert Swift matrix to C matrix.
```

let result = CFunctionConsumingMatrix(mat.cmatrix)

```
import simd

// Use initializer to convert C matrix to Swift matrix.
let mat = float4x4(CFunctionReturningMatrix())

// Use cmatrix property to convert Swift matrix to C matrix.
let result = CFunctionConsumingMatrix(mat.cmatrix)
```

# New Geometry Functions



```
simd_orient(x, y, ...)
simd_incircle(x, a, b, c)
simd_insphere(x, a, b, c, d)
```

Is a set of vectors positively oriented?

Is a set of vectors positively oriented?

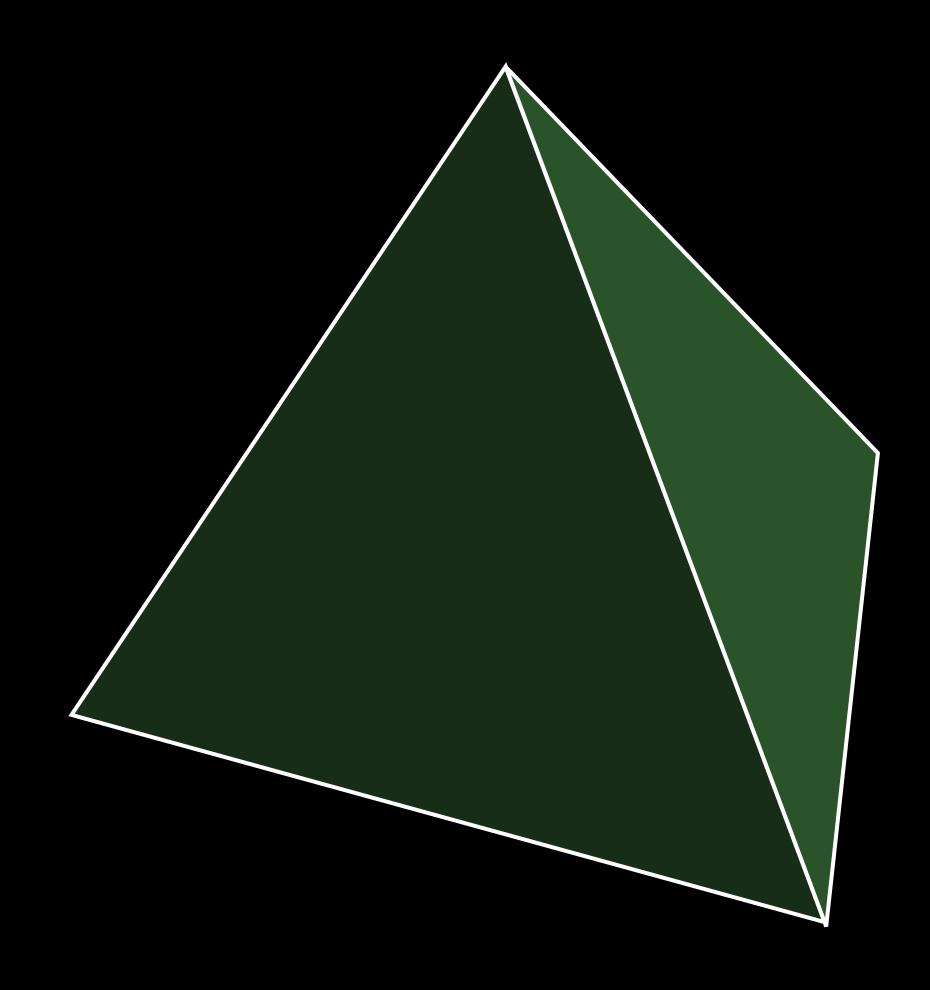
• Do they obey the right hand rule?

Is a set of vectors positively oriented?

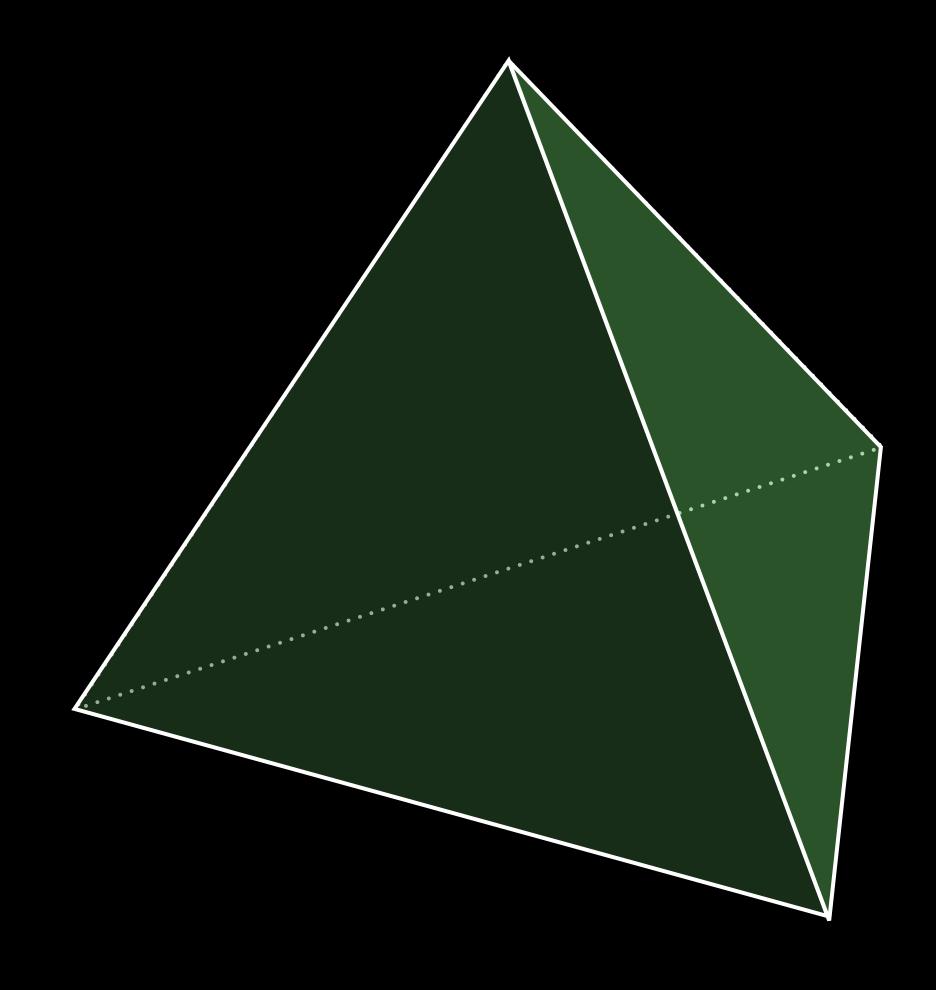
- Do they obey the right hand rule?
- Is their determinant positive?

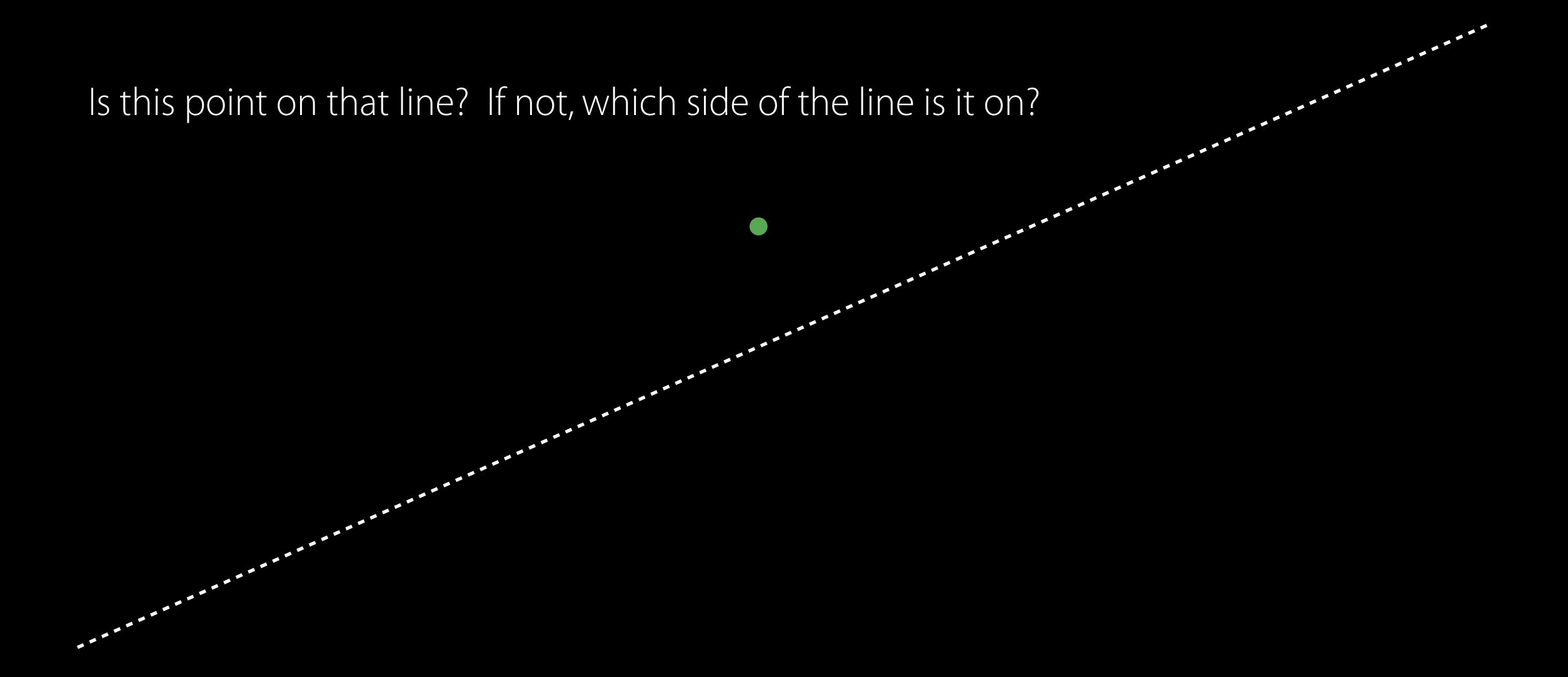
Is a triangle facing toward me or away from me?

Is a triangle facing toward me or away from me?



Is a triangle facing toward me or away from me?

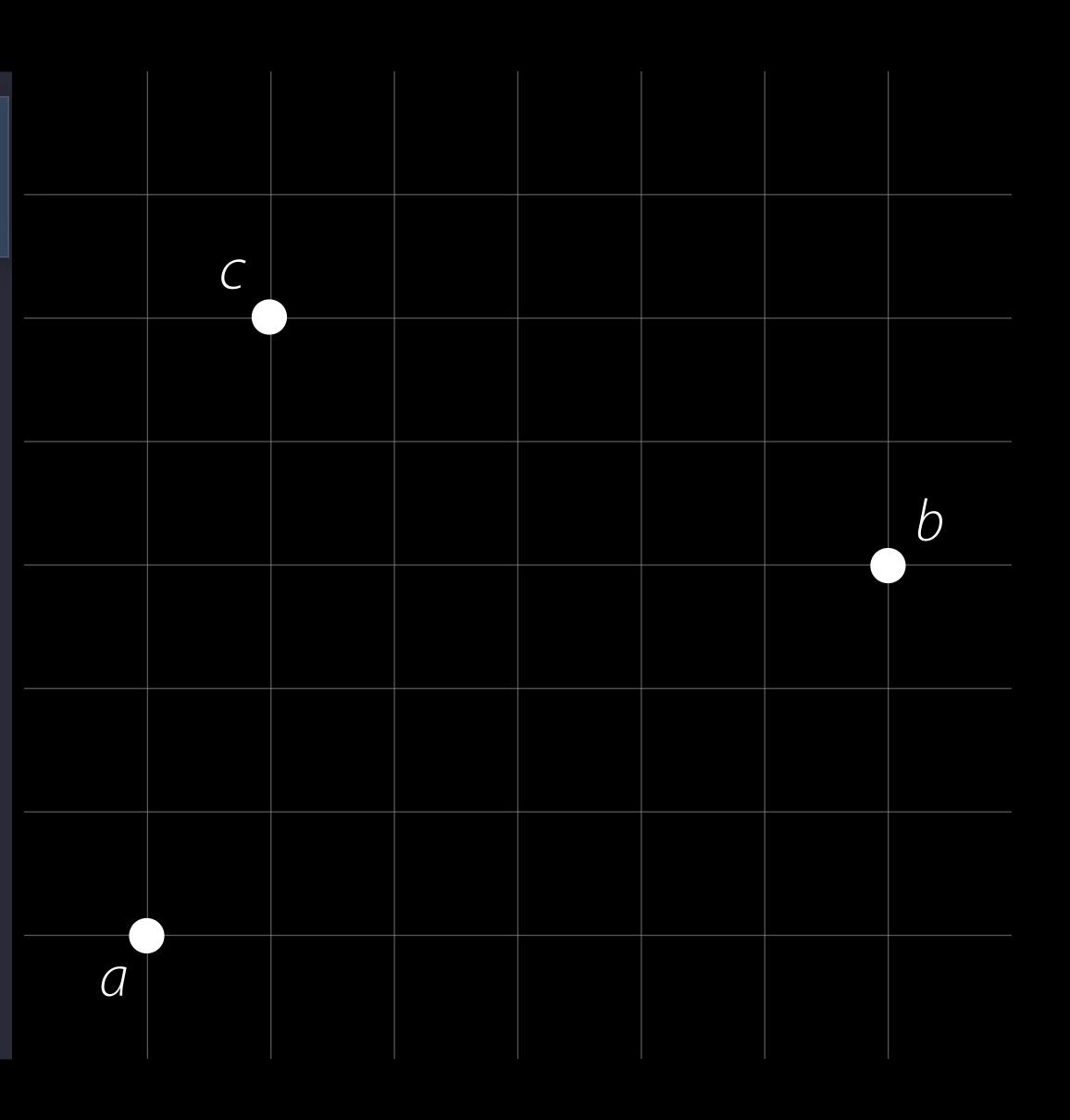




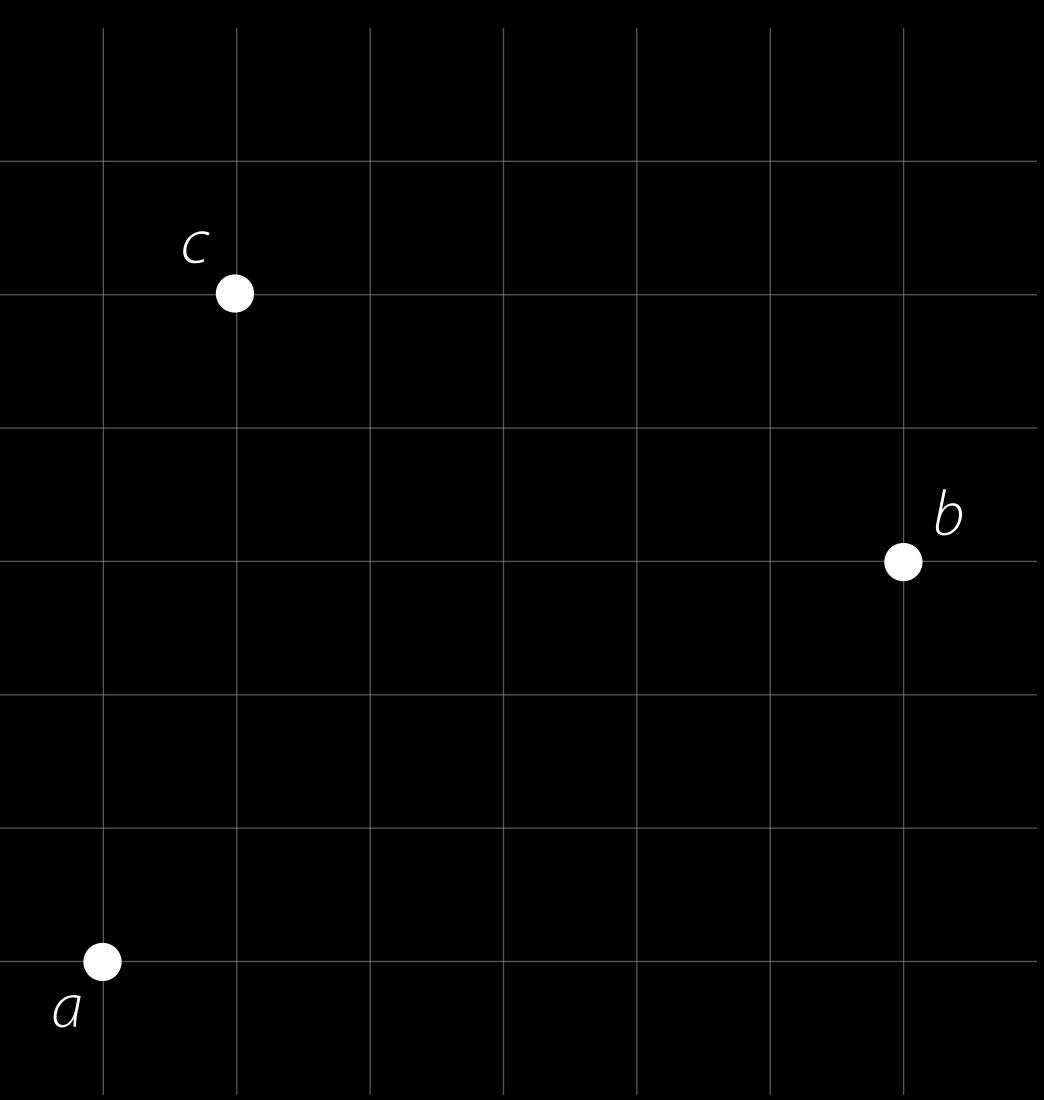
Is this point on that line? If not, which side of the line is it on? 

```
let a = float2(0,0)
let b = float2(6,3)
let c = float2(1,5)
let orientation = simd_orient(a, b, c)
if orientation > 0 {
    print("(a,b,c) is positively oriented.")
else if orientation < 0 {
    print("(a,b,c) is negatively oriented.")
else /* orientation is zero */ {
    print("(a,b,c) are collinear.")
```

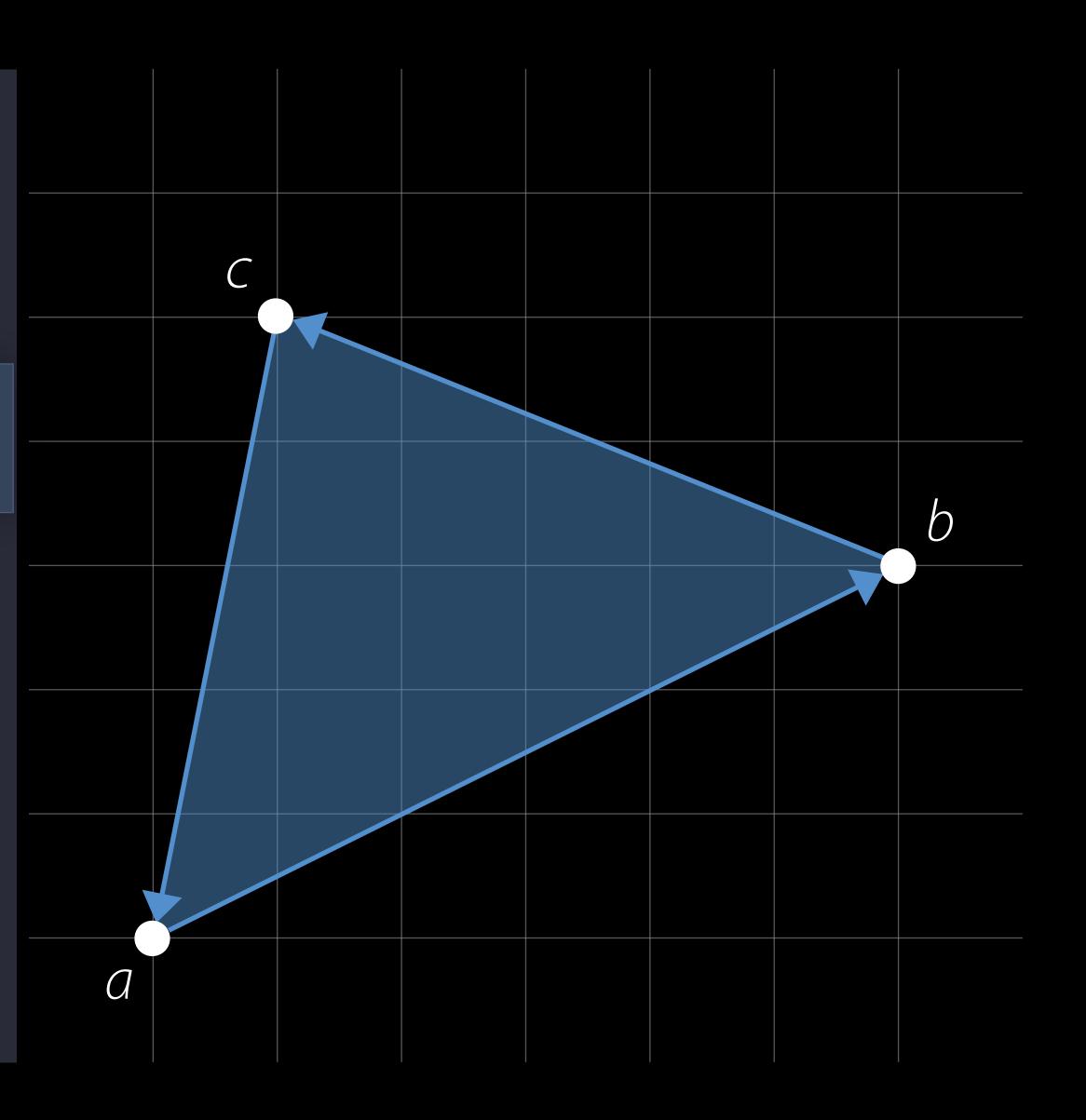
```
let a = float2(0,0)
let b = float2(6,3)
let c = float2(1,5)
let orientation = simd_orient(a, b, c)
if orientation > 0 {
    print("(a,b,c) is positively oriented.")
else if orientation < 0 {</pre>
    print("(a,b,c) is negatively oriented.")
else /* orientation is zero */ {
    print("(a,b,c) are collinear.")
```



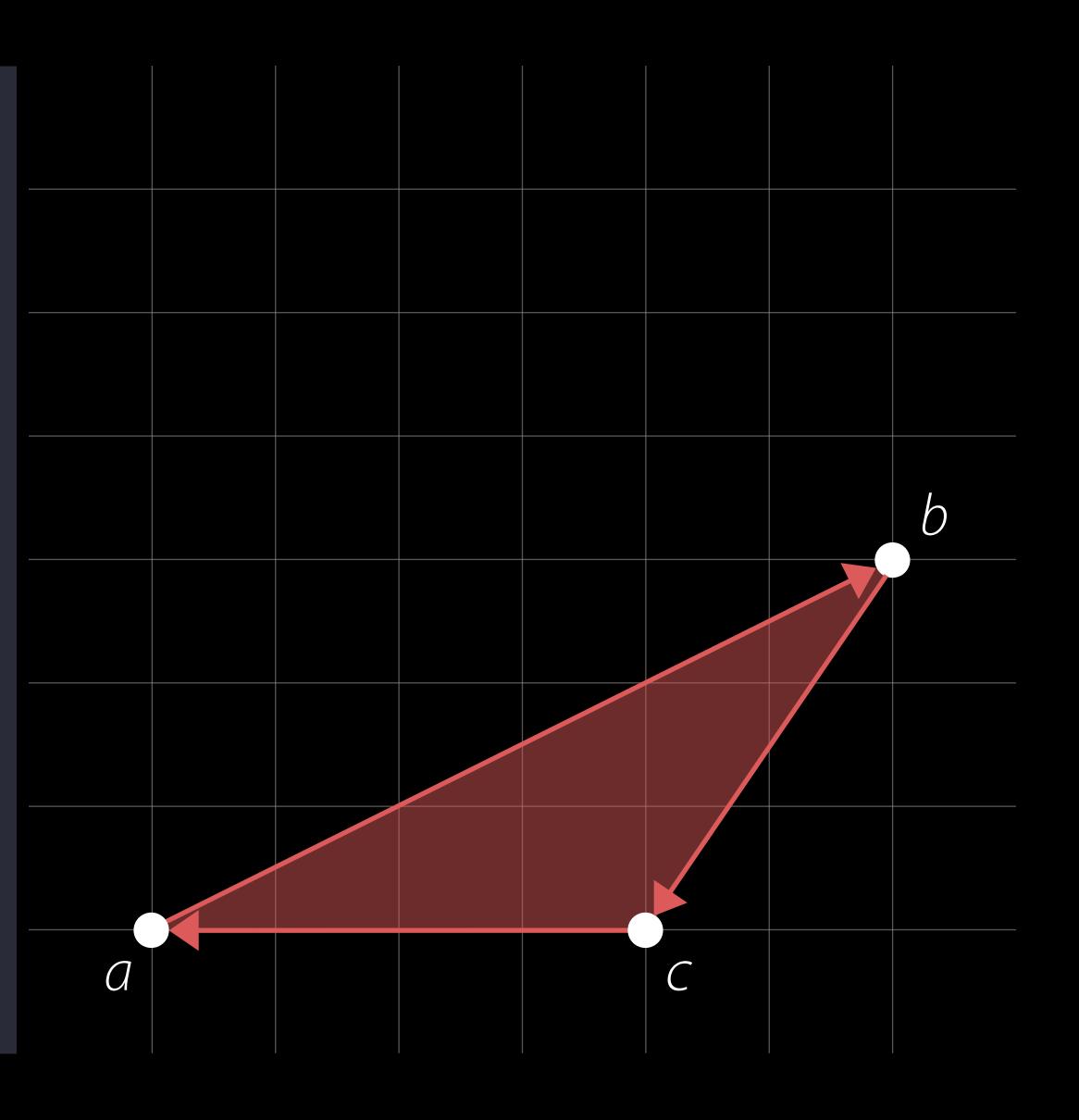
```
let a = float2(0,0)
let b = float2(6,3)
let c = float2(1,5)
let orientation = simd_orient(a, b, c)
if orientation > 0 {
    print("(a,b,c) is positively oriented.")
else if orientation < 0 {</pre>
    print("(a,b,c) is negatively oriented.")
else /* orientation is zero */ {
    print("(a,b,c) are collinear.")
```



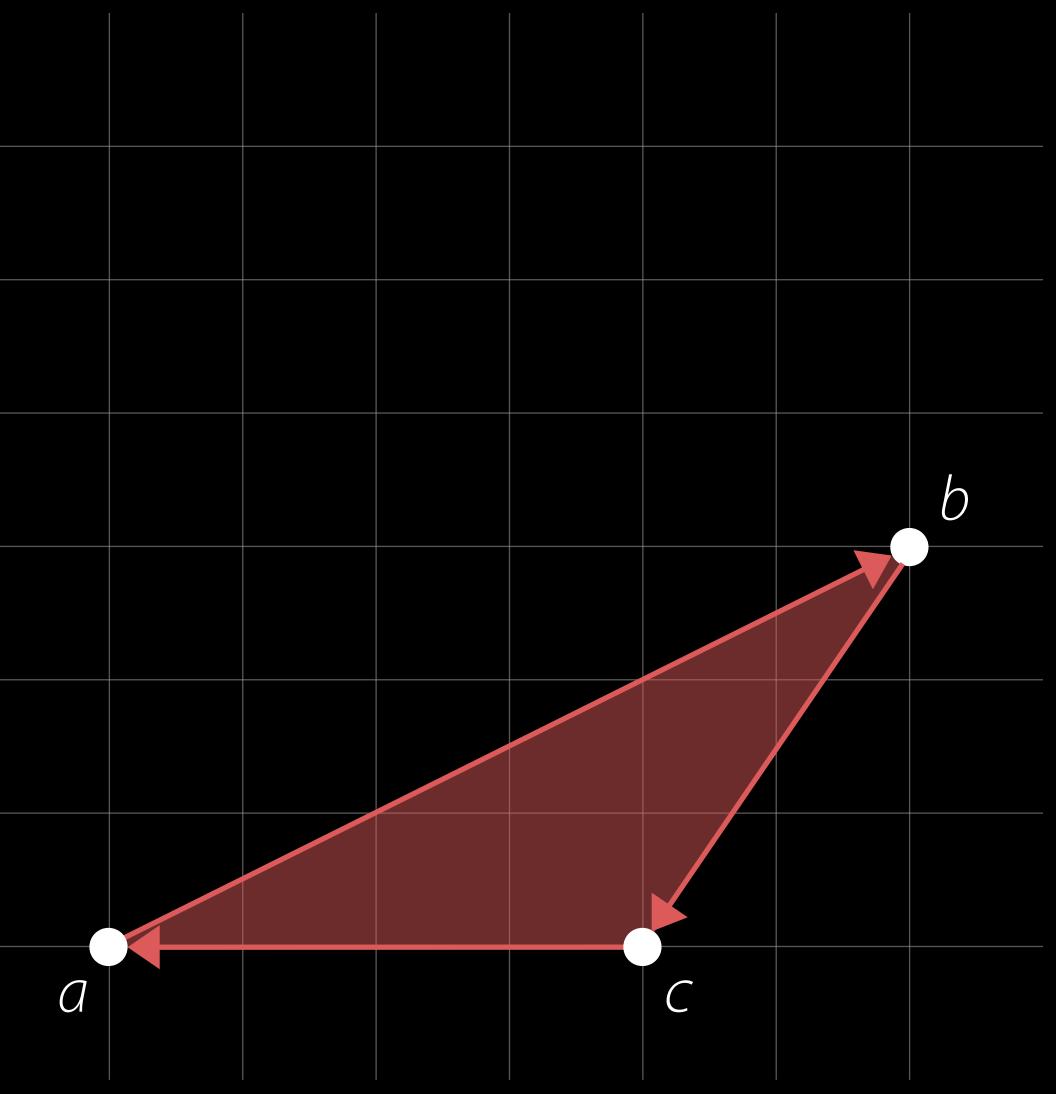
```
let a = float2(0,0)
let b = float2(6,3)
let c = float2(1,5)
let orientation = simd_orient(a, b, c)
if orientation > 0 {
    print("(a,b,c) is positively oriented.")
else if orientation < 0 {</pre>
    print("(a,b,c) is negatively oriented.")
else /* orientation is zero */ {
    print("(a,b,c) are collinear.")
```



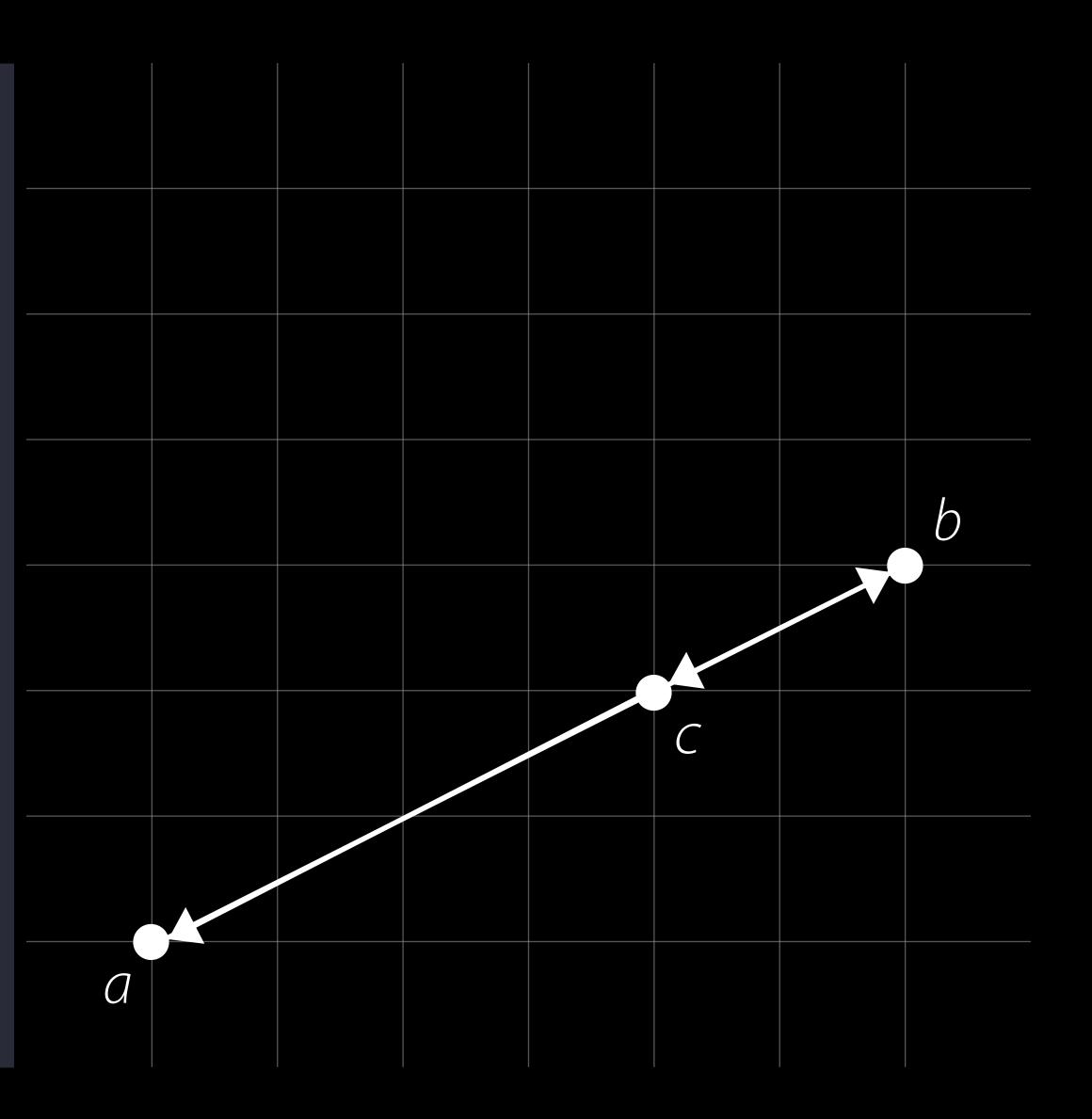
```
let a = float2(0,0)
let b = float2(6,3)
let c = float2(4,0)
let orientation = simd_orient(a, b, c)
if orientation > 0 {
    print("(a,b,c) is positively oriented.")
else if orientation < 0 {</pre>
    print("(a,b,c) is negatively oriented.")
else /* orientation is zero */ {
    print("(a,b,c) are collinear.")
```



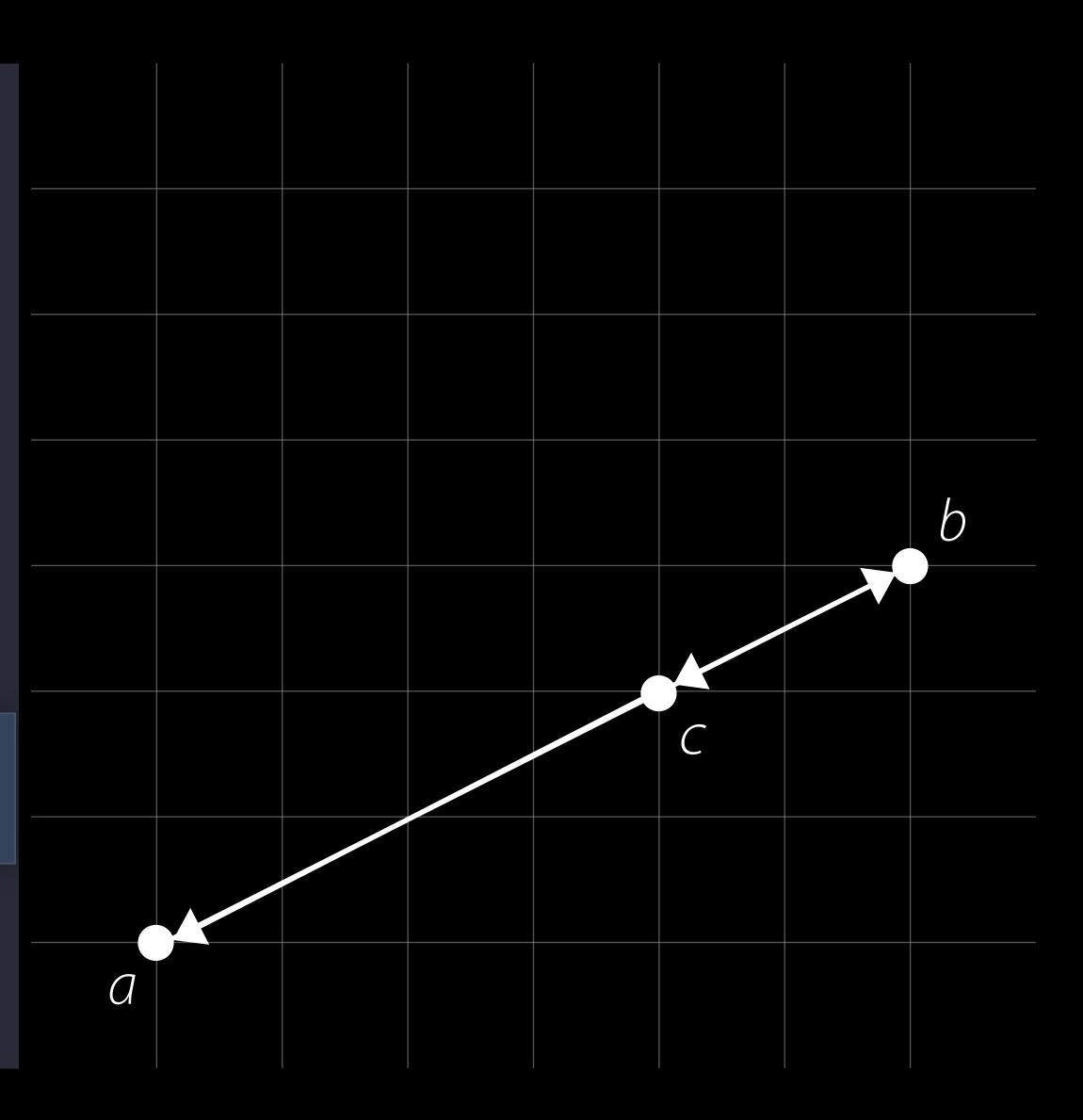
```
let a = float2(0,0)
let b = float2(6,3)
let c = float2(4,0)
let orientation = simd_orient(a, b, c)
if orientation > 0 {
    print("(a,b,c) is positively oriented.")
else if orientation < 0 {</pre>
    print("(a,b,c) is negatively oriented.")
else /* orientation is zero */ {
    print("(a,b,c) are collinear.")
```



```
let a = float2(0,0)
let b = float2(6,3)
let c = float2(4,2)
let orientation = simd_orient(a, b, c)
if orientation > 0 {
    print("(a,b,c) is positively oriented.")
else if orientation < 0 {</pre>
    print("(a,b,c) is negatively oriented.")
else /* orientation is zero */ {
    print("(a,b,c) are collinear.")
```



```
let a = float2(0,0)
let b = float2(6,3)
let c = float2(4,2)
let orientation = simd_orient(a, b, c)
if orientation > 0 {
    print("(a,b,c) is positively oriented.")
else if orientation < 0 {</pre>
    print("(a,b,c) is negatively oriented.")
else /* orientation is zero */ {
    print("(a,b,c) are collinear.")
```

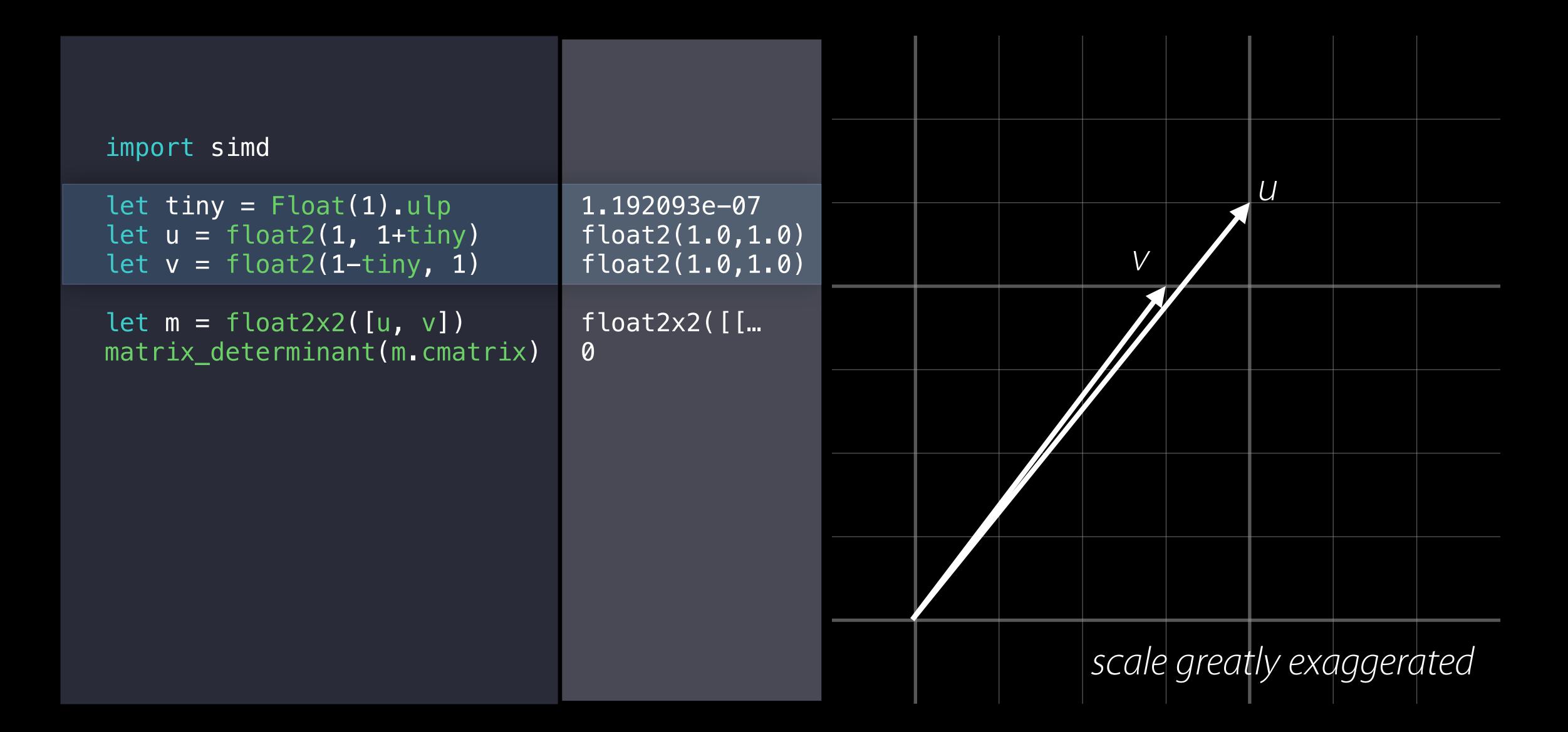


Orientation is numerically unstable

Orientation is numerically unstable

When points are nearly collinear, usual algorithms produce garbage results

```
import simd
let tiny = Float(1).ulp
                                1.192093e-07
let u = float2(1, 1+tiny)
                                 float2(1.0,1.0)
let v = float2(1-tiny, 1)
                                 float2(1.0,1.0)
let m = float2x2([u, v])
                                float2x2([[...
matrix_determinant(m.cmatrix)
                                                                   scale greatly exaggerated
```



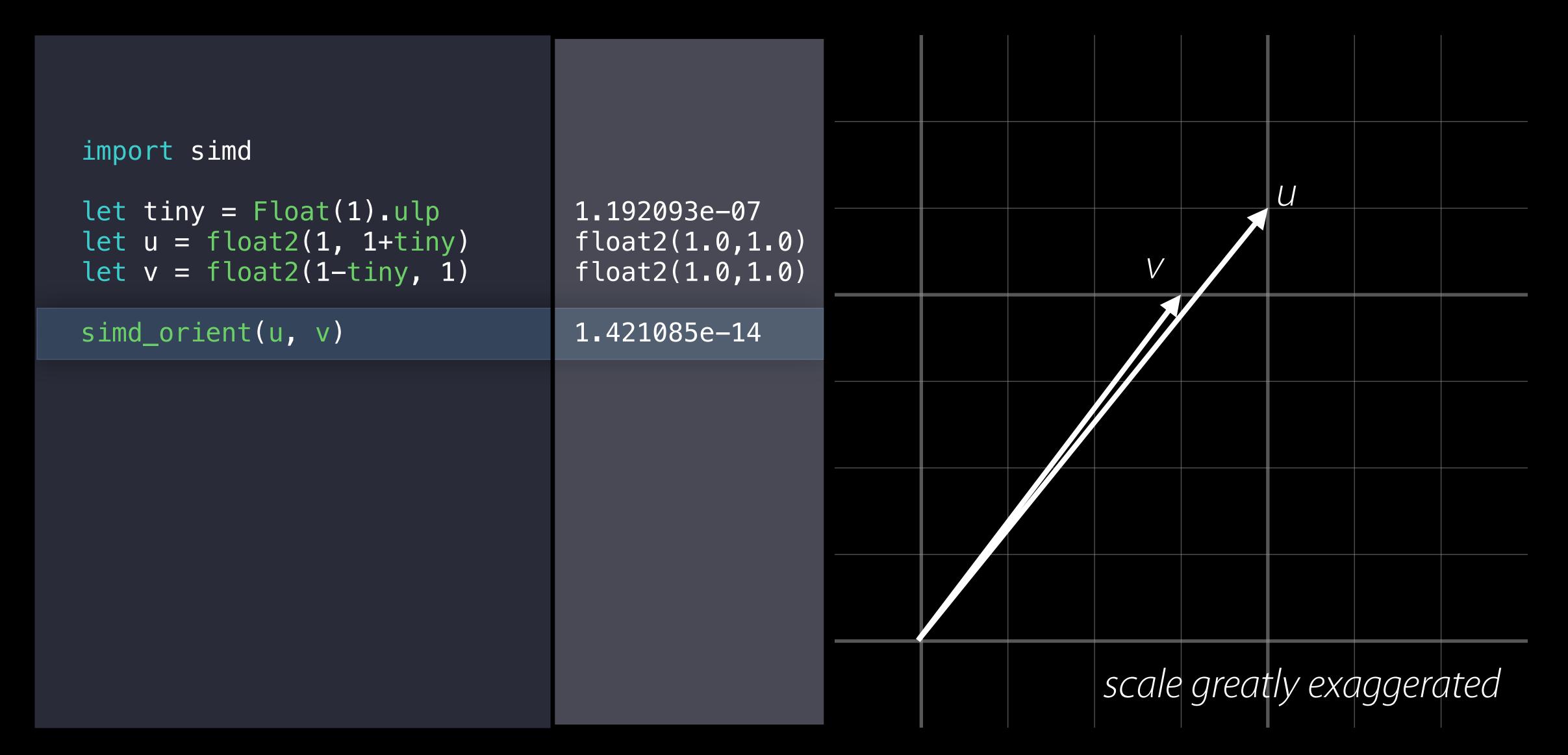


```
import simd
let tiny = Float(1).ulp
                                1.192093e-07
let u = float2(1, 1+tiny)
                                 float2(1.0,1.0)
let v = float2(1-tiny, 1)
                                float2(1.0,1.0)
let m = float2x2([u, v])
                                float2x2([[...
matrix_determinant(m.cmatrix)
                                                                   scale greatly exaggerated
```



```
import simd
let tiny = Float(1).ulp
                                1.192093e-07
let u = float2(1, 1+tiny)
                                float2(1.0,1.0)
let v = float2(1-tiny, 1)
                                float2(1.0,1.0)
simd_orient(u, v)
                                1.421085e-14
                                                                  scale greatly exaggerated
```



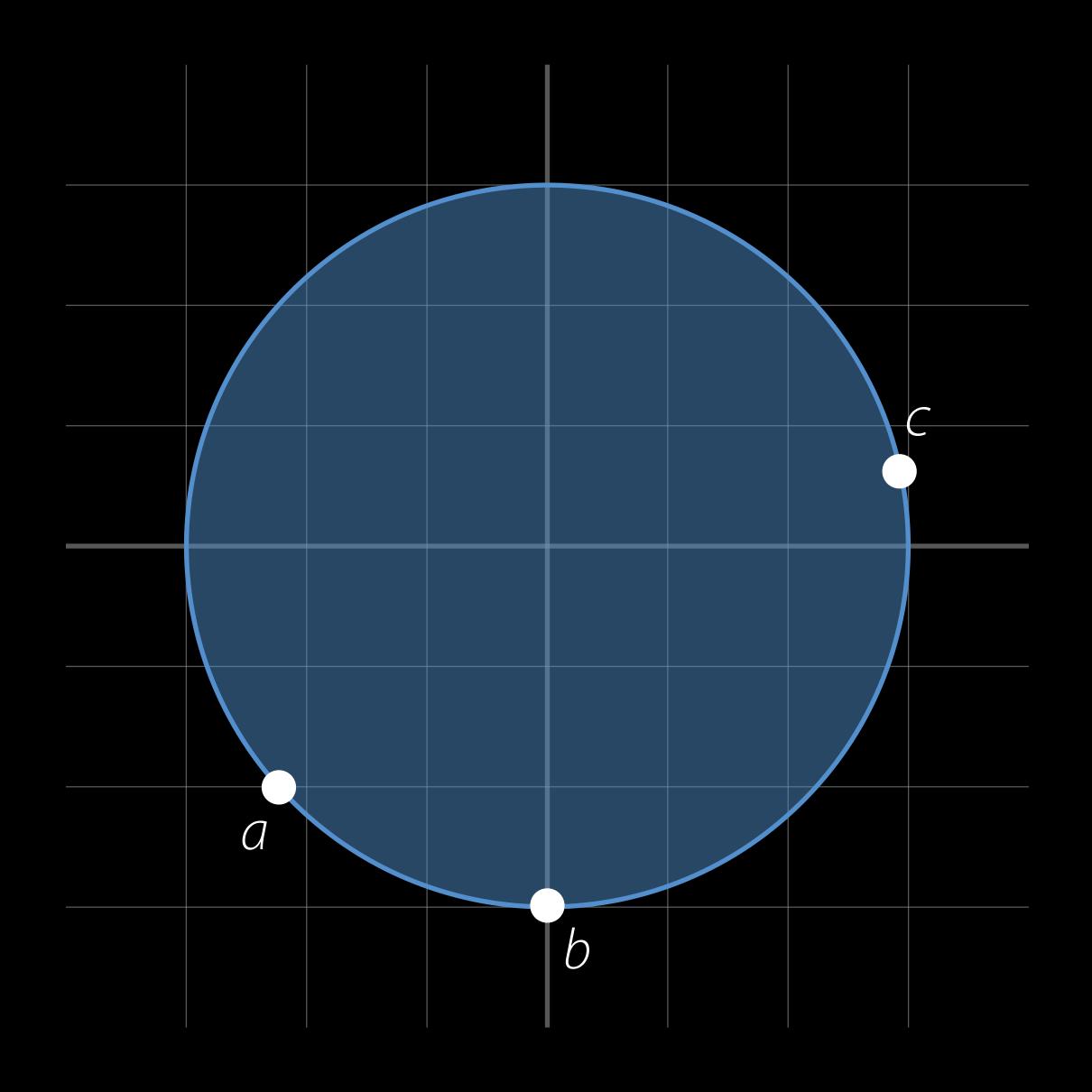


These geometric predicates use adaptive precision

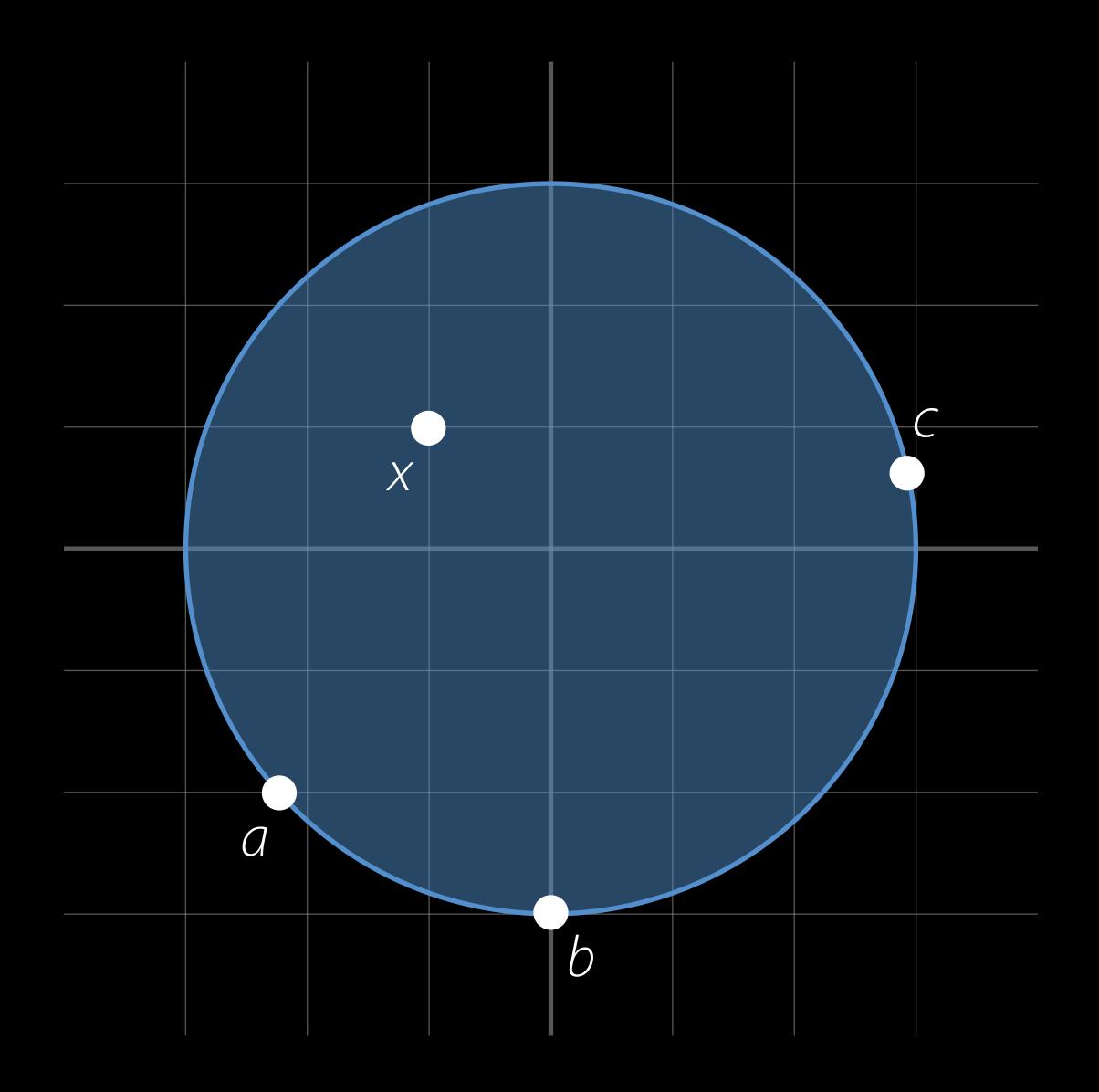
These geometric predicates use adaptive precision

Computation uses as many bits as needed to produce the correct result

Three points (a, b, c) determine a circle

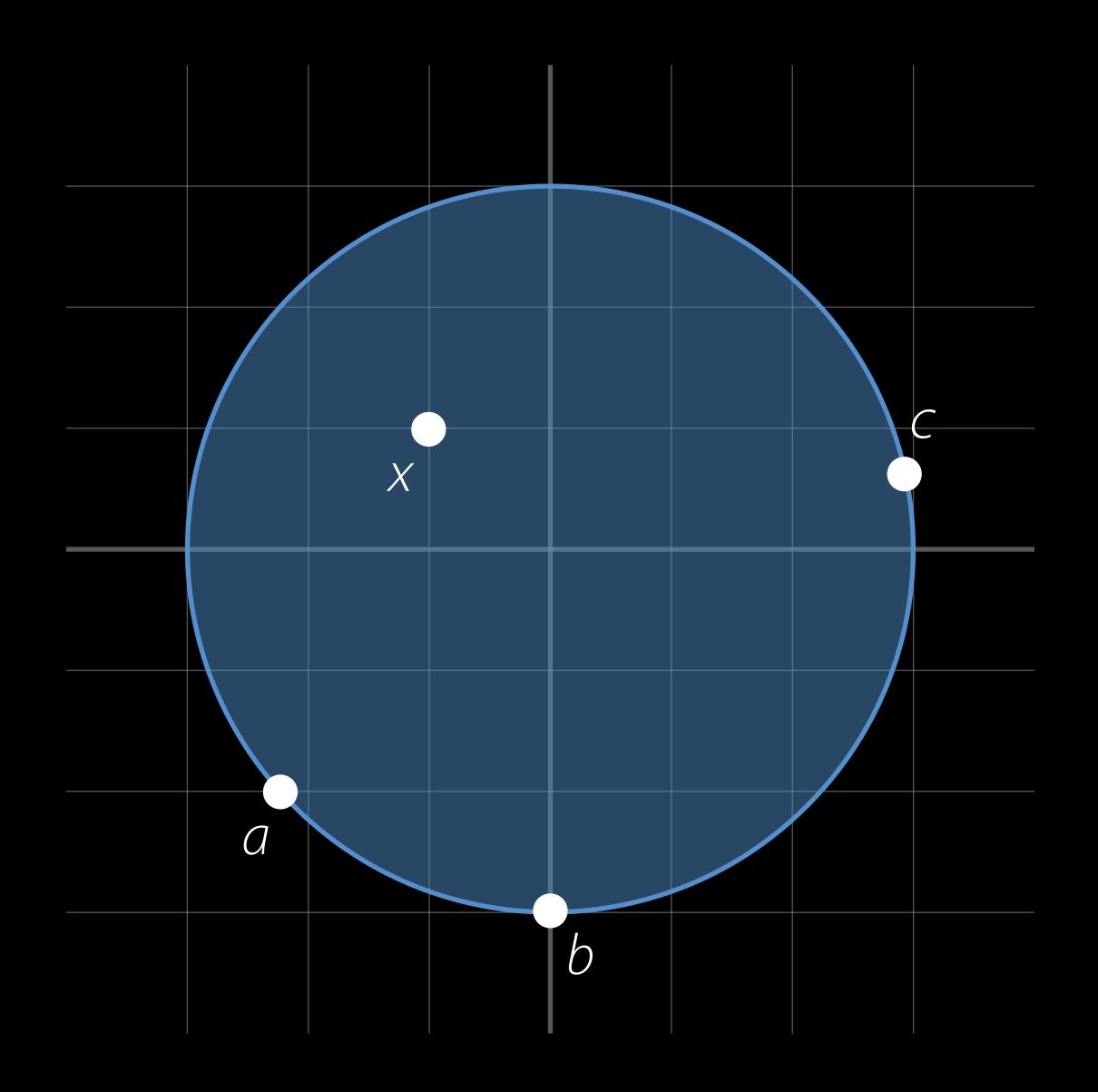


simd\_incircle(x, a, b, c)



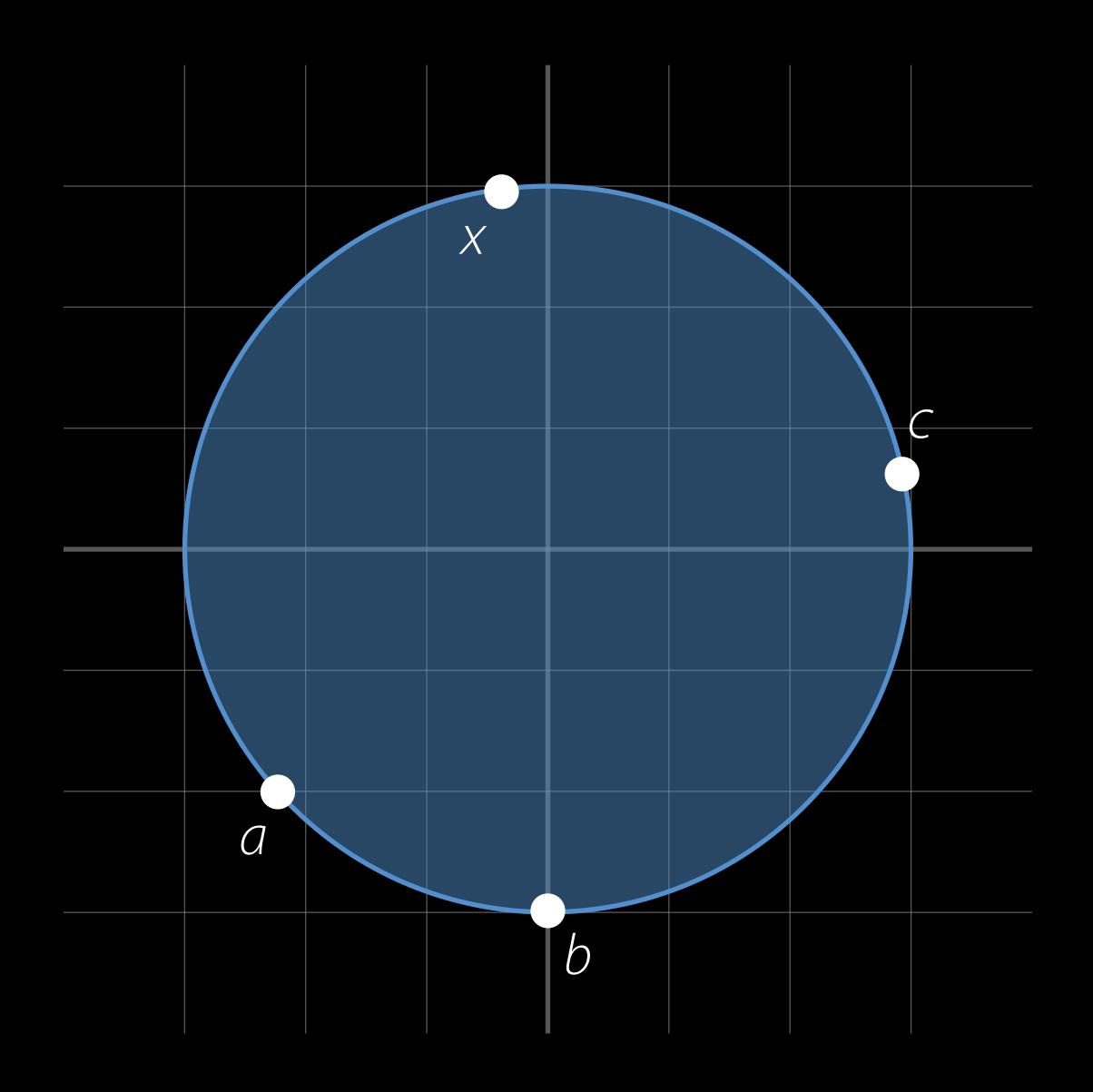
simd\_incircle(x, a, b, c)

Positive if x is inside the circle



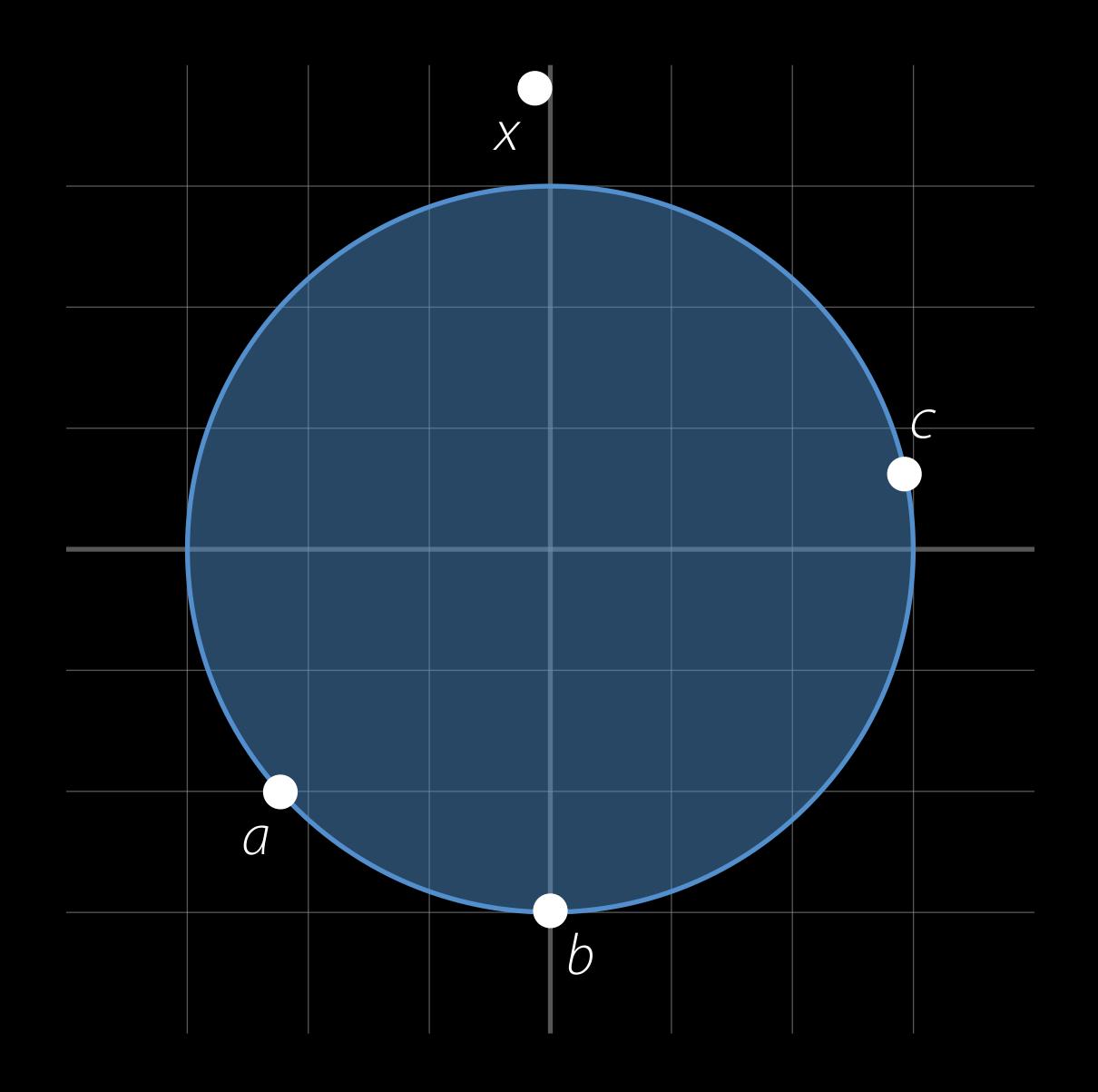
simd\_incircle(x, a, b, c)

- Positive if x is inside the circle
- Zero if x is on the circle



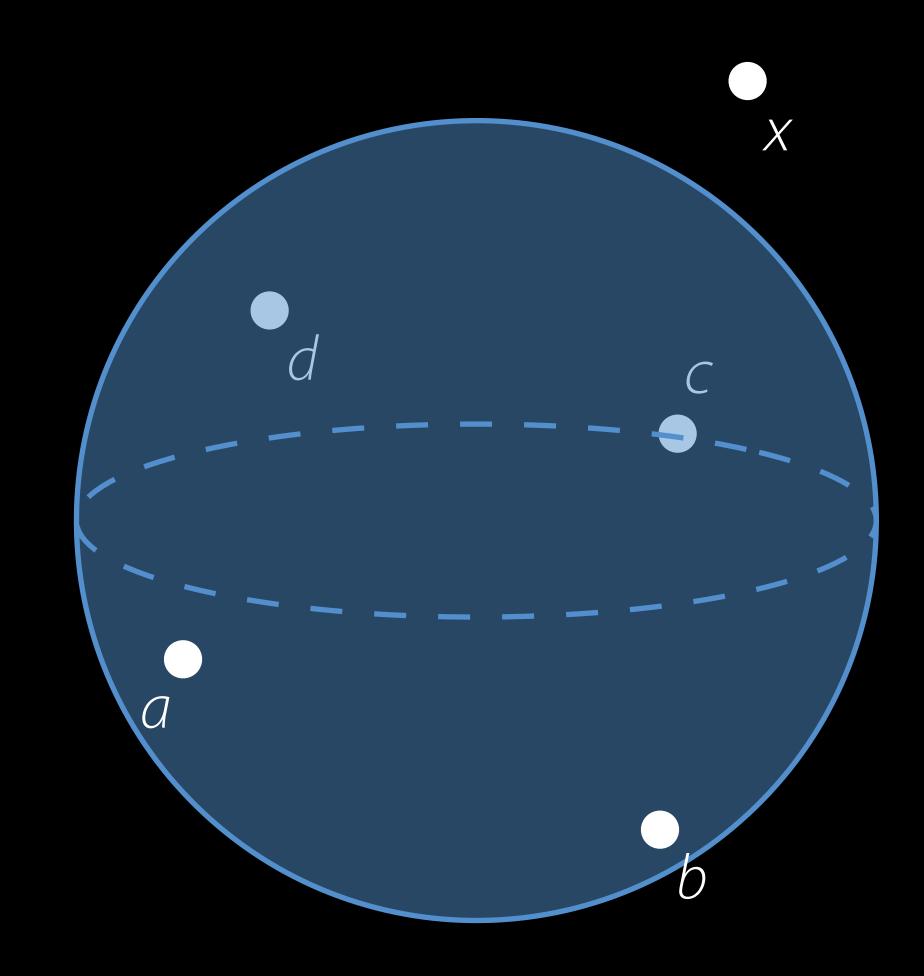
simd\_incircle(x, a, b, c)

- Positive if x is inside the circle
- Zero if x is on the circle
- Negative if x is outside the circle



# insphere

simd\_insphere(x, a, b, c, d) is the same thing in three dimensions



```
import simd
/// Simple struct representing a triangle in 3 dimensions.
struct Triangle {
    var vertices: (float3, float3)
    /// True iff `self` faces towards `camera`.
    func isFacing(camera: float3) -> Bool {
        // Vector normal to front face of triangle.
        let normal = cross(vertices.0 - vertices.2, vertices.1 - vertices.2)
        // Vector from triangle to camera.
        let toCamera = camera - vertices.2
        // If dot product is positive, the triangle faces the camera.
        return dot(toCamera, normal) > 0
```

```
import simd

/// Simple struct representing a triangle in 3 dimensions.
struct Triangle {

    var vertices: (float3, float3, float3)

    // True iff `self` faces towards `camera`.
    func isFacing(camera: float3) -> Bool {

         // Vector normal to front face of triangle.
         let normal = cross(vertices.0 - vertices.2, vertices.1 - vertices.2)

         // Vector from triangle to camera.
         let toCamera = camera - vertices.2

         // If dot product is positive, the triangle faces the camera.
         return dot(toCamera, normal) > 0
    }
}
```

```
import simd

/// Simple struct representing a triangle in 3 dimensions.
struct Triangle {

   var vertices: (float3, float3, float3)

/// True iff `self` faces towards `camera`.
func isFacing(camera: float3) -> Bool {

        // Vector normal to front face of triangle.
        let normal = cross(vertices.0 - vertices.2, vertices.1 - vertices.2)

        // Vector from triangle to camera.
        let toCamera = camera - vertices.2

        // If dot product is positive, the triangle faces the camera.
        return dot(toCamera, normal) > 0
}
```

```
import simd

/// Simple struct representing a triangle in 3 dimensions.
struct Triangle {

    var vertices: (float3, float3, float3)

    /// True iff `self` faces towards `camera`.
    func isFacing(camera: float3) -> Bool {
        return simd_orient(camera, vertices.0, vertices.1, vertices.2) > 0
    }
}
```

```
import simd

/// Simple struct representing a triangle in 3 dimensions.
struct Triangle {

    var vertices: (float3, float3)

    /// True iff `self` faces towards `camera`.
    func isFacing(camera: float3) -> Bool {
        return simd_orient(camera, vertices.0, vertices.1, vertices.2) > 0
    }
}
```

```
import simd

/// Simple struct representing a triangle in 3 dimensions.
struct Triangle {

   var vertices: (float3, float3, float3)

/// True iff `self` faces towards `camera`.
   func isFacing(camera: float3) -> Bool {
       return simd_orient(camera, vertices.0, vertices.1, vertices.2) > 0
   }
}
```

New libraries

#### New libraries

BNNS

#### New libraries

- BNNS
- Quadrature

#### New libraries

- BNNS
- Quadrature

New features

#### New libraries

- BNNS
- Quadrature

#### New features

Orientation and Incircle

#### New libraries

- BNNS
- Quadrature

New features

Orientation and Incircle

All added in response to feature requests!

vlmage geometry operations for interleaved chroma planes

vlmage geometry operations for interleaved chroma planes

Expanded supported formats for vlmage conversion

vlmage geometry operations for interleaved chroma planes Expanded supported formats for vlmage conversion

Improved performance for interleaved complex formats in vDSP

vlmage geometry operations for interleaved chroma planes
Expanded supported formats for vlmage conversion
Improved performance for interleaved complex formats in vDSP
Improved performance of level 2 BLAS operations

vlmage geometry operations for interleaved chroma planes
Expanded supported formats for vlmage conversion
Improved performance for interleaved complex formats in vDSP
Improved performance of level 2 BLAS operations

• • •

Single-stop shopping for computational operations

Single-stop shopping for computational operations

Correct

Single-stop shopping for computational operations

- Correct
- Fast

Single-stop shopping for computational operations

- Correct
- Fast
- Energy efficient

Single-stop shopping for computational operations

- Correct
- Fast
- Energy efficient

Keep the feature requests coming!

#### More Information

https://developer.apple.com/wwdc16/715

# Related Sessions

What's New in Metal, Part 2	Pacific Heights	Wednesday 1:40PM
Advanced Metal Shader Optimization	Nob Hill	Wednesday 3:00PM
Increase Usage of Your App with Proactive Suggestions	Mission	Friday 1:40PM

# Labs

Accelerate Lab	Graphics, Games, and Media Lab D	Thursday 12:00PM
Accelerate Lab	Fort Mason	Thursday 5:00PM

# ÓWWDC16