**AABB**

* Axis aligned bounding box

**Axis-Aligned Bounding Box (AABB) Explained**

An AABB (Axis-Aligned Bounding Box) is a fundamental concept in computer graphics and game development used to approximate the spatial extent of objects. It's called "axis-aligned" because the box's edges align with the coordinate axes (X, Y, and Z in 3D space), making it straightforward to work with. Below, we'll explain AABB with illustrations.

**1. 3D AABB Bounding Box**

In 3D space, an AABB is defined by two points: the minimum point (Min) and the maximum point (Max). These points represent the corners of the box, and the edges are always parallel to the coordinate axes (X, Y, and Z). The AABB encompasses the entire object, regardless of its orientation.

**2. 2D AABB Bounding Box**

In 2D space, the concept is the same. An AABB is defined by two points: the minimum point (Min) and the maximum point (Max), with edges aligned to the X and Y axes.

**Use Cases for AABB:**

* **Collision Detection**: AABBs are used for efficient collision detection between objects.
* **Bounding Volume Hierarchy**: Complex scenes use hierarchical bounding volumes like AABBs to optimize intersection tests.
* **Frustum Culling**: In computer graphics, AABBs are used to quickly remove objects outside the view frustum.

**3. AABB Representation in Code**

In code, an AABB can be represented as a structure or class containing two points: Min and Max. For example, in C++:

cppCopy code

struct AABB { glm::vec3 Min; glm::vec3 Max; };

**4. AABB vs. OBB (Oriented Bounding Box)**

While AABBs are aligned with coordinate axes, OBBs can be oriented arbitrarily in 3D space. OBBs provide a tighter fit around objects but are more complex to work with.

In summary, an AABB bounding box is a simple geometric shape used in computer graphics and game development for collision detection, optimization, and other purposes. Its key feature is its alignment with coordinate axes, making it easy to use in many scenarios.

* <https://developer.mozilla.org/en-US/docs/Games/Techniques/3D_collision_detection>

**Compute buffer**

* Of course! Think of compute shaders as tiny worker programs for your computer's graphics card (GPU). They're not focused on drawing images but are more like calculators that perform specific tasks really fast.
* Imagine you have a huge pile of math problems that need solving, and you want to do it quickly. Instead of using your CPU (the brain of your computer), which is good at handling general tasks, you ask your GPU (the graphics card) to solve these problems using a compute shader.
* So, a compute shader is a small program that runs on the GPU, taking input data, crunching numbers, and producing results. It's great for tasks like simulations, data processing, or any heavy calculations that benefit from parallel processing power.
* In a nutshell, compute shaders help offload specific tasks to your GPU, making your computer more efficient at handling certain types of work, especially when speed and parallel processing are crucial.
* Certainly, let's focus on work groups and why there are three inputs:
* Imagine you have a big task, like sorting a huge pile of books. To make this task faster, you decide to enlist the help of your friends. But, you don't want everyone to work on the same pile of books; instead, you want to divide the work to be more efficient.
* 1. \*\*Dividing the Task\*\*: You decide to divide the books into smaller piles, and each of these smaller piles represents a "work group." Now, instead of one big group sorting all the books, you have several smaller groups, and each group will sort its own pile.
* 2. \*\*Three Important Inputs\*\*: The reason you have three inputs (like gl\_NumWorkGroups, gl\_WorkGroupSize, and gl\_WorkGroupID) is to make sure this division is well-organized and efficient.
* - \*\*gl\_NumWorkGroups\*\*: This tells you how many of these smaller groups (work groups) you've created. So, you know exactly how many piles of books there are in total.
* - \*\*gl\_WorkGroupSize\*\*: This is like knowing how many friends you've assigned to each group. It helps you determine how many people are sorting each pile of books.
* - \*\*gl\_WorkGroupID\*\*: To keep things organized, you give each group a unique number or name (like Group 1, Group 2, etc.). This helps you keep track of which group is sorting which pile of books.

So, these three inputs help you efficiently divide a big task into smaller, manageable parts (work groups), making it easier to complete the task faster with the help of your friends.

* I apologize if my previous explanations were not clear. Let's try to simplify it further:
* Think of your computer's work as a big puzzle that needs to be solved quickly. The puzzle is so massive that it's impossible for one person to solve it in a reasonable time.
* 1. \*\*Work Groups (Dividing the Puzzle):\*\* To solve this puzzle efficiently, you break it into smaller sections. These sections are like mini-puzzles, and you assign each section to a group of people. These groups are called "work groups."
* 2. \*\*Why Three Inputs:\*\* Now, imagine you need a way to keep track of these groups and the people within them.
* - \*\*`gl\_NumWorkGroups`\*\*: This tells you how many of these mini-puzzles (work groups) you have in total. It's like counting how many sections you divided the big puzzle into.
* - \*\*`gl\_WorkGroupSize`\*\*: This tells you how many people are in each group working on their mini-puzzle. It's like knowing how many individuals are in each group.
* - \*\*`gl\_WorkGroupID`\*\*: To make sure each group works on a different part of the big puzzle, you label or number each group. `gl\_WorkGroupID` helps you keep track of which group is working on which mini-puzzle.

The goal is to break down a complex task into smaller, more manageable parts (work groups) and use these three inputs to organize and coordinate the work efficiently. It's like dividing a giant jigsaw puzzle into smaller pieces, having multiple groups of friends work on it, and keeping track of which group is working on which section of the puzzle.

There is a limitation of work groups that can be dispatched in a single compute shader dispatch call. This limit is defined by GL\_MAX\_COMPUTE\_WORK\_GROUP\_COUNT, which must/can be queried using the function glGetIntegeri\_v where the indices *0*, *1* and *2* corresponds to the *X*, *Y* and *Z* dimensions, respectively.

There is as well a limitation on the local size which can be queried with GL\_MAX\_COMPUTE\_WORK\_GROUP\_SIZE and another limitation of the total number of invocations within a work group, which is that the product of the X, Y and Z components of the local size must be less than GL\_MAX\_COMPUTE\_WORK\_GROUP\_INVOCATIONS.

As we define and divide the tasks and the compute shader groups sizes ourselves, we have to keep these limitations in mind.

We will bind the a 2d image in our shader as the object to write our data onto. The internal format (here rgba32f) needs to be the same as the format of the texture in the host program.

**layout**(rgba32f, binding = 0) **uniform** image2D imgOutput;

We have to use image2d as this represents a single image from a texture. While sampler variables use the entire texture including mipmap levels and array layers, images only have a single image from a texture. *Note* while most texture sampling functions use normalized texture coordinates [0,1], for images we need the absolute integer texel coordinates. Images and samplers are completely separated including their bindings. While samplers can only read data from textures, image variables can read and/or write data.

With this set up, we can now write our main function in the shader where we fill the imgOutput with color values. To determine on which pixel we are currently operating in our shader execution we can use the following GLSL Built-in variables shown in the table below:

|  |  |  |
| --- | --- | --- |
| **Type** | **Built-in name** |  |
| uvec3 | gl\_NumWorkGroups | number of work groups that have been dispatched set by glDispatchCompute() |
| uvec3 | gl\_WorkGroupSize | size of the work group (local size) operated on defined with layout |
| uvec3 | gl\_WorkGroupID | index of the work group currently being operated on |
| uvec3 | gl\_LocalInvocationID | index of the current work item in the work group |
| uvec3 | gl\_GlobalInvocationID | global index of the current work item  *(gl\_WorkGroupID \* gl\_WorkGroupSize + gl\_LocalInvocationID)* |
| uint | gl\_LocalInvocationIndex | 1d index representation of gl\_LocalInvocationID  *(gl\_LocalInvocationID.z \* gl\_WorkGroupSize.x \* gl\_WorkGroupSize.y + gl\_LocalInvocationID.y \* gl\_WorkGroupSize.x + gl\_LocalInvocationID.x)* |

Determining the optimal number of work groups to dispatch in a compute shader can be a challenging task and typically involves a balance between performance, available resources, and the specific workload you are trying to parallelize. Here are some general guidelines and steps to help you decide how many work groups to dispatch:

1. **Understand Your Problem:** Understand the nature of the computation you are parallelizing. Consider whether it is embarrassingly parallel (tasks are entirely independent) or if there are dependencies that require synchronization.
2. **Profile Your Code:** Use profiling tools to analyze the performance of your compute shader with different dispatch sizes. Profiling will help you identify bottlenecks and performance improvements.
3. **Experiment:** Start with a reasonable estimate for the number of work groups and then experiment with different values. You can gradually increase or decrease the number of work groups to find the best balance between parallelism and overhead.
4. **Hardware Limitations:** Be aware of the limitations of your target hardware. You can query these limitations using OpenGL functions like **glGetIntegeri\_v** for compute shader limits. Understand the maximum number of work groups, work group size, and available memory.
5. **Work Group Size:** Determine an appropriate work group size (the number of threads per work group) based on your algorithm and hardware. Smaller work group sizes might be more efficient if you have a large number of work groups, while larger work group sizes can be more efficient for some computations.
6. **Avoid Oversubscription:** Avoid dispatching too many work groups that oversubscribe your GPU's resources. Oversubscription can lead to performance degradation.
7. **Dependency Considerations:** If your computation has dependencies between threads, you may need to synchronize them using barriers or other synchronization mechanisms. Understand how these dependencies affect the choice of work group sizes and dispatch counts.
8. **Profile on Target Hardware:** Test your compute shader on the specific hardware and platforms you intend to target. Different GPUs may have varying performance characteristics.
9. **Balance Workload:** Balance the workload among work groups to maximize parallelism. Uneven workloads can lead to inefficient utilization of GPU resources.
10. **Iterate:** The optimal dispatch size may change as you optimize other parts of your code. Revisit your dispatch size as you make changes to other parts of your program.
11. **Use Profiling Feedback:** Some graphics APIs and tools provide feedback on GPU performance. Use this feedback to fine-tune your dispatch sizes.

In summary, there is no one-size-fits-all answer for how many work groups to dispatch in a compute shader. It requires experimentation, profiling, and consideration of hardware limitations. Optimizing compute shader dispatch is often an iterative process, and the best dispatch size may vary depending on the specific workload and target hardware.

"Invocations per work group" refers to the number of shader invocations, often represented by threads or threads per work group, that are executed concurrently within a single work group in a compute shader or parallel computing context. It's an important concept in understanding how parallel computation is organized in a GPU or similar parallel processing architecture.

Here's a breakdown:

1. **Shader Invocation:** In the context of a compute shader or similar shader program, a "shader invocation" refers to a single execution of the shader code by a thread. Each thread represents a single shader invocation.
2. **Work Group:** In parallel computing, tasks are often grouped into "work groups." A work group is a collection of threads that can collaborate and share data within a limited scope. Work groups are defined by the programmer and are a way to organize parallel computation.
3. **Invocations per Work Group:** "Invocations per work group" refers to the number of threads (shader invocations) that are executed simultaneously within a single work group. This number is determined by the shader programmer and depends on the specific parallel task being performed.

For example, if you have a compute shader where you define a work group size of 128, and you dispatch 8 work groups, it means that you have 8 work groups \* 128 invocations per work group = 1024 shader invocations running concurrently. Each of these 1024 shader invocations can process a different piece of data or perform a specific computation independently.

The concept of "invocations per work group" is crucial for optimizing parallel algorithms in compute shaders because it determines how many threads are available to perform parallel computations and how they collaborate within the defined work group. Properly balancing the number of invocations per work group and the total number of work groups is essential for efficient parallel processing and GPU utilization.

Certainly! Here's a simplified example in C++ using OpenGL Compute Shaders to demonstrate the concept of "invocations per work group." I'll provide code with comments to explain each part:

cppCopy code

#include <iostream> #include <vector> #include <GL/glew.h> #include <GLFW/glfw3.h> // Define the shader source code for the compute shader. const char\* computeShaderSource = R"( #version 430 core layout (local\_size\_x = 128) in; // 128 shader invocations per work group // This is a simple example that increments each element of an input buffer. // Each thread (invocation) operates on a different element. buffer DataBuffer { int data[]; }; void main() { uint index = gl\_GlobalInvocationID.x; // Get the global ID of the thread // Increment the value at the corresponding index in the buffer data[index]++; } )"; int main() { // Initialize GLFW and create a window (not shown in the code) // Initialize GLEW (not shown in the code) // Create a buffer for data std::vector<int> dataBuffer(1024); // 1024 elements // Create and compile the compute shader GLuint computeShader = glCreateShader(GL\_COMPUTE\_SHADER); glShaderSource(computeShader, 1, &computeShaderSource, NULL); glCompileShader(computeShader); // Create a program and attach the compute shader GLuint computeProgram = glCreateProgram(); glAttachShader(computeProgram, computeShader); glLinkProgram(computeProgram); // Use the compute shader program glUseProgram(computeProgram); // Create a buffer object and bind it GLuint dataBufferObject; glGenBuffers(1, &dataBufferObject); glBindBufferBase(GL\_SHADER\_STORAGE\_BUFFER, 0, dataBufferObject); glBufferData(GL\_SHADER\_STORAGE\_BUFFER, sizeof(int) \* dataBuffer.size(), dataBuffer.data(), GL\_DYNAMIC\_COPY); // Dispatch compute shader work groups glDispatchCompute(8, 1, 1); // 8 work groups, each with 128 shader invocations // Wait for the compute shader to finish glMemoryBarrier(GL\_SHADER\_STORAGE\_BARRIER\_BIT); // Retrieve and print the updated data glGetBufferSubData(GL\_SHADER\_STORAGE\_BUFFER, 0, sizeof(int) \* dataBuffer.size(), dataBuffer.data()); for (int i = 0; i < dataBuffer.size(); ++i) { std::cout << "Data[" << i << "] = " << dataBuffer[i] << std::endl; } // Clean up and terminate GLFW (not shown in the code) return 0; }

In this example:

* We have a compute shader that increments values in a buffer using shader invocations. Each work group consists of 128 shader invocations, as specified by **layout (local\_size\_x = 128)**.
* We dispatch 8 work groups using **glDispatchCompute(8, 1, 1)**, resulting in a total of 8 \* 128 = 1024 shader invocations running concurrently.
* The **glMemoryBarrier** function is used to ensure that the compute shader has finished its work before we retrieve and print the updated data.

This example illustrates the concept of "invocations per work group" and how you can use it to parallelize tasks in a compute shader. Each shader invocation processes a different element in parallel, making efficient use of GPU resources.

// Define the shader source code for the compute shader.

const char\* computeShaderSource = R"(

#version 430 core

layout (local\_size\_x = 128) in; // 128 shader invocations per work group

// This is a simple example that increments each element of an input buffer.

// Each thread (invocation) operates on a different element.

buffer DataBuffer {

int data[];

};

void main() {

uint index = gl\_GlobalInvocationID.x; // Get the global ID of the thread

// Increment the value at the corresponding index in the buffer

data[index]++;

}

)";

int main() {

// Initialize GLFW and create a window (not shown in the code)

// Initialize GLEW (not shown in the code)

// Create a buffer for data

std::vector<int> dataBuffer(1024); // 1024 elements

// Create and compile the compute shader

GLuint computeShader = glCreateShader(GL\_COMPUTE\_SHADER);

glShaderSource(computeShader, 1, &computeShaderSource, NULL);

glCompileShader(computeShader);

// Create a program and attach the compute shader

GLuint computeProgram = glCreateProgram();

glAttachShader(computeProgram, computeShader);

glLinkProgram(computeProgram);

// Use the compute shader program

glUseProgram(computeProgram);

// Create a buffer object and bind it

GLuint dataBufferObject;

glGenBuffers(1, &dataBufferObject);

glBindBufferBase(GL\_SHADER\_STORAGE\_BUFFER, 0, dataBufferObject);

glBufferData(GL\_SHADER\_STORAGE\_BUFFER, sizeof(int) \* dataBuffer.size(), dataBuffer.data(), GL\_DYNAMIC\_COPY);

// Dispatch compute shader work groups

glDispatchCompute(8, 1, 1); // 8 work groups, each with 128 shader invocations

// Wait for the compute shader to finish

glMemoryBarrier(GL\_SHADER\_STORAGE\_BARRIER\_BIT);

// Retrieve and print the updated data

glGetBufferSubData(GL\_SHADER\_STORAGE\_BUFFER, 0, sizeof(int) \* dataBuffer.size(), dataBuffer.data());

for (int i = 0; i < dataBuffer.size(); ++i) {

std::cout << "Data[" << i << "] = " << dataBuffer[i] << std::endl;

}

// Clean up and terminate GLFW (not shown in the code)

return 0;

}

"Invocations" in the context of computer graphics and GPU programming typically refers to the individual executions of a shader or processing unit for a specific task. Each invocation represents a single execution or thread that performs a particular operation.

* Yes, you're correct. The number of times a specific operation or shader code runs is often referred to as the "invocations." It signifies how many individual executions or threads are being used to perform a particular task.
* For example:
* In a shader, the number of "shader invocations" indicates how many times the shader code is executed. In a fragment shader, there is one invocation per pixel or fragment being rendered. In a vertex shader, there is one invocation per vertex in the input geometry.

In a compute shader, the "compute shader invocations" represent how many times the compute shader is executed. Each invocation typically processes a portion of data or performs a specific parallel computation.

Certainly, let's explain the concepts of "work groups" and "invocations per work group" in the context of compute shaders:

1. **Work Groups in Compute Shaders:**
   * **Definition:** In a compute shader, a "work group" is a collection of parallel shader invocations (threads) that are grouped together to perform a specific task or computation. Think of it as a small, self-contained unit of parallel execution.
   * **Purpose:** Work groups are used to organize and coordinate parallel computations in a compute shader. They allow you to group related tasks that can be processed simultaneously and efficiently.
   * **Example:** If you are performing a simulation of particles in a compute shader, each work group might handle the physics calculations for a subset of particles. Work groups enable you to manage and synchronize these calculations effectively.
2. **Invocations per Work Group in Compute Shaders:**
   * **Definition:** "Invocations per work group" refers to the number of shader invocations (threads) that are executed concurrently within a single work group in a compute shader.
   * **Purpose:** It specifies how many shader invocations collaborate within a work group to perform a parallel computation. The number of invocations per work group is determined by the shader programmer and depends on the specific problem and hardware capabilities.
   * **Example:** Suppose you have a compute shader where you specify that each work group should have 128 shader invocations. In this case, 128 threads within that work group will run in parallel to process data or perform a computation.

Now, let's tie these concepts together with a practical example:

**Example: Parallel Image Processing with Compute Shaders**

Let's say you're using a compute shader to process an image:

* You divide the image into work groups, and each work group processes a portion of the image.
* Each work group contains a specific number of shader invocations, for example, 128 invocations per work group.
* Within each work group, the 128 shader invocations cooperate to perform tasks like color manipulation or filtering on their assigned image portion.
* The work groups themselves can be dispatched in parallel, allowing for efficient processing of the entire image.

This way, by defining work groups and specifying the number of invocations per work group, you achieve parallelism in image processing, and the compute shader orchestrates the execution of shader invocations within each work group.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer program

Description automatically generated

A computer screen shot of a code

Description automatically generated

A computer screen shot of a program code

Description automatically generated

SSAO

* We've briefly touched the topic in the basic lighting chapter: ambient lighting. Ambient lighting is a fixed light constant we add to the overall lighting of a scene to simulate the scattering of light. In reality, light scatters in all kinds of directions with varying intensities so the indirectly lit parts of a scene should also have varying intensities. One type of indirect lighting approximation is called ambient occlusion that tries to approximate indirect lighting by darkening creases, holes, and surfaces that are close to each other. These areas are largely occluded by surrounding geometry and thus light rays have fewer places to escape to, hence the areas appear darker. Take a look at the corners and creases of your room to see that the light there seems just a little darker. Ambient occlusion is an approximation of the amount by which a point on a surface is occluded by the surrounding geometry, which affects the accessibility of that point by incoming ligh.
* An occlusion is a complete or partial blockage
* A statue of a person in a room

  Description automatically generated

Computation Challenges:

* Computing the occlusion factor can be computationally expensive, especially in offline renderers that use ray casting in normal-oriented hemispheres.
* This approach is impractical for real-time rendering.

Crysis Real-Time Solution (2007):

* Crytek introduced a real-time AO solution in Crysis, setting a benchmark for game graphics.
* It uses the depth buffer to approximate scene geometry, enabling on-GPU computation, full dynamics, and independence from scene complexity.

Crysis Method Explained:

A diagram of a circle with a triangle and a line

Description automatically generated with medium confidence

* Instead of hemisphere ray casting, Crysis samples the depth buffer using derived sphere samples.
* Each of the gray depth samples that are inside geometry contribute to the total occlusion factor; the more samples we find inside geometry, the less ambient lighting the fragment should eventually receive.
* Steps:
  1. Project sample points to screen space.
  2. Sample the depth buffer.
  3. If a sample point is behind the sampled depth (inside geometry), it contributes to occlusion.
* The result quality depends on the sample count, which must be kept low for performance, but fewer samples cause "banding" artifacts.
* Crysis mitigates this by randomly rotating the sample kernel at each pixel, trading banding for high-frequency noise that can be removed by blurring.
* The Crysis method produces a unique look: flat walls appear gray because roughly 50% of samples are inside nearby geometry. Concave corners darken, while convex ones lighten, creating a stylistic effect.
* 

Normal-Oriented Hemisphere (Enhancement):

A diagram of a circular object

Description automatically generated

* An enhancement involves using a hemisphere oriented along the surface normal instead of a spherical kernel.
* This improves the effect's appearance but requires per-fragment normal data.
* In a deferred renderer, this data is typically available, and the cost is low compared to the improved quality.
* Rather than sample a spherical kernel at each pixel, we can sample within a hemisphere, oriented along the surface normal at that pixel. This improves the look of the effect with the penalty of requiring per-fragment normal data. For a deferred renderer, however, this is probably already available, so the cost is minimal (especially when compared with the improved quality of the result).

In the context of rendering and ambient occlusion (AO), the "sphere thing" typically refers to the sampling method used to calculate ambient occlusion. Ambient occlusion is a shading technique that simulates how much ambient (indirect) light reaches a point on a surface, based on the proximity of other objects in the scene. It's used to create soft shadows and enhance the realism of 3D scenes.

To calculate ambient occlusion, we need to determine how much the point on the surface is occluded by its surroundings. One way to do this is by sampling the environment around the point in question. The "sphere thing" refers to the shape of the sampling region. Here's how it works:

1. Spherical Sampling: In many traditional AO algorithms, a spherical region of points is sampled around the point on the surface for which we want to calculate AO. This spherical region is centered at the point and extends in all directions.
2. Sample Points: Within this spherical region, multiple sample points are chosen. The number of sample points and their positions within the sphere can vary depending on the specific algorithm and quality requirements.
3. Ray Casting or Depth Testing: For each sample point within the sphere, a ray is typically cast from the surface point towards the sample point. The algorithm checks if this ray intersects with any objects in the scene (typically using ray-object intersection tests) or if the depth value at the sample point indicates that it's inside an object.
4. Calculating Occlusion: Based on the results of these tests, the algorithm calculates an occlusion factor for the surface point. If many sample points are inside or close to objects, it indicates strong occlusion, and the point is in shadow. If most sample points are in open space, it indicates less occlusion, and the point receives more ambient light.

The result of this process is a shading value that can be used in the rendering equation to calculate the final color of the point on the surface.

It's important to note that different AO algorithms may use variations of this spherical sampling approach, and they may use different shapes for the sampling region. The choice of sampling method, the number of samples, and other factors can greatly affect the quality and performance of the ambient occlusion effect in a rendering engine.

Using a hemisphere, specifically a hemisphere oriented along the surface normal at the point of interest, can offer advantages over a full sphere for ambient occlusion (AO) calculations in certain situations. Here's why:

1. Improved Realism: Hemispherical sampling can better capture the occlusion characteristics of the local environment, especially on curved or complex surfaces. This approach aligns the samples with the surface normal, making it more physically accurate.
2. Reduced Samples: A hemisphere covers only half of the sphere's area, which means you need fewer samples to achieve similar quality compared to a full sphere. This can significantly improve performance, which is crucial for real-time applications.
3. Less Noise: With fewer samples, there's less noise in the AO result. Noise in the AO can lead to flickering or visual artifacts. Hemispherical sampling, when properly implemented, often produces smoother and more stable results.
4. Consistency: Hemispherical sampling provides more consistent shading across surfaces. With full spherical sampling, points on a surface can receive contributions from samples located in unrelated directions, potentially leading to less coherent shading.
5. Easier Integration: For a deferred renderer, normal information is often readily available, making it relatively straightforward to use hemispherical sampling. In contrast, full spherical sampling requires sampling in all directions and may be more complex to integrate.

However, it's essential to note that the choice between hemisphere and full sphere sampling depends on the specific requirements of your rendering application and the desired trade-offs between quality and performance. In some cases, particularly when dealing with highly complex environments or specific artistic styles, full spherical sampling may still be preferred despite its computational cost.

In summary, hemispherical sampling is a practical compromise between accuracy and efficiency for ambient occlusion calculations, offering a good balance between realism and performance for many real-time rendering scenarios.

* <https://john-chapman-graphics.blogspot.com/2013/01/ssao-tutorial.html>
* <https://learnopengl.com/Advanced-Lighting/SSAO>