# Rendering System Overview

This document is intended to present the current design of the Hieroglyph 3 rendering system. In general, the design is constantly evolving and hence may change at a later date. However, the items discussed in this document should remain mostly unchanged for the foreseeable future, and hence are worth documenting now.

The information in this document is also intended to supplement the actual source code of the Hieroglyph 3 engine. A significant effort is made to properly document the important parts of the source code, and additional updates are made as frequently as possible. Even so, this is an open source project and is therefore not a fulltime adventure. If there is something not correctly documented or missing documentation, please consider adding a discussion topic at the project page and we will try to update it as soon as possible.

The overall design of the rendering system makes use of four major functional groups, each of which will be discussed in more detail in the following sections.

## Renderer

The true heart of the rendering system is the Renderer itself. The RendererDX11 class is the source for all of the other portions of the rendering system, and hence is the primary class that an application will interface with to do some rendering. The renderer loosely follows the same responsibilities of the ID3D11Device interface from Direct3D 11 at a low level, but also coordinates higher level activities such as scene processing and rendering batches. Here is a short description of each of the renderer’s responsibilities.

### Creation/Management of Resources

The renderer provides the ability for other classes to create and reference the various Direct3D 11 objects. The primary objects that other non-rendering system classes will be interested in are the memory resource classes, which represent buffers and textures. The renderer allows creation of these objects, but retains ownership of them and returns only an ID to the caller. That ID is then used by the caller any time they want to reference a particular resource.

This pattern is also followed for non-resource classes as well, such as shader objects, samplers, input layouts, and so on. Instead of making the other classes work with direct references to D3D11 objects, they are instead referenced by IDs. This allows the renderer to keep control over how and when the D3D11 objects are used, and minimizes the direct knowledge needed by the application to use the engine.

### Coordinates Scene Rendering

The other primary task of the renderer is to coordinate the processing and rendering of a given scene. In Hieroglyph 3, scenes are represented as spatial scene graphs of Entity3D instances. A scene rendering is encapsulated into what is referred to as a render view. We will cover render views in more detail later, but for now knowing that they represent one scene rendering is enough to get started. As a scene is processed, some special effects require additional “views” of a scene to be rendered. For example, if there is a reflective object in the scene, then a reflection map is needed that represents the reflected view of the scene. The renderer provides the interface methods for aiding in a scene to recursively process multiple render views in the proper order.

In addition to this, there is also the ability to automatically utilize the multi-threading features of Direct3D11 to process render views in parallel. When selected to use multi-threading mode, the renderer will process each render view on a separate thread. Each render view produces a single command list that holds all of its submitted API commands. After all views are processed, they are executed on the immediate context. The number of threads is selected at compile time by a constant, with the current default set to 4.

The overall processing time may or may not be reduced by this parallel processing. In particular, very small scenes may not benefit very much, or could even suffer performance penalties for using the multithreading mode. To allow traditional processing, an interface method can be used to only use a single thread and only the immediate context for processing all of the views sequentially. This allows for an easy way to check and see if you are benefitting from parallel processing or not.

There are two major sub-systems used by the renderer in carrying out its scene processing tasks. Those are the Pipeline Manager and the Parameter Manager systems.

#### Pipeline Management

Pipeline management is performed by the PipelineManagerDX11 class. In general, this class loosely follows the idea of a context in Direct3D11, although it adds additional higher level capabilities as well. It is also used for both immediate and deferred context management, and is used to hide the details of using one or the other from the rest of the system. The following sections go into more detail about what this class is used for.

##### Setting States

The primary purpose of the pipeline manager is to provide an interface to the actual rendering and computation pipelines of Direct3D 11. As such, it has methods for setting the various states such as shader programs, resource views like SRVs or UAVs, state objects, and so on.

##### Managing Pipeline State

The pipeline manager is used to monitor and update the various pipeline states. This means monitoring the API calls that are made, and reflecting those calls accordingly in its model of the pipeline. This is especially important when trying to minimize the number of API calls being used, since it can eliminate any redundant state change requests.

This is also a very important topic to consider when working with immediate and deferred contexts. The pipeline state is propagated in different ways depending on how command lists are generated and consumed, so it is very important to update the pipeline model accordingly.

##### Coordinates pipeline execution

Once the pipeline has been configured as desired, the pipeline manager also has methods for invoking the pipeline. In the case of the rendering pipeline, this means calling one of the various DrawXXX calls. For DirectCompute workloads, this involves calling one of the DispatchXXX calls. The following sections describe these in more detail.

###### Drawing

Rendering in the pipeline is performed by first configuring the pipeline, and then calling an appropriate rendering method. This can include performing simple draw calls that only use vertex buffers, indexed draw calls that use vertex and index buffers, instanced drawing that uses both per-vertex and per-index vertex buffers, or even indirect drawing that uses another buffer resource to control the draw call. When possible, a geometry class is used to encapsulate the setting of parameters for each of these draw calls. Geometry classes are discussed in more detail later in this document.

###### Dispatching

Dispatching is a much more broad activity than drawing. Since DirectCompute allows for precise control over the threading mechanisms used by the GPU, it provides the ability to perform more general purpose computation and apply it to an action that isn’t necessarily related to rendering. There are direct dispatching calls as well as indirect dispatch calls to allow using buffer resources to control the pipeline execution.

While the draw calls typically use a geometry object to configure the pipeline, this is typically not the case with dispatching. The general nature of dispatching makes it less amenable to encapsulating its state in a class, and since the compute shader is only a programmable shader stage (i.e. no fixed function stages) then we can automate most of the configuration with the parameter manager system (which we discuss next).

### Parameter System

Many of the settings that are used to configure the pipeline can be resolved to a particular text based name. For example, a shader program that uses a texture to provide the color of a model would use the name ColorMap to reference the texture. The application would likely want to specify which texture resource to use when this name is needed in a shader, as opposed to having to manually specify which shader resource view slot to bind it to. Managing the mapping from an application supplied parameter to the corresponding setting in the pipeline is handled by the parameter system in Hieroglyph 3.

The parameter system is implemented in the ParameterManagerDX11 class. As its name implies, it is responsible for coordinating the parameter data. Basically any client of the entire rendering system can acquire a handle to a parameter with its text based name, and then set that parameter’s value as it sees fit. Then during a later rendering pass, any shader parameter that is needed can easily be read out from the parameter manager. In this sense, the parameter manager can be thought of as a general storage area for named parameters where there are many writers of data and the rendering system is the reader of that data.

#### Multi-threaded Parameters

Since Hieroglyph 3 takes advantage of the multi-threading capabilities of Direct3D 11, that means that the rendering sub-systems must also take into consideration how it will support multiple threads. The parameter manager supports access from multiple threads by having each parameter store multiple copies of its data. Then each thread is assigned an index to write into, which tells it which copy of the parameter data can be written by it (and also which one to read from as well). This is depicted in Figure 1.



Figure 1: The handling of multiple threads by the parameter manager.

By allowing a parameter to support multiple clients with separate data areas, they can simultaneously allow access to multiple threads without trying to re-synchronize the data being used. This multiple copies of data are actually not inherently thread safe, but due to the sequencing of a rendering pass it becomes thread safe. We will walk through some additional details about this sequencing later on.

In addition to allowing thread safe access to the various parameters, this multiple data design allows for a single storage area to house all of the parameters. This is implemented as a single std::map<std::string, RenderParameter\*> which maps names to render parameters. This single map is accessed by all instances of a parameter manager, and then the individual data elements of each parameter are accessed according to the parameter manager’s specified index.

### The Multi-Threaded Renderer

With these two sub-systems defined in addition to the renderer itself, we can discuss how the multithreaded renderer is organized and how it performs the work that it does. As mentioned above, the general unit of GPU work in Hieroglyph 3 is called a render view. A render view is submitted directly to the renderer, where it is queued up until a later time when all of the queued render views are processed. Internally to the renderer, there are multiple copies of pipeline managers and parameter managers in use, and they are always paired together. To represent the immediate context, one pipeline manager is used and a corresponding parameter manager paired to it. Then for each additional thread that will be used in the renderer, another pipeline/parameter manager pair is created and assigned to that thread. For these subsequent pairs, a deferred context is used in the pipeline manager’s. This general concept is depicted in Figure 2.



Figure 2: The pipeline and parameter managers within the renderer.

When the renderer is configured for single threaded mode, only the first pair of pipeline/parameter managers is used to process the render views one at a time. In multithreaded mode, the render views are all processed by the additional pairs of pipeline/parameter managers to produce command lists. These command lists are then processed on the immediate context where they are submitted to the GPU/driver.

Since the pipeline/parameter managers are designed to be modular, it is trivial to increase or decrease the number of threads that are used in the renderer. In addition, the number of additional threads is determined at compile time and the manager pairs are created at startup. This means that it is possible to switch between single threaded and multithreaded rendering mode during runtime, which can be used for debugging, performance comparisons between the two modes, and switching to single threaded mode to utilize specific features which can only be used on an immediate context.

## Object System

With a general idea of who does what in the rendering system, it is time now to turn our attention to the object system and see how it fits into the overall Hieroglyph 3 rendering system. The object system is what provides the objects within a scene their spatial frames of references, as well as the interfaces for using the rendering system. The following sections will provide an introduction to the object system and how it is implemented.

### Spatial Scene Representation

The heart of the object system is the Entity3D class. This class provides a spatial representation of an object, including its position and orientation. Any object that will end up being rendered will be represented as an instance of an Entity3D. With the position and orientation data, interface methods are provided for calculating an appropriate transformation matrix to position this object in the correct place for a given scene rendering.

To provide hierarchical connections within a scene, the Node3D class is provided. This class is a subclass of Entity 3D and provides the data and methods for creating a directed acyclic graph of entities. This allows for the typical hierarchical manipulation of scenes, and provides a fairly simple access mechanism for working with a scene. Since the graph must have a root node, then the entire scene can be handled with a single node reference.

### Coordinates scene processing

This concept of a scene is used in two different aspects. First, the state of a complete scene (which is ultimately just a graph of entities) is updated with a recursive call to the method Entity3D::Update(...). This recursive call will propagate through the scene and have each entity update its world space location based on its own spatial data and that of its parent. This is performed for every frame update, and the delta time that has elapsed between frames is passed as an argument to the method.

Once the scene has been updated, it is ready to be processed by the rendering system. We have already mentioned render views several times in this article, and they are about to be discussed again... For any render view that will be providing a rendering of the scene, it must have a reference to the scene graph itself. This is typically performed by providing it with a reference to the root node of the scene. The graph structure of the scene is then used to provide a couple of methods for rendering the scene. The first is the method Entity3D::PreRender(...) method. This is a recursive call that provides each entity in the scene an opportunity to add additional render views to the renderer for processing. For example, if an object is shiny and needs a paraboloid environment map, then it would push its render view into the renderer’s queue during its pre-render method.

After all of the render views have been added to the queue, all of the render views are then processed by the renderer as detailed above (which will vary depending on if single- or multi-threaded rendering is used). Each object is processed for a particular view by calling the recursive Entity3D::Render(...) method. In this method, each entity performs the needed render pipeline configuration before actually invoking the pipeline and rendering its geometry. The exact setup of this rendering will depend on the type of view being processed, so the material system is used to provide all of the needed information for the expected view types in a given scene.

# Rendering System Details

## Entity

### Geometry

#### Provides pipeline input

### Material

#### Provides pipeline configuration for various scenarios

##### State Objects (i.e. Fixed Function)

##### Shader Objects

## Render Views

### Basic Concept

#### Provide one or more rendering operation

#### Provide logical operation groups

### Scene Processing

#### Recursive Scene Graph Traversal

#### Pre-Rendering

#### Rendering

### Specialized Processing

#### Dispatches

#### Non-scene based rendering passes

## Scene Rendering

### Camera Object

#### Provides a render view for the main window view