

Analysis of Type-Driven approach to systems programming: Implementation of OpenGL library for Rust

(Analiza programowania systemowego z wykorzystaniem systemu typów:
Implementacja biblioteki do OpenGL dla języka Rust)

Mikołaj Depta

Praca inżynierska

Promotor: dr Andrzej Łukaszewski

Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Informatyki

19 sierpnia 2024

Abstract

For the past few years in the software development industry there has been a growing interest in strongly typed languages. It manifests itself in emergence of brand-new technologies in which strong type systems were one of the core founding principles or in changes introduced to existing languages. The most common examples of modern languages with powerful type systems are TypeScript as an alternative to JavaScript in the world of web development or Rust in domain of systems programming in place of C and C++. More mature languages also had their type systems revised for example in C# 8 explicit type nullability annotations were introduced, or even dynamically typed Python has seen major improvements to its type annotation system.

This study - an implementation of the OpenGL graphics API wrapper library for Rust - will attempt to demonstrate how Rust's type system can be utilized to improve low-level software safety and maintainability as well as how it affects API design and codebase structure.

For the past few years in the software development industry there has been a growing interest in strongly typed languages. It manifests itself in emergence of brand-new technologies in which strong type systems were one of the core founding principles or in changes introduced to existing languages. The most common examples of modern languages with powerful type systems are TypeScript as an alternative to JavaScript in the world of web development or Rust in domain of systems programming in place of C and C++. More mature languages also had their type systems revised for example in C# 8 explicit type nullability annotations were introduced, or even dynamically typed Python has seen major improvements to its type annotation system.

This study - an implementation of the OpenGL graphics API wrapper library for Rust - will attempt to demonstrate how Rust's type system can be utilized to improve low-level software safety and maintainability as well as how it affects API design and codebase structure.

Contents

1	The Rust programming language	7
1.1	Data types	9
1.1.1	User defined types	9
1.2	Implementations	9
1.2.1	Inherent implementations	10
1.3	Generic types	10
1.3.1	Type parameters	10
1.3.2	Constant parameters	10
1.3.3	Lifetime parameters	10
1.4	Type level abstraction	10
1.5	Implementation Coherence	11
1.5.1	Orphan Rule	11
2	OpenGL	13
2.1	Introduction	13
2.2	OpenGL objects	13
2.2.1	Buffer object	14
2.2.2	Vertex Array object	14
2.3	History	14
2.4	Modern OpenGL	15
2.5	Graphics pipeline	15
2.5.1	Vertex Specification	15
2.5.2	Vertex Shader	16

2.5.3	Tessellation	16
2.5.4	Geometry shading	17
2.5.5	Fixed function vertex post-processing	17
2.5.6	Rasterization	17
2.5.7	Fragment processing	18
2.5.8	Fragment post processing	18
2.6	GLSL	18
2.6.1	Variables	19
2.6.2	Variable type qualifiers	19
2.6.3	Variable storage qualifiers	19
2.6.4	Variable layout qualifiers	19
Bibliography		23

Chapter 1

The Rust programming language

Introduction

Rust is an open-source, general purpose, multi paradigm, compiled, statically and strongly typed language maintained by Rust Foundation.

Although general purpose, language lends itself particularly well to systems programming, where software reliability, scalability and efficiency are paramount. These qualities can be attributed to powerful and expressive type system and ownership based memory management which guarantees memory safety without a garbage collector.

1.1 Data types

1.1.1 User defined types

Rust provides

As of version 1.79 Rust provides 3 Rust provides 3 ways to construct used defined aggregate types. These are: - **structs** - **enums** - **unions**

We focus on the first two of these.

structs Structure is a named grouping of other data types (product type). They constitute the basic building blocks for any user defined types.

1.2 Implementations

Functionality of a type is not provided inline with it's definition like in most C like languages. Instead it's associated with a type by so called *implementations*. Implementations for a type are contained within a block introduced by the **impl** keyword.

There are two types of implementations: - inherent implementations - trait implementations

All items within impl block are called associated items

Functions defined within `impl` blocks are called associated functions and can be accessed with qualification `<type-name>::<function-name>`.

Within an implementation both `self` and `Self` keywords can be used. `self` can be used in associated function definitions within an `impl` block as first parameter. Such functions are called methods and the `self` parameter denotes the receiver of method call. `self` can be additionally qualified with immutable or mutable reference `&` or `&mut`.

`Self` is a type alias that refers to implementing type.

1.2.1 Inherent implementations

We will shorthand implementation to `impl` which is common in Rust terminology.

Inherent `impl`s associate contents of an `impl` block with specified nominal type. Such blocks can contain definition of a constants or functions.

1.3 Generic types

As of version 1.79 Rust provides 3 kinds of generic parameters types can use. These are: - type parameters - constant parameters - lifetime parameters Type which uses any generic parameters is said to be generic.

1.3.1 Type parameters

Type parameters represent an abstract type

1.3.2 Constant parameters

1.3.3 Lifetime parameters

1.4 Type level abstraction

Traits provide an ability to express shared behavior in abstract way [1]. We are mostly interested in their use in *trait bounds* on types and type parameters. Trait bounds declare contracts that types must fulfil or else the program will be rejected. We used it to enforce use of valid data formats and proper sequencing of operations.

1.5 Implementation Coherence

1.5.1 Orphan Rule

Chapter 2

OpenGL

2.1 Introduction

OpenGL is an specification of an API for hardware accelerated computer graphics owned and maintained by the Khronos Group Inc.

Since it's inception and up until 2016 with release of Vulkan it has been the only widely supported cross platform graphics API.

TODO: opis opengl'a

2.2 OpenGL objects

OpenGL exposes an abstraction over GPU's resources called objects. In order to use given object it first must be bound to a var point in current OpenGL context.

Objects contain internal state which can be queried using introspection. Objects are identified by a *name* which is an unsigned 32 bit integer. There exists common object model which describes how most types of objects are managed.

Most types of object can be created with either call to

```
void Gen*s(sizei n, uint *objects)
```

which will allocate object's *name* and subsequent call to

```
void Bind*(uint name, ...)
```

which will bind given object to the context and, if given object has never been bound before, allocate their internal state. Alternatively one can use

```
void Create*s(sizei n, uint *objects)
```

which will allocate both object's *name* and its state but it will not set any context bindings. There exist separate namespace for each object type.

The most notable outliers that do not conform to the rules above are program objects, shader objects

Objects can be deleted with `void Delete*s(sizei n, uint *objects)`, bound with aforementioned `void Bind*(uint name, ...)` which usually accepts additional parameter that specifies var point.

OpenGL specification defines set of publicly available object parameters which can be queries using introspection with **GetInteger*** family of functions. One notable usage is determining compilation and linking status for shaders and programs.

2.2.1 Buffer object

Buffer objects provide means to allocate data stores in GPU memory. They can contain data of different format and purpose depending on buffer's target. Primary usage for buffers is to provide geometric information which includes vertex attribute values and indices for indexed rendering.

2.2.2 Vertex Array object

Modern OpenGL is generic over vertex format and only poses limitation on the number of such attributes and limits their values to glsl's scalar or vector types. Each attribute is assigned a zero-based index. Vertex Array object (VAO) assigns each active attribute information on how and where from to source vertex data, as well as, what is the data type of provided attribute in glsl.

This can be seen as two separate pieces of information: - in memory layout and access frequency - interpretation / conversion of the data.

Originally all of that information was specified at one with

```
void VertexAttrib*Pointer( )
```

2.3 History

IRIS GL, a proprietary graphics API, which later became OpenGL was initially developed by Silicon Graphics (SGI) during 1980's. SGI open sourced subset of their API as OpenGL due to mounting market pressure from SGI competitors (Sun Microsystems, Hewlett-Packard, IBM) who started providing their own APIs for hardware accelerated 3D graphics based on an existing open standard called PHIGS.

In 1992 OpenGL Architectural Review Board (OpenGL ARB) was established

and it was tasked with development and maintenance of the OpenGL specification. This task was passed on to Khronos Group in 2006 where it remained ever since.

2.4 Modern OpenGL

2.5 Graphics pipeline

OpenGL 4.6 models graphics pipeline as follows:

OpenGL 4.6 Pipeline Description

The OpenGL 4.6 pipeline is a sequence of stages that process and render 3D graphics onto the screen. It involves both fixed-function stages and programmable shader stages. Below is a detailed breakdown of each stage:

2.5.1 Vertex Specification

Before any rendering can begin geometric information needs to be provided as generic vertex attributes.

Vertex is an abstract composition of attributes that means to represent a point in 3D space of some object that is being modelled. Generic stands from the fact that data associated with vertices has no intrinsic meaning. Semantics of data are dictated by client provided vertex shader.

OpenGL sources data for each vertex attribute from a buffer. Each attribute is assigned an index. Association between attribute with given index and a buffer from which should that attribute source its data is established by Vertex Array Object (abbrev. VAO).

Once all vertex attributes have their data sources assigned and properly configured, vertex specification can be considered finished and one could proceed with further pipeline configuration. In this instance vertices would be interpreted sequentially as appropriate primitives. This forces vertex data to be specified redundantly for lines and especially raw triangles since each triangle shares each edge with each neighboring triangle.

To better conserve memory one can use indexed rendering. This requires additional buffer filled with indices into main vertex buffer instead of inlined vertex data. In case of basic triangle rendering (without using compressed representations like triangle fan or triangle / line strip) will still cause repetition but now only few byte wide indices instead of whole attributes which are substantially larger.

2.5.2 Vertex Shader

Vertex shader is the first programmable stage of OpenGL Pipeline and is one of two required shaders to execute a draw call, the other being the fragment shader.

Most commonly vertex shader performs 3 translations. From initial model space, world space, view space to final clip space which we will now discuss briefly.

model space - when a 3D model is created in 3D modelling software its vertex positions are specified to some local coordinate system (commonly center of an object). These positions would commonly be loaded into gpu memory. Such objects can be easily placed in broader scene by providing a so called world transform performs transformation from model's local coordinate system to scene's coordinate system.

world space - world space refers to coordinate system of a scene that uses multiple models by translating, scaling or rotating them.

view space - its common for 3D rendering applications to provide means of interacting with the scene. Whether its a 3D computer game, CAD program or medical data visualization we would like to be able to control how scene is displayed by moving a virtual camera. This can be expressed as yet another transformation of the coordinate system - we would like to transform coordinate system to align with the position of our camera. This transformation is commonly called view transform.

clip space - having accounted for model position in a scene and user interactivity all that remains is to provide vertex data in form that subsequent fixed function pipeline stage - the rasterizer - expects. Namely once vertex shader is finished fixed function processing will perform so called perspective divide and then clip all geometry to normalized device coordinates (NDC). Output of vertex shader is a 4 component vector which corresponds to a 3D position in homogenous coordinate system used in computer graphics due to its ability to represent non linear transformations using matrices. The output position of vertex is divided by the forth component in order to introduce perspective

The main responsibility of vertex shader is to transform vertices to clip space, which will be discussed in future subsection.

2.5.3 Tessellation

Tessellation stages were added as graphics hardware compute capability grew. With raw compute throughput outperforming bus throughput GPUs were equipped with hardware tessellation unit which can subdivide a larger triangle into batch of smaller ones. This allows for efficient generation of geometric detail on chip alleviating the issue of limited PCI throughput. To drive the tessellation stage two new shaders were introduced:

- **Tessellation control shader** which configures how hardware tessellator should subdivide a triangle.
- **Tessellation evaluation shader** which performs transformations on vertices generated by the tessellator.

2.5.4 Geometry shading

Geometry shader was introduced prior to tessellation stage. They operate on assembled geometric primitives and may even access primitives neighbors. Given primitive input geometry shaders output one or more primitive of the same type.

2.5.5 Fixed function vertex post-processing

Once all programmable vertex progressing has concluded, a series of fixed-function operations are applied to the vertices of the resulting primitives before rasterization. These operations include transform feedback, which captures processed vertex data, primitive queries to gather information about the primitives being processed and flat shading which applies a uniform attribute value to a whole primitive.

Primitives then get clipped against clip volume and client-defined half-spaces. The clip coordinates undergo perspective division, followed by viewport mapping to adjust for screen coordinates and depth range scaling.

2.5.6 Rasterization

If neither tessellation stage nor geometry stage was used in vertex processing, primitive assembly takes place (presence of any of the aforementioned stages would necessitate early primitive assembly). OpenGL converts geometric primitives used in currently processed draw call into base primitives which are points, lines and triangles. Mathematical representation of primitives is used during rasterization to determine if given fragment falls inside of primitive being rasterized.

Process of rasterization requires determining if given pixel position falls inside of rendered primitive. This process needs to account for point and line thickness. Polygon rasterization is obviously the most complex of the three. Prior to insidiness test face culling is performed. This optimization culls a polygon based on the sign of surface normal computed based on edge ordering as specified in vertex array. This helps reduce overdraw which can be one of two main bottlenecks in modern rendering system, the latter being insufficient memory bandwidth.

Once pixel location was deemed inside a primitive a fragment is generated. A Fragment is a collection of data corresponding to specific pixel location. Most commonly its perspective corrected barycentric interpolation of vertex data across

the primitive's surface. Tough interpolation can be disabled from within vertex shader using `flat` qualifier on output variable declaration, as well as perspective correction with `noperspective` qualifier.

Once fragments are computed early per-fragment tests take place.

- **Ownership test** - determines if pixel at location (x, y) falls into the portion of the screen that active OpenGL context owns.
- **Scissor test** - checks if pixel at location (x, y) is contained within client provided list of axis aligned rectangles
- **Early Fragment tests** - stencil test, depth test and occlusion query which are normally performed after fragment processing can optionally be performed early. We discuss them in subsection on fragment post processing.

If all tests passed fragment is submitted for programable fragment processing.

2.5.7 Fragment processing

Programable fragment processing is performed by client provided fragment shader. The most essential task that fragment shader should perform is assign pixel a color. For that purpose data interpolated from rasterization is used. Most commonly fragment shaders perform texture mapping, lighting calculations, parallax mapping to emulate geometric detail and screen space effects like ambient occlusion, use signed distance functions and implicit surface equations to render otherwise complex scenes all by itself or create volumetric effects like clouds or visualize CT scan results.

2.5.8 Fragment post processing

2.6 GLSL

GLSL is a high level shading language with c like syntax developed by OpenGL Architecture Review Board to power programable processing stages in OpenGL pipeline. GLSL code is still relevant as it can be compiled into SPIR-V and used with Vulkan API.

Shaders

Independent compilation units written in this language are called shaders. A program is a set of shaders that are compiled and linked together, completely creating one or more of the programmable stages of the API pipeline

In OpenGL 4.6 and GLSL 4.60 there exist 6 types of shaders: vertex, tessellation control and evaluation, geometry, fragment and compute. All shaders except compute shader control appropriate parts of OpenGL pipeline as described in subsections above.

Compute shaders operate completely outside of graphics pipeline. They can access same resources as fragment or vertex shader like textures, buffers, images and atomic counters but they are not expected to produce data with predetermined form or semantics. They offer general purpose compute capability on the GPU. They function similarly to other existing general purpose GPU compute APIs like CUDA or OpenCL.

2.6.1 Variables

Shaders

2.6.2 Variable type qualifiers

2.6.3 Variable storage qualifiers

2.6.4 Variable layout qualifiers

/chapterExisting solutions

There are many qualities of any software library one could consider important. In this research we focused foremost on providing minimalistic wrapper and staying as faithful as possible to original specification of the API. By this we mean that appropriate GL functions take analogous parameters as in original spec and have their names and semantics preserved. Major benefit of this approach is that we could simply follow the OpenGL specification when creating type safe facades around procedures.

Starting from these minimalistic principles we focused on providing maximal level of type safety. The main goal was to enable rejection of as many ill-formed programs at compile time as possible.

There are many levels of safety guarantees we can expect from any software package. In this analysis we devise

Here we consider alternative ways of programming computer graphics with use of OpenGL as rendering backend.

We distinguish between a language of choice and any framework at use.

/sectionNative C / C++ bindings

The simplest way one can program with OpenGL is using platform provided C bindings contained within an os provided dynamic link library (.dll for MS Windows, .so) along with appropriate function pointer loader. The requirement for the latter stems from common practice among OS vendors to officially guarantee distribution of very dated version of the specification (1.1 for Windows). This poses a requirement for manual function pointer loading at runtime, an approach that has two main benefits - abstracts away details of dynamic library loading for different platforms, - provides unified mechanism for using optional core standard extensions.

Additionally to function pointer loading one needs to initialize OpenGL context following platform defined protocol.

Most commonly there exists a library for each task, some examples for PC are GLEW for function loading and GLFW for window creation and context creation. Once these actions are accomplished one can use OpenGL in C or C++ provided appropriate attention to C interoperability.

Writing C application provides no auto

/sectionRust with unsafe bindings

Rust toolchain provides a utility for automatically generating Rust Foreign Function Interface bindings to C called `bindgen`. In this case all the setup needed for a Native C / C++ bindings application still applies. There exist appropriate counterparts to GLEW and GLFW. Once context is initialized and function pointers

loaded one can call C functions but Rust will require one to use these functions inside `unsafe` context.

Bibliography

- [1] Steve Klabnik and Carol Nichols. *The Rust programming language*. No Starch Press, 2023.