МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра автоматики   
 (полное название кафедры)   
  
  
  
  
  
  
  
  
 **ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА**  
Сизова Льва Сергеевича

#### Автор

Сизов Л.С.

(фамилия, И., О.)

АВТФ, АВТ-318

(факультет, группа)

  
 (подпись, дата)

#### Руководитель

Гунько А.В.

(фамилия, И., О.)

к.т.н., доцент

(уч. степень, уч. звание)

  
 (подпись, дата)

(фамилия, имя, отчество автора – студента-выпускника)   
  
Исследование возможностей библиотеки Vulcan для сложных математических задач

(тема работы)   
  
  
  
Направление подготовки 09.03.01. Информатика и вычислительная техника

(код и наименование направления подготовки бакалавра)

Новосибирск, 2017 г.

Оглавление

[Введение 3](#_Toc483165761)

[1. Техническое задание 6](#_Toc483165762)

[1.1. Сведения и исходные данные для разработки программы 6](#_Toc483165763)

[1.2. Требования к программе 6](#_Toc483165764)

[2. Обзор библиотеки *Vulkan* и её функций для вычислений на графических процессорах 12](#_Toc483165765)

[2.1. Общие сведения 12](#_Toc483165766)

[2.2. Обзор функций, структур данных, утилит 24](#_Toc483165767)

[3. Проектирование 35](#_Toc483165768)

[4. Реализация 36](#_Toc483165769)

[5. Тестирование 37](#_Toc483165770)

[6. Заключение 38](#_Toc483165771)

[7. Приложение 39](#_Toc483165772)

## Введение

Последние поколения процессоров показывают хорошие результаты в параллельных вычислениях, однако до видеокарт им далеко. Причина в том, что видеокарты содержат в себе много графических процессоров, предназначенных для параллельной обработки.

К сожалению, ещё не существует единого стандарта для написания программ для графических процессоров. Основными средствами для написания программ для графических процессоров являются *CUDA* от *NVidia* и *OpenCL* от *AMD*. Однако 16 февраля 2016 года группой компаний *Khronos Group* был выпущен *Vulkan API* предназначенный для программирования графики и вычислений на графических процессорах, используемых в различных платформах – от персональных компьютеров до мобильных телефонов и игровых приставок. Технология *Vulkan* разработана с заделом на будущее, поэтому она будет очень активно развиваться и поддерживаться разработчиками драйверов и графических процессоров.

Для непосредственных вычислений на графических процессорах в библиотеках *Vulkan* и *OpenCL* используется язык программирования шейдеров *GLSL*. Однако для обеспечения кроссплатформенности шейдеры компилируются в промежуточный байт-код *SPIR-V*.

На текущий момент библиотека *Vulkan* используется только такими компаниями, как *Epic Games*, *Unity Technologies* и *Valve Corporation* в их движках и инструментах для разработки компьютерных игр. Также на момент написания этого документа не было найдено ни одной библиотеки для упрощения разработки программ с использованием вычислений на графических процессорах с использованием технологии *Vulkan*. Так как разработка приложений с использованием библиотеки *Vulkan* является не самой тривиальной задачей из-за необходимости гибкой настройки графического конвейера и буферов данных, было решено разработать кроссплатформенную библиотеку для этих целей.

Новизна данной разработки заключается в том, что это будет первая библиотека для вычислений на графических процессорах, использующая вычислительные возможности видеокарт, работа с которыми происходит с использованием только вычислительных функций библиотеки *Vulkan*. С помощью данной разработки планируется изучить возможность и целесообразность использования самой библиотеки *Vulkan* для математических вычислений на графических процессорах.

Актуальность данной разработки заключается в том, что сейчас происходит лавинообразный рост объёмов информации, которые нужно обрабатывать, проводить вычисления на основе этих данных. Наличие подобной библиотеки ускорит разработку приложений для обработки этих данных. Существуют такие технологии, как *CUDA* и *OpenCL* для высокопроизводительных вычислений на графических процессорах. Однако у них есть свои минусы, которые могут отсутствовать в решениях на основе библиотеки *Vulkan*.

Так как разработчиком библиотеки *Vulkan* заявлены лучшая по сравнению с *OpenGL* поддержка многопоточности, уменьшенная нагрузка на центральный процессор и низкоуровневый контроль над графическим процессором, то было решено сравнивать производительность вычислений с программами, написанными с использованием библиотеки *CUDA SDK*, так как технология *CUDA* является лидирующей в данном классе задач.

Практическая ценность разработки будет выявлена по результатам сравнения производительности вычислений на основе разработанной библиотеки и с использованием технологии *CUDA*.

Цель работы – исследовать возможность, целесообразность и практическую ценность вычислений на графических процессорах с использованием технологии *Vulkan*.

Задачи:

* разработать библиотеку-обёртку над вызовами библиотеки *Vulkan* и предоставить в ней *API* для вычислений на графических процессорах;
* протестировать разработанную библиотеку и сравнить производительность вычислений с вычислениями на основе технологии *CUDA* с использованием аналогичных алгоритмов;

## Техническое задание

### Сведения и исходные данные для разработки программы

##### Наименованиепрограммы

Наименование программы: “Библиотека *vulkalc* для вычислений на графических процессорах для решения сложных математических задач”.

##### Сроки разработки библиотеки

Начало разработки: 19.02.2017

Окончание разработки: 10.06.2017

##### Цели создания программы

Упростить разработчикам процесс создания программ с использованием вычислений на графических процессорах с использованием технологии *Vulkan* за счёт предоставления “обёртки” над вызовами библиотеки *Vulkan* и набора функций для использования вычислительных возможностей видеокарт, запуска вычислительных шейдеров и получения результатов расчётов.

##### Назначение и область применения

Разработка предназначена для применения при написании программ для высокопроизводительных математических вычислений на графических процессорах.

### Требования к программе

##### Общие положения

Основные требования, предъявляемые к разрабатываемому продукту:

* Иметь открытый исходный код;
* Иметь открытую лицензию;
* Иметь соглашение по оформлению исходного кода;
* Иметь инструкции по сборке;
* Иметь демонстрационный проект для демонстрации примеров использования;
* Иметь инструкцию по сборке демонстрационного проекта;
* Иметь документацию на английском языке по публичным классам, методам, полям, функциям;
* Компилироваться под операционные системы *Windows* и *Linux;*
* Иметь юнит-тесты;
* Проходить все существующие тесты;
* Использовать систему контроля версий;
* Иметь комментарии в исходном коде;
* Компилироваться и запускать тесты на целевых операционных системах при каждом изменении;
* Иметь как можно меньше зависимостей;

##### Требования к техническому обеспечению

Для использования библиотеки необходим персональный компьютер:

* ОС: *Windows 7, 8, 8.1, 10* или *Ubuntu 14.04* и выше или аналогичные;
* Процессор: любой *x86* или *x86-64* процессор;
* Оперативная память: не менее 2ГБ;
* Дисковое пространство: около 2МБ для *Vulkan Runtime*, около 150МБ для *Microsoft Visual C++ Redistributable*(*Windows*), около 150МБ для видеодрайвера *Linux* или около 350МБ для видеодрайвера *Windows*;
* Дискретная видеокарта с поддержкой технологии *Vulkan;*

##### Требования к функциональности

Разрабатываемая библиотека *vulkalc* должна предоставлять следующие возможности:

* Статическое и динамическое связывание с программами на операционных системах *Windows* и *Linux;*
* Проверка наличия поддерживаемых графических процессоров и перечисление установленных в системе поддерживаемых графических процессоров;
* Проверка наличия в системе установленного *Vulkan Runtime;*
* Выбор графического процессора из списка доступных для запуска на нём вычислений;
* Запуск разработанных пользователем вычислительных шейдеров;
* Использование предоставленных разработчиком библиотеки *vulkalc* встроенных функций математических вычислений и связанных с ними типами данных;
* Замер времени выполнения вычислений и количестве операций, выполненных на графическом процессоре;
* Подготовка установленных поддерживаемых графических процессоров к вычислениям;
* Освобождение графического процессора;
* Запуск и остановка вычислений на поддерживаемых установленных графических процессорах;
* Копирование данных в видеопамять;
* Копирование данных из видеопамяти;
* Выделение видеопамяти для собственных нужд определённого размера;
* Очищение выделенной видеопамяти;
* Сообщение кода и информации об ошибке в случае её возникновения;
* Обнаружение, валидация шейдеров, компиляция шейдеров в промежуточный байт-код *SPIR-V;*

Библиотека *vulkalc* должна предоставлять следующие функции для математических операций на графическом процессоре:

* Функции работы с векторами:
  + Сложение векторов;
  + Векторизованное сложение массивов векторов;
  + Вычитание векторов;
  + Векторизованное вычитание массивов векторов;
  + Векторизованное сложение массивов векторов и скаляров;
  + Векторизованное вычитание массивов векторов и скаляров;
  + Векторизованное умножение массива векторов на массив скаляров;
  + Векторизованное деление массивов векторов и скаляров;
  + Скалярное произведение векторов;
  + Векторизованное скалярное умножение массивов векторов;
  + Векторизованное нахождение длины векторов в массиве;
  + Свёртка векторов;
* Функции работы с матрицами:
  + Проверка равенства матриц;
  + Векторизованная проверка равенства массивов матриц;
  + Сложение матриц;
  + Векторизованное сложение массивов матриц;
  + Вычитание матриц;
  + Векторизованное вычитание массивов матриц;
  + Произведение матриц;
  + Векторизованное произведение массивов матриц;
  + Умножение матрицы на число;
  + Векторизованное умножение массива матриц на массив чисел;
  + Возведение матрицы в степень;
  + Векторизованное транспонирование массива матриц;

Библиотека *vulkalc* должна предоставлять следующие вспомогательные функции для вычисления на графическом процессоре:

* Сортировка массива чисел;
* Линейный поиск в неотсортированном массиве чисел;
* Векторизованное сложение, вычитание, произведение и деление массивов чисел;
* Векторизованное возведение в степень массивов чисел;

С библиотекой *vulkalc* должна поставляться утилита, предоставляющая следующую функциональность:

* Получение технических характеристик установленных поддерживаемых графических процессоров – частота графического процессора, размер видеопамяти, разрядность шины памяти, пропускная способность памяти, количество видеочипов

##### 1.2.4. Требования к организации входных данных

При написании собственных вычислительных шейдеров программист сам определяет формат и вид входных данных. Формат и вид входных данных для предоставляемых шейдеров описан в документации проекта.

##### Требования к организации выходных данных

При написании собственных вычислительных шейдеров программист сам определяет формат и вид выходных данных. Формат и вид выходных данных для предоставляемых шейдеров описан в документации проекта.

##### Требования к временным характеристикам

Требования к временным характеристикам не предъявляются

##### Требования к информационному обеспечению

Исходный код программы должен быть написан на языке *C++*. Должна использоваться стандартная библиотека языка *C++* и библиотека *Vulkan*. Классы, структуры, методы, функции, переменные, предназначенные для использования пользователем, должны быть экспортированы в соответствии со стандартом.

##### Требования к программным средствам, используемым программой

Программные средства, используемые библиотекой *vulkalc*, которые должны быть предоставлены ОС семейства *Windows*:

* *Microsoft Visual Studio 13* или *15;*
* *CMake 2.8.7* и выше;
* *Vulkan Runtime 1.0.39.1* и выше;
* Для видеокарт *NVidia* – драйвер версии 377.06 и выше, для видеокарт *AMD* – самая новая версия драйвера;
* *Microsoft Visual C++ Redistributable 2013* и *2015;*
* *Doxygen* для генерации документации;

Программные средства, используемые библиотекой *vulkalc*, которые должны быть предоставлены ОС семейства *Linux*:

* *CMake 2.8.7* и выше;
* *Vulkan Runtime 1.0.39.1* и выше;
* *Clang 3.4* и выше или *gcc 4.6.3* и выше;
* Мета-пакет *build-essentials;*
* Пакеты *bison*, *libx11-xcb-dev*, *libxkbcommon-dev*, *libmirclient-dev*, *libwayland-dev*, *libxrandr-dev* для *Ubuntu/Debian* и аналоги этих пакетов для других ОС;
* Стандартная библиотека *C++;*
* Для видеокарт *NVidia* – драйвер версии 377.06 и выше, для видеокарт *AMD* – самая новая версия драйвера;
* *Doxygen* для генерации документации;

## Обзор библиотеки *Vulkan* и её функций для вычислений на графических процессорах

### Общие сведения

##### Терминология

В спецификации *Vulkan API* используются понятия, описанные ниже:

* *ABI* (рус. Двоичный интерфейс приложения) – Набор соглашений для доступа к операционной системе и другим низкоуровневым сервисам, спроектированный для переносимости исполняемого кода между машинами, имеющими совместимые ABI
* *API* (рус. Интерфейс программирования приложений) – Набор готовых классов, процедур, функций, структур и констант, предоставляемых приложением (библиотекой, сервисом) или операционной системой для использования во внешних программных продуктах
* *Assembler* (рус. Сборщик) – Один из этапов каждой стадии конвейеров Vulkan, который преобразует входящие данные к нужному виду
* *Command* (рус. Команда) – Непосредственные команды для графического процессора
* *Command Buffer* (рус. Командный буфер) – Структура данных, содержащая команды для графического процессора
* *Compute Shader* (рус. Вычислительный шейдер) – Программа на специальном языке для графического процессора, предназначенная для вычислений
* *CPU* (рус. ЦП) – Центральный процессор
* *Device* (рус. Устройство) – Логическая сущность, представляющая собой физическое устройство. Содержит в себе информацию об устройстве
* *Extension* (рус. Расширение) – Наборы расширения базового функционала Vulkan с помощью обратных вызовов
* *Fragment Shader* (рус. Фрагментный (пиксельный) шейдер) – Один из этапов графического конвейера
* *Geometry Shader* (рус. Геометрический шейдер) – Один из этапов графического конвейера
* *GPU* (рус. ГП) – Графический процессор
* *Host* (рус. Хост) – Программа, отправляющая команды на графический процессор
* *Physical Device* (рус. Физическое устройство) – Графический процессор, выполняющий команды, отправленные с хоста
* *Queue* (рус. Очередь) – Структура данных, в которую поступают команды для устройства
* *Queue Family* (рус. Семейство очередей) – Набор очередей, предоставляющих определённые возможности
* *SDK* – Комплект средств разработчика для создания программ для определённой платформы
* *Shader* (рус. Шейдер) – Программа на специальном языке, предназначенная для исполнения на ГП. В Vulkan представляет собой программируемую стадию графического конвейера
* *SPIR-V* – Промежуточный кроссплатформенный шейдерный байт-код
* *Tessellation* (рус. Тесселяция) – Автоматизированный процесс добавления новых выпуклых многоугольников в полигональную сетку с целью повышения детализации сетки.
* *Tessellation Shader* (рус. Шейдер тесселяции(замощения)) – Один из этапов графического конвейера
* *Validation Layer* (рус. Слой валидации) – Прослойка между драйвером и хостом для перехвата ошибок и обработок ошибок, записи логов и трассировки вызовов
* *Vertex Shader* (рус.Вершинный шейдер) – Один из этапов графического конвейера

##### Архитектурная модель

Библиотека *Vulkan* спроектирована и написана для ЦП, ГП и других аппаратных ускорителей со следующими свойствами:

* Поддержка 8, 16, 32 и 64 битных знаковых и беззнаковых целых с обратным кодом для представления отрицательных чисел
* Поддержка 32 и 64 битных чисел с плавающей запятой с точностью в 6 знаков
* Порядок байтов в двоичном представлении чисел должен быть одинаковый для хоста и физического устройства

Библиотека *Vulkan* представляет собой интерфейс программирования приложений со свободными точками входа в функции. Большая часть параметров и возвращаемых типов представляет собой структуры языка *C,* обязательные для заполнения. Все указатели по умолчанию имеют размер 64 бита, если не указан явный размер в 32 бита или на платформе *x86.*

Двоичный интерфейс приложения предоставляется производителем драйвера для конкретной платформы. Например, поставщиком *ABI* для *Android* является компания *Google.*

В исходном виде, драйвер *Vulkan* не возвращает никаких ошибок, кроме критических, и не проверяет поступающие данные. Для проверки данных используются слои, которые играют роль прослойки между хостом и драйвером. В этих слоях происходит проверка входящих данных, но никакой информации о ошибках или успехе проверок они не возвращают, а просто предотвращают критические сбои в работе программы.

Рисунки 2.1 и 2.2 демонстрируют механизм работы слоёв в случае успешного прохождения всех проверок (Рисунок 2.1) и в случае провала одной из проверок (Рисунок 2.2).

|  |  |
| --- | --- |
| C:\Users\whyami\Desktop\normal.jpg | C:\Users\whyami\Desktop\fail.jpg |
| Рис. 2.1. | Рис. 2.2. |

Базовый функционал *Vulkan* можно расширить с помощью расширений, основная часть которых поставляется вместе с *SDK*; другие расширения поддерживаются поставщиками драйверов и *ABI*. Примером расширения является базовое расширение *VK\_EXT\_debug\_report* позволяет получать детальную информацию о поведении программы и ошибках при её исполнении. Примером стороннего расширения является расширение *VK\_NV\_external\_memory\_capabilities* от *NVidia*, которое позволяет использовать память в приложении *Vulkan* память, выделенную средствами *Direct3D*.

При включенном расширении *VK\_EXT\_debug\_report* у слоя появляется возможность передавать информацию об ошибках на предыдущий слой. Тогда механизм работы слоёв меняется (Рисунок 2.3) и появляется возможность передавать отладочную информацию или информацию об ошибках при проверках в слоях назад.

##### Конвейеры *Vulkan*

Библиотека *Vulkan* предоставляет два конвейера для использования: графический и вычислительный. Они изображены на рисунке 2.4. Структуру конвейеров и порядок стадий изменить нельзя (можно только пропустить некоторые стадии). Жёлтым цветов на рисунке 2 выделены программируемые стадии конвейера, называемые шейдерами.

В шейдер передаётся необходимая информация через *Buffer* или *Image*. Где первый может содержать любые данные, а второй только информацию о пикселях в определённом изображении.

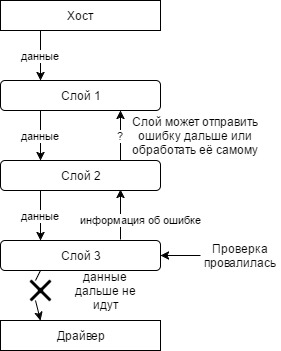


Рис. 2.3. Механизм работы слоёв со включенным расширением обработки ошибок

Перед тем как запустить шейдер на выполнение, необходимо передать ему полученную из приложения информацию. Это происходит на этапе *Assembler,* который присутствует у каждой стадии конвейера. *Assembler* занимается тем, что преобразует читает и преобразует в нужный формат полученные данные и передают дальше в шейдер.

В графическом конвейере существуют следующие стадии:

* Вершинный шейдер – производит преобразования для переноса позиций вершин из пространства моделей в пространство экрана. Он так же передает данные по вершинам далее по конвейеру;
* Шейдер тесселяции (опциональный) – позволяет разбивать трёхмерную сетку модели на основе определённых правил для повышения качества сетки;
* Геометрический шейдер – позволяет добавлять, изменять и удалять трёхмерные примитивы
* Фрагментный(пиксельный) шейдер – производит выборку из текстур и/или математические операции над цветом и значением глубины пикселей для каждого отображаемого на экране пикселя



Рис. 2.4. Графический и вычислительный конвейеры *Vulkan*

Вычислительный конвейер состоит из всего одной стадии – вычислительного шейдера. Он позволяет производить произвольные вычисления для, например, симуляции системы частиц или расчёта отражения лучей.

##### Модель выполнения

Библиотека *Vulkan* даёт доступ к физическому устройству через одну или несколько логических сущностей Устройство. Оно в своё очередь предоставляет доступ к очередям выполнения команд. Наборы поддерживаемых ГП очередей разбиты на семейства. Каждое семейство поддерживает функциональность определённого типа и может содержать несколько очередей со схожими параметрами. Существует 4 типа очередей:

* Для работы с графикой
* Для работы с вычислениями
* Для перемещения данных
* Для работы с разреженной памятью

Память на физическом устройстве может быть выделена только явным образом с хоста. Каждое устройство может создать несколько блоков памяти типа куча, которые могут размещаться как на физическом устройстве, так и на хосте, причём вне зависимости от типа кучи, эта память всегда видна хосту.

Виды памяти типа куча:

* *DLHV heap* (*device local, host visible*) – память физически размещена на физическом устройстве, видна хосту
* *HLHV heap* (*host local, host visible*) – память размещена на хосте, видна и хосту и устройству

Приложение на *Vulkan* отправляет устройствам командный буфер, содержащий записанные команды для физического устройства. После того, как создан командный буфер, он может быть поставлен в очередь на исполнение неограниченное количество раз. Спецификация не определяет, что произойдёт, если отправить несколько командных буферов в несколько разных очередей – командные буферы могут исполниться параллельно или не в том порядке, в каком они были отправлены в очереди. Но в пределах одной очереди командные буферы выполняются в том порядке, в котором они были отправлены.

Выполнение командных буферов по отношению к приложению-хосту происходит асинхронно. О синхронизации между очередями, устройством и хостом должен заботиться хост.

Командные буферы могут находиться в нескольких состояниях:

* Первичное – в этом состоянии буфер находится, когда он только был создан
* Режим записи – буфер записывает все команды для исполнения устройством, которые были вызваны функциями вида *vkCmd\**
* Исполнение – буфер записал все нужные команды и готов быть переданным на исполнение в очередь
* В ожидании – буфер команд передан в очередь и исполняется устройством. В этом состоянии нельзя изменять буфер
* Ошибка – буфер в этом состоянии может быть только освобожден, сброшен на исходное состояние или включен в режим записи

##### Операции в очередях

Очереди в *Vulkan* предоставляют возможность запуска команд на устройстве. Команды записываются в командный буфер вместе с командой *Queue submission* (англ. - отправка в очередь) и запускаются на исполнение специальными движками исполнения. После того, как команды были отправлены в очередь, хост не может никак повлиять на работу этих команд. Вместе с командным буфером можно передать набор семафоров для того, чтобы сигнализировать устройству о том, что можно начать выполнение, или для того, чтобы устройство сообщило об окончании определённой операции.

Спецификация *Vulkan* не определяет порядок выполнения команд в разных очередях – если передать два буфера команд в две разные очереди, то нельзя заранее сказать, какой из них выполнится первым.

Порядок выполнения командных буферов можно контролировать с помощью примитивов синхронизации, которые подробнее описаны в пункте 2.1.7. Не смотря на асинхронную природу выполнения команд устройством, гарантируется, что все задачи, стоящие до примитива синхронизации, будут выполнены к тому моменту, когда примитив синхронизации отправит сигнал о синхронизации устройству.

Команды в очередях могут выполнять действия (рисовать, копировать память, очищать, начинать/останавливать подпроходы отрисовки), изменять состояние (состояние подпрохода отрисовки, отправка констант, привязка буферов к очередям, конвейерам) или производить операции синхронизации.

Так как команды для устройства выполняются асинхронно, то управление может сразу же вернуться хосту после отправки командного буфера в очередь.

##### Объектная модель

Устройства, очереди и сущности *Vulkan* представлены в виде объектов. Все объекты на уровне *API* управляются через их дескрипторы. Всего существуют дескрипторы двух видов: управляемые и неуправляемые. Управляемые дескрипторы — это указатели на непрозрачные типы данных (типы данных, не объявленные в интерфейсах, скрытые от глаз). Обычно такие дескрипторы используются для перехвата вызовов *API* драйвером и чаще всего передаются в функции первым параметром. Спецификация *Vulkan* определяет, что каждый управляемый объект должен иметь один уникальный управляемый дескриптор в течение всего времени жизни объекта.

Неуправляемые дескрипторы – целочисленные 64 битные типа данных, применение которых может меняться в зависимости от реализации. Неуправляемые объекты могут иметь несколько неуправляемых дескрипторов, причём гарантируется, что уничтожение одного дескриптора не уничтожит другие дескрипторы.

Все объекты, созданные с помощью сущности логического устройства *VkDevice* видны и могут быть использованы только этим логическим устройством.

Объекты создаются с помощью функций вида *vkCreate\**, другие объекты создаются выделением памяти функцией вида *vkAllocate\**. Как только объект был создан, его структура становится иммутабельной(неизменяемой), но данные в структуре изменять не запрещено.

Предполагается, что создание и уничтожение объектов происходит нечасто, в отличии от выделения и освобождения памяти под объекты. Поэтому объектов, требующих явного выделения памяти рекомендуется использоваться пулы и кучи. Уничтожение объектов осуществляется функциями вида *vkDestroy\**, освобождение памяти – *vkFree\**.

Драйвер никак не контролирует время жизни объектов, перекладывая эту задачу на плечи хоста. Хост сам должен контролировать, чтобы объекты не уничтожались во время использования.

Существует также набор объектов, которые нельзя уничтожать при любых или определённых условиях.

Никакой командой ни при каких условиях нельзя уничтожать объекты *VkShaderModule* и *VkPipelineCache*

Нельзя уничтожать объект *VkPipelineLayout*, если какой-либо буфер команд находится в режиме записи

Если командный буфер находится в режиме ожидания, то никакой из следующих объектов, используемых командным буфером, не должен быть уничтожен: *VkEvent, VkQueryPool, VkBuffer, VkBufferView, VkImage, VkImageView, VkPipeline, VkSampler, VkDescriptorPool, VkFramebuffer, VkRenderPass, VkCommandBuffer, VkCommandPool, VkDeviceMemory, VkDescriptorSet.*

Если очередь выполняет команды, то следующие объекты, используемые в буферах команд, не должны быть уничтожены: *VkFence, VkSemaphore, VkCommandBuffer, VkCommandPool.*

Объекты типа *VkQueue* не могут быть уничтожены явно – они уничтожаются автоматически вместе с объектом устройства *VkDevice*, с помощью которого они были созданы

При уничтожении объекта пула, все объекты из этого пула уничтожаются неявным образом

Объект логического устройства *VkDevice* может быть уничтожен только тогда, когда все связанные с ним очереди находятся в режиме ожидания, т.е. не имеют команд в очереди, и уничтожены следующие объекты, связанные с очередями: *VkFence, VkSemaphore, VkEvent, VkQueryPool, VkBuffer, VkBufferView, VkImage, VkImageView, VkShaderModule, VkPipelineCache, VkPipeline, VkPipelineLayout, VkSampler, VkDescriptorSetLayout, VkDescriptorPool, VkFramebuffer, VkRenderPass, VkCommandPool, VkCommandBuffer, VkDeviceMemory.*

Объект физического устройства *VkPhysicalDevice* не должен уничтожаться явным образом, он уничтожается вместе с объектом *VkInstance* в конце цикла жизни программы

##### Механизмы синхронизации

Синхронизация доступа к ресурсам является первичной задачей приложения-хоста. Существует несколько неявных гарантий порядка запуска команд, однако в большинстве случаев это нужно делать самому, как и управление кэшем и другими оптимизациями.

*Vulkan* предоставляет 4 примитива синхронизации:

* Барьер/Забор (*Fence*) – используются для сообщения хосту о завершении команды
* Семафор – используется для контроля доступа к ресурсам из нескольких очередей
* События – используются для сообщения о наступлении какого-либо события из хоста или командного буфера.
* Конвейерный барьер (*Pipeline Barrier*) – используются для синхронизации внутри командных буферов, но только в одном месте за раз.

### Обзор функций, структур данных, утилит

В этой главе будет осуществлён обзор утилит, поставляемых вместе с *SDK Vulkan*, а также функции и структуры данных, необходимых для реализации вычислений на ГП с помощью технологии *Vulkan*. Так как основное предназначение библиотеки *Vulkan* это отрисовка графики и связанных с ней вычислений, очень многие функции будут не нужны.

##### Обзор утилит

В комплекте с *SDK Vulkan* поставляются следующие утилиты:

* *cube* и *cube\_32* – демонстрационная программа с текстурированным вращающимся кубом. Версии *x86-64* и *x86* соответственно
* *glslangValidator –* компилятор и валидатор. Проверяет валидность кода на *GLSL* и компилирует его в *SPIR-V*-совместимый байт-код
* *glslc –* компилятор *GLSL* в *SPIR-V*-совместимый байт-код
* *spirv-as* – компилирует удобочитаемый *SPIR-V* код в байт-код *SPIR-V*
* *spirv-cfg –* строит граф потока управления шейдера *SPIR-V*
* *spirv-cross –* транслирует байт-код *SPIR-V* в удобочитаемые файлы на шейдерных языках *GLSL*, *MSL* (экспериментальная версия), *C++* (экспериментальная версия).
* *spirv-dis –* дизассемблирует байт-код *SPIR-V* в удобочитаемый *SPIR-V* код
* *spirv-opt* – оптимизирует байт-код *SPIR-V*
* *spirv-remap* – сжимает и оптимизирует байт-код *SPIR-V*, удаляя неиспользуемые функции и отладочные символы
* *spirv-val –* валидатор байт-кода *SPIR-V*
* *via –* проверяет наличие и валидность установленного драйвера *Vulkan* и наличие и состояние компонентов *Vulkan*
* *vkreplay* – воспроизводит файл трассировки без необходимости в приложении, с которого была записана трассировка
* *vktrace* – трассировщик вызовов *API Vulkan*

##### Обзор структур данных

Так как библиотека *Vulkan* предоставляет *C API,* в ней нет классов и интерфейсов, только структуры. Структуры языка *С* используются как передаваемые в функции параметры и как возвращаемые типы. Названия структур, а также описание их полей и предназначения описаны в таблице 2.

Таблица 2 – Структуры библиотеки *Vulkan*

|  |  |  |  |
| --- | --- | --- | --- |
| Название структуры | Поля | Назначение поля | Назначение структуры |
| VkInstanceCreateInfo | VkStructureType sType | Тип структуры | Хранение и изменение информации о приложении |
| const void\* pNext; | 0 или указатель на следующую структуру |
| const char\* pApplicationName | Указатель на строку с названием приложения |
| uint32\_t applicationVersion | Номер версии приложения |
| const char\* pEngineName | Имя движка запуска |
| uint32\_t engineVersion | Номер версии движка запуска |
| uint32\_t apiVersion | Версия API Vulkan |
| VkDeviceCreateInfo | VkStructureType sType | Тип структуры | Параметры для создаваемого логического устройства |
| const void\* pNext | 0 или указатель на следующую структуру |
| VkDeviceCreateFlags flags | Зарезервировано |
| uint32\_t queueCreateInfoCount | Размер массива структур с информацией для создания очередей |
| const VkDeviceQueueCreateInfo\* pQueueCreateInfos | Массив структур с информацией для создания очередей |
| uint32\_t enabledLayerCount | Не используется |
| const char\* const\* ppEnabledLayerNames | Не используется |
| uint32\_t enabledExtensionCount | Количество расширений, которое нужно включить |
| const char\* const\* ppEnabledExtensionNames | Указатель на массив строк с именами расширений, которые нужно включить |
| const VkPhysicalDeviceFeatures\* pEnabledFeatures | 0 или указатель на структуру с описанием включенных |
| VkDeviceQueueCreateInfo | VkStructureType sType | Тип структуры | Информация о создаваемой очереди |
| const void\* pNext | 0 или указатель на следующую структуру |
| VkDeviceQueueCreateFlags flags | Зарезервировано |
| uint32\_t queueFamilyIndex | Индекс семейства очередей у устройства |
| uint32\_t queueCount | Количество очередей для создания |
| const float\* pQueuePriorities | Массив чисел с плавающей запятой с приоритетами для каждой создаваемой очереди |
| VkPhysicalDeviceFeatures | 55 полей типа vkBool32 | Флаги поддержки особенностей | Поддерживаемые ГП особенности |
| VkPhysicalDeviceProperties | uint32\_t apiVersion | Версия API Vulkan | Свойства физического устройства - ГП |
| uint32\_t driverVersion | Номер драйвера |
| uint32\_t vendorID | Идентификатор производителя ГП |
| uint32\_t deviceID | Идентификатор устройства |
| VkPhysicalDeviceType deviceType | Тип устройства |
| char deviceName[VK\_MAX\_PHYSICAL\_DEVICE\_NAME\_SIZE] | Название устройства |
| uint8\_t pipelineCacheUUID[VK\_UUID\_SIZE] | Уникальный идентификатор устройства |
| VkPhysicalDeviceLimits limits | Физические ограничения устройства |
| VkPhysicalDeviceSparseProperties sparseProperties | Свойства разреженной памяти устройства |
| VkPhysicalDeviceLimits | 105 полей разных типов | Физические ограничения и размеры различных параметров физического устройства |  |
| VkPhysicalDeviceMemoryProperties | uint32\_t memoryTypeCount | Количество типов памяти | Свойства памяти физического устройства ГП |
| VkMemoryType memoryTypes[VK\_MAX\_MEMORY\_TYPES] | Массив структур, описывающих доступные типы памяти |
| uint32\_t memoryHeapCount; | Количество куч памяти |
| VkMemoryHeap memoryHeaps[VK\_MAX\_MEMORY\_HEAPS] | Массив структур, описывающих кучи памяти |
| VkMemoryHeap | VkDeviceSize size | Размер области памяти типа куча | Информация о куче памяти |
| VkMemoryHeapFlags flags | Флаги памяти в виде битовой маски |
| VkMemoryType | VkMemoryPropertyFlags propertyFlags | Битовая маска с параметрами | Информация о типе памяти |
| uint32\_t heapIndex | Куче с каким индексом соответствует эта структура |

##### Обзор функций и команд

Для задач вычислений на ГП, основной интерес представляют следующие команды и функции:

* *VkResult vkAllocateMemory(VkDevice, const VkMemoryAllocateInfo\*, const VkAllocationCallbacks\*, VkDeviceMemory\*)* – выделяет память заданного размера на заданном устройстве
* *VkResult vkBeginCommandBuffer(VkCommandBuffer, const VkCommandBufferBeginInfo\*)* – переключает заданный командный буфер в режим записи команд для устройства
* *void vkCmdCopyBuffer(VkCommandBuffer commandBuffer, VkBuffer srcBuffer, VkBuffer dstBuffer, uint32\_t regionCount, const VkBufferCopy\*)* – записывает в командный буфер команду копирования содержимого одного буфера в другой
* *void vkCmdFillBuffer(VkCommandBuffer commandBuffer, VkBuffer dstBuffer, VkDeviceSize dstOffset, VkDeviceSize size, uint32\_t data)* – записывает в командный буфер команду заполнения заданного буфера фиксированными данными
* *VkResult vkCreateBuffer(VkDevice device, const VkBufferCreateInfo\* pCreateInfo, const VkAllocationCallbacks\* pAllocator, VkBuffer\* pBuffer)* – создаёт новый объект буфера
* *VkResult vkCreateDevice(VkPhysicalDevice physicalDevice, const VkDeviceCreateInfo\* pCreateInfo, const VkAllocationCallbacks\* pAllocator, VkDevice\* pDevice)* – создаёт новый объект логического устройства
* *VkResult vkCreateFence(VkDevice device, const VkFenceCreateInfo\* pCreateInfo, const VkAllocationCallbacks\* pAllocator, VkFence\* pFence)* – создаёт новый примитив синхронизации типа барьер
* *VkResult vkCreateInstance(const VkInstanceCreateInfo\* pCreateInfo, const VkAllocationCallbacks\* pAllocator, VkInstance\* pInstance)* – создаёт объект экземпляра приложения *Vulkan*
* *void vkDestroyBuffer(VkDevice device, VkBuffer buffer, const VkAllocationCallbacks\* pAllocator)* – уничтожает буфер и все данные в нём
* *void vkDestroyDevice(VkDevice device, const VkAllocationCallbacks\* pAllocator)* – уничтожает объект логического устройства
* *void vkDestroyFence(VkDevice device, VkFence fence, const VkAllocationCallbacks\* pAllocator)* – уничтожает примитив синхронизации типа барьер
* *void vkDestroyInstance(VkInstance instance, const VkAllocationCallbacks\* pAllocator)* – уничтожает объект экземпляра программы *Vulkan*
* *VkResult vkEnumeratePhysicalDevices(VkInstance instance, uint32\_t\* pPhysicalDeviceCount, VkPhysicalDevice\* pPhysicalDevices)* – перечисляет все существующие в системе физические устройства
* *void vkFreeMemory(VkDevice device, VkDeviceMemory memory, const VkAllocationCallbacks\* pAllocator)* – освобождает память, выделенную на устройстве
* *VkResult vkResetCommandBuffer(VkCommandBuffer commandBuffer, VkCommandBufferResetFlags flags)* – сбрасывает командный буфер в исходное состояние
* *VkResult vkCreateShaderModule(VkDevice device, const VkShaderModuleCreateInfo\* pCreateInfo, const VkAllocationCallbacks\* pAllocator, VkShaderModule\* pShaderModule)* – создаёт объект модуля шейдера
* *VkResult vkCreateComputePipelines(VkDevice device, VkPipelineCache pipelineCache, uint32\_t createInfoCount, const VkComputePipelineCreateInfo\* pCreateInfos, const VkAllocationCallbacks\* pAllocator, VkPipeline\* pPipelines)* – создаёт объект вычислительного конвейера для запуска вычислительных шейдеров

Общие шаги для создания, инициализации приложения *Vulkan* и запуска вычислительного шейдера:

1. Создать объект *VkInstance* с нужными параметрами с помощью *vkCreateInstance()*
2. Просмотреть существующие устройства ГП и получить идентификатор нужного с устройства *VkPhysicalDevice*
3. Создать объект логического устройства *VkDevice* с помощью метода *vkCreateDevice()*
4. Получить объект очереди VkQueue от устройства с помощью функции *vkGetDeviceQueue()*
5. Создать модуль шейдера *VkShaderModule* с помощью функции *vkCreateShaderModule(),* передав в него шейдерный код
6. Создать вычислительный конвейер *VkPipeline* с помощью метода *vkCreateComputePipelines*
7. Выделить память, записать туда данные
8. Создать командный буфер, привязать к конвейеру с помощью *vkCmdBindPipeline()*
9. Записать команду на запуск шейдера с помощью *vkCmdDispatch()*
10. Запустить командный буфер на исполнение, отправив в очередь с помощью *vkQueueSubmit()*

## Проектирование

### Положение библиотеки Vulkalc в иерархии вызовов

Библиотека Vulkalc является промежуточным звеном между пользовательским приложением, которое занимается вычислениями и средой исполнения Vulkan. Важно отметить, что библиотека не является высокоуровневой обёрткой над вызовами Vulkan, а предоставляет собственный интерфейс для удобной работы с вычислениями, скрывая непосредственную работу со средой выполнения Vulkan внутри себя.

Рисунок 3.1 демонстрирует положение библиотеки Vulkalc в иерархии вызовов функций Vulkan и работы с графическим процессором.

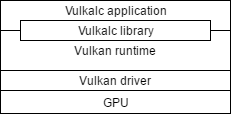


Рис. 3.1. Иерархия вызовов функций Vulkan

### Сущности Vulkalc

Библиотека Vulkalc экспортирует одно пространство имён Vulkalc для предотвращения пересечения имён. В этом пространстве имён содержатся классы, предоставляющие доступ ко всем функциям библиотеки. На рисунке 3.2 изображена структурная диаграмма сущностей.

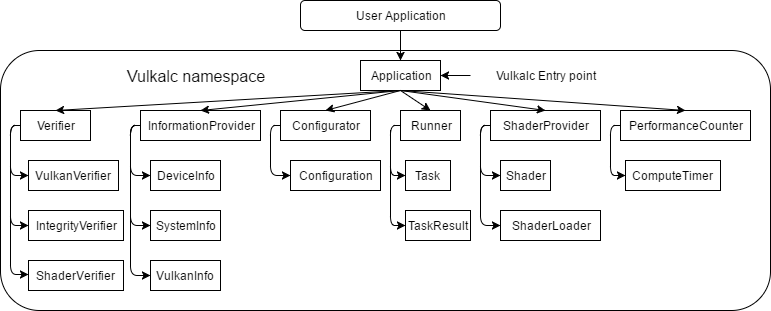


Рис 3.2. Структурная диаграмма сущностей

Точкой входа в библиотеку и предоставляемые библиотекой классы и функции является класс Application. Далее описаны все сущности и их назначение:

1. *Configurator* – конфигурирует приложение информацией о слоях валидации Vulkan, используемом графическом процессоре и другой информацией
   1. *Configuration* – содержит информацию, используемую Configurator для конфигурации приложения
2. *InformationProvider* – предоставляет информацию о различных компонентах системы, от которых зависит работоспособность и производительность библиотеки и вычислений
   1. *DeviceInfo* – информация о графическом процессоре, например, частота графического процессора, объём видеопамяти и другие.
   2. *SystemInfo* – информация о версии ОС, количестве установленных видеокарт, объёме оперативной памяти, частоте процессора
   3. *VulkanInfo* – информация о Vulkan – версия, включенные расширения и слои валидации
3. *PerformanceCounter* – счётчики производительности
   1. *ComputeTimer* – счётчик времени вычисления определённой задачи
4. *Runner* – сущность, контролирующая запуск задач на вычисление и получение результатов
   1. *Task* – сущность, выражающая собой задачу на вычисление, содержащая буфер входных данных и выбранный вычислительный шейдер
   2. *TaskResult* – результат вычисления последней задачи, содержит буфер выходных данных
5. *ShaderProvider* – предоставляет доступ к пользовательским шейдерам и шейдерам, идущим в комплекте в библиотекой, предоставляет логическую сущность шейдера
   1. *Shader* – вычислительный шейдер, загруженный с диска
   2. *ShaderProvider* – загрузчик шейдеров с диска
6. *Verifier* – предоставляет функции для различных проверок целостности и валидности
   1. *VulkanVerifier* – проверяет установленный Vulkan Runtime на совместимость, наличие необходимых слоёв валидации и расширений
   2. *IntegrityVerifier* – проверяет целостность распространяемого пакета библиотеки на наличие всех необходимых шейдеров
   3. *ShaderVerifier* – проверяет возможность компиляции загруженных шейдеров

### Описание процесса запуска задачи на исполнение

Библиотека Vulkalc предоставляет единый процесс загрузки и запуска шейдеров и для пользовательских шейдеров, и для предоставляемых библиотекой шейдеров. Процесс загрузки шейдеров продемонстрирован на рисунке 3.3.

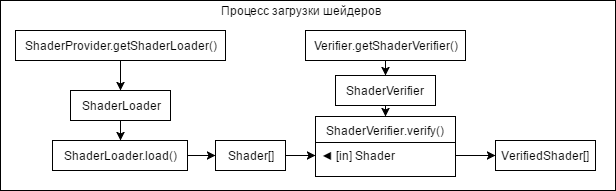


Рис. 3.3. Процесс загрузки шейдеров

Загрузка шейдеров состоит из следующих этапов:

1. Получаем экземпляр загрузчика шейдеров *ShaderLoader* с помощью метода *getShaderLoader* класса *ShaderProvider;*
2. Вызываем у экземпляра класса *ShaderLoader* метод *load*, который возвращает массив сущностей шейдера *Shader;*
3. Получаем экземпляр класса *ShaderVerifier* с помощью метода *getShaderVerifier* класса *Verifier;*
4. Выполняем компиляцию шейдеров в промежуточный шейдерный байт-код *SPIR-V* с помощью метода verify экземпляра класса *ShaderVerifier*, который возвращает массив сущностей скомпилированных шейдеров *VerifiedShader*. Стоит отметить, что шейдеры, которые не удалось скомпилировать, не попадают в этот массив. Компиляция выполняется инструментом *glslangValidator*.

После того, как шейдеры были валидированы, скомпилированы и загружены, их можно использовать для запуска задач вычисления. Процесс запуска шейдеров на выполнение продемонстрирован на рисунке 3.4 и описан ниже.

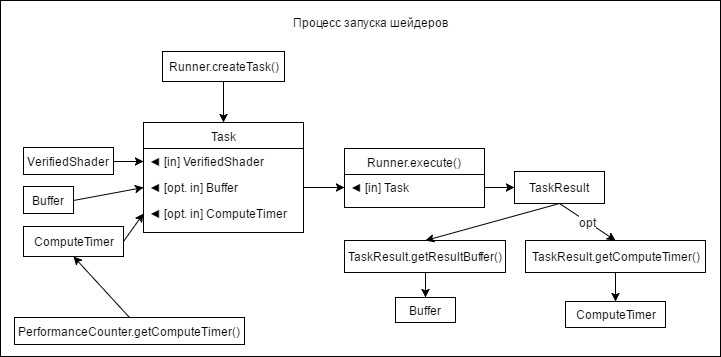


Рис. 3.4. Процесс запуска шейдера на исполнение

Этапы запуска шейдера на вычисление:

1. Создать задачу вычисления *Task* с помощью метода *createTask* класса *Runner*;
2. В задачу *Task* передать выбранный верифицированный и скомпилированный шейдер *VerifiedShader*. При необходимости передать буфер со входными данными *Buffer* и счётчик времени выполнения *ComputeTimer*, который можно получить с помощью метода *getComputeTimer* класса *PerformanceCounter;*
3. Передать задачу *Task* в метод execute класса *Runner;*
4. Получить из класса *Runner* результат вычисления *TaskResult;*
5. Если шейдер подразумевает возвращение каких-либо результатов, то их получить можно с помощью метода *getResultBuffer* у экземпляра класса *TaskResult;*
6. Если к задаче был прикреплён счётчик времени выполнения, то можно получить его с помощью метода *getComputeTimer* у экземпляра класса *TaskResult*.

## Реализация

## Тестирование

## Заключение

## Список используемых источников

## Приложение