Утверждаю
Директор Института СПИНТех
НИУ МИЭТ
Проф. _____/Гагарина Л.Г./
«____»_____ 2020 г.

Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Московский институт электронной техники»

Уманский Александр Александрович

Разработка программного модуля для анализа программ на языках C и C++ на недекларированные возможности

Специальность 09.03.04 —
«Программная инженерия»

Отчет по производственной практике
студента института СПИНТЕХ

Научный руководитель:
кандидат технических наук, доцент
Кононова Александра Игоревна

_____

Студент:
Уманский Александр Александрович

_____

Москва, г. Зеленоград — 2020

Содержание

Стр.

Список сокращений и условных обозначений

НДВ	Недекларированные возможности
ПМ	Программный модуль
БД	База Данных
ИСПДН	Информационная система персональных данных
ПО	Программное обеспечение
IDE	Интегрированная среда разработки
JSON	Формат описания структур данных в текстовом виде ключ $\rightarrow$ значение
PID	Уникальный идентификатор процесса в ОС
ПМ АПНДВ	Программный модуль анализа на недекларированные возможности

Словарь терминов

Кросс-платформенный : Программа, которая может запускаться на различных операционных системах и/или архитектурах процессоров

Программная закладка : Подпрограмма, либо фрагмент исходного кода, скрытно внедренный в исполняемый файл

Динамическая трасса : Дерево вызванных программой функций во время конкретного ее исполнения

Статическая трасса : Дерево функций программы, которые объявлены для вызова

Отладчик : Программа, в контексте которой запускается другая программа для тестирования в контролируемых условиях

Отладка : Процесс тестирования программы в контролируемых условиях

Удаленная отладка : Процесс отладки программы, запущенной вне контекста отладчика

Препроцессор : Программа-макропроцессор, обрабатывающая специальные директивы в исходном коде и запускающаяся до компилятора

Препроцессирование : Процесс обработки исходного кода препроцессором

Открытое ПО : ПО с открытым исходным кодом, который доступен для просмотра, изучения и изменения

Сериализация : Процесс перевода определеного типа данных программы в некоторый формат

Десериализация : Процесс перевода данных, находящихся в некотором формате, во внутренний тип данных программы

Скрипт : Программа, обычно на интерпретируемом языке программирования, выполняющая конкретное действие

Сигнатура функции : Объявление функции, в которое входит имя функции, количество входных параметров и их тип

Введение

Сертификация – процесс подтверждения соответствия характеристик товара определенным стандартам.

Сертификация не является универсальным способом решения всех существующих проблем в области информационной безопасности, однако сегодня это единственный реально функционирующий механизм, который обеспечивает независимый контроль качества средств защиты информации. При грамотном применении механизм сертификации позволяет вполне успешно решать задачу достижения гарантированного уровня защищенности автоматизированных систем.

Отсутствие недекларированных возможностей в скомпилированном объектном файле является ключевым аспектом сертификации ПО. Сертификация программного обеспечения необходима для подтверждения требований заказчика к защите информации, к выполнению функциональных и технических задач и к обеспечению работы ПО в целом.

Но существуют ненапрасные опасения, что на любом из этапов сборки программы из исходных кодов в ней может появиться программная закладка [1; 2]

Чтобы подобные ситуации исключить, применяется техника статичического анализа исходных кодов, динамического анализа – анализа пройденных программой трасс и последующее сравнение результатов обоих анализов.

Существуют достойные [3] разработки в области открытого программного обеспечения по проведению, например, статического анализа кода или получению динамических трасс, но не существует комбинированных решений, позволяющих провести данный процесс валидации ПО. Для каждого конкретного проекта приходится использовать различные статические анализаторы кода и динамические анализаторы исполнения программ, что приводит к переписыванию, по своей сути, кода, который проводит итоговый анализ ПО.

Целью данной работы является ускорение проведения статичического и динамического анализа программ написанных на языках программирования C/C++.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. анализ предметной области;
2. выбор используемых технологий;

3. разработка алгоритма работы программы;
4. разработка структур данных;

Практическая значимость проекта состоит в ускорении процесса сертификации ПО на отсутствие НДВ.

Полный объём отчета составляет 37 страниц, включая 8 рисунков и 10 таблиц. Список литературы содержит 49 наименований.

Раздел 1. Исследовательский раздел

Сертификация программного обеспечения проводится, когда необходимо подтвердить соответствие разрабатываемой продукции требованиям защиты информации.

1.1 Процесс сертификации ПО на отсутствие НДВ

Сертификационная процедура состоит из следующих этапов:

1) готовность документации ПО, доступность исходных текстов;
2) опредление объема исходных текстов, подлежащих анализу;
3) обращение заявителя в испытательную лабораторию с собранной информацией;
4) анализ документации;
5) разработка «Программы и методик проведения сертификационных испытаний»;
6) проведение испытаний;
7) экспертиза результатов.

Сертификация должна выявить присутствие в исполняемом файле недекларированных возможностей, которые могут являться как злым умыслом[1; 2] разработчиков компилятора, линкера и других вспомогательных программ, так и методами оптимизации ПО, которые применяются для более рационального потребления ресурсов программой.

Выявить данные расхождения между необработанными исходными кодами и поведением программы во время исполнения позволяет разработанный мной программный модуль.

Дадим определение термину «Недекларированные возможности»:

Недекларированные возможности[4] — намеренно измененная часть ПО, с помощью которой можно получить незаметный несанкционированный доступ к безопасной компьютерной среде.

1.2 Классификация НДВ

Классифицировать НДВ можно по несколькими способами, в зависимости от их целей и применения.

1.2.1 По применению

Использование НДВ может реализовываться в:

– перехвате данных;

– подмене данных;

– выводе компьютерной системы из строя;

– полном доступе к удаленной компьютерной системе.

Причем, при полном доступе к компьютерной системе, вредоносные программы программы могут быть использованы злоумышленниками для всех вышеперечисленных целей.

1.2.2 По целям

Использование НДВ может быть направлено на:

– **Персональные компьютеры и рабочие станции**
Целью могут быть как персональные компьютеры широкого числа пользователей, так и отдельные рабочие станции, которые могут являться точкой входа в защищенную компьютерную систему, так и использоваться для перехвата важной информации;

– **Серверы**
Серверы обслуживают большое количество клиентов, а значит проникновение на сервер может существенно повлиять на работу всех компьютеров, работающих с данным сервером;

– **Встраиваемые системы**
Благодаря постоянному удешелению микроконтроллеров и периферийных устройств, все больше и больше повседневных вещей обзаводятся «умной» функциональностью. Погоня производителей за прибылями отражается на безопасности прошивок умных устройств;

– **Промышленные компьютеры**
Программные закладки в такие системы чреваты шпионажем или диверсией [5] [1].

[1]Хотя данная программа является вирусом, а не программой с НДВ, случившееся ярко показывает реальное применение подобных техник для деструктивных действий

1.3 Степень опасности НДВ

Для определения опасности НДВ будем пользоваться следующими нормативными документами [6]:

– приказ ФСТЭК России от 18 февраля 2013 г. № 21;

– федеральный закон "О персональных данных"от 27.07.2006 N 152-ФЗ.

Тип угроз безопасности персональных данных определяется в зависимости от комбинаций критичности угроз в ИСПДн (табл. 1):

– наличием НДВ в системном программном обеспечении (ПО), используемом в ИСПДн;

– наличием НДВ в прикладном ПО, используемом в ИСПДн.

Таблица 1 — Тип актуальных угроз

Угрозы	Тип актуальных угроз		
	1 Тип	2 Тип	3 Тип
Наличие НДВ в системном ПО, используемом в ИСПДн	критично	некритично	некритично
Наличие НДВ в прикладном ПО используемом в ИСПДн	критично или некритично	критично	некритично

Порядок определения актуальных угроз безопасности персональных данных в ИСПДн осуществляется в соответствии с Методикой определения актуальных угроз безопасности персональных данных при их обработке в информационных системах персональных данных, утвержденных ФСТЭК России, 2008 год.

Актуальной считается угроза, которая может быть реализована в ИСПДн и представляет опасность для персональных данных. Подход к составлению перечня актуальных угроз состоит в следующем. Для оценки возможности реализации угрозы применяются два показателя:

– Y_1 - уровень исходной защищенности ИСПДн;

– Y_2 - частота (вероятность) реализации рассматриваемой угрозы;

Коэффициент реализуемости угрозы Y определяется соотношением:

$$Y = \frac{Y_1 + Y_2}{20}$$

По значению коэффициента реализуемости угрозы Y интерпретация реализуемости угрозы следующим образом:

- если $0 \leqslant Y \leqslant 0.3$, то возможность реализации угрозы признается низкой;
- если $0.3 < Y \leqslant 0.6$, то возможность реализации угрозы признается средней;
- если $0.6 < Y \leqslant 0.8$, то возможность реализации угрозы признается высокой;
- если $Y > 0{,}8$, то возможность реализации угрозы признается очень высокой.

Далее оценивается опасность каждой угрозы. Этот показатель имеет три значения:

- низкая опасность – если реализация угрозы может привести к незначительным негативным последствиям для субъектов персональных данных;
- средняя опасность – если реализация угрозы может привести к негативным последствиям для субъектов персональных данных;
- высокая опасность – если реализация угрозы может привести к значительным негативным последствиям для субъектов персональных данных.

Затем осуществляется выбор из общего перечня угроз безопасности тех, которые относятся к актуальным для данной ИСПДн, в соответствии с правилами, приведенными в табл. 2

Таблица 2 — Правила отнесения угрозы безопасности персональных данных к критичной

Возможность реализации угрозы	Показатель опасности угрозы		
	Низкая	Средняя	Высокая
Низкая	некритичная	некритичная	критичная
Средняя	некритичная	критичная	критичная
Высокая	критичная	критичная	критичная
Очень высокая	критичная	критичная	критичная

После чего выносится решение о проведении анализа ПО на НДВ в процесс сертификации или его игнорирование, как некритичного.

Сейчас анализ программы на НДВ происходит вручную:

1) с помощью специального ПО проводят статический анализ исходных кодов программного проекта;
2) с помощью отладчиков или эмуляторов проводят динамический анализ исполняемого файла, сохраняя трассы выполнения;
3) данные статического и динамического анализа приводятся к общему виду;
4) с помощью программы сравнения ищутся несовпадения или их отсутствие.

1.4 Обзор программных решений для сертификации ПО на отсутствие НДВ

На сегодняшний день не существует в комплексных разработок по сертификации программного обеспечения на предмет НДВ. Однако, существуют программы, специализирующиеся отдельно на анализе исходных кодов и отдельно исполняемого файла. Так как ПМ АПНДВ будет совмещать и расширять функционал данных программных средств, то рассмотрим их по отдельности.

1.4.1 Сравнение статических анализаторов

Таблица 3 — Сравнительная таблица статических анализаторов

Свойства \ Название программы	Microsoft Application Inspector [7]	SCI Tools Understand [8]	GNU cflow [9]
Кросс-платформенность	Да	Да	Да
Открытость исходного кода	Да	Нет	Да
Препроцессирование кода C/C++	Нет	Да	Да
Представление препроцессорных директив как вызов функций	Нет	Нет	Да
Создание графа вызовов	Нет	Да	Да
Создание обратного графа вызовов	Нет	Да	Да
Бесплатность	Да	Нет	Да
Графический интерфейс	Нет	Есть	Нет

Microsoft Application Inspector

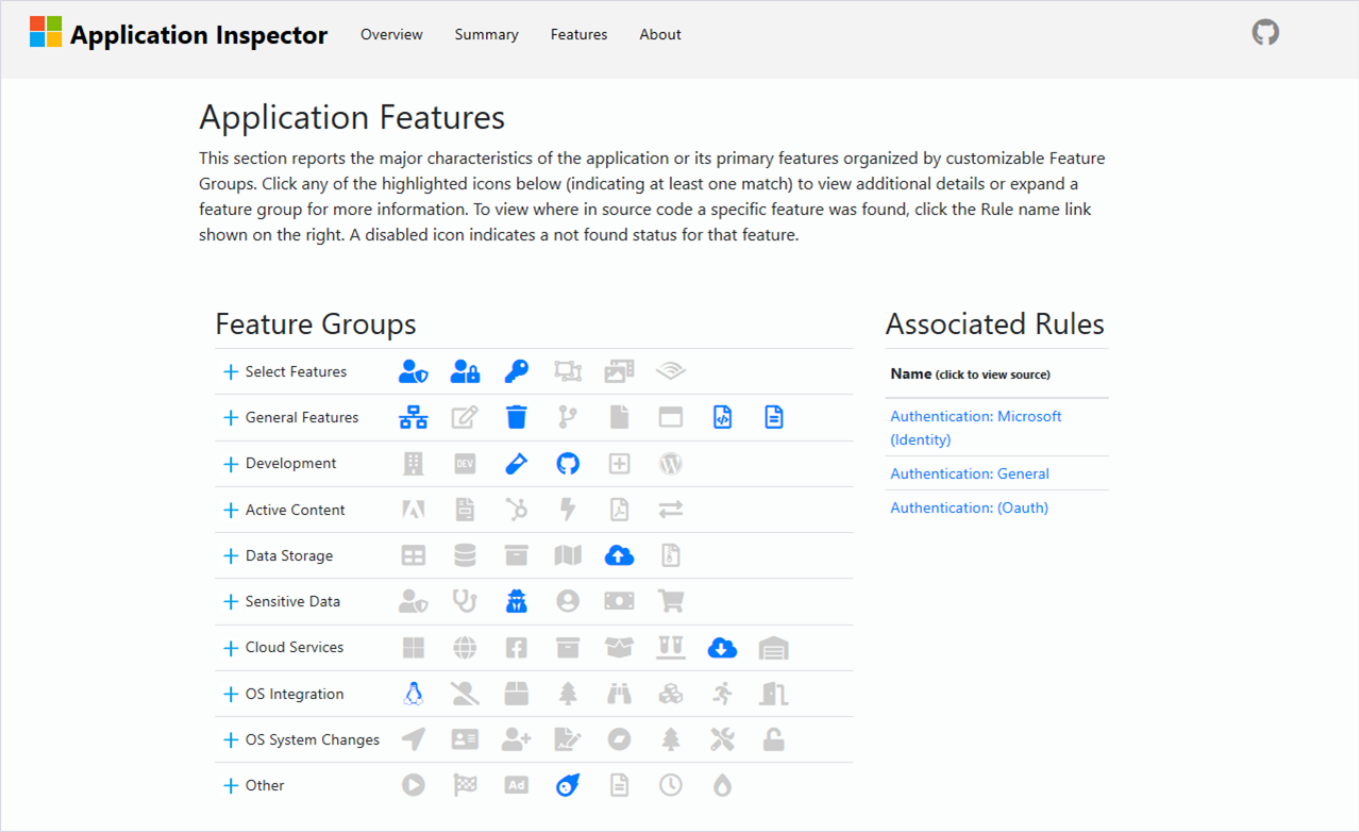


Рисунок 1.1 — Отчет Microsoft Application Inspector

Задача Microsoft Application Inspector – Систематическая и масштабируемая идентификация функций исходного кода. Анализатор написан на .NET Core [10], а это значит, что программа будет работать на всех платформах, для которых реализован .NET Core: Windows, Linux и macOS.

Распознает паттерны не только в 34 языках, но так же и в их смешениях – когда взаимодействующие части программы написаны на разных языках. Является бесплатным ПО, с открытым исходным кодом. В качестве недостатков для использования можно отметить предметный анализ функций, который ничего не говорит о последовательности их вызова.

SCI Tools Understand

SCI Tools Understand – кросс-платформенный, быстрый статический анализатор больших объемов кода, имеющий хорошие возможности в визуализации отношений модулей программы, имеет встроенный расчет различных метрик программного кода.

Поддерживает около 20 языков программирования, а так же распознает различные их редакции. Недостатки SCI Tools Understand – платность и закрытый исходный код. Но купив лицензионную копию программы, пользователь получает возможность писать скрипты манипуляции БД анализируемого проекта, генерирования отчетов и собственных метрик,

GNU cflow

Быстрый и минималистичный статический анализатор, с открытым исходным кодом, позволяющий создавать как прямые, так и обратные графы вызовов. Командный интерфейс приближен к командному интерфейсу компилятора. Поддерживает языки C и C++, а так же LEX и YACC. К достоинствам так же можно отнести удобный и емкий формат отчета, который легко разбирать регулярными выражениями.

Вывод

Так как ПМ АПНДВ ориентирован на анализ C/C++ программ, то проанализировав табл. 3 приходим к выводу, что функционал Microsoft Application Inspector не покрывает нужные сценарии использования, а SCI Tools Understand не подходит из-за своей закрытости и платности. Единственный возможнный выбор – GNU cflow.

1.4.2 Сравнение динамических анализаторов

GNU Debugger

Отладчик GDB впервые увидел свет в 1986 году и за прошедшие годы обзавелся большим количеством поддерживаемых архитектур процессоров, самые известные:

- Alpha;
- ARM;
- AVR;
- H8/300;
- Altera Nios/Nios II;
- System/370;
- System 390;
- X86 и X86-64;

Таблица 5 — Сравнительная таблица программ для динамического анализа

Свойства \ Название программы	GDB [11]	QEMU [12]
Кросс-платформенность	Да	Да
Открытость исходного кода	Да	Да
Возможность анализировать память	Да	Да
Возможность программно управлять	Да	Да
Возможность создавать собственные команды	Да	Нет
Возможность удаленной отладки	Да	Нет
Бесплатность	Да	Да
Графический интерфейс	Есть	Есть

– IA-64 "Itanium";
– Motorola 68000;
– MIPS;
– PA-RISC;
– PowerPC;
– SuperH;
– SPARC;
– VAX.

К достоинствам можно отнести возможность описания сценария отладки в командном файле, с последующим исполнением его GDB, а так же удаленную отладку.

QEMU

QEMU – Быстрый эмулятор процессоров, поддерживает множество процессорных архитектур, предоставляет возможность сохранять сгенерированный машинный код. К недостаткам можно отнести медленную, по сравнению с отладчиком работу, так как эмулятору приходится преобразовывать каждую инструкцию запущенной программы в машинный код процессора, на котором он запущен.

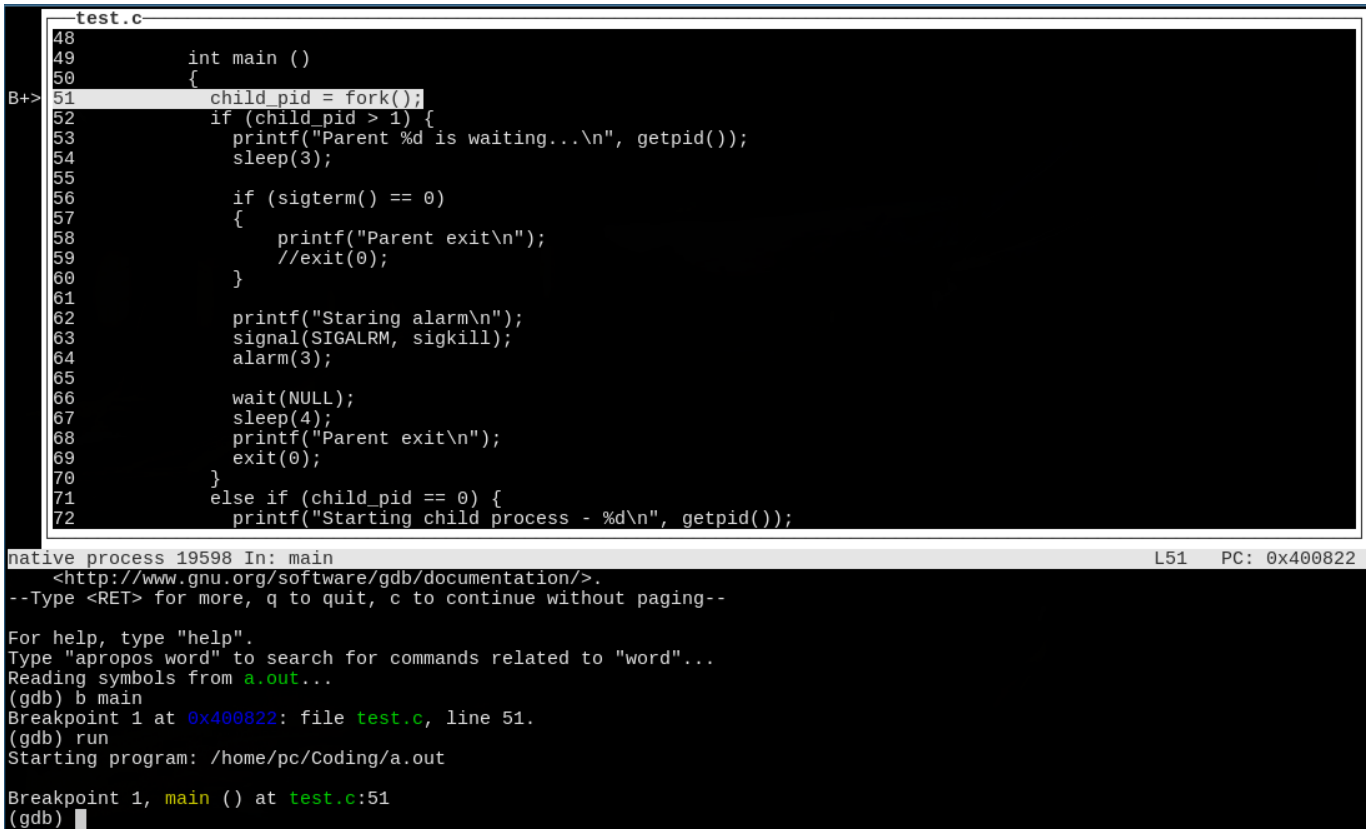


Рисунок 1.2 — Терминальный интерфейс GDB

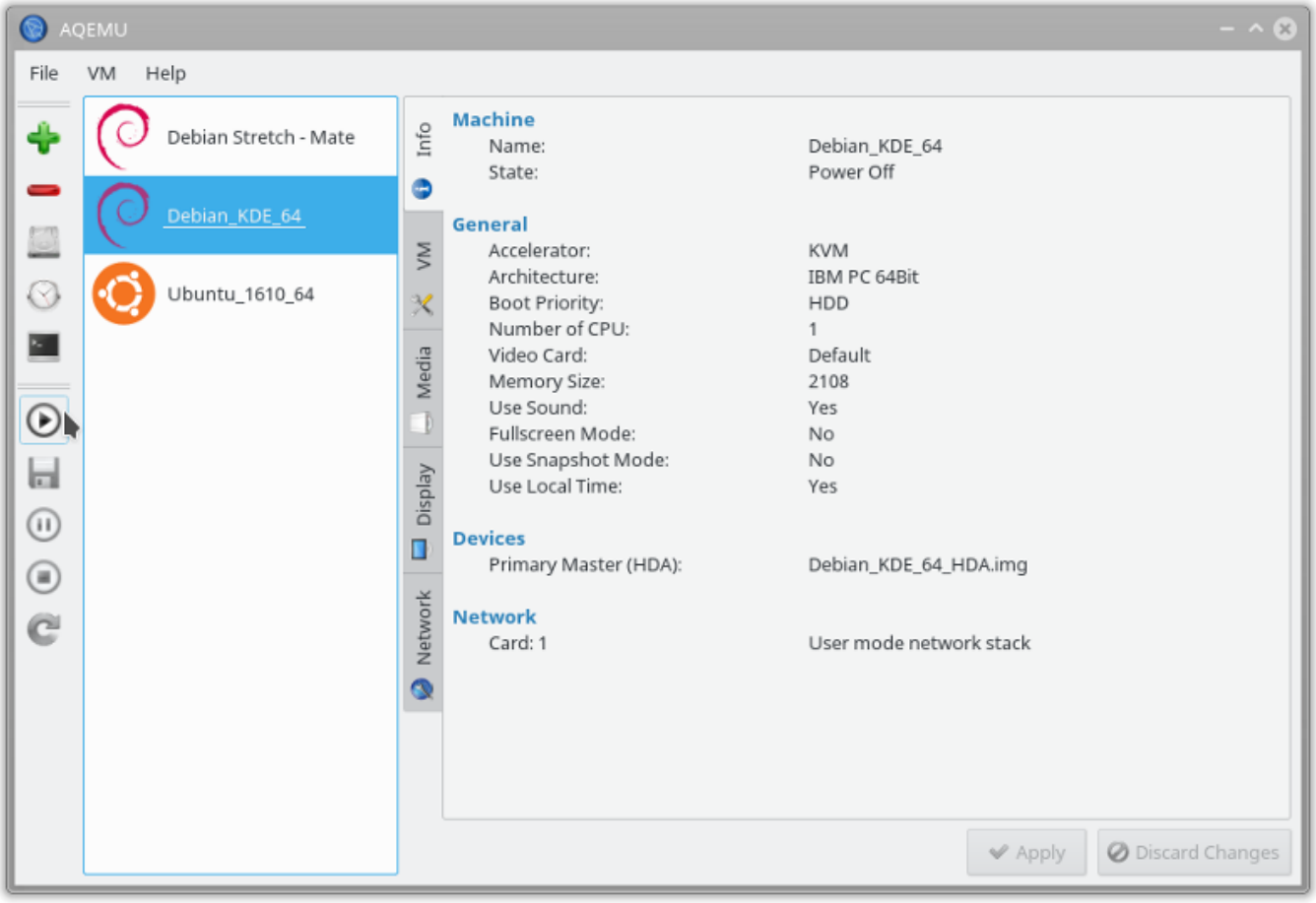


Рисунок 1.3 — Графическая оболочка AQEMU для эмулятора QEMU

Вывод

Так как для более точного выполнения задачи сертификации будет полезно получать информацию времени выполнения программы, такую, как:

1) значения регистров перед вызовом функции;
2) состояние стека перед вызовом функции;
3) стек вызовов;
4) экспертиза результатов;
5) информацию о сегментах и функциях в них определенных.

А так же расширять возможность динамического анализатора с помощью скриптов, то из табл. 5 следует, что удобнее всего это можно будет сделать с помощью отладчика GDB [11], нежели эмулятора QEMU [12].

1.5 Постановка задачи ВКР

На основе изложенного в разд. 1.1-разд. 1.4 сформированы следующие цели и задачи ВКР. Цель: сокращение времени проведения сертификации программного обеспечения, написанного на языках C и C++.

Задачи:

1) исследование предметной области (рассмотрено в разд. 1.1);
2) сравнительный анализ существующих программных решений (рассмотрено в разд. 1.4.1-разд. 1.4.2);
3) выбор языка и среды разработки;
4) разработка схемы данных ПМ АПНДВ;
5) разработка схемы алгоритма ПМ АПНДВ;
6) программирование ПМ АПНДВ;
7) отладка и тестирование ПМ АПНДВ;
8) разработка документации к ПМ АПНДВ.

Выводы по разделу

В исследовательском разделе обоснована актуальность разработки ПМ АПНДВ. Исследована предметная область и проведен анализ возможных решений, из которых был сделан вывод о том, что программы, аналогичной по функционалу ПМ АПНДВ нет на рынке. Так же поставлены задачи для дальнейшей разработки ПМ АПНДВ.

Раздел 2. Конструкторский раздел

2.1 Обоснование выбора языка программирования и среды разработки

Для удобной, быстрой и эффективной, как по срокам выполнения, так и по качеству конечного продукта, разработки ПМ АПНДВ потребуются правильные инструменты – язык программирования, на котором легче всего описать решение данной задачи и среда разработки, не только поддерживающая данный язык, но и позволяющая эффективно с ним работать.

2.1.1 Сравнение языков программирования

Для разработки ПМ АПНДВ понадобится сверхвысокоуровневый язык с кросс-платформенной стандартной библиотекой, который позволит точно и лаконично описать этапы анализа, а так же имеющий высокую скорость исполнения, для анализа больших объемов исходного кода и исполняемых файлов.

Рассмотрим подробно каждый из представленных в таблице языков.

Таблица 7 — Сравнительная таблица языков программирования

Язык программирования / Свойства	Nim [13]	Python [14]	Perl [15]	C/C++
Сверхвысокоуровневость	Да	Да	Да	Нет
Компилируется в машинный код	Да	Нет	Нет	Да
Количество функции в стандартной библиотеке	5585	638	1338	1224
Портируемость	Есть	Есть	Есть	Есть, но неудобная
Встроенная генерация документации	Есть	Есть	Есть	Нет
Статическая типизация	Есть	Нет	Нет	Есть
Автоматическое управление памятью	Есть	Есть	Есть	Есть
Обобщенное программирование	Есть	Есть	Есть	Есть
Мета-программирование	Есть	Есть	Есть	Есть
Опыт использования	Есть	Есть	Нет	Есть

C++

Мультипарадигменный высокоуровневый язык программирования, разработанный в 1983 году Бьёрном Страуструпом. Является практически полным надмножеством языка C. Статически типизирован.
Отличается высокой производительностью и неплохой гибкостью при написании кода. К минусам языка можно отнести сложность освоения и перегруженность «наследием» 80-х годов прошлого века, а так же низкую скорость компиляции, по сравнению с предшественником – C.
Портируемость языка на различные платформы обеспечивается пере- или кросс-компиляцией исходного кода под нужную платформу.

Python

Мультипарадигменный сверхвысокоуровневый язык программирования, разработанный в 1991 году Гвидо Ван Россумом. Является интерпретируемым языком, имеет слабую динамическую типизацию, что позволяет легко писать обобщенный код и использовать мета-программирование, но так же ведет к трудноулавливаемым ошибкам. Негативное влияние можно сгладить с помощью указания типов при объявлении перемнных и аргументов функций, а так же программы, проверяющей эти типы – линтера. Например pylint [16] или pyflakes [17].
Благодаря своей популярности, python так же портирован на большое количество платформ. Большим плюсом языка является его обширная стандартная библиотека, позволяющая легко писать комплексные приложения, не прибегая к установке дополнительных библиотек – такие программы, как и сам python, следуют философии «в комплекте с батарейками» («batteries included» [18]), суть которой заключается в самодостаточности программ. Помимо этого вместе с python поставляется менеджер пакетов pip [19], позволяющий удобно устанавливать требуемые библиотечные модули вместе с зависимостями.
К минусам языка можно отнести медлительность эталонного интерпретатора языка – cpython [20]. Код, исполняемый им, в определенных задачах медленнее кода на C в сотни раз. Не смотря на то, что есть более быстрые интерпретаторы: PyPy [21], Jython [22], Iron Python [23], они не смогут достичь скорости исполнения программ, компилируемых в машинный код.

На данный момент существует две, между собой несовместимые, версии языка: python 2, поддержка которого закончилась 1 января 2020 г. и python 3.

Perl

Мультипарадигменный сверхвысокоуровневый язык программирования, разработанный в 1987 году Ларри Уоллом. Является интерпретируемым языком, имеет слабую динамическую типизацию.
Полное название языка – «Practical Extraction and Report Language» («Практический Язык для Извлечения Данных и Составления Отчётов»), отражает его суть: в языке реализованы обширные возможности для работы с текстом, в синтаксис интегрированы регулярные выражения, как и в языках, которые оказали на него наибольшее влияние – AWK [24] и sed [25]. Но это же и я является его слабой частью, так как Perl скорее предназначен для однострочных команд в терминале, как AWK и sed.

Nim

Мультипарадигменный сверхвысокоуровневый язык программирования, разработанный в 2004 году Андреасом Румпфом. Является компилируемым языком, имеет строгую статическую типизацию.
Заметно, что на синтаксис языка повлиял Python, что сделало его выразительным и понятным. Язык использует промежуточную компиляцию, которая несколько замедляет процесс компиляции программ, но позволяет запускать nim-программы на различных платформах. На данный момент поддерживается компиляция в JavaScript [26] и оптимизированный C-код с несколькими моделями управления памятью:

- Сборщики мусора, основанные на:
    1) подсчете ссылок;
    2) подсчете ссылок с оптимизацией move-семантикой [27];
    3) boehm [28];
    4) go [29];
- ручном освобождении памяти;
- модель, в которой вся выделенная память высвобождается только по завершению программы (не рекомендуется к использованию).

Компиляции Nim в C означает не только высокую скорость работы, но и прозрачный программный интерфейс при взаимодействии с C библиотеками. Это значит, что можно писать Nim-код, взаимодействующий с C библиотекой так же, как если бы это была Nim-библиотека, в отличие от, например, Python.
Так же вместе с компилятором языка поставляется пакетный менеджер nimble [30] и генератор документации из комментариев, написанных на reStructuredText [31].

Вывод

Из всего вышесказанного следует, что для ПМ АПНДВ лучше всего подойдет язык Nim благодаря его скорости, выразительности и портируемости на различные платформы. Кроме того, для подготовки динамического анализа программы будут использованы утилиты, умеющие разбирать заголовки исполняемого файла, а именно objdump и readelf. Форматирование входных данных для данных утилит будет осуществляться с помощью Bash-скриптов. Не смотря на то, что данные программы имеются только на UNIX системах, есть возможность использовать их и в операционной системе Windows, через Cygwin [32].

2.1.2 Сравнение сред разработки

Для разработки на Nim существует несколько IDE и огромное количество текстовых редакторов, часть которых рассмотрим ниже:

Рассмотрим подробно каждый из представленных в таблице редакторов.

Aporia

Простая IDE, написанная на nim, с использованием GTK2. В настоящее время не поддерживается, так как большая часть Nim-программистов перешла на Visual Studio Code.

Atom

Редактор с открытым исходным кодом от GitHub Inc., написан с использованием Electron [38] – фреймворка для разработки кросс-платформенных приложений с помощью HTML, JavaScript и CSS. Из-за архитектурных и технологических решений, все программы, написанные на данном фреймворке, будут очень требовательны к ресурсам.

Таблица 9 — Сравнительная таблица IDE и редакторов кода

Свойства \ IDE/Редактор	Aporia [33]	Atom [34]	Sublime Text [35]	Visual Studio Code [36]	Vim [37]
Поддержка плагинов	Нет	Да	Да	Да	Да
Требователен к ресурсам	Нет	Да	Нет	Да	Нет
Имеет продвинутую систему редактирования текста	Нет	Нет	Нет	Нет	Да
Кросс-платформенность	Есть	Есть	Есть	Есть	Есть
Может работать без GUI	Нет	Нет	Нет	Нет	Да
Восстановление после сбоев	Нет	Есть	Есть	Есть	Есть
Возможность выделять ключевые слова с помощью регулярных выражений	Нет	Есть	Есть	Есть	Есть
Опыт использования	Нет	Нет	Есть	Есть	Есть

Sublime Text

Проприетарный текстовый редактор написан на C++ и python, возможности которого могут быть расширены с помощью плагинов на python.

Visual Studio Code

Редактор с открытым исходным кодом от Microsoft. Так же, как и Atom, написан с использованием Electron. Имеет встроенный «магазин» плагинов.

Vim

Текстовый редактор с открытым исходным кодом и большими возможностями к быстрому редактированию текстов. Является наследником редактора vi, который, в свою очередь, создавался с оглядкой на редактор ed. Управление делится на режим ввода и режим команд, благодаря чему есть возможность управлять редактором только с помощью клавиатуры, что, при должном умении, повышает скорость не только из-за отсутствия необходимости в использовании компьютерной мыши, но и более коротким сочетаниям «горячих клавиш». Легко поддается модифицированию с помощью плагинов. Есть под множество платформ.

Вывод

Из всего вышесказанного и личного опыта следует, что для разработки ПМ АПНДВ лучше всего подойдет текстовый редактор Vim, так как он поддерживает добавление плагинов, не требователен к ресурсам и позволяет очень быстро редактировать текст. В качестве расширения его функциональности использованы плагины:

1) NERDTree [39] – улучшает просмотр каталогов;
2) Tabular [40] – позволяет быстро выравнивать текст для улучшения читаемости;
3) vim-polyglot [41] – подсветка синтаксиса большого числа языков;
4) undotree [42] – просмотр истории изменений в виде дерева;
5) rainbow [43] – подсветка вложенных скобок разными цветами, для улучшения читаемости.

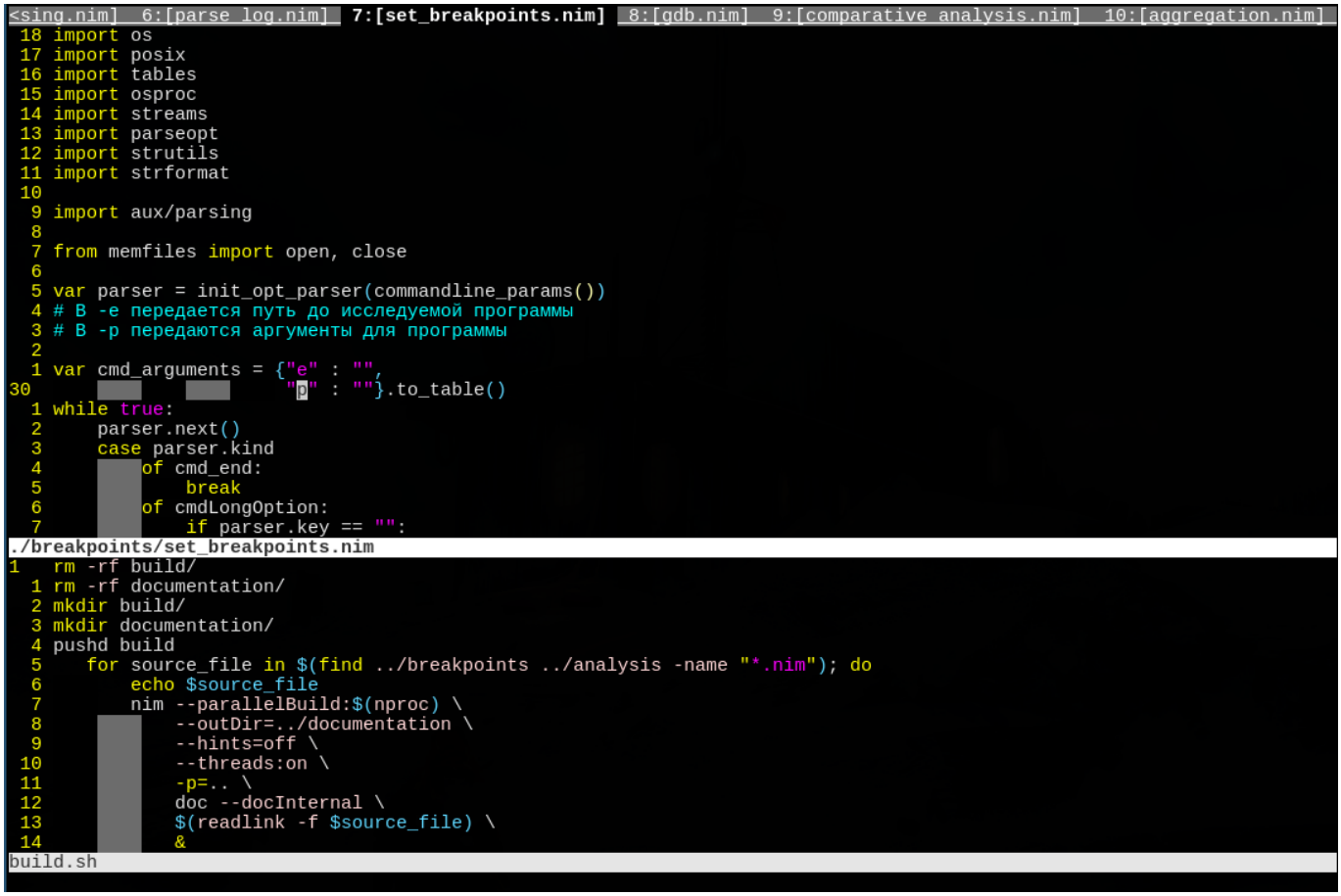


Рисунок 2.1 — Интерфейс Vim ПМ АПНДВ

2.2 Архитектура ПМ АПНДВ

Архитектура программного обеспечения это система, объединяющая внутрение компоненты, их связи между собой и с окружением, а так же принципы, использующиеся при проектировании и эволюции программы [44].

При проектировании ПМ АПНДВ была выбрана UNIX-философия [45], заключающаяся в следующих основополагающих принципах:

– создавать маленькие программы;

– программы делают одно дело, но делают его хорошо;

– хранить данные в текстовом, читаемом для людей формате.

Поэтому было принято решение разрабатывать под каждую подзадачу проведения сертификации ПО самостоятельную программу, которая была бы маленькой и хорошо бы справлялась со своим назначением.

2.2.1 Организация передачи информации между компонентами ПМ АПНДВ

Передача информации между компонентами ПМ АПНДВ осуществляется посредством сериализации внутренних структур (рис. 2.2 и рис. 2.3) конкретного модуля в формате JSON. JSON удобен тем, что является простым для чтения как человеком, так и компьютером, что позволяет оператору анализировать так же и промежуточные результаты работы, для вынесения вердикта.

Виды сериализуемых данных

В ПМ АПНДВ сериализуются данные после прохождения этапа:

– статического анализа исходных кодов;

– динамического анализа сертифицируемой программы.

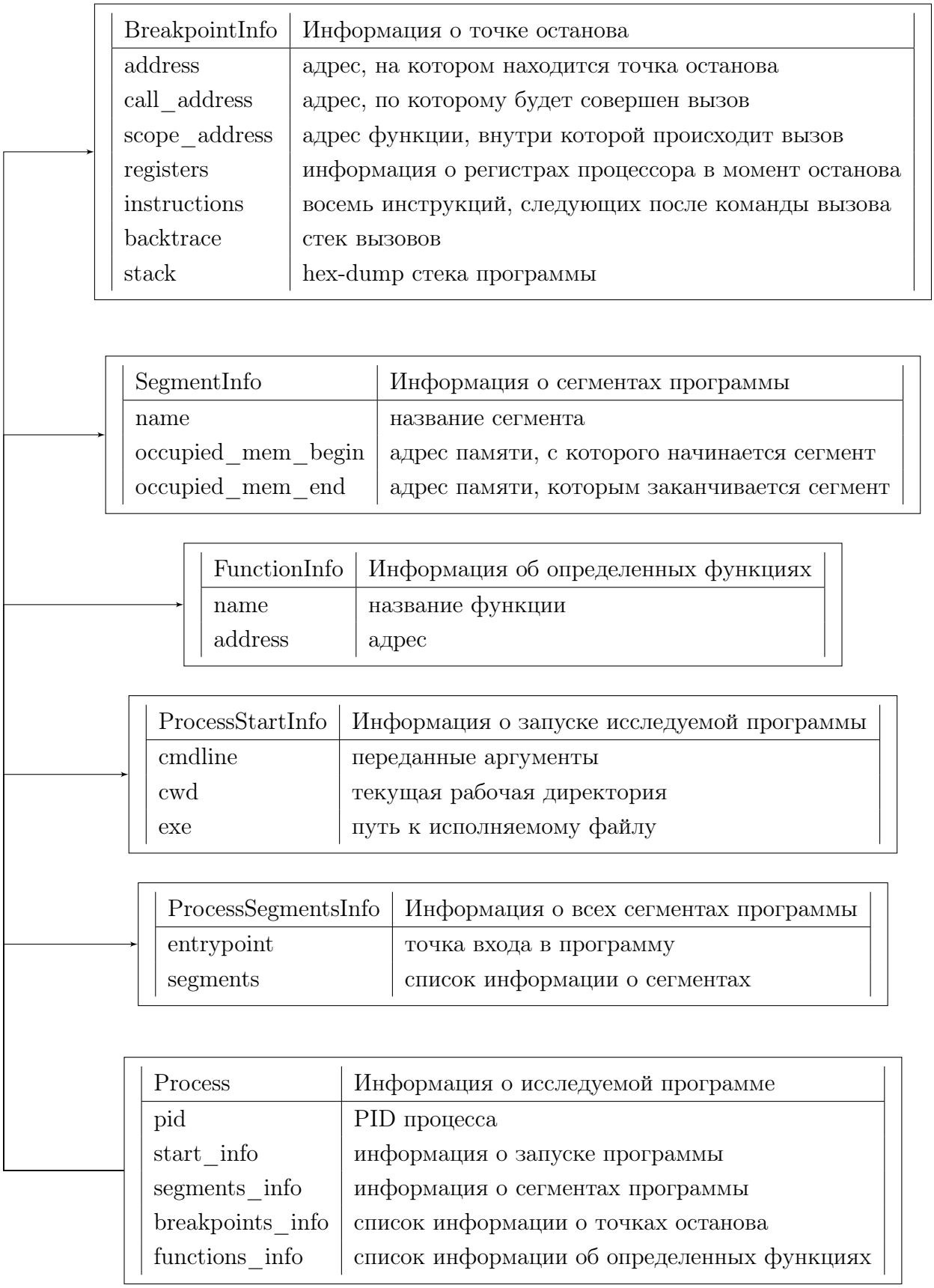
Структура данных помогает иерархически организовать доступ к собранной, во время динамического анализа, информации.

Данные с расставленных точек останова, содержатся в структуре `BreakpointInfo`, которая заполняется непосредственно во время выполнения машинных инструкций программы, а значит важно в них получить максимальное количество информации текущем мгновенном состоянии программы. В структуре содержится:

– адреса:

  – `call`-инструкции, на которой находится точка останова;

  – по которому собирается сделать вызов `call`-инструкция;



BreakpointInfo	Информация о точке останова
address	адрес, на котором находится точка останова
call_address	адрес, по которому будет совершен вызов
scope_address	адрес функции, внутри которой происходит вызов
registers	информация о регистрах процессора в момент останова
instructions	восемь инструкций, следующих после команды вызова
backtrace	стек вызовов
stack	hex-dump стека программы

SegmentInfo	Информация о сегментах программы
name	название сегмента
occupied_mem_begin	адрес памяти, с которого начинается сегмент
occupied_mem_end	адрес памяти, которым заканчивается сегмент

FunctionInfo	Информация об определенных функциях
name	название функции
address	адрес

ProcessStartInfo	Информация о запуске исследуемой программы
cmdline	переданные аргументы
cwd	текущая рабочая директория
exe	путь к исполняемому файлу

ProcessSegmentsInfo	Информация о всех сегментах программы
entrypoint	точка входа в программу
segments	список информации о сегментах

Process	Информация о исследуемой программе
pid	PID процесса
start_info	информация о запуске программы
segments_info	информация о сегментах программы
breakpoints_info	список информации о точках останова
functions_info	список информации об определенных функциях

Рисунок 2.2 — Сохраняемые структуры динамического анализа

– функции, в котором находится данная `call`-инструкция;

Которые необходимы для последующего сравнительного анализа;

– регистры, в которых могут содержаться передаваемые параметры (fastcall convention [46]);
– следующие за `call` 8 инструкций, в которых может содержаться код, обрабатывающий возвращенное значение;
– стек вызовов, позволяет посмотреть ветку исполнения исследуемой программы.

Информация о сегментах в `SegmentInfo` позволяет определить, к какому сегменту относится вызываемая, или текущая функция. Например, это может быть сегмент динамически загружаемой библиотеки.

`FunctionInfo` содержит информацию, которую предоставляет GDB при загрузке программы: список известных функций и их адреса.

`ProcessStartInfo` сохраняет параметры запуска, `ProcessSegmentsInfo` – агрегирует информацию по всем сегментам программы. Структура `Process` же агрегирует в себе всё вышеперечисленное.

UnitInfo	Информация об одном файле исходного кода
arguments	список аргументов компиляции
directory	папка с файлом исходного кода
file	имя файла

BuildInfo	Информация о сборке программы
units_info	список файлов исходного кода

CflowConstruct	Описание функции в статическом анализе
name	имя функции
nesting	уровень вложенности
signature	сигнатура функции
path	путь до файла, в котором используется функция
line	номер строки
recursive	рекурсивность функции
text_offset	отступ в сегменте .text

Рисунок 2.3 — Сохраняемые структуры статического анализа

Структуры данных, относящиеся к статическому анализу косвенно связаны друг с другом. Их можно разделить на структуры времени компиляции программы и структуры времени статического анализа. К структурам времени компиляции относятся:

- `UnitInfo` содержит информацию о сборке одного файла исходного кода; В нее входит:
  - аргументы компилятору – указание заголовочных файлов, параметры генерации машинного кода, указание макросов и т.д.;
  - папка, в котрой находится файл исходного кода;
  - название файла.
- `BuildInfo` агрегирует все `UnitInfo`, полученные при компиляции проекта и записанные в compilation database [47];

К структурам времени анализа относится `CflowConstruct`, которая содержит в себе уже разобранную и типизированную информацию, предоставляемую Cflow – программой статического анализа:

- имя функции;
- уровень вложенности вызова – уровень дерева, на котором располагается конкретная функция, относительно точки входа – функции с нулевым уровнем вложенности;
- сигнатура функции, в данном случае вместе с возвращаемым типом;
- путь до файла, в котором функция была использована;
- номер строки, где функция была использована;
- рекурсивность функции – значение принимающее либо «ложь», либо «истина», в зависимости, есть ли в определении функции вызов самой себя;
- отступ в области .text – количество в байтах от начала .text-сегмента уже скомпилированной программы до начала функции.

Все значения, кроме `text_offset`, заполняются непосредственно во время проведения статического анализа.

`text_offset` заполняется на стадии агрегации результатов линковки и результатов статического анлиза. Это необходимо, чтобы на стадии сравнительного анализа можно было сопоставить адреса вызываемых функций в динамическом и статическом анализе, полагаясь на разность между началом сегмента .text и адресом функции. Как на стадии линковки, так и в динамическом анализе для конкретной фунции он будет одинаков.

2.2.2 Схема данных

Из схемы данных на рис. 2.4 видно, что работу ПМ АПНДВ можно разбить на параллельные задачи.

2.2.3 Алгоритм работы программы

Работу ПМ АПНДВ можно разделить на функциональные этапы:

1) сборка анализируемой программы;
2) статический анализ результатов сборки;
3) динамический анализ собранной программы;
4) сравнительный анализ результатов статического и динамического анализа.

Причем п. 2) и п. 3) могут выполняться одновременно, так как не имеют зависимости по данным.

Рассмотрим подробнее каждый из этапов.

Сборка анализируемой программы

Утилита make

Make – утилита для автоматической сборки программ и библиотек из исходного кода. Работает через чтение специальных файлов – «мейкфайлов» (англ. Makefile), в которых описаны «рецепты» сборки. В мейкфайле может находиться любое количество рецептов, они могут быть как зависимы друг от друга, так и быть совершенно непересекающимися.

Отдельный рецепт имеет название, компоненты, от которых он зависит (могут остать пустыми, это будет означать, что рецепт независим) и правила сборки, они тоже могут оставаться пустыми.

Стоит заметить, что использование программы make в UNIX системах не обязательно ограничевается компиляцией программ и библиотек. В мейкфайлах с помощью рецептов так же можно описать различные сценарии, требующие последовательного выполнения команд. В большинстве программ, использующих схему распространения через компиляцию исходного кода, имеются мейкфайлы, в которых определены рецепты clean – очистить и help – помощь. Которые реализуют, соответственно, очистку директорий проекта от временных файлов, полученных в результате выполнения других рецептов мейкфайла и получения информации о доступных рецептах.

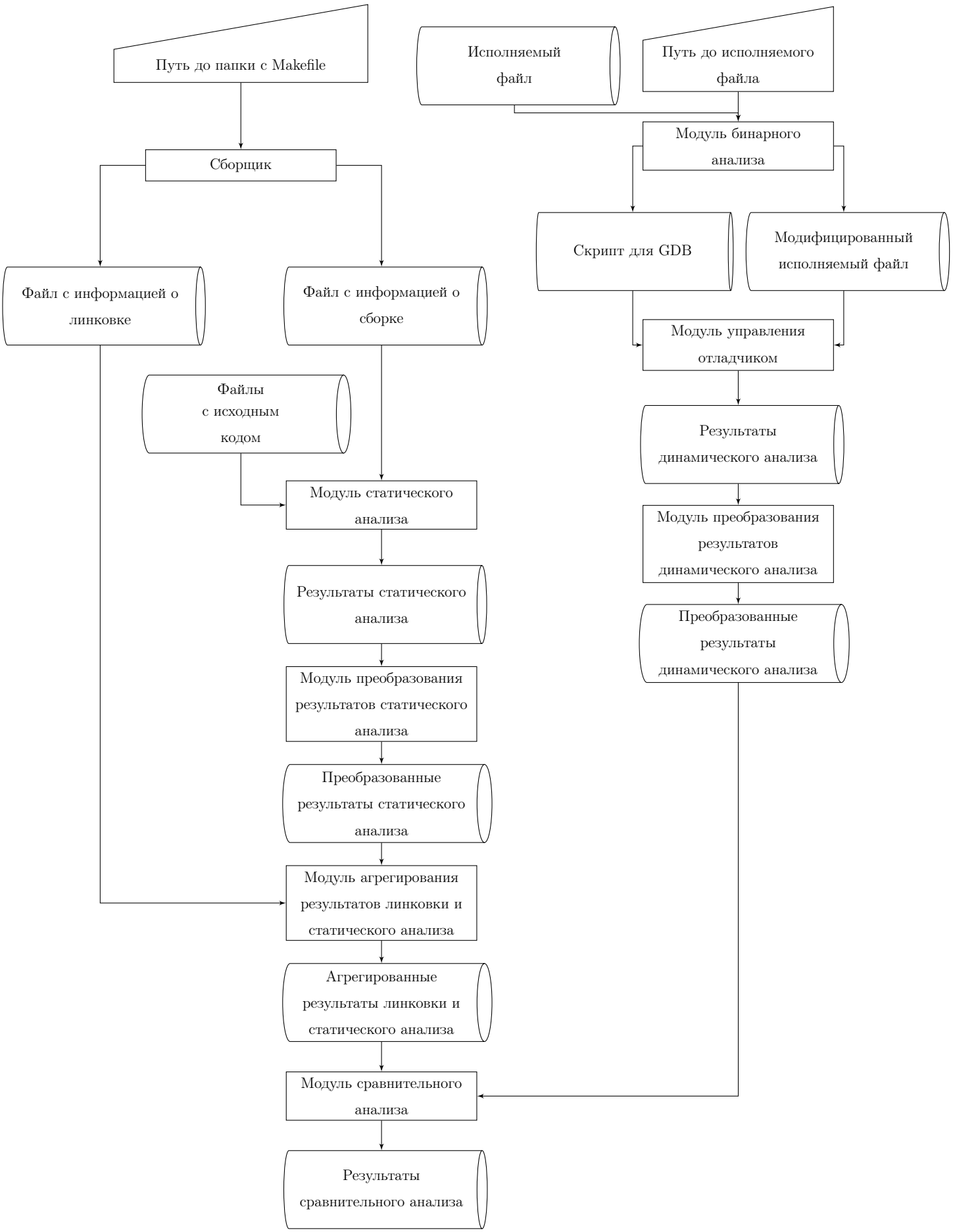


Рисунок 2.4 — Схема данных ПМ АПНДВ

По-умолчанию, make выполняет рецепты один за другим, не начиная выполнение нового рецепта, пока не закончится старый. Но при указании определенного аргумента, make может выполнять несвязанные рецепты параллельно, что значительно ускоряет процесс сборки.

Утилита BEAR

Build EAR [48], или сокращенно BEAR позволяет генерировать compilation database, указывая ей команду сборки.

Сборка анализируемой программы происходит посредством программы-обертки, повторяющей интерфейс программы make и запускающая её в контексте программы BEAR, для генерации compilation database. Помимо этого, для make указывается генерация map-файла, файла содержащего информацию о сегментах программы, относительных отступах функций внутри сегментов и др. После окончания компиляции дополнительно происходит разбор сегмента .text map-файла на предмет функций и их относительных адресов внутри сегмента. Полученные данные сохраняются на диск в JSON формате.

Статический анализ результатов сборки

Статический анализ результатов сборки производится с помощью программы Cflow, которой на вход подаются аргументы компиляции, взятые из compilation database, полученной на предыдущем шаге, а так же сами файлы с исходными кодами.

Отчет Cflow состоит из списка функций, определяемых следующим правилом, описанным в 2.1, где описания полей обрамлены косыми чертами:

Листинг 2.1 Формат записи в отчете Cflow

```
{/уровень вложенности/} /имя функции/() </сигнатура функции вместе
    с возвращаемым значением/ at /абсолютный путь до файла/:/номер
    строки в файле/>:
{/уровень вложенности вызываемой функции/} /имя вызываемой функции
   /() </сигнатура вызываемой функции вместе с возвращаемым значен
   ием/ at /абсолютный путь до файла/:/номер строки в файле/>:

    ...
```

Данный формат файла легко поддается разбору с помощью регулярных выражений. В ПМ АПНДВ использовалась библиотека регулярных выражений PCRE [49]. Не смотря на то, что Cflow умеет генерировать отчет, в которых представлен не граф вызываемых функций, а список функций, вызывавших данную, этот формат, не смотря на удобство, страдает большим количеством повторений, что в свою очередь вызывает слишком большой объем отчета и замедляет его разбор, из-за чего в ПМ АПНДВ решено было использовать стандартную версию отчета.

Листинг 2.2 Пример генерации отчета Cflow

```
{   0} printsel() <void printsel (const arg *arg) at /st/st.c:1988>:
{   1}  tdumpsel() <void tdumpsel (void) at /st/st.c:1994>:
{   2}    getsel() <char *getsel (void) at /st/st.c:590>:
{   3}      xmalloc() <void *xmalloc (size_t len) at /st/st.c:253>:
{   4}        malloc()
{   4}        die() <void die (const char *errstr, ...) at /st/st.c:654>:
```

Динамический анализ собранной программы

Подготовка к динамическому анализу собранной программы начинается сразу после завершения этапа сборки разд. 2.2.3. Путь до исполняемого файла передается в модуль расстановки точек останова для первичного модифицирования. Модифицирование заключается в том, что с помощью программ objdump и readelf, о которых говорилось в разд. 2.1.2 и небольших скриптов, написанных на bash, происходит следующее:

1) находятся все `call`-инструкции, сохраняя их относительные адреса от начала сегмента .text;
2) узнается отступ сегмента .text в байтах от начала файла;
3) сохраняется байт по адресу, полученным на предыдущем шаге;
4) заменяется байт по адресу, полученным на предыдущем шаге, на 0xCC в шестнадцатиричной системе счисления. Это машинный код инструкции `int 3` – программного прерывания, которое используется в отладчиках для установки точек останова;
5) генерируется скрипт для отладчика GDB, по расстановке точек останова на все `call`-инструкции, восстановлению изменений в файле и снятию состояний программы.

Процесс исполнения данного скрипта:

1) загружается исполняемый файл;
2) выводится информация о процессе, сегментах и обнаруженных функциях;
3) запускается исследуемая программа;
4) программа останавливается на первом байте сегмента .text, 0xCC, кодирующем программную точку останова;
5) счетчик команд уменьшается на единицу;
6) по адресу, указанном в счетчике команд записывается ранее сохраненный первый байт сегмента .text;
7) расставляются относительные точки останова;
8) программа выходит из останова и продолжает работу, собирая информацию с точек останова.

Нужно отметить, что в отладчике GDB существует команда `starti`, которая запускает программу и останавливается на первой инструкции, что позволяет отлаживать программу прямо с точки входа. Но проблема использования `starti` состоит в том, что первой инструкцией программы может оказаться не .text-сегмент, а какой-нибудь другой, а значит относительная расстановка точек будет неверной. Поэтому приходится на уровне исполняемого файла удостоверяться, что исполнение программы прервется именно на первой инструкции .text-сегмента.

Сравнительный анализ результатов статического и динамического анализа

Модуль сравнительного анализа запускается после того, как становятся готовы результаты статического и динамического анализа. Он загружает результаты с диска в описанные ранее структуры разд. 2.2.1, а так же информацию о функциях из map-файла. Это позволяет дать отчет по нескольким вариантам несовпадения:

– несовпадение функций в map-файле и функций, объявленных в статическом анализе (каких имен из множества функций, полученных из map-файла нет среди функций, определенных в исходникых текстах);
– несовпадение распознанных отладчиком GDB функций и функций, полученных в динамическом анализе (каких имен из множества функций, определенных GDB нет среди функций, полученных из map-файла);
– несовпадение функций в статическом и динамическом анализе.

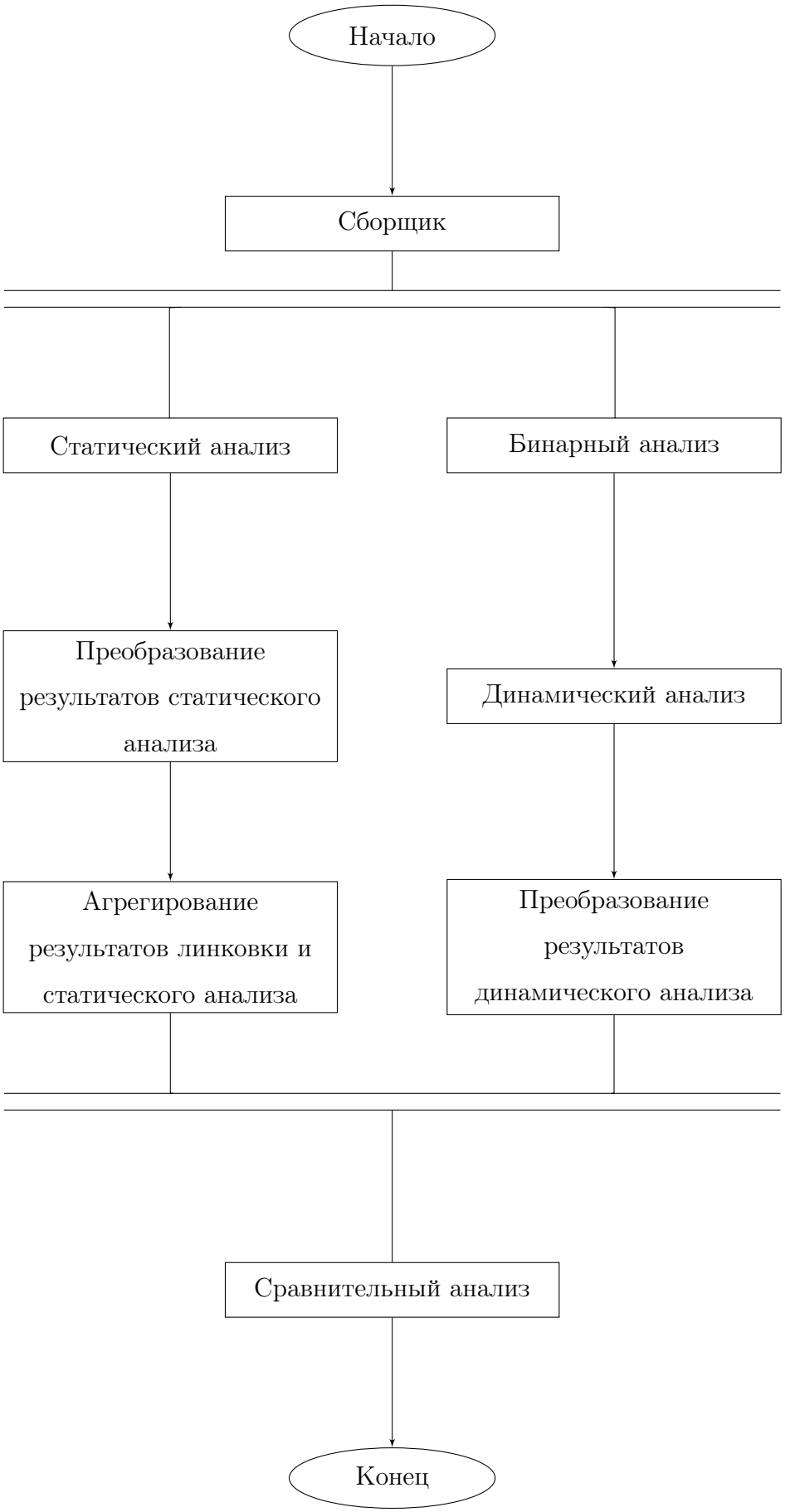


Рисунок 2.5 — Алгоритм работы ПМ АПНДВ

2.2.4 Разработка консольного интерфейса ПМ АПНДВ

Консольный интерфейс программы, или программа, поддерживающая интерфейс командной строки – компьютерная программа, обрабатывающая аргументы, переданные ей в определенном формате. Консольный интерфейс не может существовать без командного интерпретатора – другой компьютерной программы, которая обрабатывает команды компьютеру, заданные в виде текста. Один из самых старых видов взаимодействия человека и компьютера. Появившись в середине 1960-х, он используется и по сей день.

На текущий момент, для запуска ПМ АПНДВ нужно перейти в папку с собранным ПМ АПНДВ, после чего использовать bash-скрипт 2.3, который принимает следующие аргументы:

1) `$1` – путь до папки, в которой хранится мейкфайл проекта;
2) `$2` – путь до исследуемого исполняемого файла.

Результаты сравнительного анализа выводятся на экран.

Листинг 2.3 run.sh

```
pushd $1
    make clean
popd
pushd build
    ./build -C=$1
    (./set_breakpoints -e=$2 &&
     ./gdb ;
     ./parse_log &&
     ./dynamic_analysis;) &
    (./static_analysis &&
     ./aggregation) &
     wait $(jobs -p)
    ./comparative_analysis
popd
```

Если же ПМ АПНДВ еще не собран, то нужно воспользоваться скриптом 2.4:

Листинг 2.4 build.sh

```
rm -rf build
mkdir build
pushd build
   for source_file in $(find ../breakpoints ../analysis -name "*.nim"); do
       echo $source_file
       nim --parallelBuild:$(nproc) \
           --outDir=. \
           -p=.. \
           --threads:on \
           c $(readlink -f $source_file) &
   done
   wait $(jobs -p)
popd
```

2.3 Выводы по разделу

В конструкторском разделе было проведено сравнение и обоснование выбора языка программирования и среды разработки для ПМ АПНДВ. Разработана архитектура ПМ АПНДВ. Также были описаны:

1) алгоритм передачи данных между модулями ПМ АПНДВ;
2) формат данных, передающихся между модулями ПМ АПНДВ;
3) используемые сторонние программы и форматы данных, обрабатываемые ими.

Составлена схема данных, алгоритм работы ПМ АПНДВ. Подробно рассмотрены шаги выполнения процесса сертификации с помощью ПМ АПНДВ

Список литературы

1. *SophosLabs.* Compile-a-virus – W32/Induc-A [Текст] / SophosLabs. — 2009. — URL: https://nakedsecurity.sophos.com/2009/08/18/compileavirus/ (дата обр. 18.08.2009).
2. *Томпсон, К.* Ken Thompson Hack [Текст] / К. Томпсон. — 1984. — URL: http://wiki.c2.com/?TheKenThompsonHack.
3. *Алексеев, А.* Краткий обзор статических анализаторов кода на C/C++ [Текст] / А. Алексеев. — 2016. — URL: https://eax.me/c-static-analysis/ (дата обр. 11.05.2016).
4. *anti-malware.ru.* Недекларированные возможности [Текст] / anti-malware.ru. — URL: https://www.anti-malware.ru/threats/undeclared-capabilities.
5. *Ализар, А.* Stuxnet был частью операции «Олимпийские игры», которая началась еще при Буше [Текст] / А. Ализар. — 2012. — URL: https://xakep.ru/2012/06/02/58789/ (дата обр. 02.06.2012).
6. Приказ ФСТЭК России №21 [Текст]. — URL: https://fstec21.blogspot.com/2017/07/type-actual-security-threats.html.
7. *Microsoft.* Application Inspector [Текст] / Microsoft. — 2019. — URL: https://github.com/microsoft/ApplicationInspector.
8. *scitools.* Features [Текст] / scitools. — URL: https://scitools.com/features/.
9. *Позняков, С.* GNU cflow [Текст] / С. Позняков. — URL: https://www.gnu.org/software/cflow/.
10. Introduction to .NET Core [Текст]. — URL: https://docs.microsoft.com/ru-ru/dotnet/core/introduction.
11. *Free Software Foundation, I.* GNU Debugger [Текст] / I. Free Software Foundation. — URL: https://www.gnu.org/software/gdb/.
12. *Bellard, F.* QEMU [Текст] / F. Bellard. — URL: https://www.qemu.org/.
13. *Rumpf, A.* Nim [Текст] / A. Rumpf. — URL: https://nim-lang.org/.
14. *Rossum, G. van.* python [Текст] / G. van Rossum. — URL: https://www.python.org/.

15. *Wall, L.* Perl [Текст] / L. Wall. — URL: https://www.perl.org/.
16. pylint [Текст]. — URL: https://www.pylint.org/.
17. pyflakes [Текст]. — URL: https://github.com/PyCQA/pyflakes.
18. What "Batteries Included"Means [Текст]. — URL: https://protocolostomy.com/2010/01/22/what-batteries-included-means/ (дата обр. 22.01.2010).
19. pip [Текст]. — URL: https://pypi.org/project/pip/.
20. pip [Текст]. — URL: https://pypi.org/project/pip/.
21. PyPy [Текст]. — URL: https://www.pypy.org/.
22. Jython [Текст]. — URL: https://www.jython.org/.
23. Iron Python [Текст]. — URL: https://ironpython.net/.
24. AWK [Текст]. — URL: http://www.awklang.org/.
25. sed [Текст]. — URL: https://www.gnu.org/software/sed/.
26. JavaScript [Текст]. — URL: https://www.javascript.com/.
27. Move semantics [Текст]. — URL: https://nim-lang.org/docs/destructors.html#move-semantics.
28. A garbage collector for C and C++ [Текст]. — URL: https://www.hboehm.info/gc/.
29. Getting to Go: The Journey of Go's Garbage Collector [Текст]. — URL: https://blog.golang.org/ismmkeynote.
30. Nimble [Текст]. — URL: https://github.com/nim-lang/nimble.
31. reStructuredText. Markup Syntax and Parser Component of Docutils [Текст]. — URL: https://docutils.sourceforge.io/rst.html.
32. Cygwin [Текст]. — URL: https://www.cygwin.com/.
33. Aporia [Текст]. — URL: https://github.com/nim-lang/Aporia/.
34. Atom [Текст]. — URL: https://atom.io/.
35. Sublime Text [Текст]. — URL: https://www.sublimetext.com/.
36. Visual Studio Code [Текст]. — URL: https://code.visualstudio.com/.
37. Vim [Текст]. — URL: https://www.vim.org/.
38. Electron [Текст]. — URL: https://www.electronjs.org/.

39. NERDTree [Текст]. — URL: https://github.com/preservim/nerdtree.

40. Tabular [Текст]. — URL: https://github.com/preservim/nerdtree.

41. vim-polyglot [Текст]. — URL: https://github.com/sheerun/vim-polyglot.

42. undotree [Текст]. — URL: https://github.com/mbbill/undotree.

43. rainbow [Текст]. — URL: https://github.com/luochen1990/rainbow.

44. What is a software architecture? [Текст]. — URL: https://www.ibm.com/developerworks/rational/library/feb06/eeles/index.html (дата обр. 15.02.2006).

45. Unix Design Philosophy [Текст]. — 1995. — URL: https://wiki.c2.com/?UnixDesignPhilosophy.

46. __fastcall [Текст]. — URL: https://docs.microsoft.com/ru-ru/cpp/cpp/fastcall?view=vs-2019.

47. JSON Compilation Database Format Specification [Текст]. — URL: https://clang.llvm.org/docs/JSONCompilationDatabase.html.

48. Build EAR (BEAR) [Текст]. — URL: https://github.com/rizsotto/Bear.

49. PCRE - Perl Compatible Regular Expressions [Текст]. — URL: https://www.pcre.org/.