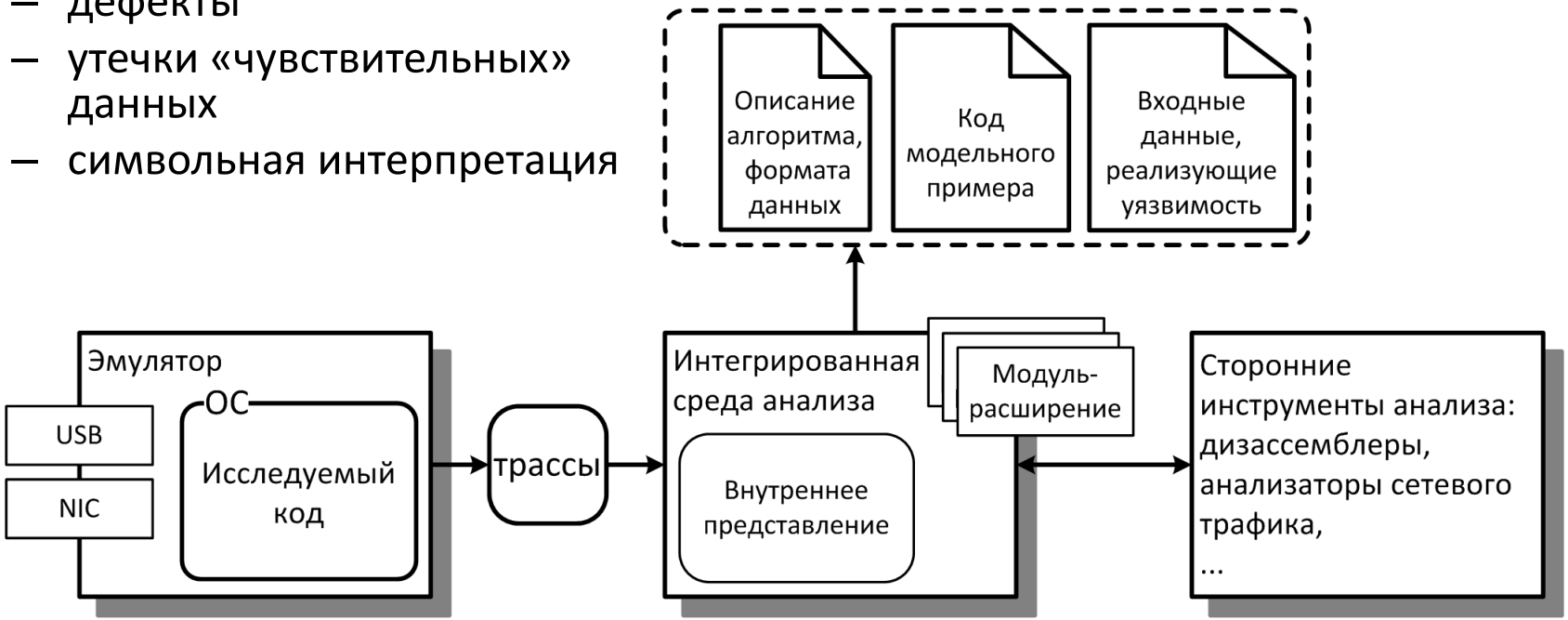
Восстановление формата данных путем анализа бинарного кода: состояние и перспективы

Гетьман А.И., Падарян В.А.
thorin@umail.ru, snoopdd@nm.ru

Анализ бинарного кода и его приложения

- Программные эмуляторы
  - отладка, сбор трасс
- Интегрированная среда анализа
  - восстановление алгоритмов и форматов данных
  - настольные компьютеры, серверы, мобильные устройства
- Поиск ошибок и уязвимостей
  - дефекты
  - утечки «чувствительных» данных
  - символьная интерпретация



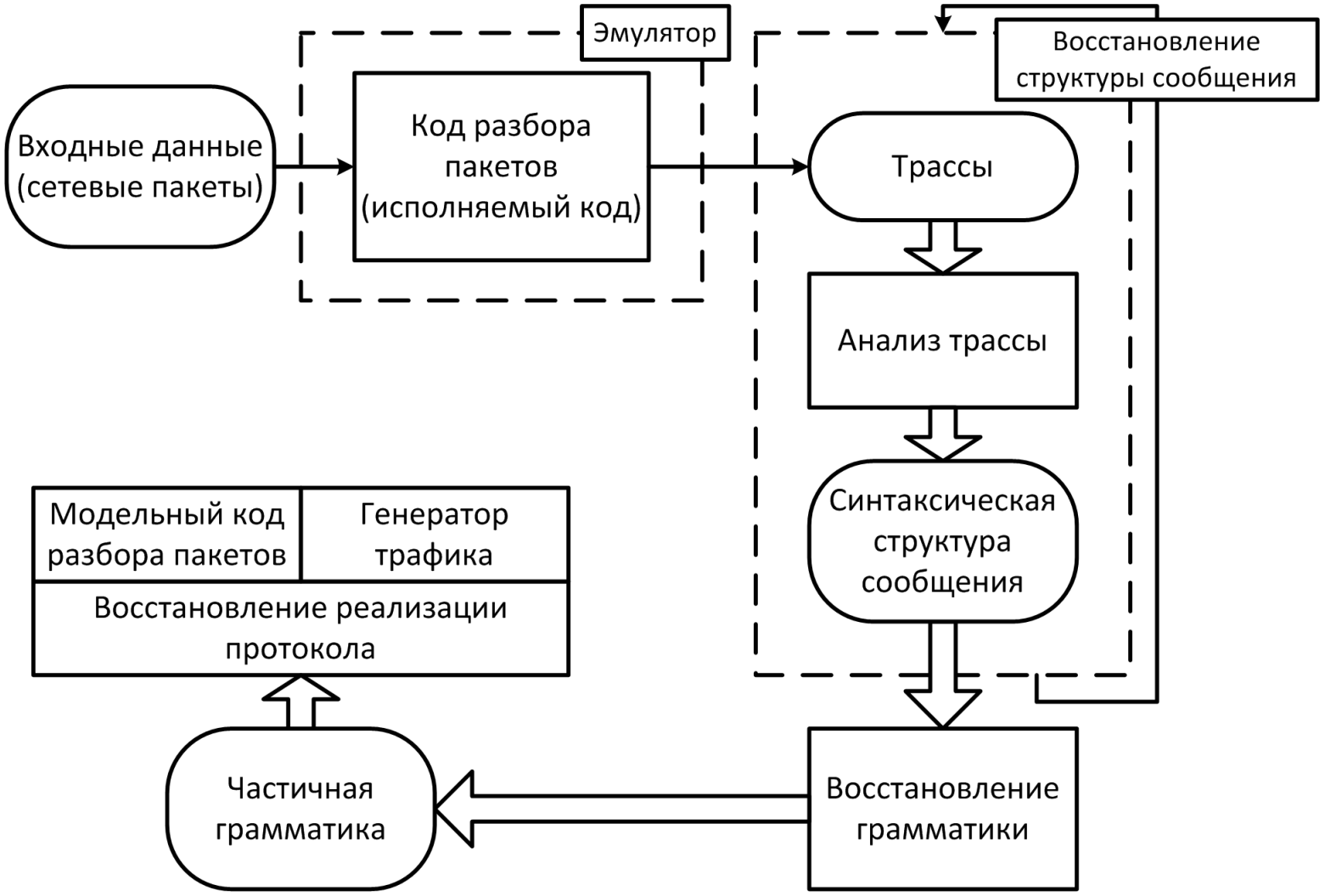
Восстановление форматов данных
Области применения

- Анализ структур данных алгоритмов
- Сравнительный анализ реализаций протоколов
- Автоматический разбор сетевых сообщений в анализаторах трафика
  - Wireshark
- Автоматическая генерация входных данных с корректной структурой в инструментах фаззинга
  - Peach
- Системы на базе DPI: межсетевые экраны, антивирусы, IDS, IPS и т.д.
  - Snort, Bro, SourceFire NGIPS/NGFW, Windriver INP

Обзор близких работ

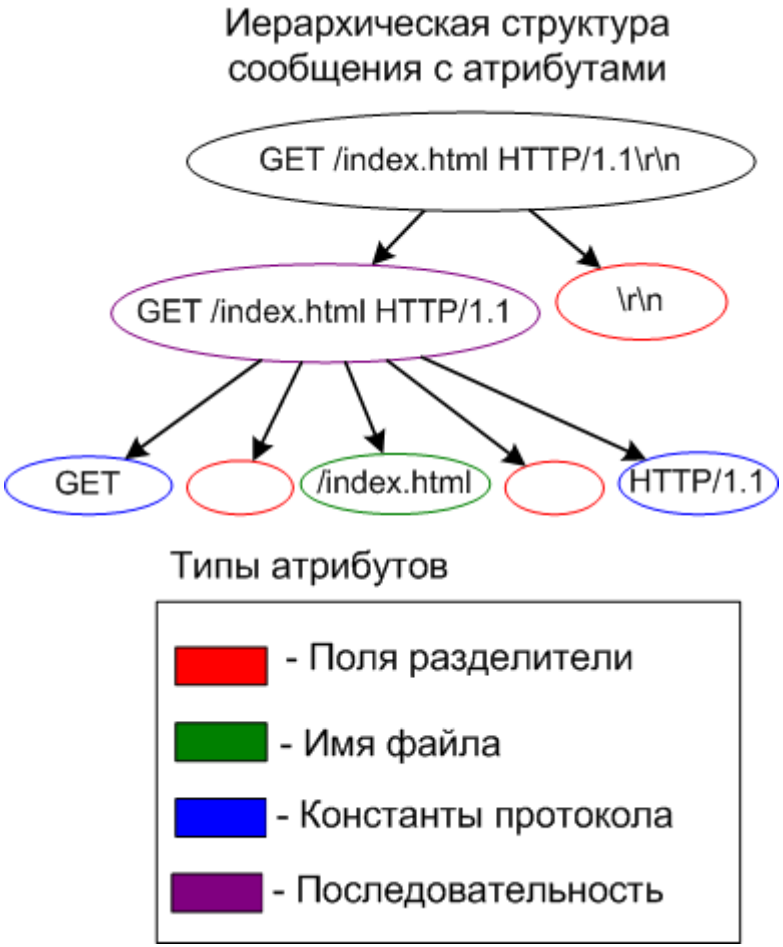
- Анализ потока сетевых пакетов с выделением повторяющихся паттернов
  - Discoverer, PEXT, ReverX
- Статический анализ кода
  - FFE/x86
- Динамический анализ кода
  - Tupni, Dispatcher, Prospex
- Российские разработки …

Динамический анализ кода, выполняющего разбор формата данных

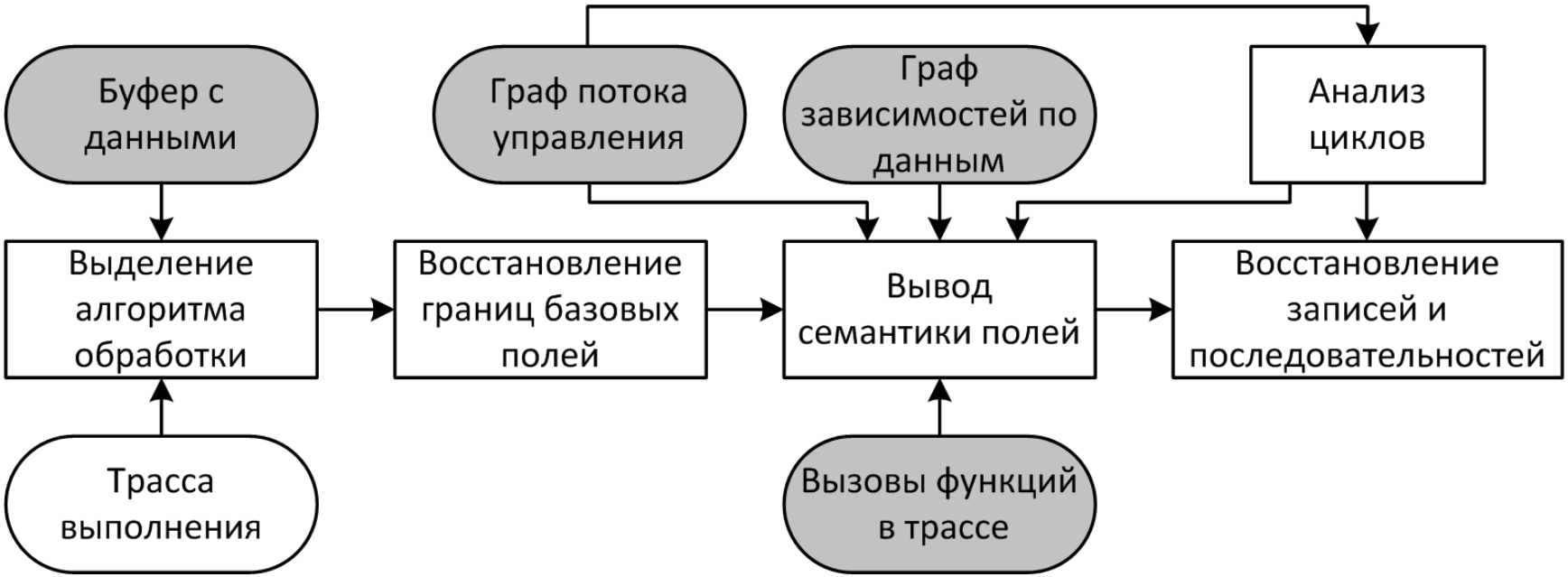


Синтаксическая структура данных

- Древовидное представление
  - Корень – всё сообщение
  - Промежуточные вершины – составные поля
  - Листовые вершины – поля базовых типов
- Атрибуты вершин
  - Тип поля
    - базовое поле
    - запись
    - последовательность
  - Семантика поля
  - Дополнительные атрибуты в зависимости от семантики



Восстановление синтаксической структуры данных (сообщения)



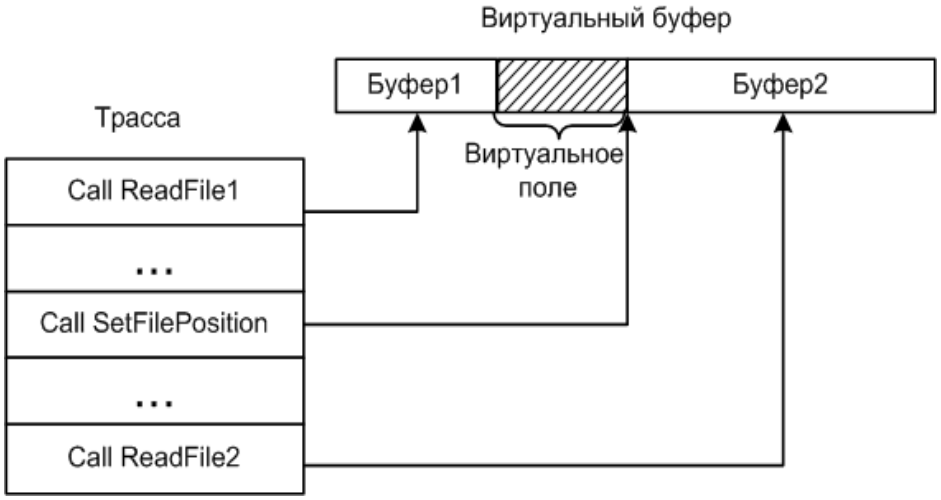
Для выделения алгоритма обработки недостаточно слайсинга трассы и анализа помеченных данных.

Осложняющие анализ факторы

- В некоторых случаях данные поступают по частям (считывание файла)
  - Требуется выделить все источники данных
- Шифрование и сжатие данных
  - Требуется выделить функции шифрования/дешифрования и параметры с данными до и после шифрования
- Для обобщения подхода вводится понятие *виртуального буфера*

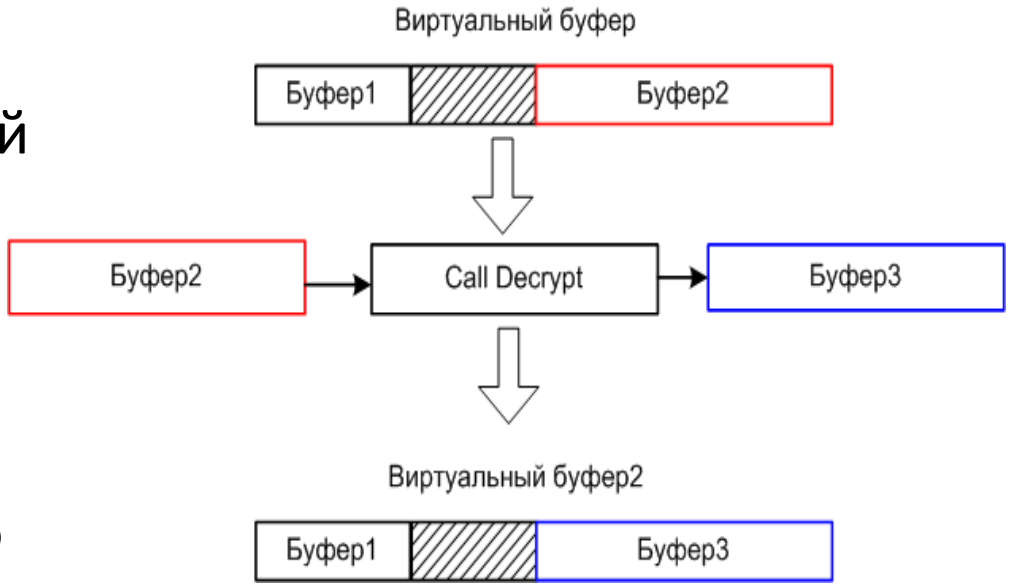
Алгоритм построения виртуального буфера

- Аннотация функций чтения, записи, позиционирования, открытия и закрытия
- Поиск всех вызовов, восстановление параметров, сопоставление вызовов отдельным файлам и сообщениям
- Для каждого вызова функции чтения/записи определяется, какому смещению в виртуальном буфере соответствует считываемый/записываемый буфер
- Части виртуального буфера, к которым не было доступа объявляются *виртуальными полями*



Алгоритм анализа шифрованных данных

- Автоматический поиск функций шифрования по высокой доле арифметических инструкций
- Уточнение списка функций и их параметров
- Вычисление частей сообщения, которым соответствуют эти параметры
- Построение виртуального буфера с учётом этих данных



Перспективное направление исследований: восстановление протокола

Возникающие подзадачи:

- Восстановление форматов всех сообщений протокола
- Выделение классов сообщений
  - буквы алфавита A
- Выделение сессий обмена данными
  - слова языка L
- Восстановление автомата протокола
  - язык L

Задача эквивалентна задаче восстановления языка по множеству входящих в него слов

Спасибо за внимание!

Вопросы?