ИССЛЕДОВАНИЕ УЯЗВИМОСТИ ФНФ ТИПА АРБИТР К КРИПТОГРАФИЧЕСКИМ АТАКАМ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

С. С. Заливако, А. А. Иванюк
Факультет электротехники и электроники, Наньянгский технологический университет
Кафедра информатики, Белорусский государственный университет информатики и радиоэлектороники
Сингапур, Сингапур
Минск, Республика Беларусь
E-mail: zali0001@e.ntu.edu.sg, ivaniuk@bsuir.by

Одним из существенных недостатков физически неклонируемой функции типа арбитр является уязвимость к криптографическим атакам с помощью машинного обучения. Этот факт обусловлен линейностью модели формирования времени распространения задержки сигнала, которая хорошо аппроксимируется линейным бинарным классификатором. В данной статье рассматривается метод, значительно затрудняющий возможность атаки с помощью машинного обучения, который основан на хешировании значений запросов. Предлагаемый метод не влияет на важнейшие характеристики ФНФ (стабильность, уникальность, случайность), поскольку ее структура остается неизменной, а преобразованию подвергаются только запросы.

Введение

В настоящее время широкое распространение получило использование физически неклонируемых функций (ФНФ) в качестве криптографических примитивов: идентификатора устройства, генератора ключей для протокола шифрования, физической односторонней функции (аппаратного хеша) и т.п. Преимуществом ФНФ по сравнению с классическими реализациями протоколов аппаратной криптографии является отсутствие необходимости хранения ключа, поскольку он генерируется на основе уникальных (неклонируемых) характеристик интегральной схемы (ИС), принимающих случайные значения. Кроме того, реализация ФНФ несет в себе относительно небольшие затраты аппаратурных ресурсов по сравнению с реализацией алгоритмов асимметричного шифрования, например, RSA. С другой стороны, к существенным недостаткам ФНФ можно отнести чувствительность стабильности генерируемых ключей к изменениям условий функционирования ИС (температура, значение питающего напряжения и т.п.), а также возможность предсказания значений ключей с помощью методов машинного обучения.

I. Физически неклонируемая функция типа арбитр

Одной из наиболее хорошо исследованных реализаций ФНФ в устройствах программируемой логики является ФНФ типа арбитр [1]. Повышение стабильности ответов ФНФ может быть осуществлено, например, с использованием кодов коррекции ошибок, мажоритарного анализа ответов, детекторов метастабильного состояния и других методов. С другой стороны, улучшение стабильности ответов ФНФ позволяет злоумышленнику воспользоваться уязвимостью множества пар запрос-ответ ФНФ типа арбитр к методам машинного обучения, осуществляющим бинарную классификацию (логистическая регрессия, метод опорных векторов, искусственные нейронные сети и т.п.) [2]. Уязвимость рассматриваемой ФНФ является следствием линейности модели формирования ответов, значения которых вычисляются как функция от разности значений временных задержек двух копий одного сигнала, которые распространяются по конфигурируемым симметричным путям. Как было показано ранее, использование упомянутых методов позволяет предсказывать с точностью 99% значения ответов 64-разрядной ФНФ типа арбитр со стабильностью ответов близкой к 100%, обладая знанием порядка 6500 пар запрос-ответ и используя их в качестве обучающей выборки. Таким образом, если реализация ФНФ типа арбитр обладает высокой стабильностью генерируемых пар запрос-ответ, то ее уязвимость к криптографическим атакам с помощью машинного обучения значительно возрастает.

II. Метод предотвращения атаки с помощью машинного обучения

Рассмотрим функцию значения ответа ФНФ $R = PUF(C)$, где C – двоичное значение запроса, а $R \in \{0, 1\}$ – значение ответа ФНФ. В свою очередь, значение ответа зависит от знака разности задержек распространения (Δ_C) двух копий исходного сигнала по конфигурируемым путям, формируемым с помощью значения запроса C, т.е. $PUF(C) \equiv PUF(\Delta_C)$. Таким образом, множество значений разностей задержек, соответствующих различным ответам арбитра, хорошо разделяется с помощью бинарного классификатора по значениям запросов C. Нелинейное преобразование множества запросов приведет к значительному ухудшению разделимости классов и, соответственно, к затруднению построения классификатора (см. Рис. 1).

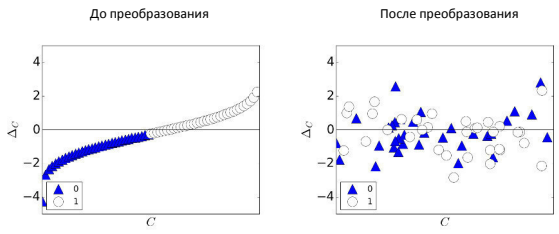

Рис. 1 – Преобразование множества пар запрос-ответ

Описанное выше преобразование может быть осуществлено, например, с помощью реализации аппаратной хеш-функции от запроса. В этом случае значение ответа арбитра будет формироваться как функция от хеш-значения запроса $R = PUF(Hash(C))$, где $Hash$ – некоторая хеш-функция. Более того, такая реализация позволяет применять методы улучшения стабильности ответов ФНФ не к значениям запросов, а к их хеш-значениям, которые не известны злоумышленнику.

III. Результаты экспериментов

Предлагаемый метод был реализован на ПЛИС ZC706, входящей в состав платы быстрого прототипирования Xilinx Zynq-7000 SoC. Одним из допущений эксперимента является то, что система, направленная на обеспечение защиты информации, по умолчанию обладает реализацией аппаратной хеш-функции и, следовательно, у разработчика ФНФ будет возможность ее повторно использовать без привлечения дополнительных аппаратурных затрат.

Экспериментальная установка представляет собой персональный компьютер (Host) и плату быстрого прототипирования (FPGA). На стороне Host был программно реализован генератор псевдослучайной M-последовательности для получения 128-битных слабо коррелированных запросов, а также алгоритм хеширования SHA256. На FPGA была произведена аппаратная реализация 128-битной ФНФ, содержащей 16 равномерно распределенных арбитров для формирования ответов. Реализованная мультиарбитральная ФНФ эмулирует поведение 16 ФНФ различной разрядности.

Для проверки предложенного метода на стороне Host было сгенерировано $N = 10^6$ запросов, которые были поданы на входы ФНФ. В результате было получено 16 наборов данных для ФНФ различной разрядности, которые, в свою очередь были разделены на обучающую (10%) и экзаменационную выборки (90%).

В соответствии с линейной аддитивной моделью представления задержки распространения сигнала ФНФ типа арбитр запросы были также преобразованы к знаковому виду [2], т.е. представлены с помощью чисел -1 и 1. Например, запрос {1, 0, 1, 1} может быть представлен в виде {-1, -1, 1, -1}. Затем было произведено обучение каждого набора данных с помощью метода опорных векторов и подсчитана точность классификатора (процент правильных классификаций) как для значений запросов, так и для соответствующих хеш-значений. Результаты эксперимента приведены на Рис. 2.

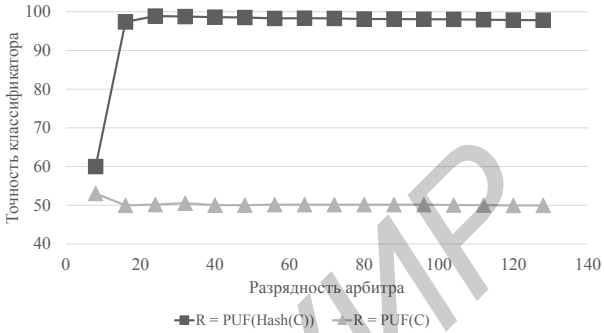

Рис. 2 – Точность классификатора на полученных наборах данных

Как показано на графике точность классификатора на запросах до хеширования значительно хуже, чем при обучении на хеш-значениях. Точность классификатора отличается незначительно только для случая арбитра разрядностью 8. Этот факт можно объяснить тем, что количество запросов относительно небольшое и было ранее показано, что ФНФ типа арбитр не обладает стабильностью, если длина путей не превышает 16. Таким образом, предлагаемый метод продемонстрировал эффективность против криптографических атак с помощью машинного обучения, поскольку классификатор, построенный на базе 10% пар запрос-ответ для арбитра каждой разрядности, показал точность не отличимую от случайного выбора или выбора константного значения (около 50%), что неприемлемо для бинарного классификатора.

IV. Заключение

Хеширование запросов ФНФ типа арбитр позволило значительно затруднить возможность криптографических атак с помощью машинного обучения, сохранив при этом характеристики стабильности, уникальности и случайности исходной реализации. Одним из направлений развития данной работы является реализация нелинейного преобразования значения запроса, со свойствами, ухудшающими работу классификаторов без допушения о том, что криптосистема содержит аппаратный хеш по умолчанию.

Список литературы

1. Suh, E.G. Physical Unclonable Functions for Device Authentication and Secret Key Generation / G.E. Suh, S. Devadas // Design Autom. Conf (DAC'07), San Diego, USA, June 2007. – P. 9–14.
2. Ruhrmair, U. PUF modeling attacks on simulated and silicon data / U. Ruhrmair, et al. // IEEE Transactions on Information Forensics and Security. -– 2013. -– № 8(11). -– P. 1876-–1891.