BitReconstructor

Data Protection

«BitReconstructor» представляет собой программный комплекс, предназначенный для восстановления файла из «обломков» файловой системы (или набора «недостоверных» копий, путем применения «бинарного голосования»). Предполагается, что «BitReconstructor» не заменяет существующие программные комплексы для восстановления данных, а лишь дополняет их. В общем случае, можно выделить следующие уровни восстановления данных с носителя:



Рисунок 1 – Структура слоев коррекции данных с точки зрения «BitReconstructor»

Физический уровень (0) отражает работу аппаратных и аппаратно-программных механизмов, позволяющих извлечь данные из структуры физических полей и вещества. Драйвер оптического диска, повторно запрашивающий данные «битого» сектора, контроллер flash-накопителя, применяющий прозрачную схему ECC – все это (с точки зрения «BitReconstructor» - предшествующий уровень, обозначаемый индексом «0», даже если он включает в себя множество подуровней). Множество уровней выше «BitReconstructor» также обозначаются одним уровнем – т.к. для пограничных уровней детализация не имеет смысла. На уровне «продвинутой коррекции данных» происходит множество сложных преобразований, возможно, применение многослойных схем коррекции ошибок, в то время как «BitReconstructor» обеспечивает первичное восстановление данных «из обломков», в ситуациях, когда высокоуровневые схемы коррекции могут быть бессильны.

«BitReconstructor» реализует два промежуточных слоя восстановления данных, которые будем называть условно «BR-L1» и «BR-L2» («BitReconstructor - Layer1» и «BitReconstructor - Layer2»). На уровне «BR-L1» (уровень неструктурированных двоичных данных) применяется технология бинарного голосования, когда потоки данных, предварительно установленные на нужные смещения от начала, проходят через блок построчного голосования, в котором из набора N-байт (по одном байту с каждого потока) выбирается наиболее часто встречающийся. Каждому потоку присваивается уровень доверия («вес»), и поток данных с уровнем доверия «2» будет голосовать дважды.



Рисунок 2 – Бинарное голосование «BR-L1»

При внесении голоса за бит «0», из счетчика будет вычитаться вес потока, а при наличии бита «1» - вес добавляется. Модуль суммы голосов по каждой битовой строке, деленный на сумму весов потоков, дает «мгновенный уровень достоверности». В случае, если хотя бы одно битовое голосование дает «0» на счетчике, это трактуется как состояние неопределенности (т.е. ошибка). Количество потенциальных битовых ошибок выводится как отдельный атрибут голосования. Минимальное количество файлов в голосовании – 2. При этом любое несовпадение бит будет давать ошибку.

На уровне «BR-L2» (уровень структурированных двоичных данных), «выжившие» виртуальные «секторы» собираются в «мозаику». Две «мозаики» могут быть рассмотрены как один выход слоя «BR-L1» и объединены.



Рисунок 3 – Структура сектора «BR-L2»

Для предотвращения чрезмерной загрузки ЦП при поиске данных в «мусоре», сектор начинается с преамбулы «BitReconstructor» (128 бит). Если активен режим шифрования, заголовок будет заполняться псевдослучайным набором данных, т.к. шифрование охватывает весь сектор – такой подход гарантирует защиту от криптоанализа, ведь различные секторы будут шифроваться одним и тем же ключом, и наличие в начале блока данных некоторой «прослойки» с неизвестным или сложно предсказуемым содержимым положительно влияет на криптографическую стойкость. С этой же целью в начале сектора размещены и три хеш-сигнатуры «Blake2b», промаркированные как «L2.x ECC» (каждая из которых вычисляется по своему смещению, все смещения разнесены на одинаковое расстояние (0, 1/3, 2/3)). Таким образом, длина «непредсказуемых» данных, поступающих на шифрование, будет обеспечивать ровно 6.5 CBC-блока (что позволит безопасно применять один и тот-же ключ).

Далее следуют поля, относящиеся к данным пользователя:

1. Относительное точное время начала создания архива;
2. Имя файла, к которому относится данный блок;
3. Порядковый номер блока данных;
4. Длина блока пользовательских данных;
5. Пользовательские данные.

Наличие нескольких наборов секторов, указывающих на файл с одним и тем же именем, не является проблемой, т.к. в этом случае принудительно активируется режим маркировки имени восстанавливаемого файла датой, временем, а также «сырым» значением количества тиков.

В конце «сектора» – блок выравнивания шифрования Rijndael-256, если используется шифрование, и хеш-сигнатура всех предшествующих данных («Blake2b/256»), если шифрование не активно. Максимально-возможный объем хеш-сигнатур обуславливается желанием максимально обезопасить пользователя от наступления события «случайный набор байт был проинтерпретирован как валидный сектор данных».