УДК 004.6.056 : 004.75

А.В. Зайцев, С.С. Гостев, П.А. Черкашин, А.Ю. Щербаков

О технологии распределенного хранения конфиденциальной информации в центрах обработки данных общего назначения*

Обсуждается проблема разработки проекта архитектуры и математической модели хранения конфиденциальной информации в центрах обработки данных общего назначения.

Ключевые слова: *центр обработки данных, распределенное хранение данных, информационная безопасность*

ВВЕДЕНИЕ

Специалисты исследовательской лаборатории армии США Александр Котт, Анантрам Свами и Брюс Вест в ноябре 2016 г. опубликовали статью «Туман войны в киберпространстве» (Kott A., Swami A., West Bruce. J. The Fog of War in Cyberspace // Article in Computer. – November, 2016. DOI: 10.1109/MC.2016.333), в которой анонсировали подход к обеспечению безопасности, когда данные разделяются на многочисленные фрагменты и постоянно распределяются по нескольким конечным вычислительным устройствам. Этот подход, как указывают авторы, может не только обеспечить большую отказоустойчивость информационных систем и предотвратить атаки, но и предоставить огромные технические проблемы. Название этой статьи породило название технологии «Кибертуман», которая нуждается в осмыслении и моделировании, а также в реальной статистической оценке. Выдающийся военный теоретик Карл фон Клаузевиц писал о «тумане войны», понимая его как фундаментальную неопределенность информации в сложной состязательности воюющих сторон (Клаузевиц К. О войне. – М.: Изд-ва: Эксмо, Мидгард, 2007. – 458 с.)

ПАРАМЕТРЫ И ЭЛЕМЕНТЫ АРХИТЕКТУРЫ ХРАНЕНИЯ

Система распределенного хранения информации (СРХИ, *далее* – система хранения) состоит из поставщиков информации, потребителей информации (в общем случае поставщики и потребители могут быть одинаковы), Центра управления распределенным хранением (ЦУРХ, *далее* – центр управления) и центров обработки данных (ЦОД). Поставщики направляют в центр управления информационные массивы, которые он распределяет по центрам обработки данных. При запросе на информационный массив центр управления «собирает» его из фрагментов, хранящихся в центрах обработки данных, и направляет потребителям. Информационный массив (или распределяемый информационный массив – РИМ) определяется именем I и длиной L, с ним также связаны свойства – конфиденциальность содержащейся в нем информации и необходимость зашифрования данного массива.

Система хранения характеризует следующие ключевые параметры:

1) общее число центров обработки данных – N;

2) число активных ЦОД на текущий момент времени – n;

3) ЦОД идентифицируется номером i, i= 1,…, N, множество активных ЦОД образует выборку мощностью n из N возможных и характеризуется вектором $A = (a_1, a_2,\ldots, a_N)$, где a_k принимает значение 0, если k-й ЦОД неактивен и 1 – если он активен, сумма компонентов вектора равна n;

4) безопасная доля информации B, $0<B<1$ (понятие, связанное с грифом конфиденциальности информации – по аналогии с инструкцией по закрытому делопроизводству: если документ имеет менее 10% (0,1) объема информации j-го грифа, то он относится к j–1-му грифу, из этого следует, что минимальное количество m ЦОД для распределения m = [1/B]+1, где в квадратных скобках результат вычисления выражения является целой частью числа, находящегося в квадратных скобках;

5) длина единичного блока e байт, на которые разбивается распределяемый информационный массив;

6) Центр управления – обязательный участник схемы, хранящий Главный файл распределенного массива (*далее* – главный файл) и описывающий расположение и свойства распределяемого информационного массива;

7) главный файл содержит: реальное имя распределяемого информационного массива, гриф конфиденциальности, ссылку на ключ шифра (если он тре-

* Публикация подготовлена в рамках работ по программе фундаментальных исследований Отделения математических наук РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения», а также работ, поддержанных РФФИ, грант № 15-07-08522.

буется), цепочку номеров центров обработки данных и идентификатор текущего обрабатываемого блока информации – ID в рамках этих центров (это может быть имя файла, в который записан текущий блок), а также индикаторы «архитектурной» избыточности.

В эту схему входят два датчика случайных чисел (ДСЧ): первый, генерирующий равномерно распределенную последовательность чисел, каждое из которых принимает значение от 1 до N, и второй, назначение которого мы поясним далее, а также код, исправляющий ошибки (КИО), который эмпирически должен создавать избыточность порядка одной трети длины e в байтах. Обозначим преобразование этого кода как $h=KIO(e)$.

С учетом возможной недоступности некоторых центров обработки данных в распределяемую информацию необходимо вводить «архитектурную» избыточность, т.е. дублировать распределяемый массив в нескольких ЦОДах. Статистическую оценку этой величины представим как $(N-n)/N$ – вероятность того, что в некоторый момент времени центр обработки данных недоступен и записанный в его массивах хранения блок не может быть прочитан – это минимальная вероятность потери блока. Соответственно, приблизительно с такой же вероятностью очередной блок надо записывать еще в один ЦОД. Таким образом, второй датчик случайных чисел (ДСЧ-2) в каждом текущем распределяемом блоке с заданной вероятностью $p>(N-n)/N$ должен давать команду на дублирование этого блока в другом центре обработки данных.

Примерный алгоритм работы Центра управления распределенным хранением:

1) информационный массив разделяется на x блоков по e байт;

2) циклически (x циклов) проводится процедура обработки массива по блокам;

3) текущему блоку вырабатывается корректирующая последовательность h;

4) вырабатывается случайное число i центров обработки данных и приводится к модулю n;

5) выбирается i-й ЦОД и в него записываются блоки e и h;

6) на i-м центре проверяется h и записывается блок e, центру управления передается ID блока, сохраненного в i-м центре обработки данных;

7) с помощью ДСЧ-2 с заданной вероятностью p вырабатывается команда архитектурного дублирования и повторяются шаги 4-6 для другого центра обработки данных;

8) в главный файл распределенного массива добавляется номер центра обработки данных и ID блока (при архитектурном дублировании два раза).

При этом должны быть предусмотрены счетчики, которые предупреждают излишнюю запись в центрах обработки данных (когда превышена безопасная доля для конкретного центра), в этом случае i-й ЦОД в векторе a становится «временно неактивным» ($a_i=0$). В случае включения шифра вопрос о безопасной доле, естественно, снимается. Описанный алгоритм и технологию далее мы предлагаем называть «Кибертуман».

ОСНОВНЫЕ ТЕНДЕНЦИИ РАЗВИТИЯ СОВРЕМЕННЫХ ТЕХНОЛОГИЙ ХРАНЕНИЯ ДАННЫХ

Более 2-х квинтиллионов байт (эксабайт) генерируются каждый день по всему земному шару, из них 85% – неструктурированные (медиа-файлы: аудио, фото, видео, электронная почта, коммерческая документация, социальные сети и т.д.). Безопасный и надежный способ хранения постоянно растущих объектов данных является одним из основных вопросов ИТ-директоров. Для решения этой проблемы многие делают выбор в пользу облачных решений.

Объем рынка облачного хранения данных увеличится с $23,76 млрд в 2016 г. до $74,94 млрд к 2021 г. (исследование Markets and Markets). Таким образом, этот рынок будет расти на 25,8% ежегодно.

Большинство современных технологий хранения данных основано на увеличении количества копий данных. Компании делают несколько резервных копий для увеличения надежности. Для предотвращения утечек конфиденциальной информации их шифруют и устанавливают к ним многоуровневый контроль доступа.

Снижение затрат

Помимо видимых преимуществ в отношении надежности, безопасности и скорости, система распределенного хранения информации экономит еще и объем памяти. Увеличение количества копий является обычной практикой для компаний, понимающих недостаток применения обычных RAID-массивов (Redundant Array of Independent Disks – избыточный массив независимых дисков).

Копии стоят дорого: каждая копия требует более 133% дополнительного объема дискового пространства, в случае использования стандартной конфигурации RAID 6. Корпорации часто прибегают к территориально-распределенному хранению данных. Некоторые компании делают по 2, 3 или даже 4 копии.

При классическом подходе увеличение затрат на 500% повышает надежность так же, как и при использовании системы распределенного хранения информации.

Превосходя крупных провайдеров облачных систем хранения данных в надежности, безопасности и скорости, система распределенного хранения информации сравнима по цене с двумя или более узлами хранения.

Масштабируемость хранилищ данных

Основной проблемой большинства облачных хранилищ является ограничение размера одного файла, как правило, это сотни мегабайт. В свете растущих потребностей хранения больших данных (BIGDATA), система распределенного хранения информации позволяет без изменения существующей инфраструктуры получить практически не ограниченное по объёму и скорости хранилище данных.

Повышение надежности и доступности систем хранения данных

В классических системах хранения данных каждая дополнительная копия дает приближение к 100% надежности. Добавление еще 30% избыточности описан-

ного алгоритма работы центра управления распределенным хранением означает, что данные будут доступны и защищены от потерь, даже если 30 из 100 узлов памяти одновременно выйдут из строя. Это соответствует 24-м девяткам надежности (99,99999999999999999999%) и 12-ти девяткам доступности.

Узлы хранения данных могут быть распределены по десяткам и сотням географических местоположений. Эта отказоустойчивая система обеспечивает беспрецедентную защиту от какого-либо внешнего или внутреннего отказа.

Безопасность

Типичные системы хранения данных применяют сложный контроль доступа и шифрование для обеспечения безопасности и предотвращения утечек. В случае несанкционированного доступа на уровне администратора все данные в системе находятся под угрозой.

«Кибертуман» гарантирует защиту данных от кражи, утечки или несанкционированного доступа. Любой файл воспринимается программой как поток байтов. Этот поток разделяется на блоки данных, которые смешиваются с закодированными маркерами для обеспечения избыточности и «распыляются» на сотни территориально-распределенных узлов. Отдельный узел хранит не весь исходный файл – а только небольшую закодированную его часть. Каждый блок имеет закрытый ключ доступа. Если случайно злоумышленники смогут взломать один или несколько узлов, то они получат только набор блоков (фрагментов) данных, недостаточных для понимания полного содержания файла. Реконструкция всех блоков (фрагментов) может быть выполнена только владельцем данных, который имеет главный ключ. Без ключа злоумышленник не сможет узнать алгоритм соединения блоков (фрагментов), так что даже если найти соответствующие фрагменты в других узлах, их корректное соединение будет крайне сложным.

Увеличение скорости передачи данных

Достижение высокой скорости передачи данных при перемещении их от одного узла хранения или от центра обработки данных маловероятно. Причина проста - один центр обработки данных имеет много клиентов. «Кибертуман» уравнивает этот дисбаланс: один клиент обслуживается десятками узлов. Каждый узел, используемый алгоритмом с точки зрения пропускной способности канала не лучше или не хуже, чем любой другой типичный современный центр обработки данных. Но при их объединении они обеспечивают увеличение скорости более чем в 10 раз по сравнению с любыми другими системами, даже если они будут использоваться независимо друг от друга.

Практические тесты скорости облачных хранилищ показали, что увеличение скорости осуществляется не только при передаче данных через множество разных каналов связи, но и при передаче файла частями с использованием множества параллельных потоков.

Дополнительный прирост скорости можно получить применив несколько операторов связи при передаче данных. Перехватить весь объём данных при передаче через один канал значительно проще, чем при передаче данных через несколько независимых каналов операторов.

Причем можно использовать гибридные виды связи (GPRG, LTE, WIFi, 3G и прочие) одновременно, что особенно актуально для регионов, в которых отсутствуют высокоскоростные каналы связи. Применение такой технологии позволяет одновременно использовать все доступные виды связи и многократно поднять скорость и безопасность передачи данных.

Преимущества технологии «Кибертуман»

1. Система хранения данных на базе технологии «Кибертуман» может быть создана с использованием существующих ресурсов государственных и коммерческих центров обработки данных или узлов хранения.
2. Для достижения максимальной территориальной распространенности описанный алгоритм работы центра управления распределенным хранением сохраняет различные блоки (фрагменты) кодированной информации в разных узлах.
3. Каждый узел содержит минимум 5 серверов хранения данных, которые изначально оснащены 5 ТБ каждый, объем легко модифицируется до 45 ТБ.
4. Общий объем 200 таких узлов хранения данных будет составлять 5 петабайт, которые легко масштабируются до 45 петабайт без добавления новых серверов.
5. Узлы хранения территориально распределены и будут обеспечивать стабильный быстрый доступ к хранилищу данных в режиме 24/7.

ЗАКЛЮЧЕНИЕ

Рассмотренный набор параметров и архитектурных решений может быть основой для разработки общих и ведомственных положений и классификаций, а также выбора и оценки технических решений в области архитектуры и математической модели хранения конфиденциальной информации в центрах обработки данных общего назначения для широкого круга информационных систем.

Материал поступил в редакцию 20.03.17.

Сведения об авторах

ЗАЙЦЕВ Андрей Викторович – заместитель директора по технической части ООО «МЕДИА», Москва
e-mail: avzajcev@mail.ru

ГОСТЕВ Сергей Сергеевич – кандидат технических наук, заместитель генерального директора по науке, «Концерн ГРАНИТ», Москва
e-mail: gostevss@mail.ru

ЧЕРКАШИН Павел Александрович – заместитель генерального директора по информационным технологиям, «Концерн ГРАНИТ», Москва
e-mail: cherkashin@granit-concern.ru

ЩЕРБАКОВ Андрей Юрьевич – доктор технических наук, профессор НИУ ВШЭ, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, Москва
e-mail: x509@ras.ru