На правах рукописи

Иваничкина Людмила Владимировна

МАТЕМАТИЧЕСКИЕ МОДЕЛИ НАДЕЖНОСТИ И МЕТОДЫ ЕЕ ПОВЫШЕНИЯ В СОВРЕМЕННЫХ РАСПРЕДЕЛЕННЫХ ОТКАЗОУСТОЙЧИВЫХ СИСТЕМАХ ХРАНЕНИЯ ДАННЫХ

Специальность 05.13.11 - Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва – 2018

Работа выполнена на кафедре информатики и вычислительной математики Московского физико-технического института (государственного университета).

Научный руководитель:

Коротаев Кирилл Сергеевич,

кандидат физико-математических наук, вице-президент по разработке, резервное копирование и хранение данных ООО "Акронис".

Официальные оппоненты:

Богатырев Владимир Анатольевич,

доктор технических наук, профессор кафедры вычислительной техники в Федеральном Государственном Автономном Образовательном Учреждении Высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», почетный работник науки и техники РФ

Незнанов Алексей Андреевич

кандидат технических наук, доцент департамента анализа данных и искусственного интеллекта факультета компьютерных наук НИУ ВШЭ, старший научный сотрудник лаборатории интеллектуальных систем и структурного анализа НИУ ВШЭ

Ведущая организация:

Автономная некоммерческая образовательная организация высшего профессионального образования «Сколковский институт науки и технологий»

Защита состоится 24 мая 2018 г. в 16 часов на заседании диссертационного совета Д 002.087.01 при Федеральном государственном бюджетном учреждении науки «Институт системного программирования им. В.П. Иванникова Российской академии наук» по адресу: 109004, Москва, ул. А. Солженицына, 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки «Институт системного программирования им. В.П. Иванникова Российской академии наук».

Автореферат разослан «___» _____ 2018 г.

Ученый секретарь
диссертационного совета Д 002.087.01,
кандидат физико-математических наук Зеленов С.В.

Общая характеристика работы

Актуальность темы исследования

Надежное хранение больших объемов информации является одним из основополагающих требований современной информационной инфраструктуры. Объем информации, производимой человечеством, удваивается каждые полтора – два года, при этом все большая часть этой информации хранится не на устройствах пользователя, а в 'облачных' хранилищах. Стремительный рост вычислительной мощности компьютеров, опережающий потребности отдельных пользователей, создает предпосылки для развития облачных сервисов, а также способствует консолидации информационной инфраструктуры предприятий на базе гиперконвергентных кластерных решений, сочетающих надежное распределенное хранилище данных с возможностью запуска виртуальных машин. Подобные решения оказываются предпочтительными, поскольку обеспечивают более высокую надежность, чем выделенные сервера, за счет развитых систем резервирования и возможности миграции данных в случае отказа оборудования.

Предъявляемые к современным системам хранения данных (СХД) требования выходят далеко за пределы возможностей единичного дискового накопителя как по объему хранимой информации, так и по надежности и производительности. Удовлетворить этим требованиям можно только на базе совершенно новых технических решений.

Методы построения отказоустойчивых СХД на базе нескольких дисковых накопителей зародились примерно 3 десятилетия назад. Это были так называемые RAID-массивы – ряд дисков, объединенных в единое хранилище. Повышение надежности такого хранилища достигалось за счет хранения избыточной информации (контрольных сумм) на одном или нескольких дополнительных дисках. При отказе одного из дисков данные можно было

восстановить, используя эту избыточную информацию. Подобные решения справлялись со своими задачами пока необходимое число дисков в хранилище не превышало десятка. Проблема, возникающая при дальнейшем увеличении числа дисков, связана с тем, что вероятность отказа одного из множества дисков в единицу времени прямо пропорциональна их количеству. Если к этому моменту не завершится восстановление после предыдущего отказа, данные могут быть потеряны.

Получение приемлемой надежности в хранилище данных, объединяющем тысячи дисков, требовало принципиально иных технических решений. Таким решением стало разбиение сохраненных данных на фрагменты, где каждый фрагмент хранится в виде набора дисковых блоков на различных дисках в СХД с необходимой для обеспечения надежности избыточностью. Увеличение надежности такого хранилища по сравнению с RAID-массивом достигается за счет того, что в случае отказа одного из дисков, поврежденные фрагменты оказываются распределены по всем дискам в такой системе. Соответственно, восстановление избыточности после потери диска также происходит одновременно на всех дисках в системе. Значит, время восстановления уменьшается с ростом числа дисков, эффективно компенсируя увеличение вероятности отказа любого из дисков с ростом их количества.

Степень разработанности темы

Построение надежного хранилища, способного не только противостоять отказам оборудования, но и обеспечивать высокую скорость доступа к данным для множества одновременно работающих с ним клиентов, является серьезной технической проблемой, не имеющей одного единственно верного решения. Всякое решение является компромиссом с точки зрения множества факторов. Успешное достижение такого компромисса возможно только на базе теоретических изысканий, имитационного моделирования и опыта практической

реализации СХД. Задача осложняется еще и тем, что математических моделей, адекватно описывающих СХД, построенную на базе описанных выше принципов, практически не существовало.

Известные до настоящего времени математические модели надежности СХД основаны на представлении системы в виде цепей Маркова. Зная набор состояний системы и вероятности переходов между ними в единицу времени, можно вычислить среднее время перехода системы в поглощающее состояние, соответствующее отказу (потере данных в нашем случае). Это время называется средним временем наработки до отказа. Мы будем использовать его в качестве количественной метрики надежности СХД. В дальнейшем для краткости изложения мы не будем делать различий между надежность и средним временем наработки до отказа (потери данных) для исследуемой системы.

Марковская модель адекватно описывает надежность RAID-массива, но совершенно не подходит для описания СХД с разбиением данных на фрагменты, поскольку они уже не могут рассматриваться, как эволюционирующие независимо друг от друга Марковские процессы. Отказ одного диска в такой системе приводит к потере сразу множества блоков данных, а их восстановление происходит не независимо, а последовательно со скоростью, которая зависит от суммарной производительности дисков, оставшихся неповрежденными.

Цель работы

Целью настоящего исследования является построение и анализ математических моделей надежности распределенного отказоустойчивого хранилища данных с разбиением данных на фрагменты и реализация такого хранилища на базе полученных в ходе исследования результатов. Особое внимание уделено методам повышения надежности СХД.

Методы исследования

Исследование начинается с изучения основных результатов, полученных в Марковской модели надежности хранения данных. Рассматривается метод моделирования стохастической системы путем сведения ее к эргодической с последующим рассмотрением ее стационарного режима. Этот метод используется для исследования разработанной нами математической модели СХД, более адекватной рассматриваемой системе, чем Марковская модель. Полученные в ходе исследования результаты верифицируются путем сравнения с результатами симуляции на моделях, имитирующих реальную СХД, а также путем сравнения со статистикой, накопленной при использовании реальной системы хранения данных, созданной нами на основе разработанных математических моделей.

Для построения основных моделей работы используются теория вероятностей, теория алгоритмов, широко используется аппарат математического анализа.

Новизна работы

1. Предложена математическая модель надежности СХД с разбиением данных на фрагменты, более адекватно описывающая реально существующие системы хранения данных, чем Марковская модель. Разработан приближенный метод оценки надежности СХД, вытекающий из свойств модели.
2. Разработан комплекс имитационных моделей надежности СХД, отражающий реальную архитектуру и особенности реализации подобных систем.
3. Исследовано влияние различных факторов на надежность и масштабируемость СХД. Впервые изучено теоретически и проверено с помощью имитационных моделей влияние различных политик размещения

дисковых блоков на надежность хранилища. Получены количественные оценки влияния скрытых повреждений на надежность СХД. Предложены методы борьбы со скрытыми повреждениями и даны оценки их эффективности.

На базе полученных теоретических результатов построена распределенная система хранения данных, гарантирующая высокую надежность за счет использования различных схем обеспечения избыточности и оптимизации надежности.

Теоретическая и практическая ценность работы

Разработанные нами методы построения математических моделей реальных СХД зарекомендовали себя удобным инструментом анализа надежности подобных систем, значительно более адекватным реальной архитектуре хранилища, чем применявшиеся ранее Марковские модели. Применение этих методов позволило нам обосновать уже известные из практики способы повышения надежности и масштабируемости СХД, а также получить новые, ранее неизвестные результаты.

Разработанные нами математические модели могут быть использованы для оценки надежности хранения данных применительно к широкому классу СХД с разбиением данных на фрагменты независимо от деталей их реализации. Применение этих математических моделей позволяет подобрать параметры хранилища, обеспечивающие оптимальное сочетание надежности и накладных расходов, с учетом потребностей конкретных пользователей СХД, а также смоделировать варианты дальнейшего повышения надежности хранилища.

Построенная на базе полученных теоретических результатов система хранения данных «Acronis Storage» показала свою высокую надежность и коммерческую эффективность при хранении данных корпоративных клиентов в

десятках стран мира. Созданная нами СХД позволяет объединить до 10,000 дисков в единое хранилище, обеспечивающее высокую надежность, автоматическое восстановление при отказах оборудования и производительность, сравнимую с производительностью локального дискового накопителя при одновременной работе тысяч клиентов. В настоящее время в различных инсталляциях «Acronis Storage» хранится более 100 петабайт пользовательских данных по всему миру. За более чем 5 лет эксплуатации системы не было ни одного случая потери данных вследствие отказа оборудования, что демонстрирует эффективность технических решений, созданных на базе полученных в ходе исследования теоретических результатов.

Положения, выносимые на защиту

1. Построена математическая модель надежности СХД с разбиением данных на фрагменты, более адекватно описывающая реально существующие системы хранения данных, чем Марковская модель.
2. Создан комплекс имитационных моделей надежности СХД, отражающий реальную архитектуру и особенности реализации подобных систем.
3. В рамках математической модели получена зависимость надежности от количества различных дисковых кортежей (групп размещения), используемых для размещения дисковых блоков.
4. Исследовано влияние скрытых повреждений на надежность СХД. В рамках математической модели получена зависимость надежности от вероятности появления скрытых повреждений в единицу времени и интенсивности проверок целостности данных (скраббинга). Доказана теорема, дающая нижнюю границу на вероятность отказа при наличии скрытых повреждений.

Апробация результатов работы

Результаты диссертационного исследования докладывались, обсуждались и получили одобрение специалистов на российских и международных научных конференциях:

- IX Международная научно-практическая конференция "Актуальные проблемы науки XXI века", г.Москва, 2016;
- Международная научно-практическая конференция «Приоритетные задачи и стратегии развития естественных и математических наук», г.Тольятти, 2016;
- IV Международная научно-практическая конференция «Наука XXI века : проблемы и перспективы», г.Уфа, 2016;
- VIII Международная научно-практическая конференция «Инновационные технологии нового тысячелетия», г.Киров, 2016;
- Конференция «Облачные вычисления. Образование. Исследования. Разработка», г.Москва, 2015;
- 58-я научная конференция МФТИ, г.Долгопрудный, 2015.

Результаты работы реализованы в виде программного комплекса по оценке надёжности СХД с помощью приближенно-аналитических математических моделей, а также вычислительных имитационных методов статистических испытаний. Исходный код доступен в публичном репозитории [11,12].

Представленные в диссертации, разработанные нами методы и программные средства повышения надёжности и высокой доступности СХД включены в реализуемые ООО «Проект ИКС» технологии и программное обеспечение распределенных и высокопроизводительных вычислительных систем для хранения и обработки больших данных.

Диссертационная работа была выполнена рамках проведения прикладных научных исследований, выполняемых по Соглашению с Министерством науки и образования о предоставлении субсидии №14.579.21.0010 от "05"июня 2014 года

по теме "Технология и программное обеспечение распределенных и высокопроизводительных вычислительных систем. Хранение и обработка больших данных". Уникальный идентификатор Соглашения RFMEFI57914X0010.

Публикации по теме диссертации

По материалам данного диссертационного исследования были опубликованы работы [1-10]. В списке изданий присутствуют рекомендованные ВАК ([4-9]), в том числе индексируемые системами WebOfScience и SCOPUS ([5-9]), два свидетельства о регистрации программы ([2,3]) и заявка на патент [10]. В работах [2-6,8], выполненных в соавторстве, автору принадлежит материал, связанный с постановкой задачи, проведениями исследований и формулировкой математических моделей. В работе [7] вклад автора заключался в разработке имитационной модели, проектировании и реализации вычислительных алгоритмов, проведении вычислений и анализе результатов. В работе [9] вклад автора заключался в разработке и реализации алгоритма распределения дисковых блоков с учетом областей отказов, проектировании и реализации имитационной модели для оценки надежности хранилища, проведении вычислений и анализе полученных результатов.

Структура и объём диссертации

Диссертация состоит из введения, четырех глав, заключения и списка использованных источников. Диссертация изложена на 130 страницах, включает 12 таблиц и 34 рисунка.

Содержание работы

Во введении обосновывается актуальность темы исследования, научная и практическая ценность работы, её научная новизна, кратко излагается содержание и структура диссертации.

В первой главе дается развернутый анализ требований, предъявляемых к современным системам хранения данных. Далее рассматриваются существующие решения и история их развития, анализируются их преимущества и недостатки. Рассматриваются известные из литературы модели надежности систем хранения данных.

Во второй главе рассматриваются математические модели надежности хранения данных. Вводится понятие *схемы хранения*, которая характеризуется двумя целыми числами n и k. Первое равно количеству дисковых блоков, которые используются для хранения одного фрагмента данных пользователя (этот набор блоков мы будем также называть *кортежем*). Второе равно минимальному количеству неповрежденных дисковых блоков, достаточному для восстановления исходных данных. Возможность СХД противостоять отказам отдельных дисков основана на том, что все n дисковых блоков сохраняются на разных дисках. При этом, данные можно восстановить при отказе до $n-k$ отдельных дисков. Это число называется *избыточностью*. Далее мы можем изучать свойства системы и делать выводы о ее надежности абстрагируясь от конкретных механизмов хранения данных, используя только известные нам параметры n и k, а также информацию о вероятности отказа диска в единицу времени и о том, насколько быстро система восстанавливается после отказа единичного диска.

Классический способ оценки надежности хранения данных с (n, k) схемой хранения – представление эволюции такой системы в виде цепи Маркова. Рассмотрим набор состояний каждого фрагмента данных, занумеровав их целыми числами соответственно количеству утраченных дисковых блоков.

Состояние 0 будет соответствовать неповрежденному кортежу из дисковых блоков. Отказ диска будет приводить к переходу из состояния i в состояние $i+1$. Восстановление утраченного дискового блока будет соответствовать переходу из состояния i в состояние $i-1$. Переход в состояние с индексом $n-k+1$ будет означать необратимую потерю данных. Такие состояния называются поглощающими, поскольку, однажды попав в него, система остается в нем навсегда. Поскольку вероятности перехода между состояниями в такой модели не зависят от предыстории, то такая система является классической цепью Маркова с непрерывным временем.

Цепь Маркова с непрерывным временем удобно описывать с помощью матрицы интенсивностей переходов $\boldsymbol{Q}$. Недиагональные элементы Q_{ij} равны вероятности перехода из состояния i в состояние j *за единицу времени.* Диагональные элементы Q_{ii} равны вероятности ухода из состояния i в любое другое (со знаком минус) *за единицу времени*. Пусть $p_i(t)$ – вероятность обнаружить систему в состоянии i в момент времени t. Тогда для $p_i(t)$ справедливо уравнение Колмогорова:

$$\frac{d}{dt}p_i = \sum_j p_j(t)Q_{ji} \qquad (1)$$

Решив уравнения Колмогорова (1), мы можем найти $p_i(t)$ в явном виде. Однако, в теории надежности нас интесует лишь среднее время до попадания системы в поглощающее состояние – так называемое среднее время наработки до отказа (Mean Time To Failure - MTTF). Зная матрицу интенсивностей переходов, это время, можно найти следующим образом. Пусть $\widetilde{\boldsymbol{Q}}$ – матрица интенсивностей переходов, в которой удалены строки и столбцы, соответствующие поглощающим состояниям, а $\boldsymbol{t}$ – результат решения системы линейных уравнений $\boldsymbol{t} * \widetilde{\boldsymbol{Q}} = [-1, 0, \dots 0]$. Тогда среднее время наработки до отказа равно:

$$T_F = \sum_i t_i \qquad (2)$$

Здесь t_i – это среднее время пребывания системы в состоянии i до того, как она попадет в поглощающее состояние.

Для нахождения среднего времени наработки до отказа мы будем пользоваться следующей методикой. Изменим нашу модельную систему так, что при попадании в состояние, где происходит потеря данных, она сразу же переходит в неповрежденное состояние 0. Можно считать, что после потери данных мы сразу же создаем новую неповрежденную систему и продолжаем наблюдать за ее эволюцией. Такая Марковская цепь будет эргодической, т.е. вероятность со временем попасть из некоторого состояния в любое другое для нее отлична от нуля. Эргодическая Марковская цепь характерна тем, что для нее существует стационарное решение, которое можно найти, решив систему уравнений

$$\boldsymbol{p} * \boldsymbol{Q} = \mathbf{0} \qquad (3)$$

Чтобы найти среднее время наработки до отказа исходной системы, нам достаточно усреднить вероятность потери данных в единицу времени для модифицированной системы и найти величину, обратную этому среднему. Пользуясь свойством эргодичности, мы можем заменить усреднение по времени усреднением по ансамблю, рассмотрев набор произвольного количества независимых реализаций модифицированной системы. Данный метод оказывается более удобным, чем рассмотрение эволюции системы в терминах вероятностей попадания в различные состояния, поэтому он широко используется в нашем исследовании.

Хотя формула (2) позволяет нам найти среднее время наработки до отказа для произвольной Марковской цепи в аналитическом виде, количество членов в таком выражении растет экспоненциально с ростом количества состояний. Поэтому нами предложен простой метод нахождения приближенно-асимптотического решения для случая, когда среднее время между дисковыми отказами много больше времени восстановления после отказа. Введем понятие

уровня деградации некоторого состояния, как минимальное количество отказов, которое приводит в него из неповрежденного состояния. Можно показать, что уровень деградации определяет нижнюю и верхнюю границу вероятности обнаружить систему в данном состоянии. Это позволяет нам игнорировать переходы, вносящие пренебрежимо малый вклад в уравнения для стационарного состояния (3). В частности, для произвольной (n, k) схемы хранения данных можно получить простое выражение для среднего времени наработки до отказа T_F в зависимости от среднего времени до отказа любого из дисков в хранилище T_1 и среднего времени, необходимого для восстановления утраченного дискового блока T_R:

$$T_F = T_1 \frac{(k-1)!}{n!} \left[\frac{T_1}{T_R}\right]^{n-k} \qquad (4)$$

Формула (4) демонстрирует несколько важных закономерностей. Во-первых, выигрыш в надежности по сравнению с хранением без избыточности (когда $n = k$) оказывается пропорционален отношению среднего времени до отказа диска к времени восстановления в степени, равной избыточности. То есть, чем больше избыточность и чем меньше время восстановления, тем выше надежность. Во-вторых, коэффициент пропорциональности быстро падает с увеличением количества дисковых блоков n примерно, как n в степени, равной избыточности. Максимальную надежность имеет схема $(n, 1)$, например, когда данные сохраняются в n идентичных копиях (в этом случае их называют *репликами*). Но она имеет и самую низкую из всех схем с равной избыточностью эффективность использования дискового пространства, поскольку данные при хранении занимают в n раз больше места, чем их размер. Увеличение эффективности использования дискового пространства возможно при увеличении количества блоков и сохранении избыточности. Но при этом страдает надежность хранилища. Следующий рисунок иллюстрирует эту зависимость для различных схем хранения с избыточностью 2. Надежность

показана в условных единицах – за единицу на вертикальной шкале принята надежность хранения в репликах.

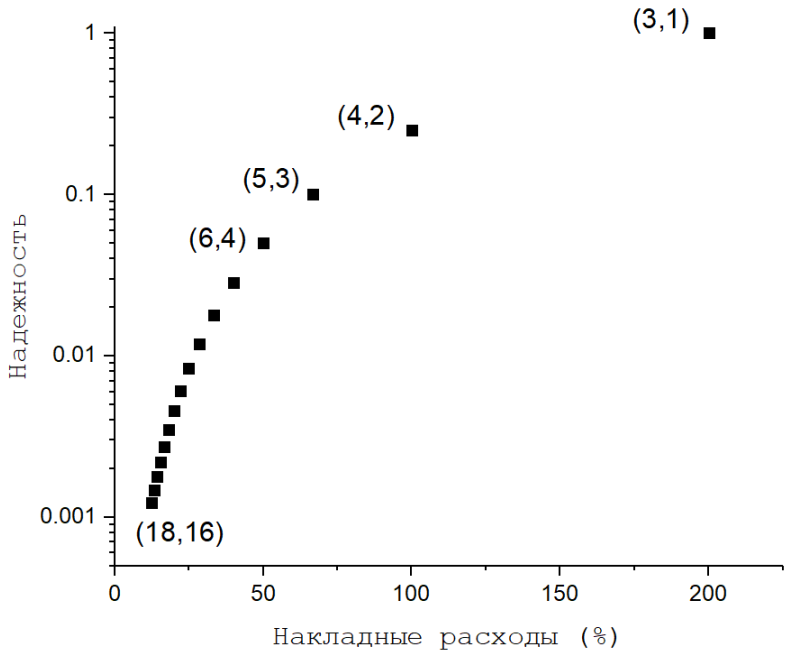

Рисунок 1. Взаимосвязь надежности и накладных расходов для различных схем хранения данных с избыточностью 2.

Формула (4) была получена нами для одного фрагмента данных, нас же будет интересовать среднее время до потери любого фрагмента данных в хранилище. Можем ли мы использовать формулу (4) для нахождения этого времени, считая, что количество фрагментов C нам известно? Хотя в литературе такой подход встречается повсеместно, он является в корне неправильным. Проблема заключается в том, что отказ одного диска приводит к повреждению всех хранящихся на нем блоков, а их восстановление происходит не независимо, а последовательно со скоростью, ограниченной производительностью дисков. Поэтому T_R в формуле (4) уже не может считаться постоянной и известной нам величиной. Следующий рисунок иллюстрирует процесс восстановления в системе с 3 дисками и схемой хранения с $n = 2$. Как видно из рисунка, в процессе восстановления после отказа одного из дисков участвуют все оставшиеся работоспособными диски.

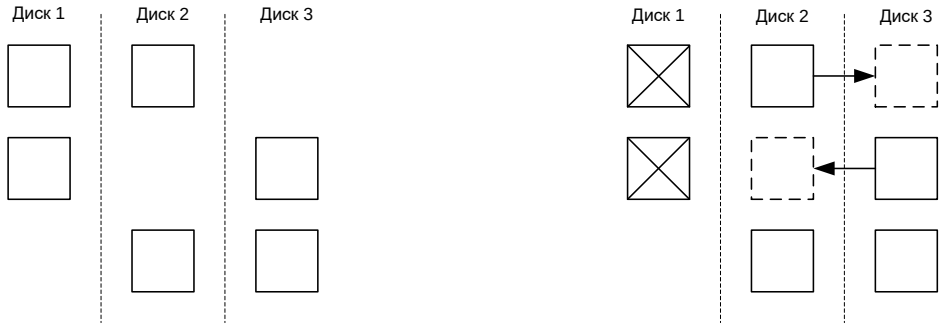

Рисунок 2. Восстановление после потери диска в многодисковом хранилище ($n = 2$).

Чтобы проверить, насколько формула (4) может соответствовать надежности реальной системы, мы построили две имитационные модели, соответствующие поведению реальной СХД с различной степенью точности. Как показали проведенные на этих моделях эксперименты, формула (4) правильно описывает зависимость надежности хранилища от среднего времени до отказа диска T_1. Однако при этом она не позволяет нам сделать никаких выводов о зависимости надежности от прочих параметров реальной СХД, например, от количества дисков в системе. Между тем, эта зависимость является ключевой для построения *масштабируемого* хранилища, то есть такого, надежность которого не ухудшается с ростом количества дисков (до определенного предела).

Чтобы построить модель надежности, более адекватно описывающую реальную СХД, мы рассмотрели систему, содержащую C фрагментов данных, распределенных по N дискам, где эволюция каждого фрагмента уже не описываются моделью Марковской цепи. Вместо моделирования состояния отдельного фрагмента, мы будем описывать общее состояние системы следующим образом. Присвоим каждому фрагменту уровень деградации, равный количеству утраченных дисковых блоков этого фрагмента. Пусть $p_i(t)$ – количество фрагментов на уровне деградации i в момент времени t. Мы будем описывать состояние системы двумя наборами параметров: $P_i = \langle p_i(t) \rangle_t$ – среднее количество фрагментов на уровне деградации i, и $F_i = \Pr(p_i(t) > 0)$ – вероятность обнаружить ненулевое количество фрагментов на уровне

деградации i. Пусть λ – вероятность отказа любого диска в единицу времени. Обозначим избыточность схемы хранения как $m = n - k$. Вероятность потери данных в единицу времени p_F может быть найдена из вероятности обнаружить ненулевое количество чанков на уровне деградации $m = n - k$ следующим образом:

$$p_F = F_m * \lambda \qquad (5)$$

Будем считать, что пока $p_i(t) > 0$, восстановление с уровня i на уровень $i - 1$ идет с постоянной скоростью φ, которая определяется суммарной производительностью неповрежденных дисков. Тогда средняя скорость восстановления просто равна φF_i. Рассмотрим стационарное состояние системы, считая, что мы компенсируем потерянные фрагменты добавлением в систему новых. Для такого стационарного состояния можно получить следующие уравнения, связывающие параметры системы на разных уровнях деградации:

$$F_i = \frac{\lambda}{\varphi}\frac{n-i}{N}P_i \qquad (6)$$

$$P_i = \frac{\alpha_{i-1}}{2}\frac{\lambda}{\varphi}\left[\frac{n-i+1}{N}\right]^2\frac{P_{i-1}^2}{F_{i-1}} \qquad (7)$$

здесь α_{i-1} – коэффициент, зависящий от формы функции распределения $p_i(t)$. Можно показать, что $\alpha_0 = 1$, $\alpha_1 = 4/3$. Уравнения (5), (6), (7) позволяют нам найти вероятность потери данных в единицу времени p_F, при условии, что параметры нулевого уровня деградации нам известны: $F_0 = 1, P_0 = C$.

Чтобы проверить разработанную нами математическую модель, мы сравнили ее предсказания с результатами имитационного моделирования. Результаты демонстрируют хорошую степень соответствия в пределах погрешности имитационной модели.

Результаты, полученные в рамках математической модели, позволяют нам определить степень масштабируемости системы в зависимости от избыточности, то есть выяснить вид зависимости ее надежности от количества дисков. Как показывает проведенный нами анализ, использование избыточности,

равной 2, гарантирует сохранение надежности с увеличением количества дисков, а использование избыточности 3 гарантирует увеличение среднего времени наработки до отказа с ростом числа дисков. Это объясняет, почему в реальных СХД избыточность менее 2 практически не применяется. На практике масштабирование системы рано или поздно начинает ограничиваться другими факторами, оставшимися за рамками нашего анализа. Например, пропускная способность сетевой инфраструктуры с ростом числа серверов в системе может начать ограничивать скорость восстановления, что приведет к падению надежности.

В третьей главе рассматриваются несколько важных факторов, влияющих на надежность СХД, такие как использование групп размещения, обнаружение скрытых ошибок и учет особенностей аппаратной инфраструктуры.

Группы размещения были изначально предложены как способ повысить надежность хранения данных путем ограничения числа возможных способов распределения дисковых блоков по дискам в системе. Кажется очевидным, что использование меньшего количества вариантов размещения блоков (их называют группами размещения) снижает вероятность потери данных в результате отказа $n-k+1$ дисков при использовании (n,k) схемы хранения просто потому, что для произвольного набора отказавших дисков может не найтись соответствующей группы размещения. Нами впервые было показано, что с учетом зависимости скорости восстановления от количества групп размещения, этот вывод является неправильным. Надежность всего хранилища падает при ограничении числа возможных способов размещения дисковых блоков. С другой стороны, надежность хранения определенной части данных, например, принадлежащей конкретному пользователю, может и увеличиться, если ограничить количество групп размещения для этой части. Кроме того, использование групп размещения позволяет уменьшить количество метаданных,

описывающих размещение дисковых блоков. Поэтому, группы размещения являются полезным инструментом, требующим вдумчивого использования.

Чтобы получить количественные оценки влияния групп размещения на надежность хранилища, мы добавили их в нашу математическую модель. Фактически, учет групп размещения сводится к тому, что на уровне деградации i дисковые блоки оказываются распределены не по всем оставшимся дискам, а по некоторому подмножеству от общего числа работоспособных дисков размером S_i. Это число мы будем называть разбросом дисковых кортежей. Мы получили аналитические оценки для S_i в зависимости от первоначального количества групп размещения, что позволило нам вычислить количество групп размещения, достаточное для сохранения надежности хранилища на приемлемом уровне.

Другим чрезвычайно важным фактором обеспечения надежности хранилища является *учет скрытых повреждений дисков*. Дисковые ошибки не всегда сводятся к полному отказу диска. Вторая категория ошибок – ошибки чтения сектора или группы секторов, которые чаще всего являются следствием врожденных дефектов или приобретенных повреждений поверхности вращающегося диска, либо запоминающих транзисторов твердотельного носителя информации. Отличительной особенность скрытых ошибок является то, что они остаются незамеченными до попытки чтения данных с диска, поэтому они и называются скрытыми или латентными. Понятно, что если данные не читаются с диска, то рано или поздно мы можем столкнуться с ситуацией, когда данные, хранящиеся в (n, k) схеме, уже невозможно восстановить, поскольку более чем $n - k$ блоков потеряно из-за скрытых ошибок. А при наличии дисковых отказов возможна ситуация, когда после отказа p блоков ($p \leq n - k$) данные невозможно восстановить вследствие ошибок чтения других q блоков, так что $p + q > n - k$. Поэтому, для надежного хранения данных критически важно постоянно проверять целостность блоков данных, сохраненных на каждом

диске. Обычно это делается путем последовательного чтения блоков данных и сравнения их содержимого с контрольной суммой, которая как правило хранится отдельно от данных. При этом фиксируются как ошибки, приводящие к невозможности чтения данных, так и ошибки, приводящие к чтению искаженных данных. Этот процесс обычно называется скраббингом. В рамках нашей математической модели мы получили количественные оценкаx влияния скрытых ошибок на надежность хранилища. Полученные результаты мы сформулировали в виде следующей теоремы.

Теорема о скрытых ошибках

Пусть T_B - среднее время до возникновения ошибки чтения любого блока на диске, T_d - среднее время до отказа диска целиком. Пусть T_R – время, необходимое для восстановления утраченного блока, τ – время, необходимое для проверки блока на наличие скрытых ошибок, причем блоки восстанавливаются и проверяются последовательно один за другим. Пусть вероятность потери данных в единицу времени вследствие одних только дисковых отказов, без учета скрытых ошибок, равна p_F. Обозначим вероятность потери данных в единицу времени с учетом скрытых ошибок как $\tilde{p}_F$. Тогда при использовании (n, k) схемы хранения справедливо следующее неравенство:

$$\tilde{p}_F > p_F \left(1 + \frac{k}{k+1} * \frac{T_d}{T_B} * \frac{\tau}{T_R}\right) \qquad (8)$$

Эта теорема имеет важное следствие для практической реализации СХД. Предположим, что восстановление происходит с максимальной скоростью, ограниченной производительностью диска, чтобы минимизировать p_F. Поскольку скорость проверки диска на наличие скрытых ошибок тоже ограничена производительностью диска, τ не может быть меньше, чем T_R. Как показывает практика, среднее время наработки до отказа конкретного сектора превышает время наработки до отказа всего диска на несколько порядков. Однако при этом среднее время наработки до возникновения ошибки чтения

любого из множества секторов диска оказывается в несколько раз меньше времени наработки до отказа диска. Отсюда с учетом (8) нетрудно сделать следующий вывод – *в реальных системах хранения данных скрытые ошибки будут основным источником потери данных*.

Как показывает проведенный нами анализ, скрытые ошибки влияют также и на масштабируемость СХД. Действительно, с ростом числа дисков возрастает вероятность обнаружения такого количества скрытых повреждений в кортеже дисковых блоков, которое сделает его восстановление невозможным. При этом скорость восстановления скрытых повреждений никак не меняется, что в итоге приводит к падению надежности хранилища. Уменьшить роль этого фактора падения надежности можно только уменьшив размер дискового блока, что уменьшит вероятность появления в нем скрытых повреждений. Однако уменьшение размера блока оказывается нежелательным, поскольку оно приводит к увеличению общего объема метаданных, описывающих распределение блоков по дискам. Нами предложен метод увеличения надежности, приводящий к тому же результату, что и уменьшение размера блока. Он сводится к делению блока на более мелкие фрагменты только на стадии восстановления с тем, чтобы иметь возможность восстановления утраченного блока при наличии скрытых повреждений отдельных фрагментов. Таким образом, противодействие скрытым повреждениям дисков является нетривиальной задачей, требующей принятия целого комплекса мер для сохранения надежности хранилища на приемлемом уровне.

Отказы оборудования не всегда сводятся к одним только дисковым отказам. Например, у сервера может отказать система питания или материнская плата, что приведет к недоступности всех подключенных к нему дисков. Аварии в системе питания или охлаждения, либо пожар могут привести и к более масштабным отключениям - целой стойки, на которой расположен сервер, либо целого ряда стоек, либо всего машинного зала центра обработки данных (ЦОД).

При этом менее масштабные отключения гораздо более вероятны, чем более масштабные. Учет этого обстоятельства при распределении блоков данных по дискам может существенно улучшить надежность СХД. Для этого мы ввели понятие *области отказов* как совокупности оборудования, подверженную одновременному отказу вследствие некоторых катастрофических событий. Области отказа образуют иерархию, где области отказа нижнего уровня (например, диски) являются частью областей отказа более высокого уровня (например, сервера). На каждом уровне иерархии области отказов образуют некоторое конечное множество (дисков, серверов, стоек и т.д.). Каждое такое множество областей отказов мы будем называть доменом отказов. Дисковый кортеж *устойчив к отказам* домена D, если все входящие в него диски принадлежат различным областям отказов из домена D. Это означает, что выход из строя любой области отказа домена D приведет к отказу не более одного диска в таком дисковом кортеже. Чем выше уровень домена D в иерархии, тем большую надежность обеспечивает дисковый кортеж, устойчивый к отказам из D, при равной избыточности схемы хранения. Следовательно, домен отказов является важным параметром при распределении блоков данных по дисковым кортежам.

При создании дисковых кортежей не следует ориентироваться ни на домен отказов самого нижнего – дискового – уровня (поскольку отказ сервера является достаточно частым событием), ни на домен самого верхнего уровня (поскольку рост загруженности сетевой инфраструктуры может нивелировать все преимущества в устойчивости к редким катастрофическим событиям). Оптимальным по соотношению надежности и накладных расходов способом хранения данных является создание кортежей, устойчивых к отказам серверного домена. При необходимости пользователи системы должны иметь возможность устанавливать необходимый им домен отказов на уровне папок или отдельных файлов также, как и схему хранения данных.

В четвертой главе рассматриваются результаты внедрения полученных в ходе исследования теоретических результатов при создании СХД «Acronis Storage», реализованного автором в составе коллектива разработчиков.

Для доступа к дисковым блокам в распределенной системе необходим специальный сервис, предоставляющий сетевой интерфейс для этого. В нашей системе этим сервисом является сервис чанков или CS (chunk service). Каждый диск в системе связан с единственным CS, который имеет монопольный доступ к этому диску и предоставляет сетевой доступ к нему для клиентов. Для того, чтобы клиенты могли обратиться к CS для доступа к данным, где-то должна храниться информация о соответствии между файлами, дисковыми блоками и CS, на которых они хранятся. Эта информация, очевидно, не может храниться на самом CS. Для ее хранения используется отдельный сервис метаданных или MDS (Meta Data Service). Он же используется и для хранения дерева имен файловой системы, а также для координации доступа к файлам и для восстановления дисковых блоков после сбоев. Все CS периодически посылают MDS регистрационные сообщения, так что MDS имеет полную информацию о том, какие CS находятся в рабочем состоянии, и знает их сетевые адреса. Эту информацию он использует для координации восстановления после сбоев, а также предоставляет клиентам в составе *карты блоков*, описывающей физическое расположение блоков определенного фрагмента файла пользователя. Наконец, третий компонент системы – *агент*, предоставляющий пользователям доступ к файловой системе посредством монтирования ее дерева имен в локальную файловую систему пользователя. Таких агентов в кластере может быть произвольное количество – как минимум по одному на каждый компьютер, имеющий доступ к СХД. Следующий рисунок иллюстрирует основные компоненты СХД «Acronis Storage» и их взаимодействие.

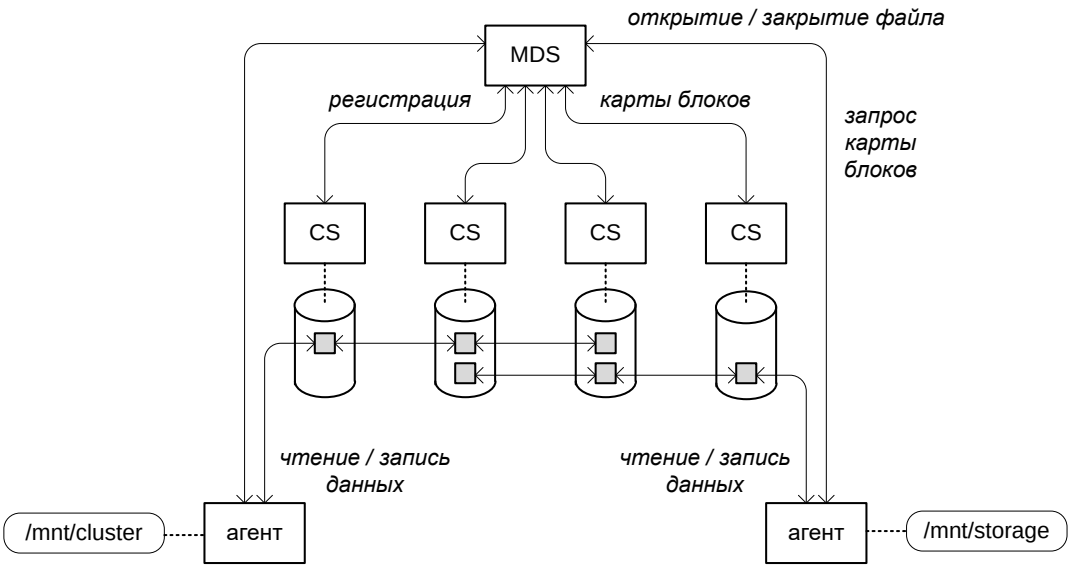

Рисунок 3. Основные компоненты СХД «Acronis Storage».

В процессе реализации системы хранения данных «Acronis Storage» мы провели всестороннюю проверку выбранных при ее проектировании параметров с учетом разработанной нами теоретической модели. В результате был принят ряд важных с точки зрения оптимизации надежности хранилища технических решений:

1. Мы отказались от использования схем хранения с избыточностью 1, как не обеспечивающих масштабирования системы.
2. В СХД реализовано кодирование с избыточностью 2 и 3, последнее рекомендуется для использования при наличии большого количества дисков.
3. Процедура восстановления после дисковых отказов обеспечивает последовательное восстановление утраченных дисковых блоков, так что каждый диск в любой момент времени участвует в восстановлении не более одного блока. Это позволяет достичь максимально возможной производительности дисковых операций, соответственно сократив время восстановления и повысив надежность хранилища.

При этом первыми восстанавливаются фрагменты данных с максимальным количеством утраченных блоков.

4. Используется максимально возможное количество групп размещения.
5. В СХД используется непрерывный скраббинг со скоростью порядка 1/10 от максимальной производительности дисков для раннего обнаружения скрытых повреждений дисков.
6. Схема хранения данных, а также домен отказов для их размещения в хранилище, настраиваются пользователем на уровне папок на файловой системе, так чтобы при необходимости обеспечить желаемый баланс между эффективностью хранения и надежностью в зависимости от цены потенциальной утраты конкретных данных пользователя.

Вышеперечисленные технические решения позволили нам реализовать отказоустойчивое хранилище данных, обеспечивающее высокую скорость доступа к файлам. В настоящее время СХД «Acronis Storage» широко используется сотнями корпоративных клиентов по всему миру, что подтверждает его эффективность и надежность.

В заключении указаны основные результаты и выводы диссертации.

Основные результаты и выводы диссертации

1. Предложена математическая модель надежности хранения данных в многодисковом хранилище с разбиением данных на фрагменты с заданной избыточностью более полно описывающее реальные СХД, чем модель Марковских цепей.
2. Исследовано влияние избыточности на масштабирование надежности хранилища с ростом числа дисков. Показано, что для сохранения надежности при добавлении дисков в систему необходимо использование значения избыточности не менее 2.

3. На базе созданной математической модели исследовано влияние ограничения числа возможных вариантов размещения дисковых блоков (групп размещения) на надежность хранилища. Показано, что такое ограничение уменьшает надежность хранилища. Найдена нижняя граница для количества групп размещения, обеспечивающее сохранение надежности хранилища на приемлемом уровне.
4. На базе созданной математической модели исследовано влияние скрытых повреждений на надежность хранилища. Доказана теорема, дающая нижнюю границу на вероятность отказа при наличии скрытых повреждений. Предложены методы борьбы со скрытыми повреждениями и дана оценки их эффективности.
5. Полученные результаты использованы при создании комплекса программ реализующего отказоустойчивое, масштабируемое хранение данных.

Список публикаций по теме диссертации

1. Иваничкина, Л.В. Методы обеспечения надежности хранения сверхбольших объемов данных в распределенной системе. / Л.В. Иваничкина // 58 научная конференция МФТИ. сб. науч. тр. — М.: МФТИ, 2015.
2. Иваничкина, Л. В., Бухтияров, П.А., Вареник, Р.В., Науменко, С.А., Непорада, А.Л. Имитационная модель надежности хранения данных, выполняющая симуляцию сбоев и процессов восстановления данных в распределенной системе хранения данных. Свидетельство о государственной регистрации программы для ЭВМ № 2015618800. 2015.
3. Иваничкина, Л.В., Бухтияров, П.А., Вареник, Р.В., Науменко, С.А., Непорада, А.Л. Имитационная модель процессов сбоев и восстановления данных для различных схем размещения данных в распределенной системе хранения. Свидетельство о государственной регистрации программы для ЭВМ № 2016613180. 2016.
4. Иваничкина, Л.В. Модель надежности распределенной системы хранения данных в условиях явных и скрытых дисковых сбоев. / Л.В. Иваничкина, А.П. Непорада // Труды Института системного программирования РАН, том 27, выпуск 6, 2015 г. стр. 253–274.
5. Ivanichkina, L. The reliability model of a distributed data storage in case of explicit and latent disk faults / L. Ivanichkina, A. Neporada // Journal of Engineering and Applied Sciences, 2015, – Т.10, №20, С. 9713—9724. http://www.arpnjournals.org/jeas/research_papers/rp_2015/jeas_1115_2928.pdf
6. Ivanichkina, L. Mathematical methods and models of improving data storage reliability including those based on finite field theory / L. Ivanichkina, A. Neporada // Contemporary Engineering Sciences. – Т.7. – №28. – 2014. – С. 1589 – 1602.

7. Ivanichkina, L. Computer Simulator of Failures in Super Large Data Storage / L. Ivanichkina, A. Neporada // Contemporary Engineering Sciences. – 2015. – 8(28). – C. 1679–1691. http://www.m-hikari.com/ces/ces2015/ces33-36-2015/p/ivanichkinaCES33-36-2015.pdf
8. Ivanichkina, L. Comparative study of LRC and RS codes / L. Ivanichkina, V. Vinnikov // Contemporary Engineering Sciences, 2016, vol. 9, no. 21, 1015-1029.
9. Ivanichkina, L. Failure resilient data placement policies for distributed storages / L. Ivanichkina, K. Korotaev, A. Neporada // Contemporary Engineering Sciences, 2016, vol. 9, no. 30, 1463-1489.
10. O. Volkov, A. Zaitsev, A. Kobets, L. Ivanichkina, K. Korotaev. Management of garbage data in distributed systems. US patent application 15/445,858. 2017.
11. [Электронный ресурс] https://github.com/ludmilai/markov_model
12. [Электронный ресурс] https://github.com/ludmilai/storage_model