Модель надежности распределенной системы хранения данных в условиях явных и скрытых дисковых сбоев

[1]Л. В. Иваничкина <livanichkina@odin.com>
[2]А.П. Непорада <Andrew.Neporada@acronis.com>
[1]ООО Проект ИКС, 127566, Россия, г. Москва, Алтуфьевское ш, дом 44
[2]ООО Акронис, 127566, Россия, г. Москва, Алтуфьевское ш, дом 44

Аннотация. В настоящей работе рассматривается подход к расчёту надёжности хранилища данных, учитывающий как явные дисковые сбои, так и скрытые битовые ошибки, а также процедуры их выявления. Для расчёта надёжности предлагается новая математическая модель аналитического описания распределенного хранилища, описывающая динамику потерь и восстановления данных на основе Марковских цепей, соответствующих разным схемам избыточного кодирования. Рассматриваются преимущества разработанной модели по сравнению с классическими моделями для традиционных RAID-массивов. Исследуется влияние скрытых ошибок жестких дисков без учёта других битовых сбоев, возникающих в остальных аппаратных компонентах вычислительной машины. Оценка надёжности осуществляется согласно новым аналитическим формулам на основе критерия среднего времени до отказа, при котором утрата данных превышает порог восстанавливаемости, определяемый параметрами избыточного кодирования. Приводятся новые аналитические зависимости среднего времени жизни хранилища от среднего времени полной проверки данных в нём.

Ключевые слова: Среднее время до отказа (MTTF); цепи Маркова; избыточное кодирование; задача Гюйгенса о разорении игрока; распределенное хранилище данных; процедура скраббинга; контрольные суммы; MTTDL распределенного хранилища данных; дисковые сбои; скрытые битовые ошибки; секторные ошибки.

DOI: 10.15514/ISPRAS-2015-27(6)-16

Для цитирования: Иваничкина Л.В., Непорада А.П. Модель надежности распределенной системы хранения данных в условиях явных и скрытых дисковых сбоев. Труды ИСП РАН, том 27, вып. 6, 2015 г., стр. 253-274. DOI: 10.15514/ISPRAS-2015-27(6)-16.

1. Введение

Хранилища данных петабайтного объема состоят из большого количества накопителей на жёстких магнитных дисках. Сохранность данных в распределенной системе в значительной степени зависит от надежности и рабочих характеристик используемых жестких дисков. Обеспечение надёжности хранилища в целом является важной прикладной задачей, сложность решения которой возрастает по мере увеличения количества задействованных накопителей.

Будучи сложным техническим устройством, жесткий диск подвержен сбоям, которые вызываются комплексом разнообразных случайных факторов и носят стохастический характер. С некоторой степенью приближения вероятность отказа жесткого диска можно описать статистическим законом, а затем обобщить процесс возникновения отказов на совокупность всех дисков хранилища.

Модель надёжности хранилища, описывающая серии отказов и замен жестких дисков до выполнения определенного условия, соответствует некоторой задаче Гюйгенса о разорении игрока. В классе задач Гюйгенса рассматривается игра с ограниченной начальной суммой при фиксированной ставке и известным математическим ожиданием выигрыша на каждом раунде, проводимая либо до приумножения начального капитала, либо до его полной потери. Широко используются два эквивалентных подхода к решению этой задачи, а именно: испытания Бернулли и случайное блуждание. Так, в методе случайного блуждания рассматривается стохастическое перемещение игрока по дискретным состояниям в зависимости от исхода очередного раунда.

В модели надёжности хранилища при допущении равнозначности и независимости всех дисков дискретные состояния соответствует количеству исправных дисков. Переход из состояния в состояние определяется явным сбоем, который обнаруживается сразу после возникновения. К таким сбоям относят, в частности, отказ программно-аппаратной части жесткого диска.

Поскольку в реальных приложениях применяется резервирование, сохранность данных обеспечивается при выходе из строя некоторой части накопителей. Максимальный порог отказавших дисков, выше которого происходит безвозвратная потеря данных, определяется схемой и параметрами избыточного хранения.

Еще одной технической особенностью жестких дисков является возникновение скрытых битовых ошибок, не выявляемых непосредственно на момент их возникновения. Битовые ошибки искажают хранимые данные, но не влияют на физическое функционирование устройства. Поскольку процедура восстановления запускается только после обнаружения ошибок, скрытые битовые ошибки отрицательно влияют на надёжность хранилища. Этот класс ошибок относится к невосстановимым ошибкам чтения (irrecoverable read errors) и должен приниматься во внимание при проектировании хранилища данных.

Основным способом борьбы с такими ошибками является вычисление контрольных сумм для фрагментов каждого блока и их дальнейшая проверка. При этом возможны несколько подходов к проверке данных.

В первом методе проверка контрольных сумм выполняется только при непосредственном запросе данных клиентом. По результатам проверки в случае несовпадения контрольных сумм инициируется процесс восстановления. В таком случае важным параметром, влияющим на надежность хранилища, будет средняя интенсивность доступа к данным. На практике интенсивность доступа к данным разного типа может сильно отличаться, поэтому риск потери данных для большого архивного файла, к которому почти нет обращений, может быть значительным. Исходя из этого, первый вариант не обеспечивает в достаточной мере требуемый уровень надежности хранения для всех файлов.

Второй подход к проверке контрольных сумм предполагает наличие в хранилище данных централизованного сервиса, который управляет непрерывным процессом проверки контрольных сумм фрагментов, этот подход также известен как скраббинг (scrubbing). При этом централизованному сервису необязательно самостоятельно заниматься пересчетом контрольных сумм, достаточно, чтобы он организовывал хранение информации, необходимой для проверки данных, а также посылал команды проверки данных сервисам хранения. Таким образом, нагрузка, связанная с проверкой данных, будет распределена по всему хранилищу. Интенсивность работы такого процесса принято оценивать через ожидаемое время проверки всего объема данных, которые содержатся в хранилище, и может составлять от нескольких суток до нескольких недель.

В настоящей работе предлагается новая математическая модель надежности хранения данных с оригинальным аналитическим описанием переходов между состояниями, учитывающая возникновение скрытых дисковых ошибок и непрерывный процесс скраббинга.

2. Обзор литературы

Классические модели надёжности хранения данных, построенные на основе Марковских цепей в непрерывном времени, дают ориентировочное представление о среднем времени до потери данных (MTTDL) хранилища, но не учитывают возникающие на дисках битовые ошибки. Эти модели рассмотрены в ряде работ (например, [1,2,3,4]) по исследованию надежности RAID-массивов в условиях явных дисковых сбоев. За последние годы всё больше научных работ посвящается построению расширенных Марковских моделей надёжности, включающих математические описания скрытых дисковых ошибок и процессов их обнаружения. Так, например, недостатки классических моделей с одномерными Марковскими цепями без учёта памяти подробно описаны в широко известной работе [5]. Контраргументы, в которых показана пригодность Марковских цепей для моделирования надёжности,

приводятся в работе [6], где эффект памяти воспроизводится в рамках более точного детального описания системы увеличенным количеством состояний и переходов между ними. Оригинальная модель для RAID-массива из SSD-накопителей с ростом интенсивности ошибок по мере износа предложена в работе [7]. Например, в работах [8,9,10,11] рассмотрено последовательное обобщение модели надёжности MTTDL для RAID-массивов на случаи возникновения скрытых невосстановимых секторных ошибок чтения, и применения процедуры проверки контрольных сумм. В опубликованных работах исследование свойств двумерной Марковской модели в силу её сложности осуществляется имитационными численными методами, включающими методы статистического моделирования. Несмотря на сложность описания двумерных цепей, Марковские модели остаются привлекательными для использования, поскольку позволяют оценивать надежность хранения данных для произвольной схемы избыточного кодирования, учитывать различные типы дисковых сбоев, различные состояния компонентов хранилища, а также различные политики замены сбойных компонентов и восстановления данных, такие как параллельное и последовательное восстановление, разные интенсивности восстановления для разного количества потерянных фрагментов. В отличие от имитационного моделирования, основанного на статистических испытаниях Монте-Карло, данный метод позволяет получить точное аналитическое выражение для надежности хранилища в рамках данной модели, а аналитическое выражение может быть использовано для выявления некоторых функциональных зависимостей между параметрами модели. Кроме того, Марковские модели сохраняют значительное преимущество в производительности и скорости расчётов (до 150 раз, см. [6]) по сравнению с полномасштабным имитационным моделированием.

Представленная модель развивает идеи описания традиционных аппаратных средств локального хранения данных в RAID-массивах из отдельных дисков и рассматривает перспективные распределенные системы хранения данных с избыточным кодированием [12], где отдельные фрагменты блоков данных не привязаны к дискам и перемещаются при репликации или балансировке нагрузки. В разработанной математической модели переопределяется семантика классических моделей для RAID-систем — состояния соответствуют фрагментам блока с избыточным кодированием, а не дискам. Аналогичным образом переходы из состояния в состояние описывают потерю или восстановление фрагмента данных, а не выход из строя индивидуального накопителя.

3. Базовая математическая модель

Пусть имеется блок данных, составленный из n закодированных фрагментов, в которых k фрагментов соответствуют исходным данным, а остальные $(n-k)$ фрагментов — контрольным суммам. При этом пусть схема

кодирования такова, что для восстановления данных достаточно любых k фрагментов. Блок данных характеризуется следующим набором состояний S_i: состояние S_0 – все n фрагментов доступны и ни одного фрагмента не повреждено, состояние S_1 – поврежден один фрагмент, состояние S_2 – повреждено 2 фрагмента, …, состояние S_{n-k} – повреждено $(n-k)$ фрагментов, состояние S_{n-k+1} – повреждено больше, чем $(n-k)$ фрагментов, что означает невозможность восстановить блок, т.е. потерю данных блока. Состояние $(n-k+1)$ в дальнейшем удобно обозначить названием DL (“data loss”) для выделения среди всех остальных состояний. Состояние DL является абсорбирующим, поскольку для него отсутствуют возвратные переходы в другие состояния системы. Согласно модели система за некоторое время вне зависимости от начального состояния с вероятностью 1 перейдет в абсорбирующее состояние. Для таких систем важной практической характеристикой является среднее время работы до перехода в абсорбирующее состояние.

Переход между состояниями системы описывает потерю фрагментов данных в блоке из-за сбоев дисковых накопителей. В качестве первого упрощающего допущения можно принять, что дисковые сбои подчиняются Пуассоновскому распределению, согласно которому вероятность выхода диска из строя в течение какого-либо периода времени не зависит от возраста диска. Выход из строя диска означает полную потерю хранящихся на нем данных, таким образом, потерянные фрагменты могут быть восстановлены только с использованием фрагментов, которые располагаются на работающих дисках.

Пусть интенсивность дисковых отказов равна λ, это означает, что ожидаемое время функционирования диска до отказа составляет $1/\lambda$. Для преобладающего большинства моделей дисковых накопителей интенсивность отказов составляет порядка 10^{-9} секунды. Вероятность сбоя диска в течение интервала времени $[0,t)$ определяется интенсивностью дисковых отказов и равна $\Pr(t_F < t) = 1 - e^{-\lambda t}$.

Вторым общепринятым допущением является предположение о независимости в совокупности сбоев индивидуальных дисков. В этом случае для n работающих дисков суммарная интенсивность отказов равна $n\lambda$, а ожидаемое время до сбоя (MTTF) хотя бы одного диска из n равно $1/(n\lambda)$.

Отказ диска определяет переход модели из состояния S_l с l неработающими дисками в состояние S_{l+1} с $(l+1)$ неработающими дисками. Поскольку в моделях с непрерывным временем совпадение моментов выхода из строя двух любых дисков считается случайным событием нулевой меры, можно считать, что непосредственные переходы между состояниями, различающимися более чем на один работающий диск, имеют нулевую вероятность, несмотря на априорную возможность самого такого события. Таким образом в модели определяющими событиями являются единичные дисковые сбои.

Следующим важным фактором модели является среднее время до восстановления фрагмента данных $1/\mu$. В отличие от интенсивности

дисковых отказов этот параметр определяется не только физическими свойствами используемых носителей данных, но и архитектурой конкретного распределенного хранилища, и может зависеть от многих факторов. Среднее время восстановления фрагмента данных складывается из времени, прошедшего от дискового сбоя до его обнаружения, и времени, прошедшего от обнаружения сбоя до момента, когда восстановленный фрагмент записан на один из дисков системы. Как правило, архитектура распределенного хранилища данных предполагает наличие сервиса мониторинга, который следит за состоянием всех блоков в хранилище данных и запускает процедуру восстановления данных при обнаружении сбоев, поэтому время до обнаружения сбоя даже при большом объеме данных, которые хранятся в системе, можно считать достаточно малым – порядка нескольких минут. Далее рассматривается обобщение классической модели на сценарии возникающих скрытых дисковых сбоев и процедур их обнаружения.

4. Расширенная модель

В этом разделе раскрыты характеристики стилей, используемых в данном документе.

4.1 Частота невосстановимых ошибок чтения

Как правило, частота невосстановимых ошибок чтения указывается производителем в спецификации диска и составляет порядка одной ошибки на 10^{14} битов или порядка одной ошибки на 11 Тб прочитанных с диска данных. С целью учета этого параметра в предложенной модели необходимо оценить, насколько часто такие ошибки будут происходить в течение жизни хранилища, а для этого требуется оценка среднего количества данных, которое в среднем считывается с одного диска в датацентре на протяжении некоторого характерного интервала времени, например, в течение года.

Среднюю интенсивность доступа к диску можно оценить, исходя из средней загруженности диска, указанную в спецификации. Пусть ожидаемая загруженность диска составляет порядка 20%, т.е. в течение 20% времени происходит доступ к диску, а 80% времени диск простаивает. Тогда установившаяся скорость последовательного чтения с диска равна 140 Мб/с, тогда в течение года с диска будет прочитано $(140 \cdot 0.2 \cdot 3600 \cdot 24 \cdot 365)/1024 = 862312$Гб или около 842 Тб. Исходя из этого, частоту невосстановимых битовых ошибок для одного диска по порядку величины можно оценить как 77 ошибок в год. Пусть размер диска составляет 2 Тб и он заполнен наполовину, а характерный размер одного фрагмента данных в хранилище равен 50 Мб, тогда средняя частота ошибок при чтении выбранного фрагмента блока данных будет равна приблизительно 0.00367 ошибок в год или около 1 ошибки в течение 272 лет.

В разработанную модель добавлен параметр c — частота невосстановимых ошибок чтения для одного фрагмента блока данных, в частном случае равная

$272лет^{-1}$. Этот параметр сравним по величине с интенсивностью дисковых сбоев, которую для дисков с MTTF=200000 часов можно оценить как $23лет^{-1}$, и должен быть учтен при оценке MTTDL хранилища данных.

Необходимо отметить, что этот параметр был рассчитан для сценария сравнительно высокой степени использования дисков. В общем случае частота невосстановимых ошибок чтения существенно зависит от интенсивности использования диска. Средняя ожидаемая загруженность диска, указанная в спецификации, является ориентировочным значением для оценок, но для конкретного хранилища данных лучше брать средние оценки загруженности дисков для данного хранилища. Например, архивное хранилище данных может иметь более низкую загруженность дисков. и в общем случае будет зависеть от интенсивности и паттернов доступа к данным для реального датацентра. Тем не менее, рассчитанное значение параметра позволит оценить снизу MTTDL хранилища с учетом наличия битовых ошибок, а также характер их влияния на надежность хранения данных.

4.2 Ожидаемое время полной проверки данных в датацентре

Среднее время выполнения полной проверки данных в датацентре, определяется средней интенсивностью проверки данных, т.е. средним объемом данных, проверяемым за единицу времени. Физически реалистичные значения для средней интенсивности проверки данных будут определяться как степенью загруженности хранилища и гарантиями производительности, которые оно предоставляет клиентам, так и экономическими затратами на выполнение проверки контрольных сумм.

В зависимости от загруженности хранилища пользовательские задачи доступа к данным и сервисные процессы работы с данными могут претендовать на одни и те же ресурсы и провоцировать конфликты доступа. В зависимости от конкретной архитектуры хранилища и ее задач такими ресурсами могут быть процессорное время, дисковые операции ввода-вывода, а также пропускная способность сети. В таком случае, как правило, ресурсы, используемые сервисными процессами, ограничиваются так, чтобы обеспечить клиентам гарантированный уровень производительности (гарантированное время отклика, гарантированное полное количество операций ввода-вывода).

В случае, когда реальная загруженность хранилища далека от предельных значений, а пользовательские задачи и системные процессы не требуют эксклюзивного доступа к одним ресурсам, интенсивность процесса проверки данных полностью определяется балансом между требованиями обеспечения надежности хранения данных и предельными затратами в процессе проверки контрольных сумм. Эти затраты определяются, во-первых, возможной необходимостью вывода диска, который содержит проверяемый фрагмент, из режима ожидания, в котором снижено его энергопотребление. Во-вторых, затраты связаны с необходимостью расходования процессорного времени на вычислительно требовательный алгоритм подсчета контрольных сумм и также

возможным выводом одного из ядер процессора из режима пониженного энергопотребления.

В качестве ориентировочного значения времени полной проверки данных можно выбрать характерный для в датацентра период времени от одной недели до одного месяца. Среднее время полной проверки также должно быть меньше, чем среднее время между запросами одних и тех же данных пользователя хранилища. значение интенсивности проверки данных как α.

В классической модели состояние распределенной системы хранения полностью определяется количеством вышедших из строя дисковых накопителей. В новой модели состояние системы зависит не только от явных поломок, но и от скрытых поломок, определяющих количество поврежденных фрагментов. Таким образом, расширенная модель, построена на двумерном, а не одномерном пространстве состояний Марковской цепи, в отличие от классической модели.

Пусть состояние системы (l, m) характеризуется явными сбоями l дисков, а также m скрытыми повреждениями фрагментов. При этом скрытые повреждения фрагментов учитываются только для работоспособных дисков, т.к. данные на вышедших из строя дисках считаются недоступными для чтения.

Можно отметить, что значение суммы $(l + m)$ для состояния, в котором данные блока все еще можно восстановить, ограничено сверху количеством фрагментов, которые можно утратить в данной схеме помехоустойчивого кодирования без потери данных, т.е. для такого состояния выполняется неравенство $0 \le l + m \le n - k$. Для состояний с условием $l + m > n - k$, данные блока восстановлению не подлежат, поэтому вне зависимости от соотношения скрытых и явных повреждений все эти состояния можно объединить в одно состояние DL («data loss»).

Таким образом общее количество состояний в системе, не считая состояния DL, будет равно

$$\sum_{i=1}^{n-k+1} i = (n - k + 1)\frac{n - k + 2}{2}.$$

Первый тип переходов между состояниями описывает дисковые сбои. Для состояния (l, m) возможны два варианта переходов. Во-первых, может выйти из строя диск, соответствующий одному из поврежденных фрагментов и система с интенсивностью $m\lambda$ перейдет в состояние $(l + 1, m - 1)$. Во-вторых, может отказать диск, соответствующий одному из неповрежденных фрагментов, и система с интенсивностью $(n - l)\lambda$ перейдет в состояние $(l + 1, m)$.

Второй тип переходов — это переходы из-за битовых ошибок в состояния со скрытыми повреждениями. Из-за битовых ошибок возможен переход из состояния (l, m) в состояние $(l, m + 1)$ с интенсивностью $(n - l - m)c$. При таком переходе не учитывается возникновение скрытых битовых ошибок в

уже вышедших из строя дисках, а также не рассматривается возможность появления новых битовых ошибок в уже поврежденных фрагментах, т.к. эти события не меняют состояния системы.

Переходы между событиями, связанные с восстановлением данных, происходят как из-за процесса восстановления данных при обнаружении явных сбоев, так и из-за процесса детектирования скрытых ошибок в данных. Пусть при восстановлении после явных сбоев происходит проверка контрольных сумм для фрагментов на работоспособных дисках, и наоборот при обнаружении скрытых ошибок происходит не только перезапись поврежденных фрагментов, но и восстановление фрагментов, потерянных в результате явных дисковых сбоев. Пусть также процесс восстановления данных всегда переводит систему в состояние (0, 0), в котором отсутствуют явные и скрытые повреждения. Переход в состояние (0, 0) обоснован тем, что значения MTTDL для процессов последовательного и параллельного восстановления данных не отличаются в случае, когда интенсивности процессов восстановления намного выше интенсивности дисковых сбоев. В зависимости от состояния системы интенсивность перехода в состояние $(0,0)$ меняется следующим образом – для состояний вида $(l,0), l > 0$ интенсивность равна μ, для состояний вида $(0,m), m > 0$ интенсивность равна α, а для состояний вида $(l,m), l > 0, m > 0$ интенсивность равна $(\alpha + \mu)$.

Переходы в состояние DL в новой модели отличаются от переходов в другие состояния, т.к. состояние DL объединяет в себе множество состояний. В состояние DL система может перейти из состояний вида $(i, n-k-i)$, где $0 \le i \le n-k$. Интенсивность перехода в состояние DL для любого из этих состояний равна $k(\lambda + c)$.

4.3 Расчет MTTDL с учетом невосстановимых битовых ошибок и процесса проверки контрольных сумм

До перехода к описанию модели с произвольными параметрами n и k для наглядност целесообразно рассмотреть практически важные частные случаи расчета MTTDL со значениями $n-k=1$ и $n-k=2$. Прикладная эффективность этих наборов параметров для использования в локально восстановимых кодах (Locally Repairable Codes, LRC) показана в работе [12]. Схемы состояний системы и переходов между ними для этих случаев приведены на рис. 1 и рис. 2. Алгоритм расчета в целом аналогичен расчетам для классической модели.

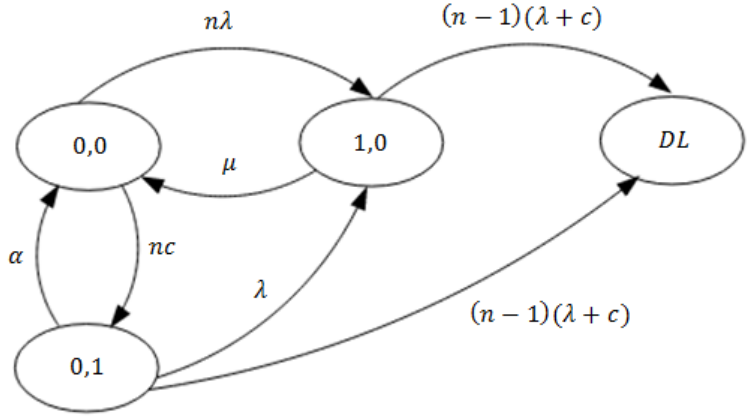

Рис. 1. Интенсивности переходов в Марковской цепи для случая n - k = 1

Пусть $Q_{l,m}$ — вероятность того, что система перейдет из начального состояния (l,m) в состояние DL до того, как окажется в состоянии $(0,0)$, где отсутствуют скрытые или явные сбои. Из определения $Q_{l,m}$ следует, что $Q_{0,0}=0$, а $Q_{DL}=1$.

Значение $Q_{1,0}$ выражается через вероятности $Q_{0,0}$ и Q_{DL} для состояний, в которые можно перейти из состояния $(1,0)$:

$$Q_{1,0}=\frac{1}{\mu+(n-1)(\lambda+c)}\left(\mu Q_{0,0}+(n-1)(\lambda+c)Q_{DL}\right).$$

В этом пункте и далее при проведении расчетов предполагается, что значения параметров λ и c пренебрежимо малы по сравнению со значениями параметров μ и α. Учитывая, что $Q_{0,0}=0$, $Q_{DL}=1$ и отбрасывая пренебрежимо малые члены, можно получить:

$$Q_{1,0}\approx\frac{(n-1)(\lambda+c)}{\mu}.$$

Значение $Q_{0,1}$ выражается следующим образом:

$$Q_{0,1}=\frac{1}{\alpha+\lambda+(n-1)(\lambda+c)}\left(\alpha Q_{0,0}+\lambda Q_{1,0}+(n-1)(\lambda+c)Q_{DL}\right).$$

Можно заметить, что член $\lambda Q_{1,0}\sim\lambda(\lambda+c)$, а член $(n-1)(\lambda+c)Q_{DL}\sim(\lambda+c)$, следовательно, $\lambda Q_{1,0}$ является пренебрежимо малым по отношению $(n-1)(\lambda+c)Q_{DL}\sim(\lambda+c)$:

$$Q_{0,1}\approx\frac{(n-1)(\lambda+c)}{\alpha}.$$

В данной модели возможны две варианта начала цикла – цикл может начаться либо с дискового сбоя, т.е. из состояния $(1,0)$, или с невосстановимой ошибки чтения, т.е. из состояния $(0,1)$.

Ожидаемое количество циклов до потери данных в случае начала из состояния $(1,0)$ будет равно

$$n_{e,(1,0)} = \frac{1}{Q_{1,0}} \approx \frac{\mu}{(n-1)(\lambda + c)}.$$

Ожидаемое количество циклов до потери данных в случае начала из состояния $(0,1)$ будет равно

$$n_{e,(0,1)} = \frac{1}{Q_{0,1}} \approx \frac{\alpha}{(n-1)(\lambda + c)}.$$

Оценка среднего времени цикла выполняется следующим образом. Пусть $T_{l,m}$ ожидаемое время, которое пройдет прежде, чем система перейдет из состояния (l, m) в состояние $(0,0)$ или в состояние DL. По определению $T_{0,0} = 0, T_{DL} = 0$.

Значения $T_{l,m}$ для состояний системы вычисляются как

$$T_{1,0} = \frac{1}{\mu + (n-1)(\lambda + c)}\left(\mu T_{0,0} + (n-1)(\lambda + c)T_{DL}\right) + \frac{1}{\mu + (n-1)(\lambda + c)};$$

$$T_{1,0} = \frac{1}{\mu + (n-1)(\lambda + c)};$$

$$T_{1,0} \approx \frac{1}{\mu};$$

$$T_{0,1} = \frac{1}{\alpha + \lambda + (n-1)(\lambda + c)}\left(\alpha T_{0,0} + \lambda T_{1,0} + (n-1)(\lambda + c)T_{DL}\right) + \frac{1}{\alpha + \lambda + (n-1)(\lambda + c)};$$

$$T_{0,1} = \frac{1}{\alpha + \lambda + (n-1)(\lambda + c)}\frac{\lambda}{\mu} + \frac{1}{\alpha + \lambda + (n-1)(\lambda + c)};$$

$$T_{0,1} \approx \frac{1}{\alpha + \lambda + (n-1)(\lambda + c)} \approx \frac{1}{\alpha}.$$

Пусть $t_{0,0\to1,0}$ и $t_{0,0\to0,1}$ — средние времена перехода из состояния $(0,0)$ в состояния $(1,0)$ и $(0,1)$ соответственно.

Для средних времен цикла $t_{e,(1,0)}$ и $t_{e,(0,1)}$ при начале из состояний $(1,0)$ и $(0,1)$ выполняются соотношения:

$$t_{e,(1,0)} = t_{0,0\to1,0} + T_{1,0};$$

$$t_{e,(0,1)} = t_{0,0\to0,1} + T_{0,1};$$

$$t_{e,(1,0)} \approx \frac{1}{n\lambda} + \frac{1}{\mu};$$

$$t_{e,(0,1)} \approx \frac{1}{n\mathrm{c}} + \frac{1}{\alpha}.$$

Учитывая, что значения $\frac{1}{\mu}$ и $\frac{1}{\alpha}$ пренебрежимо малы по сравнению с $\frac{1}{(\lambda+c)}$, можно получить

$$t_{e,(1,0)} \approx \frac{1}{n\lambda}$$
$$t_{e,(0,1)} \approx \frac{1}{nc}.$$

Пусть $p_{e,(1,0)} = \frac{\lambda}{\lambda+c}$ и $p_{e,(0,1)} = \frac{c}{\lambda+c}$ вероятности начать цикл из состояний (1,0) и (0,1) соответственно.

Из вышеприведенных соотношений видно, что оценка MTTDL хранилища для новой модели равна наименьшему значению из ожидаемых времен до потери данных в случае начала из одного из состояний $(1,0)$ или $(0,1)$:

$$MTTDL_{n-k=1} = p_{e,(1,0)} n_{e,(1,0)} t_{e,(1,0)} + p_{e,(0,1)} n_{e,(0,1)} t_{e,(0,1)} \approx \frac{\mu}{n(n-1)(\lambda+c)^2}, \frac{\alpha}{n(n-1)(\lambda+c)^2}.$$

Далее рассматривается случай $n-k=2$.

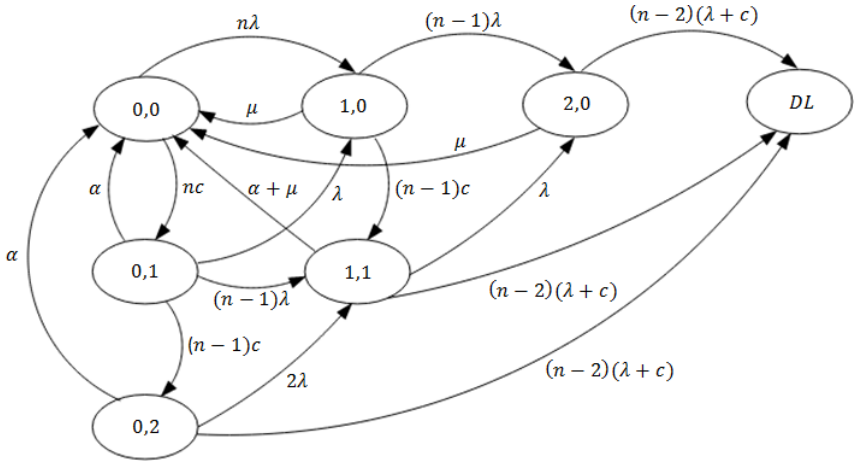

Рис.2. Интенсивности переходов в Марковской цепи для случая $n-k=2$

Значения $Q_{l,m}$ вычисляются, начиная с состояний на диагонали $l+m=n-k$.

$$Q_{2,0} = \frac{1}{\mu + (n-2)(\lambda+c)}\left(\mu Q_{0,0} + (n-2)(\lambda+c)Q_{DL}\right);$$

$$Q_{2,0} \approx \frac{(n-2)(\lambda+c)}{\mu};$$

$$Q_{1,1} = \frac{1}{\alpha+\mu+\lambda+(n-2)(\lambda+c)}\left((\alpha+\mu)Q_{0,0} + \lambda Q_{2,0} + (n-2)(\lambda+c)Q_{DL}\right);$$

$$Q_{1,1} \approx \frac{(n-2)(\lambda+c)}{\alpha+\mu};$$

$$Q_{0,2} = \frac{1}{\alpha+2\lambda+(n-2)(\lambda+c)}\left(\alpha Q_{0,0} + 2\lambda Q_{1,1} + (n-2)(\lambda+c)Q_{DL}\right);$$

$$Q_{0,2} \approx \frac{(n-2)(\lambda+c)}{\alpha}.$$

Далее рассматриваются состояния на диагонали $l+m=n-k-1$:

$$Q_{1,0} = \frac{1}{\mu+(n-1)\lambda+(n-1)c}\left((n-1)\lambda Q_{2,0} + (n-1)cQ_{1,1} + \mu Q_{0,0}\right);$$

$$Q_{1,0} \approx \frac{n-1}{\mu}\left(\lambda\frac{(n-2)(\lambda+c)}{\mu} + c\frac{(n-2)(\lambda+c)}{\alpha+\mu}\right);$$

$$Q_{1,0} \approx \frac{(n-1)(n-2)(\lambda+c)}{\mu}\left(\frac{\lambda}{\mu} + \frac{c}{\alpha+\mu}\right);$$

$$Q_{0,1} = \frac{1}{\alpha+n\lambda+(n-1)c}\left(\lambda Q_{1,0} + (n-1)\lambda Q_{1,1} + (n-1)cQ_{0,2} + \alpha Q_{0,0}\right).$$

Учитывается пренебрежимая малость $\lambda Q_{1,0}$ по сравнению с $\lambda Q_{1,1}$ и $cQ_{0,2}$:

$$Q_{0,1} \approx \frac{1}{\alpha}\left((n-1)\lambda Q_{1,1} + (n-1)cQ_{0,2}\right);$$

$$Q_{0,1} \approx \frac{1}{\alpha}\left((n-1)\lambda\frac{(n-2)(\lambda+c)}{\alpha+\mu} + (n-1)c\frac{(n-2)(\lambda+c)}{\alpha}\right);$$

$$Q_{0,1} \approx \frac{(n-1)(n-2)(\lambda+c)}{\alpha}\left(\frac{\lambda}{\alpha+\mu} + \frac{c}{\alpha}\right).$$

Вычисляется ожидаемое количество циклов до потери данных в случае начальных состояний $(1,0)$ и $(0,1)$:

$$n_{e,(1,0)} = \frac{1}{Q_{1,0}} \approx \frac{\mu}{(n-1)(n-2)(\lambda+c)}\left(\frac{\lambda}{\mu} + \frac{c}{\alpha+\mu}\right)^{-1};$$

$$n_{e,(0,1)} = \frac{1}{Q_{0,1}} \approx \frac{\alpha}{(n-1)(n-2)(\lambda+c)}\left(\frac{\lambda}{\alpha+\mu} + \frac{c}{\alpha}\right)^{-1}.$$

Оцениваются средние времена цикла для начальных состояний $(1,0)$ и $(0,1)$. Значения $T_{l,m}$ вычисляются в той же последовательности, что и соответствующие значения $Q_{l,m}$:

$$T_{2,0} = \frac{1}{\mu+(n-2)(\lambda+c)};$$

$$T_{2,0} \approx \frac{1}{\mu};$$

$$T_{1,1} = \frac{1}{\alpha+\mu+\lambda+(n-2)(\lambda+c)}\lambda T_{2,0} + \frac{1}{\alpha+\mu+\lambda+(n-2)(\lambda+c)};$$
$$T_{1,1} \approx \frac{1}{\alpha+\mu};$$
$$T_{0,2} = \frac{1}{\alpha+2\lambda+(n-2)(\lambda+c)} \cdot 2\lambda T_{1,1} + \frac{1}{\alpha+2\lambda+(n-2)(\lambda+c)};$$
$$T_{0,2} \approx \frac{1}{\alpha};$$
$$T_{1,0} = \frac{1}{\mu+(n-1)\lambda+(n-1)c}\left((n-1)\lambda T_{2,0} + (n-1)cT_{1,1}\right) + \frac{1}{\mu+(n-1)\lambda+(n-1)c};$$
$$T_{1,0} \approx \frac{1}{\mu};$$
$$T_{0,1} = \frac{1}{\alpha+n\lambda+(n-1)c}\left(\lambda T_{1,0} + (n-1)\lambda T_{1,1} + (n-1)cT_{0,2}\right) + \frac{1}{\alpha+n\lambda+(n-1)c};$$
$$T_{0,1} \approx \frac{1}{\alpha}.$$

Для средних времен цикла $t_{e,(1,0)}$ и $t_{e,(0,1)}$ для начальных состояний $(1,0)$ и $(0,1)$ соотношения принимают вид:

$$t_{e,(1,0)} = t_{0,0\to1,0} + T_{1,0};$$
$$t_{e,(0,1)} = t_{0,0\to0,1} + T_{0,1};$$
$$t_{e,(1,0)} \approx \frac{1}{n\lambda};$$
$$t_{e,(0,1)} \approx \frac{1}{nc}.$$

MTTDL хранилища для данной модели равен

$$MTTDL_{n-k=2} = p_{e,(1,0)}n_{e,(1,0)}t_{e,(1,0)} + p_{e,(0,1)}n_{e,(0,1)}t_{e,(0,1)} \approx \frac{\mu\left(\frac{\lambda}{\mu}+\frac{c}{\alpha+\mu}\right)^{-1}}{n(n-1)(n-2)(\lambda+c)^2} + \frac{\alpha\left(\frac{\lambda}{\alpha+\mu}+\frac{c}{\alpha}\right)^{-1}}{n(n-1)(n-2)(\lambda+c)^2}.$$

4.4 Обобщение расширенной модели на произвольные параметры $(\mathbf{n}, \mathbf{k})$

Соотношения, полученные для частных случаев $n-k=1$ и $n-k=2$ можно обобщить для произвольных параметров (n,k).

Схема вычисления значений $Q_{l,m}$ и $T_{l,m}$, изображенная на рис. 3, предполагает последовательное вычисление значений вдоль диагоналей $l+m=const$,

начиная с диагонали, соответствующей состояниям (l, m), для которых $l + m = n - k$, при движении сверху вниз вдоль диагоналей.

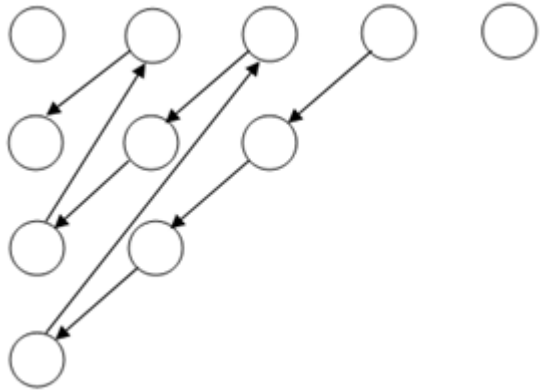

Рис. 3. Схема вычисления значений $\boldsymbol{Q_{l,m}}$ *и* $\boldsymbol{T_{l,m}}$

Средние времена цикла $t_{e,(1,0)}$ и $t_{e,(0,1)}$ для модели с произвольными параметрами (n, k) вычисляются следующим образом.

$$T_{l,m} = \begin{cases} \dfrac{(n-l-m)\lambda T_{l+1,m} + (n-l-m)cT_{l,m+1} + 1}{\mu + m\lambda + (n-l-m)(\lambda + c)}, l \neq 0, m = 0 \\ \dfrac{m\lambda T_{l+1,m-1} + (n-l-m)\lambda T_{l+1,m} + (n-l-m)cT_{l,m+1} + 1}{\alpha + \mu + m\lambda + (n-l-m)(\lambda + c)}, l \neq 0, m \neq 0 \\ \dfrac{m\lambda T_{l+1,m-1} + (n-l-m)\lambda T_{l+1,m} + 1}{\alpha + m\lambda + (n-l-m)(\lambda + c)}, l = 0, m \neq 0 \end{cases}$$

Отбрасывая пренебрежимо малые члены, можно получить:

$$T_{l,m} \approx \begin{cases} \dfrac{1}{\mu}, l \neq 0, m = 0 \\ \dfrac{1}{\alpha + \mu}, l \neq 0, m \neq 0 \\ \dfrac{1}{\alpha}, l = 0, m \neq 0 \end{cases}$$

Из выведенного соотношения видно, что значение $T_{l,m}$ зависит только от положения состояния (l, m) внутри диагонали $l + m = const$, но не зависит от диагонали.

Таким образом, средние времена цикла равны:

$$t_{e,(1,0)} \approx \frac{1}{n\lambda}, \qquad t_{e,(0,1)} \approx \frac{1}{nc}.$$

Вычисление $Q_{l,m}$ в произвольном случае производится несколько сложнее.

$$Q_{l,m} = \begin{cases} \dfrac{(n-l-m)\lambda Q_{l+1,m} + (n-l-m)cQ_{l,m+1}}{\mu + m\lambda + (n-l-m)(\lambda + c)}, l \neq 0, m = 0, \\ \dfrac{m\lambda Q_{l+1,m-1} + (n-l-m)\lambda Q_{l+1,m} + (n-l-m)cQ_{l,m+1}}{\alpha + \mu + m\lambda + (n-l-m)(\lambda + c)}, l \neq 0, m \neq 0, \\ \dfrac{m\lambda Q_{l+1,m-1} + (n-l-m)\lambda Q_{l+1,m} + (n-l-m)cQ_{l,m+1}}{\alpha + m\lambda + (n-l-m)(\lambda + c)}, l = 0, m \neq 0. \end{cases}$$

Если учесть, что $\lambda Q_{l+1,m-1}$ пренебрежимо мало по сравнению с $\lambda Q_{l+1,m}$, а $cQ_{l,m+1}$ — значения $Q_{l,m}$ в рамках одной диагонали $l + m = const$ сравнимы между собой по величине, в то время как значения $Q_{l,m}$ на внутренних диагоналях пренебрежимо малы по сравнению со значениями $Q_{l,m}$ на диагоналях, внешних по отношению к этим диагоналям, то выполняются условия:

$$Q_{l_1,m_1} \ll Q_{l_2,m_2}, \text{где } l_1 + m_1 < l_2 + m_2;$$

$$Q_{l,m} \approx \begin{cases} \dfrac{(n-l-m)\lambda Q_{l+1,m} + (n-l-m)cQ_{l,m+1}}{\mu}, l \neq 0, m = 0, \\ \dfrac{(n-l-m)\lambda Q_{l+1,m} + (n-l-m)cQ_{l,m+1}}{\alpha + \mu}, l \neq 0, m \neq 0, \\ \dfrac{(n-l-m)\lambda Q_{l+1,m} + (n-l-m)cQ_{l,m+1}}{\alpha}, l = 0, m \neq 0. \end{cases}$$

Пусть «простым» путем из состояния Q_{l_1,m_1} в состояние Q_{l_2,m_2} называется такая последовательность переходов между состояниями системы, в которой нет событий восстановления данных, а также нет переходов между состояниями, располагающимися на одной диагонали. Пути, в которых есть переходы между состояниями на одной диагонали, можно не учитывать в силу их маловероятности по отношению к «простым» путям.

Можно заметить, что значение $Q_{l,m}$ в состоянии (l, m) определяется как сумма слагаемых S_i, вычисляемых вдоль всех «простых» путей из состояния $Q_{l,m}$ в состояние Q_{DL}. Число таких «простых» путей для состояния $Q_{l,m}$ равно $2^{n-k-(l+m)}$, т.к. длина всех «простых» путей из состояния $Q_{l,m}$ в состояние Q_{DL} фиксирована и равна $n - k - (l + m)$, и в каждом промежуточном состоянии есть два варианта выбора дальнейшего направления пути – вверх или вниз.

Из рекуррентных соотношений для $Q_{l,m}$ следует, что значение S_i, соответствующее некоторому фиксированному «простому» пути, определяется как произведение интенсивностей переходов между состояниями вдоль данного пути, поделенное на произведение интенсивностей восстановления данных в каждом из состояний пути, кроме состояния DL.

Определим вспомогательные функции $I(l, m)$ и $P(l, m)$:

$$I(l,m) = \begin{cases} \mu, l \neq 0, m = 0 \\ \alpha + \mu, l \neq 0, m \neq 0 \\ \alpha, l = 0, m \neq 0 \end{cases}$$

$$P(l,m) = \prod_{i=0}^{(n-k)-(l+m)} (n - (l+m) - i)$$

Таким образом, значение $Q_{l,m}$ для произвольного состояния (l,m) определяется следующим выражением:

$$Q_{l,m} \approx P(l,m)\left(\sum_{\beta_1,\dots,\beta_{(n-k)-(l+m)}=0}^{1} \frac{(\lambda + \mathrm{c})\prod_{i=1}^{(n-k)-(l+m)} \lambda^{\beta_i} c^{1-\beta_i}}{I(l,m)\prod_{i=1}^{(n-k)-(l+m)} I(l+\sum_{j=1}^{i}\beta_i, m+\sum_{j=1}^{i}(1-\beta_i))}\right)$$

$$Q_{l,m} \approx \frac{P(l,m)}{I(l,m)}\left(\sum_{\beta_1,\dots,\beta_{(n-k)-(l+m)}=0}^{1} \prod_{i=1}^{(n-k)-(l+m)} \frac{(\lambda + \mathrm{c})\lambda^{\beta_i} c^{1-\beta_i}}{I(l+\sum_{j=1}^{i}\beta_i, m+\sum_{j=1}^{i}(1-\beta_i))}\right)$$

Ожидаемое количество циклов $n_{e,(1,0)}$ и $n_{e,(1,0)}$ определяется следующими выражениями

$$Q_{1,0} \approx \frac{P(1,0)}{I(1,0)}\left(\sum_{\beta_1,\dots,\beta_{(n-k)-1}=0}^{1} \prod_{i=1}^{(n-k)-1} \frac{(\lambda + \mathrm{c})\lambda^{\beta_i} c^{1-\beta_i}}{I(1+\sum_{j=1}^{i}\beta_i, \sum_{j=1}^{i}(1-\beta_i))}\right)$$

$$Q_{0,1} \approx \frac{P(0,1)}{I(0,1)}\left(\sum_{\beta_1,\dots,\beta_{(n-k)-1}=0}^{1} \prod_{i=1}^{(n-k)-1} \frac{(\lambda + \mathrm{c})\lambda^{\beta_i} c^{1-\beta_i}}{I(\sum_{j=1}^{i}\beta_i, 1+\sum_{j=1}^{i}(1-\beta_i))}\right)$$

$$n_{e,(1,0)} = \frac{1}{Q_{1,0}} \approx \left(\sum_{\beta_1,\dots,\beta_{(n-k)-1}=0}^{1} \frac{P(1,0)}{I(1,0)} \prod_{i=1}^{(n-k)-1} \frac{(\lambda + \mathrm{c})\lambda^{\beta_i} c^{1-\beta_i}}{I(1+\sum_{j=1}^{i}\beta_i, \sum_{j=1}^{i}(1-\beta_i))}\right)^{-1}$$

$$n_{e,(0,1)} = \frac{1}{Q_{0,1}} \approx \left(\sum_{\beta_1,\dots,\beta_{(n-k)-1}=0}^{1} \frac{P(0,1)}{I(0,1)} \prod_{i=1}^{(n-k)-1} \frac{(\lambda + \mathrm{c})\lambda^{\beta_i} c^{1-\beta_i}}{I(\sum_{j=1}^{i}\beta_i, 1+\sum_{j=1}^{i}(1-\beta_i))}\right)^{-1}$$

Таким образом, аналитическое выражение для MTTDL системы для произвольных параметров (n,k):

$$MTTDL_{n,k} \approx p_{e,(1,0)} n_{e,(1,0)} t_{e,(1,0)} + p_{e,(0,1)} n_{e,(0,1)} t_{e,(0,1)}$$

$$= \frac{1}{n(\lambda + c)^2\left(\sum_{\beta_1,\dots,\beta_{(n-k)-1}=0}^{1} \frac{P(1,0)}{I(1,0)} \prod_{i=1}^{(n-k)-1} \frac{(\lambda + \mathrm{c})\lambda^{\beta_i} c^{1-\beta_i}}{I(1+\sum_{j=1}^{i}\beta_i, \sum_{j=1}^{i}(1-\beta_i))}\right)}$$

$$+ \frac{1}{n(\lambda + c)^2\left(\sum_{\beta_1,\dots,\beta_{(n-k)-1}=0}^{1} \frac{P(0,1)}{I(0,1)} \prod_{i=1}^{(n-k)-1} \frac{(\lambda + \mathrm{c})\lambda^{\beta_i} c^{1-\beta_i}}{I(\sum_{j=1}^{i}\beta_i, 1+\sum_{j=1}^{i}(1-\beta_i))}\right)}$$

5. Результаты расчётов по разработанной модели

В этом разделе раскрыты характеристики стилей, используемых в данном документе.

5.1 Сравнительный анализ классической и расширенной моделей

Сравнение результатов расчётов по новой разработанной модели с итогами вычислений по классической модели приведено на рис. 4. В новой модели интенсивность битовых ошибок определялась средней интенсивностью доступа к дискам в хранилище. По графику видно, что результаты, полученные в классической модели, учитывающей только явные дисковые сбои, отличаются от уточненных результатов, полученных в рамках предложенной модели. Наблюдаемая разница является значительной и подтверждает практическую значимость предложенной модели. Полученные результаты соответствуют приведенным в работе [11] выводам о необходимости учёта скрытых ошибок и проведении проверок контрольных сумм.

Важно отметить, что MTTDL хранилища в значительной степени зависит от средней интенсивности использования дисков, увеличиваясь с уменьшением интенсивности использования и приближаясь к более простой модели, не учитывающей битовые ошибки. Этот результат подтверждает существенную зависимость частоты невосстановимых ошибок чтения от интенсивности использования диска. Таким образом интенсивность использования диска может служить эмпирическим критерием оценки частоты возникновения скрытых дисковых сбоев.

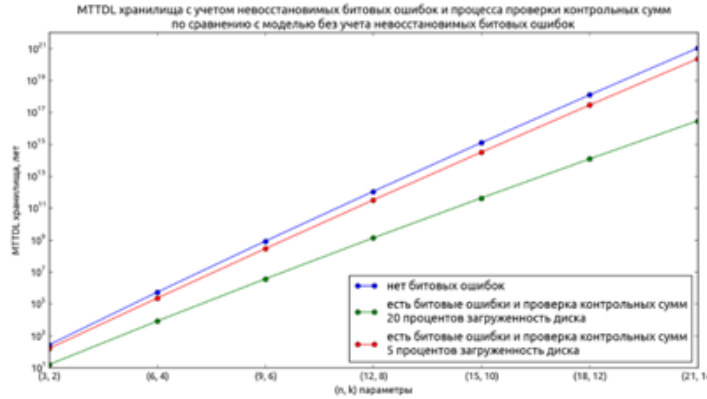

Рис. 4. Сравнение результатов, полученных в модели, учитывающей невосстановимые битовые ошибки и процесс проверки контрольных сумм, и в более простой модели

Из графика зависимости MTTDL хранилища данных от среднего времени полной проверки данных в хранилище видно, что MTTDL сначала резко убывает на начальном участке, а потом становится практически постоянной и не зависит от интенсивности проверки данных, приближаясь к значению MTTDL хранилища, в котором присутствуют скрытые битовые ошибки, но отсутствуют процессы проверки контрольных сумм.

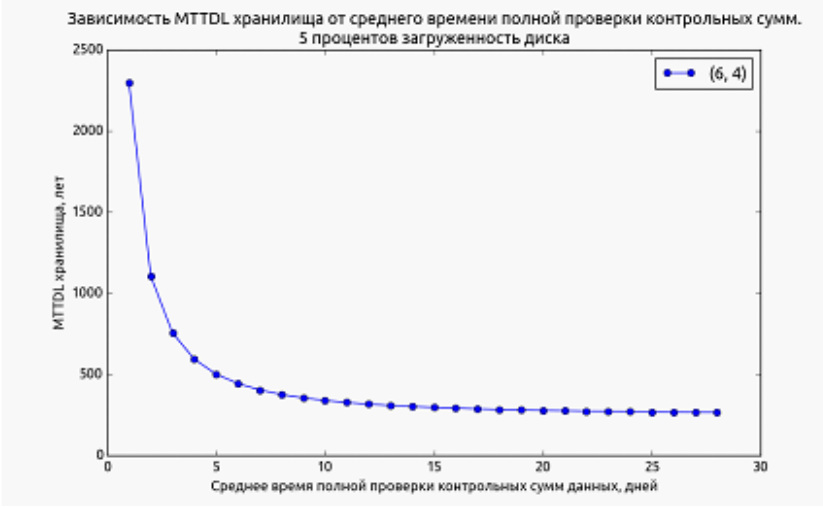

Рис. 5. Зависимость MTTDL хранилища данных от среднего периода полной проверки данных в хранилище

Зависимость между MTTDL хранилища данных и интенсивностью проверки контрольных сумм позволяет находить баланс между надежностью хранения данных и экономическими затратами на непрерывный процесс проверки данных.

5. Выводы

В настоящей работе была представлена новая математическая модель надёжности распределенных хранилищ с новым аналитическим описанием с основе двумерных Марковских цепей. Предложен новый метод расчёта MTTDL хранилища с помощью приближенно аналитических выражений для произвольных параметров (n, k). Наглядно показано, что невозможность учёта битовых ошибок чтения в классических моделях приводит к завышенным оценкам MTTDL. Из этого следует, что скрытые сбои существенно влияют на надежность хранения данных, и этим фактором не следует пренебрегать.

6. Примечания

Статья опубликована в рамках выполнения прикладных научных исследований при финансовой поддержке Министерства образования и науки Российской Федерации. Соглашения о предоставлении субсидий № 14.579.21.0010. Уникальный идентификатор Соглашения RFMEFI57914X0010.

Литература

[1]. Patterson D. A., Gibson G., and Katz R. H. A Case for Redundant Arrays of Inexpensive Disks (RAID), Proc. of ACM SIGMOD, 1988.

[2]. Reibman A. and Trivedi K. S. A. Transient Analysis of Cumulative Measures of Markov Model Behavior. Communications in Statistics-Stochastic Models, 1989, vol. 5, pp. 683–710.

[3]. Schultz M., Gibson G., Katz R., and Patterson D. How Reliable is a RAID? Proceedings of CompCon, 1989, pp. 118–123.

[4]. Malhotra M. and Trivedi K. S. Reliability Analysis of Redundant Arrays of Inexpensive Disks. Journal of Parallel and Distributed Computing — Special issue on parallel I/O systems, 1993, vol.17, – no. 1-2., pp. 146-151.

[5]. Greenan K. M., Plank J. S. and Wylie J. J. Mean Time To Meaningless: MTTDL, Markov models, and Storage System Reliability, Proceedings of the 2nd USENIX conference on Hot topics in storage and file systems, 2010, pp. 1–5.

[6]. Karmakar P. and Gopinath K. Are Markov Models Effective for Storage Reliability Modelling? arXiv:1503.07931v1, 2015.

[7]. Li Y., Lee P. P. and Lui J. Stochastic analysis on raid reliability for solid-state drives, IEEE 32nd International Symposium on Reliable Distributed Systems (SRDS). IEEE, 2013, pp. 71–80. (http://arxiv.org/pdf/1304.1863.pdf)

[8]. Mann S. E., Anderson M. and Rychlik M. On the Reliability of RAID Systems: An Argument for More Check Drives, arXiv:1202.4423v1, 2012.

[9]. Pâris J.-F., Schwarz T., Amer A. and Long D. D. E. Improving Disk Array Reliability Through Expedited Scrubbing, Proceedings of the 5th IEEE International Conference on Networking, Architecture, and Storage, 2010, pp. 119–125.

[10]. Xin Q., Miller E. L., Schwarz T. J., Long D. D. E., Brandt S. A. and Litwin W. .Reliability mechanisms for very large storage systems, Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies, 2003, pp. 146–156.

[11]. Elerath J.G. and Pecht M. Enhanced Reliability Modeling of RAID Storage Systems, 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2007, pp. 175–184.

[12]. Ivanichkina L. and Neporada A. Mathematical methods and models of improving data storage reliability including those based on finite field theory, Contemporary Engineering Sciences, 2014, vol. 7, no. 28, 1589-1602 http://dx.doi.org/10.12988/ces.2014.411236.

The Reliability Model of a Distributed Data Storage in Case of Explicit and Latent Disk Faults

[1]L. Ivanichkina <ivanov@ispras.ru>
[2]A. Neporada <Andrew.Neporada@acronis.com>
[1]OOO Proekt IKS, Altufievskoe sh, 44, Moscow, 127566, Russian Federation
[2]OOO Acronis, Altufievskoe sh, 44, Moscow, 127566, Russian Federation

Abstract. This work examines the approach to the estimation of the data storage reliability that accounts for both explicit disk faults and latent bit errors as well as procedures to detect them. A new analytical math model of the failure and recovery events in the distributed data storage is proposed to calculate reliability. The model describes dynamics of the data loss and recovery based on Markov chains corresponding to the different schemes of redundant encoding. Advantages of the developed model as compared to classical models for traditional RAIDs are covered. Influence of latent HDD errors is considered, while other bit faults occurring in the other hardware components of the machine are omitted. Reliability is estimated according to new analytical formulas for calculation of the mean time to failure, at which data loss exceeds the recoverability threshold defined by the redundant encoding parameters. New analytical dependencies between the storage average lifetime until the data loss and the mean time for complete verification of the storage data are given.

Keywords: Mean time to failure (MTTF), Markov chains, redundant encoding, Huygens' gambler's ruin problem, distributed data storage, scrubbing procedure, checksums, MTTDL of distributed data storage, disk faults, irrecoverable bit errors, latent sector errors.

DOI: 10.15514/ISPRAS-2015-27(6)-16

For citation: Ivanichkina L., Neporada A. The Reliability Model of a Distributed Data Storage in Case of Explicit and Latent Disk Faults. Trudy ISP RAN/Proc. ISP RAS, vol. 27, issue 6, 2015, pp. 253-274 (in Russian). DOI: 10.15514/ISPRAS-2015-27(6)-16

References

[1]. Patterson D. A., Gibson G., and Katz R. H. A Case for Redundant Arrays of Inexpensive Disks (RAID), Proc. of ACM SIGMOD, 1988.
[2]. Reibman A. and Trivedi K. S. Transient Analysis of Cumulative Measures of Markov Model Behavior. Communications in Statistics-Stochastic Models, 1989, vol. 5, pp. 683–710.
[3]. Schultz M., Gibson G., Katz R., and Patterson D. How Reliable is a RAID? Proceedings of CompCon, 1989, pp. 118–123.

[4]. Malhotra M. and Trivedi K. S. Reliability Analysis of Redundant Arrays of Inexpensive Disks. Journal of Parallel and Distributed Computing — Special issue on parallel I/O systems, 1993, vol.17, – no. 1-2., pp. 146-151.

[5]. Greenan K. M., Plank J. S. and Wylie J. J. Mean Time To Meaningless: MTTDL, Markov models, and Storage System Reliability, Proceedings of the 2nd USENIX conference on Hot topics in storage and file systems, 2010, pp. 1–5.

[6]. Karmakar P. and Gopinath K. Are Markov Models Effective for Storage Reliability Modelling? arXiv:1503.07931v1, 2015.

[7]. Li Y., Lee P. P. and Lui J. Stochastic analysis on raid reliability for solid-state drives, IEEE 32nd International Symposium on Reliable Distributed Systems (SRDS). IEEE, 2013, pp. 71–80. (http://arxiv.org/pdf/1304.1863.pdf)

[8]. Mann S. E., Anderson M. and Rychlik M. On the Reliability of RAID Systems:An Argument for More Check Drives, arXiv:1202.4423v1, 2012.

[9]. Pâris J.-F., Schwarz T., Amer A. and Long D. D. E. Improving Disk Array Reliability Through Expedited Scrubbing, Proceedings of the 5th IEEE International Conference on Networking, Architecture, and Storage, 2010, pp. 119–125.

[10]. Xin Q., Miller E. L., Schwarz T. J., Long D. D. E., Brandt S. A. and Litwin W. .Reliability mechanisms for very large storage systems, Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies, 2003, pp. 146–156.

[11]. Elerath J.G. and Pecht M. Enhanced Reliability Modeling of RAID Storage Systems, 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2007, pp. 175–184.

[12]. Ivanichkina L. and Neporada A. Mathematical methods and models of improving data storage reliability including those based on finite field theory, Contemporary Engineering Sciences, 2014, vol. 7, no. 28, 1589-1602 http://dx.doi.org/10.12988/ces.2014.411236.