

Гончаренко В. А., Хомоненко А. Д., Абу Хасан Р.
V. A. Goncharenko, A. D. Khomonenko, R. Abu Khasan

ГЕНЕТИЧЕСКИЙ АЛГОРИТМ ВЫБОРА ОПТИМАЛЬНОЙ КОНФИГУРАЦИИ RAID-МАССИВОВ СИСТЕМЫ ХРАНЕНИЯ ДАННЫХ ПРЕДПРИЯТИЯ

GENETIC ALGORITHM FOR SELECTING THE RAID ARRAY OPTIMAL CONFIGURATION OF THE ENTERPRISE DATA STORAGE SYSTEM

Гончаренко Владимир Анатольевич – кандидат технических наук, доцент кафедры «Информационные и вычислительные системы» Петербургского государственного университета путей сообщения Императора Александра I, преподаватель кафедры информационно-вычислительных систем и сетей Военно-космической академии им. А. Ф. Можайского (Россия, Санкт-Петербург). E-mail: vlango@mail.ru.

Vladimir A. Goncharenko – PhD in Engineering, Associate Professor, Information and Computing Systems Department, St. Petersburg State Transport University named after Emperor Alexander I, Lecturer of Information and Computing Systems and Networks Department, Mozhaisky Military Space Academy (Russia, St. Petersburg). E-mail: vlango@mail.ru.

Хомоненко Анатолий Дмитриевич – доктор технических наук, профессор кафедры «Информационные и вычислительные системы» Петербургского государственного университета путей сообщения Императора Александра I, профессор кафедры математического и программного обеспечения Военно-космической академии им. А. Ф. Можайского (Россия, Санкт-Петербург). E-mail: khomon@mail.ru.

Anatoly D. Khomonenko – Doctor of Technical Sciences, Professor, Information and Computing Systems Department, St. Petersburg State Transport University named after Emperor Alexander I, Professor of Mathematical and Software Engineering Department, Mozhaisky Military Space Academy (Russia, St. Petersburg). E-mail: khomon@mail.ru.

Абу Хасан Рахед – магистр, аспирант кафедры «Информационные и вычислительные системы» Петербургского государственного университета путей сообщения Императора Александра I (Россия, Санкт-Петербург). E-mail: ragheb1997@yandex.ru.

Ragheb Abu Khasan – Master, Postgraduate Student, Information and Computing Systems Department, St. Petersburg State Transport University named after Emperor Alexander I (Russia, St. Petersburg). E-mail: ragheb1997@yandex.ru.

Аннотация. Сформулирована постановка задачи оптимизации выбора конфигурации RAID-массивов из нескольких уровней в качестве системы хранения данных предприятия по критерию надёжности и производительности. Разработан генетический алгоритм её решения. Охарактеризованы основные шаги алгоритма: инициализация популяции, оценка фитнес-функции, отбор лучших особей для дальнейшего размножения путём турнирного или рулеточного отбора, применение кроссовера к отобранным особям для создания новых особей-потомков, применение мутации к новым особям для изменения уровня RAID или количества дисков, замена худших особей текущей популяции новыми потомками, повторение шагов до тех пор, пока не будет достигнуто максимальное число итераций или оптимальное решение. Приведены примеры расчёта оптимальных вариантов RAID-массивов из нескольких уровней. Выполнены расчёты фитнес-функций по генетическому алгоритму с учётом стохастической параметрической неопределённости характеристик надёжности и производительности RAID-массивов. В задаче выбора RAID-массива для системы хранения данных неопределённость касается характеристик надёжности и производительности различных уровней RAID. В рамках генетического алгоритма это улучшает предсказуемость производительности и надёжности системы, особенно в ситуациях, когда параметры надёжности и производительности подвержены значительным колебаниям.

Summary. The problem of optimizing the choice of a RAID array configuration from several levels as an enterprise data storage system according to the criteria of reliability and performance is formulated. A genetic algorithm for its solution has been developed. The main steps of the algorithm are described: initialization of the population, assessment of the fitness function, selection of the best individuals for further reproduction by tournament or roulette selection, application of crossover to selected individuals to create new offspring, application of mutation to

new individuals to change the RAID level or the number of disks, replacement of the worst individuals of the current population with new descendants, repetition of steps until the maximum number of iterations or the optimal solution is reached. Examples of calculating optimal RAID array options from several levels are given. Calculations of fitness functions using a genetic algorithm have been performed, taking into account the stochastic parametric uncertainty of the reliability and performance characteristics of RAID arrays. In the task of choosing a RAID array for a data storage system, the uncertainty concerns the reliability and performance characteristics of various RAID levels. Within the framework of the genetic algorithm, this improves the predictability of system performance and reliability, especially in situations where reliability and performance parameters are subject to significant fluctuations.

Ключевые слова: RAID-массивы, генетический алгоритм, надёжность, производительность, стохастическая параметрическая неопределённость.

Key words: RAID arrays, genetic algorithm, reliability, performance, stochastic parametric uncertainty.

УДК 004.75

Введение. В условиях стремительного роста объёмов данных и требований к их обработке высокое качество хранения информации становится критически важным для организаций, занимающихся анализом данных и управлением данными [1–3]. Крупнейшие российские и зарубежные компании, такие как Сбербанк, «Газпром», «Яндекс», «Ростелеком», Google, Amazon Web Services, Facebook и др., для эффективного и надёжного хранения данных используют технологии RAID (Redundant Array of Independent Disks – избыточный массив независимых дисков). Это подчёркивает важность выбора оптимальной конфигурации RAID-массива для повышения отказоустойчивости и увеличения скорости доступа к данным [4; 5].

Выбор подходящей комбинации уровней RAID-массива системы хранения данных (СХД) представляет собой сложную задачу, требующую учёта множества факторов, таких как надёжность, производительность, стоимость и объём хранения [6; 7].

Вопросам оценки уровня надёжности и выбора наилучшего уровня RAID для хранилищ данных посвящён ряд работ [8; 9]. При этом в задачах оценивания и выбора решений возникают проблемы учёта стохастической или нечёткой неопределённости исходных данных [9; 10]. Необходимость учёта параметрической неопределённости в характеристиках надёжности и производительности различных уровней RAID связана с эксплуатационными условиями, качеством оборудования и другими факторами, что добавляет дополнительный уровень сложности к процессу принятия решения [11; 12]. В [13] предложен композиционный подход к моделированию систем с неопределённостью исходных данных, когда параметры представляются в виде случайных величин.

В условиях неопределённости даже небольшие изменения в параметрах могут существенно влиять на общую эффективность системы хранения данных. В связи с этим возникает необходимость разработки методов, способных учитывать стохастические аспекты этих параметров и находить оптимальные решения даже в условиях неопределённости. Одним из наиболее перспективных подходов к решению подобных задач являются генетические алгоритмы.

В [14] реализован метод представления битовых массивов для оптимизации структурной топологии с использованием генетических алгоритмов. В [15] представлен многоцелевой генетический алгоритм для оптимизации расположения данных при декластеринге чётности в сетевом RAID, в том числе с учётом распределённой рабочей нагрузки восстановления и распределённой чётности, что может обеспечить лучшую компоновку данных для сетевого хранилища.

В статье рассматривается вопрос выбора оптимальной комбинации RAID-массивов в СХД на основе генетических алгоритмов с учётом многоцелевого хранения информации с различными требованиями по надёжности и производительности, а также с учётом возможной параметрической неопределённости исходных данных.

Постановка задачи. Пусть для построения системы хранения данных предприятия имеется K типов (уровней) RAID-массивов (например, RAID 0, RAID 1, RAID 10, RAID 5, RAID 6) и N дисков, которые можно распределить между различными уровнями RAID. Необходимо выбрать оптимальную комбинацию уровней RAID и количества дисков для каждого уровня, чтобы максимизировать целевую функцию, учитывающую показатели надёжности и производительности, при соблюдении бюджетных и ресурсных ограничений.

Переменные:

- n_k – количество дисков, выделяемых под уровень RAID k , где $k = 0, 1, 2, \dots, K$.
- c_k – стоимость одного диска уровня RAID k .
- r_k – показатель надёжности уровня RAID k .
- p_k – показатель производительности уровня RAID k .
- B – общий бюджет.
- N – общее количество дисков СХД.

Целевая функция представляет собой взвешенную сумму нормализованных показателей надёжности, производительности и эффективности использования дисков:

$$F(n) = w_p \cdot P^n(n) + w_v \cdot V^n(n) + w_e \cdot E^n(n),$$

где w_p , w_v и w_e – весовые коэффициенты для показателей надёжности, производительности и эффективности использования СХД соответственно, отражающие относительную важность того или иного показателя для СХД, причём $w_p + w_v + w_e = 1$; $P^n(x)$, $V^n(x)$, $E^n(x)$ – нормализованные общие показатели надёжности, производительности и эффективности использования СХД соответственно.

Поскольку показатели надёжности, производительности и эффективности использования дискового пространства имеют разные размерность и физический смысл, их нужно нормализовать, т. е. привести к единому масштабу и уравновесить их влияние на целевую функцию. Нормализованные значения показателей надёжности, производительности и эффективности для уровней RAID обычно получают методом мини-макс нормализации:

$$P^H_{\text{RAID}k}(n_k) = (P(n_k) - P_{\min}) / (P_{\max} - P_{\min});$$

$$V^H_{\text{RAID}k}(n_k) = (V(n_k) - V_{\min}) / (V_{\max} - V_{\min});$$

$$E^H_{\text{RAID}k}(n_k) = (E(n_k) - E_{\min}) / (E_{\max} - E_{\min}).$$

Примем, что максимальное значение показателя надёжности RAID-массива $P_{\max} = 1$, минимальное значение – $P_{\min} = P_{\text{RAID}0}(n)$, поскольку именно в RAID0 отсутствует резервирование, максимальное значение показателя производительности $V_{\max} = V_d \cdot n$ также достигается в RAID0 за счёт технологии чередования, минимальное значение $V_{\min} = V_d$, т. е. равно скорости чтения-записи одиночного диска, максимальное значение показателя эффективности использования дисков $E_{\max} = 1$, или 100 %, которое достигается в RAID0, минимальное значение $P_{\min} = 0.5$.

Общий показатель надёжности СХД рассчитывается как произведение показателей надёжности отдельных уровней RAID:

$$P(n) = \frac{\sum_{k=1}^K P_k(n_k) \cdot n_k}{\sum_{k=1}^K n_k}.$$

Общий показатель производительности системы рассчитывается как средневзвешенный показатель производительности всех уровней RAID:

$$V(n) = \frac{\sum_{k=1}^K V_k(n_k) \cdot n_k}{\sum_{k=1}^K n_k}.$$

Общий показатель эффективности использования дискового пространства системы рассчитывается как средневзвешенный показатель эффективности всех уровней RAID:

$$E(n) = \frac{\sum_{k=1}^K E_k(n_k) \cdot n_k}{\sum_{k=1}^K n_k}.$$

Ограничения:

1. Бюджетное ограничение:

$$\sum_{k=1}^K c_k \cdot x_k \leq B.$$

2. Ресурсное ограничение:

$$\sum_{k=1}^K x_k = N.$$

Рассмотрим основные параметры различных уровней RAID-массивов и требования к их организации.

1. Требования по количеству дисков и применение RAID:

- RAID 0 (чередование). Требуется минимум 2 диска, но не предоставляет избыточности данных. Отказ одного диска ведёт к потере всей информации. *Применение:* для быстрого доступа к временным или некритичным данным, когда требуется высокая производительность и можно восстановить информацию.

- RAID 1 (зеркалирование). Требуется минимум 2 диска, данные дублируются на каждом диске, количество дисков должно быть чётным. RAID-массив остаётся работоспособным, пока работает хотя бы один диск. *Применение:* для максимального повышения надёжности хранения и доступности данных.

- RAID 10 (RAID 1 + RAID 0). Требуется минимум 4 диска, количество дисков должно быть чётным. *Применение:* для работы с корпоративными базами данных, веб-серверами и другими высокотранзакционными нагрузками.

- RAID 5. Требуется минимум 3 диска, производится распределение контрольных сумм по всем дискам массива, позволяет восстанавливать данные при отказе одного диска. *Применение:* для некритичной информации с небольшой нагрузкой на диски, например, в системах видеонаблюдения.

- RAID 6. Требуется минимум 4 диска. Аналогичен RAID 5, но допускает отказ двух дисков. *Применение:* для надёжного хранения критически важных данных, где не важна высокая транзакционная производительность (системы безопасности, архивные системы, системы резервного копирования).

2. Критерии оценки фитнес-функции:

- показатель надёжности – вероятность безотказной работы RAID-массива – P_{RAIDi} (для различных уровней RAID будет отличаться);

- показатель производительности – средняя скорость чтения/записи, IOPS (Input/Output Operations Per Second) – $V_{RAIDi} = (V_{r_i} + V_{w_i})/2$;

- показатель эффективности использования дискового пространства – $E_{RAIDi} = n_e/n$, где n_e – эффективное количество дисков, n – общее количество дисков под RAID-массив.

3. Вероятность безотказной работы уровня RAID:

- RAID 0: $P_{RAID0}(n) = (1 - q)^n$, где q – вероятность отказа одного диска за определённый период (обычно за год), n – количество дисков массива.

- RAID 1: $P_{RAID1}(n) = (1 - q^2)^{n/2}$. Для 2 дисков: $P_{RAID1}(2) = 1 - q^2$.

- RAID 10: $P_{RAID10}(n) = (1 - q^2)^{n/2}$. Для 4 дисков: $P_{RAID10}(4) = (1 - q^2)^2$.

- RAID 5: $P_{RAID5}(n) = (1 - q)^{n-1}(1 + (n - 1)q)$. Для 3 дисков: $P_{RAID5}(3) = (1 - q)^2(1 + 2q)$.

– RAID 6: $P_{\text{RAID6}}(n) = (1 - q)^{n-2}(1 + (n - 2)q + (1 - 1,5n + 0,5n^2)q^2)$. Для 4 дисков: $P_{\text{RAID6}}(4) = (1 - q)^2(1 + 2q + 3q^2)$.

4. Производительность (для различных уровней RAID будет отличаться):

– RAID 0: $V_{\text{RAID0}}(n) = V_d \cdot n$, где V_d – скорость ввода-вывода одиночного диска, n – количество дисков RAID0. Для 2 дисков: $V_{\text{RAID0}}(2) = 2V_d$. Для 3 дисков: $V_{\text{RAID0}}(3) = 3V_d$. Для 4 дисков: $V_{\text{RAID0}}(4) = 4V_d$.

– RAID 1: $V_{\text{RAID1}}(n) = (2V_d + V_d/2)/2$. Для 2 дисков: $V_{\text{RAID1}}(2) = 1.25V_d$.

– RAID 10: $V_{\text{RAID10}}(n) = (V_d \cdot n + V_d \cdot n/2)/2$. Для 4 дисков: $V_{\text{RAID10}}(4) = 3V_d$. Для 6 дисков: $V_{\text{RAID10}}(6) = 4.5V_d$.

– RAID 5: $V_{\text{RAID5}}(n) = (V_d \cdot n + V_d \cdot n/4)/2$. Для 3 дисков: $V_{\text{RAID5}}(3) = 1.875V_d$. Для 4 дисков: $V_{\text{RAID5}}(4) = 2.5V_d$. Для 5 дисков: $V_{\text{RAID5}}(5) = 3.125V_d$. Для 6 дисков: $V_{\text{RAID5}}(6) = 3.75V_d$.

– RAID 6: $V_{\text{RAID6}}(n) = (V_d \cdot n + V_d \cdot n/6)/2$. Для 4 дисков: $V_{\text{RAID6}}(4) = 2.33V_d$. Для 5 дисков: $V_{\text{RAID6}}(5) = 2.917V_d$. Для 6 дисков: $V_{\text{RAID6}}(6) = 3.5V_d$.

5. Эффективность:

– RAID 0: $E_{\text{RAID0}}(n) = n_e/n = 1$, где n_e – эффективное количество дисков, n – общее количество дисков под RAID-массив.

– RAID 1: $E_{\text{RAID1}}(n) = 0.5$.

– RAID 10: $E_{\text{RAID10}}(n) = 0.5$.

– RAID 5: $E_{\text{RAID5}}(n) = (n - 1)/n$. Для 3 дисков: $E_{\text{RAID5}}(3) = 0.67$. Для 4 дисков: $E_{\text{RAID5}} = 0.75$. Для 5 дисков: $E_{\text{RAID5}} = 0.8$.

– RAID 6: $E_{\text{RAID6}}(n) = (n - 2)/n$. Для 4 дисков: $E_{\text{RAID6}}(4) = 0.5$. Для 5 дисков: $E_{\text{RAID6}} = 0.6$. Для 6 дисков: $E_{\text{RAID6}} = 0.67$.

Характеристика и математическая модель генетического алгоритма. Генетические алгоритмы, основанные на эволюционных подходах, представляют собой мощный инструмент для решения задач оптимизации, позволяя находить приемлемые решения в сложных и многомерных пространствах поиска. В статье представлен генетический алгоритм, который нацелен на выбор оптимального варианта RAID-массива с учётом надёжности, производительности, эффективности использования дисков, а также стохастической природы входящих данных.

Основные шаги алгоритма включают инициализацию популяции возможных решений, оценку их приспособленности по заданным критериям, отбор лучших особей, генетические операции (кроссовер и мутацию) и повторение этих шагов до достижения заданного критерия остановки. Особое внимание уделяется модификации стандартного алгоритма для учёта стохастической неопределённости, что позволяет повысить его предсказуемость и надёжность.

Рассмотрим задачу оптимизации функции $f(x)$, где x – вектор переменных, который нужно оптимизировать. Задача заключается в поиске вектора x , который минимизирует или максимизирует функцию $f(x)$ в зависимости от постановки задачи.

Элементы генетического алгоритма:

1. **Хромосомы (индивид)** есть представление потенциального решения задачи. В нашем случае хромосома представляет собой вектор x , состоящий из n элементов (генов): $x = (x_1, x_2, \dots, x_n)$.

2. **Популяция:** множество хромосом (индивидов), представляющих возможные решения задачи. Размер популяции обозначим через P .

3. **Фитнес-функция $F(x)$** оценивает качество каждого индивида в популяции: в случае задачи минимизации $F(x) = f(x)$, а в случае задачи максимизации $F(x) = -f(x)$. Для учёта стохастической неопределённости параметров вводится их интервальное представление в виде диапазона возможных значений.

4. **Операции над популяцией:**

– **Инициализация популяции:** на первом шаге создаётся начальная популяция размером P , состоящая из случайно сгенерированных хромосом.

– **Оценка фитнес-функции:** для каждой хромосомы в популяции вычисляется значение фитнес-функции.

– *Отбор (селекция)*: из текущей популяции выбираются лучшие индивиды для создания следующего поколения. Существует несколько методов селекции, например, турнирный отбор, пропорциональный отбор и др.

– *Скрещивание (кроссовер)*: два выбранных родителя обмениваются частями своих хромосом, создавая новое поколение потомков.

– *Мутация*: случайные изменения в некоторых генах потомков вносятся для поддержания разнообразия популяции.

– *Замещение*: новое поколение заменяет старое либо частично обновляет его.

В рамках представленной работы приводится пример расчёта оптимального варианта RAID-массива, иллюстрируя эффективность предложенного подхода на основе исходных данных. Результаты исследования могут быть полезны для современных компаний, стремящихся к оптимизации своих систем хранения данных в условиях растущих требований к надёжности и производительности.

Описание генетического алгоритма расчёта конфигурации RAID. Приведём описание алгоритма и пример расчёта конфигурации RAID с учётом, что типы RAID включают: RAID 0, RAID 1, RAID 10, RAID 5, RAID 6, а также с учётом стохастической параметрической неопределённости.

Рассмотрим генетический алгоритм сначала без учёта, затем с учётом стохастической параметрической неопределённости. Эта модификация предполагает, что параметры надёжности и производительности для каждого типа RAID могут варьироваться случайным образом, что требует учёта этой вариации при расчёте фитнес-функции и выборе оптимальной конфигурации.

1. **Представление хромосомы**: каждая хромосома представляет собой список целых чисел, соответствующих количеству дисков, назначаемому каждому уровню RAID. Например, хромосома [2, 2, 6, 0, 0] означает, что два диска назначаются RAID 0, два диска – RAID 1, шесть дисков – RAID 10, и ни одного диска – RAID 5 и RAID 6.

2. **Инициализация популяции**: сначала создаётся начальная популяция из случайных хромосом, каждая из которых удовлетворяет ограничениям на бюджет и количество дисков.

3. **Оценка фитнес-функции (целевой функции)**: для каждой хромосомы вычисляется фитнес-функция, которая учитывает надёжность, производительность и эффективность использования.

Показатели надёжности, производительности и эффективности использования СХД считаются как средние значения для каждого типа RAID, умножаемые на количество дисков, назначенное этому типу. Решения, нарушающие бюджетные или ресурсные ограничения, могут получать штрафы в фитнес-функции, но при этом учитываются стохастические флуктуации этих параметров.

4. **Отбор**: используется турнирный отбор, где из трёх случайно выбранных хромосом остаётся та, у которой наибольшее значение фитнес-функции.

5. **Кроссовер (скрещивание)**: при скрещивании двух хромосом производится обмен частями хромосомы, что создаёт новую хромосому-потомка.

6. **Мутация**: с небольшой вероятностью одна из хромосом подвергается мутации, что приводит к изменению количества дисков, назначенного какому-то типу RAID.

7. **Замещение (создание новой популяции)**: новое поколение хромосом замещает текущее (потомки занимают места своих родителей).

8. **Повторение и останов**: процесс повторяется, пока не будет достигнуто условие остановки, например, достижение определённого числа поколений или стабилизация фитнес-функции. Популяция не производит потомство, которое значительно отличается от предыдущего поколения.

Пример расчёта. Предположим, что у нас есть *следующие параметры*:

Бюджет: 30 000 рублей.

Стоимость одного диска: 1600 рублей.

Максимальное количество дисков: 16.

$16 \cdot 1600 = 25\,600 \leq 30\,000$.

В случае недостаточности бюджета либо уменьшаем количество дисков, либо ищем другого поставщика, либо корректируем бюджет.

Вероятность отказа одного диска в течение года – $q = 0.05$.

Количество операций ввода/вывода в секунду (IOPS) – $V_d = 100$.

Коэффициенты важности: $w_p = 0.5$; $w_v = 0.4$; $w_e = 0.1$.

Характеристики RAID (минимальное количество):

- RAID 0 (2 диска): надёжность 0.9025, производительность 200, эффективность 100 %.
- RAID 1 (2 диска): надёжность 0.9975, производительность 125, эффективность 50 %.
- RAID 10 (4 диска): надёжность 0.995, производительность 300, эффективность 50 %.
- RAID 5 (3 диска): надёжность 0.9928, производительность 187.5, эффективность 67 %.
- RAID 6 (4 диска): надёжность 0.9995, производительность 233, эффективность 50 %.

Количество дисков для определённого уровня RAID также может быть равно нулю, если это поможет улучшить фитнес-функцию за счёт других уровней RAID.

Шаг 1: Инициализация популяции. Случайная инициализация хромосом:

- хромосома 1: [4, 0, 10, 0, 0];
- хромосома 2: [0, 2, 0, 3, 5];
- хромосома 3: [3, 0, 6, 4, 0];
- хромосома 4: [2, 2, 0, 5, 6];
- хромосома 5: [0, 0, 8, 0, 7];
- ...

Шаг 2: Оценка фитнес-функции. Рассчитаем фитнес-функцию для первой хромосомы [4, 0, 10, 0, 0]:

- RAID 0 (4 диска): $P_{RAID0}(n) = (1 - q)^n = 0.815$; $V_{RAID0}(n) = V_d \cdot n = 400$; $E_{RAID0}(n) = n_e/n = 1$.

Нормализованные значения: $P^H_{RAID0}(4) = 0$; $V^H_{RAID0}(4) = 1$; $E^H_{RAID0}(4) = 1$;

- RAID 1 (0 дисков): $P^H_{RAID1}(0) = 0$; $V^H_{RAID1}(0) = 0$; $E^H_{RAID1}(0) = 0$;

– RAID 10 (10 дисков): $P_{RAID10}(10) = (1 - q^2)^5 = 0.988$; $V_{RAID10}(10) = (V_d \cdot 10 + V_d \cdot 10/2)/2 = 750$; $E_{RAID10}(n) = 0.5$.

Нормализованные значения: $P^H_{RAID10}(10) = 0.969$; $V^H_{RAID10}(10) = 0.722$; $E^H_{RAID10}(10) = 0.5$;

- RAID 5 (0 дисков): $P^H_{RAID5}(0) = 0$; $V^H_{RAID5}(0) = 0$; $E^H_{RAID5}(0) = 0$;

- RAID 6 (0 дисков): $P^H_{RAID6}(0) = 0$; $V^H_{RAID6}(0) = 0$; $E^H_{RAID6}(0) = 0$.

Суммарная надёжность: $(0 \times 4 + 0.988 \times 10)/16 = 0.6056$.

Суммарная производительность: $(1 \times 4 + 0.722 \times 10)/16 = 0.7014$.

Суммарная эффективность использования: $(1 \times 4 + 0.5 \times 10)/16 = 0.5625$.

Фитнес-функция: $F_1(n) = 0.5 \times 0.6056 + 0.4 \times 0.701 + 0.1 \times 0.563 = \mathbf{0.6396}$.

Аналогично находим фитнес-функции 2-5-й хромосом:

$F_2(n) = 0.5 \times 0.611 + 0.4 \times 0.263 + 0.1 \times 0.4375 = \mathbf{0.4543}$.

$F_3(n) = 0.5 \times 0.5955 + 0.4 \times 0.575 + 0.1 \times 0.5625 = \mathbf{0.584}$.

$F_4(n) = 0.5 \times 0.7749 + 0.4 \times 0.5098 + 0.1 \times 0.75 = \mathbf{0.6664}$.

$F_5(n) = 0.5 \times 0.9173 + 0.4 \times 0.582 + 0.1 \times 0.625 = \mathbf{0.7539}$.

Шаг 3: Отбор. Проведём турнирный отбор среди первых пяти хромосом:

- Хромосома 1: [4, 0, 10, 0, 0] – **0.6396**.
- Хромосома 2: [0, 2, 0, 3, 5] – **0.4543**.
- Хромосома 3: [3, 0, 6, 4, 0] – **0.584**.
- Хромосома 4: [2, 2, 0, 5, 6] – **0.66646**.
- Хромосома 5: [0, 0, 8, 0, 7] – **0.7539**.

Победителем тура станет хромосома 5.

Шаг 4: Кроссовер. Проведём кроссовер между хромосомой 5 и другой случайно выбранной хромосомой, например, хромосомой 4:

- хромосома 4: [2, 2, 0, 5, 6];
- хромосома 5: [0, 0, 8, 0, 7].

После одноточечного кроссовера в точке 5 получаем хромосомы-потомки:

- хромосома 4: [2, 2, 0, 5, 7];
- хромосома 5: [0, 0, 8, 0, 6].

Шаг 5: Мутация. С вероятностью 0.05 произойдёт мутация в пятой хромосоме, например, во втором типе RAID:

- Мутация: изменяется количество дисков для RAID 1 – с 0 на 2.
- Новая хромосома: [0, 2, 8, 0, 6].

Шаг 6: Замещение. Новое поколение хромосом замещает старое, и процесс продолжается до достижения максимального числа поколений или стабилизации фитнес-функции.

Итоговый результат. По завершении работы алгоритма может быть найдена оптимальная конфигурация RAID:

- **Итоговая хромосома:** [0, 0, 10, 0, 6] с фитнес-функцией **0.8068**. У остальных хромосом значительно меньше.
- **Конфигурация RAID:** 10 дисков для RAID 10, 6 дисков для RAID 6, 0 дисков для RAID 0, RAID 1 и RAID 5.

При выборе других коэффициентов важности победителем в итоге может оказаться другая хромосома с большей по значению фитнес-функцией. Кроме того, на выбор может также повлиять учёт неопределённости исходных данных, таких как вероятность отказа или скорость чтения/записи.

Учёт параметрической неопределённости. При наличии возможных колебаний или неточном задании параметров, таких как вероятность отказа одного диска и скорости чтения-записи диска, изменение параметров может значительно повлиять на финальный результат вычисления значений фитнес-функций и последующий выбор. Допустим, неопределённость параметра q задаётся интервальным диапазоном: $q \in [0.03, 0.07]$ либо в процентах отклонения – $\pm 40\%$. Неопределённость параметра V_d задаётся интервальным диапазоном: $V_d \in [80, 120]$ либо в процентах отклонения – $\pm 20\%$.

Рассмотрим, как неопределённость может повлиять на оценки и, соответственно, на выбор конфигураций.

Шаг 1. Определение интервалов неопределённости.

Для вероятности отказа одного диска: $0.03 \leq q \leq 0.07$.

Для скорости чтения-записи: $80 \leq V_d \leq 120$.

Это означает, что при вычислении фитнес-функций для каждой хромосомы необходимо учитывать оценки в границах этих интервалов. Мы можем сделать несколько расчётов с разными значениями q и V_d в пределах заданных границ.

Шаг 2. Пересчёт фитнес-функции.

Предположим, что для каждого параметра у нас имеются значения, соответствующие различным сценариям. Например, мы можем рассмотреть следующие значения:

- для q : 0.03, 0.05, 0.07;
- для V_d : 80, 100, 120.

Используя различные комбинации этих значений, мы можем пересчитать фитнес-функцию для каждой хромосомы при различных комбинациях параметров, а затем проанализировать, как это повлияет на финальный выбор конфигурации RAID.

Для начальной хромосомы 5: [0, 0, 8, 0, 7]:

При $q = 0.03$ и $V_d = 80$:

суммарная надёжность $P(n) = (0.9834 \times 8 + 0.9804 \times 7)/16 = 0.9206$,

суммарная производительность $(0.9286 \times 8 + 0.6840 \times 7)/16 = 0.7635$,

суммарная эффективность использования $(0.5 \times 8 + 0.8571 \times 7)/16 = 0.625$,

фитнес-функция $F_5(n) = 0.5 \times 0.9206 + 0.4 \times 0.7635 + 0.1 \times 0.625 = \mathbf{0.8282}$.

При $q = 0.05$ и $V_d = 100$:

суммарная надёжность $P(n) = (0.9704 \times 8 + 0.9875 \times 7)/16 = 0.9173$,

суммарная производительность $(0.7143 \times 8 + 0.5139 \times 7)/16 = 0.5820$,

суммарная эффективность использования $(0.5 \times 8 + 0.8571 \times 7)/16 = 0.625$,

фитнес-функция $F_5(n) = 0.5 \times 0.9173 + 0.4 \times 0.5820 + 0.1 \times 0.625 = \mathbf{0.7539}$.

При $q = 0.07$ и $V_d = 120$:

суммарная надёжность $P(n) = (0.9558 \times 8 + 0.9906 \times 7)/16 = 0.9113$,

суммарная производительность $(0.5714 \times 8 + 0.4005 \times 7)/16 = 0.4609$

суммарная эффективность использования $(0.5 \times 8 + 0.8571 \times 7)/16 = 0.625$,

фитнес-функция $F_5(n) = 0.5 \times 0.9113 + 0.4 \times 0.4609 + 0.1 \times 0.625 = \mathbf{0.7025}$.

Шаг 3. Обработка результатов.

Анализ результатов показывает, что возможен существенный разброс значений фитнес-функции. Кроме того, симметричное отклонение значений параметров от среднего ($q = 0,05 \pm 0,02$, $V_d = 100 \pm 20$) тем не менее приводит к несимметричному отклонению фитнес-функций от исходного невозмущённого значения:

($q = 0.07$) и ($V_d = 120$),

($q = 0.05$) и ($V_d = 100$),

($q = 0.03$) и ($V_d = 80$),

$F_5^-(n) = 0.7025$,

$F_5(n) = 0.7539$,

$F_5^+(n) = 0.8282$,

отклонение от $F_5(n)$: $\Delta^- = 0.0743$,

$(F_5^-(n) + F_5^+(n))/2 = 0.7654$,

отклонение от $F_5(n)$: $\Delta^+ = 0.0514$.

При тех же неопределённых параметрах фитнес-функция другой хромосомы [2, 2, 0, 5, 6] будет показывать следующие значения:

($q = 0.07$) и ($V_d = 120$),

($q = 0.05$) и ($V_d = 100$),

($q = 0.03$) и ($V_d = 80$),

$F_4^-(n) = 0.6259$,

$F_4(n) = 0.6664$,

$F_4^+(n) = 0.7230$,

отклонение от $F_5(n)$: $\Delta^- = 0.0405$,

$(F_5^-(n) + F_5^+(n))/2 = 0.7654$,

отклонение от $F_5(n)$: $\Delta^+ = 0.0566$.

Правая граница диапазона $F_4^+(n)$ перекрывается левой границей диапазона $F_5^-(n)$, а это значит, что в процессе работы генетического алгоритма может быть сделан выбор в пользу другой хромосомы, что приведёт к выбору несколько другой конфигурации СХД.

Шаг 4. Учёт влияния неопределённости на выбор конфигурации.

Получаемые фитнес-функции будут меняться с изменением параметров q и V_d , а также в случае уточнения коэффициентов важности в процессе эксплуатации и пополнения статистики. Например, увеличение вероятности отказа диска, а также рост скорости чтения/записи могут положительно сказаться на одной конфигурации и негативно на другой. В результате конфигурация, которая изначально выглядела менее предпочтительной, может оказаться более устойчивой с учётом этих неопределённостей.

Учёт неопределённости позволяет получать решения, более устойчивые к изменениям параметров, а не рассчитывать на их статические значения, которые могут измениться из-за различных обстоятельств. Например, если в условиях интенсивной эксплуатации и высокой вероятности отказов оборудования выясняется, что определённые конфигурации RAID оказываются менее надёжными, можно улучшить свой выбор, увеличив количество дисков на уровне RAID 6 вместо RAID 5, что обеспечит лучшими показателями, даже если производительность может немного падать (что может быть в рамках допустимого). Таким образом, используя диапазоны значений и учитывая неопределённость, можно получить более сбалансированное, обоснованное и рыночное решение.

Заключение. Предложенный генетический алгоритм эффективно справляется с задачей выбора оптимальной конфигурации RAID-массива, принимая во внимание стохастическую природу параметров надёжности и производительности [8–10]. Такой подход позволяет минимизировать риски, связанные с непредсказуемостью поведения систем, и обеспечить максимальную надёжность и производительность при заданных ограничениях. Приведённый пример показывает, как генетический алгоритм может использоваться для поиска оптимальной конфигурации RAID-массива с учётом ограничений на бюджет и количество дисков, а также неопределённости параметров надёжности и производительности.

Учет неопределённости параметров при расчёте фитнес-функций обеспечивает более обоснованный и надёжный выбор конфигураций. Он позволяет гибко адаптироваться к изменяющимся условиям, повышая вероятность выбора наиболее эффективного решения в конечном итоге. Стратегический подход к определению параметров и их колебаний может дать концептуальное преимущество в условиях неопределённости.

ЛИТЕРАТУРА

1. Савельев, Б. А. Моделирование каскадной системы повышения качества хранения информации / Б. А. Савельев, Г. В. Бобрышева // Труды Международного симпозиума «Надёжность и качество». – 2005. – Т. 1. – С. 205-207.
2. Cottrill C. D., Derrible S. Leveraging big data for the development of transport sustainability indicators // Journal of Urban Technology. 2015. Т. 22. № 1. Р. 45-64.
3. Иванцов, Д. С. Подход к обеспечению устойчивости и оперативности функционирования распределённых хранилищ данных о безопасности информации / Д. С. Иванцов, И. Б. Саенко // XIX Санкт-Петербургская международная конференция «Региональная информатика (РИ-2024)», Санкт-Петербург, 23-25 октября Р32 2024 г. – СПб.: СПОИСУ, 2024. – С. 108.
4. Лазарева, С. В. Системы хранения данных RAIDIX: российские инновации на мировой арене / С. В. Лазарева, А. В. Маров // Инновации. – 2016. – № 5 (211). – С. 25-30.
5. Thomasian A., Tang Y. Performance, reliability, and performability of a hybrid RAID array and a comparison with traditional RAID1 arrays // Cluster Computing. 2012. Т. 15. Р. 239-253.
6. Савин, И. В. Выбор оптимального уровня RAID для повышения надёжности хранения информации / И. В. Савин // Современные тенденции развития образования, науки и технологий: сборник научных трудов по материалам V междунар. науч.-практ. конф. / под общ. ред. А. В. Туголукова. – М.: Индивидуальный предприниматель Туголуков Александр Валерьевич, 2018. – С. 216-219.
7. Anderson E. et al. Selecting RAID Levels for Disk Arrays // FAST. 2002. Т. 2. Р. 189-201.
8. Атрощенко, В. А. К вопросу выбора наилучшего уровня RAID для хранилищ данных информационной системы, обеспечивающей быструю обработку больших данных / В. А. Атрощенко, А. И. Тымчук // Современные наукоёмкие технологии. – 2017. – № 4. – С. 12-16.
9. К оценке уровня надёжности и управлению конфигурацией RAID-массивов системы хранения данных на основе нечёткого вывода / Гончаренко В. А. [др.] // Вестник Ростовского государственного университета путей сообщения. – 2024. – № 4 (96). – С. 94-106.
10. Гончаренко, В. А. Обоснование выбора производительности вычислительных систем при неизвестных параметрах распределений рабочей нагрузки / В. А. Гончаренко, А. Н. Соколовский, А. В. Калюжный // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. – 2021. – № 4. – С. 152-160.

11. Zhao Y., Wu H., Yang C., Liu Z., Cheng Q. New Reliability Modeling Methods for Structural Systems with Hybrid Uncertainty. *Quality and Reliability Engineering International*. 2020. Vol. 36. no. 6. P. 1855-1871.
12. Гончаренко, В. А. Модели и методы анализа систем массового обслуживания с параметрической неопределённостью / В. А. Гончаренко // *Интеллектуальные технологии на транспорте*. – 2017. – № 4 (12). – С. 5-11.
13. Гончаренко, В. А. Композиционный подход к имитационному моделированию систем массового обслуживания со случайными параметрами / В. А. Гончаренко, А. Д. Хомоненко, Р. Абу Хасан // *Информатика и автоматизация*. – 2024. – Т. 23. – № 6. – С. 1577-1608.
14. Wang S. Y., Tai K. Structural topology design optimization using Genetic Algorithms with a bit-array representation // *Computer Methods in Applied Mechanics and Engineering*. Vol. 194, Issues 36–38, 23 September 2005, p. 3749-3770.
15. Liu X., Wang G., Liu J. A Double-Objective Genetic Algorithm for Parity Declustering Optimization in Networked RAID // *Algorithms and Architectures for Parallel Processing (ICA3PP 2007)*. Available at: <https://www.researchgate.net/publication/221312608> (accessed 26 January 2025).